2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

무제한 단어인식 시스템을 위한 VCCV분할에 관한 연구

윤재선*, 정광우**, 홍광석*
*성균관대학교 전기전자컴퓨터공학부 HCI연구실
** 한국철도대학교 운전기전과

A Study on VCCV Segmentation in Unrestricted Word Recognition System

Jeh-Seon Youn*, Kwang-Woo Chung**, Kwang-Seok Hong*
*HCI Lab, Electrical & Computer Engineering, Sungkyunkwan University
**Dept. of Operation-Mechatronics, KOREA Railroad College

sunhci@ece.skku.ac.kr, ckw@chuldo.krc.ac.kr, kshong@yurim.skku.ac.kr

요 약

무제한 인식 시스템을 구현하기 위해서는 적절한 인식 단위, 훈련 데이터 베이스의 확보, 인식단위의 분할, 인식 알고리즘과 같은 문제점을 모두 해결하여야 한다.

따라서 본 논문에서는 무제한 음성인식 시스템의 인식의 기본 단위로 모음의 안정구간을 검출하여 분할하는 CV(Consonant-Vowel), VC(Vowel-Consonant), VCCV(Vowel-Consonant-Consonant-Vowel)단위와 분할 파라미터를 제안하고, 분할 실험을 통해 그 유효성을 확인하고자 한다.

I. 서 론

음성인식에 사용되고 있는 인식 단위로는 음소, 음절, 단어 등의 언어적으로 정의된 단위들과 음향, 음성학적인 유사도에 기반한 유사음소단위(Phoneme Likely Unit)와 같은 단위들이 사용되고 있다. 단어를 기본 단위 모델로 사용할 경우에는 그 자체만으로 음향학적 변이성이 주로 단어의 처음과 끝부분에서만 일어나기 때문에 언어적 의미가 음향학적으로 잘 표현된다고 할 수 있다. 단어 사전을 구성하지 않아도 되기 때문에 구성상 간단한 점은 있으나, 모든 단어에 대한 모델을 전부 구성해야 하기 때문에 연속음성으로의 확장은 불가능하다. 그러므로 대어휘 연속음성 인식시스템으로의 구현을 위해서는 보다 적은 규모의 subword 단위의 인식 모델이 필요하게 된다. 현재 사용되고 있는 음성인식의 기본단위는 단어보다 작은 음소 또는 2~3개 음소를 결합한 구조를 인식의 기본 단위로 한다. 이들 단위로 이용할 경우 인식에 필요한 총 개수는 약 수십 개에서 수만 개에 이르며, 단위가 길고 복잡해짐에 따라 총 수는 증가한다. 일반적으로 사용되는 단위로는 음소, 반음절, 음절, 다이폰 등이 사용되고, 다이폰에서 실현되지 않는 조음결합에 의한 음운변동을 흡수하기 위해서 CVC나 VCV와 같은 더 큰 단위도 이용하고 있다.[1][2]

따라서 본 논문에서는 음성 분할 및 레이블링 작업을 수월하기 위해 안정된 모음 영역을 분할하여 인식단위로 사용하는 복합음소단위인 CV, VC, VCCV단위를 선택하였다. 일반적으로 한국어는 초성 + 중성 + 종성으로 구성되어져 있으므로 주파수 영역에서 안정된 모음 영역을 찾는 것이 음소경계를 찾는 방법보다는 비교적 수월하며, 또한 음성학 전문가가 아니더라도 자동분할된 경계 위치에 수정작업을 할 수 있는 장점이 있다.[3]

II. CV, VCCV, VC분할 방법

대용량 어휘 연속음성 인식 시스템에서는 DB의 구축 및 하드웨어의 제한된 메모리나 계산 속도로 인한 시스템의 실시간 구현이 어렵다. 이에 대한 해결책으로 단어보다 더 기본적인 음성단위를 이용하여 인식을 수행하는 연구들이 활발히 이루어지고 있다.[4]

