대역에너지를 이용한 잡음음성의 끝점검출 알고리즘

°박 기 상, 석 수 영, 정 호 열, 정 현 열
영남대학교 대학원 정보통신공학과

An Endpoint Detection Algorithm for Noise Speech using Band Energy

°Ki-Sang Park, Su-Young Suk, Ho-Youl Jung, Hyun-Yeol Chung
Department of Information and Communication Eng., Yeungnam University

요 약

음성인식 시스템의 실용화를 위해서 우선적으로 해결되어야 될 문제중 하나로 잡음환경하에서의 끝점검출을 들수 있다. 잡음이 존재하지 않는 환경에서는 기존의 에너지 파라미터만으로도 어느정도 신뢰성있는 끝점 구간을 검출할 수 있으나 도심 소음과 같은 실제 잡음환경하에서는 대부분 좋지 않은 결과를 보인다. 본 논문에서는 도심환경의 배경잡음을 제거하는 방법으로 입력되는 음성에 대하여 주변소음에 의해 손상된 음성스펙트럼의 크기 성분만을 제거하는 전처리 기법인 Bark scale에 기반한 스펙트럼 차감법을 사용하고, 인간의 청각특성을 고려하여 음성의 주파수 대역을 3개의 대역으로 분리한 후, 대역별로 세밀한 에너지 문턱치값을 설정하여 음성의 끝점을 탐색하는 방법을 제안한다. 제안한 방법의 유효성을 확인하기 위해 실제 사무실 및 지하철역 등의 잡음환경하에서 녹음된 데이터베이스를 이용하여 끝점검출을 수행한 결과 기존의 에너지와 영교차율을 이용한 방법에 비해 평균 46%의 오차율 감소와 대역에너지만을 사용한 경우에 비해 평균 17%의 오차율 감소를 나타내어 제안한 방법의 유효성을 확인할수 있었다.

1. 서 론

음성인식시스템을 상용화하기 위해서는 아직 많은 문제점들을 해결해야 하지만 그 중에서도 정확한 음성구간 검출에 관한 문제가 인식에 커다란 과제로 남아있다. 이는 음성구간의 정확한 검출은 인식률 향상에 큰 영향을 끼치기 때문이다.

잡음환경에 강한 음성인식기의 구현을 위해서는 인식 시스템에 전처리 단계를 두어 음성에 포함된 잡음을 제거함으로써 음성끝점 검출성능향상 및 이로 인한 음성인식시스템의 성능저하를 최소화할 수 있다. 잡음제거를 위한 연구 예를 살펴보면 잡음보상법으로는 Boll등[1]에 의한 SS(Spectral Subtraction)법, Hermansky등[2]에 의한 켑스트럼 고역통과필터링에 의한 RASTA(RelAtive SpecTrAl)처리법, 켑스트럼 영역에서의 CMN(Cepstral Mean Subtraction)처리방법 등이 있으나 대부분의 연구가 부가잡음, 채널왜곡을 분리해 처리하고 있다. 일반적으로 사용하는 음성의 끝점검출 파라미터로는 단구간 에너지와 영교차율 등이 있다. 여기서 단구간에너지는 음성구간과 묵음구간을 구분하는데 이용되며, 영교차율은 음성의모음과 자음구간을 구분하는데 이용된다. 하지만 음성끝점검출에 있어서 배경잡음이 존재하는 경우 높은 SNR에서는 끝점검출이 비교적 효과적으로 수행될 수 있지만 SNR이 낮아질수록 효과적으로 음성을 검출하지 못하는 문제점이 있다.

따라서 본 연구에서는 음성을 효율적으로 검출하기 위해 음성의 배경잡음을 Bark Scale에 기반을 둔 스펙트럼 차감법으로 제거하고 단구간 에너지와 영교차율을 이용하는 대신 SNR에 근거한 대역에너지를 세분화하여 음성의 끝점을 탐색하는 방법을 도입한다.

2장에서는 음성신호의 잡음제거를 위해 사용된 스펙트럼 차감법에 대해 살펴보고, 3장에서는 3개의 세분화된 대역에너지를 이용한 끝점 검출 알고리즘에 대해 기술한다. 4장에서는 제안한 방법의 실험 결과를 분석한후 5장에서는 결론 및 향후 연구방향에 대해 기술한다.

