인식 단위로서의 한국어 음절에 대한 연구

A Study on the Korean Syllable As Recognition Unit

김 유 진*, 김 회 린**, 정 재 호*
(Yu-Jin Kim*, Hoi-Rin Kim**, Jae-Ho Chung*)

요 약

본 논문에서는 한국어 대용량 어휘 인식 시스템에 적합한 인식 단위에 대하여 연구 및 실험하였다. 특히 현재 인식 시스템의 인식 단위로 주로 사용되는 음소와 한국어의 특징을 잘 나타내는 음절을 선택하고, 인식 실험을 통해 음절이 한국어 인식 시스템의 인식 단위로서 적합한가를 음소와 비교하였다.

객관적인 비교 인식 실험 결과를 제시하기 위하여 동일한 남성 화자의 음성 데이터를 수집하고, 수작업 음소 경계 및 레이블링 과정을 거친 음성 데이터 베이스를 구축하였다. 또한 각 인식 단위에 동일한 HMM 기반의 훈련 및 인식 알고리즘을 적용하기 위해 Entropic사의 HTK(HMM Tool Kit) 2.0을 사용하였다. 각 인식 단위의 훈련을 위해 5상태 3출력, 8상태 6출력 HMM 모델의 연속 HMM(Continuous HMM)을 적용하였고, PBW 3회분, POW 1회분을 훈련에 사용하고 PBW 1회분을 각 인식 단위로서 인식하는 화자 종속 단어 인식 실험을 구성하였다.

실험 결과 8상태 6출력 모델을 사용한 경우 음소 단위는 95.65%, 음절 단위는 94.41%의 인식률을 나타내었다. 한편 인식 속도에서는 음절이 음소보다 약 25% 빠른 것으로 나타났다.

ABSTRACT

In this paper, study and experiments are performed for finding recognition unit fit which can be used in large vocabulary recognition system. Specifically, a phoneme that is currently used as recognition unit and a syllable in which Korean is well characterized are selected. From comparisons of recognition experiments, the study is performed whether a syllable can be considered as recognition unit of Korean recognition system.

For report of an objective result of the comparison experiment, we collected speech data of a male speaker and processed them by hand-segmentation for phoneme boundary and labeling to construct speech database. And for training and recognition based on HMM, we used HTK(HMM Tool Kit) 2.0 of commercial tool from Entropic Co. to experiment in same condition. We applied two HMM model topologies, 3 emitting state of 5 state and 6 emitting state of 8 state, in Continuous HMM on training of each recognition unit. We also used 3 sets of PBW(Phonetically Balanced Words) and 1 set of POW(Phonetically Optimized Words) for training and another 1 set of PBW for recognition, that is "Speaker Dependent Medium Vocabulary Size Recognition."

Experiments result reports that recognition rate is 95.65% in phoneme unit, 94.41% in syllable unit and decoding time of recognition in syllable unit is faster by 25% than in phoneme.

I. 서 론

1950년대부터 시작된 음성 인식 연구는 인간과 자연스럽게 대화하는 기계 구현을 목표로 지난 40여년 동안 진행되어 왔다. 아직도 임의의 화자의 음성을 어떤 환경에서도 인식할 수 있는 가장 자연스런 음성 인식 기술은 완성되지 못했지만 음성 인식 기술은 다양한 응용 분야에 적용되어지고 있고 궁극적인 음성 인식 기술의 가능성을 보여주고 있다.[1][2]

대표적인 응용 분야로서 전화선을 통한 각종 정보 제공 서비스, 개인용 컴퓨터를 들 수 있다.[3][4] 아직도 화자 독립 연속어 인식 기술의 난제가 완벽하게 해결되지 않았기 때문에 인간의 자연스런 발성을 인식하지 못하지만 이러한 응용 분야에서 고립 단어와 같은 제한된 발성 방법은 사용자의 이해, 숙달 등을 통해서 극복될 수 있다. 따라서 어느 정도의 화자 독립 인식 기술과 접목된 대용

*인하대학교 전자공학과 디지탈 신호처리 연구실
**한국전자통신연구소 음성언어연구실

접수일자 : 1997년 1월 24일

량 어휘 인식 기술만으로 성공적인 응용 분야를 찾을 수 있을 것이다.

대용량 어휘 인식 시스템을 구현하기 위해 연구되어져야 할 분야 중의 하나는 인식 단위에 대한 연구이다. Kai-Fu Lee는 이러한 인식 단위의 선택 조건으로 문맥 민감성, 훈련성, 공유성의 3가지를 제시하였다.[1]

각 인식 단위의 3가지 조건을 비교한 결과는 표 1과 같다.

표 1. 대용량 어휘 인식 시스템을 위한 인식 단위 비교[1]
Table 1. Comparison of recognition units for large vocabulary recognition[1]

인식 단위	문맥 민감성 (Context Sensitivity)	훈련성 (Trainability)	공유성 (Sharability)
단어	좋다	나쁘다	나쁘다
음소	나쁘다	좋다	필요 없다
다음소 모델	좋다	어렵다	나쁘다
전이 모델	좋다	어렵다	나쁘다
단어 종속 음소	좋다	어렵다	좋다
문맥 종속 음소	좋다	어렵다	좋다

결론적으로 음소 단위는 훈련성에서 뛰어난 장점을 가지고 있지만 문맥의 조음 현상은 극복하지 못한다. 따라서 이러한 단점을 극복하기 위해 Hunt[5], Rosenberg[6] 등에 의해 조음 현상을 포함할 수 있는 좀 더 긴 다음소(multi-phoneme) 모델에 대한 연구가 진행되었다. 그러나 이들 단위의 중심 음소를 제외한 시작, 끝 음소는 여전히 문맥에 따른 조음 현상에 둔감하며 음절은 약 2만 여개, 반음절은 약 1,000여 개의 인식 단위를 훈련해야 하는 치명적인 단점을 가지고 있다.[1][7] 따라서 조음 현상에 민감하고 공유성을 통해 훈련성의 단점을 극복할 수 있는 triphone과 같은 문맥 종속 단위(contextual dependent unit)가 현재로선 가장 우수한 인식 단위로 알려져 있다.

