2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

전화 음성의 Segmentation 및 Labeling에 관한 연구

어 범석*, 최 갑근*, 김 학진*, 김 순협*
*광운대학교 컴퓨터공학과

A Study on the Segmentation and Labeling of telephone-based Speech

Bumsuk Eo* Kabkeun Choi*, hakjin Kim*, Soonhyob Kim*
*Dept. of Computer Eng. Kwangwoon University

요 약

상용 가능한 대규모 음성인식 시스템의 개발을 위해서는 음성 데이터베이스 구축이 중요한 과제의 하나로써, 많은 시간과 노력이 요구되며 특히 세그멘테이션과 라벨링은 그 노력의 상당부분이 된다. 본 논문은 ARS 주식 거래 시스템에서 사용되는 대용량 음성 DB의 효과적 구축을 위해 세그멘테이션 및 라벨링의 자동화에 대한 연구를 하였다. 본 연구를 위해 20대 성인 남녀를 대상으로 증권거래와 관련한 15개의 문장을 발성하도록 하였으며 Dialogic사의 D/41ESC보드를 장착하고, Window NT4.0 플랫폼에서 음성을 수집하였다. 또한 자동 Segmentation과 labeling은 Aligner를 사용하였으며 수동과 비교하기 위해 CSLU speech Tool Kit을 사용하였고 수작업은 숙련도가 있는 전문가가 하도록 하였다. [2][3][4][6][7]

1. 서 론

연속음성인식에서 발성사전에 따른 세그멘테이션과 라벨링은 가장 기초적인 작업이며 가장 번거롭고 중요한 작업이다. 이러한 작업은 대부분 음절의 경계를 기준으로 세그멘트하고 또 어떤 시스템에서는 자음 또는 소멸음등을 기준으로 자른다. 하지만 이러한 대다수의 시스템의 성능은 대단히 불만족스럽고 그 기준이 음향적 모델에 매우 민감하므로 올바른 세그멘테이션에 장애가 된다. 따라서 대개는 수작업으로 이 일을 처리하나 ARS 주식 거래 시스템과 같이 대용량 음성 DB를 사용해야하는 경우에는 수작업이 곤란하다. 그러므로 본 연구에서는 세그멘테이션 및 라벨링의 자동화를 위해 음절단위의 세그멘테이션과 라벨링을 시도하였다. 자동화를 위해 먼저 라벨링을 위한 한국어 음소사전의 작성하였고 이 작업은 영어로 되어있는 기존의 음소사전을 한글로 변환하는 작업이었으며 이러한 한국어 음소사전은 [표1-1]과 같다.

표 1

한국어음소사전										
음 소	ㄱ	ㅣ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅊ
*KWPLU	K	N	T	R	M	P	S	NX	C	CH
음 소	ㅋ	ㅌ	ㅍ	ㅎ	ㄲ	ㄸ	ㅃ	ㅆ	ㅉ	ㅏ
KWPLU	KH	TH	PH	H	KK	TT	PP	SS	CC	AA
음 소	ㅓ	ㅗ	ㅜ	ㅡ	ㅣ	ㅐ	ㅔ	ㅚ	ㅟ	ㅑ
KWPLU	AX	OW	UW	WW	IY	EH	EY	OI	UI	JA
음 소	ㅕ	ㅛ	ㅠ	ㅒ	ㅖ	ㅘ	ㅝ	ㅙ	ㅞ	ㅢ
KWPLU	JX	JO	JU	JH	JE	WA	WX	WH	WE	WI

(*KWPLU:KwangWoonPhonemeLikeUnit)

2. 음성신호의 분석

음성신호는 음성 여기원에 따라 유성음, 무성음, 혼합음으로 구분할 수 있다. 무성음의 경우에는 백색 가우시안 불규칙 시퀀스가 그 여기원이므로 주기성은 나타나지 않지만, 주로 3kHz 근방에서 첫 번째의 공진 봉우리를 갖기 때문에 유성음에 비해 평균 영교차율이 크다. 유성음은 폐에서 올라온 공기가 성문을 통하여 배출될 때 진동되고, 성도의 공명으로 인하여 에너지가 크고 준 주기적인 형태의 신호가 된다. 이를 주파수 영역에서 살펴보면 성도의 공명봉우리에 음성신호의 기본 주파수 F_0가 세세하게 나타나고 있다. 성도 공명 봉우리의 주파수들을 제1포만트라고 한다.

