工學硏究 第 27 輯 (1997) pp. 99~104
全北大學校 工業技術硏究所

대용량 음성인식을 위한 음소분할 알고리즘에 관한 연구

문 인 섭* · 박 기 영* · 김 종 교**

A Study on Phoneme Segmentation Algorithm for Large Vocabulary Speech Recognition

Is-Seob Moon,* Kee-young Pak,* and Chong-Kyo Kim**

ABSTRACT

Recently, in the speech recognition, an advanced algorithm is needed to recognize large vocabulary without degradation of the recognition rate. The performance of recognition algorithm using words itself as a recognition units is seriously degraded when the number of recognition objects is increased. However, the speech recognition by phonemes has shown comparatively good results in large vocaburary. In order to increase the recognition rate, we should investigate the detailed characteristics of phonemes as the recognition units.

In this paper, we proposed an algorithm of phoneme segmentation for the large vocabulary speech recognition. For phoneme segmentation, we used energy, LCR and 2nd-PVR for speech feature parameters. The speech data (27 phoneme balanced words) for phoneme segmentation are collected from 5 male speakers in Lab with various noises. By the proposed algorithm, we obtained the segmentation rate of 89.45%.

1. 서 론

음성 인식 분야에서 궁극적으로 추구하는 것은 기계에 의하여 사람이 발성한 음성을 완벽하게 이해하는 데 있다. 기존의 연구성과에 의해 단어 단위의 인식이 실용화할 정도의 성능향상을 이룸에 따라 실질적인 목표인 대용량 음성을 인식하는 문제에 관심이 쏠려 있다. 이를 위해서 해결해야 할 최우선적인 문제는, 다양한 문장을 모두 모델로 선택하여 훈련시키는 기존의 방법은 더 이상 사용할 수 없기 때문에, 음절 단위나 음소 단위로 분할하여 제한된 모델로 음성을 구성하는 것이 필요하다. 대부분의 연구가

* 전북대학교 전자공학과 박사과정
** 전북대학교 전자공학과 교수

음소단위로 이루어지고 있으며, 대표적인 접근방법은 세가지 정도로 나눌 수 있다. 첫째로 통계적인 이론을 바탕으로 전개된 것으로 Divergence Test, GLR (generalized likelihood ratio) 그리고 Original Pulse 방법[1] 등을 적용하여 음성에 대해 음소단위로 특징짓는 파라미터를 구하여 분할하는 것이다. 둘째로 언어가 구성하고 있는 대표음소의 갯수만큼 특징을 구하여 기준패턴으로 구성하고 분할하고자 하는 음성을 프레임 단위로 나누어 기준패턴과 매칭시켜 음소단위로 나누는 방법이 있다. 세번째는 위의 두가지를 혼합하는 방식으로 확률적인 방법을 기준으로 하고 잘못 분할된 프레임 영역에서 매칭을 취하여 음소로 분할하는 방법이다. 그러나 위의 방법들에서의 기본적인 개념은 음성신호를 주파수영역으로 변환하여 LPC 파라미터를 구하고 음소를 특징짓는 계수[4]를 얻기 때문에 계산량에 있어서 효율성을 기대하기 어렵다. 이렇듯 대부분의 이론 전개가 주파수영역에서 이루어지는 이유는 시간영역에서 분할을 수행할 때 발생하는 몇 가지 문제점 때문이다. 즉, 잡음의 영향에 약하고 음소를 특징짓는 파라미터를 찾기 어려운 단점이 있다. 그렇지만 시간영역에서 처리할 때 주파수 영역에서보다 많은 계산량을 줄일 수 있기 때문에 이에 대한 연구가 필요하다.

그리하여, 본 논문에서는 시간영역에서의 음소분할 알고리즘을 구현하여 계산량을 줄이고자 한다. 그러기 위해 기존의 파라미터를 기초로 음소의 특성을 잘 나타내는 파라미터로 변환하여 음소분할 알고리즘의 성능을 향상시킨다. 주로 사용되는 시간영역 파라미터는 에너지, 영교차율, 레벨교차율(LCR), 정규화한 자기상관 계수, 피치(pitch), PVR(Peak Valley Rate)[5] 등이 있다. 음소분할을 하는데 있어서 중요한 또하나의 문제가 문턱값을 어떻게 정하는가이다. 이를 위해 각 음성의 크기를 고려하여 각각의 파라미터에 대해 독립된 문턱값을 구하고자 한다.

