2002년도 한국음향학회 학술발표대회 논문집 제21권 제1(s)호

자동 음성 분할 시스템의 성능 향상

김무중*, 권철홍**

*㈜언어과학 음성공학연구소, ** 대전대학교 컴퓨터정보통신공학부

An improved automatic segmentation algorithm

Kim, Mu Jung*, Kwon, Chul Hong**.

*Eoneo Inc., **Daejeon University.

*donaldos@eoneo.co.kr, **chkwon@dju.ac.kr

요 약

본 논문에서는 한국어 음성 합성기 데이터베이스 구축을 위하여 HMM을 이용하여 자동으로 음소경계를 추출하고, 음성 파라미터를 이용하여 그 결과를 보정하는 반자동 음성분할 시스템을 구현하였다. 개발된 시스템은 16KHz로 샘플링된 음성을 대상으로 삼았고, 레이블링 단위인 음소는 39개를 선정하였고, 음운현상을 고려한 확장 모노폰도 선정하였다. 그리고 언어학적 입력방식으로는 음소표기와 철자표기를 사용하였으며, 패턴매칭 방법으로는 HMM을 이용하였다. 유성음/무성음/묵음 구간 분류에는 ZCR, Log Energy, 주파수 대역별 에너지 분포 등의 파라미터를 사용하였다.

개발된 시스템의 훈련된 음성은 정치, 경제, 사회,문화, 날씨 등의 코퍼스를 사용하였으며, 성능평가를 위해 훈련에 사용되지 않은 문장 데이터베이스에 대해서 자동 음성 분할 실험을 수행하였다. 실험 결과, 수작업에 의해서 분할된 음소경계 위치와의 오차가 10ms 이내가 87%, 30ms 이내가 91%가 포함되었다.

1. 서 론

음성 데이터를 음소 단위로 분할 및 레이블링하는 작업은 음성합성 및 음성인식에서 기반이 되는 일이다. 코퍼스 기반 음성합성에서 각 음소의 특징 파라미터 및 지속시간을 정확히 추출하는데 음소 분할된 음성 데이터베이스가 필요하며, 음성인식 시스템의 훈련과정에서 음소단위의 시간정보가 정확히 입력되면, 인식성능의 향상을 가져온다[1].

코퍼스 기반 접근 방식의 대용량 음성 데이터베이스에서 음소와 같은 기본 단위들로 분할하고 레이블링하는 과정은 사람이 직접수행 할 수 있지만, 반복되는 스펙트로그램 및 파형의 판독 등은 음성학적 지식을 요하며, 반복되는 작업으로 많은 시간을 소비하게 되며, 작업자에 따라 음소 경계 선정에서 주관적인 작업으로 일관성을 보장 받지 못하게 된다. 부분적으로 음성분할이 자동적으로 수행될 수 있다면 위에서 언급한 문제들을 해결할 수 있으며 작업 시간을 크게 줄일 수 있다[2].

본 논문에서는 다양한 음운현상을 고려한 음소를 설정하여, HMM 모델을 생성 후 자동 음성 분할 후, 유성음/무성음/묵음 특징을 추출 후 분할 결과를 보정하여 음성 분할 결과에 향상을 가져왔다.

본 논문의 구성은, 2장에서는 자동 음성 분할 기술에 대한 내용과, 3장에서는 자동 음성 분할 시스템 구성 요소인 HMM 모델 생성 그리고 음성 파라미터 추출 및 임계치 선정과 구간 선정 그리고 보정 방법에 대한 내용과, 4장에서는 실험 결과 그리고 마지막 5장에서는 결론을 맺는다.

2. 자동 음성 분할 기술

음성 분할 기술은 크게 언어학적 정보를 사용하는 방식과 사용하지 않는 방식으로 크게 나뉜다. 다음은 언어학적 정보를 사용하는 HMM 모델을 이용한 음성분할 기술과, 언어학적 정보를 사용하지 않는 음향학적 분할 기술에 대한 개요이다.

2.1. HMM 모델을 이용한 음성 분할 기술

언어학적 정보를 사용하는 음성 분할 과정은 발화자가 음소의 빈도수를 고려한 문장을 발화하여 생성된 음성 데이터에서 음소를 기준으로 하여, 각각의 음소들에 대한 통계적 모델을 생성한다. 분할될 음성과 음소 열이 입력되면, 입력된 음소에 대해 각 음소들의 모델들을 연결해서 입력음성과 매칭시키는 과정에서 음소경계 정보가 얻어진다. 이러한 통계적 패턴 매칭 방법 중에서 대부분의 음성 분할 기술은 음성신호의 발생과정을 확률과정으로 가정한 모델인 HMM(Hidden Markov Model)모델에 기반을 둔다[2].

