한국음향학회지 제21권 제2호 pp. 199~210 (2002)

# 결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘을 이용한 음성인식에 관한 연구

## A Study on Speech Recognition Using the HM-Net Topology Design Algorithm Based on Decision Tree State-clustering

오 세 진*, 황 철 준*, 김 범 국*, 정 호 열**, 정 현 열**
(Se-Jin Oh*, Chul-Joon Hwang*, Bum-Koog Kim*, Ho-Youl Jung**, Hyun-Yeol Chung**)

*대구과학대학 디지털정보통신계열, **영남대학교 전자정보공학부
(접수일자: 2001년 11월 7일; 채택일자: 2002년 1월 7일)

본 논문은 한국어 음성인식에서 음향모델의 성능개선을 위한 기초적 연구로서 결정트리 상태 클러스터링에 의한 HM-Net (Hidden Markov Network)의 구조결정 알고리즘을 이용한 음성인식에 관한 연구를 수행하였다. 한국어는 다른 언어와 비교하여 많은 문법과 변이음이 존재하는데, 국어 음성학에서 정의한 다양한 변이음을 조사하고, 음소결정트리를 위한 음소 질의어 집합을 작성하였다. 본 논문의 HM-Net 구조결정 알고리즘의 아이디어는 SSS (Successive State Splitting) 알고리즘의 구조를 가지면서 미리 작성해 둔 문맥의존 음향모델의 상태를 다시 분할하는 방법이다. 즉, 모델의 각 상태위치마다 음소 질의어 집합에 의해 음소결정트리를 생성하고, PDT-SSS (Phonetic Decision Tree-based SSS) 알고리즘에 의해 문맥의존 음향모델의 상태열을 다시 학습하는 방법이다. 결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘의 유효성을 확인하기 위해, 국어공학센터 (KLE)의 452단어와 항공편 예약에 관련된 YNU200 문장을 대상으로 음성인식 실험을 수행하였다. 인식실험 결과, 음소, 단어, 연속음성인식 실험에서 상태분할을 수행한 후 상태수의 변화에 따라 인식률이 점진적으로 향상됨을 확인하였다. 상태수 2,000일 때 음소, 단어 인식률이 평균 71.5%, 99.2%를 각각 얻었으며, 연속음성인식률은 상태수 800일 때 평균 91.6%를 얻었다. 또한 HM-Net 구조결정 알고리즘의 파라미터 공유관계를 비교하기 위해 상태공유를 수행하는 HTK를 이용한 단어인식 실험을 수행하였다. 실험결과, HTK를 이용한 문맥의존 음향모델에 비해 평균 4.0%의 인식률 향상을 보여, 본 논문에서 적용한 결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘의 유효성을 확인하였다.

**핵심용어:** HM-Net 구조결정 알고리즘, SSS 알고리즘, 음소결정트리, 상태 클러스터링, 문맥의존 음향모델, 상태대응확률

**투고분야:** 음성처리 분야 (2.5)

In this paper, we carried out the study on speech recognition using the HM-Net topology design algorithm based on decision tree state-clustering to improve the performance of acoustic models in speech recognition. The Korean has many allophonic and grammatical rules compared to other languages, so we investigate the allophonic variations, which defined the Korean phonetics, and construct the phoneme question set for phonetic decision tree. The basic idea of the HM-Net topology design algorithm is that it has the basic structure of SSS (Successive State Splitting) algorithm and split again the states of the context-dependent acoustic models pre-constructed. That is, it have generated the phonetic decision tree using

책임저자: 정현열 (hychung@yu.ac.kr)
712-749 경북 경산시 대동 214-1
영남대학교 전자정보공학부
(전화: 053-810-2496; 팩스: 053-814-5713)

the phoneme question sets each the state of models, and have iteratively trained the state sequence of the context-dependent acoustic models using the PDT-SSS (Phonetic Decision Tree-based SSS) algorithm. To verify the effectiveness of the above algorithm, we carried out the speech recognition experiments for 452 words of center for Korean Language Engineering (KLE452) and 200 sentences of air flight reservation task (YNU200). Experimental results show that the recognition accuracy has progressively improved according to the number of states variations after perform the splitting of states in the phoneme, word and continuous speech recognition experiments respectively. Through the experiments, we have got the average 71.5%, 99.2% of the phoneme, word recognition accuracy when the state number is 2,000, respectively and the average 91.6% of the continuous speech recognition accuracy when the state number is 800. Also we have carried out the word recognition experiments using the HTK (HMM Toolkit) which is performed the state tying, compared to share the parameters of the HM-Net topology design algorithm. In word recognition experiments, the HM-Net topology design algorithm has an average of 4.0% higher recognition accuracy than the context-dependent acoustic models generated by the HTK, implying the effectiveness of it.

