1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

변형된 Teager 에너지에 기초한 음성끝점검출 알고리듬에 관한 연구

A Study on the Endpoint Detection Algorithm Based on a Modified Teager Energy

이재한, 백승준, 성굉모
서울대학교 전기공학부

Jachan Lee, SeongJoon Baek, Koeng-Mo Sung
School of Electrical Engineering, Seoul National University

E-Mail : ljh@acoustics.snu.ac.kr

요 약

본 논문에서는 변형된 Teager에너지를 이용하여 음성의 끝점을 검출하는 알고리듬을 제안하였다. 기존의 방법에서는 대부분 음성신호의 에너지와 영교차율을 이용하거나 이 파라미터들과 함께 다른 여러 파라미터들을 사용하여 끝점을 검출하였다. 여러 파라미터들을 사용하는 알고리듬의 경우 계산량이 많아지게 되는데, 이에 비해 본 논문에서는 하나의 파라미터를 이용하기 때문에 계산량이 기존의 알고리듬보다 적다. 그리고 이 알고리듬에서 사용한 변형된 Teager 에너지는 음성신호의 진폭뿐만 아니라 주파수까지 고려한 파라미터이다. 일반적으로 마찰음은 진폭이 작아 검출하기가 어려운데, 본 논문에서는 이러한 마찰음에 대해 실험을 했고, 그 결과를 통해 제안한 알고리듬이 기존의 다른 여러 알고리듬보다 성능이 우수하다는 것을 확인할 수 있었다.

I. 서 론

일반적으로 음성인식에 있어서 인식률을 저하시키는 요인 중에 큰 비중을 차지하는 것이 바로 음성의 끝점검출이다. 또한, 음성신호의 시작점과 끝점을 정확하게 찾지 못함으로 인해 그만큼 계산해야 할 양이 많아지게 되어 처리시간도 길어진다. 따라서 음성신호의 시작점과 끝점을 정확하게 찾아내는 것은 음성인식에 있어서 인식률을 증가시킬 뿐만 아니라 처리시간도 단축시킬 수 있다.

대부분의 알고리듬은 음성신호의 에너지와 영교차율의 검출에 토대를 두고 있다[3, 4, 5, 6, 7, 8]. 유성음과 무성음, 그리고 주변잡음을 구분시킬 수 있는 요소 중의 하나가 에너지이고 음성신호의 주파수성분을 나타내주는 것이 영교차율이다. 대개 무성음의 경우가 영교차율이 가장 높고(특히 마찰음의 경우가 더 높다), 유성음의 경우는 영교차율이 낮다. 주변잡음의 영교차율은 무성음과 유성음의 중간정도에 해당된다. 그리고 에너지의 경우는 대개 유성음, 무성음, 주변잡음 순으로 에너지가 높지만 주변환경에 따라서 무성음이 주변환경에 묻힐 수도 있다. 이런 경우는 무성음보다 주변잡음의 에너지레벨이 더 높다. 이 두 가지 파라미터를 주로 이용하여 기존의 알고리듬은 음성의 끝점을 검출한다.

음성신호를 모델링함에 있어서 Teager는 음성신호의 에너지 계산에 대한 새로운 알고리듬을 고안해 냈고[1], Kaiser는 이 알고리듬을 Teager's Energy Algorithm으로 제안하였다. 본 논문에서는 이 알고리듬에 기초하여 음성의 시작점과 끝점을 검출한다.

II 장에서 Teager에너지에 대해 간단히 살펴본 다음, III장에서 음성의 끝점검출을 위해 변형된 Teager에너지 파라미터를 제안하고 IV장에서는 이를 이용한 알고리듬의 성능을 기존의 여러 알고리듬과 비교했다.

