Journal of the Research Institute of Industrial Technology, Vol. 18, 2001

주파수영역 에너지와 SFM을 이용한 잡음환경에서의 음성구간 추출에 대한 연구

정성일[1], 유광주[1], 신옥근[2]

Voice Activity Detection Based Spectral Energy and SFM In Noisy Environments

S. I. Jung, G. J. You, O. K. Shin

요 약

본 논문에서는 잡음이 섞인 음성신호에서 음성구간을 찾는 새로운 방법을 제안한다. 잡음에 강인한 검출성능을 위해 주파수 스펙트럼의 각 성분이 얼마나 평활한지를 나타낼 수 있는 *SFM*(Spectral Flatness Measure)과 주파수 영역에서의 에너지 평균을 이용한다. 제안한 음성구간 검출방법은 특정한 환경에만 한정되는 것이 아니라, 다양한 잡음환경에서도 우수한 검출성능을 갖는 것으로 확인되었다.

1. 서 론

음성구간 추출(Voice Activity Detection)은 입력 신호에서 음성과 묵음 혹은 잡음구간을 구분하는 과정을 말한다. 음성구간 추출은 음성인식(Speech Recognition)뿐만 아니라 음성코딩(Speech Coding), 음성합성(Speech Synthesis), 음성 DB작업등의 전처리 과정으로 여러 가지 음성 관련 분야에서 필수적인 요소이다. 잡음이 없는 신호에서는 전력밀도[1]나 영교차율(Zero Crossing Rate), LPC(Linear Predictive Coding)[2] 혹은 HMM(Hidden Markov Model)[3]등의 파라미터만 이용하여 음성구간을 추출하여도 만족할 만한 결과를 얻을 수 있다. 그러나 잡음 환경에서의 음성구간 추출은 무성음과 잡음이 유사한 경우가 많아 정확한 시작점과 끝점 추출이 매우 어렵다. 특히 음성인식을 위한 음성구간 추출은 음성코딩의 경우보다 정확해야하며, 잘못 추출한 음성구간은 오인식의 원인이 된다[4].

본 논문에서는 잡음에 강인한 음성구간 추출을 위해서, 주파수 스펙트럼의 각 성분이 얼마나 평활한지를 나타낼 수 있는 *SFM*[5]과 에너지 평균을 이용하였다. 제안하는 음성구간 추출방법의 성능을 검사하기 위하여, 두 가지 실험을 하였다. 먼저 잡음이 없는 환경에서 추출실험을 하였으며, 두 번째는 여러 가지 종류의 잡음환경에서 추출실험을 하였다.

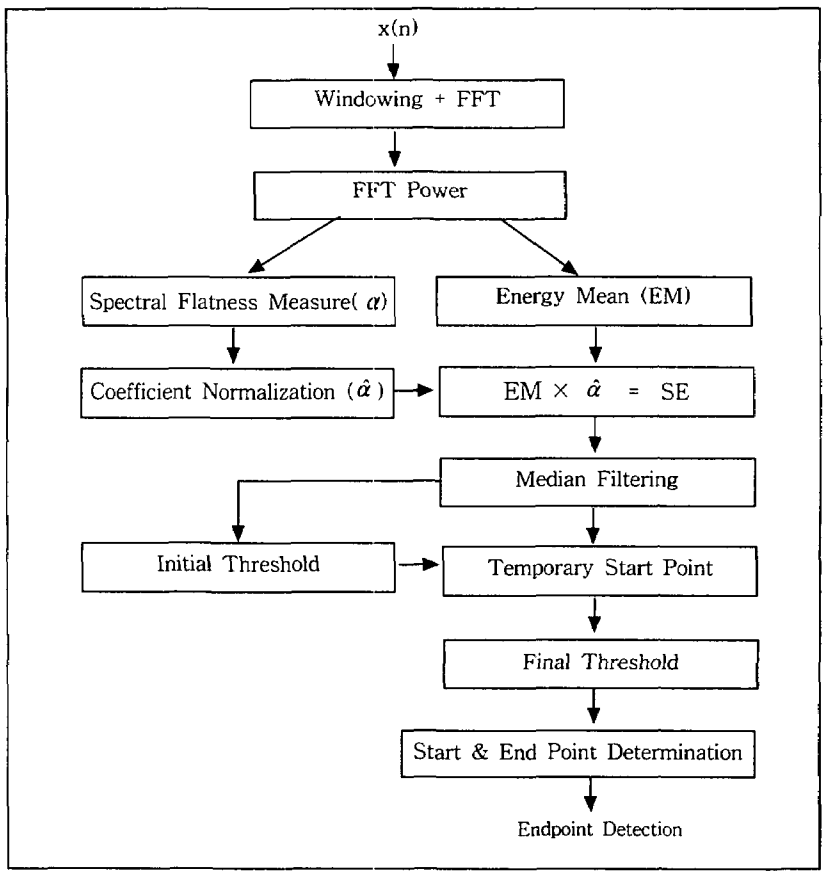
잡음이 없는 깨끗한 음성신호에서의 추출율은 평균 96.5%였으며, 백색 또는 Pink 잡음이 섞인 음성신

