2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

LPC Smoothed Log Amplitude Spectra를 이용한 자동 음성 분할

김 도 한*, 이 상 운*, 이 기 정**, 홍 재 근*
*경북대학교 전자공학과
**포항1대학 컴퓨터응용과

Automatic Segmentation Using LPC Smoothed Log Amplitude Spectra

DoHan Kim*, SangWoon Lee*, KiJung Lee**, JaeKeun Hong*
*Dept. of Electronic Engineering, Kyungpook National University
**Dept. of Computer Technology and Application, Pohang College
E-mail ; dododo@speech.knu.ac.kr

요 약

연속음 인식과 음성 합성을 위해서는 정밀한 음성학적 모델과 연속 음성에 적용 가능한 언어 모델의 개발이 중요하다. 이를 위해서는 음성 데이터 베이스에 대한 인식 단위, 혹은 합성 단위의 분할이 필요한데, 수동 음성 분할은 일관성의 유지가 어렵고 긴 시간이 소요되므로 최근에는 자동 분할 기술이 많이 연구되고 있다. 자동 음성 분할 기법으로는 시간 영역이나 주파수 영역 특징 벡터의 천이를 분석하는 방법과 특징 벡터간의 상관도를 구하여 경계를 추출하는 방법이 있다.

LPC smoothed log amplitude spectra는 음성의 주파수 영역의 특징을 잘 나타내며, 동일 음소 내의 상관도가 서로 다른 음소의 상관도보다 더 크고, 음소의 경계 구간에서 급격한 상관도의 변화를 보인다. 이 특성을 이용하여 이웃 프레임에 대한 상관도의 방향성이 특정 조건을 만족하는가를 검사하여 음소의 경계를 구하는 방법을 찾았다. 또한 LPC 이득 인자만으로 묵음 구간을 검출하는 방법을 제시한다. 이렇게 하면 묵음 구간 검출과 음소 경계 검출의 일관성을 향상시키고 수행 시간을 단축시킬 수 있다.

제안한 기법으로 허용 오차 20ms 이내에서 연속음성에 대한 음소 경계 검출 실험을 수행한 결과, 수작업으로 행한 경계 검출 지점의 약 88%를 정확히 검출하였다.

I. 서론

근래의 음성 인식 시스템은 연속음성을 대상으로 인식을 수행하는 경우가 대부분이다. 외국의 경우 TIMIT 데이터베이스와 같이 잘 가공된 음성 데이터 베이스를 구성하고 보급함으로써 동일한 데이터를 토대로 한 연구가 가능했고, 그 결과 연속음성 인식에 많은 발전을 가져왔다. 이에 반하여 국내에는 이렇다 할 표준이 될 만한 음성 데이터베이스가 존재하지 않기 때문에 표준적인 비교가 불가능한 형편이다.

수작업에 의한 음성 분할 및 레이블링 작업은 스펙트로그램 판독 및 반복되는 청취 평가를 통해서 이루어지는데, 이 경우 숙련된 전문가가 필요할 뿐만 아니라 많은 시간을 요하게 되며, 주관적인 요소를 배제할 수 없어 오류가 발생한다.

자동 음소 분할에는 음소에 대한 사전 정보가 없이 음성 데이터 그 자체, 혹은 특징 벡터의 변화분을 계산하여 음성의 경계를 검출해 내는 맹목 분할(blind segmentation)과 음소의 수, 표기(transcription)같은 음운학적 지식을 가지고 경계를 검출하는 지식 기반 분할(knowledge-based segmentation)이 있다. 지식 기반 분할 방법은 표준적이고 공개된 음성 데이터 베이스가 없는 우리말 환경에 적용하기엔 많은 제약이 따른다. 맹목 분할 방법은 음성 인식을 위한 데이터 베이스 구축을 위해서 뿐만 아니라 합성 단위 자동 생성을 위한 음소 분할기와 연속음 인식을 위한 전처리기의 역할도 수행할 수 있다. 이에 본 논문에서는 맹목 분할 방식으로 음성의 특징을 추적하고 음소의 경계를 검출하는 기법을 고찰하였다.

II. 경계 검출

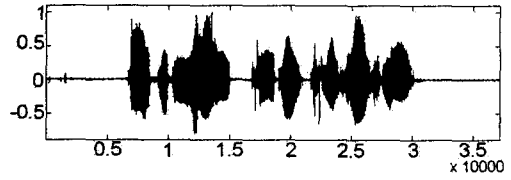
1. LPC 이득 인자를 이용한 묵음 검출

묵음 구간 검출을 위한 특징 벡터는 일반적으로 차분 단구간 에너지와 영교차율이 이용된다[1][2]. 이러한 특징 벡터는 묵음 구간에 부가된 잡음이 저주파이거나 에너지의 크기가 클 경우 성능이 저하된다. 이에 비하여 LPC 이득 인자[4]는 입력 신호의 LPC 계수와 자기상관계수에 따라 변하는 값이므로 저주파나 에너지의 크기가 큰 경우에도 차분 단구간 에너지에 비하여 효과가 있으며 LPC smoothed log amplitude spectra를 사용한 음소 분할에서 경계 검출의 일관성을 유지할 수 있다.

