강인한 끝점 추출과 에너지 정규화

고기원, 정원용
경남대학교 정보통신공학과

Robust Endpoint Detection and Energy Normalization

Gi-won Go, Won-yong Chong
Information & Telecommunication Engineering, Kyungnam University

요 약

자동 음성 인식(ASR) 시스템에, 끝점 추출과 에너지 정규화는 중요한 역할을 하게 된다. 그러나 낮은 SNR이나 nonstationary 환경에서, 기존 방법은 끝점 추출과 에너지 정규화에 있어서 자주 실패하게 되며, ASR을 급격히 열화시키곤 한다.

ASR을 수행하기 위해, 최적의 필터에 3상태 천이도를 사용하고, 필터는 정확성과 강인함을 확실히 하기 위해 여러 이론들을 이용하여 설계하였고 여러 가지 잡음이 있는 음성 신호환경에서 거의 일정한 응답을 주었다.

검출된 끝점은 곧바로 에너지 정규화에 적용된다.

실험 결과는 제안된 알고리즘이 낮은 SNR에서 에러율을 크게 감소시키고 있다는 것을 보여준다.

Ⅰ. 서 론

음성 인식은 음성과 침묵과 여러 잡음으로 이루어진 발화문 신호를 처리해야 한다. 끝점 추출은 수 십년동안 연구 되어왔고, 다양한 알고리즘이 개발되어져 왔다. 일반적으로 여러 조건이나 상황에 따라 각기 서로 다른 여러 알고리즘을 필요로 하게 된다.

끝점 추출은 정확성과 빠른 계산 속도 등의 여러 이유들로 ASR 시스템의 성능에 크게 영향을 끼치므로 중요하다. 음성 인식에 있어서, 잡음은 끝점 추출을 상당히 어렵게 만들고 있다. 특히, 인간이 만들어 내는 비음성 부분에 해당되는 여러 인공 산물은 음성인식을 자주 불명료하게 만든다.

최근 무선 통신이나 핸드폰 등의 SNR은 이전보다 더욱 낮고 그 잡음들은 전화선에서 나타나는 잡음보다 불안정하다. 이러한 이유로 인해, 끝점 추출의 신뢰가 떨어지기 때문에 음성 인식의 성능은 크게 떨어진다.

음성 인식에 있어서 주어진 발화문의 가장 큰 에너지 레벨을 어떤 상수나 영 이하로 만들기 위해서 에너지 정규화를 한다. 이것은 batch-mode 처리에선 문제가 되지 않으나, 실시간 처리에서 심각한 문제가 될 수 있다. 그 이유는 환경이 바뀌어지는 상황에서 짧은 데어터 버퍼를 가지고 발화문의 에너지 최대값을 측정하기 어렵기 때문이다. 실제로 이 논문의 뒷 부분에서 이 에너지 정규화가 끝점 추출과 관계있다는 것을 알려 줄 것이다. 즉, 끝점 추출이 더 정확 할수록 실시간 에너지 정규화를 더 잘 수행할 수 있게 된다.

이 논문에서 제안하는 알고리즘은 아래의 요구를 만족해야 한다. 요구 조건은 끝점의 정확한 위치, 낮은 계산적 복잡성, 빠른 응답 등이다.

이 논문의 구성은 Ⅱ에서 끝점 추출을 위한 필터를 소개하겠고, Ⅲ에서 여러 상황에서 ASR을 위한 끝점 추출과 에너지 정규화에 대한 알고리즘을 제안 하겠다. Ⅳ에선, 여러 잡음이 있는 상황에서의 실험 결과를 보여주겠으며 Ⅴ에서 이 논문에 대한 결론 부분을 서술하도록 하겠다.