한국어 음성 인식에서도 음소와 문맥 종속 음소가 인식 단위로 사용되어 만족할 만한 인식률을 보이고 있다.[3][4][8] 그러나 한국어에서도 음소 이외에 한국어의 표기 특성을 잘 나타내는 단위인 음절에 대해서 많은 관심을 보이고 있으며 일부 연구가 진행되었다.[9][10] 한국어의 음절 단위는 훈련성에 있어서 음소보다 불리하고, 조음 현상에 둔감하다는 단점과 2~3음소의 결합으로 발생하는 조음 현상을 포함하는 장점을 영어의 음절 단위와 동일하게 지니고 있다. 그러나 한국어 음절 단위는 한국어의 특성상 음소보다 훈련과 인식에서 음절을 분리해내기 쉽고, 음소보다는 많지만 영어의 음절보다는 적은 약 1000여 개의 모델로 한국어를 모두 인식할 수 있다는 가능성을 가지고 있다.[9][11] 따라서 한국어 음절은 대용량 어휘 인식을 위한 인식 단위로서 그 가능성을 고려해볼 수 있다고 사료된다.

본 논문에서는 지금까지 음절에 대한 음성 인식 연구와는 달리 이러한 음소, 음절 인식 단위에 대해 동일한 인식 실험을 수행 함으로서 인식 단위를 선택함에 있어 좀 더 객관적인 근거를 제시하고자 하였다. 두 가지 인식 단위에 대해 동일한 인식 실험을 수행하기 위해 훈련 및 인식에 약 5600 여 단어로 구성된 음성 데이터 베이스를 사용하였고 동일한 훈련 및 인식 알고리즘의 적용을 위해 HTK 2.0을 사용하였다. 또한 이러한 실험에 사용된 음성 데이터와 훈련 및 인식 결과를 비교함으로써 각 인식 단위의 장·단점을 구체적으로 제시하고자 노력하였다.

본 논문의 전체적인 구성은 다음과 같다. 2절에서는 각 인식 단위의 특징을 음운론적인 관점에서 설명하고 비교하였다. 3절에서는 인식 단위 비교를 위해 수행된 실험의 전체적인 방법과 과정에 대해 설명하고 4절에서는 실험에 대한 결과와 결과에 대해 고찰한다. 마지막으로 5절에서 결론을 맺는다.

Ⅱ. 인식 단위 고찰

신호처리 관점에서의 음성 인식은 의사소통의 수단인 언어보다는 의미를 전달하는 수단인 음성 즉 음성 신호를 다룬다. 따라서 음성 인식에서의 인식 단위는 곧 음성 기관에서 음성이 만들어지는 과정을 연구하는 조음 음성학에 따른 음성 분류와 관련된다. 좀 더 구체적으로 음소는 조음적, 음향적인 단위라고 말할 수 있으며 음절은 음성학적, 음운론적 단위라고 할 수 있다.[11] 각 인식 단위에 대해 구체적으로 알아보면 다음과 같다.

2.1 음소[11]

음성학적인 구분에 의한 자음, 모음 등의 음들 가운데 서로 다른 소리로 인식하지 않는 음들을 묶어서 음소(phoneme)라고 부른다. 다시 말해 음소는 서로 구별되어 쓰이지 않는 음들의 집합이라고 말 할 수 있다. 한편 한 음소를 이루는 음들을 변이음(allophone)이라고 하며 이러한 변이음들에 대한 정의는 학파와 학자에 따라 다르다.

일반적으로 국어의 자음 체계에 따른 19개의 자음은 다음과 같다.

ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ

또한 8개의 모음 및 2개의 반모음(/w/, /j/)과 모음의 결합으로 만들어지는 12개의 이중 모음은 다음과 같다.

ㅣ ㅡ ㅜ ㅔ ㅓ ㅗ ㅐ ㅏ
ㅟ ㅞ ㅝ ㅛ ㅒ ㅑ ㅟ ㅖ ㅙ ㅘ ㅢ

인식 단위로서의 음소는 지금까지의 정의한 기본적인 음소 외에 유성음화 된 자음과 초성과 종성의 위치에 따라 구분되는 음성학적인 음소를 추가할 수 있다. 이는 두드러진 변이음들을 구분함으로써 개개 음소의 변별력을 높이고자 함이 목적이며 동시에 인식률을 향상시키기 위함이다. 또한 일반적인 발성 특성에 의해 '외', '왜', '웨'

'ㅗ'의 음소와 '에', '애' 등의 음소는 동일한 음소로 취급한다. 이상의 정의에 따라 대략 인식에 사용되는 한국어 음소의 개수는 약 40~50개 정도이다.

2.2 음절[11]

자음이나 모음과 같은 분절음이 이어지면 분절음보다 큰 음운론적 단위가 생겨나는데 그 중에서 순수한 음성학적·음운론적 단위로서 음절을 정의할 수 있다. 음절은 다음과 같은 특성을 가진다.

첫째, 음절은 하나 이상의 분절음으로 ·, 성된다.

둘째, 음절은 더 이상 쪼갤 수 없는 최소의 발음 가능한 단위이다.

세째, 음절은 '(초성) +중성 +(종성)'의 구조를 가지며 중성은 필수적인 성분이다.

네째, 음절은 음성학적으로 공명도가 큰 분절음을 중심으로 그 앞에서는 공명도가 점점 커지고 그 뒤에서는 점점 작아지는 모습을 하고 있다.

다섯째, 음절은 운율적 요소가 걸리는 가장 일반적인 단위이다.

음절의 구성을 살펴보면 다음과 같다.

C^iVC^f [C:자음, V:모음, C^i:초성, V:중성, C^f:종성]

[C^i] 의 위치에는 18개의 자음이 올 수 있다. (초성 'ㅇ' 제외)

[V] 의 위치에는 8개의 단모음과 12개의 이중모음이 올 수 있다.

[C^f] 의 위치에는 ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ 등의7개의 자음만이 올 수 있다.

이상의 정의에 의해 발생 가능한 음절의 수를 계산해 보면 다음과 같다.