일반적으로 유성음 구간에서는 F1의 에너지 봉우리는 다른 포만트 보다 10dB이상 높기 때문에 이를 시간 영역의 파형으로 표현하면 F1의 영향이 주로 나타난다. F1이 주파수 영역에서 다른 포만트들보다 훨씬 높은 에너지 봉우리를 갖기 때문에 F1만을 고려하여 근사적인 방법으로 성도를 분석할 수 있다.[7]

3. 음성검출 파라미터

세그멘테이션을 하기 위해서는 유성음을 검출하여 음성 신호를 유성음 구간과 무성음 구간으로 구분하기 위해 유성음 검출 파라미터 Vi를 이용한다. 일반적으로 유성음은 저주파 부분에 에너지가 밀집되고 무성음은 고주파 영역에 에너지를 많이 포함하고 있으므로, 저주파 부분의 에너지를 추출하면 유성음과 무성음의 구분이 가능하다. 따라서 LPC 대수 스펙트럼의 형태로 표시된 Spectrum envelop Xi(k)의 기본 주파수 대역 부분의 평균치가 유성음 검출에 유효하므로 이것을 유성음 검출 파라미터 Vi로서 사용한다. [1]

$$Vi = \frac{1}{5}\sum_{k=1}^{5} Xi(k)$$

1) 유성음 검출파라미터

유성음을 검출하여 음성신호를 유성음 구간과 무성음 구간으로 구분하기 위해 유성음 검출 파라미터 Vi를 이용한다. 일반적으로 유성음은 저주파 부분에 에너지가 밀집되고 무성음은 고주파 영역에 에너지를 많이 포함하고 있으므로, 저주파 부분의 에너지를 추출하면 유성음과 무성음의 구분이 가능해 진다.

따라서, LPC 대수 스펙트럼의 형태로 표시된 Spectrum envelop Xi(k)의 기본 주파수 대역 부분의 평균치가 유성음 검출에 유효하므로 이것을 유성음 검출파라미터 Vi로서 이용한다.

$$Vi = \frac{1}{5}\sum_{k=1}^{5} Xi(k)$$

2) 영차 LPC CEPSTRUM 계수

유성음 구간에서 모음과 모음, 모음과 비음 또는 모음과 유음이 연이어 발음될 때 유성음 분리 작업해야 할 것이다. 따라서, 저주파 영역에 대부분의 언어정보가 많이 포함되어 있는 점을 고려해 스펙트럼의 저 역 부분에 weight를 둔, 영차 LPC cepstrum 계수를 이용한다. i번째 프레임의 영차 LPC cepstrum 계수 Ci 는

$$Ci = Xi[0] = \frac{1}{2\pi}\int_{-\pi}^{\pi} Xi(\Omega)d(\Omega)$$

으로 표시된다. 영차 LPC cepstrum 계수 Ci는 직선 주파수 눈금 상에 표시된 LPC 대수 스펙트럼에 큰 weight를 붙인 평균치다.

여기서

$$W(\Omega) = \frac{1+\alpha^2}{1-2\alpha\cos+\alpha^2}$$

$$W(\Omega) = \frac{(1+\alpha^2)}{(1-2\alpha\cos\Omega+\alpha^2)}$$

3) 영차 LPC cepstrum 시간 변화 파라미터

영차 LPC cepstrum 시간 변화 파라미터를 사용한다. i 번째 분석 프레임을 중심으로 하는 영차 LPC cepstrum 계수의 시계열 Ci+n (|n|≤M)에 대해서 직선식을 적용해 회귀계수 Ai에 의해서 영차 cepstrum $X = A_i n + B$계수의 시간 변화의 양을 표현하는데, 직선 적용 식에서 시계열 구간 양단의 절단 영향을 작게 하기 위해 평가 함수인 가중된 2승 평균 오차를 이용한다. [2][3][4][6]

$$\varepsilon = \frac{1}{2M+1} \sum_{n=-M}^{M} W_n \ (\ W_i + B - \ C_i + n)^2$$

여기서 Wn 은 |n| > M에서 0 이 되는 음함수 window 함수이다. 직선 적용의 계수 Ai는 가중된 2승 평균 오차 ε를 최소로 하는 조건에서 Ai 에 대해 1차 편미분 하면 아래와 같이 된다.

$$A_i = K_M \sum_{n=-M}^{M} W_n n C_i + n,$$
$$\left(K_M = \left(\sum_{n=-M}^{M} W_n n^2 \right)^{-1} \right)$$

4) ZCR 파라미터

스펙트럼에서 에너지가 집중되는 주파수를 찾는데 유용한 특징파라미터로 사용되는 영 교차율은 무성음 구간에서의 유성자음과 무성자음 그리고 묵음을 검출하기 위해 사용된다.

