HMM 및 보정 알고리즘을 이용한 자동 음성 분할 시스템*

An Automatic Segmentation System Based on HMM and Correction Algorithm

김 무 중** · 권 철 홍***
Mu Jung Kim · Chul Hong Kwon

ABSTRACT

In this paper we propose an automatic segmentation system that outputs the time alignment information of phoneme boundary using Viterbi search with HMM (Hidden Markov Model) and corrects these results by an UVS (unvoiced/voiced/silence) classification algorithm. We selecte a set of 39 monophones and a set of 647 extended phones for HMM models. For the UVS classification we use the feature parameters such as ZCR (Zero Crossing Rate), log energy, spectral distribution. The result of forced alignment using the extended phone set is 11% better than that of the monophone set. The UVS classification algorithm shows high performance to correct the segmentation results.

Keywords: Automatic Segmentation, HMM, Correction Algorithm

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]

음성 분할 기술은 언어학적 정보를 사용하는 HMM을 이용한 음성분할 방식과, 언어학적 정보를 사용하지 않는 음향학적 음성분할 방식으로 나뉜다. 언어학적 정보를 사용하는 음성분할 방식은, 발화자가 음소의 빈도 수를 고려한 문장을 발화하여 생성된 음성 데이터에서, 음소를 기준으로 하여 각각의 음소들에 대한 통계적 모델을 생성한다. 음성데이터와 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 모델에 기반을 둔

* 본 연구는 한국과학재단 목적기초연구(R01-2002-000-00283-0) 지원으로 수행되었음.
** (주) 언어과학 음성공학연구소
*** 대전대학교 컴퓨터정보통신공학부

다.[2] 음소표기 등의 언어학적 정보가 입력되지 않는 음향학적 음성분할 방식은, ZCR, SVF (Spectral Variation Function), MFCC (mel-frequency cepstral coefficients), LPC 계수, 에너지 그리고 RASTA (relative spectral processing) 등을 이용하여 음성신호에 포함된 음향학적 정보들을 추출하여, Navie Bayesian 알고리즘과 Back-propagation 알고리즘을 이용하여 음성을 분할한다.[3]

본 논문에서는 먼저 언어학적 정보를 사용하는 음성분할 방식을 이용하여 음성을 분할한 뒤, 음향학적 음성 분할 방식으로 분할 결과를 보정하는 2 단계 자동 음성분할 기법을 제안한다. 첫 번째 단계에서는 모노폰 및 다양한 음운현상을 고려한 단순화된 트라이폰을 분할단위로 설정하여 HMM 모델을 생성하여 자동 음성 분할을 수행하고, 두 번째 단계에서는 유성음/무성음/묵음 특징을 추출 후 분할 결과를 보정하여 음성 분할 결과에 향상을 가져온다.

본 논문의 구성은, 2 장에서는 본 논문에서 제안한 자동 음성 분할 시스템을 설명하고, HMM 모델의 생성 방법, 그리고 보정 알고리즘에 사용된 음성 파라미터의 추출 및 임계치 선정과 보정 방법에 대한 내용을, 3 장에서는 실험 결과를 기술하고, 그리고 4 장에서 결론을 맺는다.

2. 제안된 자동 음성 분할 시스템

본 장에서는 제안한 자동 음성 분할 및 레이블링 시스템의 구성에 대하여 설명한다.

2.1 음성신호 전 처리과정 및 음향 모델 구성

음성 전사 단위는 자음 19 개, 모음 18 개, 묵음(silence) 및 짧은 휴지부(short pause) 등 총 39 개 모노폰으로 분할 단위를 선정하여 HMM 모델을 생성하였다. 그런데, 동일한 음소일지라도 좌우 음운환경에 따라 음소의 특징이 상이하다고 알려져 있다. 즉, 자음은 여러 가지 음운 현상에 의해 변이음들이 나타나며, 모음도 앞뒤 음운환경의 영향으로 지속구간이 상이한 특징을 보여 주므로, 이를 HMM 모델 생성에 반영하기 위해 39 개 모노폰의 앞뒤 음운환경을 고려한 단순화된 트라이폰 647 개를 선정하여 HMM 모델을 생성하였다. 예를 들어, '감'은 'gx+ax+mx'로 전사된다. 'ax'의 앞 음소 'gx'는 무성자음이고 뒤 음소 'mx'는 유성자음이다. 이 경우에 모노폰 'ax'는 단순화된 트라이폰 'ccaxvc'로 변환된다. 여기에서 cc는 무성자음을 vc는 유성자음을 의미한다. 이와 같이 단순화된 트라이폰의 앞, 뒤 문맥은 무성자음(cc), 유성자음(vc), 유성모음(vv), 무성모음(cv) 등 4 가지 종류가 올 수 있다.

