웨이브렛 변환을 이용한 음성신호의 끝점검출

Endpoint Detection of Speech Signal Using Wavelet Transform

석 종 원*, 배 건 성*
(Jong Won Seok*, Keun Sung Bae*)

※ 본 연구는 한국과학재단의 핵심전문연구비(과제번호 : 971-0917-103-2) 지원으로 수행되었습니다.

요 약

본 논문에서는 잡음이 포함된 음성의 시작점과 끝점을 효율적으로 검출할 수 있는 알고리듬에 대하여 연구하였다. 이를위해, 웨이브렛 영역에서의 에너지 분포를 고려함으로써 잡음환경하에서도 음성을 검출할 수 있는 새로운 검출 파라미터를 제안하였다. 제안된 끝점검출 파라미터는 웨이브렛 영역에서 세 번째 coarsed 스케일의 표준편차와 가중치를 곱한 첫 번째 detailed 스케일의 표준편차의 합으로 정의하였다. 제안된 끝점검출기의 성능평가를 위해서 다양한 SNR에서 기존방식과 비교하여 시작점과 끝점의 정확도 실험을 수행하였고 HMM 음성인식시스템을 이용하여 인식실험도 수행하였다.

ABSTRACT

In this paper, we investigated the robust endpoint detection algorithm in noisy environment. A new feature parameter based on a discrete wavelet transform is proposed for word boundary detection of isolated utterances. The sum of standard deviation of wavelet coefficients in the third coarse and weighted first detailed scale is defined as a new feature parameter for endpoint detection. We then developed a new and robust endpoint detection algorithm using the feature found in the wavelet domain. For the performance evaluation, we evaluated the detection accuracy and the average recognition error rate due to endpoint detection in an HMM-based recognition system across several signal-to-noise ratios and noise conditions.

I. 서 론

음성인식, 합성 및 분석등 음성공학의 거의 모든 분야에서 음성신호의 시작점 및 끝점을 주변잡음(Environmental Noise)과 분리하여 정확하게 알아내는 일은 매우 중요하다. 특히 음성신호의 경계점 또는 끝점의 검출은 고립단어 인식 시스템 개발에는 반드시 선결되어야만 하는 과제이며 끝점 검출기의 성능은 고립단어인식 시스템의 최종 인식률에 직접적인 영향을 주게 된다. 대부분의 고립단어 인식시스템들이 음성부 경계점 정보를 이용한 DTW(Dynamic Time Warping)나 신경회로망(Neural Network), 또는 HMM-(Hidden Markov Model) 기법을 사용하고 있다는 사실이 이를 뒷받침해 준다. 그러므로, 믿을 수 있는 끝점검출 알고리듬이 존재한다면 불필요한 묵음구간을 사전에 제거하므로써 단어 인식에 소요되는 시간을 줄일 수도 있는 것이다. 또한, 끝점검출 알고리듬은 대용량 음성 데이타베이스의 효율적인 구축에도 크게 기여할 수 있다.

고립단어의 시작점과 끝점은 화자가 발성시에 만드는 artifact에 의해 구분이 어려워지며 또한 nonstationary한 주변잡음에 의해 더욱 어렵게 된다. 그리고 대부분의 응용이 실시간 구현을 목적으로 하기 때문에 이에 적합하여야 한다. 따라서 고품질의 음성부 검출기를 실현한다는 것은 쉽지 않은 일이라는 것은 널리 알려져 있는 사실이다[1,2]. 신호 대 잡음비가 충분히 큰 경우는 가장 작은 에너지 레벨을 갖는 음성신호라 할지라도 주변잡음보다는 큰 에너지 값을 가지므로 에너지 함수만 이용하여도 성능이 좋은 음성부 검출기를 쉽게 구현할 수 있다. 일반적으로 신호 대 잡음비가 30dB를 넘는 경우에는 에너지와 영교차율(zero crossing rate)을 이용하여 간단하게 음성부 검출 알고리듬을 실현할 수 있다고 알려져 있다[1-3]. 그러나 잡음이 심한 환경에서는 잡음의 일부분이 검출된 음성구간에 포함되거나 검출구간에 있어야 할 음성신호 부분이 포함되지 않는 경우가 많이 발생할 수 있다. 그러므로 정확한 끝점검출을 위해서는 주변잡음을 정확히 모델링 해야하며, 효과적인 파라미터와 결정규칙을 갖고 있어야 한다. 주변 환경은 수시로 변하므로 끝점검출에 사용되는 문턱값(threshold)은 일정한 값으로 고정시켜서는 안되고 주변잡음 환경 변화에 적응시킬 필요가 있다.

대부분의 끝점검출 알고리듬은 음성신호의 영교차율과 에너지의 조합을 바탕으로 하고 있다. 대표적인 것으로는 Rabiner/Sambur의 에너지와 영교차율을 이용한 음성부

* 경북대학교 전자 · 전기공학부
접수일자 : 1999년 3월 25일

검출 알고리듬[1], Lamel의 레벨 등화기(level equalizer)를 이용한 음성부 검출 알고리듬[2], 그리고 Teager 에너지를 이용하는 방법[4,5] 등이 있다. 그러나 에너지와 영교차율을 이용하는 알고리듬들은 신호대 잡음비가 작을 경우 신뢰성 있는 경계점을 얻기에 불충분하다. 특히 진폭이 작은 마찰음이나 파열음으로 시작하는 경우는 정확한 시작점 검출이 불가능하다. Teager 에너지는 전화음성 인식과 같이 좁은 대역폭을 갖는 경우의 음성 끝점 검출에 효과적으로 사용될 수 있다. 그러나 Teager 에너지는 잡음에 민감한 성질을 갖고 있고, 특히 wideband 잡음과 큰 에너지를 가지는 잡음에서는 사용에 주의를 하여야 한다.