II. Teager 에너지 알고리듬

뉴턴의 운동 법칙에 있어서 에너지는 위치에너지(potential energy)와 운동에너지(kinetic energy)의 합으로 표현된다. 오실레이션 운동을 하는 물체의 신호가 $x(t)=A\cos(\omega t+\phi)$ 으로 주어진다면 에너지는 다음과 같이 된다.

$$E=\frac{1}{2}kx^2+\frac{1}{2}mx'^2, \qquad \omega=\sqrt{\frac{k}{m}}$$

여기에서 m은 오실레이션 운동하는 물체의 질량이고 k는 탄성계수이다. 위 식에다 x(t)를 대입하면 다음과 같은 식을 얻을 수 있다.

$$E = \frac{1}{2} m\omega^2 A^2$$

or $$E \propto A^2\omega^2$$

결과적으로 여기에서 구한 에너지는 진폭뿐만 아니라 주파수에도 비례함을 알 수 있다. Teager는 이 개념을 음성신호의 에너지를 계산하는데 이용했는데, 그 방법을 살펴 보면 다음과 같다.

$$x_n = A\cos(\Omega n + \phi)$$

으로 주어지는 샘플에 대해서 생각해 보자. 그렇다면

$$x_{n+1} = A\cos[(n+1)\Omega + \phi]$$

$$x_{n-1} = A\cos[(n-1)\Omega + \phi]$$

이다. 이 식들을 적당히 변형시키면 다음과 같은 식을 얻을 수 있다.

$$x_n^2 - x_{n+1}x_{n-1} = A^2\sin^2(\Omega)$$

그리고 Ω 의 값이 작으면 $\sin(\Omega) \approx \Omega$ 가 성립하므로

$$x_n^2 - x_{n+1}x_{n-1} \approx A^2\Omega^2$$

이다. 따라서 위의 식을 에너지 파라미터로 사용할 수 있다. 즉,

$$TE(n) = x_n^2 - x_{n+1}x_{n-1} = A^2\sin^2(\Omega) \approx A^2\Omega^2$$

이다. 이 에너지 파라미터를 이용하는 알고리듬을 Teager's Algorithm이라 한다.

Ⅲ. 제안된 알고리듬

앞에서 살펴본 바와 같이 Teager에너지 파라미터는 음성신호의 진폭뿐만 아니라 그 신호의 주파수까지 고려한다는 것을 확인할 수 있다. 무성음의 경우 마찰음이나 파열음은 진폭이 작은 반면에 5kHz이상의 주파수에서도 에너지가 분포되어 있다. 따라서 이 두 가지를 같이 고려해 주는 Teager에너지를 사용하면 무성음의 경우도 높은 레벨의 에너지 값을 얻을 수 있다.

무성음과 주변잡음을 비교해 보면 무성음의 경우가 영교차율이 더 크게 나타나는데, 이것에서 무성음의 주파수가 더 높다고 볼 수 있다. 따라서 이 주파수 성분을 Teager가 제안한 에너지 파라미터보다 더 키워준다면 무성음의 에너지 파라미터값이 주변잡음의 그것에 비해서 더 크게 나타날 것이다. 본 논문에서는 Teager에너지 파라미터를 변형해서 주파수 성분을 더 많이 고려해 줄 수 있는 파라미터를 제안한다. 이 변형된 파라미터는 다음과 같이 유도된다.

즉,

$$x_n = A\cos(\Omega n + \phi)$$

$$MTE(n) = x_n^2 - x_{n+k}x_{n-k} = A^2\sin^2(k\Omega) \approx A^2(k\Omega)^2$$

이다. 이 식에서 보면 k가 변함에 따라 주파수 성분이 그만큼 더 커진다는 것을 알 수 있다. 본 논문에서는 여러 가지 k값을 가지고 실험을 해 본 다음 k값을 40으로 설정했고 이 에너지 파라미터를 변형된 Teager 에너지(Modified Teager Energy)라 부르기로 한다.

그리고 다음은 이 에너지 파라미터를 사용하여 음성의 시작점과 끝점을 찾는 알고리듬을 도시한 것이다.