본 연구의 목적은 무제한 TTS 시스템의 음성 DB를 구축하기 위하여 1인 여성 화자의 발성 파일에 대한 자동 음성 분할 시스템의 구축이다. 1 인 여성 화자의 HMM 모델 생성을 위한 텍스트 및 음성 데이터는 정치, 경제, 사회, 문화, 스포츠, 날씨 관련 문장 등 총 1,593문장을 선정하였다. 본 연구에서는 음소경계의 오차범위를 5 ms 정도를 목표로 하여 매 5 ms마다 25 ms 구간의 음성신호로부터 음성특징을 추출하였으며, 특징 파라미터는 인간의 청각특성을 반영하는 MFCC 계수를 사용하여 12 차 MFCC, delta coefficients, acceleration coefficients, 에너지, delta 에너지, acceleration 에너지 등 총 39 개의 파라미터를 사용하였다.

음성인식에 있어서 음향모델은 음성신호가 어떤 형태로 표현될 수 있는지를 나타낸다. 음향모델의 기본 단위는 음소 또는 유사음소 단위이다. HMM에서 각 모델은 하나의 음향모델 단위를 나타내며 보통 3 개의 상태로 구성되며, 주로 좌에서 우로의 상태간 천이만 허용된다. 각 상태에서의 음성특징 벡터의 관측 확률은 이산 확률분포 또는 연속 확률밀도함수로 표현된다. 그림 1은 본 연구에서 채택한 음소에 대한 HMM 음향 모델이다.

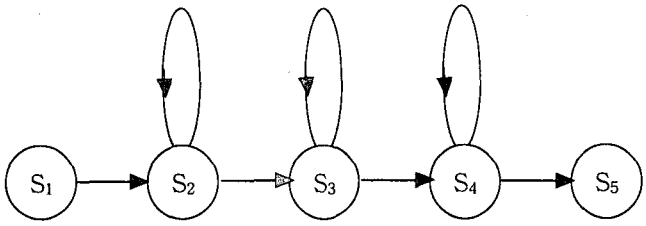

그림 1. 3-state 음향모델

2.2 HMM 모델을 이용한 자동 음성 분할 방식

그림 2는 본 논문에서 제안한 HMM 모델을 이용한 음성 분할 방식의 흐름도이다. 각 단계를 살펴보면, 텍스트를 발음변환 프로그램을 이용하여 발음 열 형태로 전환 후, 발음 열 형태를 HMM 모델에서 사용된 음향 모델 단위인 음소 형태로 변환하고, 발화된 음성 데이터에서 HMM 모델에서 사용되는 특징 파라미터를 추출하여 일정한 특징 벡터열로 저장 후, 입력된 음소 표기와 특징 파라미터를 기반으로 훈련된 음소기반 HMM 모델에서 비터비(Viterbi) 탐색 및 정렬을 이용하여 자동으로 음소단위 경계 검출을 한다.

