문맥종속 반음소단위에 의한 음운 자동 레이블링 시스템의 성능 개선*

박 순철, 김 봉완, 이 용주(원광대)

<차 례>

<Abstract>

Improvement of automatic phoneme labeling system using context-dependent demiphone unit

Soon-Cheol Park, Bong-Wan Kim, Yong-Ju Lee

To improve the performance of automatic labelling system, the context-dependent demiphone unit was proposed. A phone is divided into two parts: a left demiphone that accounts for the left side coarticulation and a right demiphone that copes with the right side context. Demiphone unit provides a better training of the transition between phones.

In this paper, If the length of the phone is less than 120 msec, it is split into two demiphones. If the length of the phone is greater than 120 msec, it is divided into three parts.

In order to evaluate the performance of the system, we use 452 phonetically balanced words(PBW) database for training and testing phoneme models.

According to the experiment, the system using proposed demiphone unit compared with that using old demiphone unit gains 3.83% improved result(71.63%) within 10ms of the true boundary, and 2.20% improved result(86.41%) within 20ms of the true boundary.

* 본 연구의 일부는 1998년도 원광대학교 교비연구의 지원으로 이루어진 것임.

1. 서 론

음성인식 기술은 음성합성 기술과 함께 인간의 가장 편리한 의사전달 수단인 음성을 통해 인간이 컴퓨터와 대화할 수 있도록 해주는 도구로서 정보화의 진전과 더불어 그 필요성이 더욱 증대하고 있다. 미국, 일본, 유럽 등 선진각국에서는 1970년대 이전부터 음성인식에 대한 연구를 추진해 왔으며, 특히 국가 주도 형태의 대규모 프로젝트를 통해 많은 기술적 진보를 가져왔다. 그 결과 한정 어휘의 인식 및 합성 시스템들은 실용화 단계에 접어들고 있다. 그러나, 인간이 자연스럽게 발음한 연속음성의 인식기술은 아직까지 한정 어휘의 인식시스템에 비해 성능면에서 크게 뒤떨어져 있다[15].

임의의 어휘를 대상으로 하는 연속음성인식은 발성자에 따른 개인차는 물론이고 전후에 발성되는 음소의 영향에 의한 조음 결합에 따라 그 특성이 크게 변화한다. 이러한 개인차 및 조음결합의 현상을 분석하기 위해서는 많은 사람이 발성한 다양한 음성데이터를 수집한 후, 음성 데이터를 음소와 같은 음성의 기본단위로 분할하고 레이블링하여 그 다음 단계에서 통계적 처리가 가능하도록 가공하는 작업이 필수적으로 요청된다[11,21]. 실제로 미국 등 선진국에서는 TIMIT 데이터베이스와 같이 잘 가공된 음성 데이터를 구성하고 보급함으로써, 많은 연구 그룹들이 양질의 동일한 음성 데이터를 토대로 한 경쟁적인 연구를 수행함으로써 음성인식기술에 많은 발전을 가져왔다. 따라서 한국어 음성처리기술의 발전을 위해서도 대용량의 음소분할 및 레이블링된 음성 데이터베이스의 구축은 필수적인 과제이다. 또한 대용량 음성 데이터베이스 구축작업의 원활한 진행을 위해서는 이미 개발된 자동 음성분할 및 레이블링 시스템의 성능 향상이 중요한 역할을 할 것이다.

일반적으로 음성 데이터베이스를 구축하는 과정은 다음과 같다. 먼저 음성 데이터베이스의 목적에 따라 발성내용 및 발성방법, 데이터의 수집환경 등을 정의하고, 다수의 발성자를 참여시켜 실제 음성을 수집한다. 수집된 음성 데이터는 단어 또는 문장 등, 일정한 단위로 분할한 후, 레이블링 기준에 따라 레이블링하게 된다. 레이블링된 음성데이터는 수정작업을 거친 후 음성전문가 그룹이나 관련연구자들의 객관적 평가 및 검증을 거쳐 배포나 이용이 용이하도록 데이터베이스화한다[5].

음성 데이터베이스를 구축하는 과정 중, 음성분할 및 레이블링 작업은 일반적으로 사람의 수작업에 의해서 수행된다. 그러나 수작업에 의한 음소분할 및 레이블링 작업은 다음과 같은 문제점을 지닌다[2,4,5]. 첫째, 스펙트로그램의 판독 및 반복적인 청취평가를 통해서 이루어지므로 매우 지루한 작업일 뿐 아니라, 많은 시간이 소요된다[4]. 또한 반복되는 작업으로 인한 판단오류가 발생할 수 있다. 둘째, 수작업에 의한 음소 분할은 높은 수준의 음성학적 지식을 요하기 때문에 소

수의 음성학 전문가에게 의존할 수밖에 없다. 셋째, 음소경계 선정을 위한 구체적인 판단 기준을 미리 정해놓더라도 상당부분 주관적인 판단을 피할 수 없으며, 이에 따라 음성경계 선정과정에서의 일관성을 보장할 수 없다. 따라서 서로 다른 음성학 전문가들이 동일한 음성 데이터를 분할할 경우는 물론이고, 동일한 사람이라 할지라도 동일한 음성 데이터를 분할하는 데 있어 추출된 음성경계에는 상당한 차이가 날 수 있다.

이와 같은 문제점을 해결하기 위하여 음운 자동 레이블링 시스템에 관한 연구가 진행되고 있으며[1,7,8,10,12,14,16,22,23], 현재 특정언어에 국한된 것이기는 하지만 상품화된 자동 음소분할 및 레이블링 시스템이 나오기에 이르고 있다[22].

자동 음성분할 및 레이블링 시스템에서 인식의 단위로 monophone, biphone, triphone, diphone, demiphoe(반음소)등 여러 단위가 사용될 수 있다.

monophone의 경우 훈련할 모델의 수가 적기 때문에 훈련하기 쉽고 모델 당 훈련량이 많은 반면, 전·후 음소에 의한 조음효과를 표현하지 못하는 단점이 있다. 이러한 단점을 극복하기 위해 triphone과 같이 전·후 음소에 의한 조음효과를 반영하는 문맥종속 단위가 제안된 이후[24], 문맥종속 단위에 대한 연구가 활발히 진행되었다[3,6,17]. 그러나, triphone의 경우 인접한 음소들에 의하여 음향학적 특성이 변화되는 각 음소들마다 별도의 모델을 사용하기 때문에 모델의 수가 크게 증가하게 되어 학습데이터 양이 부족하게 되는 단점이 있다[18,19]. 따라서, 이러한 tirphone이 가지는 단점을 보완하면서 문맥조음 효과를 잘 반영할 수 있는 인식단위로 반음소가 제안된 바 있다[3,17].

본 논문에서는 기존에 제안된 반음소의 모델을 개선하여 음운 자동 레이블링 시스템을 구축하고, 성능을 평가하였다.

2장에서는 음운 자동 레이블링 시스템의 개요에 대해서 설명하고, 3장에서는 인식 단위로 사용한 반음소의 정의와 레이블링 시스템의 성능을 향상시키기 위한 반음소 단위의 모델개선 방법에 대해서 서술하고, 4장에서는 본 논문에서 구현한 시스템을 이용하여 음소분할을 수행한 실험의 결과를 보이고, 마지막으로 5장에서 결론 및 향후 연구방향을 제시한다.

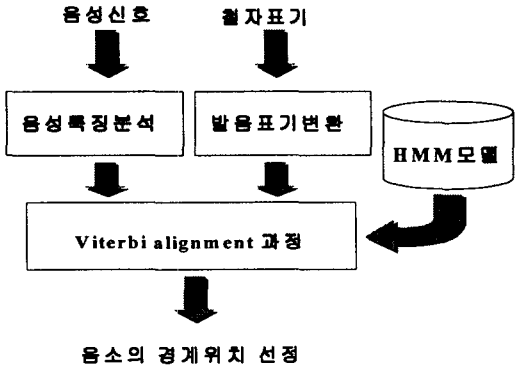
2. 음운 자동 레이블링 시스템

본 장에서는 음운 자동 레이블링 시스템 구축 시 고려사항들을 바탕으로 본 논문에서 구현한 음운 자동 레이블링 시스템에 대해서 기술한다.

