한국어 음성인식 시스템에서 음소 경계 검출을 위한 Branch 알고리즘

서영완[0] , 한승진 , 장홍종 , 이정현
인 하 대 학 교 전 자 계 산 공 학 과
syw@nlsun.inha.ac.kr

Branch Algorithm for Phoneme Segmentation in Korean Speech Recognition System

Young-Wan Seo[0], Sung-Jin Han, Hung-Jong Jang , Jung-Hyun Lee
Dept. of Computer Science & Engineering, Inha University

요 약

음소 단위로 구축된 음성 데이타는 음성인식, 합성 및 분석 등의 분야에서 매우 중요하다. 일반적으로 음소는 유성음과 무성음으로 구분되어 진다. 이러한 유성음과 무성음은 많은 특징적 차이가 있지만, 기존의 음소 경계추출 알고리즘은 이를 고려하지 않고 시간 축을 기준으로 이전 프레임과의 매개변수(스펙트럼) 비교만을 통하여 음소의 경계를 결정한다.

본 논문에서는 음소 경계 추출을 위하여 유성음과 무성음의 특징적 차이를 고려한 블록기반의 Branch 알고리즘을 설계하였다. Branch 알고리즘을 사용하기 위한 스펙트럼 비교 방법은 MFCC(Mel-Frequency Cepstrum Coefficient)를 기반으로 한 거리 측정법을 사용하였고, 유성음과 무성음의 구분은 포만트 주파수를 이용하였다. 실험 결과 3~4음절 고립단어를 대상으로 약 78%의 정확도를 얻을수 있었다.

1. 서론

음소분할은 음성인식, 음성분석 등의 음성신호처리 분야에서 중요한 문제 중의 하나이다. 연속 음성 인식기에서 사용되는 분석의 기본 단위는 단어, 어절, 음소 등이 사용될 수 있으나, 음소 단위는 단어 및 음절 단위보다 그 종류가 작고 음향적인 특성을 인식기에 고르게 반영할 수 있기 때문에 많이 사용한다. 특히 40여 개의 음소를 갖는 한글에서는 이들 음소들을 인식 단위로 사용함으로써 효율적인 음성 인식 시스템을 만들 수 있을 것이다. 그러나 음성을 음소 단위로 정확하게 분할한다는 것은 수작업으로 하는 경우에도 쉽지 않은 작업으로 수작업에 의한 음소 분할은 많은 시간이 소요되며, 일관성이 보장되지 않는 문제점을 안고 있다[1]. 따라서 음성을 정확히 음소 단위로 분할할 수 있는 자동 음소 분할기의 구현은 음성 인식 시스템의 인식률을 높일 수 있는 한 방안으로 연구되고 있다.

본 논문에서는 한국어의 특성을 고려하여 유성음과 무성음에 차등 임계치를 적용한 자동 음소 경계 검출기를 설계하였다.

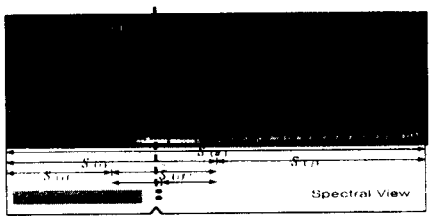
본 논문에서 사용한 스펙트럼 비교 방법은 수정된 MFCC 거리 측정법을 사용하였고, 분석구간 결정을 위해 Branch 알고리즘을 제안하였으며, 유/무성음 구분을 위해 포만트 주파수를 사용하였다.

2. Branch 알고리즘

음성은 시간에 따라 변화하는 특성을 가지고 있으나, 20~30msec정도의 짧은 시간간격 동안은 변하지 않는다고 가정할 수 있다. 따라서 이를 바탕으로 발음기관의 모양과 위치 정보를 포함하고 있는 성도 파라미터의 추출이 중요하다. 일반적으로 이런 시간적 특성을 고려하여 10msec로 중첩된 20msec 프레임 간격으로 특징벡터를 추출한다. 본 논문에서는 입력된 음성신호에서 추출된 특징벡터를 Branch 알고리즘에 스펙트럼 비교법을 적용하여 입력음성을 음소별로 분할하고 군집화하여 음소의 경계를 추출하였다. 또한 구간별 유/무성음 판단을 위하여 포만트 주파수를 이용하였다.

2.1 Branch 알고리즘 구조

일반적으로 음성신호의 분석은 스펙트럼의 차이를 이용하여 음성의 변화를 측정할 수 있다[2]. [그림 1]은 본 논문에서 설계한 Top-Down 방식의 블록기반 Branch 알고리즘에 대한 개념도이다.