1) 한국해양대학교 대학원
2) 한국해양대학교 기계·정보공학부

호에서는 성능이 평균 약 90% 였으며, F16 또는 공장 잡음이 섞인 음성신호에 대해서는 상대적으로 낮은 추출율을 나타내었다. 본 논문 II장에서는 음성의 시작점과 끝점을 추출하는 알고리즘을 전반적으로 소개하고, III장에서는 실험 결과와 고찰, 마지막으로 IV장에서는 결론의 순서로 본 논문을 기술한다.

2. 음성 구간 추출

본 논문에서는 *SFM*과 에너지 평균을 잡음이 섞인 음성신호에 적용하여 음성 구간을 검출하며, 제안한 방법의 흐름도를 그림 2.1에 나타내었다.



<그림 2.1> 제안한 음성구간 추출 방법의 흐름도

2.1 에너지의 평균

잡음이 섞인 음성신호를 프레임 단위로 FFT한 다음, 파워 스펙트럼 $P(w)$를 구한다. 프레임의 스펙트럼 파워를 합하여, i번째 프레임의 에너지(FE_i)를 (식 2.1)과 같이 구한다. 그리고 에너지 평균 (EM_i)을 (식 2.2)와 같이 구한다.

$$FE_i = \sum_{w=1}^{FFTPOINT} P_i(w) \tag{2,1}$$

$$EM_i = \frac{FE_i}{FFTPOINT} \tag{2.2}$$

2.2 Spectral Flatness Measure(*SFM*)

SFM_{dB}는 파워 스펙트럼의 산술 평균 (A_m)과 기하 평균 (G_m)의 비(Ratio)를 dB 단위로 나타낸 것이며, (식 2.3)과 같이 정의한다.

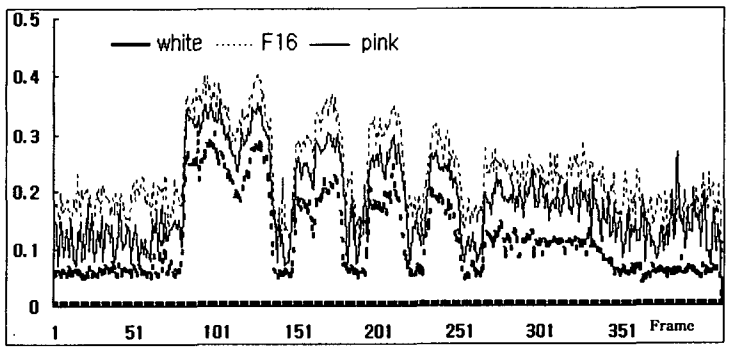
$$SFM_{dB} = 10\log_{10}\frac{Gm}{Am} \tag{2.3}$$

입력신호가 잡음성분과 유사한지, 아니면 음성성분과 유사한지를 나타내는 파라미터 α는 (식 2.4)와 같이 정의한다.

