화자인식을 위한 강인한 끝점 검출 알고리즘

정대성, 김정곤, 김형순
부산대학교 전자공학과

Robust Endpoint Detection Algorithm For Speaker Verification

Dae Sung Jung, Jung Gon Kim, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
E-mail : {jungds7, gnome, kimhs}@pusan.ac.kr

Abstract

In this paper, we propose a robust endpoint detection algorithm for speaker verification. Proposed algorithm uses energy and cepstral distance parameters, and it replaces the detected endpoints with endpoints of voiced speech, when the estimated signal-to-noise ratio (SNR) is low.

Experimental results show that proposed algorithm is superior to energy-based endpoint detection algorithm.

Ⅰ. 서 론

끝점 검출기는 음성인식 및 화자인식에 있어서 음성 앞 뒤의 불필요한 묵음구간을 제거해주며, 인식성능과 계산량에 상당한 영향을 준다. 일반적으로 끝점 검출에 주로 사용되는 특징 파라미터는 에너지이다. 에너지의 경우 비음성 구간의 에너지와 음성구간의 에너지의 차이가 큰 조용한 환경에서는 끝점 검출이 원활하게 이루어지지만, SNR이 낮은 환경에서는 음성 구간과 비음성 구간의 에너지의 차이가 크지 않아서 끝점 검출이 실패하는 경우가 자주 발생한다.

본 논문에서는 끝점 검출시 널리 사용되는 에너지 파라미터[1]와 음성과 비음성의 스펙트럼의 차이를 이용한 cepstral distance 파라미터[2]를 이용해서, 음성구간의 시작점과 끝점을 검출하였다. 또한, 검출된 음성구간의 시작점과 끝점에서 SNR을 추정하여 SNR이 낮을 경우에는 시작점 혹은 끝점 검출이 잘못되었다고 판단하고 유성음의 시작점과 끝점을 출력한다.

본 논문의 구성은 다음과 같다. 2절에서는 기존의 energy에 기반한 음성 검출 알고리즘[1]에 대해서 살펴보고, 3절에서 본 논문에서 제안한 끝점 검출 알고리즘에 대해서 설명한 후, 4절에서는 화자 인식 실험 및 결과를 정리하며, 5절에서 결론을 맺는다.

Ⅱ. 에너지를 이용한 음성 검출 알고리즘[1]

이 절에서는 본 논문에서 baseline으로 하고 있는 에너지 기반의 끝점 검출 알고리즘에 대해서 설명한다.

2.1 특징 파라미터

여기서 사용되는 에너지는 mel-filter bank 에너지이며, 음성 앞 뒤의 잡음 레벨이 달라서 잡음 구간을 음성구간으로 판단하는 문제점을 해결하기 위해 음성 구간의 에너지를 이용해서 SNR을 예측한 후 SNR에 따라서 잡음 에너지의 문턱치가 가변적으로 변하도록 하였다.

음성 검출에 사용된 에너지 파라미터는 다음과 같다.

$$Current\_E = E_{l+1} \tag{1}$$

$$Forward\_E = \frac{1}{L+1}\sum_{l=L+2}^{2L+1} E_l \tag{2}$$

$$Energy\_Ratio = \frac{\sum_{l=L+2}^{2L+1} E_l}{\sum_{l=1}^{L} E_l} \tag{3}$$

$$Q(0.5)\_E = Sorted\_E[(2L+1)0.5] \tag{4}$$

$$Q(0.9)\_E = Sorted\_E[(2L+1)0.9] \tag{5}$$

$$Speech\_E = \begin{cases} if\ \ 2nd\ max\ speech\ energy = 0 \\ \quad max\ speech\ energy \\ else \\ \quad (max\ speech\ energy + 2nd\ max\ speech\ energy)/2 \end{cases} \quad (6)$$

$$Noise\_E = \begin{cases} Background\_noise\_E & if\ Speech\_E = 0 \\ \dfrac{Speech\_E}{Estimated\_SNR} & if\ Speech\_E > 0 \end{cases} \quad (7)$$

$$Background\_noise\_E = \frac{sum\ of\ silence\ frames'\ energy}{number\ of\ silence\ frame} \quad (8)$$

$$Estimated\_SNR = \frac{(Speech\_E - Background\_noise\_E)}{Background\_noise\_E} \quad (9)$$

위 음성 검출 시스템에서 총 2L+1개의 프레임을 버퍼에 저장한 후 중앙(L+1번째)의 프레임에 대해서 음성/비음성 여부를 판단하게 된다.