음운 자동 레이블링은 입력음성에 포함된 음소열에 대한 대부분의 정보를 미리 알고 있는 경우이므로, 음소 분할과 레이블링이 단일과정으로 수행된다. 훈련과정을 통해 미리 각각의 음소들에 대한 대표 패턴들 또는 통계적 모델들을 연결

시켜서 이들과 입력음성을 매칭시키는 과정에서 음소경계가 자동적으로 추출된다. 이러한 패턴 또는 모델 매칭 방법으로는 DTW(Dynamic Time Warping)방법과 HMM(Hidden Markov Model)방법 등이 사용될 수 있다. 그러나, DTW방법에서는 복수 개의 대표패턴들 만으로 음성신호에 내재된 변화요인들에 대처하는 데 한계가 있기 때문에, 이러한 변화요인들을 확률모델 형태로 다루는 HMM방법이 널리 사용되고 있다.

<그림 1>은 본 논문에서 사용한 HMM방식에 의한 음운 자동 레이블링 시스템의 일반적인 구성도이다. 음성신호가 들어오면 음성 특징 분석과정에서 음성 특징 계수들을 추출하고 미리 구성된 시퀀스와 비터비 검색 알고리즘에 의해 최적의 시간 경계를 설정하는 과정에서 각 음소의 경계위치가 얻어진다.

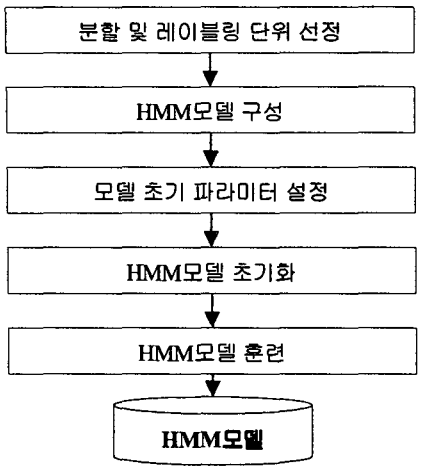


<그림 1> HMM을 이용한 음운 자동 레이블링 시스템의 구성도

<그림 2>는 HMM모델의 일반적인 생성절차이다. HMM모델을 생성하기 위해서는 먼저 음성을 분할하기 위한 단위와 레이블링 단위를 결정하여야 한다. 그리고, HMM모델의 형태와 HMM에서 사용할 특징 파라미터를 결정한 후, HMM모델을 초기화하게 된다. 초기화된 HMM모델은 반복적인 훈련과정을 거쳐 자동 음소 분할 및 레이블링 시스템의 입력으로 사용될 최종적인 HMM모델로 생성된다.

2.1. 음성신호 전처리

음성신호의 전처리 과정에서는 음운 자동 레이블링 시스템의 입력으로 들어가는 음성분석의 시간단위와 특징파라미터를 결정해야 한다.



<그림 2> HMM모델 생성 절차

2.1.1. 음성분석의 시간단위

일반적으로 음성인식의 경우 매 10ms마다 20ms구간의 음성신호를 분석하여 특징파라미터를 추출하는 방식이 널리 사용되고 있다. 그러나, 음운 자동 레이블링 시스템에서는 정교한 음성분할을 위해서 음소경계 검출의 정밀도가 10ms수준인 것은 바람직하지 않으며 정밀도가 5ms를 넘지 않아야 좋을 것으로 판단된다. 참고로 TIMIT음성 데이터베이스의 경우에는 2.5ms 시간단위를 사용한 것으로 알려져 있다[22]. 이러한 관점에서 본 논문에서 구현한 시스템에서는 10ms단위의 해밍윈도우를 5ms간격으로 이동하면서 분석하였다.

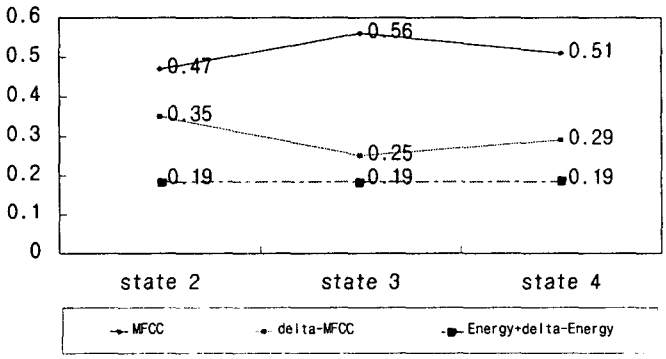
2.1.2. 특징파라미터

음성분할 시스템에서 음성 신호의 특징을 표시하는 특징파라미터의 선정은 매우 중요하며, 시스템의 성능을 좌우하는 요소가 된다. 따라서, 음소간의 변별력이 뛰어나면서 음성학적으로는 중요하지 않은 변화요인들에 대해서 둔감한 특징을 가지는 파라미터를 선정해야 한다.

지금까지 음성인식에서 효과적으로 사용되어온 음성특징분석 파라미터들로는 성도의 공진 특성을 표현하는 LPCC(Linear Predictive Cepstral Coefficient)와 청각기관의 주파수 선택특성을 표현하는 MFCC(Mel-Frequency Cepstral Coefficient)를 들 수 있다. 그리고, 이들의 시간에 따른 변화요인을 나타내는 시간축 미분값이 있다.

이외에도 음성의 단구간 에너지 및 그 미분값도 중요한 정보로 활용될 수 있다. 이러한 특징파라미터들은 일정 차원 수를 가지는 벡터의 형태로 표현된다[20].

그리고, 각 특징벡터들을 병합하여 사용하는 것보다는 독립적으로 사용하고, 신호성분을 표현하는 정도에 따라 가중치를 부여하는 방법들이 보다 좋은 성능을 나타내는 것으로 알려져 있다. 본 논문에서는 12차의 MFCC와 그 시간축 미분값 그리고, 정규화된 단구간 에너지와 그 미분값을 사용하였다. 본 논문에서 구현한 시스템에 사용된 특징 파라미터 상태별 가중비율은 <그림 3>과 같다[9].



<그림 3> 특징 파라미터 상태별 가중 비율

2.2. 인식을 위한 단위

자동 음소분할 및 레이블링 시스템에서 사용하는 인식의 기본단위로는 음소, 유사음소(phoneme-like unit), 변이음(allophone)등이 사용될 수 있다.

본 논문에서 구현한 레이블링 시스템에서 인식의 기본단위는PBW(Phonetically Balanced Words) 452 단어에 사용된 유사음소를 선정하여, triphone과 demiphone으로 확장하여 사용하였다.

2.3. HMM모델의 구성

HMM에 의해 음소모델을 구성하기 위해서는 HMM에서의 관찰확률분포, 상태수와 천이방식 등의 HMM형태(topology)를 구성해야 한다.

2.3.1. 관찰확률분포

HMM에서 관찰확률분포는 연속확률분포, 이산확률분포, 준연속확률분포 등이 사용된다. 이산확률분포를 사용하는 이산HMM은 피쳐(feature)공간을 일정한 개수로 나눈 대표 값으로 영역을 표현하는 벡터 양자화 기법을 이용하여 관측 심벌을 표현하게 되며, 입력은 이러한 관찰 심볼 중의 하나로 변환된다. 그러나, 대표 값을 사용하기 때문에 입력 표현 시 벡터 양자화 오류가 발생되며, 경우에 따라서 이러한 오류는 치명적인 결과를 가져오기도 한다.

이러한 표현상의 오류를 극복하기 위하여 연속적인 피쳐 공간의 실수값을 그대로 입력값으로 사용하는 연속확률분포를 사용하는 연속HMM이 등장하였다. 연속HMM은 음성 신호의 특징 추출 단계를 거쳐 생성된 특징 파라미터를 관측열(observation sequence)로서 직접 모델링하므로 양자화 과정에서 오차가 발생하는 문제점을 막을 수 있으나, 피쳐 공간의 정확한 근사가 어렵고, 계산이 복잡하며, 시간이 오래 걸리는 단점이 있다.

본 논문에서 구현한 시스템에서는 연속확률분포를 사용하는 연속HMM을 사용하고 있다.

2.3.2. HMM형태(topology)

<그림 4>는 서로 구별되는 HMM의 3가지 형태를 보이고 있다. 음성인식 문제나 온라인 필기인식과 같은 문제에서는 입력의 시간적인 순서가 중요한 정보로 이용된다. 따라서, 시간적 제약을 구조 자체 내에 담고 있는 left-right형태의 HMM을 많이 사용한다. 이러한 HMM은 맨 왼쪽의 상태가 초기상태(inital state), 맨 오른쪽의 상태가 마침상태(final state)가 되며, 상태 전이는 항상 자기 자신이나 우측 방향, 즉 자신의 오른쪽에 있는 상태로 제한된다. 이러한 구조에서는 시간이 경과함에 따라 상태열이 오른쪽으로 이동하므로, 음성인식에서의 시간의 경과에 따른 입력을 모델링할 수 있다.