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dBMax}}, 1\right) \tag{2.4}$$

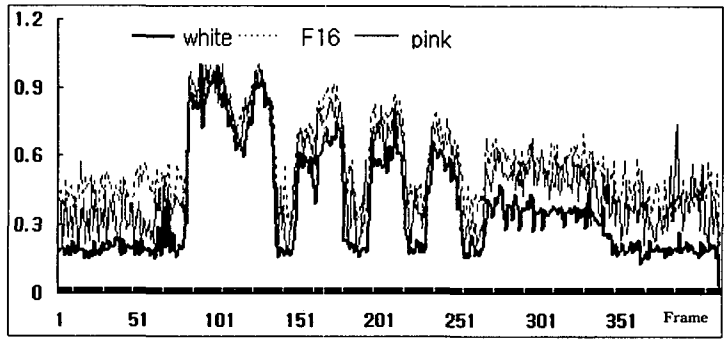
여기서 SFM_{dBMax}는 $-60dB$로 설정하였으며, SFM_{dB}가 SFM_{dBMax} 에 가까울 때의 프레임은 음성과 비슷한 스펙트럼 분포를 나타내며, 반대로 $0dB$에 가까우면, 잡음과 비슷한 스펙트럼의 분포를 나타낸다. 예를 들어, SFM_{dB}가 $-60dB$이상이면 α=1이 되고, $-30dB$이면 $\alpha = 0.5$이다.

주파수 스펙트럼의 분포가 거의 일정한 백색잡음이 섞인 음성신호의 경우, 잡음구간과 음성구간을 파라미터 α로 분명하게 구분할 수 있지만, 주파수 스펙트럼의 분포가 일정하지 않은 pink 또는 F16잡음인 섞인 음성 신호의 경우에는 그림 2.2에서 보는 것처럼 α만으로 잡음구간과 음성구간을 구분하기에 충분치가 않다.



< 그림 2.2 > 잡음 음성의 α

본 논문에서는 α를 음성구간 판별을 위한 파라미터로 사용하기 위해서, α 값들을 최대 값인 α_{Max}로 나누어 정규화 하였으며, 정규화 된 $\alpha(\hat{\alpha})$를 그림 2.3에 보인다.

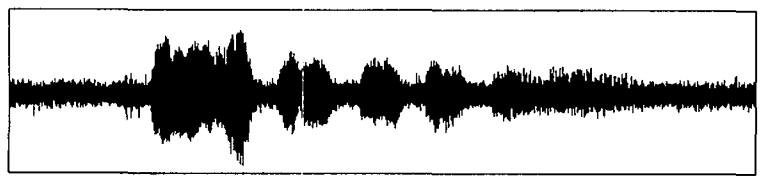


< 그림 2.3 > 정규화 된 $\hat{\alpha}$ 값

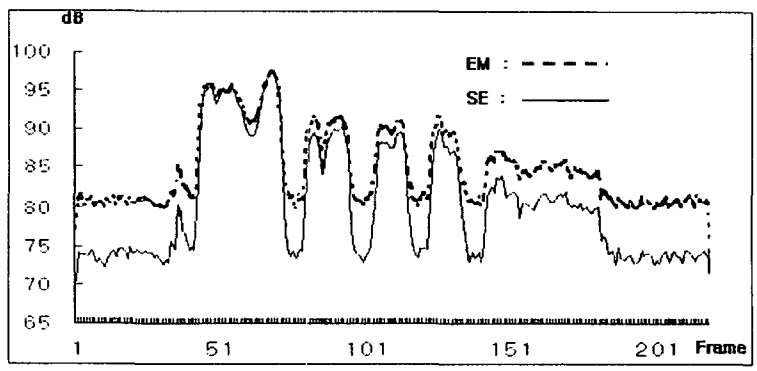
2.3 $\hat{\alpha}$과 EM의 곱

그림 2.4는 백색잡음이 첨가된 음성신호를 나타낸 것이고, 그림 2.5는 이 신호의 에너지 평균(EM)과 $SE(\hat{\alpha} * EM)$를 나타낸 것이다. 그림 2.5에서는 잡음과 음성구간을 $\hat{\alpha}$ 또는 EM만으로 구별하는 것보다 SE로 구별하는 것이 변별력이 높음을 볼 수 있다.