여기서 Q(0.5)_E는 2L+1개의 에너지 값들을 올림 차순으로 정렬 했을 때의 중간값(median)이고, Q(0.9)_E는 (2L+1)·0.1번째로 큰 에너지 값이다. 그리고, Noise_E는 입력 신호의 SNR에 따라서 변화하는 값으로서, 음성 앞과 뒤의 SNR이 다른 환경에서 음성의 끝부분을 찾을 때, 음성 앞부분에서 구한 Background_noise_E를 문턱치로 사용하는 것에 비해서 보다 신뢰성 있는 음성 검출을 가능하게 해 준다. Speech_E는 음성 구간에서 제일 큰 peak와 두 번째로 큰 peak를 찾아서 평균값을 사용하였고, 두 번째 peak가 없는 경우에는 첫 번째 peak값을 사용한다.

2.2 음성 검출 알고리즘

그림 1은 앞 절에서 설명한 에너지 파라미터들을 사용하여 구현한 음성 검출기의 전체 흐름도이다.

Voice activity check에서는 현재 프레임이 잡음에서 음성으로 넘어가는 단계인지를 확인하는 부분으로 Forward_E와 Current_E가 Noise_E의 k배 이상이면 음성이 시작할 가능성이 있는 것으로 판단한다.

Speech start condition check에서는 이전에 음성을 검출한 적이 있는지 없는지에 따라서 다른 조건을 주게 되는데, 이전에 음성을 검출한 적이 있으면 SNR에 따라서 가변적으로 변화하는 Noise_E를 기준으로 음성의 시작 여부를 판단하게 되고, 이전에 음성을 검출한 적이 없으면 Energy_Ratio를 사용해서 현재 프레임이 음성의 시작점인지 아닌지를 결정한다.

Speech hold on check에서는 현재 프레임이 음성의 끝점 이거나 끝점일 가능성이 있는지를 보게 되는데, Q(0.9)_E와 Noise_E의 비가 미리 정한 문턱치 이하로 떨어지게 되면 음성이 끝났다고 판단하고 hang over를 적용하고, Q(0.9)_E와 Noise_E의 비가 문턱치 이상이라 하더라도 Forward_E와 Noise_E의 비가 문턱치 이하인 프레임들이 연속해서 N개 이상 나타날 때에는 hang over로 넘어가게 된다.

Return to speech condition check에서는 hang over 기간 중에 다시 음성을 시작할 가능성이 있는지를 보게 되는데 Q(0.9)_E/Noise_E, Forward_E/Noise_E, Energy_Ratio, Speech_E/Current_E 값들을 이용해서 다시 음성이 시작했는지를 판단하게 된다.