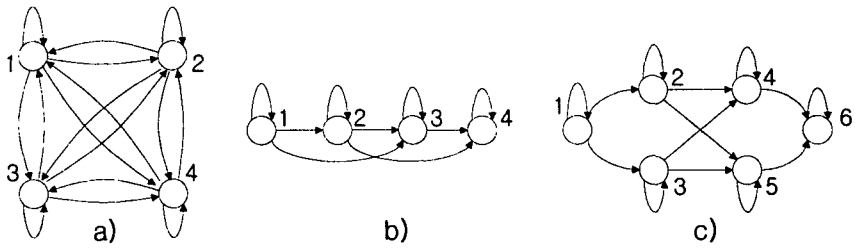
본 논문에서 구현한 시스템에서는 5상태 7천이를 가지는 left-right모델을 사용하였으며, <그림 5>에서 그 구조를 보이고 있다.

2.4. HMM모델의 훈련

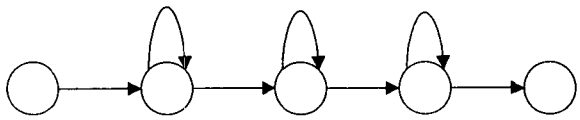
2.4.1. 음성 데이터베이스

본 논문에서 구현한 시스템의 훈련은 PBW(Phonetically Balanced Word) 452 단어로 구성된 데이터베이스의 남성화자 30명분의 데이터를 사용하였다. PBW 음성

데이터베이스는 방음부스에서 Senheizer HMD224X를 사용하여 녹음되었으며, DAT(Digital Audio Tape)에 저장된 데이터를 KAY CSL 4300B를 사용하여 AD/DA 변환하였다. 16kHz로 샘플링(Sampling)되어 있으며, 16Bits로 양자화 되어있다[8].



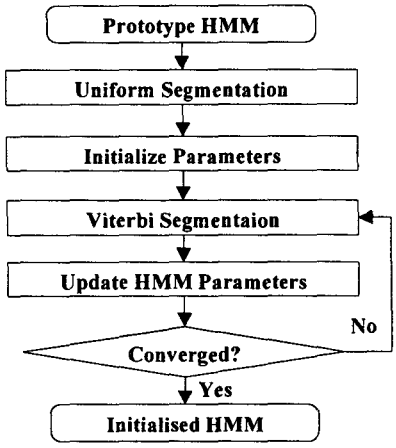
<그림 4> 서로 구별되는 HMM의 3가지 형태(topology)
a) 4-상태(state) ergodic model
b) 4-상태(state) left-right model
c) 6-상태(state) 병렬경로(parallel path) left-right model



<그림 5> 5상태 7천이를 가지는 left-right 모델

2.4.2. HMM모델의 초기화

HMM모델의 초기화는 비터비 분할(Viterbi Segmentation)에 의해서 이루어진다. 비터비 알고리즘은 '최적원리(Principle of Optimality)'에 입각한 동적 프로그래밍(dynamic Programming)으로, 임의상태에 이르는 경로의 비용 또는 확률을 선행상태의 비용과 그 곳에서 현재 상태로의 전이 비용을 합하는 방식으로 순환적으로 계산하는 방식이다. <그림 6>은 비터비 알고리즘을 이용하여 HMM모델을 초기화하는 과정을 보이고 있다.



<그림 6> HMM모델 초기화 과정

2.4.3. HMM모델 훈련

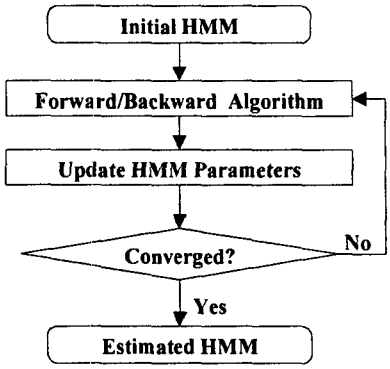
<그림 7>은 Baum-Welch 알고리즘을 이용하여 초기화된 HMM모델을 훈련하는 과정을 보이고 있다. Baum-Welch 알고리즘은 모델 파라미터 재추정(model parameter re-estimation)문제에 대한 해결방법으로 많이 사용되고 있으며, 최대우도 측정(Maximum Likelihood Estimation, MLE)방식이다. 즉, 현재의 모델 파라미터를 이용하여 훈련 데이터에 대한 모델 정보를 계산하고, 이를 이용하여 훈련 데이터의 정보가 반영되도록 파라미터를 수정하는 과정을 반복적으로 적용하여 모델과 훈련 데이터 사이에 유사도(Likelihood)값을 높이는 방식이다. 위와 같은 훈련과정을 모든 훈련 데이터에 대하여 실행하면, 최종적으로 원하는 HMM모델을 얻을 수 있다.

3. 반음소(Demiphone)의 정의 및 모델 개선

본 장에서는 본 논문에서 구현한 반음소 단위의 자동 레이블링 시스템에서 인식의 단위로 사용된 반음소와, 반음소의 모델 개선방법에 대해서 서술한다.

3.1. 반음소(Demiphone)의 정의

일반적으로 음소는 정상시점을 기준으로 선행음소의 영향을 받는 전반부와, 후



<그림 7> HMM모델 훈련 과정

속음소의 영향을 받는 후반부로 분류할 수 있다. 이와 같이, 음소를 두 부분으로 나눌 수 있는 이유는, 많은 경우 선행음소와 후속음소가 미치는 영향이 음소의 전반부 및 후반부에 국한된다는 점에서 음소를 서로 성질이 다른 두 부분으로 구분지을 수 있다[6,17].

반음소는 음소 또는 변이음을 전·후 음소의 영향을 받지 않는 정상상태시점에 중점을 경계로 양분함으로써 얻어지는 음성단위이다. 이렇게 나누어진 반음소는 선행음소에 의한 조음효과를 포함하는 전반음소(Left-Demiphone)와 후속음소에 의한 조음효과를 포함하는 후반음소(Right-Demiphone)의 두 부분으로 나눌 수 있다. 즉, 음소의 경계와 diphone의 경계를 동시에 포함하는 단위라고 할 수 있다[6,17].

예컨대 파열음(stop) 특히 외파음(explosive)은 정지갭(stop gap)과 후속되는 정지버스트(stop burst)로 구성되며, 파찰음(affricate)은 정지갭과 후속되는 강한 마찰잡음(strong fricatoin noise)으로 이루어진다. 따라서 이들은 자연스럽게 전반음소와 후반음소로 이분된다. 또, 공명음(sonorant)은 포먼트 구조를 가진 성대진동음(voicing)으로 구성되어 있는 것이 특징이다. 즉 모음(vowel)과 활음(glide)은 3개 이상의 뚜렷한 포먼트를 가진 유성음(voice sound)으로 구성되며, 스펙트럼 영점(spectral zero)을 가지지 않는다. 모음과 활음의 중점 부근에 정상상태(steady state)가 존재하는데 선행음소들과 후속음소들의 영향은 정상상태 시점까지로 국한된다. 기타의 음소의 경우에도 모음이나 활음의 경우와 같지는 않지만 유사한 성질이 있다. 비음(nasal)은 포먼트들과 같이 1kHz부근의 주파수에 스펙트럼 영점을 가지는 유성음인 네이절 머머(nasal murmur)로 구성된다. 유음(liquid) 특히 설측음(lateral)은 네이절 머머와 유사하나, 그보다 에너지가 강하고 포먼트 위치가 다르며 스펙트럼 영점이 훨씬 더 높은 주파수쪽에 존재하는 유성음으로 구성된다. 강마찰음(strong fricative)은 포먼트(formant) 구조를 가지는 강한 마찰잡음(strong

frication noise)만으로, 무성 약마찰음(unvoiced weak fricative)은 넓은 스펙트럼 범위(broad spectral range)에 고르게 분포된 약한 마찰잡음(weak frication noise)으로, 그리고 유성 약마찰음(voiced weak fricative)은 성대진동음(voicing)에 의해 변조된 것처럼 보이는 약한 마찰잡음(weak frication noise)으로 구성된다. 이들 음소의 전반부는 주로 선행음소들의 영향을, 후반부는 주로 후속음소들의 영향을 받는다. 따라서 이들 음소들도 전반음소와 후반음소로 이분된다[20].

3.2. 반음소의 구성

<그림 8>은 나비라는 음절의 음소, 다이폰(diphone), 트라이폰(triphone), 반음소의 경계를 보이고 있다. <그림 8>에서 '#'기호는 묵음(silence)를 의미하고, '~'기호는 다이폰을 의미하며, '-'와 '+'는 선행음소와 후속음소를 의미한다.