2. 특징 파라미터 변환

2.1 에너지

$$E_n = \sum_{m=0}^{N-1} x^2(m)/N \qquad (1\text{-a})$$

$$E_n(dB) = 20 \log E_n \qquad (1\text{-b})$$

음성이 자음+모음 형태의 음절일 때 자음과 모음 사이의 변화는 충분히 급격하게 변하여 음소 경계를 알 수 있다. 이를 효과적으로 사용하기 위해 일반적인 자승 평균 대수에너지 대신에 신호파형의 변화 정도를 크게 하여 문턱값에 여유를 준다. Fig 1에 에너지의 예를 보인다.

Fig. 1. Energy of the "Gong Nam Kong"

2.2 레벨교차율(LCR)

$$L_n = \frac{1}{2} \sum_{m=0}^{N-1} |sgn[x(n-m) - TH] - sgn[x(n-m-1) - TH]| \qquad (2)$$

where, $sgn[s(n)] = \begin{cases} 1, & s(n) \geq 0 \\ -1, & otherwise \end{cases}$

음성의 배경잡음의 영향이 신호 파형에 포함되어 있기 때문에 이를 제거하기 위해 레벨 문

턱값(TH)을 구하여야 한다. 이는 시작점·끝점 검출에서 설명한다.

2.3. 2차 Peak Valley rate(PVR)

일반적인 PVR 파라미터에 대한 식은 다음과 같다.

$$PVR_n = \sum_{m=n}^{n+N-1} (1 - u[\Delta x(m) - \Delta x(m+1)]w(n-m) \quad (3)$$

$$\Delta x(m) = x(m) - x(m-1),$$

where,

$$u(n) = \begin{cases} 1, & n \geq 0 \\ 0, & otherwise \end{cases}$$

이것을 음성에 적용한 파라미터의 파형은 Fig 2와 같이 얻을 수 있다.

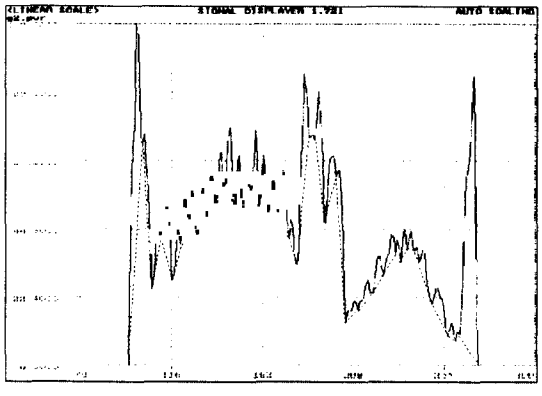

Fig. 2. PVR(real line) and 2nd-PVR(dot line) of the "Gong Nam Kong"

위의 그림과 같이 자음에서의 값은 크고 모음에서의 값은 상대적으로 작다. 또한 자음에서 모음으로 또는 반대로 변하는 경계점에서는 값의 변화가 더욱 커서 효과적인 음소분할을 하는데 유용하다. 그러나 이 파라미터는 하나의 모음 안에서의 변화 역시 상당히 크게 나타나기 때문에 문턱값을 정하기가 어렵고 단어의 발성에 따라 변화 정도가 심하여 그 자체로는 사용할 수 없다. 이를 해결하기 위해 PVR의 valley 값 만을 추적하여 연결시킴으로써 복잡한 파형을 단순화시킨다. 그렇게 하면 그림 2에서 보는 바와 같이 경계점에서 확연한 변화를 얻을 수 있다. 이를 2차 PVR이라 정의한다. 파라미터의 특성상 음의 기울기에서 양의 기울기로 변하는 점이 정해진 문턱값을 넘으면 음소의 경계일 가능성이 대단히 크게 된다는 사실을 이용하여 음소를 분할한다.

결론적으로 그림에서 보여진 것처럼 같은 음소 프레임에서 변화를 압축시키고 음소 경계에서의 변화를 그대로 유지함으로써 안정적인 문턱값을 정의할 수 있으며 이로 인해 음소 경계를 명확하게 검출하는 데 훌륭한 정보를 제공한다.