본 논문에서는 최근 신호처리 분야에서 활발히 연구되고 있는 웨이브렛 변환을 이용하여 잡음이 포함된 음성의 시작점과 끝점을 효율적으로 검출할 수 있는 알고리듬에 대하여 연구하였다. 시간-주파수 영역에서의 에너지 분포를 고려함으로써 잡음환경하에서도 음성을 검출할 수 있는 새로운 검출 파라미터를 웨이브렛 변환영역에서 정의하고, 이를 이용하여 잡음이 포함된 음성을 대상으로 끝점검출을 수행하여 기존의 방식과 비교 분석하였다.

본 논문의 구성은 2장에서는 기본적인 웨이브렛 이론에 대하여 살펴보고 3장에서는 새로운 끝점검출 파라미터에 대해 소개한다. 4장에서는 전체적인 끝점검출 알고리듬을 설명하고 5장에서는 제안된 알고리듬의 성능 및 기존방식과 비교한 실험결과를 보인다. 그리고 6장에서 결론을 맺는다.

II. 웨이브렛 변환

웨이브렛 이론은 응용수학에서 처음 소개된 후 최근 컴퓨터 비젼 분야에서 연구되어 온 다중해상도 표현과 연관성이 있음이 밝혀졌으며 이산 웨이브렛 변환 이론은 이산신호의 subband 분해 방법과도 연관성이 존재한다 [6-9]. 연속 웨이브렛 변환은 다음과 같이 정의된다.

$$CWT(t,a) = \int f(t)\ \varphi^*_{a,\tau}\ dt \tag{1}$$

$$\varphi_{a,\tau} = \frac{1}{\sqrt{a}}\ \varphi\left(\frac{t-\tau}{a}\right) \tag{2}$$

식 (1)에서 $\varphi(t)$는 원형(prototype) 웨이브렛 이며 $\varphi_{a,\tau}$는 $\varphi(t)$를 이동(shift)과 확장(dilation)하여 구해진다. 식 (2)를 살펴보면 scale factor a가 증가할 수록 창함수 $\varphi\left(\frac{t-\tau}{a}\right)$의 길이가 길어짐을 알 수 있다. 따라서, 짧은 지속시간을 갖는 고주파신호에 대해서는 짧은 창함수를 사용하고 긴 지속시간을 갖는 저주파신호에 대해서는 긴 창함수를 이용하는 결과가 되어 주파수 영역에 따른 다중해상도를 갖게 된다.

음성 신호와 같이 저역과 고역의 주파수 성분을 함께 가지고 있으며, 특히 신호가 짧은 지속시간을 가지는 고주파 성분이나 긴 저주파 성분 또는 이의 합성으로 구성되어 있는 경우 앞에서 언급한 특성의 기저함수(basis function)를 갖는 변환은 고정된 시간-주파수 해상도를 갖는 short-time Fourier 변환에 비해 더욱 유효하리라는 것을 알 수 있다. 식 (1)의 연속 웨이브렛 변환은 시간과 scale factor가 연속인 값을 갖기 때문에 이를 실제 이용 가능한 이산적인 형태로 표시하면 다음과 같다.

$$d_{j,k} = \int f(t)\ \varphi^*_{j,k}(t)\ dt \tag{3}$$

$$\varphi_{j,k}(t) = a_0^{\frac{-j}{2}}\ \varphi(a_0^{-j}t - kT) \tag{4}$$

식 (4)에서 $a_0 \approx 1$이고 샘플링 주기 T가 작을 경우는 식 (2)의 근사식이 된다. 웨이브렛 변환을 이용한 계수 구현을 더욱 용이하게 하기위해 a_0=2로 하는 dyadic 웨이브렛 변환은 식 (5)와 같이 주어진다.

$$d_{j,k} = \frac{1}{\sqrt{2^j}} \int f(t)\ \varphi^*\left(\frac{t}{2^j} - kT\right)\ dt \tag{5}$$

그림 1은 이산 웨이브렛의 dyadic형태를 구현하기 위해 사용되는 트리(tree)형태의 필터뱅크를 나타내고 있다. 여기에서 $h_0(n)$과 $h_1(n)$은 각각 저역과 고역통과 필터를 나타내고 ↓2는 2배로 다운 샘플링 한다는 의미이다.

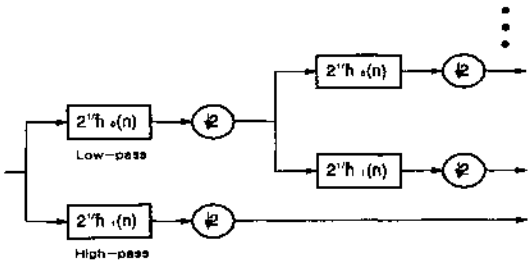


그림 1. 이산 웨이브렛 변환을 위한 트리구조의 필터뱅크
Fig. 1. Tree-structured filter bank for discrete wavelet transform.

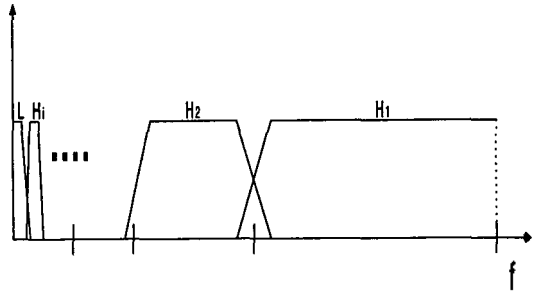


그림 2. 웨이브렛 변환의 주파수 해상도
Fig. 2. Frequency resolution of wavelet transform.