2. 잡음이 부가된 음성의 처리

음성신호에 배경잡음 혹은 채널왜곡이 존재하는 경우 잡음을 제거하는 방법으로, 음성이 없는 구간에서 부가

잡음을 추정하여 열화된 음성에서 추정된 잡음을 제거하는 스펙트럼 차감법(SS)[3][4], RASTA, CMN이 대표적이다. 이 중에서 SS법은 비교적 간단하면서도 어느정도의 잡음 성분을 효과적으로 제거하는 것으로 잘 알려져 있다. 이하 이에 대해 간략한다.

2.1 스펙트럼 왜곡

실제 환경에서의 음성신호에는 다양한 종류의 부가잡음 및 채널왜곡이 혼입된다. 음성의 왜곡은 주변소음이 음성에 산술적으로 더해진다는 가정과 음성을 인지하는 청각의 특성은 음성의 주파수 성분별 위상정보보다는 크기정보에 더 많은 영향을 받는다는 가정하에 다음을 고려할 수 있다. 잡음이 섞인 음성신호 $y(m)$은 식(1)과 같이 표현된다.

$$y(m)=x(m)+n(m) \tag{1}$$

여기서, $x(m)$ 은 잡음이 섞이지 않은 원래의 음성신호이고, $n(m)$은 부가잡음이다. 일반적으로 $x(m)$과 $n(m)$은 상관성이 없으며(uncorrelated) $n(m)$은 정재적(stationary)이거나 $x(m)$에 비해 매우 천천히 변화한다고 가정한다. $Y(\omega)$, $X(\omega)$, $N(\omega)$ 를 신호 $y(m)$, $x(m)$, $n(m)$ 의 단구간 전력스펙트럼밀도(power spectral density)라 하면, 신호와 잡음은 서로 상관성이 없으므로 각각의 주파수대역 ω_k에 대해 다음과 같은 관계식이 성립한다.

$$Y(\omega_k)=X(\omega_k)+N(\omega_k) \tag{2}$$

여기서 ω_k 는 k 번째의 subband를 나타낸다. 따라서, 특정 프레임에서 잡음 음성의 전력스펙트럼 밀도 $\hat{Y}(\omega)$가 구해지고 잡음의 전력 스펙트럼 밀도 $\hat{N}(\omega)$가 추정되면, 식(2)로부터 잡음이 제거된 음성의 전력 스펙트럼 밀도는 다음과 같이 추정할 수 있다.

$$\hat{X}(\omega_k)=\hat{Y}(\omega_k)-\hat{N}(\omega_k) \tag{3}$$

2.2 스펙트럼 차감법(Spectral Subtraction)

대부분의 음성인식시스템에서 사용하는 음성특징파라미터들은 스펙트럼의 크기정보만을 사용하므로 스펙트럼 차감법은 음성인식시스템에 효과적인 방법이라 할 수 있다. 음성신호 $s(k)$에 잡음신호 $n(k)$가 더해져 있을 때 잡음에 의해 손상된 음성신호 $x(k)$에 대해서 스펙트럼 영역에서 식(4)와 같이 차감할 수 있다. 여기서 $S(\omega)$, $X(\omega)$, $\mu(\omega)$ 는 각 신호들의 주파수 성분을 나타내는 것이다.

$$|S(\omega)| = |X(\omega)|-\mu|(\omega)| \tag{4}$$
$$\mu|(\omega)| = \frac{1}{M}\sum_{i=1}^{M}|N_i(\omega)|$$

여기서, $S(\omega)$는 음성신호의 스펙트럼, $X(\omega)$는 잡음에 의해 손상된 음성신호의 스펙트럼, $N_i(\omega)$는 i프레임에서의 잡음스펙트럼을 나타낸다. 소음이 제거된 음성신호의 시간축파형을 얻기 위해서는 위상정보가 중요한 변수가 아니라는 사실을 이용하여 $\theta_s(\omega)$를 손상된 음성의 위상정보 $\theta_x(\omega)$로 대치한다. 결과적으로 개선된 음성의 주파수영역에서의 표현은 식(5)와 같이 나타낼 수 있으며, 그림2에 스펙트럼 차감법의 전체 처리 과정을 나타낸다.