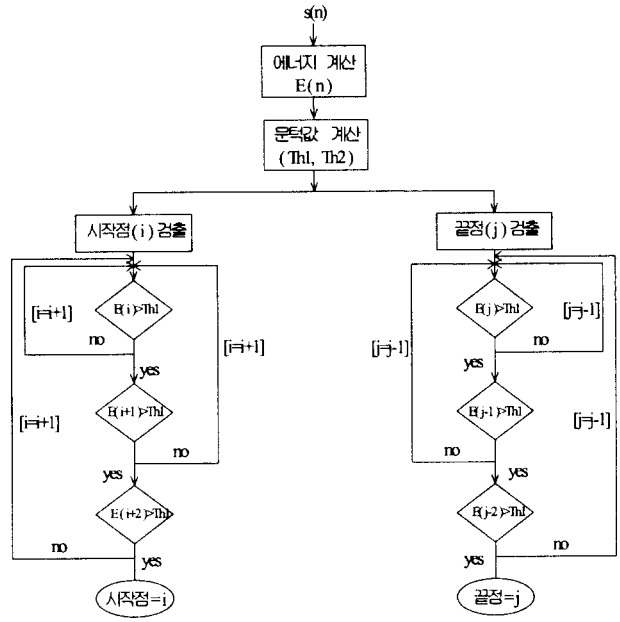

그림 1. 제안된 알고리듬의 블록도

처음에 시작점과 끝점을 찾는 방법은 Sambur[3]가 제안한 알고리듬을 이용하였고 최종적인 시작점과 끝점을 찾는 것은 시작점 뒤 2프레임과 끝점 전 2프레임을 조사한 뒤 모두 문턱값보다 클 때 그 프레임을 시작점과 끝점으로 선택했다. 이것은 변형된 Teager에너지 파라미터 값을 통계적으로 조사해 본 결과로 얻은 것이다. 위의 그림에서 문턱값은 두 개를 사용했는데, 이것은 Sambur가 제안한 알고리듬의 개념과 같지만, 기준은 그것과 다르다. 본 논문에서는 처음 10프레임과 마지막 10프레임을 silence 구간으로 설정했는데, 첫 10프레임과 끝 10프레임의 에너지값 평균과 표준편차의 합을 Th1으로 설정하고 Th1의 2배를 Th2로 설정했다.

Ⅳ. 실험결과

본 논문에서 사용한 음성데이터베이스는 원광대 휴먼 인터페이스 연구실에서 제작한 DB이다. 이 음성데이터베이스의 음성은 방음부스에서 Sennheizer HMD224X를 사용하여 녹음되었으며, 발성된 데이터는 디지털 오디오 테이프에 저장되었다. A/D는 PC환경에서 실시하였고, AD/DA Module은 KAY CSL4300B를 사용하였다. 그리고 16kHz로 샘플링하고 16bits로 양자화하였다. 이 실험에서 사용한 음성은 에너지레벨이 일반적으로 낮지만 주파수 성분이 강한 마찰음(fricative : ㅅ, ㅎ, ㅈ, ㅊ)으로 시작하는 단어이고, 화자는 남녀 아나운서 각각 1명과 일반인 8명(남:4명, 여:4명)을 대상으로 했으며, 각 화자당 10개의 단어, 총 100개의 단어에 대해서 실험을 했다. 일반적으로 마찰음은 다른 유·무성음에 비해 진폭이 작아 끝점검출이 힘들기 때문에 본 논문에서 실험대상으로 삼았다.

표 1. 음성끝점검출 실험 조건

sampling frequency	16,000 Hz
quantization	16bits
frame length	160 sample(10ms)
overlap	no
음성데이터 수	100
시료	마찰음으로 시작하는 단어
silence frame	시료 처음과 끝의 10frame

본 실험에서는 제안된 알고리듬(A) 이외에 Frame based Teager Energy Measure를 사용한 알고리듬(B)[1], Teager에너지를 이용한 알고리듬(C)[1], 그리고 개선된 Sambur알고리듬(D)을 사용했고, 여기에서 얻은 결과를 eye-detection으로 얻은 결과와 비교했다. 알고리듬 B의 에너지는 음성신호의 파워스펙트럼을 구한 다음 각 포인트에서의 값에다 그 포인트에서의 주파수의 제곱을 곱해서 얻은 값을 한 프레임동안 더해서 얻은 파라미터의 제곱근값이다. 그리고 개선된 Sambur 알고리듬은 기존의 Sambur 알고리듬을 변형시키서 좀 더 나은 결과를 얻을 수 있게 한 알고리듬이다. error계산은 eye-detection으로 얻은 결과와 앞에서 언급한 알고리듬에서 구한 결과와의 차이를 평균한 것인데, 그 결과가 다음의 표에 있다.