웨이브렛 변환시 각 스케일은 각각 바로 위의 스케일에서 factor 2로 decimation하여 구해지게 되므로 각 scale의 샘플수는 상위 스케일의 반이 된다. 신호를 웨이브렛 변환을 이용하여 분할하기 위해서는 트리형태의 필터뱅크를 구성한다. 입력신호가 저역통과 필터와 고역통과 필터를 거치게 되면 한 번의 웨이브렛 변환이 이루어지며, 이러한 과정을 반복적으로 수행하여 웨이브렛 변환된 신호를 얻을 수 있다. 그림 2는 이와같은 과정을 거친 신호의 주파수 해상도를 나타내고 있고 이것이 바로 dyadic 이산 웨이브렛 변환의 주파수 해상도이다.

III. 끝점검출 파라미터

3.1 제안된 끝점검출 파라미터

일반적으로는 끝점검출에 많이 사용되는 방법은 단구간 에너지와 문턱값을 비교해서 대략적인 끝점을 찾은 뒤에 영교차율로 정확한 끝점을 찾아내는 방법이다. 먼저 N개 데이타를 한 구간으로 하여 얻어진 단구간 에너지에 대수를 취한 값을 사용한다. 이때 신호 크기는 음성마다 달라지므로 음성의 크기에 따라 적응적으로 정규화하는 것이 필요하다. 이러한 정규화된 에너지와 영교차율을 사용하여 시작점과 끝점을 검출하게 된다. 하지만 이러한 검출방법은 잡음이 없는 음성신호에 대해서는 신뢰할 수 있는 결과를 보이지만 잡음이 조금이라도 타게 되면 그 결과의 신뢰성을 보장받을 수 없게된다[10-11].

시작이나 끝부분에 존재하는 파열음이나 마찰음의 경우는 신호의 에너지가 유성음구간에 비해 작아서 잡음환경하에서 검출하기가 용이하지 않으며 시작점/끝점 검출 실패의 주요한 이유중 하나가 된다. 하지만 마찰음이나 파열음의 경우 비록 신호의 에너지는 유성음구간에 비해 상대적으로 작지만 주파수 영역에서 고주파 부분에 많은 에너지를 가지게 된다. 따라서, 주파수 영역에서의 에너지 분포를 고려함으로써 잡음환경하에서도 음성을 검출할 수 있는 새로운 검출 파라미터를 웨이브렛 변환영역에서 제안하였다. 그림 3은 /start/라는 음성에 대한 첫 번째 detailed와 세 번째 coarse 스케일에서의 웨이브렛 계수의 표준편차 분포를 보여주고 있다. 그림에서도 볼 수 있듯이 세 번째 coarse 스케일은 저주파를 잘 반영하는 유성음 구간을 대표한다고 볼 수 있으며 첫 번째 detailed 스케일은 마찰음이나 파열음의 존재유무를 알려주게 된다.

이러한 성질을 이용하여 본 과제에서는 웨이브렛 영역에서 음성검출을 위한 새로운 검출 파라미터를 개발하였다. 즉 식 6과 같이 세 번째 coarse와 가중치를 곱한 첫 번째 detailed 스케일의 합으로 검출 파라미터를 정의하였다.

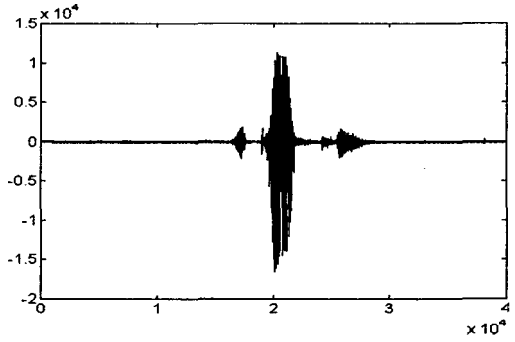
$$PA = \sigma_{d3} + \lambda\sigma_{d1} \tag{6}$$

σ_{d1} : 첫 번째 detailed 스케일의 표준편차

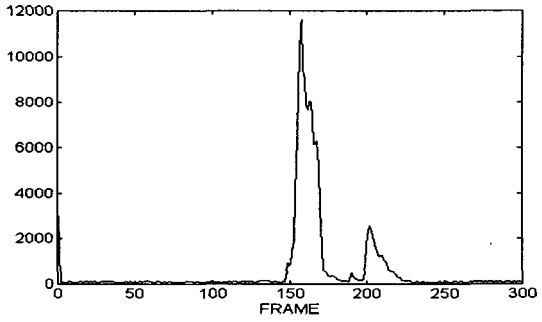
σ_{d3} : 세 번째 coarse 스케일의 표준편차

λ : weighting factor

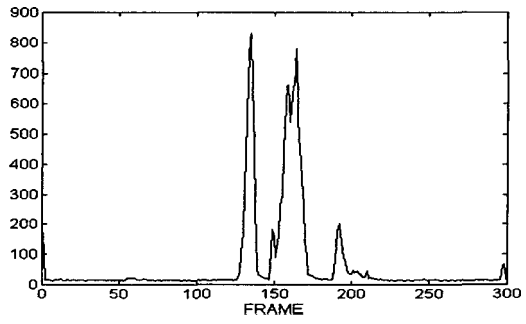
그림 4는 기존의 끝점검출 파라미터들과 본 연구를 통해 제안된 검출 파라미터의 성능을 비교하기 위해 자동차 잡음환경에서 이들 파라미터의 프레임별 변화를 나타낸 것이다. 그림 5는 제안된 검출 파라미터를 이용하여 잡음이 없는 음성과 자동차 잡음환경에서 SNR이 0dB와 10dB일 경우에 검출 파라미터의 프레임에 따른 변화를 나타낸 것이다. 그림 4와 5에서 수직선은 실제 끝점을 나타내고 있다.