V			20 개
CV	18 개×20개	=	360 개
VC	20개×7개	=	140 개
CVC	18개×20개×7개	=	2520 개

따라서 이론적으로 계산되는 음절의 수는 3,040개이나, 실제로 쓰이는 수는 음소의 연결에 제약이 있기 때문에, 이보다 훨씬 더 적다고 알려져 1,096개로[주] 알려져 있다.

2.3 인식 단위의 비교 및 정의

음소는 가장 기초적인 음성 단위이고 풍부한 훈련 데이터를 얻을 수 있다는 장점 때문에 대부분의 HMM 기반의 음성 인식 시스템에서 적용되어져 왔다. 그러나 음소는 가장 기본적인 단위이므로 수 많은 문맥이(context) 존재할 수 있고 이러한 이유로 동일한 음소라도 전후의 환경에 따라 발생하는 다양한 음향학적 변화에 민감하다는 단점이 있다.[1][7][11] 따라서 이러한 단점을 극복할 수 있는 문맥 종속 음소를 정의하기에 이르렀고 음소의 앞, 뒤의 음소를(변이음) 고려한 triphone을 인식 단위로 사용하고 있다. 그러나 문맥 종속 음소는 문맥에 따라 모든 음소를 다른 모델로 정의하므로 많아지는 훈련 모델과 그에 따른 훈련 데이터가 부족한 문제가 발생하게 된다.[1][7][12]

특히 대용량 어휘 인식에서 인식 대상이 되는 어휘의 개수는 한정될 수 있지만 음소의 문맥 환경은 제한 할 수 없으므로 훈련에 필요한 문맥 종속 음소의 수는 무한 어휘 인식에 필요한 그것과 크게 다르지 않으리라 예상된다. 음절은 이에 비해 2~3음소의 결합으로 구성되는 동시 조음 현상을 포함하며 특히 고립 단어 인식에서는 단어를 쉽게 음절 단위로 분해하고 구성할 수 있다는 장점이 있다. 또한 음절은 문맥 종속 음소와는 달리 인식 대상 어휘에 따라 그 개수가 한정될 수 있다.

예를 들어 445 단어로 구성된 PBW(Phonetically Balanced Words)를 인식하기 위해 훈련되어져야 하는 문맥 종속 음소의 수는 이론적으로 대략 1,807개이다. 하지만 필요한 음절은 대략 450개로 조사되었다. 따라서 음소의 문맥에 따른 동시 조음 현상을 극복하기 위해 문맥 종속 음소가 아닌 음절을 고려해보는 것은 타당한 것이라 할 수 있다.

2.4 인식 단위의 정의

본 논문의 실험에서 인식 단위로서 사용한 음소는 일반적인 한국어 음소의 분류에 따라 정의했다. 따라서 모음 20개와('ㅚ'와 'ㅞ'는 동일한 음소로 취급) 초성에서의 자음 18개를 사용했으며 종성에서의 자음 7개를 추가하였다. 묵음을 음소에서 제외시킬 경우 총 45개의 음소 단위를 정의했다.

인식 단위로서의 음절은 한국어의 경우 약 1,000여 개로 볼 수 있지만 본 연구에서 수집한 음성 데이터로는 모든 음절들을 충분히 훈련시킬 수 없다. 따라서 인식 대상 어휘를 구성하는 음절만을 선택하여 음절 모델을 정의하였다. 각 음절의 표기는 간단히 음소 표기법을 조합하여 쉽게 정의할 수 있다. 예를 들면 다음과 같다.

BE(베), BEr2(벨), BU(쁘), Ba(빠), BaN(빵)

본 연구에서 사용된 음절 모델은 인식용 PBW 1set의 단어를 인식하기 위한 458개의 음절만을 사용하였다.

Ⅲ. 실험 방법 및 과정

본 논문의 실험은 동일한 화자가 발성한 PBW 단어 483개를 인식하는 화자 종속 중규모 어휘인식 실험으로 구성하였다. 이는 본 연구가 인식 알고리즘이나 시스템의 구현보다는 인식 단위에 대한 기초적인 연구에 대해 초점을 맞추고 있기 때문이며 객관적인 화자 독립 인식 결

[주] 허웅, 국어 음운학, pp.213, 정음사, 1982.

과를 내기 위해서는 필요한 대량의 음성 데이터를 수집하고 가공해야 하기 때문이다.

3.1 음성 데이터 베이스

본 논문의 실험에서는 동일한 음성에서 음소와 음절의 2 가지 음성 단위를 구분하여 인식 실험 및 비교를 수행하므로 음성을 수집하고 각 음성 단위로 분할 및 레이블링한 음성 데이터 베이스가 필수적이다.

실험을 위해 수집한 음성 데이터는 대학 교육 방송국의 아나운서가 발성한 단어 집단이다. 녹음된 단어 집단은 한국 전자 통신 연구소에서 음성 인식 연구를 위해 제안한 PBW와 POW(Phonetically Optimized Words)이다.[13] PBW는 2 음소열을 고려한 445개, POW는 triphone을 고려한 3,848개의 단어로 구성되어 있다.

음성 데이터는 동일한 남성 화자가 방송국 녹음실에서 PBW 4회, POW 1회를 발성한 음성을 DAT(Digital Audio Tape) 매체를 이용하여 44.1Khz 샘플링 주파수로 A/D 변환하여 녹음되었다. 각 단어들은 단어 사이에 묵음을 두고 연속적으로 발성되었다. 수집된 음성 데이터는 다시 16KHz, 16 비트의 해상도를 갖는 디지털 신호로 A/D 변환하였다. A/D 변환 작업은 수집된 단어가 대략 5628개로 매우 많은 양이고 추후 음소 분할 및 레이블링 작업의 편이를 위해 단어 단위가 아닌 평균 50여 개의 단어 단위로 변환되었다. A/D 변환된 음성 데이터들은 실제적인 데이터 베이스를 구성하기 위한 처리 과정을 거친다. 이 처리 과정은 첫번째로 모든 데이터들을 음소 경계로 분할하고, 두 번째로 분할된 음소의 시작점과 끝점을 표기하는 레이블링(labeling) 순으로 이루어진다.