인식 단위의 경계추출은 무성음과 유성음사이에서는 뚜렷하게 나타나는 경우가 있지만, 유성음과 유성음 혹은 유성음과 무성음들 사이에서 상호간의 조음현상 때문에 정확한 경계를 찾기가 어려울 뿐만 아니라 이러한

경계 영역의 천이구간에서의 음성 데이터들은 음소단위 인식에서는 별다른 의미를 갖지 못한다. 왜냐하면 이러한 천이 구간의 데이터는 조음현상에 의하여 인접 음소의 영향을 받으므로 인접 음소에 따라 특성이 달라지기 때문이다. 따라서 이러한 음소의 경계를 정확히 찾는 것보다는 특성의 변화가 적은 안정된 대표구간 즉 모음 영역을 찾는 것이 더 효율적이라 할 수 있다.

본 논문에서는 언어학적 특성인 음절의 개수와 모음 정보를 이용하여 모음의 안정된 위치를 찾는 알고리즘을 제안한다. 먼저 음성신호의 유음구간과 무음구간으로 나눈 후, 제안한 파라미터를 이용하여 안정된 모음의 위치를 구한다.

음성 분할을 위한 파라미터로서는 정적인 segmentation parameter와 동적인 파라미터를 이용한다. 정적 파라미터는 음성신호의 단시간 power, 주파수 대역별 power 그리고 신호의 영교차율이 이용되고 동적 파라미터로서는 spectral envelop의 시간 변화나 power의 시간 변화등이 이용되나, 본 시스템에서는 안정된 모음 구간을 검출하기 때문에 정적 파라미터를 이용하여 모음 영역을 구했다.

음성분할 작업의 첫 번째 작업은 음성 신호를 유음 구간과 무음 구간으로 나누는 것이다. 이 두 구간은 class간의 음성신호의 특성이 매우 상이하게 다르며, 모음열이 유음구간에만 존재하기 때문에 경우의 수를 줄일 수 있다.

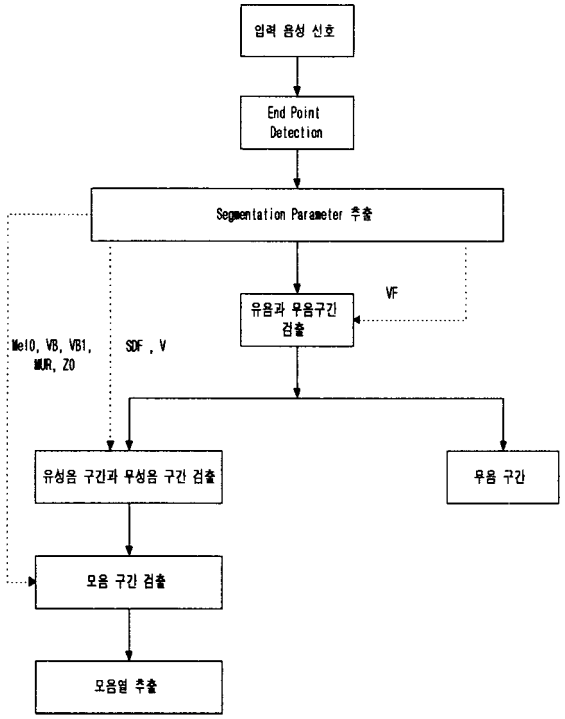

그림 1 분할 시스템의 순서도

모음 분할에 사용되는 파라미터로는 유음구간과 무음구간을 검출하는 파라미터 Volume Function VF, 유음구간의 유성음구간과 무성음구간을 검출하는 Sonorant Detection Function SDF, Voiced Region Parameter V, Mel cepstrum 0차 Mel_0 ,모음의 제 1 formant와 제 2 formant가 위치하고 있는 주파수 대역 power인 Voice Band parameter VB_1, 모음 '이' 모음 검출을 위한 파라미터 VB, 유성종성자음 검출을 위한 파라미터 MUR, 영교차율 Z_0, 각 음절별 Mean과 Covariance를 이용하여 음성 분할 시스템을 구성하였으며, 이들 분할 파라미터를 이용한 segmentation system의 순서도는 그림 1에 나타내었다.