[그림 1] Branch 알고리즘 개념도

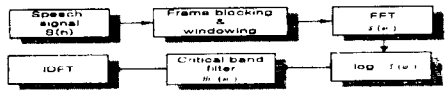
[알고리즘 1]은 본 논문에서 설계한 Branch 알고리즘의 수행 단계를 나타낸다. 알고리즘의 수행은 우선, 입력된 신호의 음성구간을 검출한 후 분석구간을 결정한다. 결정된 구간을 분할한 후 매개변수를 비교하여 일정 임계치 이하이면 같은 음소가 연속적으로 발생되었다고 판단하고, 더 이상 분할을 하지 않고 다음 구간으로 이동을 한다. 이때, 만약 비교치가 임계치 이상이면 서로 다른 음소가 혼합되었다고 가정하고, 분할된 구간을 임계치 이하일 때까지 계속 분할을 수행한다. 음성의 정적구간인 10msec까지 분할이 계속되면 그 중간값을 취하고 분할을 멈춘다.

[알고리즘 1] Branch 알고리즘

```
전체입력음성:S[N], FSIZE:160 //최종 프레임 크기(10msec)
mfc_1[ORDER],mfc_2[ORDER]
Input() ; //전체 입력음성
count = 0;
Range_Index[count]=0;
Branch_Split(0, N-1);
Begin
 Branch_Split(int Index1, int Index2)
 { fsize=Index2-Index1;
   if (fsize > FSIZE) {
     Alpha=(Index2-Index1)/2;
     Beta = Index2;
     mfc_1=MFCC(Index1, Alpha);
     mfc_2=MFCC(Alpha+1,Beta);
     mfcc_diff = mfcc_dist(mfc1,mfc2);
     if (mfcc_diff > mfcc_dist_Threshold){
        Branch_Split(Index1, Alpha);
        Branch_Split(Alpha+1,Index2);
     } else{ Range_Index[++count]=Index2; }
   } else{ Range_Index[++count]=Index2 ; }
end
```

2.2 수정된 MFCC 거리 측정

성분이 서로 다른 파형의 스펙트럼 차이를 잘 나타내기 위해서는 FFT를 바탕으로 한 스펙트럼 비교법이 필요하다. 이러한 FFT를 이용한 거리측정 방법 중 본 논문에서는 인간의 청각 특성을 적용한 MFCC 거리를 적용하였다. 인간의 청각 특성은 1KHz 정도까지의 저주파 영역의 신호에는 민감하여 선형 스케일(linear scale)을 보이지만, 고주파 영역의 신호에는 민감하지 못하여 로그 스케일(log scale)의 특성을 보인다. 이런 특성을 고려한 것이 로그값을 취한 멜 스케일(mel scale)이다[3]. 이러한 멜 켑스트럼은 [그림 2]와 같은 과정을 통해서 구하게 된다[4,6].



[그림 2] 음성 신호로부터 멜켑스트럼을 구하는 과정

본 논문에서 제안한 i번째 블록과 j번째 블록의 MFCC 거리 $d_c^2(L)$는 다음 식과 같다.

$$d_c^2(L) = \sum_{n=1}^{L}\left(\sum_{k=i}^{(i-j)} \hat{c}_{km} - \sum_{k=j}^{K} \hat{c}'_{km}\right)^2$$

2.3 포만트 주파수

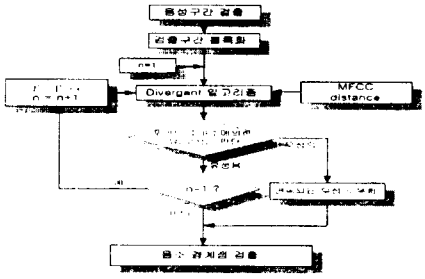
일반적으로 포만트 주파수는 유성음 에서 에너지가 집중적으로 나타나는 영역을 의미하기 때문에 포만트 주파수를 이용하면 음소의 유/무성음을 쉽게 구분할 수 있다. 포만트 주파수와 대역폭에서 크기의 순서에 의해 3개를 선택하여 이를 포만트 주파수로 설정한다. 여기서 포만트 1차부터 3차까지의 포만트 주파수 변화율을 계산하고 3개의 포만트가 모두 변하는 구간을 판단하여 변화율이 임계치를 넘는 구간에서 유성음이 발생되었다고 판단할 수 있다. 주파수 변화율은 다음 식에 의해 계산할 수 있다[5].

$$M(n) = M(n-1) + \sum_{i=1}^{m}[Fi(n+1) - Fi(n)]$$

3. Branch 알고리즘을 적용한 음소경계검출

음소 경계 검출 시스템은 입력음성을 음소 단위로 분류하기 위해 음소 경계 구간을 정확히 검출하기 위한 시스템이다. 기존의 음소 경계 검출기는 유/무성음을 구분하지 않고 스펙트럼 비교만을 통하여 음소의 경계를 찾으려고 한다. 본 논문에서는 앞 절에서 제시한 Branch 알고리즘을 적용하여 1차 음소 경계 후보를 추출한 후 포만트 주파수를 이용하여 구간별로 유성음과 무성음을 판단한다. 이때, 스펙트럼 비교 거리