$$S'(\omega) = [X(\omega)-\mu(\omega)]\exp(-j\theta_x(\omega)) \tag{5}$$

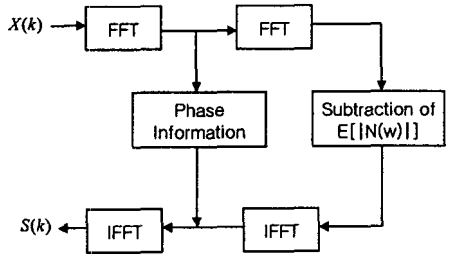


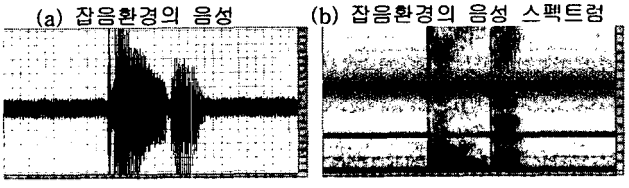
그림 2. 스펙트럼 차감 처리 과정

이와 같은 스펙트럼 차감법은 아래와 같은 두 가지 조건을 만족하는 환경하에서 사용할 수 있다.

1) 배경잡음의 스펙트럼형태를 미리알고 있거나 소음의 스펙트럼을 추정하기에 충분한 묵음구간이 주어져야 한다.
2) 배경소음은 최소한 부분적으로 연속적인 특성을 가져야 하며 통계적 특성이 서서히 변화하는 환경에서는 음성이 존재하는 구간과 소음만이 존재하는 구간을 검출할 수 있는 방법이 필요하다.

따라서 본 논문에서는 과거의 복수개의 프레임으로부터 구한 단구간 스펙트럼의 이동 평균으로 추정된 잡음레벨을 사용하는 연속 스펙트럼 차감법을 이용한다. 그림 3은 스펙트럼 차감법을 이용한 잡음음성의 처리예를 나타내고 있다.

(a) 잡음환경의 음성 (b) 잡음환경의 음성 스펙트럼



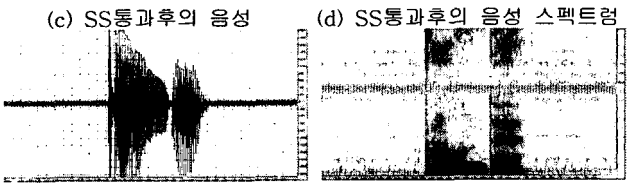


그림2. 스펙트럼 차감법을 이용하기 전후의 잡음음성

3. 끝점 검출 알고리즘

기존의 음성구간을 검출하기 위한 끝점 검출 알고리즘은 영교차율, 에너지등을 이용한 방법과, 잡음의 특성을 이용한 방법들도 연구되고 있다. 일반적으로 단구간 에너지와 영교차율[5][6][7]을 사용하는 끝점검출 방법에서는 먼저 단구간 에너지를 사용하여 음성의 안정적인 구간을 찾은 후에 영교차율을 이용하여 검출된 안정적인 음성구간 양단의 자음 부분을 교정하여 최종적인 음성구간을 결정한다. 그러나, 잡음이 포함된 실제환경에서는 단구간 에너지만으로 안정적인 음성구간을 검출하는 것이 어려우며, 영교차율을 이용하여 음성구간의 처음과 끝 부분에서 자음과 모음을 구분하는 데에도 어려운 점이 있다. 따라서, 본 논문에서는 음성의 대역을 3개의 주파수 대역으로 세분화 한후 5ms 단위의 대역별 에너지 문턱치를 이용하여 시작점과 끝점을 개략적으로 검출한 후 전향탐색과 후향탐색을 이용하여 정교한 음성 구간을 검출하는 방법을 이용한다. 그림3에 이 과정을 나타낸다.