#	나	비	#

a) 음절단위

#	n	a	b	i	#

b) 음소단위

#~n	n~a	a~b	b~i	i~#

c) 다이폰(diphone) 단위

#	#-n+a	n-a+b	a-b+i	b-i+#	#

d) 트라이폰(triphone) 단위

#	#-n	n+a	n-a	a+b	a-b	b+i	b-i	i+#	#

e) 반음소(demiphone) 단위

<그림 8> 음절, 음소, 다이폰, 트라이폰, 반음소의 경계
a) 음절, b) 음소, c) 다이폰, d) 트라이폰, e) 반음소

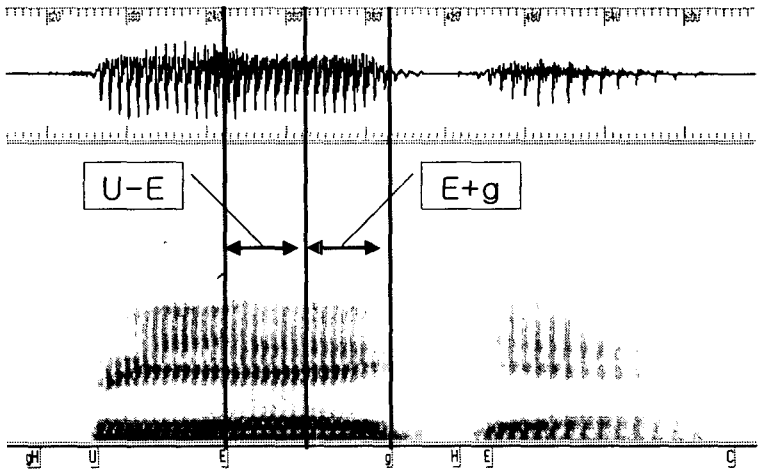
<그림 8>에서 알 수 있듯이 반음소는 음소의 경계와 다이폰의 경계를 모두 포함하고 있는 단위이다. 따라서 반음소의 경계는 음소경계와 다이폰의 경계를 규정함으로써 설명될 수 있다.

음소는 음운론적으로 볼 때, 어휘 의미의 변별을 초래하는 단위이다. 하나의 음소는 음성 환경에 따라 서로 다른 음가의 소리로 실현되는데, 이것을 변이음(allophone)이라 한다. 음소간의 경계는 파형 및 스펙트로그램 상에서 쉽게 구별할

수 있다. 그러나, 모음과 모음, 파열음과 파열음처럼 조음방식이 비슷한 음소들간의 경계는 쉽게 구별이 되지 않는다. 이러한 경우는 에너지 차와 포먼트 천이 정보에 의해 구별짓는다.

다이폰은 <그림 8>에서 보는 것과 같이 한 음소에서 다음 음소로 천이하는 구간을 포함하는 음소단위이다. 그러므로, 다이폰은 선행음소의 일부와 후속음소의 일부를 같이 포함하게 된다.

<그림 9>는 '그에게'라는 단어에서 실제 반음소단위 경계로 분할된 예를 보이고 있다. <그림 9>에서 'E'음소의 정상시점을 중심으로 선행하는 음소인 'U'의 조음효과를 포함하는 전반음소인 'U+E'와, 후속음소인 'g'의 조음효과를 포함하는 후반음소인 'E-g'로 이분된 경계를 볼 수 있다.



<그림 9> 반음소의 경계

3.3. 반음소의 특성

자동레이블링 시스템에서 인식의 단위로 반음소를 사용함으로써 얻을 수 있는 장점은 아래와 같다.

첫째, 확장성이 좋다. 훈련 시 발생하지 않은 모델에 대해 인식 시 새로운 모델을 생성하기가 유리하다. 또한, 전·후반음소를 잘 결합하면 음소단위나, 다이폰 단위, 트라이폰 단위와 같은 효과를 가져올 수 있다.

둘째, 전·후 음소에 의한 조음현상을 잘 반영하는 모델로는 트라이폰 등이 있다. 반음소의 경우에는 훨씬 적은 수의 모델을 사용하면서도 전·후 음소에 대한 문맥조음효과를 잘 반영할 수 있다. 실험에 사용된 PBW(Phonetically Balanced

Words) 452단어를 기준으로 했을 때, 트라이폰의 경우 10823개의 모델이 필요한 반면, 반음소는 1874개의 모델이 필요하다.

셋째, 반음소의 하나의 음소가 전·후반음소로 나누어진다. 이와 같이 하나의 음소에서 두 개의 모델이 생성되기 때문에 훈련 량이 두배가 된다. 즉, 하나의 음소에서 전·후반음소 두 개의 모델을 동시에 훈련시킬 수 있다.

즉, 반음소는 문맥의 조음효과를 잘 반영하면서 훈련효과를 높일 수 있는 장점을 가지고 있는 모델이다.

3.4. 반음소 단위의 모델 개선

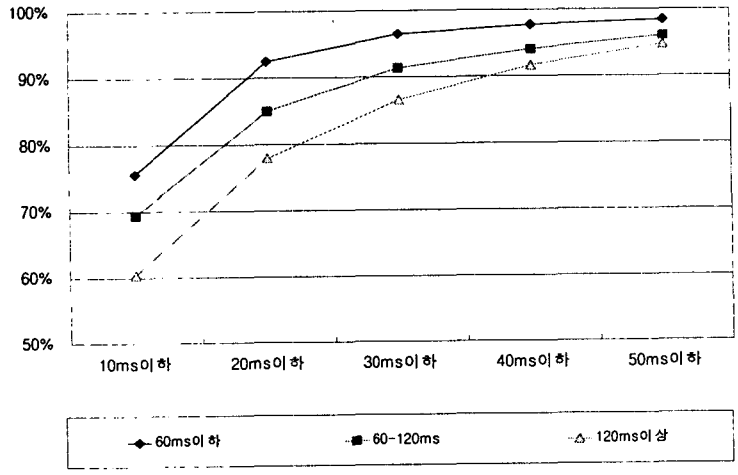
반음소의 구성은 하나의 단위에서 전·후 음소와의 천이구간과 안정화 구간의 두 부분으로 이루어진다. 여기에서 천이구간은 전·후 음소와의 상호 문맥조음효과를 반영하는 구간이다.

그러나 모음과 같은 경우에 음소의 길이가 길어지게 되면, 음소의 전·후 천이구간보다는 음소의 안정화 구간이 상대적으로 길어지게 된다. 결과적으로, 전·후 음소와 상호 문맥조음효과를 포함하는 천이구간보다 음소의 안정화구간이 길어지게 되어 천이구간의 특성을 잘 반영하는 반음소단위의 특성이 약화될 우려가 있다.

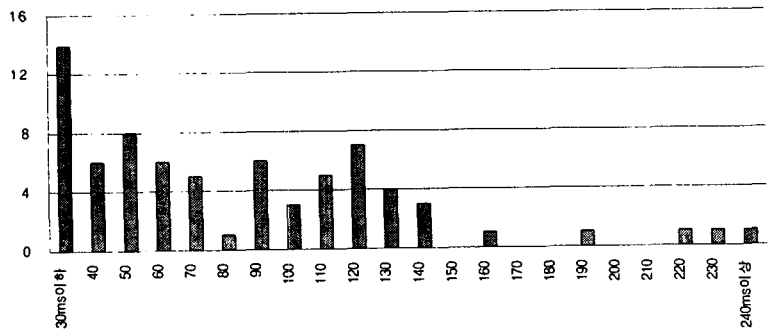
<그림 10>은 본 논문에서 구현한 기존의 반음소 단위를 사용한 자동 레이블링 시스템에서 음소의 지속시간별 인식률을 보인 것이다. <그림 10>에서 보는 바와 같이 지속시간의 길이가 120ms 이상인 음소의 경계인식률이 음소의 지속시간이 짧은 음소에 비해 성능이 저하됨을 알 수 있다. 이와 같은 이유는 여러 가지가 있을 수 있지만, 반음소 단위의 자동 레이블링 시스템에서 천이구간의 특징을 잘 반영하는 반음소 단위 특성이 약화되었기 때문이라고 생각해볼 수 있다.

따라서, 본 논문에서는 이러한 점을 개선하기 위하여 음소의 지속시간이 길어지는 음소에 대해서 안정화 구간을 별도의 단위로 모델링하였다. <그림 11>은 단위 음소별 평균지속 시간을 나타내는 그래프이다. 본 논문에서는 <그림 12>에서 보는 바와 같이 평균 지속시간의 길이가 120ms이상 되는 음소에 대해 별도의 단위로 모델링하였다. 즉, 하나의 음소에서 선행·후속 음소와의 천이구간을 전·후 각 60ms로 보고 지속시간이 120ms 이상인 모음에 대해서 <그림 12>와 같이 전·후 각 60ms를 전·후반 음소로 구성하고 나머지 가운데 부분을 음소의 안정구간으로 구성하였다.