HMM 모델을 생성하기 위한 방법에는 flat start 방식과 bootstrap 방식으로 나눌 수 있다. 본 시스템에 사용된 모델은 위의 두 방식을 모두 사용하여 두 단계를 통해 생성된 것이며, 그림 3과 같은 처리방법을 통해 생성하였다. Flat start 방식을 사용한 첫 번째 단계에서는, 1,593문장에 대한 음성데이터 파일과 시간정보가 없는 음소기반의 전사파일을 입력으로, 음성데이터의 전체 평균과 분산을 계산하여 HMM 초기 프로토타입 모델을 생성한다. 생성된 프로토타입 HMM 모델을 각 음소의 HMM 모델로 확장하고, 각 음소 모델에 대해 Baum-Welch Re-estimation을 이용하여 파라미터 재추정 단계를 통하여 HMM 파라미터 훈련 과정을 5 회 반복하고, 생성된 묵음 HMM 모델에 대해 상태 천이 과정을 추가하고, 짧은 휴지부 모델에 대하여 1-state HMM 모델을 생성한다. 이렇게 생성한 묵음, 짧은 휴지부 및 전체 음소 HMM 모델에 대하여 Baum-Welch Re-estimation을 이용하여 파라미터 재측정 단계를 통하여 HMM 파라미터 훈련 과정을 5 회 반복하고, Mixture 수를 7 개까지 늘려가며 훈련하여 1 차 HMM 모델을 생성한다.

첫 번째 단계에서 생성된 모델을 기반으로 비터비 디코딩 과정을 수행하여 음소 정렬과 시간정보 추출과정을 거쳐, 1,593 개의 음성 데이터 파일의 시간정보를 포함한 전사정보를 구하여, 이 시간정보와 전사정보를 수작업을 통하여 보정한다.

Bootstrap 방식을 사용한 두 번째 단계에서는, 앞에서 구한 시간정보를 포함한 전사파일과 음성 데이터 파일을 입력으로, 음소의 시간정보를 이용하여 분할 후, 각 음소에 대하여 평

균과 분산을 구하고, segmental K-means 알고리즘을 적용하여 파라미터 값의 초기 셋을 계산하며, 비터비 알고리즘을 이용하여 각 훈련 데이터에 일치하는 가장 유사한 상태 열을 찾고 파라미터 값들을 재조정하여, 이러한 과정을 반복적으로 수행하여 수렴하는 과정을 거쳐 HMM 초기 처리과정을 수행한다. 초기 생성된 모델을 Baum-Welch 방식을 이용하여 각 파라미터에 대해 재측정 과정을 거치며, Forward-Backward 알고리즘을 이용하여 각 시간 프레임에서 각 상태의 천이 확률을 계산하여 모델을 생성한다. 생성된 모델에 Baum-Welch 알고리즘을 적용하여 최종 모델을 생성하게 된다. 최종 생성된 모델을 기반으로 비터비 디코딩 과정을 통해 최종 음성 세그멘테이션 정보를 얻게 된다.