그림 2는 분할 시스템을 이용하여 남성 화자가 발성한 /개인심리학/의 음성 신호 파형, spectrogram과 segmentation parameter를 이용하여 CV /개/, VCCV /애이/, /인시/, /임리/ /이하/, VC /악/의 인식 단위로 분할한 결과를 나타내고 있다.

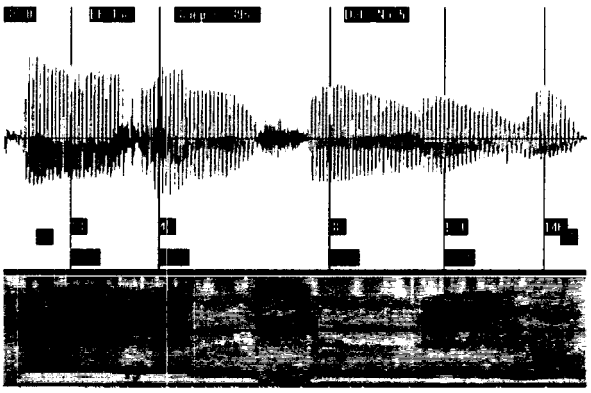

그림 2 /개인심리학/ CV, VCCV, VC의 분할 결과

(1) Volume Function parameter(VF)

각 프레임의 loudness의 크기 또는 acoustic volume의 특징을 나타내는 volume function은 SDF , MUR, VB1, VB parameter의 기본이 되는 함수로서 식 1과 같다.

$$V(i) = \frac{1}{N_i}\sqrt{\sum_{m=A}^{B}\left|H_i(e^{j\pi\frac{m}{128}})\right|} \qquad (1)$$

여기서 i는 현재 프레임의 번호이며, N_i는 프레임의 샘플길이, A는 bandpass filter의 낮은 주파수의 번호이며, B는 bandpass filter의 높은 대역의 cutoff 주파수이다. A = 17, B = 127 의 값을 가지는 single volume function은 유음과 무음을 검출하는데 사용한다. Low cutoff index 값 17은 주파수 732.1Hz에 상응

하며, High cutoff 번호 127의 주파수는 5512.5Hz이다. 이때 A = 17 로 설정한 이유는 유성 마찰음의 민감도를 낮추고, 환경상의 바람에 의한 잡음 등의 영향을 적게 받기 위함이다.

(2) Sonorant Detection parameter(SDF)

공명음(Sonorant)의 특성은 낮은 주파수가 높은 주파수에 비해 큰 값을 가지는 특징이 있다. 공명음에 속하는 음은 모음, 비음, 반모음 등이 속하며, 비공명음(non-sonorant)은 무성 파찰음, 무성 마찰음 등이 속한다. 낮은 주파수 대역의 volume function은 A = 2(86.2Hz), B = 23(991.3Hz) 값을 가지는 Low-frequency Volume Function(LVF)이며, 높은 주파수 대역의 volume function은 A = 94(4051.4Hz), B = 127(5512.5Hz) 값을 가지는 High-frequency Volume Function(HFV)를 사용한다.

공명음을 검출하기 위한 SDF는 식 2와 같다.

$$SDF(i) = \frac{LFV(i)}{HFV(i)} \tag{2}$$

(3) Voiced Region Parameter(V)