Ⅱ. 끝점 추출을 위한 필터

먼저 계산의 복잡성을 줄이기 위해 입력 데이터를 1차 단구간 에너지로 변환하여 사용한다.[4]

$$g(t) = 10 \log_{10} \sum_{j=n_t}^{n_t+I-1} o(j)^2 \qquad (1)$$

o(j) 입력 데이터
t 프레임 번호
g(t) 데시벨로 나타낸 프레임 에너지
I 윈도우 길이
n_t 윈도우에서 첫 번째 입력 데이터의 번호

검출된 끝점은 자동적으로 ASR 특징으로 사용될 수 있고 계산은 음성율에서 프레임율로 감소된다.

끝점 추출의 정확성과 강인함을 위해서 에너지 특징으로부터 모든 가능한 끝점을 검출할 수 있는 검출기를 필요로 하게 된다. 또한, 이 검출기의 출력은 FA(False Acceptance)를 포함 할 수 있으므로 FAR(FA Rate)을 줄이기 위해 최종 결정 모듈을 이 검출기의 끝단에 설치해야 한다.

여기서, 우리는 하나의 발화문이 숨을 쉬는 부분을 통해 여러 음성 세그먼트로 나뉘어 진다고 가정하겠다. 각각의 세그먼트는 시작점과 끝점이라 불리는 한 쌍의 끝점들을 검출함으로써 결정 될 수 있다.

발화문의 에너지 윤곽(contour)에선, 항상 시작점을 뒤이

어서 상승하는 에지와 끝점에 앞서서 하강하는 에지가 존재하는데 이것들을 시작 에지, 끝 에지라고 부르겠다. 이들에 관해서 그림 4에 나타낸다.

끝점들이 항상 에지들과 함께 하므로, 우리는 첫 번째로 에지들을 검출한 뒤 정확한 끝점들을 찾도록 하겠다.

여기서 사용하게 될 끝점 추출 검출기는 Canny에 의해 처음 제안 되었다.[3] 그는 최적의 step-에지 검출기를 유도했으며, 그 후 Spacek이 Canny가 제안한 조건들을 조합하여 성능 측정 파라메터를 형성 시켰으며 최적의 step- 에지 검출기에 대한 솔루션을 제공하였다. 그 후 Petrou와 Kittler가 이 검출기를 ramp-에지 검출에 관한 것으로 확장 하였다. 음성에 있어서 에지들은 step-에지보다 ramp-에지에 더 가까우므로 이 논문에선 Petrou와 Kittler가 만든 필터[2]를 적용하였다.

위에서 논의된대로 에너지 윤곽에서의 시작 에지가 아래의 함수로 모델되어지는 ramp-에지라고 가정한다.

$$c(x)=\begin{cases} 1-e^{-sx}/2 & \text{for } x\geq 0 \\ e^{sx}/2 & \text{for } x\leq 0 \end{cases} \qquad (2)$$

x 프레임 번호

s 여러 에지에 대해 조절될 수 있는 상수

검출기는 moving-average 필터로써 동작될 수 있는 1차 필터 f(x)이다. [1]

이 필터는 아래의 요구조건을 만족하여야 한다.

1) antisymmetric. 즉, f(x)=-f(-x). f(0)=0.
   이것은 시작과 끝 에지에 민감하게 반응하기를 원하기 때문이다.
2) 짧은 시간 지연과 예측을 위해, 필터의 끝부분에선 영으로 부드럽게 확장되어 가야 한다.
   f(±w)=0, f'(±w)=0, f(x)=0 for |x|≥w
   여기서, w는 이 필터의 절반 길이에 해당된다.
3) 제한된 응답레벨(|k|)을 가져야 한다; $f(x_m)=k$.
   여기서, x_m은 $f'(x_m)=0$이고, x_m은 구간 (-w, 0)에서 존재한다.