(a) 음성신호



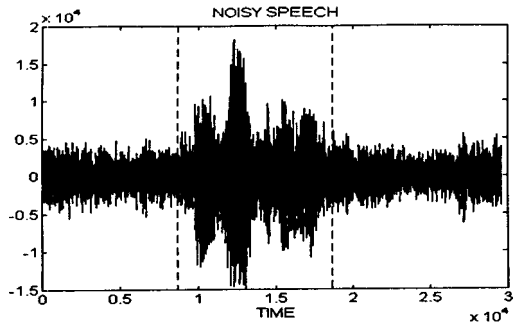
(b) 세 번째 coarse 스케일



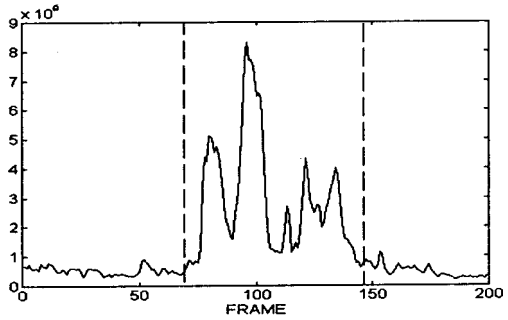
(c) 첫 번째 detailed 스케일

그림 3. 음성신호 /start/에 대한 웨이브렛 계수의 표준편차의 프레임에 따른 분포

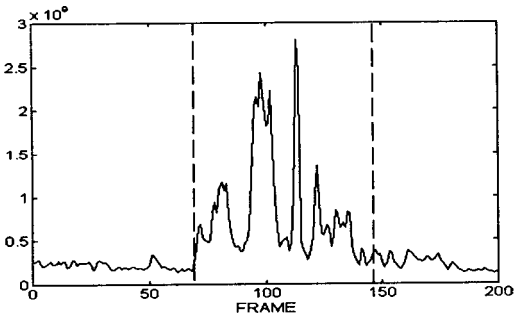
Fig. 3. Frame-based distribution of standard deviation of wavelet coefficients for the utterance /start/.



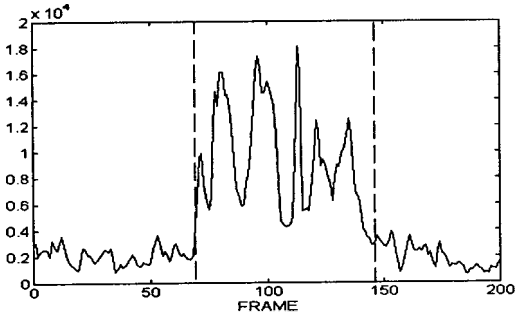
(a) 잡음음성(SNR=10dB)



(b) Square energy

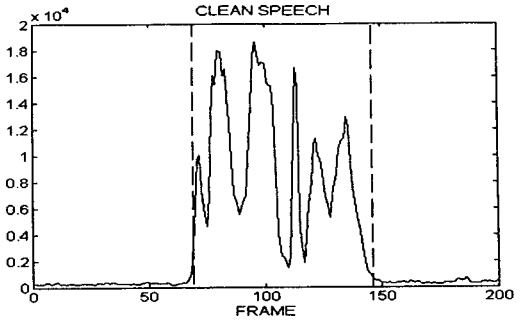


(c) Teager 에너지

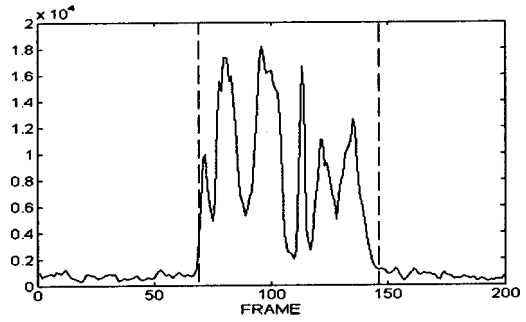


(d) 제안된 파라미터

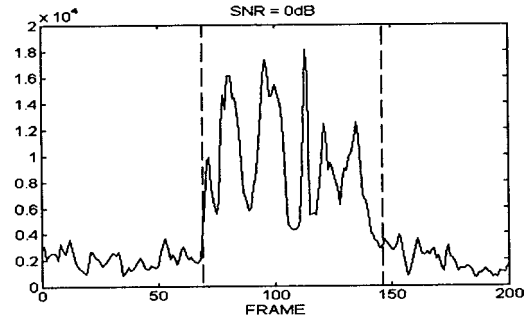
그림 4. 끝점검출 파라미터의 비교
Fig. 4. Comparison of endpoint detection parameter.