$$SE_i = \hat{\alpha}_i \times EM_i \qquad (2.5)$$



< 그림 2.4 > 백색 잡음이 첨가된 음성신호



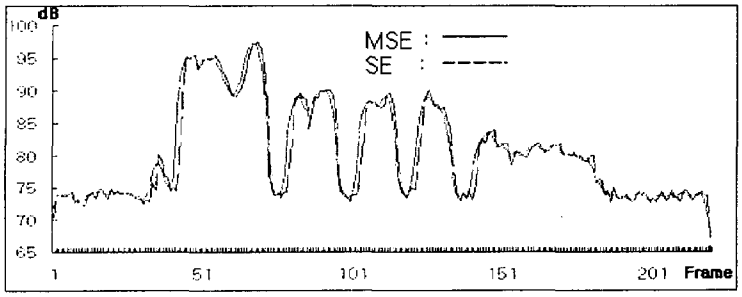
< 그림 2.5 > 그림 2.4로부터 추출한 EM과 SE

2.4 Median Filter

잡음구간 속에서 짧은 구간동안(Pause, Stop 구간)에 큰 에너지를 갖는 잡음성분의 영향으로 SE가 갑자기 커지는 현상을 제거하기 위해, (식 2.6)의 Median Filter[6]를 이용하여 SE를 평활화 한다.
식(2.6)의 MSE는 Meadian Filter를 거친 SE를 나타내고, l은 Median Filter를 적용할 프레임 수를 나타낸다.

$$MSE_i = \frac{1}{2l+1}\sum_{k=i-l}^{i+l} SE_k \quad , \quad l=1 \tag{2.6}$$

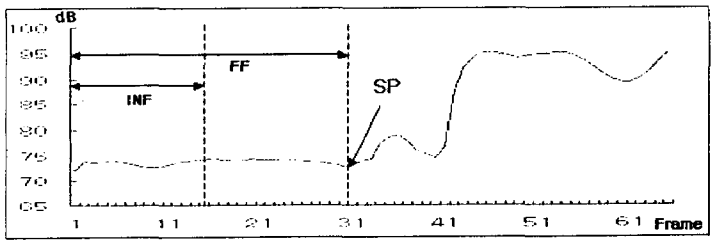
그림 2.6은 그림 2.4로부터 추출한 SE와 MSE를 나타낸 것으로, MSE가 SE보다 평활함을 알 수 있다.



< 그림 2.6 > 그림 2.4로부터 추출한 SE와 MSE

2.5 음성구간을 결정하기 위한 임계값의 결정

화자가 음성을 입력할 때 시작과 동시에 발화가 이루어지는 것이 아니라 약간의 시간이 경과 된 후에 이루어지는 것이 일반적이다. 본 논문에서는 음성이 입력되지 않는 구간을 80ms로 가정하여 임계치를 결정하며, 그림 2.7에 임계치 결정 방법을 보인다.



< 그림 2.7 > 임계치 설정 방법

그림 2.7의 INF는 음성이 입력되지 않는 80ms를 프레임 수로 나타낸 것이며, 이 구간의 평균 MSE가 초기 임계치(ITH)이다. ITH는

$$ITH = \frac{\sum_{i=1}^{INF} MSE_i}{INF} \tag{2.7}$$

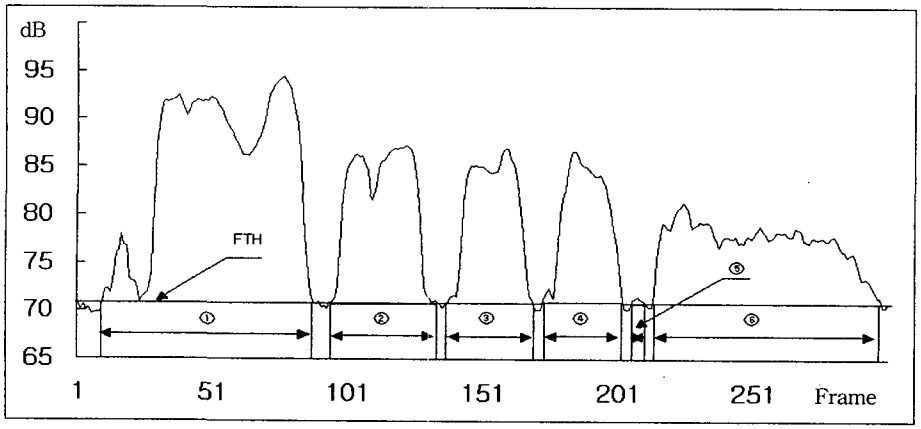
으로 정의되며, 이 임계치를 이용하여 그림 2.7에 나타낸 첫 시작점(*SP*)을 추출한다. 그림 2.7의 ***FF***는 음성신호의 처음부터 첫 시작점까지의 프레임 수를 나타낸 것으로, 이 구간에서 최종적인 임계치 (*FTH*)결정하며 *FTH*는