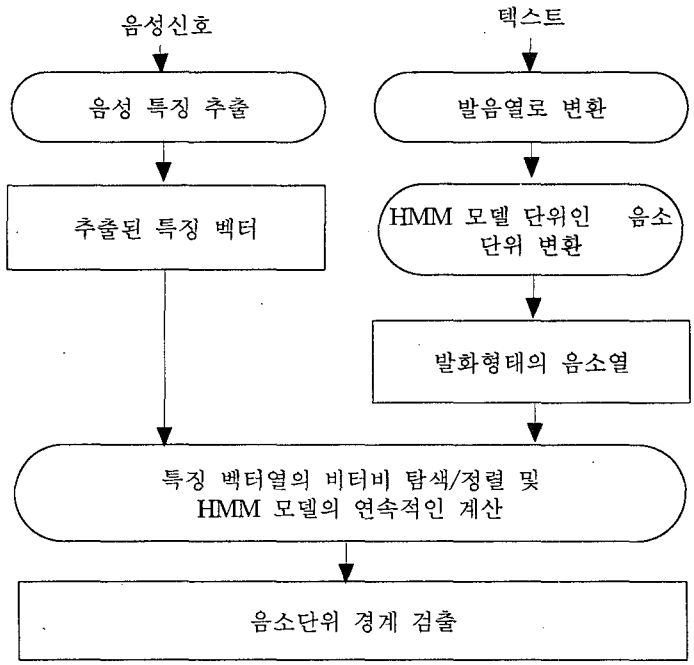

그림 2. HMM을 이용한 자동 음소 경계 검출 흐름도

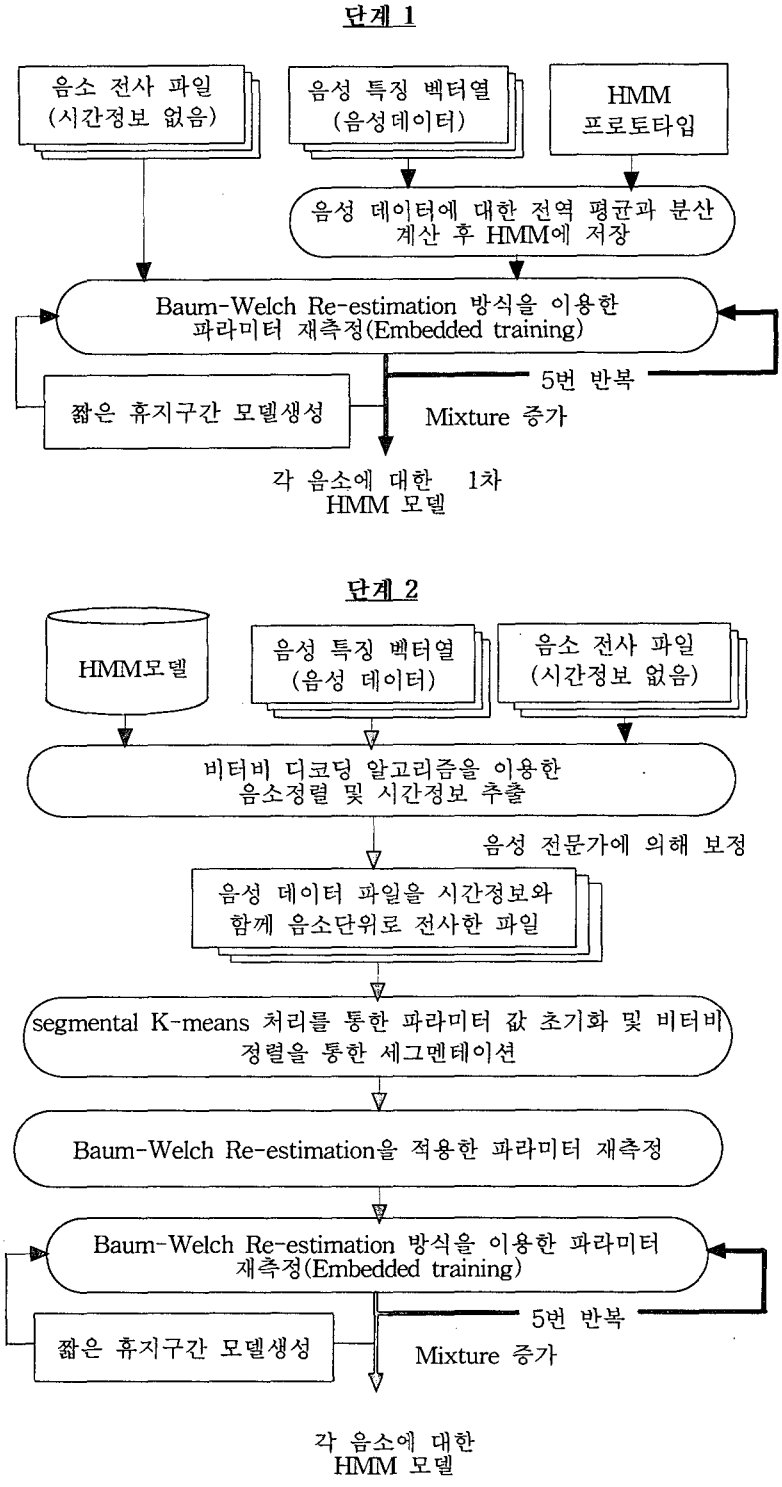

그림 3. 제안된 HMM 모델 생성 과정

2.3 HMM 자동 분할 결과 보정 알고리즘

본 절에서는, 앞에서 제시한 방식으로 자동으로 음소 분할한 결과에, 유성음/무성음/묵음 구간을 분리하는 특징 파라미터를 이용하여 분할 결과를 보정하는 방법을 제안한다. 유성음/무성음/묵음을 분류하는 파라미터를 살펴보면,[3][4] ZCR, 로그 에너지, Level Crossing Rate, Normalized Autocorrelation Coefficient at unit sample delay, Spectral Distribution 등이 있다.