유음 구간을 검출한 후, 유성음 구간과 무성음 구간을 나누기 위해 Voiced Region Parameter V를 이용한다. 일반적으로 유성음은 저주파 부분에 에너지가 밀집되고, 무성음은 고주파 영역에 에너지가 많이 포함하고 있으므로 저주파 대역의 에너지를 추출하면 유성음과 무성음의 구분이 가능해 진다. 따라서 대수 FFT 스펙트럼의 형태로 표시된 spectrum envelop $X(k)$의 기본주파수(pitch) 대역의 평균치가 유성음 검출에 유효하므로 이것을 유성음 검출 파라미터 V로써 이용한다

$$V(i) = \frac{1}{9}\sum_{k=1}^{9} X(k) \tag{3}$$

여기서 한 샘플간의 주파수 f_T는 11.025KHz로 샘플링하였으므로

$$f_T = \frac{11025}{256} = 43.1\,\mathrm{Hz} \tag{4}$$

이다. 그러므로 주파수 양단의 주파수 ω_1, ω_2는 각각 43.1Hz와 387.9Hz이다.

(4) Voice Band Parameter1(VB1)

모음 삼각도를 구성하는 모음들의 제 1 Formant와 제 2 Formant는 2500Hz 미만에 존재한다. 따라서 Voice 영역은 A = 16(689.6Hz) , B = 60(2586 Hz)의 volume function으로 구성한 VB1을 사용한다.

(5) '이' 모음 Detection parameter(VB)

일반적으로 모음과 무성종성이 결합된 음절인 경우, 종성의 길이는 매우 짧고 모음 주파수 영역에 큰 영향을 주지 않기 때문에 모음 구간을 검출하는 하는 방법은 초성 + 중성인 경우와 같은 방법으로 모음구간을 분할한다. 그러나 모음과 유성종성이 올 경우 특히 /이/모음 계열의 /인/, /일/, /임/, /잉/의 음절인 경우에는 /이/의 모음영역이 짧아지고, 유성종성의 주파수 성질과 유사한 특성을 가지고 있다.

따라서, 유성자음이 존재하는 /이/계열의 음절을 보다 정확하게 모음분할을 위해 제안한 VB는 다음과 같다.

$$VB(i) = \frac{VB2(i)}{VB1(i)} \tag{5}$$

이 파라미터는 /이/모음과 유성종성자음사이를 분할 정보를 제공한다. 그림 3은 /인 일 임 잉/의 VB 파라미터 값이 모음영역과 유성종성영역의 구별능력을 보여 주고 있다.

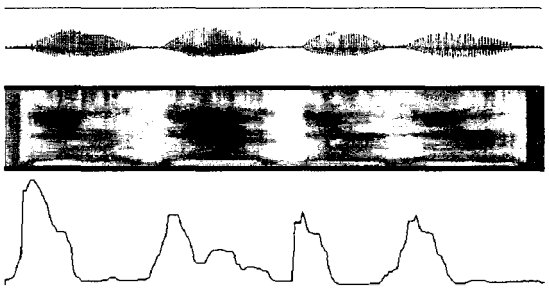

그림 3 /인 일 임 잉/의 파형, 스펙트로그램, VB

(6) 각 음절의 평균과 공분산

먼저 성명데이터(1145개, 62명), 단음절데이터(521개, 53명)을 이용하여 655개의 음절 데이터로부터 각 음절의 안정구간의 모음 영역의 평균과 공분산을 구한다. 입력된 언어학적 정보를 이용하여 분할하고자 하는 각 음절의 평균과 공분산을 이용한 pdf값을 사용하여 분할 파라미터들 의해 결정된 모음 영역의 안정된 위치를 결정한다. 이때 평균과 공분산이 없는 음절인 경우에는 초성은 /ㅇ/으로 종성은 종성법칙을 적용하여 적절하게 변환한 후, 음절과 공분산 정보를 이용하여 각 음절의 pdf값을 구한다.

III. 실험 및 결과

실험에 사용한 데이터는 남성화자 10명이 발성한 단음절 521개 데이터를 사용하였다. 단어의 구성은 초성+

중성+종성이 고르게 분포하도록 구성하였다. 음성 데이터 녹음은 일반 사무실 환경에서 녹음을 실시하였으며, 11.025KHz 샘플링 주파수와 16bit 양자화로 저장하였다. 데이터의 안정된 모음 구간을 검출하고 분할하기 위해서는 음절의 특성에 따라 종성이 없는 CV, 무성 종성 자음이 있는 CVC, 유성 종성 자음이 있는 CVC로 나누어 분할 작업을 수행했다.