그림 1은 음성학적 정보를 사용한 HMM 모델을 이용한 음성 분할 기술의 흐름도이다. 각 단계를 살펴보면, 텍스트를 발음변환 프로그램을 이용하여 발음 열 형태로 전환 후 발음 열 형태를 HMM 모델에서 사용된 음소형태로 전환하고, 발화된 음성 데이터에서 HMM 모델에서 사용된 파라미터로 특징 파라미터를 추출하여 일정한 포맷으로 저장 후 입력된 음소와 특징을 기반으로 동일 음소기반 HMM 모델에서 비터비 탐색 및 정렬을 이용하여 자동으로 음소단위 경계 검출을 한다. 그러나 출력된 경계에 대한 음소 경계는 음성학 전공자들이 수작업으로 지정한 경계와 크게는 30ms 이상 차이

가 난다.

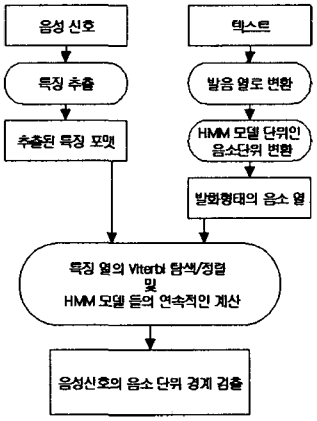

그림 1. HMM을 이용한 음소 경계 검출 방법

2.2. 음향학적 음성 분할 기술

음소표기 등의 언어학적 정보가 입력되지 않은 음성 분할 기술은, 음성신호에 포함된 음향학적 정보들을 ZCR(Zero Crossing Rate), SVF(Spectral Variation Function), MFCC(mel-frequency cepstral coefficients), LPC 계수, 에너지 그리고 RASTA(re ative spectral processing) 등을 이용하여 추출하여, Navie Bayesian 알고리즘과 back-propagation 알고리즘을 이용하여 분할 하는 기법이 있다[3].

3. 자동 음성 분할 시스템의 구성

본 장에서는 본 논문에서 구현한 한국어 자동 음성 분할 및 레이블링 시스템의 구성과 관련된 기술적인 사항이다.

3.1. 발화문장 선정 및 레이블링 단위 선정

HMM 모델 구성을 위한 텍스트 및 음성 데이터 작업에서 정치, 경제, 사회, 문화, 스포츠, 날씨 관련 문장을 수집한 1593문장을 선정하였으며, 음소 열 변환은 본 연구소에서 선정한 39개 음소 셋을, 음소 앞뒤의 음운현상을 고려한 624개의 확장 음소 셋을 선정하였다.

3.2. 음성신호 전 처리과정

발화된 음성 데이터는 약 10 음절로 이루어진 1593 문장을 발화하여 샘플링 주파수 16KHz, 부호화 비트 16bit 형식으로 스튜디오에서 녹음하였다. 일반적으로 음성인식에서는 음소경계 오차범위를 5ms 정도를 목표로 매 5ms마다 25ms 구간의 음성신호로부터 음성특징을 추출하였으며, 특징추출은 인간의 청각특성을 반영하는 특징 표현, 다양한 잡음환경/화자/채널 변이에 강인한 특징, 시간적인 변화를 잘 표현하는 특징의 추출의 관점 중 청각특성을 반영하여 12 차의 MFCC 와 delta coefficients, acceleration coefficients, 에너지, delta 에너지, acceleration 에너지의 총 39차의 차수를 가지고 있다.

3.3. 음향 모델의 구성

음성인식에 있어서 음향모델은 음성신호가 어떤 형태로 표현할 수 있는지를 나타낸다. 최근 음성 인식기에서 가장 널리 사용되는 음향모델은 HMM에 기반한 것이다. 음향모델의 기본 단위는 음소 또는 유사음소 단위이다. 각 모델은 하나의 음향모델 단위를 나타내며 보통 3개의 상태(state)로 구성된다. 주로 좌에서 우로의 상태간 천이만 허용된다. 각 상태에서의 음성특징 벡터의 관측 확률은 이산 확률분포 또는 연속 확률밀도 함수로 표현된다.