2.3.1 특징 파라미터 추출

훈련 데이터 중 500 문장을 선정하여, 5 ms 단위로 시간 도메인에서 ZCR, 로그 에너지를 추출하였으며, 주파수 도메인에서는 5 개 대역의 스펙트럼 에너지를 추출하였다. 다음은 특징 추출에 사용된 파라미터에 대한 설명이다.

1) ZCR

ZCR은 음성신호가 영점을 교차하는 한 프레임 내 횟수이다. ZCR은 무성음 구간에서 유성음과 묵음보다 크며, 유성음 구간이 일반적으로 잡음이 없는 묵음 구간보다 약간 많이 나타난다.[5]

2) 로그 에너지

$$E_m = \log \sum_{n=1}^{N} s_m^2(n) \tag{1}$$

유성음의 에너지는 무성음의 에너지보다 높으며 무성음의 에너지는 묵음의 에너지 보다 높으므로, 에너지는 유성음과 무성음 그리고 음성구간과 묵음구간을 분류하는데 중요한 역할을 한다.

3) Spectral Distribution

음성 5 ms 구간마다 1024-point FFT를 이용하여 스펙트럼 에너지를 구하였다. 주파수 대역을 5 개로 분리하여 각 대역별로 에너지를 구했는데, 각 대역은 200-800 Hz, 800-1800 Hz, 1800-3000 Hz, 3500-4500 Hz, 4500-8000 Hz 등이다. 스펙트럼 분석에서 유성음의 스펙트럼은 1 KHz 이하에서 대부분의 에너지가 나타나며, 무성음은 2.5 KHz 이상에 에너지가 집중되어, Spectral Distribution은 유성음/무성음 구간을 선명하게 구분하는 파라미터로 이용할 수 있다.

유성자음과 유성모음의 특징은 ZCR는 크게 차이가 없으나, 유성자음은 유성모음에 비해 로그 에너지가 낮으며, 주파수 대역에서의 에너지 분포도 저주파 대역의 에너지가 유성모음에 비해 상당히 낮은 특징을 보여준다.

2.3.2 음소 경계 보정 알고리즘

500 문장에 대하여 위 3 가지 파라미터를 추출하여 평균과 분산을 구하여, 무성자음, 유성

자음, 유성모음, 무성모음, 묵음 구간을 구분하는 임계치를 설정하고, 먼저 묵음 구간(묵음 구간이 200 ms 이상 연속인 경우)을 판정한 뒤, 무성자음/유성자음/유성모음/무성모음 구간을 판정하였다.

다음은 음소 경계 보정 알고리즘에 대한 설명이다.

단계 1. 유성음/무성음/묵음 분류 알고리즘으로 검출된 묵음구간을 HMM을 이용한 자동 음소 경계 검출 결과에 100% 적용하여 경계 위치를 보정한다.

단계 2. 자동 음소 분할 결과의 짧은 휴지부 지점에서, 로그 에너지 및 주파수 대역 에너지가 현저하게 작은 프레임을 좌우 5 ms 간격으로 검색하여 존재하면 경계 위치를 조정한다.

단계 3. 자동 음소 분할 결과의 무성음+유성음, 유성음+무성음 경계 위치에서, 좌우15 ms 구간을 검색하여 유성음/무성음/묵음 분류 알고리즘에서 도출된 유성음/무성음 경계 정보가 존재하면 경계 위치를 보정한다.

단계 4. 자동 분할 결과에서 유성음+유성음 경계 위치에서 스펙트럼 에너지가 급격히 변하는 구간을 탐색하여 존재하면 경계 위치를 조정한다.

3. 실험 결과

자동 음소 분할 및 보정 알고리즘에 대한 실험을 위하여, 훈련 데이터는 1,593 문장을 선정하였으며, 테스트 데이터는 훈련 데이터로 사용되지 않은 11,000 문장으로 구성하였고, 각 문장은 평균 10 음절로 이루어져 있고, 샘플링 주파수 16 KHz, 부호화 비트 16 bit 형식으로 녹음하였다. 실험 방법은 모노폰 및 단순화된 트라이폰 모델을 이용하여 자동 분할한 1 차 결과를 구하고, 이 중에서 성능이 좋은 단순화된 트라이폰 모델을 이용한 자동 분할 결과에 보정 알고리즘을 적용하여 2 차 결과를 얻었다. 그리고 성능 분석을 위하여 자동 분할한 결과를 4 인의 실험음성학 전문가들이 수동 보정하여 수동 분할 결과를 구하였다. 음성학 전문가들의 수동 분할은 본 실험실의 규정과 서로의 의견교환 및 동일한 툴을 사용하여 작업의 일관성을 유지하였다.