(a) 원 음성



(b) SNR=10dB



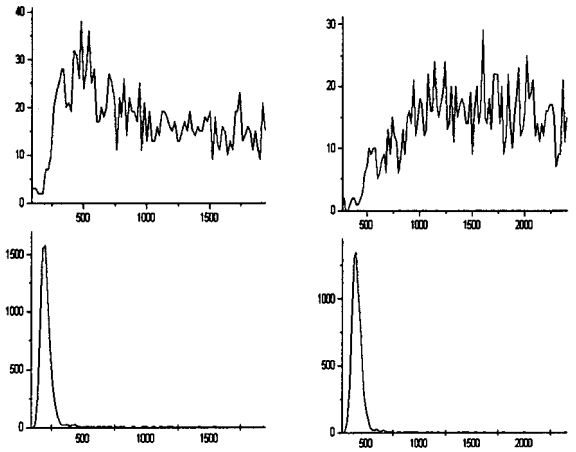
(c) SNR=0dB

그림 5. 음성신호 /신호처리/에 대한 제안된 파라미터의 프레임에 따른 분포
Fig. 5. Frame-based distribution of proposed parameter for the utterance /신호처리/.

3.2 가중치 λ값의 결정

검출 파라미터에서 정의한 가중치 λ를 정하기 위해 음성구간(유성음과 무성음)과 배경잡음부분에 해당하는 묵음구간의 히스토그램을 이용하였다. 사용된 음성은 조용한 실험실에서 녹음한 음성을 대상으로 하였다. 그림 6은 λ의 값이 각각 1과 5일때의 검출 파라미터분포에 따른 히스토그램을 나타내고 있다. 검출 파라미터에서 λ의 의미는 고주파영역을 강조하기 위한 것으로 λ의 값이 증가함에 따라 음성구간과 묵음에서의 검출 파라미터값 역시 증가함을 그림 6을 통해 확인할 수 있다. 그리고 λ 값이 증가함에 따라 음성구간 히스토그램의 무게중심이 뒤쪽으로 이동하는 것을 알 수 있다. 정확한 λ의 값을

결정하기 위해서 각각 다른 λ값($1 \le \lambda \le 10$)에 대해 음성구간과 묵음구간의 겹치는 프레임수를 측정하였다. 그림 7은 λ값에 따라 음성구간과 묵음구간의 겹치는 프레임 수를 나타낸 것이다. 그림에서도 확인 할 수 있듯이 대략 λ = 5~7에서 음성구간과 묵음구간의 겹침이 작았으며 7 이상의 값에서는 겹치는 프레임 수가 다시 증가하기 시작해서 10 이상에서는 급격히 증가 하였다. 본 연구에서는 실제 음성검출을 위해서 λ는 6으로 설정하였다.



(a) $\lambda=1$ (b) $\lambda=5$

그림 6. 음성구간과(상단) 배경잡음구간(하단)의 히스토그램
Fig. 6. Histogram of speech(upper part) and background(lower part) region.

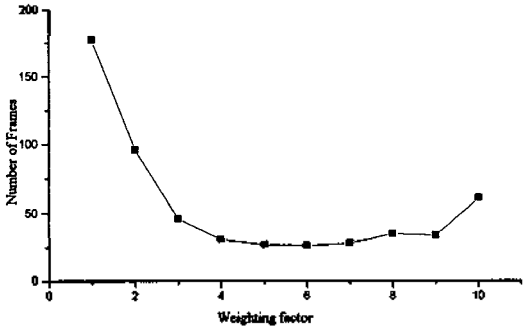


그림 7. λ값에 따른 음성구간과 배경잡음구간의 겹쳐지는 프레임 수
Fig. 7. Number of overlapping frames depending on weighting factor λ.

IV. 끝점검출 알고리듬

4.1 적응적 문턱치 결정

웨이브렛 영역에서 세 번째 coarsed 스케일의 표준편차가 잡음이 없을 경우 첫 번째 detailed 스케일에 비해 최소 다섯배 이상이 된다는 점과 잡음이 존재할시에는 거의 비슷한 값을 보인다는 점을 확인할 수 있었다. 따라서 초기 10프레임에서 세 번째 coarsed 스케일의 표준편차의 평균이 첫 번째 detailed 스케일의 그것보다 크면 잡음이 비교적 작은 음성이라 간주하고 반면에 첫 번째 detailed 스케일의 표준편차의 평균이 크면 잡음이 첨가된 음성이라 판별하게 된다. 이러한 성질을 이용하여 잡음의 존재를 자동으로 판별하여 잡음의 존재유무에 따라 음성검출을 위한 문턱치를 적응적으로 결정할 수 있게 하였다. 문턱치 결정과정은 다음과 같다.

(i) 프레임 단위로 웨이브렛 변환을 수행한다.
(ii) σ_{d1} 과 σ_{d3} 를 구한다.
(iii) (i)과 (ii)과정을 처음 연속적인 10프레임에 대해 구하고 σ_{d3} 과 $\lambda \times \sigma_{d1}$ 의 평균을 구한다.
(iv) 만일 σ_{d3} 이 $\lambda \times \sigma_{d1}$ 보다 크다면 $4 \times \sigma_{d3}$ 를 문턱치로 정한다. 그렇지 않을 경우 $2 \times \sigma_{d1}$ 를 문턱치로 정한다.

4.2 끝점검출 과정

초기 10개의 프레임을 대상으로 문턱치를 결정한 다음 입력되는 프레임에 대하여 검출 파라미터를 구하고 미리 정한 state에 따라 입력된 데이타가 음성구간인지 아닌지를 판별하게 된다. 끝점검출을 위해 6개의 state를 가정하였으며 각 state는 다음과 같다.

START : 문턱치를 결정하기 위한 초기 10 프레임의 묵음구간
INSILENCE : 묵음구간
SIGNAL : 음성구간
MAYBEEND : 잠정적인 끝점이 정해진 구간
END : 실제 끝점이라 판명된 구간
DISCARD : START state에서 한 프레임내에 신호의 레벨이 모두 0인 구간 이거나 미리 정해 놓은 일정 레벨 이상의 값을 가지는 경우

끝점검출 과정은 아래와 같다.