3. 음성의 시작점·끝점 검출

음성의 시작점과 끝점을 정확히 검출하기 위해 사용하는 파라미터는 에너지와 레벨 교차율이다[2,3,6]. 먼저, 녹음시 발생하는 전원 잡음, 배경 잡음 등을 제거하기 위해 음성 데이터의 처음 부분과 끝 부분에서 두 프레임(20ms) 에너지의 평균을 구한다. 이 때 파형의 순간 피크값을 제거하기 위해 파형의 일반적인 배경 잡음의 크기×여유분(1.3)을 일정값으로 정하여 이 값보다 큰 순간 피크는 프레임의 에너지 계산에서 제외시킨다. 이렇게 구한 배경 잡음값의 1.3배(무성음의 에너지 보다 작고 묵음의 에너지보다 큰 값)로 문턱값을 정하고 음성 전체의 에너지를 비교하여 이 값보다 큰 값들을 구한다. 일부 묵음 구간이 일시적으로 크게 나타나는 비음성을 제외시키기 위해 30ms 이내의 연속적인 에너지가 나타나고 그 다음 프레임에서는 에너지 문턱

값보다 낮은 값을 갖는 경우 이를 비음성으로 판정하여 시작점과 끝점에서 제외시킨다. 또한 파형의 레벨 교차율을 구하기 위한 레벨 문턱값을 구하기 위해 신호 파형의 처음 부분과 끝 부분에서 에너지의 문턱값을 구하는 방법과 동일하게 파형의 평균값을 구한다. 이 값의 1.3배를 기준으로 레벨 교차율을 구하고 비음성 부분의 순간적인 피크 영향을 같은 방법으로 제거시킨다. 이렇게 하여 구한 에너지와 레벨 교차율의 시작점과 끝점을 비교하여 차이가 30ms 이하이면 시작점에서는 두 개의 값 중에서 작은 것을 취하고 끝점에서는 큰 것을 선택하여 시작점과 끝점으로 정한다. 차이가 30ms를 초과하는 경우는 위에서 설명한 전처리 과정에서 거의 모두 제거되지만 초과한다면 에너지를 기준으로 구한 값을 시작점과 끝점으로 한다. 왜냐하면 에너지는 프레임 단위의 평균값으로 레벨 교차율에 비해 음성 파형의 변화에 둔감하기 때문이다.

4. 음소분할 알고리즘

4.1 에너지에 의한 음소분할

에너지를 이용하여 음소분할을 수행하기 위해 현재 프레임과 이전 프레임의 변화가 어느정도인가를 문턱값으로 지정한다. 문턱값이 너무 작아서 경계점을 과분할하는 것보다 적당히 크게 조정하여 에너지의 큰 변화를 우선 분할한다. 이 방법은 모음과 다음 음절의 음소가 모음으로 시작하는 경우, 에너지의 변화가 아주 작아서 경계점을 찾기가 힘들다. 이를 해결하기 위해 에너지의 평균을 이용한다. 시작점과 끝점을 기준으로 처리가 이루어지므로 에너지의 평균은 보통 모음 에너지의 0.6~0.7배 정도의 값을 가진다. 이 값은 자음부분을 제외시키고 모음부분을 포함하기 때문에 제2의 작은 문턱값을 정하여 모음사이의 경계를 찾는 데 사용할 수 있다.