표 1은 화자 10명을 대상으로 얻어진 데이터를 사용하여 손으로 직접 분할작업을 한 데이터를 기준으로 자동으로 분할된 데이터와 비교하였다.

표 1 화자별 분할 정확도

발성자	정확도
화자 1	86.6%
화자 2	81.4%
화자 3	84.3%
화자 4	96.4%
화자 5	96.5%
화자 6	78.7%
화자 7	91.0%
화자 8	91.7%
화자 9	92.5%
화자 10	60.3%

화자에 따른 정확도는 수작업을 거친 분할 위치와 자동으로 생성된 분할 위치의 10ms 이내의 정확도를 나타내었다.

그림 4는 각 구간별에 따른 정확도를 나타내었다. 분할 위치에 따라 정확도를 고려하였고, 구간은 12ms이하, 24ms이하, 36ms이하, 48ms이하로 나누었다.

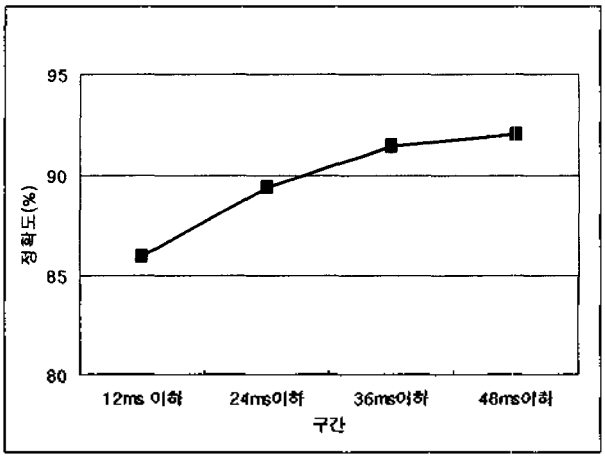

그림 4 구간별 정확도

그림 4에서 알 수 있는 바와 같이 30ms이하에서 자동분할로 얻어진 데이터는 91.5%의 정확도를 보여주고 있다.

Ⅳ. 결론

본 논문에서는 비교적 분할작업을 수월하게 할 수 있고, 인식시 분할작업에서 발생되는 약간의 오류를 포괄할 수 있으면서, 무제한 인식 시스템을 구축할 수 있는 CV, VC, VCCV 단위를 제안하였다. 이 복합음소단위는 비교적 넓은 영역을 포함하고 있는 안정된 모음 영역에서 분할하기 때문에 다른 인식 단위에 비해 분할 오류에 따른 인식률의 저하를 피할 수 있을 것이며, 훈련데이터에 포함되어 있지 않는 VCCV모델을 VC와 CV모델 결합으로 사용할 수 있기 때문에 67032 개의 VCCV모델의 구성 없이도 무제한 인식 시스템을 구현할 수 있는 인식단위이다.

참고 문헌

[1] 김순협 외 4인, "음소 단위에 의한 한국어 연속 숫자음 인식에 관한 연구," 한국음향학회지 제 8권 3호, pp. 5-15, 1989.

[2] 박현상 외 3인, "Diphone단위의 hidden Markov model을 이용한 한국어 단어인식," 한국음향학회지 제 13권 1호, pp. 14-23, 1994.

[3] 윤재선, 홍광석, "반음절 단위HMM을 이용한 연속 숫자 음성인식," 한국음향학회지 제17권 제5호, pp.73-78, 1998.

[4] 김태환, 박순철, "문맥종속 반음소 단위 모델을 이용한 자동 음소분할 및 레이블링 시스템의 구현," 한국음향학회, 제 17권 2호, 1998.