Canny에 따르면 최적의 검출기의 요구조건은 좋은 SNR, 좋은 위치 검출, 잘못된 응답에 대한 최대 억제이다.[3]

이 조건들을 조합하여 만든 Spacek의 시스템 성능 파라메터는 식 (3)과 같다.

$$P=(SLC)^2=\frac{s^4}{w^2}\frac{\left|\int_{-w}^{0} f(x)[1-e^{sx}]dx\int_{-w}^{0} f(x)\,e^{sx}dx\right|^2}{\int_{-w}^{0}|f(x)|^2dx\int_{-w}^{0}|f''(x)|^2dx}. \qquad (3)$$

S SNR 비율에 관한 파라메터

L 위치(Locality) 측정 파라메터

C 잘못된 에지의 억제에 관계된 파라메터

P의 값을 최대로 하기 위해 라그랑쥐 승수법을 사용하면 최적의 필터는 다음과 같이 된다.

$$f(x)=e^{Ax}[K_1\sin(Ax)+K_2\cos(Ax)]+e^{-Ax}[K_3\sin(Ax)+K_4\cos(Ax)]+K_5+K_6e^{sx} \qquad (4)$$

K_1-K_6와 A는 필터 파라메터에 해당하고, 위의 f(x)는 필터의 절반에 해당한다.

즉, i가 정수이고 w=W일 때, 실제 필터는 대칭성을 이용하여

$$h(i)=\{-f(-W\leq i\leq 0),\ f(1\leq i\leq W)\} \qquad (5)$$

이고, 필터는

$$F(t)=\sum_{i=-W}^{W}h(i)g(t+i) \qquad (6)$$

g(·) 에너지 특징

t 현재 프레임 번호

로써 이동 평균 필터로 수행된다.

필터의 한 예를 그림 2에서 보여주고 있다. 직관적으로 시작 에지에서 정(正)응답을 가지게 되고, 끝 에지에선 음(陰)응답을 가짐을 알 수 있으며, silence에선 거의 0에 가까워짐을 알 수있다. 이 응답은 기본적으로 서로 다른 잡음에 대해 0에 가까운 응답을 가지므로 큰 변화가 없다.

Ⅲ. ASR을 위한 실시간 끝점 추출과 에너지 정규화

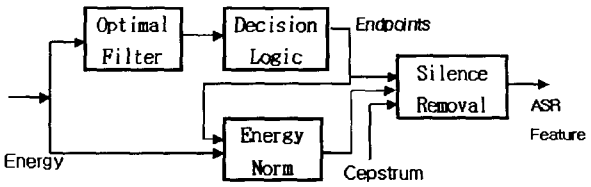

그림 3. 실시간 ASR을 위한 끝점 추출과 에너지 정규화

그림 1에서 실시간 끝점 추출을 사용한 예를 보여주고 있다. 이 시스템을 이행하는데 최적의 필터를 사용하며 최종 결정 모듈로써 3상태 천이도[1]를 사용한다. 추출된 끝점에 대한 정보는 실시간 에너지 정규화에 사용된다.

결국, 모든 silence 프레임들은 제거되고 Cepstrum을 포함하는 speech 프레임과 정규화된 에너지를 인식기에 보낸다.

1. 시작과 끝 에지 검출을 위한 필터

필터 사이즈 W=11으로 정하고, 필터 계수로써 논문 [2]을 바탕으로 s=7/W=0.6364, A=0.41s=0.2609로 정했으며, $[K_1 \ldots K_6]$는 W=7에 있는 것을 그대로 사용하였다.

성능 P는 대략 0.054604가 나온다.

필터를 h/11으로 정규화 하였으며, 그림 1에 보여주고 있고, 이 필터는 moving-average 필터로써 위의 3가지 요구

조건을 만족하여야 한다.