(i) 프레임 단위로 웨이브렛 변환을 수행한다.
(ii) 검출파라미터 $PA = \sigma_{d3} + \lambda\sigma_{d1}$ 를 구한다.
- 시작 10프레임 구간 이내이면 START state
- 시작 10프레임 구간 밖이면 INSILENCE state

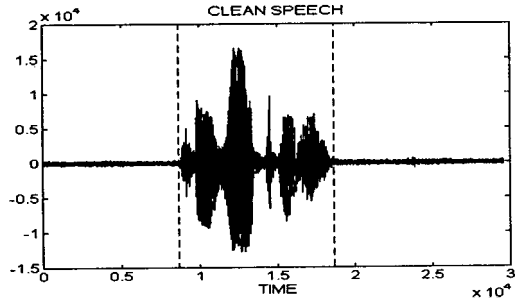
(iii) 음성의 시작점을 찾는다.
- PA가 문턱치 보다 큰 값이면 SIGNAL state
- PA가 문턱치 보다 작으면 INSILENCE state
- 연속된 5프레임 이상이면 음성 시작구간이라 간주하고 시작점으로 정한다.

(iv) 음성의 끝점을 찾는다.
- PA가 문턱치의 반보다 작은값이면 MAYBEEND state
- PA가 문턱치의 반보다 작은값이 아니면 SIGNAL state

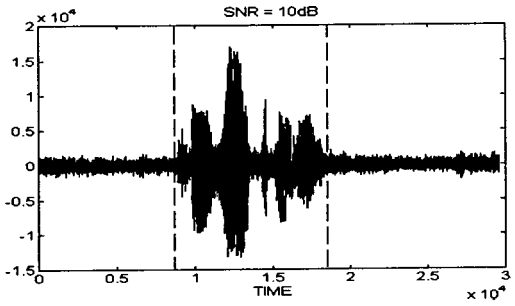
- PA가 문턱치의 반보다 작은값이 연속된 20프레임 이상이면 음성의 끝구간이라 간주하고 END state

(v) 만일 검출된 음성구간이 20프레임 미만이면 검출된 부분은 짧은 잡음성분이 존재하는 구간이라 간주하고 시작점을 다시 찾기위해 (i)과정으로 돌아간다.

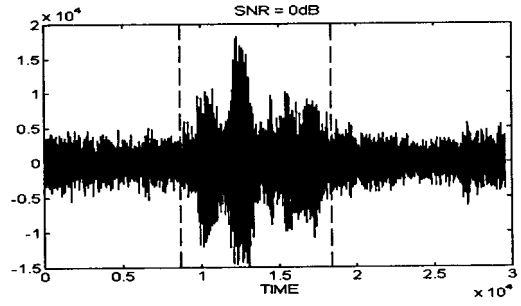
그림 8은 /신호처리/라는 실제음성을 대상으로 하여 잡음이 없는 경우, SNR이 10dB 그리고 SNR이 0dB의 경우에 실제 검출결과를 보여주고 있다. 그림에서도 확인할 수 있듯이 잡음환경하에서도 제안한 알고리듬이 음성을 잘 검출할 수 있음을 볼 수 있다.



(a) 원 음성



(b) SNR=10dB



(c) SNR=0dB

그림 8. 음성신호 /신호처리/에 대한 실제 끝점검출 결과
Fig. 8. The result of actual endpoint detection for the utterance /신호처리/.

V. 실험 및 고찰

음성 데이타의 시작점과 끝점을 모두 3가지 방법으로 검출하여 검출결과를 비교하고 인식실험에 이용하였다. 각 방법은 첫째 manual한 끝점 검출방법, 둘째, EZ 방식(Energy 와 Zero-crossing 이용)[8], 세째 본 연구에서 개발한 웨이브렛 변환을 이용한 검출방법 이다.

먼저 manual하게 시작점/끝점을 검출하는 방법은 개개의 데이타에 대하여 수작업으로 직접 음성의 시작점과 끝점을 확인하여 세그먼테이션을 함으로써 각 데이타에 대하여 시작점과 끝점을 구하였다. 이 방법으로 검출된 시작점과 끝점은 다른 방법으로 검출한 시작점과 끝점의 정확도에 대한 비교 기준으로 사용하였다. EZ 방식에서는 사용될 문턱값을 수집된 데이타에 맞도록 조절하여 실험하였으며, 웨이브렛 변환을 이용한 방법에서는 자체적으로 개발한 검출 알고리즘을 이용하였다. 검출 정확도 및 인식결과를 제시함에 있어서, 서술상의 편의를 위하여 EZ 방식의 검출 알고리즘을, "EZ 방식"이라 하며 웨이브렛 방식으로 검출한 경우를 "웨이브렛 방식"이라고 하기로 한다. 한편 이 두가지 방법과 비교하게될 manual하게 시작점과 끝점을 검출한 경우를 "manual"이라고 하기로 한다.

5.1 음성 DB 수집

음성 DB는 ETRI의 445DB에서 단어의 사용빈도를 고려해서 모든 음소를 골고루 포함하도록 40개의 단어를 선정하였다. 환경적인 요인과 화자의 변동성을 배제하기 위해 스피커를 통해 음성 파일을 재생하면서 노트북 컴퓨터 및 데스크 탑 컴퓨터를 이용하여 음성신호를 녹음하여 데이터 수집을 하였다. 화자의 선정은 445DB에 명시된 6명의 남성화자와 4명의 여성화자로 구성된 테스트 화자 10명으로 하였다. 녹음된 음성 DB는 16kHz로 샘플링 되었고 16bits의 resolution을 가진다.