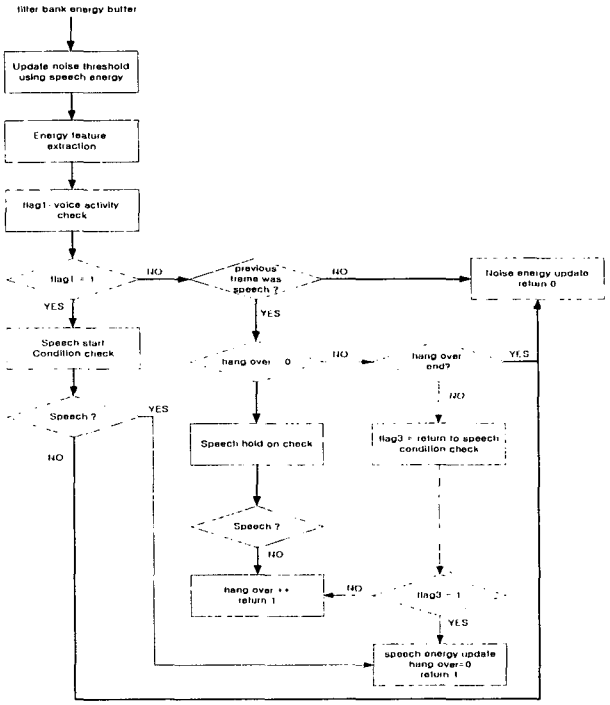

그림 1. 음성 검출기의 흐름도

Ⅲ. 제안한 끝점 검출 알고리즘

2절에서 언급한 음성 검출 알고리즘을 기반으로 하여 새로운 끝점 검출 알고리즘을 제안한다.

3.1 Cepstral Distance[2]를 이용한 음성 검출 알고리즘 수정

에너지만으로 신뢰성 있는 음성 검출이 어렵기 때문에, 음성과 비음성의 스펙트럼 특성이 다르다는 사실을 이용한 cepstral distance를 함께 사용하였다.

cepstral distance를 구하는 방법은 다음과 같다. 우선, 잡음구간에서 다음 두 파라미터를 구한다.

$$C_{mean}(i) = \frac{1}{N_f} \sum_{j=1}^{N_f} c_j(i) \quad (10)$$

$$D_{sil} = \frac{1}{N_f} \sum_{k=1}^{N_f} \sum_{i=1}^{p} (c_j(i) - C_{mean}(i))^2 \quad (11)$$

여기서, N_f는 비음성의 특징을 추정할 프레임 수이며, $c_j(i)$는 j번째 프레임의 i번째 cepstrum 계수이고, p는 cepstrum의 차수이다. C_{mean}은 비음성구간의 평균적인 spectral 특성을 나타내며, D_{sil}는 첫 N_f프레임 내의 cepstral vector와 C_{mean}과의 평균 Euclidian distance를 나타낸다.

다음으로 입력음성의 각 프레임마다 Euclidian distance d_k를 구한다.

$$d_k = \sum_{i=1}^{p} (c_k(i) - C_{mean}(i))^2 \qquad (12)$$

여기서, k는 프레임 인덱스이다.

그림 2는 동일한 clean 음성에 NOISEX DB중 babble 잡음과 Volvo 잡음이 섞인 경우에 대해서 cepstral distance d_k를 나타낸 것이다. 그림에서 볼 수 있듯이 Volvo 잡음에서는 cepstral distance가 음성검출에서 좋은 파라미터로 작동할 수 있지만, babble 잡음과 같이 음성과 비슷한 스펙트럼을 가지는 잡음에 대해서는 cepstral distance가 효과적이지 못하다. 본 논문에서는 D_{sil}값이 정해진 임계값보다 큰 값을 가지면, cepstral distance를 적용하지 않았다.

수정된 음성 검출기는 다음과 같다.

(1) if D_{sil}>Th1

then VAD(k)=Energy_Based_VAD(k)