<그림 12>는 '거액이'라는 단어의 예를 보이고 있다. <그림 12>의 'v'와 같이 음소의 지속시간이 120ms 이상일 때, 'v'음소를 선행음소와의 경계에서부터 60ms, 후속음소와의 경계구간으로부터 60ms, 그리고 나머지 구간을 안정화 구간으로 3등분한다. 최종적으로 'v'음소는 선행음소와의 경계구간인 'gH-v', 'v', 'v+E'와 같



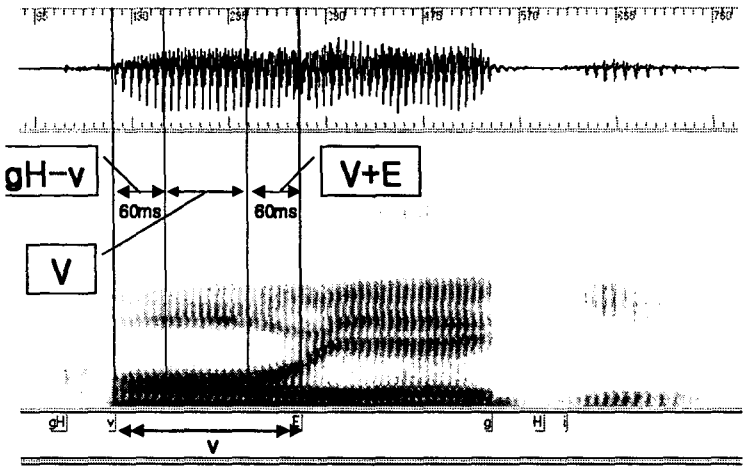
<그림 10> 음소의 지속시간별 인식률 비교



<그림 11> 음소별 평균 지속시간

이 3개의 반음소 단위로 모델링된다.

본 논문에서 제안한 반음소 단위의 성능을 비교 평가하기 위하여, 기존의 반음소 단위와 새로 제안한 반음소 단위를 사용하여 레이블링 시스템을 구축한 후 성능을 비교 평가하였다.



<그림 12> 제안한 반음소 단위의 경계 (단어 '거액이'의 예)

4. 자동 레이블링시스템의 성능평가

본 논문에서 구현된 문맥종속 음운 자동 레이블링 시스템의 성능평가를 위해 PBW(Phonetically Balanced Words) 452 단어로 구성된 남성화자 데이터 중 훈련에 사용하지 않은 4명분의 데이터를 평가에 사용하였다.

레이블링 시스템의 훈련은 음소단위로 구성된 PBW 452 단어의 남성화자 30명분의 데이터를 기존의 반음소단위와 개선된 반음소 단위로 확장하여 사용하였다.

4.1. 음성 분할을 위한 단위

본 논문에서 구현된 레이블링 시스템의 인식 단위를 선정하기 위하여 수작업으로 레이블링된 PBW 452 단어의 음성 데이터베이스를 분석하여 총 93개의 단위로 분류했다. 분석결과를 바탕으로 거의 출현하지 않은 음소들은 레이블링 시스템의 인식의 단위에서 제외하였다.

레이블링 단위로 선정된 유사음소의 목록 중 유기 파열음(p, t, k), 무기 파열음(B, D, G), 유기 파찰음(c), 무기 파찰음(Z)의 폐쇄/파열 구간이 유성음화된 부분, 치조 마찰음(s)이 유성음화된 음소, 불파음화, 모음 ' ㅖ'와 ' ㅒ', 'ㅚ'와 'ㅐ', 'ㅞ'를 하나의 음소로 하여, 총 73개의 음소단위를 분할을 위한 단위로 선정하였다. 최종적으로 선정된 레이블링 시스템에서 인식을 위하여 사용될 기본단위들은 <표 1-4>

에 나타낸 것처럼 72개의 유사음소와 1개의 묵음을 포함하여 총 73개의 단위이다.

일반적으로 반음소는 앞장에서 정의한 바와 같이 음소의 정상상태를 기점으로 하여 나눈다. 그러나, 음소의 중점을 기점으로 나누어도 음소의 정상상태를 기점으로 하는 경우와 통계적으로 거의 일치한다. 따라서, 본 논문에서 사용한 반음소 단위는 선정된 유사음소의 중점을 경계로 하여 전반음소와 후반음소의 두 부분으로 나누었으며, 음소의 길이가 120ms이상인 경우에는 앞장에서 설명한 방법으로 전·후 경계 부분을 기준으로 60ms를 전·후반음소로 나머지를 음소의 안정화 구간으로 하는 3등분으로 분할하였다.

<표 1> PBW 452 단어 음성 DB에서 선정된 레이블링 기호 목록(모음/묵음)

모음	음소	ㅏ	ㅓ	ㅗ	ㅜ	ㅡ	ㅣ	ㅔ, ㅐ	ㅢ	
	기호	a	v	o	u	U	i	e	Ui	
이중 모음	음소	ㅑ	ㅕ	ㅛ	ㅠ	ㅖ, ㅒ	ㅝ	ㅟ	ㅚ,ㅞ,ㅙ	ㅘ
	기호	ja	jv	jo	ju	je	wv	wi	we	wa
묵음	C									

<표 2> PBW 452 단어 음성 DB에서 선정된 레이블링 기호 목록(파열음/파찰음)

	음소	폐쇄구간	폐쇄구간의 유성음화	파열/마찰구간	파열/마찰구간의 유성음화	폐쇄구간의 공명음화	파열/마찰구간의 공명음화
파열음	ㄱ	g	gV	gH	gHV	gR	gHR
	ㄲ	G		GH			
	ㅋ	k		kH			
	ㄷ	d	dV	dH	dHV	dR	dHR
	ㄸ	D		DH			
	ㅌ	t		tH			
	ㅂ	b	bV	bH	bHV	bR	bHR
	ㅃ	B		BH			
	ㅍ	p		pH			
파찰음	ㅈ	z	zV	zH	zHV	zR	zHR
	ㅉ	Z		ZH			
	ㅊ	c		cH			

<표 3> PBW 452 단어 음성 DB에서 선정된 레이블링 기호 목록(마찰음/비음)

	음소	마찰성분	유성음화	공명음화		음소	기호
마찰음	ㅅ	s			비음	ㅁ	m
	ㅆ	S				ㄴ	n
	ㅎ	h	hV	hR		ㅇ	N

<표 4> PBW 452 단어 음성 DB에서 선정된 레이블링 기호 목록(유음)

	음소	폐쇄구간	폐쇄구간의 유성음화	기식음	기식음의 유성음화	폐쇄구간의 공명음화	기식음의 공명음화	설측음
유음	ㄹ	r	rV	rH	rHV	rR	rHR	l

PBW 452 단어에서 선정한 73개의 유사음소를 바탕으로 한 문맥종속 반음소 단위들 중에서 출현빈도수가 낮은 반음소 단위들은 훈련에 참여하는 빈도가 낮아 훈련이 어렵다. 훈련에 참여하는 반음소의 단위의 출현빈도수가 50회 미만인 경우 훈련하기에 충분한 양이 아니기 때문에, 훈련량이 충분하지 않아 전체적으로 시스템의 성능을 떨어뜨리게 된다. 따라서, <표 5>에서처럼 음소를 17개의 전·후 문맥 정보로 분류하여 선행·후속음소 대신에 문맥정보를 사용하였다.