4.2 시작점과 끝점의 정확도 비교

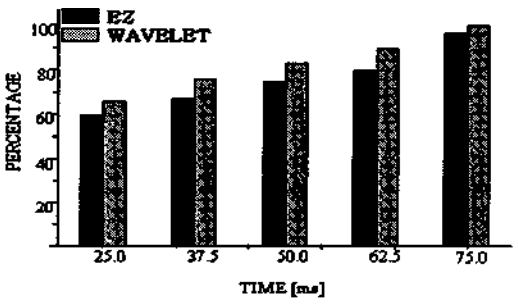
수집된 음성 DB에 white Gaussian noise, computer noise, car noise를 첨가하여 다양한 SNR을 가지는 잡음음성을 만들어 이를 대상으로 실험을 수행하였다. 정확도 비교를 위해서 EZ 방식과 웨이브렛 방식을 이용하여 얻은 결과를 manual하게 검출한 결과와 비교하였다. 표 1은 clean 음성과 SNR이 각각 10dB, 20dB인 경우에 대하여 시작점/끝점 검출 정확도를 보여주고 있다. 즉, 표 1에서는 manual하게 검출된 시작점과 끝점을 기준으로 에러 범위를 25ms부터 75ms 사이에서 12.5ms 간격으로 변화시키면서 EZ방식과 웨이브렛 방식을 이용하여 검출한 시작점/끝점의 결과가 허용오차범위내에 들어가는 경우를 백분율로 표시한 것이다.

EZ 방식의 경우를 살펴보면 SNR이 10dB에서 실제 시작점/끝점과의 차이가 75ms 안에 들어가는 경우는 전체 400개의 DB에서 시작점의 경우 89.9%이고 끝점의 경우는 86.1%에 해당된다. 반면에 웨이브렛 방식의 경우 동일한 SNR에서 시작점의 및 끝점 경우의 99%가 75ms

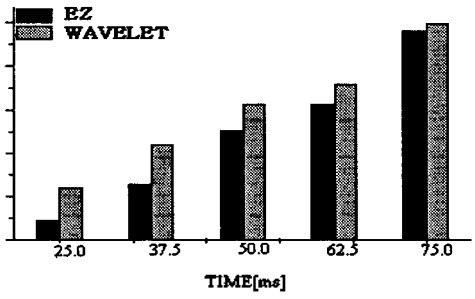
의 에러범위 안에 포함 되었음을 볼 수 있다. 전체적으로 EZ 방식보다는 웨이브렛 방식이 더 나은 성능을 보임을 확인할 수 있었다. 그림 9는 SNR이 15dB에서의 시작점과 끝점의 정확도를 그림으로 나타낸 것이다.

표 1. SNR에 따른 끝점검출 정확도 결과
Table 1. The results of endpoint detection accuracy across the several SNRs.

EZ 방식						
Time*	10dB		20dB		Clean	
[ms]	시작점	끝점	시작점	끝점	시작점	끝점
25.0	49.1%	6.9 %	63.9%	11.2%	65.9%	13.1%
37.5	56.5%	18.8%	70.2%	33.3%	71.8%	39.5%
50.0	64.4%	37.4%	77.8%	57.4%	80.0%	60.8%
62.5	70.0%	48.6%	82.0%	69.2%	84.4%	72.3%
75.0	89.9%	86.1%	98.7%	95.5%	100%	98.7%
웨이브렛 방식						
Time	10dB		20dB		Clean	
[ms]	시작점	끝점	시작점	끝점	시작점	끝점
25.0	59.2%	11.2%	73.0%	40.7%	86.2%	70.8%
37.5	71.5%	23.9%	82.0%	64.1%	94.9%	86.4%
50.0	80.8%	48.7%	89.2%	81.4%	97.2%	94.1%
62.5	86.1%	69.3%	92.7%	88.0%	97.4%	95.4%
75.0	99.8%	99.0%	100%	99.7%	100%	100%



(a) 시작점



(b) 끝점

그림 9. SNR 15dB에서 끝점검출 정확도 결과
Fig. 9. The result of endpoint detection accuracy for SNR 15dB.

5.3 음성인식 실험

인식실험은 크게 두 부분으로 나누어서 시행되었다. 첫째는 음성데이타를 인식기에 입력시킬 때 필요한 시작점과 끝점을 일괄적인 인식실험의 수행을 위하여 사전에 인식에 사용될 각 음성데이타에 대하여 검출하여서는 텍스트 파일로 저장하는 부분이고, 둘째는 음성데이타를 인식기에 직접 입력시켜서 인식시키는 부분으로 인식엔진을 구동시킬 때 별도로 끝점검출을 하지 않고 미리 검출하여 저장해놓은 텍스트 파일을 이용하여 시작점과 끝점을 읽어서 인식실험에 이용하였다.

인식실험 결과를 제시함에 있어서 크게 두가지 경우를 인식 실패한 것으로 간주하고 이에 따른 결과를 제시하였다. 첫째는 인식기에 음성데이타를 입력하기 전에 끝점검출 부분에서 끝점검출에 실패한 경우에 해당한다. 인식기에 입력되는 음성데이타는 검출한 시작점과 끝점을 기반으로 하여 해당되는 음성부분을 인식하게 되므로 시작점과 끝점을 검출하지 못하는 경우는 인식에 실패하게 된다. 그러므로 이 경우를 인식실패에 포함시켰다. 둘째는 시작점과 끝점을 검출하여 해당하는 음성부분을 인식기에 입력시켰을 때 인식기에서 다른 단어로 오인식하여 인식에 실패하였는 경우이다. 인식실험에 사용된 인식엔진은 HMM(Hidden Markov Model)알고리듬을 기반으로 하는 인식시스템을 이용하였다.