본 연구에서는 음소의 경계점 외에 음절의 경계점도 알아야 한다. 따라서 음절을 구성하는 음소들을 모아서 다시 레이블링 하는 다단계 레이블링 과정을 거쳤다. 물론 음절의 경계 분할은 음절의 첫 음소와 마지막 음소의 시작, 끝 위치로 경계를 삼았다. 이러한 레이블링이 필요한 이유는 각 음소, 음절 모델로 HMM을 초기화할 때 정확한 음소 및 음절의 데이터를 제공하는 것이 전체 인식률에 큰 영향을 미치기 때문이다. 특히 음절의 경우 레이블링 오류로 인한 모델 초기화의 오류 및 훈련에서의 오류는 인식에서 치명적인 인식 오류를 발생시킨다.

3.2 HTK(HMM Tool Kit)[14][15][16]

인식 단위 훈련 및 인식은 HMM을 기반으로 한 음성 인식 시스템 구현 및 실험을 위한 상용 도구인 HTK를 사용하였다. HTK는 특별히 음성 인식 분야에서 HMM을 도구로서 사용할 수 있도록 정형화된 구조를 가지고 있으며 HTK가 제공하는 기능들을 적절하게 사용할 때 대량의 음성 데이터를 다루는 인식 실험을 정확하고 빠르게 수행할 수 있는 장점을 가지고 있다. HTK는 크게 음성 데이터와 관련된 각종 데이터를 처리하는 도구, 훈련을 위한 도구 그리고 인식 및 분석을 위한 도구들로 나누어져 있다. 각 도구들은 인식 시스템 구현 및 실험을 위한 각 단계별로 사용하도록 구성되어 있다.

3.3 훈 련

인식 실험을 위한 훈련 과정은 크게 특징 파라미터 추출, HMM 위상(topology) 정의 그리고 HMM 모델, 즉 인식 단위 모델 초기화 및 재추정으로 나눌 수 있다.

본 실험에서는 음성 데이터에 20ms 길이의 해밍 윈도우를 씌우고, 10ms마다 12차 LPC 켑스트럼, 12차 LPC 델타 켑스트럼 그리고 에너지와 델타 에너지 등의 특징 파라미터를 구하여 구성한 26차 특징 벡터를 음성 신호를 표현하기 위해 사용했다. HTK에서 음성의 특징 파라미터 추출은HCopy 도구에 의해 수행할 수 있다.

본 논문의 실험에서는 일반적으로 음성 신호의 특성을 잘 표현하는 left-to-right 형태를 가진 5상태 3출력 HMM 모델과 8상태 6출력의 2가지 HMM 모델을 사용하였으며(그림 1) 특히 동일한 HMM 위상에서 도약 경로(skip path)를 변화시켜 인식하는 실험을 수행하였다. 즉 5상태 3출력 HMM 위상에서는 3개, 1개의 도약 경로로 나누어 실험하였고 8상태 6출력 HMM 위상에서는 6개, 3개의 도약 경로를 가진 위상으로 나누어 실험하였다. 그림 1에서 굵게 표시된 도약 경로는 각각 5상태 3출력 HMM 위상에서 1개 도약 경로 일 때, 8상태 6출력 HMM 위상에서의 3개의 도약 경로 일 때를 나타낸다. HTK에서 HMM 위상 정의는 텍스트 파일로 정의된다. 본 실험에서 사용된 가우시안 분포는 단일 mixture를 가진 분포이며 이는 초기 실험 결과 mixture의 개수가 인식률에 거의 영향을 미치지 않았기 때문이다. 또한 HTK에서는 diagonal covariance로 표현되는 가우시안 분포만 지원된다.

HTK에서는 음성 데이터의 레이블링 정보를 이용하여 하위 단어 단위 HMM 모델의 초기화를 수행하는 HInit, 재추정을 수행하는 HRest 도구를 제공한다. HInit를 통한 초기화 과정에서, 훈련 데이터의 특징 벡터와 모델을 구성하는 각 상태 사이의 대응은 균등 상태 분할(uniform state segmentation)로부터 시작하여 Viterbi 알고리즘을 통한 최대 유사 상태열에(maximum likelihood state sequence) 근거한 확률값을 계산하고 다시 특징 벡터에 대응하는 상태를 분할하게 된다. 이 과정은 확률값이 문턱값(threshold) 이하로 수렴되었을 때 완료된다.[14][17]

한편 초기화 과정을 거친 하위 단어 단위 HMM 모델의 재추정은 전향 및 후향 확률을 이용한 Baum-Welch 알고리즘을 통해 수행된다. 이 과정은 최대 유사 상태열이 아닌 모든 상태열을 고려한 확률값의 수렴에 의해 완료된다.[14][17]

또한 HTK는 인식 단위의 경계 정보가 없는 연속으로 발성된 음성 데이터를 하위 단어 단위 HMM 모델의 훈련에 사용하도록 HERest 도구를 제공한다. 이 도구는 연속으로 발성된 음성 데이터의 transcription정보와 이미 초기화 또는 재추정된 하위 단어 단위 HMM 모델을 이용

하여 여러 개의 하위 단어 단위 HMM 모델로 구성된 합성(composite) HMM 모델을 생성한다. 생성된 합성 HMM 모델에 수정된 Baum-Welch 알고리즘을 적용하여 합성 HMM 모델을 1회에 한하여 재추정한다. 이 도구는 인식 단위의 경계 정보가 없이도 음성 데이터를 훈련에 사용할 수 있다는 장점 외에 인식 대상 단어를 구성하는 하위 단어 단위 간의 전이 현상을 하위 단어 단위 HMM 모델에 포함시킬 수 있다는 장점이 있다. 그러나 지나친 횟수의 재추정은 HERest에 사용된 훈련 데이타에 편향된(over-training) 하위 단어 단위 HMM 모델을 생성하게 될 가능성이 있다.[14] 이러한 문제점은 HERest의 수행 결과 계산되는 훈련 데이터의 프레임당 전체 확률값의 변화를 살핌으로써 더 이상의 향상이 없는 회수를 찾아내어 해결했다.