4.2 2차 PVR에 의한 음소분할

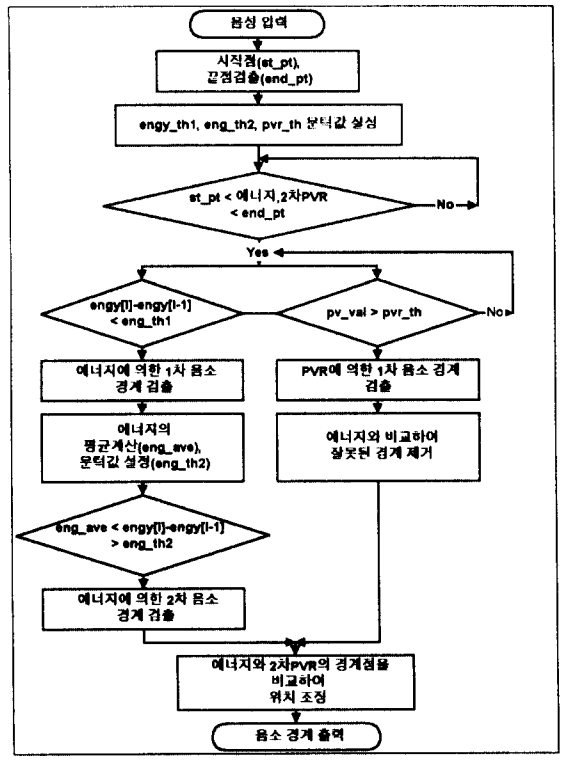

Fig. 3. Flowchart of the Phoneme Segmentation algorithm

PVR에서 구한 2차 PVR은 에너지에 비해 변화 정도가 커서 문턱값을 쉽게 정할 수 있다. 또한 파형의 변화에 민감하게 나타나므로 자음과 모음사이 뿐만 아니라 모음과 모음 사이에서 확연히 경계를 검출할 수 있다. 한가지 단점은 보통 자음의 길이는 40~70ms 정도지만 모음의 지속시간은 자음의 몇 배이므로 같은 모음 구간에서도 파형의 변화가 두세번 반복되는 경우가 나타난다. 특히 단어의 끝음절에서 심하게 나타난다. 그래서 이 변화 부분을 음소의 경계로 검출되는 오류가 발생한다. 이를 보완하기 위해서

검출된 음소 경계점에서 에너지의 변화 유무를 판정하여 음소 경계를 찾는 것이 필요하다.

4.3 음소 경계점 결정

에너지와 2차 PVR에 의한 음소분할 알고리즘을 그림 3에 나타낸다. 위의 에너지에 의한 음소경계와 2차 PVR에 의한 음소경계를 비교하여 음소의 경계점을 마지막으로 조정한다. 여기서 그림 3의 두 번째 단계에서 구한 에너지는 2차 PVR에 종속되어 사용된다. 즉, 에너지의 두 번째 단계에서의 경계점은 존재하지만 2차 PVR에서는 존재하지 않을 경우 오검출로 판정하고 경계점에서 제외시킨다.

첫 번째 단계의 에너지에 의한 경계점과 2차 PVR에서의 경계점의 차이가 30ms 이내에 있는 경우 이들 값의 중간 위치를 경계점으로 정한다.

Table. 1 Spccch data

Phoneme Balanced Words(음소 수)				
가운데/ 가운데(6)	강남콩/ 강남콩(9)	계획표/ 계획표(7)	꽃송이/ 꽃송이(7)	페인트/ 페인트(6)
나그네/ 나그네(6)	너희들/ 너희들(7)	다람쥐/ 다람쥐(7)	도움지/ 도움찌(6)	하여금/ 하여금(6)
라디오/ 라디오(5)	마음씨/ 마음씨(6)	무조건/ 무조껀(7)	받침대/ 받침때(8)	효과적/ 효파적(7)
방위판/ 방위판(7)	붙잡다/ 붇짭다(8)	사용법/ 사용뻡(7)	스탠드/ 스탠드(7)	
아직도/ 아직또(6)	약수터/ 약쑤터(6)	온도계/ 온도계(6)	장발장/ 장발짱(9)	
중요성/ 중요썽(7)	태권도/ 태꿘도(6)	토마토/ 토마토(6)	특별히/ 특뼈리(7)	

Table. 2 Error rate of the phoneme Segmentation

화자\에러율	삽입에러(%)	삭제에러(%)
A	9.3	1.1
B	8.24	1.65
C	10.9	1.65
D	5.49	1.1
E	10.4	2.75
평균 에러율	8.9	1.65

5. 실험 및 고찰

음성 데이터는 27개의 음소 균형된 단어(phoneme balanced word)를 사용하여 남성화자 5명을 대상으로 녹음하였다. 음성 데이터는 표 1과 같다. 각각의 화자마다 발성한 음소수는 182개이다.

음성 녹음은 speech station board(Ariel corp. PC56D)를 이용하여 차단 주파수가 4.6KHz인 LPF를 통과시키고, 10KHz로 샘플링하여 실험실 잡음환경에서 행해졌다. 해석 프레임은 10ms, 쉬프트 간격은 5ms로 하였다. 계산은 펜티엄급(100MHz) 컴퓨터에서 수행되었고 처리시간은 1.5초 이하이었다. 실험 결과는 표 2와 같다. 여기서 삽입과 삭제 에러는 허용오차를 25ms로 하여 구하였다. 에러의 비교 대상은 스피커를 통해서 들리는 음과 육안으로 보이는 파형의 변화를 확인하여 수작업으로 분할하였다.