<표 5> 문맥정보를 위한 음소 분류

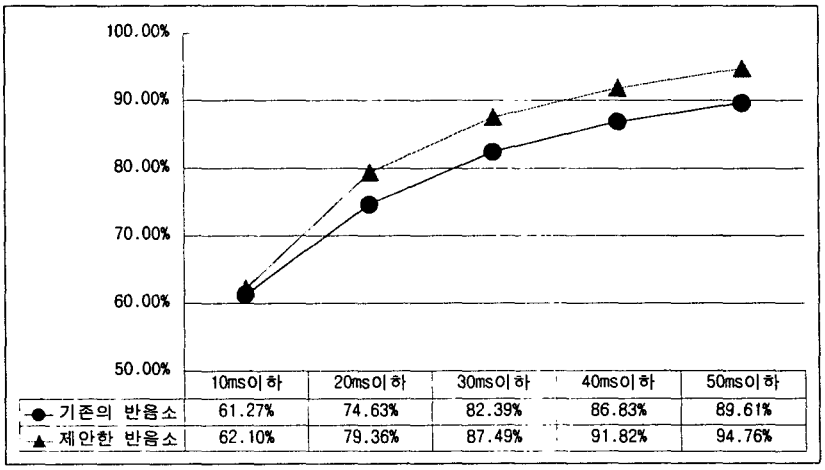
음소 분류	기호	음소 분류	기호
파열음의 폐쇄구간	sS	파찰음의 폐쇄구간	sA
파열음 폐쇄구간의 유성음화	sVS	파찰음의 폐쇄구간의 유성음화	sVA
파열음의 파열구간	bS	파찰음의 마찰구간	bA
파열음의 파열구간의 유성음화	bVS	파찰음의 마찰구간의 유성음화	bVA
파열음의 공명음화	sRS	파찰음의 공명음화	sRA
마찰음	F	비음	N
성문 마찰음	hF	모음	V
유음	L	이중모음	yV
유음의 공명음화	RL	묵음	sil
설측음	lL		

4.2. 실험결과

본 논문에서는 음소의 지속시간길이가 긴 경우에 대해, 120ms를 기준으로 하여 120ms 이상인 음소에 대해 3등분으로 분할하였다. <그림 13>은 이러한 경우 즉, 음소의 지속시간이 120ms이상 되는 음소의 경계 분할 적중률을 비교한 것이다. 지속시간이 120ms 이상인 총 10730개의 음소에 대해 개선된 반음소단위가 전체적으로 성능이 우수함을 볼 수 있다. 이러한 결과로 볼 때 지속시간이 긴 음소의 경우에 기존의 반음소 단위보다 본 논문에서 개선한 반음소 단위가 효과적으로 음소

의 천이구간이 모델링 되었다는 것을 알 수 있다.

<그림 14>는 60ms이하, 60-120ms사이, 120ms이상 되는 단위 음소별로 기존의 반음소 단위와 제안한 반음소 단위의 경계 분할 적중률을 비교한 것이다. 본 논문에서 제안한 반음소 단위의 성능이 높게 나타남을 볼 수 있는데, 이러한 결과 역시 제안된 모델이 기존의 반음소 단위보다 천이구간의 특성이 잘 반영되었다는 것을 알 수 있다.

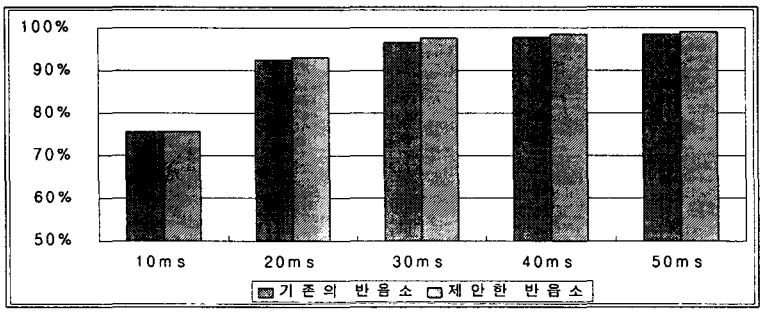


	10ms이하	20ms이하	30ms이하	40ms이하	50ms이하
기존의 반음소	61.27%	74.63%	82.39%	86.83%	89.61%
제안한 반음소	62.10%	79.36%	87.49%	91.82%	94.76%

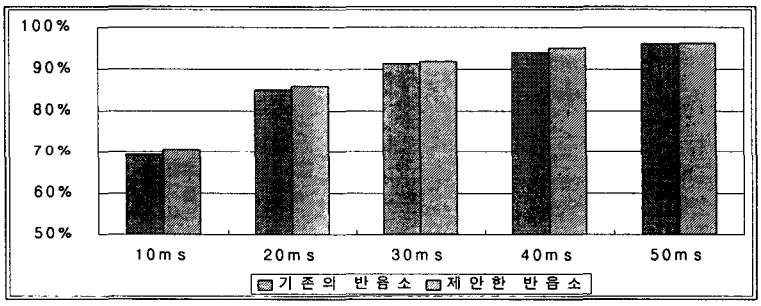
<그림 13> 지속시간 120ms이상 음소의 경계 분할 적중률 비교

<그림 15>는 음소군별 경계 분할 적중률을 비교한 것이다. 기존의 반음소 단위와 비교해 볼 때 본 논문에서 제안한 반음소 단위가 파찰/마찰 구간의 유성음화 부분, 모음과 같이 음소의 안정화 구간이 비교적 길게 나타나는 음소들의 경계 분할 적중률이 향상된 것을 볼 수 있다.

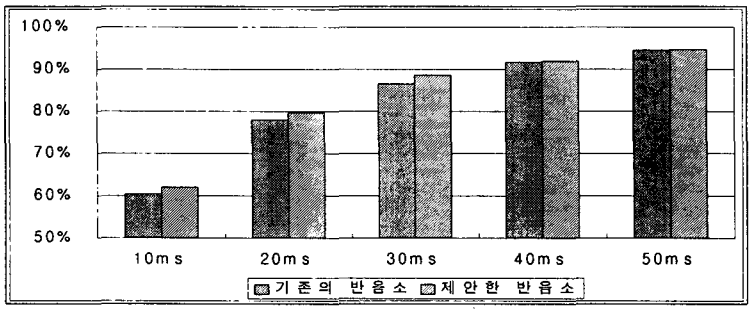
<그림 16>은 기존의 반음소 단위와 본 논문에서 제안한 반음소 단위의 음소군별 천이구간에서의 인식률을 나타낸 것이다. <그림 16>을 살펴보면 경계오차는 대부분 모음과 모음, 묵음과 모음, 모음과 이중모음 등 모음과의 경계부분에서 많이 발생하고 있다. 그러나, 기존의 반음소 단위에 비해 본 논문에서 제안한 반음소 단위가 모음과의 경계구간에서 경계 분할 적중률이 향상된 것을 볼 수 있다.