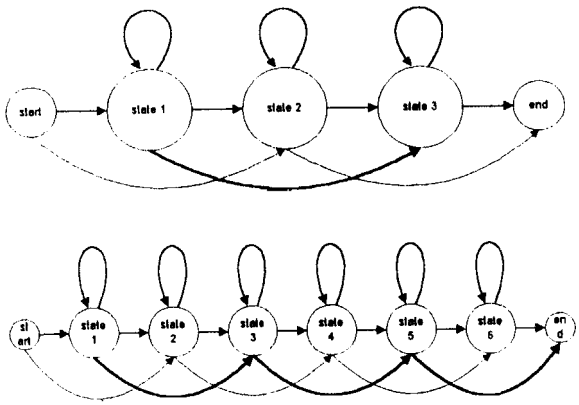


그림 1. 실험에 사용된 HMM 위상
Fig 1. HMM topology

3.4 인 식

본 논문의 인식 실험은 미리 인식용 PBW 1set을 선정한 후 이들 인식 대상 단어에 대해서만 인식을 수행하는 망을 구성하며 다시 각 단어의 인식은 사전을 통해 단어를 구성하는 하위 단어 HMM 모델들에 대해서만 복호화(decoding)를 시도하게 된다. 이러한 사전과 망을 이용한 음성 데이터의 복호화(decoding), 인식 과정은 HTK의 HVite도구에 의해 수행된다. 인식될 단어들로 구성되는 망은 단어 루프(Word Loop) 망으로 정의하였다. HTK에서 이러한 망의 구성은 SLF(Standard Lattice Format)로 표현된다. SLF는 노드(node)와 노드를 연결하는 경로(arc)로 구성되며 각 노드들은 단어를, 경로들은 단어 사이의 전이를 나타낸다.[14] 그러나 인식 대상 단어들을 직접 SLF로 표현하는 것이 쉽지 않으므로 단어망(word network)을 텍스트 파일로 정의한 후 HParse 도구를 이용하여 SLF로 변환할 수 있다. 단어 사전의 경우 본 연구의 실험이 음소와 음절의 2가지 하위 단어 단위로 인식을 수행하므로 각각 음소와 음절로 이루어진 단어 사전을 구성해야 한다. 사전은 텍스트 파일로 구성되며 '아파트'라는 단어를 음소와 음절 단위로 각각 정의한 예는 다음과 같다.

음소 단위	apatU	[아파트] a p a t U
음절 단위	apetU	[아파트] a pa tU

각각 5개의 음소와 3개의 음절로 '아파트'라는 단어가 정의되고 출력 심볼은 '아파트'임을 알 수 있다. 출력 심볼은 생략될 수 있는데 이 경우 출력 심볼은 단어로 대체된다. 본 연구의 실험에서는 PBW 4회분 중 1회를 선택하여 발생된 단어들로 망과 사전을 구성하였다.

음성의 복호화 과정을 위한 인식 도구인 HVite는 Viterbi 알고리즘을 구현한 도구로서 인식 결과는 단어 사전에 정의된 심벌로 구성된 transcription파일이다. 이 transcription파일은 레이블링 파일의 형식과 동일하고 미리 준비된 인식용 음성 데이터의 레이블링 파일과 인식 결과 만들어진 레이블링 파일을 동적 프로그래밍 기법을 이용해 비교하는 Hresults 도구를 이용하여 인식률을 분석하게 된다.

Hvite는 SLF로 표현된 단어망과 사전 그리고 하위 단어 단위 HMM 모델들을 참조하여 인식망을 생성한다. 다시 말해 인식망은 단어, 망 그리고 하위 단어 단위 HMM 모델 순으로 구성되어 각 모델의 상태들이 연결된 망을 형성하게 된다. 입력된 음성의 모든 관찰 벡터에 대해 구성된 인식망을 통해 복호화 과정을 거친 후 결과를 SLF 형태와 레이블링 파일의 형태로 출력한다.[14]

Ⅳ. 실험 및 결과

본 논문의 실험에서는 훈련 및 인식 알고리즘과 특징 파라미터를 동일하게 정하고 훈련 데이터의 양과 음성을 표현하는 HMM 모델의 위상을 변화시키면서 각 인식 단위에 의한 인식률 및 인식 속도를 비교하였다.

4.1 실험의 구성

첫번째 실험은(실험 1) PBW 3회분을 인식 단위의 초기화, 재추정 그리고 합성 HMM의 재추정에 사용하고 훈련에 사용되지 않은 PBW 1회분을 인식하는 실험으로 구성하였다. 이때 각 인식 단위의 특성에 맞는 HMM 모델을 찾아내기 위해 음소 및 음절을 표현하는 HMM 모델의 상태와 도약 경로의 수를 변화시키면서 실험하였다. 상태의 수를 변화시킨 것은 음절이 보통 2~3개의 음소로 구성되는 인식 단위이므로 인식 단위를 표현하기 위해 필요한 상태(state)의 수도 비례해야 한다고 예상하고 이를 확인해 보기 위함이다. 또한 도약 경로의 수를 변화시킨 것은 각 인식 단위의 평균 길이가 다르므로 각 인식 단위를 모델링하기 위해 필요한 최소 상태수의 변화에 따른 인식률의 변화를 살펴보기 위함이었다. 실험 1의 세부 실험을 다시 정리하면 다음과 같다.

표 2. 실험 1의 구성
Table 2. Experiment 1

실험	상태(state)의 수	도약 경로의 수
S3S3	3	3
S3S1	3	1
S3S0	3	0
S6S6	6	6
S6S3	6	3
S6S0	6	0

두 번째 실험은(실험 2) 첫번째 실험을 통해 각각의 인식 단위에서 가장 좋은 인식률을 나타낸 HMM 위상을 선택하여 POW 1회분을 인식 단위 훈련에 추가시키는 실험이다. 이때 POW는 인식 단위의 초기화와 재추정에만 사용하고, 실제 인식 단어에 대한 훈련인 합성 HMM의 재추정에는 첫번째 실험과 동일하게 PBW 3회분만을 사용했다. 이 실험은 훈련 데이터를 추가함으로써 훈련 데이터가 늘어났을 경우 각 인식 단위의 인식률 향상을 비교해보는 것이 목적이다.