$$FTH = \frac{\sum_{i=1}^{FF} MSE_i}{FF} \tag{2.8}$$

으로 정의된다.

2.6 음성의 시작점과 끝점 추출방법

그림 2.8은 음성의 후보 시작점과 후보 끝점 추출방법을 보인 것으로, FTH는 임계치를, 수평 화살표 구간은 MSE가 임계치보다 큰 구간들을 나타낸다. 이 그림의 ①,②,③,④,⑥번 구간과 같이 구간의 프레임 수가 10 프레임 이상이면 각 구간의 시작점을 후보 시작점으로, 끝점을 후보 끝점으로 간주하지만, ⑤번 구간과 같이 프레임 수가 10 프레임 미만이면 구간의 시작점과 끝점을 후보 시작점과 끝점으



< 그림 2.8 > 후보 시작점과 끝점 추출방법

로 간주하지 않는다.

위에서 기술한 방법으로 음성구간을 추출하면, 여러 개의 후보 시작점과 끝점이 추출된다. 이 후보 점들 중에서 최종적인 음성의 시작점을 찾기 위해 본 연구에서는 후보 시작점과 끝점의 좌우의 *MSE*의 크기를 비교하여 최종적인 시작점과 끝점을 찾기로 한다. 일반적으로 화자의 연속 음성 내에 있을 수 있는 짧은 공백(Pause)의 길이는 대부분(99.56%) 150ms보다 짧은 것으로 알려져 있다.[7] 먼저 후보 시작점 좌측 150ms까지 *MSE*의 평균을 *LMSE*, 우측의 150ms까지 *MSE*의 평균을 *RMSE*라 정의하며, 각각 (식2.9), (식 2.10)과 같이 나타낸다.

$$LMSE_k = \frac{\sum_{i=k}^{k+L} MSE_i}{L} \tag{2.9}$$

$$RMSE_k = \frac{\sum_{i=k}^{k+L} MSE_i}{L} \tag{2.10}$$

여기서, L은 150ms에 해당하는 프레임 수이며, k는 후보 시작점의 인덱스이다.

본 논문에서는 후보 시작점의 좌측 150ms내에 있는 모든 프레임의 ***MSE***가 ***FTH***에 1.3을 곱한 값인 ***UFTH***보다 크고, 위에서 기술한 ***LMSE***와 ***RMSE***가 각각 (식 2.11)과 (식 2.12)를 만족한다면 해당 후보 시작점을 최종적인 음성의 시작점으로 간주한다.

$$0.85 * FTH \leq LMSE \leq 1.15 * FTH \quad (2.11)$$

$$RMSE \geq 1.15 * FTH \quad (2.12)$$

최종적인 음성의 끝점(***FEP***)은 시작점 결정 방법과 비슷하게 후보 끝점에 (식 2.13)와 (식 2.14)를 적용하고, 후보 끝점의 우측 150ms내에 있는 모든 프레임의 ***MSE***와 ***UFTH***를 비교하여 판별한다.