a) 지속시간 60ms이하 단위음소의 경계 분할 적중률 비교

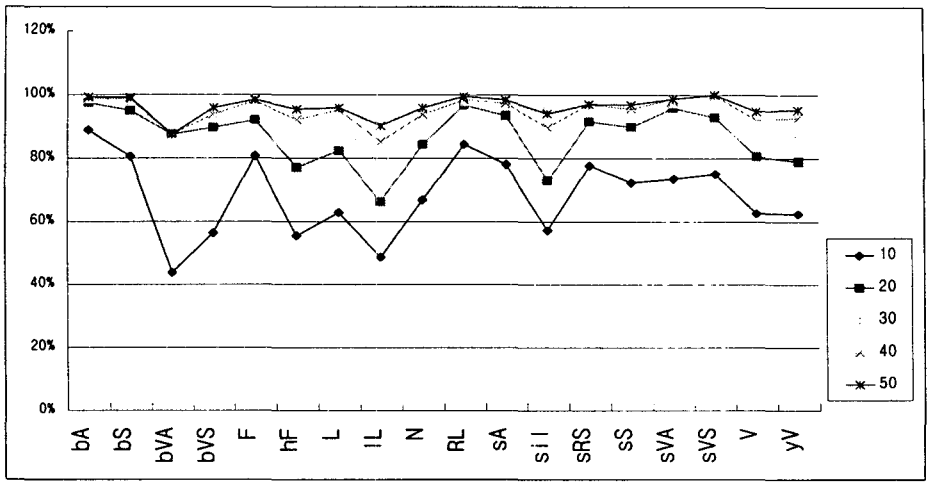


b) 지속시간 60-120ms 단위음소의 경계 분할 적중률 비교

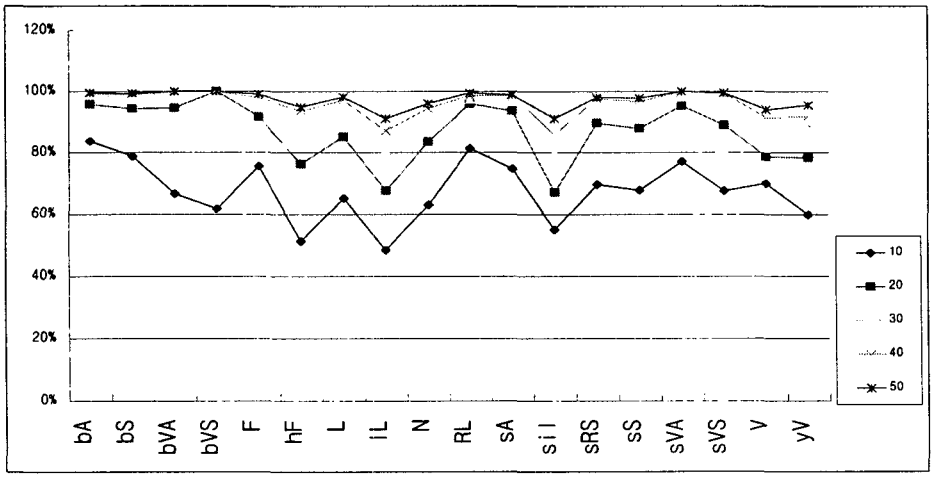


c) 지속시간 120ms이상 단위음소의 경계 분할 적중률 비교

<그림 14> 지속시간 기준 단위음소별 경계분할 적중률 비교

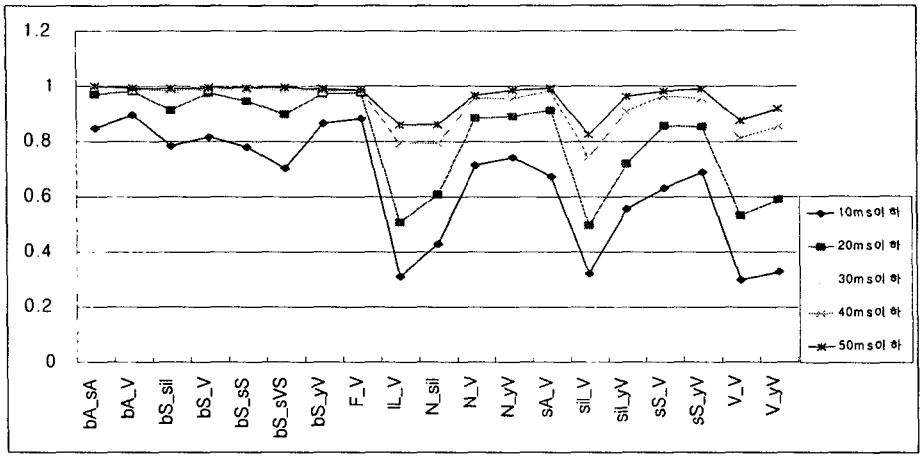


a) 반음소 단위의 음소군별 경계 분할 적중률

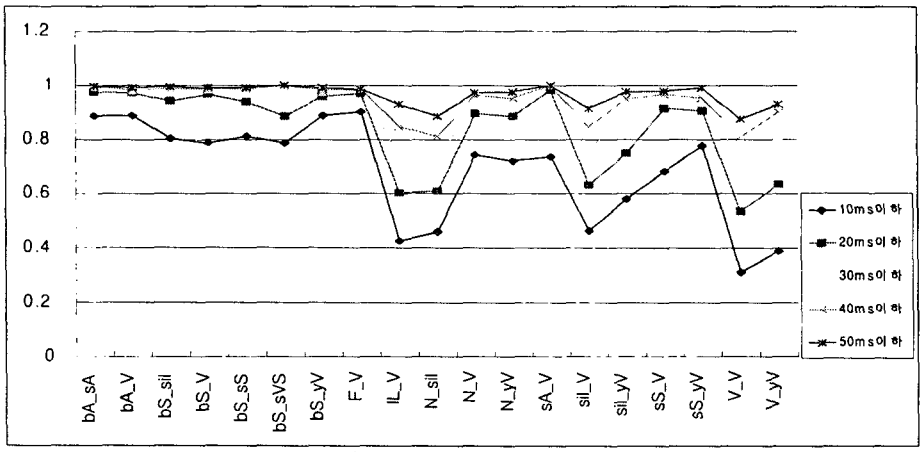


b) 제안한 반음소 단위 음소군별 경계 분할 적중률

<그림 15> 음소군별 경계 분할 적중률 비교



a) 반음소 단위 음소군별 천이에서의 경계 분할 적중률

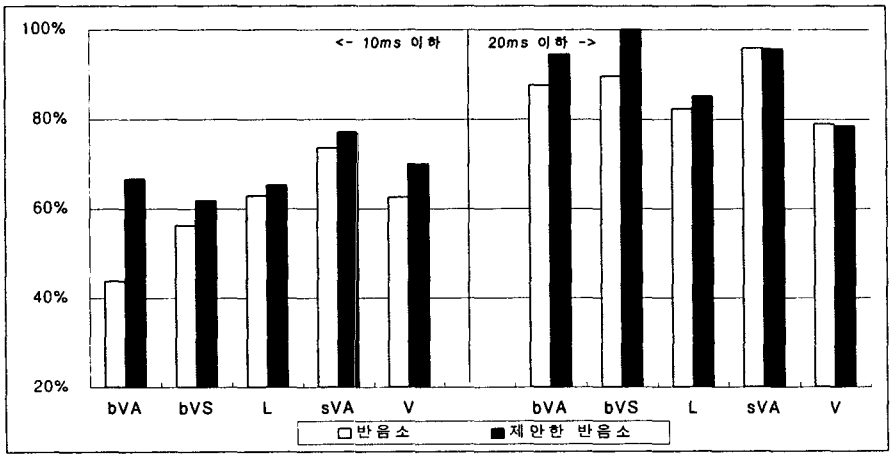


b) 제안한 반음소 단위 음소군별 천이에서의 경계 분할 적중률

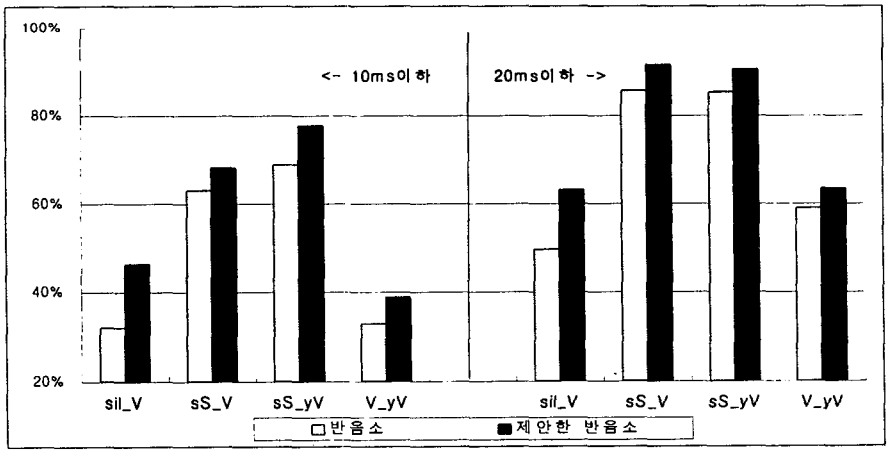
<그림 16> 음소군별 천이에서의 경계 분할 적중률

<그림 17>은 <그림 15>와 <그림 16>에서 성능향상이 많은 음소그룹을 뽑은 것이다. <그림 17>을 보면 음소의 평균지속시간이 비교적 긴 모음과 모음과의 경계에서 성능이 향상되었음을 알 수 있다. 이와 같은 결과로 제안한 반음소 단위가

천이구간의 특성을 기존의 단위보다 효율적으로 모델링하였다는 것을 알 수 있다.



a) 음소별 경계오차 비교



b) 음소 그룹별 천이구간에서의 경계오차 비교

<그림 17> 기존의 반음소 단위와 제안한 반음소 단위의 경계오차 비교

<표 6>은 기존의 반음소 단위와 본 논문에서 제안한 반음소 단위의 전체 경계 분할 적중률을 비교한 것이다. 본 논문에서 제안한 반음소 단위를 사용했을 때 기존의 반음소 단위보다 10ms 이하에서 3.83%, 20ms 이하에서 2.20%의 인식률 향상

을 얻을 수 있었다. 이러한 결과로 볼 때 기존의 반음소 단위보다 본 논문에서 제안한 반음소 단위가 효율적이라는 것을 알 수 있다.

<표 6> 기존에 제안한 단위와 개선된 반음소 단위의 인식률 비교

	기존의 반음소단위	제안한 반음소 단위
10ms이하	67.80 %	71.63 %
20ms이하	84.21 %	86.41 %
30ms이하	90.97 %	92.58 %
40ms이하	94.13 %	95.40 %
50ms이하	96.27 %	97.19 %

전체적으로 경계 분할 적중률이 낮은 경우는 다음과 같다.

- ▸ 모음과 성문마찰음, 비음과 성문마찰음의 경우 유성음 사이에 나오는 성문마찰음이 약화되거나 발생하지 않는 경우
- ▸ 성문 마찰음이 약화된 경우
- ▸ 성문마찰음이 발성되지 않은 경우
- ▸ 불파음화

하지만, 이러한 경우들은 출현빈도수가 낮기 때문에 성능에는 많은 영향을 미치지 않는다.