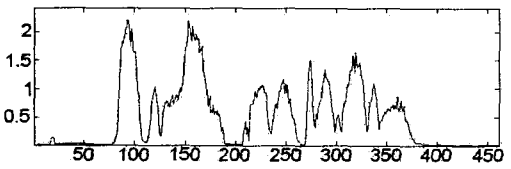
묵음 구간 검출은 이 LPC 이득 인자 σ의 묵음 구간 문턱값 S_{th}를 결정한 후, 문턱값 이하의 값을 가지는 부분을 검출하여 수행된다. 묵음 구간 문턱값 S_{th}는 음성과 묵음에 대한 SNR과 데이터 베이스의 전체적인 보정 계수 C_s로부터 구하였다.[2]

$$S_{th} = 1/10 \sum_{i=1}^{10}(\sigma\sqrt{1+SNR^2/C_s}) \qquad (1)$$

그림 1은 앞과 뒤에 묵음 구간을 가지는 음성 데이터의 파형과 그에 해당하는 프레임 단위 LPC 이득 인자의 파형이다. 그림 1에서 알 수 있듯이 같은 묵음이라도 음성 신호 부근의 짧은 묵음에 대한 LPC 이득 인자가 녹음 후 강제 첨가한 긴 묵음 구간의 LPC 이득 인자보다 더 크게 나타난다.



(a)



(b)

그림 1. 음성의 (a) 시간 영역에서의 파형과 (b) LPC 이득 인자 σ

묵음 구간으로 검출된 프레임과 그 좌우 프레임은 음소 경계 검출 프레임에서 제외하였다.

2. 음소 경계 검출

음소 경계 검출은 인접 프레임에 사이의 주파수 영역 특성 변화가 동일 음소 내에서는 적으며, 음소 경계에서는 현저히 크다는 점을 이용하여 수행하였다. 주파수 영역에서 음소의 변화를 표현할 특징 벡터로는 LPC smoothed log amplitude spectra을 사용하였다. 이 특징 벡터는 좌우 프레임과의 상관계수 값이 동일 음소 내에서는 자기상관계수 값과 유사한 값을 가지며, 음소의 경계 지점에서는 자기상관계수 값에 비하여 급격하게 떨어진다[4].

LPC smoothed log amplitude spectra를 사용한 음성 분할기의 음소 경계 검출의 조건으로 기존의 방법에서는 현재 프레임의 앞 음소에 대한 상관도가 문턱값 이하로 떨어지는 부분과 뒤 음소에 대한 상관도가 문턱값 이상으로 올라가는 부분을 검출하는 방법을 사용하였다[3]. 그러나 이 방법은 음성 데이터 베이스에 따라 서로 다른 문턱치를 실험으로 구하여야 하는 단점이 있다. 이에 음소의 경계 부분에서 LPC smoothed log amplitude spectra의 상관도는 일정한 방향성을 가진다는 점을 이용하여, 프레임의 좌우 음소에 대한 상관도의 경향을 조사하여 상관도의 경향이 앞 음소에서 뒤 음소로 넘어가는 프레임을 음소의 경계로 검출하였다. 또한 음성 데이터의 불균일한 크기에 따른 영향을 최소화하기 위해 상관도에 대한 정규화를 해 주었다.

경계 검출을 위한 상관도 경향의 비교는 다음 수식과 같이 비교 영역내의 상관도의 합과 프레임 단위의 상관도 비율을 이용하였다.

$$S_p(i) = \sum_{j=i-3}^{i-1} C_{i,j}, \quad i = 1,2,\ldots,K \qquad (2)$$

$$S_n(i) = \sum_{j=i+1}^{i+3} C_{i,j}, \quad i = 1,2,\ldots,K \qquad (3)$$

$$R(i) = \frac{S_n(i)}{S_p(i)}, \quad i = 1,2,\ldots,K \qquad (4)$$

$C_{i,j}$는 i번째 프레임과 j번째 프레임간의 상관계수이고 K는 비교 대상 프레임 수이다. $S_p(i)$는 i번째 프레임과 그 앞의 세 개 프레임과의 상관 계수의 합으로 검색 구간내의 앞 음소에 대한 전체적인 상관도를 조사한 것이다. $S_n(i)$는 i번째 프레임과 그 뒤 세 개 프레임과의 상관 계수의 합으로 뒤 음소에 대한 전체적인 상관도를 조사한 것이다. 음소의 경계 지점은 $S_n(i)$가 $S_p(i)$보다 커지는 지점이다. $S_p(i)$ 및 $S_n(i)$의 비교 대상 프레임의 수를 3으로 한 것은 프레임 이동값이 5ms일 때 현 프레임으로부터 15ms 지점까지의 상관도를 구하기 위함이다. 이 수치는 비교 실험결과 가장 높은 경계치 검출을 나타내는 구간이며