표 1은 모노폰으로 훈련된 HMM 모델을 이용한 음소 분할 결과와 단순화된 트라이폰으로 훈련된 HMM 모델을 이용한 음소 분할한 결과를 보여준다. 표 1의 결과를 살펴보면, 모노폰 모델을 사용한 음소분할 결과보다 단순화된 트라이폰 모델을 사용한 음소분할 결과가 최대 11%까지 성능이 향상된 것(cc+cv)을 볼 수 있다. 특히 유성음과 무성음의 경계(cc+cv, vv+cc, cv+cc)와 같이 경계 구분이 뚜렷한 음운환경의 경우는 다른 음운환경의 경계보다 더 정밀한 결과를 도출할 수 있었다. vc+vv, vc+vc, vv+vv, vv+vc인 경우는 유성음과 유성음의 경계이므로 비교적 낮은 결과를 도출해 냈으며, 수작업시 음성학 전문가들의 견해가 상이하여 오차 범위를 20 msec로 규정하였다.

이 결과에 유성음/무성음/묵음 구간 분류 알고리즘을 이용하여 보정한 결과를 살펴보면 표 2와 같다. 보정 결과에서, 구간의 구분이 용이한 무성음과 유성음의 경계(cc+vv, cc+vc,

vc+cc, vv+cc, cv+vc)와 묵음과 음성의 경계(silence+speech)에 보정 결과는 크게 향상되었으나, 유성음과 유성음(vc+vv, vv+vv) 및 무성음과 무성음의 경계(cc+cv, cc+cc, cv+cc)에서는 보정 결과 향상 정도가 미흡하였다. 이는 수동 보정에서도 유성음과 유성음 경계 그리고 무성음과 무성음의 경계 구분이 음성학 전문가 사이에서도 견해가 상당히 다름을 시사하기도 한다.

표 3은 자동 음소 분할 결과에 보정 방법을 반영한 정도를 보여주는 음운환경별 보정 비율을 나타내고 있다. 보정 알고리즘의 단계별로 살펴보면, 묵음 구간은 100% 보정이 적용되었으며, 무성음+유성음 경계는 ZCR, 주파수별 에너지의 구분이 뚜렷하기 때문에 72%의 보정 결과가 있었으며, 유성음+무성음 경계의 경우는 무성음이 유성음화 되어 경계 검출이 쉽지 않으므로 53% 정도의 보정 결과를 가지고 왔다. 유성자음과 유성모음의 경계구간은 유성자음 특징 중 ZCR은 유성모음과 비슷한 특징을 보이나, 로그 에너지와 주파수 대역에서 모음 성분에 비해 에너지가 낮다는 특징을 고려하여 보정하였는데 39%의 보정 비율을 보였다. 반면에 유성모음+유성모음의 경우 보정 비율은 약 12%를 나타냈으나, 수작업 세그멘테이션 결과와 상이한 경향이 있어 최종 결과에는 반영하지 않았다.

표 1. 모노폰 및 단순화된 트라이폰 HMM 모델을 이용한 자동 음성 분할 결과

음운현상	오차범위	모노폰	단순화된 트라이폰
silence+speech	10msec <=	92 %	93 %
cc+vv	10msec <=	79 %	82 %
cc+cv	10msec <=	67 %	78 %
cc+vc	10msec <=	73 %	77 %
cc+cc	10msec <=	77 %	81 %
vc+vv	20msec <=	73 %	77 %
vc+vc	20msec <=	74 %	81 %
vc+cc	10msec <=	79 %	82 %
vv+vv	20msec <=	70 %	73 %
vv+vc	20msec <=	72 %	73 %
vv+cc	10msec <=	74 %	83 %
cv+vv	10msec <=	77 %	82 %
cv+vc	10msec <=	76 %	79 %
cv+cc	10msec <=	73 %	81 %