그러나 모음과 모음, 모음과 묵음의 경우는 출현빈도수가 높기 때문에 성능에 많은 영향을 미친다. 이러한 경우 발생된 경계오차는 파열음의 파열성분이나 파찰음의 마찰성분 그리고 치조 마찰음 뒤에 나오는 모음이 무성음화 된 경우 무성음화 된 모음은 다음에 나오는 음소에 영향을 미쳐 전혀 엉뚱한 곳에서 경계를 분할하게 되고, 후속음소까지 영향을 받게된다. 또한 앞에서 언급했듯이 성문 마찰음이 거의 나타나지 않거나 탈락된 경우에도 이러한 오차가 발생한다. 따라서, 성능을 향상시키기 위한 향후 연구 시 이러한 부분의 보완이 있어야 될 것이다.

5. 결론 및 향후 연구방향

본 논문에서는 문맥종속 반음소 단위 자동 레이블링 시스템의 성능을 개선하기 위하여, 음소의 지속시간이 길어지는 경우에 음소를 3등분하는 새로운 반음소 단위를 도입하여 자동 레이블링 시스템을 구현하였다.

레이블링 단위는 72개의 유사음소와 1개의 묵음 등 총 73개의 레이블링 단위를 사용하였으며, 패턴매칭 기법으로는 연속 HMM을 사용하였으며, 음소 모델은 5상태와 7개의 천이를 가지는 left-right모델을 사용하였다. 특징파라미터는 12차의 MFCC, MFCC의 시간축 미분값, 에너지와 에너지의 미분값을 상태별 파라미터별로 가중 적용하여 각각을 독립적인 벡터로 사용하였다.

시스템의 성능평가를 위해 PBW(Phonetically Balanced Words) 452 단어로 구성된 데이터베이스의 남성화자 30명분의 데이터를 훈련에 사용하여 자동 레이블링 시스템을 구축한 후, 훈련에 사용하지 않은 남성화자 4명분의 데이터에 대해 자동 레이블링 시스템의 성능을 평가하였다.

평가결과 기존의 문맥종속 반음소 단위의 자동 레이블링 시스템에 비해 본 논문에서 새롭게 제안한 반음소 단위를 사용한 자동 레이블링 시스템이, 10ms 이하에서 3.83%, 20ms 이하에서 2.20%의 성능향상을 얻었다. 이러한 이유는 본 논문에서 제안한 반음소 단위가 음소의 지속시간이 길어질 경우에 대해 안정화 구간을 별도로 모델링하였기 때문에 천이구간의 특성이 잘 반영되었다는 것을 알수 있다.

본 논문에서는 기존의 문맥종속 반음소 단위의 레이블링 시스템과의 비교를 위해 특징파라미터와 HMM음소모델들을 기존의 시스템과 동일하게 사용하였으나, 앞으로 성능향상을 위해 이러한 부분을 최적화 해야 할 것이다. 또한, 본 논문에서는 기존의 반음소 단위를 개선하기 위하여 음소의 지속시간이 120ms 이상인 경우에 음소의 앞·뒤 경계에서 60ms를 전·후반음소로 하고 나머지를 안정화 구간으로 취하였으나, 보다 정확한 음소간의 경계값을 선정하기 위해서는 문맥에 따른 음소의 천이구간 지속시간에 대한 추가적인 연구와 사용자 인터페이스에 대한 연구도 있어야 할 것이다.

참고 문헌

[1] 김석수(1996), 「HMM을 이용한 자동 음성 분절에 관한 연구」, 울산대학교 전자공학과 석사학위 논문.

[2] 김종진, 김봉완, 이용주(1996), 한국어 음성데이터베이스 구축을 위한 한국어 레이블링 기준에 관한 연구, 「제 13 회 음성통신 및 신호처리 워크샵 논문집」, 제 13권 1호, 250-255, 한국음향학회.

[3] 김태환(1998), 「문맥종속 반음소단위를 이용한 자동 음소분할 및 레이블링 시스템의 구현」, 원광대학교 컴퓨터공학과 석사학위 논문

[4] 김형순(1995), 음소 자동 레이블링의 현황 및 과제, 「제 12회 음성통신 및 신호처리 워크샵 논문집」, 제 12권 1호, 297-304, 한국음향학회.

[5] 이용주, 이숙향 외(1996), 「한국어 음성 데이터베이스의 구축에 관한 연구」, 한국 과학기술원 2차년도 최종보고서.

[6] 이종락(1993), 반음소 : 새로운 음성합성 및 인식단위, 「제 10회 음성통신 및 신호처리 워크샵 논문집」, 208-212, 한국음향학회.

[7] 송석일(1996), 「한국어 음성 인식을 위한 음소 분할에 관한 연구」, 한양대학교 전자통신공학과 석사학위 논문.

[8] 성종모, 김형순 외(1996), 한국 음성 데이터베이스 구축을 위한 반자동 음성분할 및 레이블링 시스템 구현, 「제 13 회 음성통신 및 신호처리 워크샵 논문집」, 제 13권 1호, 161-166, 한국음향학회.

[9] 최환진, 한무성, 박재득(1997), 음성인식에 있어서 특징파라미터의 기여도에 기반한 상태별 특징파라미터가중, 「한국음향학회 논문집」, 제 16권 2호, 71-74.

[10] 홍성태(1997), 「자동 음성분할과 레이블링 시스템의 성능향상」, 부산대학교 전자공학과 석사학위 논문.

[11] A. Komatsu, A. Ichikwa, D. Nakota, Y. Asakawa, H. Matsuzaka(1982), Phoneme recognition in the continuous speech, in *Proc. ICASSP*, 883-886.

[12] A. Ljolie, M. D. Riley(1991), Automatic segmentation and labeling of speech, in *Proc. ICASSP*, 473-476.

[13] B. Eisen, H. Tillmann, C. Draxler(1992), Consistency of judgements in manual labeling of phonetic segments: the distinction between clear and unclear cases, in *Proc. ICSLP*, 871-874.

[14] B. Wheatley, G. Doddington, C. Hemphilland J. Godfrey(1992), Robust automatic time alignment of orthographic transcriptions with unconstrained speech, in *Proc. ICASSP*, I-553-556.

[15] D.B. Roe, J.G. Wilpon(1994), *Voice Communication between Humans and Machines*, National Academy Press.

[16] J. P. Van Hemert(1991), Automatic segmentation of speech, in *Proc. ICASSP*, vol. 39-4, 1008-1012

[17] José B. Mariño, Albino Nogueiras, Antonio Bonafonte(1997), The Demi-phone: an efficient subword unit for continuous speech recognition, in *Proc. ICASSP*, 1997.

[18] K. F. Lee(1998), *Large-vocabulary speaker-independent continuous speech recognition : The SPHINX system*, Ph.D. dissertation, Comput. Sci. Dep., Carnegie Mellon Univ.

[19] K. F. Lee(1988), On large-vocabulary speaker-independent continuous speech recognition, *J. Euro. Assoc. Signal Processing(Speech Communications)*, no.7, 375-379.

[20] L. Rabiner, B. H. Juang(1993), *Fundamentals of Speech Recognition*, Prentice Hall.

[21] S. B. Davis, P. Mermelstein(1980), Comparison of parametric recognition for monosyllable word recognition in continuous spoken sentences, in *Proc. ICASSP*, Vol. Asp-28-4. 357-366.

[22] *The Aligner(1994), A System for Automatic Time Alignment of English Text and Speech*, Entropic Research Laboratory, Inc.

[23] Y. Gong, J. P. Haton(1992), DTW-based phone labeling using explicit phoneme duration constraints, in *Proc. ICSLP*, 863-866.

[24] Y. L. Chow, R. M. Schwartz, S. Roucos, O. Kimball, P. Price, F. kubala, M. Dunham, M. Krasner, J. Makhoul(1986), The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system, in *Proc. ICASSP.*

접수일자: 1999년 4월 3일
게재결정: 1999년 6월 14일

▶ 박순철(Soon-Cheol Park)
주소: 전라북도 익산시 신용동 344-2
소속: 원광대학교 컴퓨터공학과
전화: 0653) 850-6885
E-mail: bluejini@gaebyok.wonkwang.ac.kr

▶ 김봉완(Bong-Wan Kim)
주소: 전라북도 익산시 신용동 344-2
소속: 원광대학교 컴퓨터공학과
전화: 0653) 850-6885
E-mail: tacanemo@gaebyok.wonkwang.ac.kr

▶ 이용주(Yong-Ju Lee)
주소: 전라북도 익산시 신용동 344-2
소속: 원광대학교 컴퓨터공학과
전화: 0653) 850-6885
E-mail: yjlee@wonnms.wonkwang.ac.kr