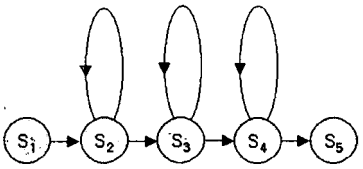

그림 2. 음소 모델링을 위한 HMM 구조

그림 2는 본 논문에서 사용한 모델의 형태이며, 가장 널리 사용되는 음소에 대한 HMM 모델을 나타낸다. S_1과 S_5는 각각 모델의 시작과 끝을 나타내는 가상적인 상태이다.

3.4. 모델 생성

모델을 훈련하는 과정을 살펴보면, 첫번째, flat start 방식을 이용하여 모델을 생성하여, HMM 모델을 이용한 자동 정렬 후 정렬된 데이타를, 음성학 전공자들에 의해 음성분할 및 레이블링 작업을 하여, boot strap 방식을 이용하여 모델을 생성하였다. 다음 그림 3은 모델 생성 과정을 나타낸 것이다.

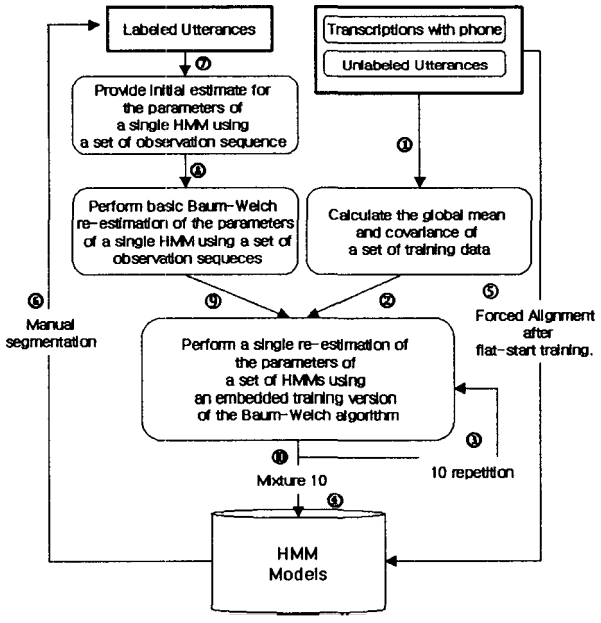

그림 3. 음소기반 HMM 모델 생성과정

3.5. HMM 자동 분할 결과 보정 알고리즘

본 논문에서는 3.4 절의 방식으로 생성된 HMM 모델을 통하여 음소분할 한 결과에 유성음/무성음/묵음 분리 파라미터를 이용하여 후처리 적용, 분할 결과를 향상시켰다.

성음/무성음/묵음을 분류하는 파라미터를 살펴보면, 크게 Log Energy, Zero Crossing Rate, Level Crossing Rate, Normalized Autocorrelation Coefficient at unit sample delay, Spectral Distribution 등이 있다. 특히 스펙트럼 분석에서 유성음의 스펙트럼은 1KHz 이하에서 대부분의 에너지가 나타나며, 무성음은 2.5KHz 이상에 집중 되어 있어 유성음/무성음/묵음 프레임을 선명하게 구분 지을 수 있다[3][4].

3.5.1. 특징추출

훈련 데이터 중 500개 샘플을 취하여, 5ms 길이로 시간 도메인에서는 ZCR, Log Energy을 추출하였으며, 주파수 도메인에서는 5개 대역의 스펙트럼 에너지를 추출하였다. 다음은 특징 추출에 사용된 수식 및 개념이다[4].

1) ZCR (Zero Crossing Rate)

$$Sign[s(n)\times s(n+1)] < 0 \qquad \text{식 3.1}$$

2) Log Energy

$$E_m = \log \sum_{n=1}^{N} s_m^2(n) \qquad \text{식 3.2}$$

3) Spectral Energy

음성을 5ms 구간으로 1024-point FFT을 이용하여 구하였다. 각 대역은 150Hz - 500Hz, 1000Hz - 1400Hz, 2500-2900Hz, 3500Hz - 3900Hz, 4500 Hz - 4900Hz 대역에서 추출하였다.