마지막으로 각 인식 단위의 가장 좋은 인식률을 나타내는 실험에서 인식 속도를 측정하는 실험이다. 이상의 실험을 HTK를 이용한 과정으로 나타내면 그림 2와 같다.

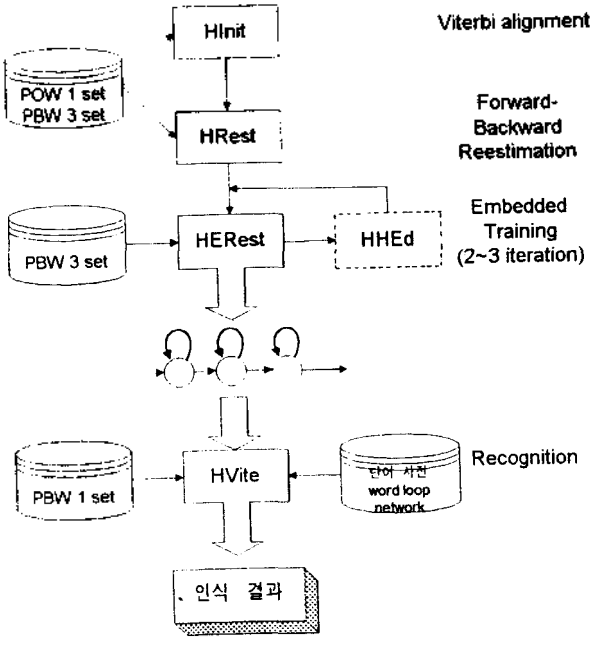


그림 2. 인식 실험 과정
Fig 2. Experiments procedure

4.2 훈련 데이터 통계

동일한 음성 데이터를 사용하더라도 각 인식 단위에 따라 훈련 데이터 양이 다르다. 이는 동일한 단어를 구성하는 음소와 음절의 수가 틀리기 때문이다. 따라서 인식 결과와 함께 각 인식 단위의 상대적인 훈련 데이터 양을 비교해 볼 필요가 있다.

그림 3은 실험1의 훈련에 사용된 PBW 3회분에서 발생된 음소와 음절 단위의 발생 분포도이고, 그림 4는 실험 2의 훈련에 사용된 PBW 3회분과 POW 1회분을 합한 음성 데이터에서 발생한 음소와 음절 단위의 분포도이다. 각 단위의 발생 횟수는 로그 척도(log scale)로 나타내었다. 분석 결과 훈련해야 할 하위 단어 단위의 수에서 음절이 음소보다 10배 가량 많으며 훈련량에 있어서는 음소가 음절보다 PBW 3회분에서는 약 19배, POW 1회분의 추가의 경우에는 약 22배 가량 많은 것으로 나타났다. 따라서 동일한 훈련 데이터가 주어질 경우 음소가 압도적으로 훈련량에서 유리함을 확인 할 수 있다.

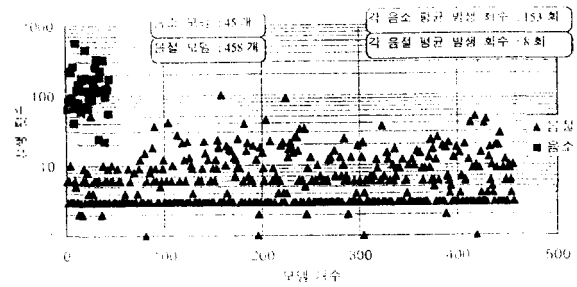


그림 3. PBW 3회분의 음소, 음절 분포도
Fig 3. Phonemes and syllables distributions of PBW 3sets

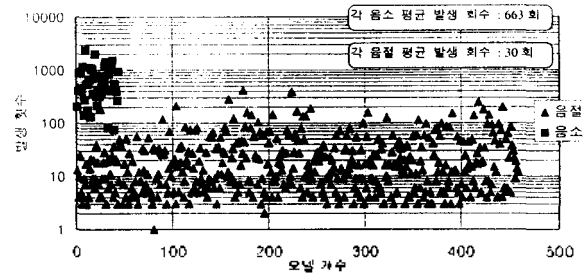


그림 4. PBW 3회분과 POW 1회분의 음소,음절 분포도
Fig 4. Phonemes and syllables distributions of PBW 3sets and POW 1set

특히 POW 1회분이 추가된 통계를 살펴보면(그림 4) 음소와 음절의 발생 빈도 차이가 더욱 커지는 것을 볼 수 있다. 이는 POW가 음소를 인식 단위로 사용하는 인식 시스템을 위해 고안된 단어 집단이기 때문으로 사료된다. 특히 POW는 음소 인식 단위의 단점인 문맥에 따른 민감성을 보완하기 위해 고안된 단어 집단이므로 훈련량의 증가와 함께 음소 단위에 최적화된 훈련 데이터로서 의미가 더욱 크다고 할 수 있다. 따라서 POW 1회분이 추가되었을 때의 인식률 향상은 음소가 클 것으로 예상할 수 있다.

4.3 실험 결과

HTK의 HResults 도구에 의해 얻어지는 인식률은 총 단

어 수(N), 인식된 단어 수(H), 삭제 오류(D), 대체 오류(S), 삽입 오류(I)의 횟수를 통해 다음과 같이 계산된다.[14]

$$Correct(\%) = \frac{N-(D+S)}{N} \times 100$$

$$Accuracy(\%) = \frac{N-(D+S+I)}{N} \times 100$$

4.3.1 실험 1-HMM 위상의 변화

실험 1의 결과 나타난 인식 오류와 인식률을 나타내면 다음과 같다.

실험 1의 결과 음소는 6 상태 6 도약 경로, 음절은 도약 경로가 없는 6 상태로 구성된 HMM 모델에서 가장 좋은 인식률을 나타내었다. 따라서 상태 수와 각 인식 단위의 인식률은 비례한다는 것을 알 수 있다. 또한 음소보다 음절이 상태의 수가 늘어났을 때의 인식률 향상이 큰 것으로 나타나 음절이 음소보다 많은 상태를 할당하는 것이 타당함을 확인할 수 있다.