표 2에서 보는 바와 같이 삽입·삭제에 의한 전체에러율은 10.55%이다. 주로 음소의 경계를 과분할하여 이들 중 오검출된 경계를 제외시키는 방식으로 알고리즘을 구현하였기 때문에 삭제에러는 작다. 이 에러는 주로 종성 폐쇄음에서 나타났다. 왜냐하면 이 음의 변화는 에너지의 변화가 급격하고 2차 PVR 역시 작은 변화율을 갖기 때문에 에러율의 많은 비중을 차지했다.

이와 반대로 삽입에러는 크게 나타났는데, 그것은 하나의 모음을 두 개 이상으로 나눔으로써 발생하는 에러가 대부분이다. 앞 절에서 설명한 것처럼 같은 모음에서는 성도의 변화에 의해 파형의 흔들림(ripple)이 여러번 나타날 수 있기 때문에 파라미터 또한 안정된 값을 갖지 못한다. 그러므로 끝음절에서 모음 부분은 에너지와 비교하여 경계를 재조정해야 한다. 음성 데이터 중에 하나의 예를 스펙트로그램으로 Fig 4에 나타낸다.

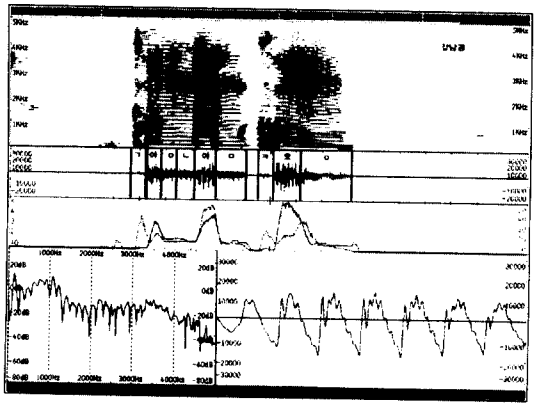

Fig. 4. Jpectrogram and phoneme of the "Gong Nam Kong"

6. 결 론

연속 음성 인식에 반드시 필요한 과정이 음소분할이다. 음소분할은 대부분 음성의 특징을 잘 나타내는 LPC 파라미터에 의해 수행되어진다. 본 논문에서는 시간영역에서 사용되는 파라미터를 이용하여 실시간으로 처리할 수 있는 음소분할 알고리즘을 구현하였다.

사용한 파라미터는 에너지, 레벨 교차율 그리고 2차 PVR이다. 에너지와 레벨 교차율을 사용하여 음성의 시작점과 끝점을 검출하였다. 음소분할을 위해서 파형의 변화에 민감한 2차 PVR과 에너지를 결합하여 경계점을 구하였다. 그 결과 89.45%의 정확성을 얻을 수 있었다. 삭제에러는 종성 폐쇄음에서 주로 나타났고 삽입에러는 끝음절의 모음에서 주로 나타났다.

삽입에러를 줄이고 종성폐쇄음을 찾기 위해 좀더 알고리즘을 효율적으로 구현할 필요가 있다. 또한 음성인식과 연계하여 높은 인식율을 얻기 위해 경계점을 조정해야 하며 특히, 전이구간이 인식에 미치는 영향을 고려하여 음소분할을 수행해야 한다.

참 고 문 헌

1. Regine Andre-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-36, No. 1, pp. 29-40, Jan. 1988.
2. L. R. Rabiner and M. R., "Samb Algorithm for Determining the Endpoints of Isolated Utterances," *Bell Syst. Tech. Journal*, Vol. 54, No. 2, pp. 297-315, Feb. 1975.
3. L. F. Lamel et. al., "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-29, No. 4, pp. 777-785, 1981.
4. L. J. Siegel, A. C. Bess "Voiced /Unvoiced/Mixed Excitation Classification of Speech," *IEEE Trans. Acoust., Speech, and Signal Processing*, Vol. ASSP-30, No. 3, pp. 451-460, June 1982.
5. 전홍렬 외 3인, "PVR-파라미터를 이용한 유성/무성/혼합여기음 검출에 관한 연구", 대한전자공학회 하계 학술발표논문집, Vol. 17, No. 1, pp. 287-290, 1994, 7.
6. M. H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals," Speech Communication, pp. 45-60. 1989.