EZ 방식의 경우 SNR이 높을 경우(20dB 이상)는 manual하게 검출한 경우와의 인식율 차이가 5%를 넘지 않는것으로 나타났다. 하지만 낮은 SNR의 경우는 그 차이가 30%이상 나타났다. 하지만 웨이브렛 방식의 경우 낮은 SNR에서도 manual 방식과의 차이가 3% 정도로 거의 비슷하게 나타남을 확인할 수 있었다. 그림 10은 각 방식별 인식율을 SNR에 따라 보여주고 있다.

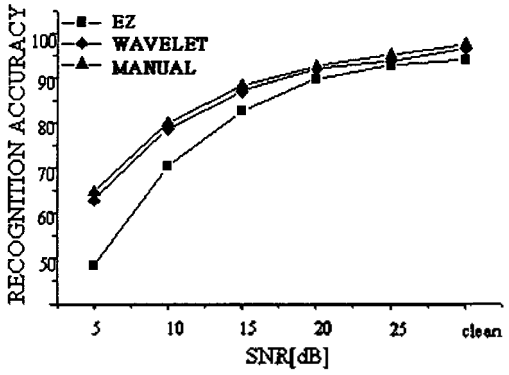


그림 10. SNR에 따른 인식성능 비교
Fig. 10. Comparison of recognition performance across several SNRs.

VI. 결 론

본 논문에서는 잡음이 포함된 음성의 시작점과 끝점을 효율적으로 검출할 수 있는 알고리듬에 대하여 연구하였다. 이를 위해, 시간-주파수 영역에서의 에너지 분포를 고려함으로써 잡음환경하에서도 음성을 검출할 수 있는 새로운 검출 파라미터를 제안하였다. 제안된 끝점검출 파라미터는 웨이브렛 영역에서 세 번째 coarsed 스케일의 표준편차와 가중치를 곱한 첫 번째 detailed 스케일의 표준

편차의 합으로 정의하였고, 음성의 시작이나 끝부분에 존재하는 파열음이나 마찰음 같이 신호의 에너지가 유성음 구간에 비해 작아서 잡음환경하에서 검출하기가 용이하지 않은 경우에도 음성부분을 효율적으로 검출할 수 있었다.

시작점과 끝점의 정확도 면에서는 기존의 방식의 경우, SNR이 10dB에서 실제 시작점과 끝점과의 차이가 75ms 안에 들어가는 경우는 전체 시작점의 경우 89.9%이고 끝점의 경우는 86.1% 정도인 반면에 제안된 방식의 경우 동일한 SNR에서 시작점의 및 끝점 경우의 99%가 75 ms의 에러범위 안에 포함 되었음을 확인할 수 있었다. 인식 실험 결과에 있어서도 SNR이 낮은 경우 제안된 방식은 기존 방식에 비해 30% 정도의 인식을 향상이 있었다.

참 고 문 헌

1. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., Vol. 54, No. 2, pp. 297-315, February 1975.
2. L.F. Lamel et. al., "An improved endpoint detector for isolated word recognition", IEEE Trans. Acoust., Speech, and Signal Processing, Vol, ASSP-29, No.4, pp.777-785, 1981.
3. Jean-Claude Junqua, Brian Mak, and Ben Reave, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE Trans. Acoust., Speech, and Signal Processing, Vol 2, No. 3, pp.406-412, 1994.
4. J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", In Pro. IEEE ICASSP-90, pp.381-384, Apr. 1990.
5. G.S. Ying, C.D. Mitchell, L.H. Jamieson, "Endpoint detection of isolated utterancesbased on a modified teager energy measurement", In Proc. IEEE ICASSP - 93, pp732-735, 1993.
6. I. Daubechies, "The Wavelet Transform Time FrequencyL-ocalization and Signal Analysis", IEEE Trans. on Info. Theory, Vol. 36, No. 5.
7. O. Rioul and M. Vetterli, "wavelet and Signal Processing", IEEE Signal Processing Magazine, pp. 14-38, Oct. 1991
8. M. Vetterli, "Multi-dimensional Sub-band Coding", Signal Processing, Vol. 6, No. 2, pp. 97-112, 1984.
9. I. Daubechies, Ten Lectures on wavelets, SIAM, 1992.
10. M. H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals", Speech Communication, Vol. 8, No. 1, pp. 45-60, March 1989.
11. M. Han and C. K. Park, "An improved speech detection algorithm for isolated Korean utterances", Proc. IEEE ICASSP, pp. 525-528, 1992.

▲석 종 원(Jong Won Seok)

1993년 2월 : 경북대학교 전자공학과 (공학사)
1995년 2월 : 경북대학교 전자공학과 (공학석사)
1995년 2월~현재 : 경북대학교 전자공학과 박사과정 재학중

※주관심분야 : 디지털 신호처리, 음성신호처리, 웨이브렛 이론

▲배 건 성(Keun Sung Bae)

1977년 2월 : 서울대학교 전자공학과 (공학사)
1979년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)
1989년 5월 : University of Florida (공학박사)
1979년 3월~현재 : 경북대학교 전자공학과 교수

※주관심분야 : 음성분석 및 인식, 디지털 신호처리, 디지털 통신, 음성 부호화, 웨이브렛 이론 등