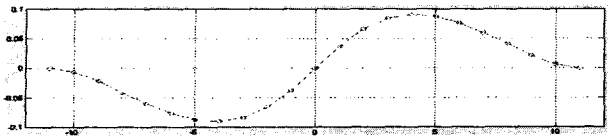

그림 5 설계된 최적화 필터 윤곽

실시간 검출을 위해, H(i)=h(i-11) 로 두면 이 필터는 H(0), H(22)=0 이기 때문에 20프레임을 예측할 수 있는 총 23 points의 필터가 된다. 따라서, 이 필터는 이동 평균 필터로써

$$F(t)=\sum_{i=2}^{20} H(i)g(t+i-2) \qquad (7)$$

로 수행된다.

F(t)의 출력은 최종 끝점 결정을 위해 3상태 천이도에서 계산된다.

2. 상태 천이도

끝점에 대한 결정은 미리 정해놓은 경계치를 가지고 F(t)의 출력과 비교함으로서 이루어진다.

실시간에서의 결정 과정의 복잡성과 순차적인 특성 때문에 3상태 천이도를 사용한다.

그림 3에서 보듯이 3상태 천이도는 silence, in-speech, leaving-speech로 구성된다. 과정은 silence나 in-speech 상태에서 시작 할 수 있으며 어느 상태에서든 끝이 날 수 있다.

본 실험에선 silence에서 시작한다는 가정에서 출발하도록 하겠다. 입력은 F(t)가 되며 출력은 시작과 끝 프레임의 프레임 번호가 된다. 상태 천이 조건은 각 상태사이에 표시되어 있으며, 각각의 천이 이벤트에 대한 출력은 괄호 안에 표시해 두었다. "Count"는 프레임 카운터에 해당하며, T_U와 T_L은 $T_U > T_L$의 조건을 가진 경계치이다. 그리고, "Gap"은 검출된 끝점과 실제 입력 신호의 끝점사이의 프레임 차이를 가리키는 정수이다.

"GO"에 대한 입력 신호의 에너지를 그림 4(a)에서 그래프로 보여주고 있으며, 그림 4(c)에서 필터의 출력부분을 보여주고 있다.

상태 천이도는 F(t)가 그림 4(b)에서 S점에 도달하기 전까지 초기 상태로 silence에 있게된다. 만약, 이 상태에서 $F(t) \geq T_U$ 가 발생하면 시작 에지가 검출되었다는 것을 의미하며 그 즉시 시작 지점을 출력하고서 in-speech 상태로 천이하게 되어 그림 4(b)에서 A점에 도달하기 전까지 in-speech 상태에 있게 된다. 만약, $F(t) < T_L$가 발생하면 leving speech 상태로 천이하게 되고 Count=0로 설정하게 된다. Counter는 B지점까지 Gap값과 경계치들과 비교하여 여러번 재설정된다. 끝지점(Counter=Gap=20)에선 그림 3에 표시된 경계치 조건을 만족하면서 끝 지점를 출력하고서 초기상태인 silence 상태로 천이하게 된다. 만약 leving speech상태에 있는 동안 $F(t) > T_U$가 발생하면 이것은 또 다른 시작 에지를 검출했다는 것을 뜻하므로 다시 in-speech 상태로 천이하게 된다.

여기서 주목해야 할 것은 경계치를 입력 신호의 에너지가 아니라 필터를 거친 출력에 대해서 적용하였다는 것이다. 필터의 출력은 어떤 잡음에 대해서도 안정적이기 때문에 검출된 끝점은 더욱 더 신뢰할 수 있다.

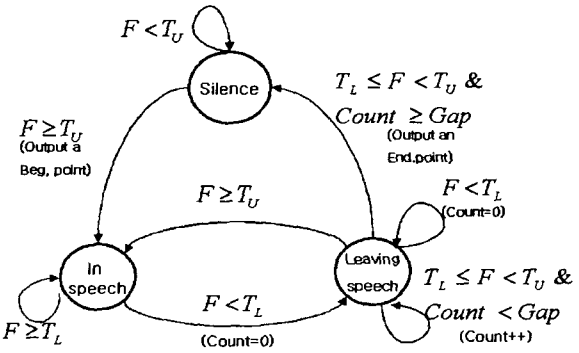

그림 4. 최종 끝점 결정을 위한 3상태 천이도

여기서 사용된 상수들(Gap, T_U, T_L)은 여러번 실험을 통해서 구해진다.