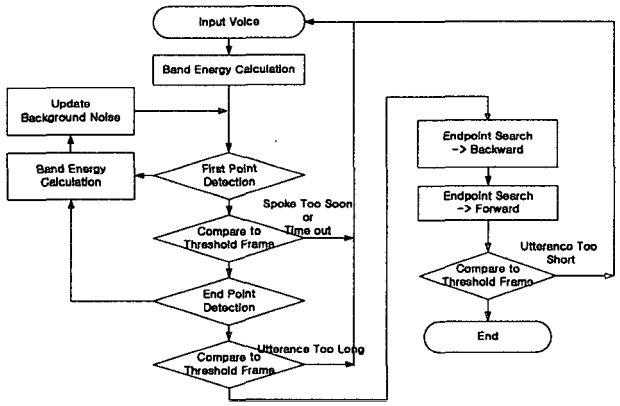


그림 3. 전체음성 끝점검출 알고리즘의 구성도

3.1 전향 탐색과 후향 탐색

전향 탐색에서는 입력되는 음성에서 대략적인 음성구간을 설정하는 단계로써 SNR에 근거한 상위문턱치와 비교하여 설정한다. 상위와 하위문턱치(Threshold)와의 비교는 Bark scale에 근거한 3개의 대역에 대하여 식(3)과 같이 각각 비교한다.

$$\frac{BE^{i}_{inst}(n)}{N^{i}_{est}(n-1)} > Thresh_{high}(SNR^{i}_{est}(n-1)) \quad (3)$$

- $i=0,1,2$

여기서, n은 음성의 입력 프레임, $BE^{k}_{inst}(n)$은 일시적인 대역에너지, $N^{k}_{est}(n)$은 배경잡음, $S^{k}_{est}(n)$은 신호에너지 추정값, $SNR^{k}_{est}(n)$은 대역 k와 프레임 n의 신호대 잡음비 추정값을 나타낸다. 앞에서 기술한 3개의 대역은 인간의 청각특성을 고려하여 각 에너지 대역의 총합이 균등하게 분포할 수 있도록 분할하였다. 8K 음성의 경우 대역의 폭은 다음과 같다.

표 1. 대역별 주파수 구간

대역	주파수 대역
Band 1	175Hz ~ 889Hz
Band 2	889Hz ~ 1797Hz
Band 3	1797Hz ~ 4000Hz

3.2 잡음의 추정과 문턱치

입력 음성의 SNR추정은 10ms의 해밍창을 통해 5ms의 프레임 단위로 이동하면서 추정한다. 입력된 초기 15ms를 배경잡음으로 정의하고 이후의 프레임에 대하여 SNR을 추정한다. 입력음성과 배경잡음 에너지는 식 (4)와 같다.

$$S^{k}_{est(n)} = \max[f^{s}_{decay} * S^{k}_{est}(n-1),\ BE^{k}_{inst(n)}]$$
$$N^{k}_{est(n)} = \max[f^{n}_{decay} * N^{k}_{est}(n-1),\ BE^{k}_{inst(n)}] \quad (4)$$

- $k=0,1,2$

$$f^{n}_{decay} = 1.03 \text{ for } non-speech frame$$
$$= 1.01 \text{ for } speech\ frame$$
$$f^{s}_{decay} = 0.97$$

$$SNR^{k}_{est}(n) = 10 * \log 10\left(\frac{S^{k}_{est}(n)}{N^{k}_{est}(n)}\right) \quad (5)$$

추정된 SNR을 근거로 설정된 상위 문턱치는 전체 입력음성의 잡음구간과 음성구간의 에너지에 대한 평균으로 실험에 의해 결정한다. 만약 3개의 대역과 분석프레임에서 계산한 SNR이 3개의 대역에 대응되는 문턱치와 비교하여 높은 SNR을가지면 그 프레임을 음성으로 간주한다.

4. 실험 및 고찰

본 논문에서 제안한 Bark scale에 기반을 둔 스펙트

럼 차감법을 적용시킨 후, 3개의 대역 에너지로 분리하여 탐색하는 음성끝점검출 알고리즘의 유효성을 확인하기 위해 잡음환경에서 채록한 600명에 의한 20단어 발성 중 5명의 발성을 대상으로 기존의 에너지와 영교차율을 이용한 방법과 비교하였다. 이 데이터베이스는 일반 사무실환경, 시내 도심환경, 지하철 환경이 포함되어 있다.

실험에서는 수작업으로 작성한 Labeling정보를 기준으로 에너지와 영교차율을 이용한 방법, 대역에너지만을 이용한 방법, 스펙트럼 차감법과 대역에너지를 이용한 경우를 비교하였다.