일반적으로 유성음의 에너지는 무성음의 에너지보다 높으며 무성음의 에너지는 묵음의 에너지 보다 높으므로 에너지는 유성음과 무성음 그리고 음성구간과 묵음 구간을 분류하는데 중요한 역할을 한다. 그리고 ZCR은 무성음에서 유성음과 묵음보다 많이 나타나며, 유성음 구간이 일반적으로 잡음이 없는 묵음구간보다 약간 많이 나타난다[5].

3.5.2. 임계치 설정 및 경계검출

500개의 샘플 데이터를 통하여, 유성음 구간, 무성음 구간, 묵음 구간의 위의 3가지 파라미터를 추출하여 평균과 표준 편차를 이용하여 유성음, 무성음, 묵음 프레임 판단 임계치를 설정하고, 판정된 유성음, 무성음, 묵음 프레임과 샘플데이타를 비교하여 23개 구간 내 프레임 패턴을 분석하여, 제일 먼저 묵음구간(묵음프레임이 200ms 이상 연속)을 설정 후 유성음/ 무성음 프레임구간을 설정하였다. 그림 4는 음성특징 추출과정과 추출된 특징을 통해 유성음/무성음/묵음 분류를 위한 파라미터 임계치 설정 과정이다.

3.5.3. HMM 모델을 이용한 음성분할 결과 보정 알고리즘

다음은 결과 보정 단계를 요약한 내용이다.

step 1. 묵음 구간 검출 결과를 HMM 모델을 이용한 음성 분할 결과에 100% 가중치로 보정.

step 2. 자동 분할 결과에서 [무성음+유성음], [유성음+무성음] 경계 점에서 음성 파라미터를 이용한 유성음/무성음 구간 정보를 ±15ms 구간으로 탐색 및 보정.

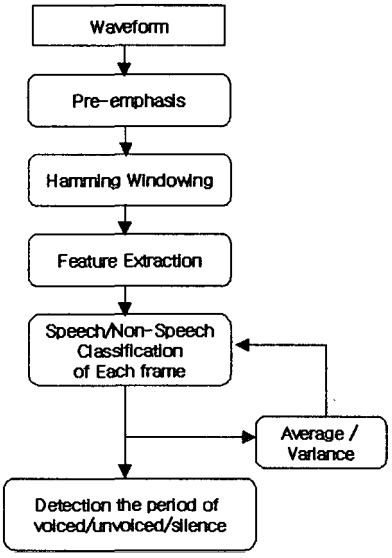

그림 4. 음성 특징 파라미터 추출 및 임계치 설정 과정

step 3. 유성음+유성음 구간에서 spectral distribution 정보로 급격히 값이 변하는 구간을 탐색 및 보정.

4. 실험 결과

본 논문에서 음소의 기본 단위로 확장 모노폰 셋을 선정하여 HMM 모델을 생성하였으며, 생성된 HMM 모델을 이용하여 자동 음소 분할 후, 음성 특징 파라미터를 추출하여 유성음/무성음 구간을 분류하여 자동 음소 분할 결과를 보정하는 시스템을 제안하였다.

자동 음소 분할 및 보정 알고리즘의 실험은 훈련 데이터로 사용되지 않은 외국 관련 뉴스 문장,증권 관련 경제 문장, 교통, 모음으로 이루어진 단어, 외래어 및 기구축된 음성 합성기에 사용 될 DB 중 빈도수가 낮은 트라이폰을 추출하여 생성된 단어들을 대상으로 테스트 데이터를 구성하였으며, 스튜디오에서 16KHz, 16Bit PCM 데이터로 디지털화 하였다. 성능비교에서는 자동 음소 분할 후, 수작업으로 분할한 음소 경계를 비교한 내용이다.

표 1은 모노폰으로 훈련된 HMM 모델을 이용한 음소 분할 결과와 확장 모노폰으로 훈련된 HMM 모델을 이용한 음소 분할 결과를 보여주고 있다.