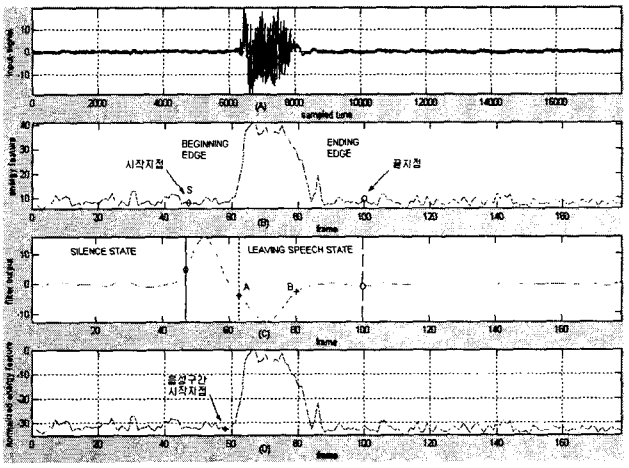

그림 6. (a) "GO"에 대한 입력 신호. (b) 프레임 단위의 에너지. (c) 필터 응답 (d) 에너지 정규화

3. 실시간 에너지 정규화

입력 신호의 최대 에너지 값을 g_{max}라고 하자.

에너지 정규화는 입력 신호의 최대 에너지 값을 0에 가까이 두기 위해 $g'(t)=g(t)-g_{max}$를 수행한다.

실시간에서 데이터를 수집하는 동안 순차적으로 g_{max}를 측정해야 하며 계속해서 변화하기 때문에 하나의 변수로써 $g_{max}(t)$로 표시하도록 하겠다.

이제 g_{max}에 대해 좀더 나은 측정을 하기 위해 검출된 끝점을 사용할 수 있다.

먼저 g_{max}를 초기에 g_0로 설정한다.

그림 4(b)에서 시작 점에 도달할때까지 초기값 g_0를 사용하여 정규화를 수행한다. 시작지점에서

$$E(g(t);S\le t\le S+2W)\ge g_m \quad (8)$$

조건을 만족하는지를 알아본다.

식 (8)에서 g_m은 새로운 g_{max}가 검출할 음성에 해당됨을 확실히 해주기 위한 경계치가 된다.

식 (8)을 만족하면

$$g_{max}(t)=\max(g(t);S\le t\le S+2W) \quad (9)$$

을 수행한다.

이후로,

$$g_{max}(t)=\max(g(t+2W),\ g_{max}(t-1);\forall t>S) \quad (10)$$

을 수행해서 갱신한다.

Ⅳ. 실 험 및 결 과

실험은 lab에서 실시 되었으며, 프로그램은 matlab으로 구현하였다.

프로그램에 사용된 각 정보를 보면,

1. 입력 신호
   샘플율=10000/s, 트리거 당 샘플율=20000/s
   사운드카드 사용,
2. 1차 단구간 에너지
   해닝창를 사용.
   프레임 크기=0.1msec, 변이=0.01msec
3. 끝점 결정 및 에너지 정규화
   Tu=3, Tl=-3, gap=20, g_0=40, g_m=30
4. 잡음
   그림 5 "폭포 소리", 그림 6 "소나기 소리"

실험을 위해 미리 잡음을 취득해 준비 하여 입력 음성과 합성되어 실험에 사용되었다. 실험 결과 잡음이 첨가되면서 끝점 검출에 영향을 주고 있으나, 거의 비슷하게 출력되고 있으며 에너지 정규화의 경우에도 좋은 성능을 보이고 있다.

그림 5는 실험 결과를 보여주고 있다.