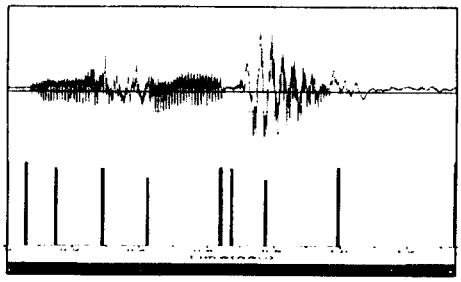
가 낮은 유성음 구간은 가변 임계치를 사용하여 2차 음소 경계 후보를 찾는다. 한국어의 특성상 연속되는 무성음은 없기 때문에 연속된 무성음은 하나의 음소로 통합한다. 그리고, 구간별 MFCC를 구하여 인접 구간의 거리를 측정하여, 측정된 결과가 임계치 이하이면 음소 경계 후보에서 제외시켜 최종 음소 경계를 추출한다. [그림 3]은 음소경계검출의 블록도이다.



[그림 3] 음소경계검출 블록도

4. 실험 및 평가

본 논문에서 제안한 음소경계검출의 성능을 평가하기 위해서 남녀 대학생 3명이 3개씩 발성한 고립단어를 기반으로 실험을 수행했다. 이때의 실험 환경으로 표본화 주파수는 16KHz, 그리고 A/D 양자화 해상도는 16bit이고 창 함수는 해밍 창을 사용하였다. 프레임 크기는 20ms, 창 이동 크기는 10ms간격으로 특징을 추출하였다.



[그림 4] 음소 경계 검출 예 /인하대/

[그림 4]와 [표 1]은 음소 경계 검출기의 예를 나타내었다. 입력음성은 "인하대"이고, 거리측정은 이전 구간과 이후 구간의 스펙트럼 차이를 나타낸 것이다.

[표 1] 음소 경계 검출기 예 /인하대/

	Time (hms)	distance (%)		Time (hms)	distance (%)
1	0.065	98.6	5	0.642	92.6
2	0.150	92.5	6	0.674	91.3
3	0.288	91.9	7	0.777	78.2
4	0.423	80.6	8	0.995	93.1

[표 2]의 성능평가로는 전체 프레임 당 오류 프레임의 비율로써 정확도를 평가하였다.

[표 2] 음소 경계 검출기의 성능평가

%	Euclidean MFCC distance[6]				Branch 알고리즘을 적용한 음소경계검출기			
화자	data1	data2	data3	평균	data1	data2	data3	평균
1	68.78	71.63	70.37	70.26	74.07	79.28	79.63	77.66
2	66.67	67.61	68.75	67.67	74.19	78.87	78.75	77.27
3	70.05	67.86	70.55	69.49	77.00	77.85	78.53	77.79
4	71.96	71.22	69.81	70.99	80.42	79.86	79.87	80.05
5	75.00	73.72	70.81	73.18	78.19	81.02	79.50	79.57
6	70.74	72.86	70.00	71.20	77.65	80.00	79.37	79.01
남자	68.50	69.03	69.89	69.14	75.09	78.67	78.97	77.58
여자	72.57	72.60	70.21	71.79	78.75	80.29	79.58	79.54
전체	70.53	70.82	70.05	[illegible]	76.92	79.48	79.28	[illegible]

5. 결론 및 향후 연구과제

음성분석에 있어서 정확한 음소경계추출은 중요한 부분이다. 본 논문에서는 한국어 고립단어를 대상으로 음성신호를 음성의 최소단위인 음소 단위로 분할하는 알고리즘을 제안하였다. 자동 음소 경계 추출을 하는 알고리즘으로 유/무성음의 특징차이를 고려한 블록기반의 Branch 알고리즘을 사용하였다. 제안한 자동 음소 경계 검출의 최종 성능 평가 결과 3~4음절 고립단어를 대상으로 약 78%의 정확도를 보였다. 개선점으로, 본 논문에서는 고립단어를 대상으로 MFCC 거리측정과 포만트 주파수만을 가지고 유/무성음을 구분하였다. 따라서 비교방법 및 음성특성을 나타내는 다양한 방법을 복합적으로 적용하여 대어휘 연속 음성에 대한 연구가 향후 필요하다.

참고 문헌

[1] B. Eisen, H. G. Tillman, and C. Draxler, "Consistency of judgments in manual labeling of phonetic segments: The distinction between clear and unclear cases," Proc of the ICSLP (Banff), pp.871-874, 1992.

[2] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signals," Prentice Hall.1978.

[3] Hesham Tolba, Douglas O'Shaughnessy, "Automatic Speech Recognition based on Cepstral Coefficients and A Mel-Based Discrete Energy Operator" Proc. ICASSP, vol. 2, 1998, pp.973-976

[4] J. R. Deller J. R. J. G. Proakis, and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Prentice Hall, 1987

[5] J. D. Markel and A. H. Gray, Jr. "Linear Prediction of Speech" Springer-Verlag, New York, 1976.

[6] Lawrence Rabiner, Biing-Hwang Juang "Fundamentals of speech recognition," Prentice Hall, 1993.