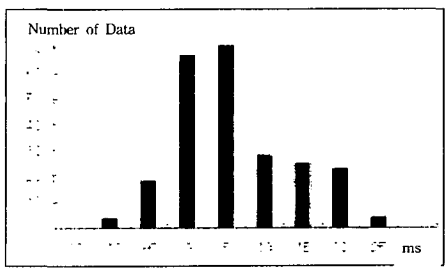
$$0.85 \times FTH \leq RMSE \leq 1.15 \times FTH \quad (2.13)$$

$$LMSE \geq 1.15 \times FTH \quad (2.14)$$

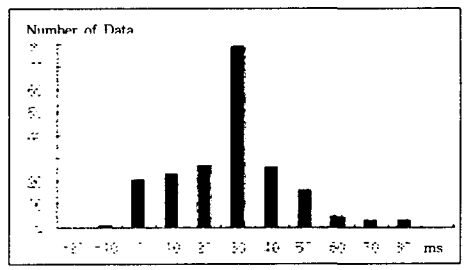
3. 실험 및 결과 분석

3.1 잡음이 없는 음성신호에서의 음성구간 추출 실험

본 논문에서 제안한 음성구간 추출성능을 알아보기 위해 먼저 조용한 환경에서 녹음한 음성신호 240개에 대하여 음성구간 추출 실험을 수행하였다. 제안한 방법으로 추출한 시작점과 끝점을 비교할 기준은 수작업을 통하여 준비하였다. 시작점의 오차허용 범위는 음성 시작점의 좌로 20ms, 우로 10ms로 설정을 하였으며, 끝점의 오차허용 범위는 음성 끝점의 좌로 10ms, 우로 50ms로 설정을 하였다.
잡음이 없는 신호는 시간축상에서의 적은 에너지를 갖는 파찰음이나 마찰음의 경우에도 우수한 추출 성능을 보였으며, 시작점은 97%, 끝점은 95%의 정추출율의 성능을 나타내었다. 그리고 시작점의 평균 오차는 5.6ms, 끝점은 15.3ms로 나타났다. 그림 3.1은 허용 오차 범위 내에서의 시작점추출 분포이며, 그림 3.2은 음성 끝점 추출 분포를 나타낸다.