cc: 무성자음, vc: 유성자음, vv: 유성모음, cv: 무성모음, silence: 묵음, speech: 음성구간

표 2. 유성음/무성음/묵음 구간 분류 알고리즘을 이용하여 보정한 결과

음운현상	오차범위	단순화된 트라이폰	결과 보정후
silence+speech	10msec <=	93 %	100 %
cc+vv	10msec <=	82 %	97 %
cc+cv	10msec <=	78 %	79 %
cc+vc	10msec <=	77 %	92 %
cc+cc	10msec <=	81 %	82 %
vc+vv	20msec <=	77 %	81 %
vc+vc	20msec <=	81 %	93 %
vc+cc	10msec <=	82 %	92 %
vv+vv	20msec <=	73 %	73 %
vv+vc	20msec <=	73 %	78 %
vv+cc	10msec <=	83 %	94 %
cv+vv	10msec <=	82 %	86 %
cv+vc	10msec <=	79 %	87 %
cv+cc	10msec <=	81 %	82 %

표 3. 음운환경에 따른 보정 비율

음운환경	보정 비율	비고
묵음 구간	100%	
짧은 휴지부 구간	82%	
무성음+유성음 경계	72%	
유성음+무성음 경계	53%	
유성자음+유성모음 경계	39%	
유성모음+유성모음 경계	12%	미반영

4. 결 론

본 연구 결과 단순화된 트라이폰 모델을 선정하고 HMM 모델을 생성한 자동 음성 분할 방식이 좋은 결과를 얻었다. 이는 모노폰 모델을 선정하여 생성한 HMM 모델을 이용한 자동 음성 분할 결과보다 최대 11%의 성능 향상을 가져 왔으며, 이 결과에 음성 변화 특성의 정보를 잘 표현하는 ZCR, 로그 에너지, Spectral Distribution을 이용한 유성음/무성음/묵음 구간 설정으로 보정하여 추가적인 성능 향상을 얻을 수 있었다.

본 시스템은 방대한 양의 음성인식 및 음성 합성을 위한 데이터베이스 구축에 이용할 수 있으며, 정교한 음소 모델 구현에 큰 역할을 할 것으로 판단된다. 추후 단순화된 트라이폰 셋을 일반적인 또는 군집화된 트라이폰으로 확장하여 시스템을 구축하고, 유성음과 유성음 경계검출 파라미터를 연구하여 적용하면 더 나은 성능을 얻을 수 있을 것이다.

참 고 문 헌

[1] Svendsen, T. & F. K. Siong. 1987. "On the automatic segmentation of speech signal." *Proc. of IEEE ICASSP 87*, 77-80.

[2] Brugnara, F. et al. 1993. "Automatic segmentation and labeling of speech based on hidden Markov model." *Speech Communication*, Vol. 12, 357-370.

[3] Rabiner, L. R. & M. R. Sambur. 1995. "An algorithm for determining the endpoints of isolated utterances." *The Bell System Technical Journal*, Vol. 54, No. 2, 297-315.

[4] Sarikaya, R. & H. L. John. 1998. "Robust speech activity detection in the presence of noise." *Proc. of ICSLP 98*, Vol. 4, 1455-1458.

[5] 윤석현, 유창동. 2001. "시간-주파수 영역에서 음성/잡음 우세 결정에 의한 새로운 잡음처리." *한국음향학회지*, 20(3), 48-55.

접수일자: 2002. 10. 29.
게재결정: 2002. 12. 3.

▲ 김무중
서울 관악구 봉천4동 882-5 관악전화국 2층 (우: 151-716)
(주) 언어과학 음성공학연구소
Tel: +82-2-887-8125 (ext:310) Fax: +82-2-887-8127
E-mail: donaldos1024@hotmail.com

▲ 권철홍
대전광역시 동구 용운동 96-3 (우: 300-716)
대전대학교 컴퓨터정보통신공학부
Tel: +82-42-280-2555 Fax: +82-42-284-0109
E-mail: chkwon@dju.ac.kr