표 2. 제안된 알고리듬(A)의 Error

	시작점	끝점	평균
Error(ms)	7.1	14.9	11.0

표 3. Frame based Teager Energy Measure를 사용한 알고리듬(B)의 Error

	시작점	끝점	평균
Error(ms)	11.4	31.9	21.7

표 4. Teager 에너지를 이용한 알고리듬(C)의 Error

	시작점	끝점	평균
Error(ms)	40.4	35.6	38.0

표 5. 개선된 Sambur 알고리듬(D)의 Error

	시작점	끝점	평균
Error(ms)	23.7	37.9	30.8

위의 표를 살펴 보면, 제안된 알고리듬이 기존의 여러 알고리듬에 비해 약 10ms에서 20ms 즉, 1 프레임에서 2 프레임 정도 더 적은 error를 갖고 있음을 알 수 있다. 이것은 음성인식이나 분석에 있어서 그 만큼 인식률과 계산량을 개선시킬 수 있음을 나타내주는 것으로 볼 수 있다. 그리고 실험결과에 따르면 시작점검출에서 보다 끝점검출에서 error가 더 많이 나오는 것을 확인할 수 있는데, 이는 실험의 시료가 주파수성분이 강한 마찰음을 대상으로 한 것과 관련이 있음을 보여 주는 것이다. 음성신호의 시작이 처음 10프레임 안에서 시작되면 error의 값이 커지는데, 이것은 처음 10프레임의 silence 구간으로부터 정확한 문턱값을 제대로 구할 수 없기 때문이다. 이런 경우를 제외하고는 제안된 알고리듬이 시작점과 끝점을 꽤 정확하게 찾았다. 그리고 음성시료 자체에 사람의 입에서 나오는 미립소리같은 잡음이 섞여 있었기 때문에 기존의 알고리듬의 결과가 기대했던 것에 비해 좋지 않았다.

Ⅴ. 결론

본 논문에서는 Teager 에너지를 변형시켜서 주파수성분을 더 많이 고려해 줄 수 있는 에너지 파라미터를 제안하고 이를 이용한 알고리듬을 마찰음의 끝점검출에 적용하여 실험을 했다. 기존의 여러 알고리듬과 비교해 본 결과, 제안한 알고리듬이 계산량이 더 적을 뿐만 아니라 성능도 우수하다는 것을 알 수 있었다. 백색잡음과 같은 환경에서 이 알고리듬의 성능이 얼마나 되는지에 대해 실험해 볼 필요가 있고, 마찰음이 아닌 파열음(plosive)이나 기타 다른 무성음과 유성음에 대해서도 제대로 동작하는지에 대해 확인해 볼 필요가 있다. 그리고 문턱값을 최적화하는 방법도 계속 고려해 보아야 할 사항이다.

감사의 글 : 이 실험에서 사용할 수 있도록 음성 DB를 제공해 주신 원광대 이용주 교수님께 감사드립니다.

< 참고문헌 >

[1] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," *Proc. ICASSP-93*, pp. 732-735, 1993.

[2] James F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. ICASSP-90*, pp. 381-384, 1990.

[3] 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, 1975.

[4] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. ASSP*, Vol 24, No. 3 pp. 203-212, June 1976.

[5] M. Hahn and C. K. Park, "An improved speech detection algorithm for isolated korean urtterance," *Proc. ICASSP-92*, pp. I 525-528, 1992.

[6] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. ASSP*, Vol 29, pp. 777-785, August 1981.

[7] L. J. Siegel and A. C. Bessey, "Voiced/unvoiced/ mixed excitation classification of speech," *IEEE Trans. ASSP*, Vol 30, pp. 451-460, June 1982.

[8] M. H. Savoji, "A robust algorithm for accurate endpointing of speech signals," *Speech Communication*, Vol 8, No. 1, pp. 45-60, March 1989.