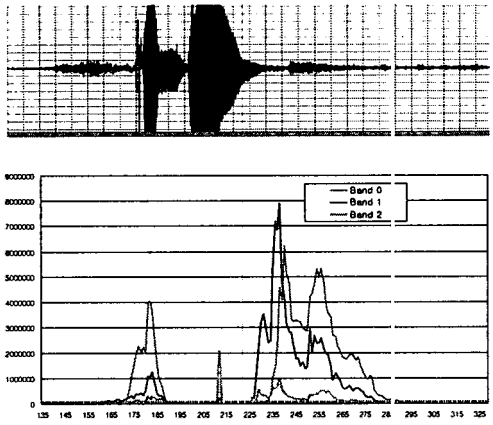


그림 4. SS통과후 음성과 대역 에너지의 예

그림 4는 '변경'이라는 발성에 대해 스펙트럼 차감법을 이용하여 전처리한 후 각 프레임별로 계산된 3개의 대역에너지를 검출한 일예이다. 대역별로 에너지가 균등하게 나타날 수 있도록 고려되어 있음을 확인할 수 있다.

표 2에서와 같이 스펙트럼 차감법과 대역에너지를 이용한 경우 에너지와 영교차율을 이용한 경우와 대역에너지만을 이용한 경우에 비해 평균 각각 46%와 17%의 오차율 감소를 나타내어 제안된 방법의 유효성을 확인할 수 있었다.

표 2. 에너지와영교차율, 대역에너지, 스펙트럼 차감법과 대역에너지를 이용한 경우의 시작점과끝점 평균오차

	시작점	끝점
Energy & Zero Crossing rate	39 msec	58 msec
Band Energy	28 msec	37 msec
SS+Band Energy	22 msec	32 msec

5. 결 론

음성인식 시스템의 실용화를 위해서는 도심소음과 같은 실제환경에서 음성의 끝점 검출이 필요하나 기존의 에너지 파라미터, 영교차율 등을 이용한 방법으로는 만족할 만한 성능을 나타내지 못하였다. 이를 개선하기 위해 본 논문에서는 배경잡음을 제거하는 방법으로 Bark scale에 기반한 스펙트럼 차감법을 이용하여 전처리한 후, 인간의 청각특성을 고려하여 음성의 주파수 대역을 3개의 대역으로 분리하여, 대역별로 세밀한 에너지 문턱치값을 설정하여 음성의 끝점을 탐색하는 방법을 도입하였다. 제안한 끝점 검출 알고리즘의 유효성을 확인하기 위해 기존의 단구간 에너지와 영교차율을 이용한 음성의 끝점 추출결과와 비교한 결과 기존의 에너지와 영교차율을 이용한 방법에 비해 평균 46%의 오차율 감소와 대역에너지만을 사용한 경우에 비해 17%의 오차율 감소를 나타내어 제안한 방법의 유효성을 확인할수 있었다.

향후 현재까지 검토한 결과를 바탕으로 여러 가지 환경에서 채록한 음성자료를 이용하여 본 논문에서 도입한 끝점검출알고리즘의 성능향상 실험을 통하여 다양한 잡음환경에 강건한 음성인식 시스템을 구축하고자 한다.

참고 문헌

[1] L.R Rabiner, B.H. Juang, "Fundamentals of Speech Recognition," Prentice Hall, 1993.

[2] Shingo Kuroiwa, Masaki Naito, Seiichi Yamamoto and Norio Higuchi, "Robust speech detection method for telephone speech recognition system," Speech Commun. vol.27, pp.135-148,1999.

[3] J.A.N. Flores, S.J. Young, "Continuous Speech Recognition on Noise using Spectral Subtraction and HMM Adaptation," ICASSP, pp.409-412, 1994.

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustic., Speech Signal Processing, vol. ASSP-27, No. 2, pp.113-120, April 1979.

[5] L.R. Rabinar and M.R. Sambur, "An algorithm for Determining the End points of Isolated Utterances," Bell system Tech. J. vol. 54 No. 2 pp.297-315, 1975

[6] Yiying Zhang, Xiaoyan Zhu, Yu Hao, Yupin Luo, " A robust and fast endpoint detection algorithm for isolated word recognition," Intelligent Processing Sysems, 1997. ICIPS '97. 1997 IEEE International Conference on, vol.2 1997, vol.2, pp.1819-1822.

[7] L.F.Lamel, L.R.Rabiner, A.E.Rosenberg, and J.G.Wilson, "An Improved endpoint detector for isolated word recognition," IEEE ASSP Mag., vol 29, pp.777-785, Aug. 1981.