동시에 한국어 한 음소의 최소 크기이기도 하다. 그림 2는 음소 경계 구간의 $R(i)$값을 나타낸 것이다. 그림 2에서 60번째 프레임에서는 앞 음소에 강한 상관도를 가지고, 61번째 프레임에서는 뒤 음소에 강한 상관도를 가짐을 볼 수가 있다.

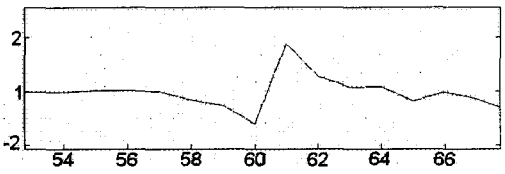


그림 2. 음소의 경계 지점에서 $R(i)$의 값

경계 검출의 조건은 i번째 프레임에 대한 $S_b(i)$, $S_n(i)$ 및 $R(i)$를 사용하여 아래 수식과 같이 구한다.

첫 번째 경계 조건은 전향 탐색 방법(B_f)으로

$$\sum_{j=1}^{2} S_b(i-j) < \sum_{j=1}^{2} S_n(i+j) \quad (5)$$

을 만족하는 i 프레임을 경계로 결정한다.

두 번째 경계 조건은 후향 탐색 방법(B_b)으로

$$\sum_{j=1}^{2} S_b(i-j) > \sum_{j=1}^{2} S_n(i+j) \quad (6)$$

을 만족하는 i 프레임을 경계로 결정한다.

세 번째 경계 조건은 비례 탐색 방법(R_c)으로

$$[R(i-j)<1] \text{ and } [R(i+j)>1] \quad (7)$$

을 만족하는 i 프레임을 경계로 결정한다.

전향 탐색과 후향 탐색은 i 프레임의 비교 구간 내 전체적인 상관도의 경향을 조사하는 것이고, 비례 탐색은 비교 구간 내 개별적인 프레임의 상관도 경향을 조사하는 것이다. 비교 구간은 실험을 통하여 가장 좋은 결과를 내는 두 개 프레임으로 설정하였다.

III. 실험 결과

실험에서 사용된 데이터 베이스는 110명의 화자가 110 종류의 문장과 299어절의 단어를 1회씩 발성하여 문장 단위로 분할한 43,890개의 문장 중 임의로 400문장을 선별하여 100개씩 네 그룹으로 나눈 것이다. 데이터 베이스에는 아무런 음성학적 정보도 주지 않았다.

녹음된 음성은 16Bits, 16KHz로 샘플링되어 있으며, 파일의 앞뒤 구간에 균일하게 400ms의 묵음 구간이 들어있다. 제안한 방법의 정확도를 측정하기 위하여, 전문가가 스펙트로그램 판독을 통해 400개 문장의 음소 경계를 수동으로 추출하였다.

LPC smoothed log amplitude spectra는 16차 LPC 계수를 사용하여 구하였다. 음소 경계 구간 검출은 프레임의 크기를 15ms에서 30ms까지 5ms 단위로 변화시키며 각 오차 범위에 대하여 수행하였다. 고주파 영역 강조, 윈도우 함수 통과 등의 전처리는 주파수 상호간의 상관도에 영향을 미치므로 하지 않았다.

1. 묵음 구간 검출

문장 앞뒤의 묵음 구간과 문장 중간의 휴지구간은 음성인식기에서 동일한 묵음 모델로 모델링이 가능하다. 따라서 문장 앞뒤의 묵음 구간뿐만 아니라 문장 중간의 휴지 구간 검출도 가능해야 한다. 묵음 구간 검출 실험은 기존의 차분 단구간 에너지 측정 방법과 비교하여 실행하였다.

표 1은 차분 단구간 에너지를 이용한 묵음 구간 검출과 LPC 이득 인자를 이용한 묵음 구간 검출의 정확도이다. LPC 이득 인자를 이용한 방법이 차분 단구간 에너지를 이용한 방법에 비하여 묵음 구간 검출과 휴지 구간 검출에 있어서 1.2%와 0.5%의 향상이 있었다.