(2) if D_{sil}<Th1 & Energy_Based_VAD(k)=1 &

d_k>Th2* D_{sil}

then VAD(k)=1

(3) if D_{sil}<Th1 & d_k>Th3

then VAD(k)=1

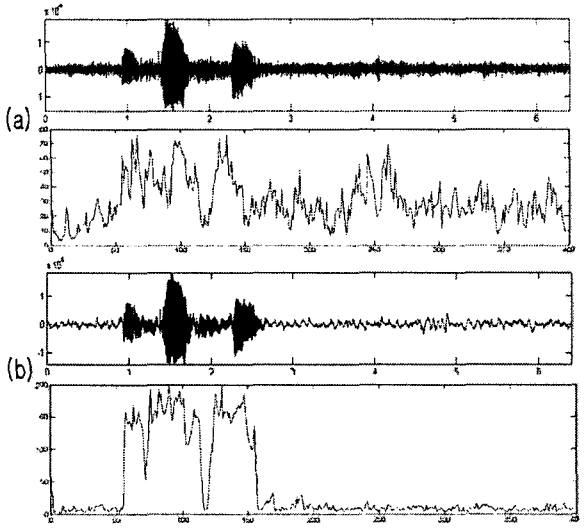

그림 2. 잡음섞인 음성파형과 cepstral distance
(a) babble noise (b) Volvo noise

(1)에서 D_{sil}이 임계값 보다 크면 cepstral distance를 적용하지 않고, 에너지 기반 음성 검출기를 그대로 적용한다. (2)에서는 에너지 기반 음성 검출기에서 음성구간이라고 판단되더라도 cepstral distance로 재확인한다. (3)에서는 cepstral distance가 Th3보다 크면 에너지 기반 음성 검출기의 결과에 관계없이 음성 프레임이라고 판단한다.

3.2 유성도(voicedness) 측정

수정된 음성검출 알고리즘을 통해서 음성구간이라고 판단된 부분에 대해서 3-level center clipping[3] 방법을 통해 유성구간을 검출하였다.

3.3 끝점 검출 알고리즘

유성음의 시작점과 끝점은 최초 5프레임이상 유성구간이라고 검출된 지점을 시작점으로 하였고, 50프레임이상 무성구간이 계속될 때 마지막 유성구간을 끝점으로 하였다. 음성의 시작점과 끝점도 음성 검출기의 결과를 이용해 같은 방법으로 적용했다.

3.4 SNR에 따른 가변적 끝점 검출

끝점 검출기에 의해 구해진 시작점과 끝점의 SNR은 다음 식으로 구해지며, 이들이 미리 정해진 임계값보다 작을 경우 검출된 시작점 혹은 끝점이 잘못됐다고 판단하고, 유성음의 끝점 검출 결과를 따른다.

$$SNR_{begin} = 10 * \log \frac{\frac{1}{10} \sum_{i=begin\_frame}^{begin\_frame+9} E(i)}{Background\_noise\_E} \qquad (13)$$

$$SNR_{end} = 10 * \log \frac{\sum_{i=end\_frame-9}^{end\_frame} E(i)}{\sum_{i=end\_frame+1}^{end\_frame+10} E(i)} \qquad (14)$$

여기서 E(i)는 i번째 프레임의 에너지이며, Background_noise_E는 식(8)에 의해 구해지며, begin_frame과 end_frame은 끝점 검출기에 의해 구해진 음성의 시작 및 끝 프레임이다.

Ⅳ. 실험 및 결과

4.1 음성 데이터 베이스

본 논문에서 사용한 데이터 베이스는 20대 남성 19명과 여성 10명을 대상으로 4연 숫자, 단어, 짧은 문장을 3개월간 한달 간격으로 3회 녹음한 것으로, 발성 목록은 아래 표와 같다. 화자 인식에 있어서 끝점 검출의 성능을 살펴보기 위해서, 문장형태는 실험에서 제외하였다.

표1. 화자 확인 DB의 발성 목록

4연 숫자	단 어	문 장
1565 2529	우리나라	하나를 보면 열을 안다
	안하무인	목마른 놈이 우물 판다
	남가일몽	형만한 아우 없다

수집환경은 비교적 잡음이 많은 사무실 환경이며 각 문장당 6회씩 발성하였다. 음성신호는 16 kHz로 sampling되어 있으며 16 bit로 양자화 되어 있다.

음성 데이터베이스 중 잡음을 섞지 않은 첫 달 분량은 화자확인 시스템의 사용자 모델의 훈련을 위해서 사용하였고, 나머지 2달 분량의 데이터는 잡음환경을 시뮬레이션 하기 위해서 ITU P.56 소프트웨어[4]를 사용하여 잡음 데이터 베이스인 NOISEX에 있는 음성 잡음인 babble 잡음과 자동차 잡음인 Volvo 잡음을 SNR 20dB, 15dB, 10dB이 되도록 녹음된 음성에 더해서 테스트 데이터 베이스를 구성하였다.