< 그림 3.1 >오차 범위에서 시작점 분포

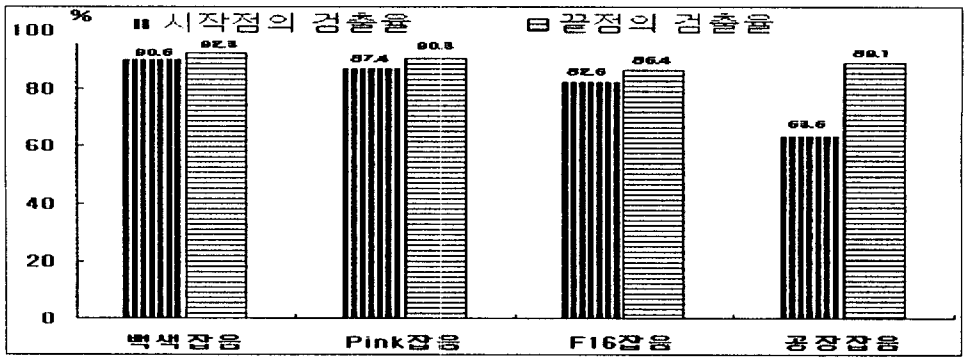


< 그림 3.2 >오차 범위에서 끝점 분포

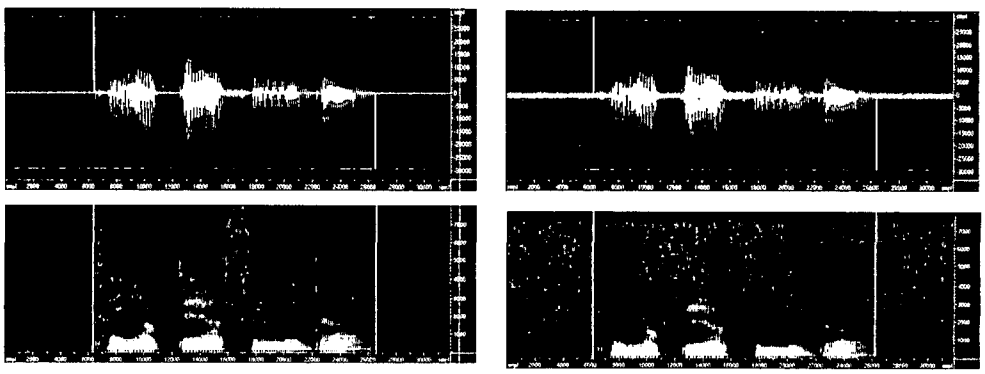
3.2 잡음 환경에서의 음성구간 검출 실험

SNR 20dB의 백색 , Pink , F16, 공장잡음이 섞여있는 음성신호 각각 240개씩에 대하여 음성추출 실험을 하였다. 이 경우 잡음에 의해서 무성 자음성분이 손상을 입게 되는데, 특히 잡음 성분과 비슷한 마찰음이나 파찰음의 경우에 자음과 쉽게 구분할 수 없어, 시작점과 끝점의 추출이 어려웠다. 이 실험에서도 잡음이 없는 음성 신호에 적용하였던 오차 허용범위를 이용하였다.

백색, Pink 잡음의 경우 잡음이 없는 데이터의 성능에 비해서 크게 뒤떨어지지 않는 결과를 나타내었으며, 백색잡음이 섞인 음성의 시작점은 90.6%, 끝점은 92.3%, Pink잡음이 섞인 음성의 시작점은 87.4%, 끝점은 90.3%의 추출성능을 나타내었다. 그러나 F16 또는 공장 잡음이 섞인 음성에 대해서는 상대적으로 낮은 추출율을 나타내었다. F16잡음이 섞인 음성의 시작점은 82.6%, 끝점은 86.4%, 공장잡음이 섞인 음성의 시작점은 63.6%, 끝점은 89%의 추출성능을 나타내었다. F16잡음과 공장잡음이 섞인 음성신호는 음성신호와 비슷한 주파수대역에 많은 에너지를 가지고 있는데, 이것이 추출율을 낮게하는 중요한 요인으로 분석된다. 그림 3.3은 잡음에 대한 추출 성능의 백분율이며, 그림 3.4는 잡음이 없는 음성과 백색 잡음이 섞인 음성 파형에서 음성구간을 찾은 예이다.



< 그림 3.3 > 잡음 음성의 데이터 분포정추출율



< 그림 3.4 >잡음이 없는 음성과 잡음음성의 음성구간 추출 예

4. 결 론

본 논문에서는 *SFM*과 주파수 영역에서의 에너지 평균을 이용하여 잡음이 섞인 음성신호에 대해서 음성구간을 찾는 새로운 방법을 제안하였다. 잡음이 없는 음성에 대해서는 우수한 추출성능을 나타내었으며, 신호 대 잡음의 비(SNR)가 20dB인 잡음이 섞인 음성에서도 비교적 강인한 추출성능 나타내었다. 제안한 검출 알고리즘은 신뢰성, 강인함, 정확성, 적응성과 잡음에 대한 사전 정보 없이 잡음평

가를 해서 음성구간을 찾는다는 점에서 장점을 가지고 있다. 그러나 신호 대 잡음비가 낮은 음성신호에 대해서는 자음성분과 잡음을 정확히 구별하기 어려웠다. 따라서 잡음과 자음에 대한 변별력이 높은 파라미터와 이를 이용한 알고리즘에 대한 연구가 더 필요하다.

참고 문헌

[1] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," IEEE ASSP Mag,. vol. 29, pp.777-785, 1981.

[2] C. Tsao and R. M.. Gray, "An endpoint detector for LPC speech using residual error look-ahead for vector quantization applications," in Proc. ICASSP-84, pp.18b. 7. 1-4. 1984.

[3] J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov medels to automatic speech endpoint detection," Computer, Speech, Language, vol. 2, pp.321-341, 1987.

[4] B. H Juang. "Speech Recognition in adverse enviroments," Computer Speech and Language. pp.275-294 1991.

[5] J. D, Johnston, "Transform coding of audio signal using perceptual noise criteria," IEEE J. Select. Areas Commun., vol 6. pp.314-323, Feb. 1988.

[6] J. C. Junqua, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," IEEE Trans. ASSP. pp.406-412, Aug. 1994

[7] P. T. Brady, "A Technique for Investigating the On-Off Patterns of Speech", BSTJ, Jan., p. 1. 1965.