$$Z = \sum_{m=1}^{M} \frac{[1 - sgn(X_m + 1) sgn(X_m)]}{2}$$

여기서

$$sgn[X(m)] = \begin{cases} 1 & X(m) \geq 0 \\ -1 & X(m) < 0 \end{cases}$$

5) Viterbi algorithm을 이용한 음소탐색

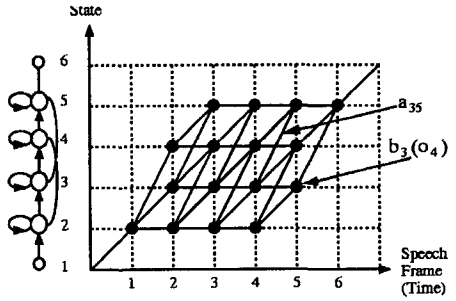


그림 1 Viterbi Search

음소모델을 찾기 위한 최대 확률 값은 다음 수식에 의해 얻어진다. [8]

$$\phi_j(t) = \max\{\phi_i(t-1)a_{ij}\} b_j(o_t).$$

여기서

$$\phi_1(1) = 1$$

$$\phi_j(1) = a_{1j} b_j(o_1)$$

for 1<j<N. 최대 유사도 확률 $\hat{P}(O|M)$는 다음과 같이 주어진다.

$$\phi_N(T) = \max\{\phi_i(T)a_{iN}\}$$

재추정시 직접적인 Viterbi 연산은 underflow를 유발할 수 있기 때문에 log file을 이용하는 것이 유리하며 이를 위한 수식은 다음과 같다.

$$\psi_j(t) = \max\{\psi_i(t-1) + log(a_{ij})\} + log(b_j(o_t)).$$

4. Labeling 알고리즘

Aligner 소프트웨어는 기본적으로 음소 단위 Labeling system이며 HMM을 기반으로 구성되어 있다. 또한 완벽한 Labeling을 위해 두 개의 데이터베이스를 사용한다. 하나는 십만 개 이상의 단어사전과 발음사전을 이용하고 또 하나는 각 음소의 음향학적 모델을 사용한다.

Viterbi탐색 법은 음성 파형과 각 음소의 음향학적 모델 사이에서 최대 확률 값을 찾는다. [5]

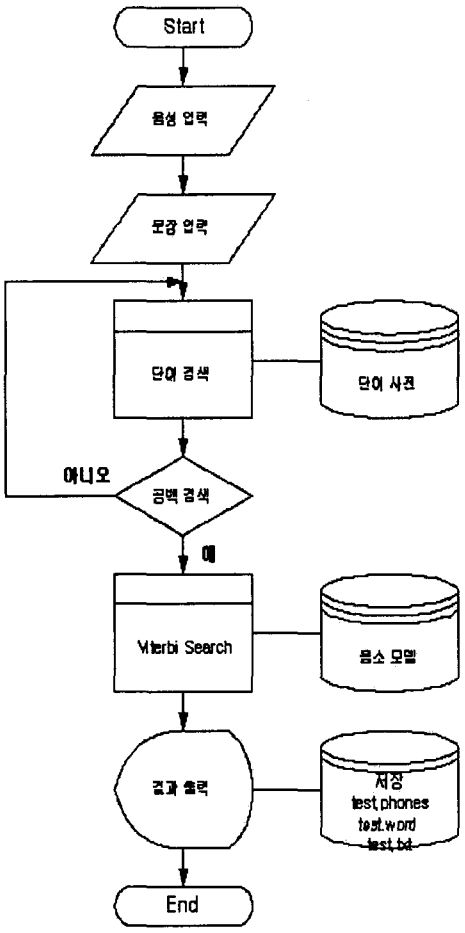


그림 2 Labeling Algorithm

5. 실험

1) 실험환경

표 2. 실험환경

운영체제	Sun Solaris
시스템 사양	Sun Sparc 20
Sampling Rate	8KHz
음성형식	16Bit PCM
발성장소	사무실환경
발성매체	전화기

2) 대상어휘

표 3. 대상어휘

구 분	예 제
문장음	한양증권 매도호가가 얼맙니까? 하한가에 백칠십주를 매수하겠습니다 통일중공업 거래량이 얼마나 됩니까
숫자음	920209-7125891 148,200 7470718
일반종목	강원산업 금호석유우
코스닥종목	골드뱅크 가산전자

3) 실험방법

실험은 두 명의 20대 남녀 화자에게 각각 15문항씩 문장과 숫자음을 주고 발성한 것을 녹취해 사용했으며 다음에 보여질 두 그림은 자동 세그멘테이션과 수동 세그멘테이션의 오차를 보여주며 " 현재가 "를 발음한 것이다.