4.2 실험 및 결과

실험에 사용한 음성 특징 파라미터로는 12차의 MFCC를 사용하였다. 훈련 시에는 clean 데이터를 이용해서 모델링된 HMM 파라미터와 각 등록 음성에 대한 likelihood score를 사용자 모델과 함께 저장한다. 테스트 시에는 User ID를 받으면 사용자 모델 database에서 HMM 파라미터와 threshold를 가져오고, 테스트 음성의 score를 구해서 임계값 보다 크면 본인이라고 판단하고 그렇지 않으면 사칭자라고 판단하게 된다.

그림 3은 여러 가지 끝점 검출에 의한 화자 확인 시스템의 EER을 보여주고 있다. Manual은 음성 데이터 베이스의 각 발화를 수작업에 의해 시작점과 끝점을 labeling한 것이고, VAD는 2절에서 언급한 에너지 기반 음성 검출 알고리즘을 사용한 것이며, VAD+Cepstral distance는 3.1절에서 설명한 에너지 기반 음성 검출 알고리즘에 cepstral distance를 함께 사용한 것이다. SNR-dependent endpoint는 3.4절에서 설명한 SNR에 따른 가변적 끝점 검출을 사용한 것이다.

Cepstral distance를 사용했을 때 기존 에너지 기반 음성 검출 알고리즘에 비해 자동차 잡음에서 성능 향상이 큼을 알 수 있고, SNR에 따른 가변적 끝점 검출을 한 경우는 cepstral distance를 사용한 끝점 검출 알고리즘에 비해 전체적으로 성능향상이 있었다. 또한 끝점 검출에 의한 화자인식 실험의 upper limit으로 볼 수 있는 Manual 실험결과에 아주 근접함을 알 수 있다

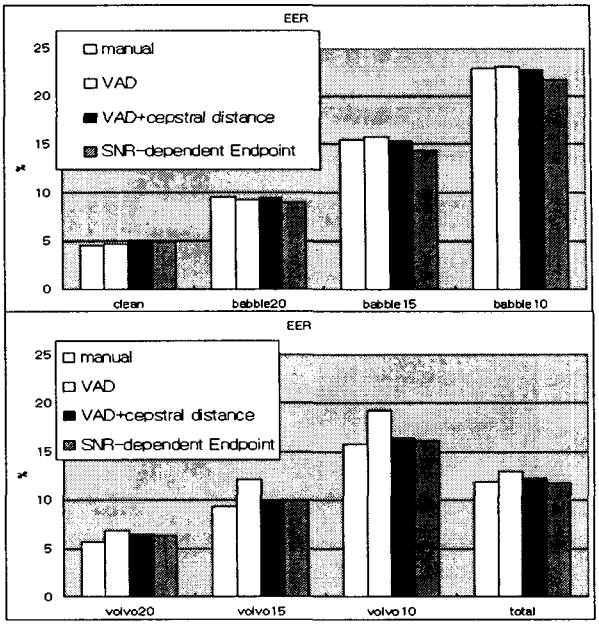

그림 3. 끝점 검출 방법에 따른 EER

V. 결 론

본 논문에서는 에너지와 cepstral distance를 함께 사용하여 음성의 끝점을 검출하고, 시작점과 끝점에서 SNR이 낮을 경우에는 유성구간의 시작점과 끝점으로 대체하는 끝점 검출 알고리즘을 제안하였다. 실험 결과, 다양한 환경에 대해 화자 확인 시스템의 성능을 전반적으로 향상시킴을 확인하였다.

본 논문은 한국전자통신연구원 위탁연구과제 결과의 일부입니다.

VI. 참 고 문 헌

[1] 김정곤, 김형순 외, "주변 잡음에 강인한 화자인식 알고리즘 성능 분석 연구," 최종연구보고서, 한국전자통신연구원, 2003년 1월.

[2] S. E. Bou-Ghazale, K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," *Proc. ICASSP*, vol.4, pp. 3803-3811, 2002.

[3] L .R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice Hall, 1978.

[4] ITU recommendation P.56, "Objective measurement of active speech level", 1993.