표 1. 묵음 구간 검출의 정확도 (단위 : %, 오차 범위 : 20ms)

	차분 단구간 에너지	LPC 이득 인자
묵음 구간	98.3	99.5
휴지 구간	98.1	98.6

2. 음소 경계 검출

음소 경계 검출은 프레임의 크기와 오차 범위를 변화시키며 실험하였다.

프레임의 크기는 한국어 음소의 최소 길이인 15ms를 기준으로 5ms씩 증가 혹은 감소시켰다. 프레임의 크기가 15ms 이하인 경우는 주파수 해상도 떨어져서 낮은 경계 검출 능력을 보였다.

표 2는 프레임의 크기를 변화시키며 음소 경계 검출의 정확도를 측정한 값이다. 프레임의 크기가 커질수록 주파수 해상도가 좋아져서 더 좋은 결과를 나타냄을 알 수가 있다. 그러나 프레임의 크기가 20ms를 넘어서면 20ms 이하의 음소 단위들의 경계가 무시되므로 프레임의 크기는 20ms로 결정하였다. 15~20ms 사이의 한국어 음소들의 수는 전체 음소 수에 비하여 적은 양이므

로 무시하기로 하였다.

표 2. 프레임 크기 변화에 따른 음소 경계 검출의 정확도 (단위 : %)

허용오차 / 프레임크기	10ms	20ms	30ms
10ms	45.26	60.49	71.26
12ms	54.34	66.36	75.06
15ms	66.93	81.59	90.34
20ms	77.03	88.65	95.29

표 3은 프레임의 크기를 20ms로 고정한 뒤에 허용 오차의 값을 변화시키며 정확도를 측정한 값이다. 전향 탐색과 후향 탐색은 비교 구간 내의 전체적인 정확도를 측정하는 것인데 반해서, 비례 탐색 방법은 인접하는 비교 구간 내의 개별 프레임에서 조건을 만족해야 하므로 후보군의 수가 적어서 낮은 정확도를 보인다.

표 3. 검색 방법별 음소 경계 검출의 정확도 (단위: %, 프레임 크기 : 20ms, B_f: 전향 탐색 방법, B_b: 후향 탐색 방법, R_c: 상관도 비율 탐색 방법)

오차 / 방법	5ms	10ms	15ms	20ms	25ms	30ms	35ms	40ms
B_f	67.61	77.03	83.46	88.65	91.87	95.29	97.54	99.95
B_b	65.54	73.66	81.11	86.20	90.77	93.94	96.97	98.55
R_c	51.65	60.06	67.80	73.52	78.42	83.03	86.68	89.57

IV. 결론

본 논문에서는 LPC 이득 인자를 사용하여 묵음 구간을 검출하고 LPC smoothed log amplitude spectra에 대한 상관도의 방향성을 비교하여 음소 경계를 검출하는 방법을 제안하였다. 제안한 방법은 음성의 경계 구간 결정에 상관도의 절대값을 사용하지 않아 새로운 데이터 베이스가 주어지더라도 경계 구간 문턱치의 값을 새로 계산할 필요가 없으며, 묵음 구간 검출에 LPC 이득인자를 이용하여 음소 경계 검출과의 일관성을 유지할 수 있는 장점이 있다. 음절핵 추출 등의 후처리 기법을 도입하거나 음소의 경계에 더욱 민감하게 변화하는 특징 벡터를 이용한다면 자동 분절의 오류를 더욱 줄일 수 있을 것으로 생각된다. 또한 프레임의 크기를 조절하면 음소 경계뿐만 아니라 음절 단위의 경계 측정도 가능하다.

V. 참고문헌

[1] A. Ganapathiraju, L. Webster, J. Trimble, K. Bush, and P. Kornman, "Comparision of energy-based endpoint detections for speech signal processing," *Proceedings of the IEEE Southeastcon '96. Bringing Together Education, Science and Technology*, pp. 500-503, 1996.

[2] Duanpei Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, and X. Menendez-Pidal, "A robust speech detection algorithm for speech activated hands-free applications," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 2407-2410, 1999.

[3] Jan P. van Hemert, "Automatic segmentation of speech," *IEEE Transactions on Signal Processing*, Vol. 39, pp. 1008-1012, 1991.

[4] J. D. Markel and A. H. Gray, Jr., *Linear prediction of Speech*, Springer-Verlag, Berlin Heidelberg, New York, 1982.

[5] 박은영, 김상훈, 정재호, "합성단위 자동 생성을 위한 자동 음소 분할기 후처리에 관한 연구," *한국음향학회지*, 제17권, 제7호, pp. 50-56, 1998.