도약 경로의 변화에 따른 결과를 살펴보면 음소는 최소 3개의 상태만을 거치는 것을 허용할 때, 음절은 6개의 상태에서 도약 경로가 없을 때 가장 좋은 인식률을 나타내었다. 이는 음소가 6개의 상태를 허용하면서 동시에 최소 3개의 상태를 거치도록 하는 유연성을 허용하여 적절한 모델링이 가능한 반면, 음절은 음소보다 긴 인식 단위이므로 6개의 상태수로는 이러한 유연성을 허용하지 못함을 알 수 있다. 따라서 음절은 7개 이상의 상태를 할당하고 도약 경로를 허용할 때 음소와 비슷한 모델링 효과를 보임을 예상할 수 있다. 또한 각 인식 단위 모두 도약 경로를 추가할 때 삽입 에러가 증가하는 것으로 나타났으며, 특히 음절 단위의 경우 하나의 음절 모델을 두개의 음절 모델로 인식하는 삽입 에러가 많은 것으로 나타났다.

결과적으로 비록 근소한 차이지만 음절이 음소보다 대체 오류가 6개 많고, 엄밀한 인식률인 accuracy 인식률에서 1.24% 뒤지는 것으로 나타났다. 그러나 훈련량에서 음절이 음소의 약 1/20 보다 적은 점을 감안한다면 음절 단위는 훈련량의 열세를 음소 단위보다 문맥에 따라 둔감하다는 점과 상대적으로 긴 인식 단위라는 장점으로 보완했다고 사료된다.

4.3.2 실험 2-POW의 추가

실험 2의 결과 나타난 인식 오류와 인식률을 나타내면 다음과 같다.

표 4. 실험 2의 인식 결과
Table 4. Results of experiments 2

실험	H	D	S	I	N	Correct(%)	Accuracy(%)
음소	466	0	17	2	483	96.48	96.07
음절	456	0	27	2	483	94.41	93.58

인식 단위의 초기화 및 재추정에 POW 1회분을 투입한 결과 음소의 인식률은 0.42% 향상됐지만 음절의 경우 오히려 0.83%가 떨어지는 것으로 나타났다. 이는 POW가 앞,뒤 음소와의 조음 현상을 훈련할 수 있도록 최적화된 단어 집단임을 고려할 때 조음 현상에 민감한 단점이 보완된 반면 음절은 상대적으로 인식 대상이 아닌 다른 단어를 구성하는 음절의 조음 현상을 모델링함으로써 인식률이 저하된 것으로 사료된다. 또한 투입된 POW 1회분의 단어가 PBW 3회분의 단어보다 약 2.8배 가량 많으므로 PBW보다는 POW에서 발생되는 음소, 음절의 영향을 많이 받았다고 할 수 있다.

4.3.3 인식 속도

Sun Sparc 5기종에서 가장 좋은 인식률을 보인 실험의 인식 속도를 측정한 결과는 다음과 같다.

측정 결과 음절이 음소보다 약 25% 인식 속도가 빠른 것으로 나타났다. 이는 상대적으로 인식 단어를 구성하

표 3. 실험 1의 인식 결과
Table 3. Results of experiment 1

실험		H	D	S	I	N	Correct(%)	Accuracy(%)
음소	S3S3	447	0	36	29	483	92.55	86.54
	S3S1	449	0	34	19	483	92.96	89.03
	S3S0	453	0	30	12	483	**93.79**	**91.30**
	S6S6	465	0	18	3	483	**96.27**	**95.65**
	S6S3	462	0	21	2	483	95.65	95.24
	S6S0	461	0	22	0	483	95.45	95.45
음절	S3S3	429	0	54	54	483	88.82	77.64
	S3S1	437	0	46	40	483	90.48	82.19
	S3S0	442	0	41	33	483	**91.51**	**84.68**
	S6S6	455	0	28	7	483	94.20	92.75
	S6S3	460	0	23	7	483	**95.24**	93.79
	S6S0	459	0	24	3	483	95.03	**94.41**

표 5. 인식 속도 비교
Table 5. Decoding time

인식 단위	측정 실험	인식 시간
음소	실험 1의 S6S6	1시간 40분
음절	실험 1의 S6S0	1시간 15분

는 모델이 음소보다 음절이 적기 때문으로 풀이된다.

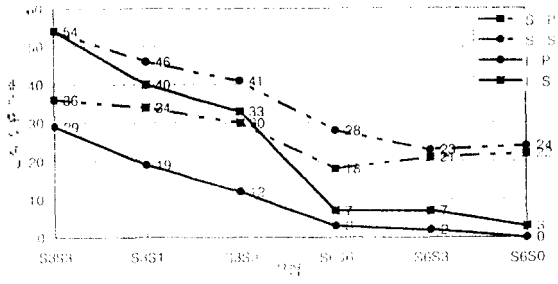


*그림 5. 실험 1의 인식 오류 비교
Fig 5. Recognition errors of experiment 1
S-P: 음소의 대체 오류(Substitution errors of phoneme),
S-S: 음절의 대체 오류(Substitution errors of syllable)
I-P: 음소의 삽입 오류(Insertion errors of phoneme),
I-S: 음절의 삽입 오류(Insertion errors of phoneme)

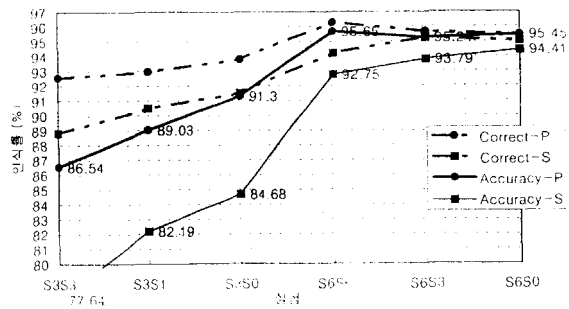


그림 6. 실험 1의 인식률 비교
Fig 6. Percent accuracy and correct of experiment 1
Correct-P: 음소의 Correct(Correct of phoneme),
Correct-S: 음절의 Correct(Correct of syllable)
Accuracy-P: 음소의 Accuracy(Accuracy of phoneme),
Accuracy-S: 음절의 Accuracy(Accuracy of syllable)