표1. 모노폰 및 확장 모노폰 HMM을 이용한 자동 음성 분할 결과

음운현상	오차범위	모노폰모델	확장모노폰모델
Sil + speech	10ms <=	92%	93%
cc+vv	10ms <=	79%	82%

cc+cv	20ms <=	67%	78%
cc+vc	10ms <=	73%	77%
cc+cc	10ms <=	77%	81%
vc+vv	20ms <=	73%	77%
vc+vc	20ms <=	74%	81%
vc+cc	10ms <=	79%	82%
vv+vv	20ms <=	70%	73%
vv+vc	20ms <=	72%	73%
vv+cc	10ms <=	74%	83%
cv+vv	10ms <=	77%	82%
cv+vc	10ms <=	76%	79%
cv+cc	10ms <=	73%	81%

cc : 무성자음, vc: 유성화 자음
vv : 유성모음, cv: 무성화 모음.
sil : 묵음 speech : 음성구간

위의 결과를 살펴보면 모노폰 모델을 사용한 음소분할 결과보다 확장 모노폰 모델을 사용한 음소분할 결과가 최대 11%까지 성능이 향상된 것을 볼 수 있다. 특히 자음과 모음의 경계는 다른 음운환경의 경계보다 더 정밀한 결과를 도출할 수 있었다. 이 결과를 음성 파라미터를 추출하여 분류한 유성음/무성음/묵음 구간 결과로 보정한 결과를 살펴보면 표 2와 같다.

표 2. HMM을 이용한 자동 음성 분할 후 보정 결과

음운현상	오차범위	확장모노폰모델	보정후
sil + 음성	10ms <=	93%	100%
cc+vv	10ms <=	82%	97%
cc+cv	20ms <=	78%	79%
cc+vc	10ms <=	77%	92%
cc+cc	10ms <=	81%	82%
vc+vv	20ms <=	77%	81%
vc+vc	20ms <=	81%	93%
vc+cc	10ms <=	82%	92%
vv+vv	20ms <=	73%	73%
vv+vc	20ms <=	73%	78%
vv+cc	10ms <=	83%	94%
cv+vv	10ms <=	82%	86%
cv+vc	10ms <=	79%	87%
cv+cc	10ms <=	81%	82%

보정 후 결과에서 음성의 파라미터의 구분이 정확하게 발생하는 무성음과 유성음의 경계 / 묵음과 음성의 경계에 보정 결과는 크게 향상되었으나, 유성음과 유성음 무성음과 무성음 경계는 보정 결과 향상정도가 미비하였다. 이는 수정 보정에서도 유성음과 유성음 경계 그리고 무성음과 무성음의 경계 구분이 음성학자 견해가 상당히 다름을 시사하기도 한다.

5. 결론

본 논문에서는 확장 모노폰 모델을 선정하고 HMM 모델을 생성하여 자동 음성 분할 결과를 얻었다. 이는 모노폰 모델을 선정, 생성한 HMM 모델을 이용한 자동 음성 분할 결과보다 최대 11%의 성능향상을 가지고 왔으며, 이 결과에 음성 변화 특성의 정보를 잘 표현하는 ZCR, Log Energy, Spectral Distribution을 이용한 유성음, 무성음, 묵음 구간 설정으로 보정하여 추가적인 성능 향상을 얻을 수 있었다.
본 시스템은 음소단위로 분할된 방대한 양의 음성합성기 데이터베이스를 구축하는데 기존의 수작업에 의한 방식에 비해서 많은 시간과 비용을 절감할 수 있었으며, 이러한 환경을 통해서 앞으로 방대한 양의 음성인식기 및 음성 합성기 데이터 베이스를 구축한다면 정교한 음소 모델구현에 큰 역할을 할 것으로 판단되며, 추후 확장 모노폰 셋을 이용하여 triphone으로 확장하여 시스템을 구축하며, 유성음과 유성음 경계검출 파라미터를 연구하여 적용하면 더 나은 성능을 얻을 수 있을 것 이다.

참고문헌

[1] T. Svendsen and F.K. Siong, "On the automatic segmentation of speech signal", in Proc. IEEE Int, Conf. Acoust., Speech Signal Processing, pp. 77-80, Apr., 1987.
[2] F. Brugnara, et al., "Automatic sementation and labeling of speech based on hidden Markov model", Speech Communication, Vol. 12, pp. 357-370, 1993.
[3] L.R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances". The Bell system technical journal, Vol. 54, No. 2, pp. 297-315, Feb., 1995.
[4] R. Sarikaya and H.L. John, "Robust speech activity detection in the presence of noise", ICSLP '98, Vol. 4, pp. 1455-1458, 1998.
[5] 윤석현, 유창동, "시간-주파수 영역에서 음성/잡음 우세 결정에 의한 새로운 잡음처리," 음향학회지 20권 3호, pp. 48-55, April, 2001.