실험을 위해 먼저 수동 세그멘테이션에 참여하는 작업자는 음성DB 구축의 경험이 있는 숙련자가 하도록 하여 가급적 세그멘테이션의 정확도를 기했으며 특히 불확실한 음소단위의 세그멘테이션은 지양하고 그 효과가 확연한 음절단위로 구분했다. 구분기준은 Center for Spoken Language Understanding(CSLU is an Oregon Graduate Institute of Science and Technology Research Center that focuses on Spoken Language Technologies)에서 제공하는 speech Tool kit을 사용하여 포만트 분포, 에너지, 피치 등을 시간 축으로 음성파형과 정합 시켜 피치기준으로 세그멘테이션 했고 그 실험결과는 다음의 [그림 3] 에 나타나 있다.

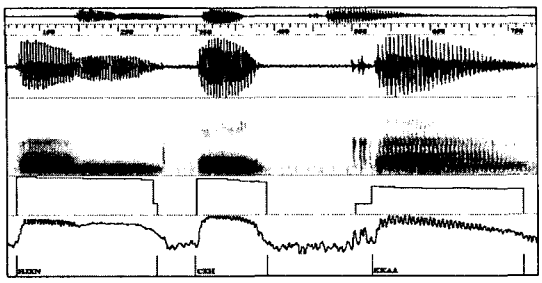
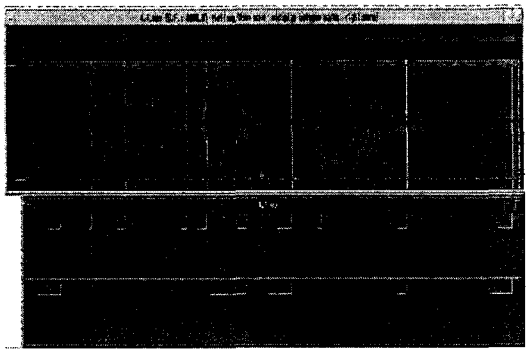


그림 3 CSLU Tool kit을 이용한 Labeling

다음 그림은 Entropic 사의 Aligner를 이용한 자동 세그멘테이션 및 라벨링 작업을 한 것을 보이고 있다. [그림 4] 의 결과는 먼저 Entropic사에서 제공하는 영어식 음소사전에 한국어 발음과 유사한 영어식 발음을 음소단위로 정리하여 사상시켰고 그 사상시킨 음소사전(KWPLU:Kwangwoon University Phoneme Like Unit)을 추가해 얻은 결과로 "현재가"가 음소별로 구분, 세그멘테이션 되고 있는 것을 볼 수 있다.



[그림 4 : Aligner를 이용한 자동 세그멘테이션]

6. 결론

본 연구를 통해 자동 세그멘테이션을 이용해 상용 음성DB구축의 가능성을 확인하였고 특히 적절한 Tool과 한국어에 적합한 음소사전의 구축이 중요하다는 것을 알 수 있었다. 자동 세그멘테이션의 성능측정을 위해 수동 세그멘테이션과 비교하였고 수동 세그멘테이션을 기준으로 1ms의 간격으로 두고 오차범위를 10ms두었으며 오차를 음성구간과 묵음구간 기준으로 확인하였다.

7. 참고문헌

[1] DHMM과 어휘 해석을 이용한 한국어 연속 숫자음 인식 / 최성호 저. 석사학위 논문 1990.

[2] Mermelstein, P. "Automatic segmentation of speech into syllabic units." JASA 58(4):880-883, October 1975.

[3] A. Kurematsu and K. Takeda, "ATR Japanese speech database as a tool of speech recognition and synthesis," Proc. ESCA'89, pp.2.3.1-4, 1989

[4] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T.Umeda, and H. Kuwabara, "A large-scale Japanese speech database," Proc. ICSLP'90, pp. 1089-1092, 1990

[5] Entropic, Inc. Wave+ Manual

[6] cslu speech language technology site, http://cslu.cse.ogi.edu

[7] 전처리된 가변대역폭 LPF에 의한 피치검출법 / 한진희, 정세현, 배명진, 김명제, 김삼룡, 제 12회 음성통신 및 신호처리 워크샵.

[8] Entropic, Inc. HTK manual / Steve young