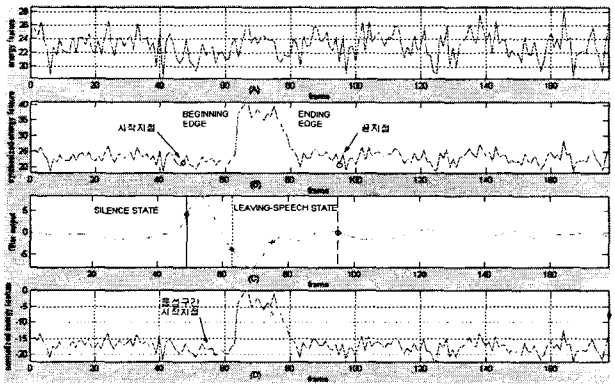

그림 7. (A) 폭포 소리에 해당하는 잡음 에너지
(B) 잡음을 추가한 "GO"에 대한 에너지
(C) 필터 응답 (D) 에너지 정규화

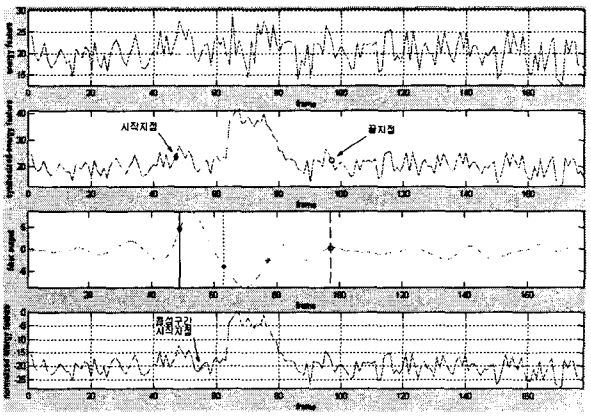

그림 8. (a) 소나기 소리에 해당하는 잡음 에너지
(b) 잡음을 추가한 "GO"에 대한 에너지
(c) 필터 응답 (d) 에너지 정규화

그림 4는 음성 입력시에 백색 잡음을 가지고 있는 20dB SNR경우이고, 그림 5는 인위적으로 잡음을 추가한 15dB SNR경우이며, 그림 6은 10dB SNR 경우이다.

그림 5, 6과같이 SNR이 더 작은 경우에 추출된 끝점 위치가 조금 차이가 나고 있으나 거의 유사한 결과를 보이고 있으며, 정규화에도 좋은 결과를 보여 주고 있다.

Ⅴ. 결 론

논문에서 실시간 끝점 추출에 대해 제안 하였다. 끝점 추출을 위해 필터를 사용하였고 3상태 천이도를 이용하여 필터의 출력을 결정 하였다.

필터의 응답이 여러 다른 잡음에도 강인하도록 설계되었으므로 낮은 SNR에서 제안된 알고리즘이 믿을 수 있는 결과를 주고 있다는 것을 알았다. 또한, 이 끝점은 실시간에서의 에너지 정규화에도 직접적으로 영향을 미치게되어 좀 더 빠르고 정확한 수행을 하게 된다.

제안된 알고리즘은 실시간 ASR 시스템에 좀 더 큰 향상을 기대하며, 앞으로 batch-mode에서의 적용을 하고자 한다.

참 고 문 헌

[1] Jinsong zheng, Augustine Tsai, Qiru zhou, Qi Li, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition", IEEE Trans. Speech Audio Processing, vol. 10, No. 3, pp. 146-157, MAR. 2002

[2] Maria Petrou and Josef kittler, "Optimal Edge Detectors for Ramp Edges", IEEE Trans. Speech Audio Processing, vol. 13, No. 5, pp. 483-491, MAY. 1991

[3] J. Canny,"A computational approach to edge detection", IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-8, pp. 679-698, Nov.1986.

[4] 오영환, "음성언어 정보처리", 홍릉 출판사