V. 결 론

본 논문에서는 HMM 기반의 대용량 어휘 고립 단어 인식 시스템을 위해 적합한 하위 단어 단위에 대해 연구하였다. 특히 음소와 음절 단위를 선택하여 동일한 음성 데이터와 동일한 훈련 및 인식 알고리즘을 적용한 비교 실험을 수행하여 각 인식 단위의 장·단점 및 인식률을 비교하였다. 인식을 위해 음소는 45개, 음절은 458개의 HMM 모델을 훈련했으며 동일한 훈련 데이터에서 음소는 음절에 비해 약 20배 가량 훈련량이 많았다. 인식 결과는 음소가 음절에 비해 인식률이 1.24%~2.49% 앞서는 것으로 나타났으며 인식 속도는 음절 단위로 인식 했을 경우가 약 25% 빠른 것으로 나타났다.

인식 결과, 동일한 훈련 데이터에서 음소는 압도적으로 우세한 것으로 나타났으며 상대적으로 음절보다 훈련과 인식이 용이함을 확인할 수 있었다. 그러나 음절 단위는 음소 단위보다 훈련량에서 압도적으로 불리하다는 단점에도 불구하고 비교할 만한 인식률을 나타내었다. 이는 인식 단위로서의 음절이 음소보다 문맥에 따른 영향에 둔감하다는 장점과 음소보다 긴 인식 단위이기 때문에 나타난 결과라고 사료된다.

음절은 훈련을 위한 음성 데이터의 처리 작업이 쉽고, 의미 및 운율을 가진 최소 인식 단위이며, 인식 속도가 빠르다는 장점을 가지고 있다. 앞으로 이러한 장점을 고려한 연구를 진행하고, 음절 단위의 훈련에 적합한 단어 집단 및 음절의 훈련량 부족을 보완할 수 있는 방법을 연구해나갈 것이다. 이러한 연구를 통해 음절에 대한 기초 연구가 진행된 후, 음절 단위에 적합한 HMM 위상 및 지속 시간 모델링의 방법에 대해서도 연구해야 할 것이다.

참 고 문 헌

1. K. F. Lee, *Automatic speech recognition: the development of the SPHINX system*, Kluwer Academic Publisher, Boston, 1989.
2. John R. Deller, Jr., John G. Proakis, John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.
3. 김재인, 구명완, "음성인식 증권정보시스템의 개발 및 시험 운용결과 분석," 제 13회 음성통신 및 신호처리 워크샵 13권 1호, pp. 185-191, 1996.
4. 이황섭, 김희린, 이정철, 김상훈, "PC에서의 어휘 독립 및 화자 독립 단어 인식기 구현," 제13회 음성통신 및 신호처리 워크샵 논문집 13권 1호, pp. 192-194, 1996.
5. Hunt, M. J., Lennig, M., Mermelstein, P., "Experiments in Syllable-Based Recognition of Continuous Speech," *IEEE International Conference of Acoustics, Speech and Signal Processing*, pp. 880-883, April, 1980.
6. Rosenberg, A. E., Rabiner, L. R., Wilpon, J., Kahn, D., "Demisyllable-Based Isolated Word Recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31(3), pp. 713-726, June, 1983.
7. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition, New Jersey*, Prentice Hall, 1993.
8. 김희린, 이황섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험", 제13회 음성통신 및 신호처리 워크샵 논문집 13권 1호, pp. 127-130, 1996.
9. 이영호, 정궁, "음절을 기반으로한 한국어 음성인식", 전자공학회논문집 제31권 B편 제1호, pp. 11-22, 1994.
10. 김주성, 이양우, 허강인, 안점영, "세그먼트 차원압축을 이용한 HMM의 음절인식," 한국음향학회지, Vol. 15, No. 2,

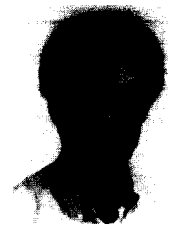
pp. 40-48, 1996.
11. 배주채, 국어 음운학 개설, 신구 문화사, 1996.
12. 이호영, 지민제, 김영송, "동시조음에 의한 변이음들의 음향적 특성," 한글 제 220호 별쇄본, 한글 학회, 1993년 6월.
13. 임연사, 이영직, "Large scale word recognizer를 위한 음성 database-POW," 제12회 음성통신 및 신호처리 워크샵 논문집, pp. 291-294, 1995.
14. Steve Young, *The HTK Book*, Entropic, 1996.
15. Entropic Research Laboratory, "Using HTK To Design A Speaker-Independent Connected-Digit Recognition System," 1993.
16. Steve Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognisers", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-569-572, 1992.
17. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceeding of the IEEE*, Feburuary, 1989.

▲김 유 진(Yu-Jin Kim)

1995년 2월 : 인하대학교 전자공학과 학사 졸업
1997년 2월 : 인하대학교 전자공학과 석사 졸업
1997년 2월~현재 : LG반도체 SD 연구소 DSP Gr. 연구원
※주관심분야 : 음성 합성, 인식/화자 확인, 인식/실시간 처리

▲김 희 린(Hoi-Rin Kim) : 음향학회지 16권 2호 참조

▲정 재 호(Jae-Ho Chung)



1982년 : 美 University of Maryland, College Park Campus(BSEE)
1984년 : 美 University of Maryland, College Park Campus(MSEE)
1990년 : 美 Georgia Institute of Technology(Ph.D.)
1984년~1985년 : 美 국방성 산하 해군 연구소, 신호처리실, 연구원
1991년~1992년 : 美 AT&T Bell Laboratories, 음성신호처리 연구실, 연구원(MTS)
1992년~현재 : 인하대학교 전자공학과, (현)부교수
1995년~현재 : 한국전자통신연구원 음성언어처리 연구실, 초빙 연구원