***Keywords:*** *HM-Net topology design algorithm, SSS (Successive State Splitting) algorithm, Phonetic decision tree, State clustering, Context dependent acoustic models, State correspondence probability*

***ASK subject classification:*** *Speech signal processing (2.5)*

# I. 서 론

최근 대어휘 연속음성인식에 관한 연구에서는 통계적 음향모델이 널리 사용되고 있으며 높은 인식성능을 나타내고 있다[1,2]. 대어휘 연속음성인식을 위한 강건하고 정밀도가 높은 음향모델을 작성하기 위해 고려할 사항으로는 첫 번째, 음소의 문맥 환경에 의존하는 HMM (Hidden Markov Model)[14]을 이용하는 방법이 있으며, 두 번째는 많은 음성 데이터를 이용하여 모델의 파라미터를 추정하는 방법을 들 수 있으며, 세 번째는 음성 데이터의 양에 대해 적절하게 모델의 정밀도를 향상시킬 수 있는 방법 등이 일반적이라고 할 수 있다. 이러한 방법들이 고려되는 이유는 대규모 음성 데이터를 사용할 수 있을 경우 음소의 문맥환경에 의존하는 모델 파라미터의 공유에 의해 모델의 정밀도를 자유롭게 다룰 수 있기 때문이다. 이는 대규모 음성 데이터를 이용한다면 정밀한 음향모델을 작성할 수 있음을 의미한다. 하지만 대량의 음성 데이터를 이용할 수도 있으나 실제로 음성 데이터가 한정된 경우 임의의 파라미터 추정과 정밀한 모델을 작성하기 위해서는 많은 시행착오를 거쳐야 하는 문제점이 있다. 따라서 소규모의 음성 데이터를 효율적으로 사용하여 음소의 문맥환경 의존 HMM 모델을 작성하기 위해서는 적절한 파라미터의 공유관계를 추정하고 데이터 양에 적합한 모델의 규모를 선택하는 것이 중요하다고 할 수 있다.

소규모 음성 데이터를 이용하여 문맥 환경에 의존하는 음향모델을 작성하는데 있어 모델 파라미터의 공유 방법에 관한 연구가 많이 수행되고 있다. 그 중에서 정밀한 모델을 작성한 후 파라미터를 클러스터링 (Clustering) 하는 방법[3-6]과 파라미터를 계속 증가시키는 방법[7-12] 등을 예로 들 수 있다.

첫 번째 방법의 대표적인 연구는 음소결정트리에 의한 상태 클러스터링이 있다. 이 방법은 우선 정밀한 모델 (triphone, quinphone 등)을 학습해 두고, 중심음소가 동일한 HMM의 상태위치 등에 음소결정트리에 의한 상태 클러스터링을 수행하는 방법이다. 이 방법은 공유관계를 미리 작성해 둔 모델의 추정 정밀도에 크게 의존하기 때문에 많은 학습 음성 데이터를 이용해야 한다는 전제 조건이 있지만, 클러스터링 자체에 소요되는 시간은 비교적 적기 때문에 고속으로 대규모의 음향모델을 작성할 수 있는 방법이다. 음성학적 질의어를 이용한 yes와 no의 분할에 의해 미지 음소 문맥환경의 모델을 추정할 경우 여러 가지 장점이 있다. 하지만 상태의 위치마다 독립성을 가져야 한다는 가정이 있기 때문에 모델 전체로서 적

절한 공유관계를 구한다는 것은 확신할 수 없는 문제점이 있다.

두 번째 방법의 대표적인 연구는 SSS (Successive State Splitting) 알고리즘[10]을 들 수 있다. SSS 알고리즘은 처음부터 정밀한 음향모델을 작성하지 않고 상태의 분할과 파라미터의 추정을 반복학습에 의해 모델을 점진적으로 정밀하게 학습하는 방법이다. SSS 알고리즘에서는 HMM의 상태 공유관계를 은닉 마르코프망 (Hidden Markov Network; HM-Net)[10]이라고 하는 상태 네트워크로 표현할 수 있다. 비교적 적은 양의 음성 데이터를 이용하여 적절한 공유관계 (HM-Net 구조)를 구할 수 있으며 시간방향의 상태분할에 의해 음소의 문맥환경에 의존하는 상태의 길이를 설정할 수 있는 특징이 있다. 그러나 학습 음성 데이터를 이용하여 상태분할과 파라미터 추정을 반복하기 때문에 최종 모델을 학습하는데 필요한 계산량이 상대적으로 증가하는 문제점이 있다. 따라서 불특정 화자의 대규모 학습 음성 데이터 (수천에서 수만 문장)를 이용할 경우에는 크게 문제가 되지 않지만 대체로 모델의 구조를 결정하기 위해서는 특정화자의 학습 음성 데이터 (수백 문장 또는 수백 단어)를 이용하여 기본적인 파라미터를 추정한 후, 각 상태의 분포만을 불특정 화자의 대규모 학습 음성 데이터에 의해 재추정하는 방법을 이용하고 있다. 그러나 기본모델의 구조결정에 이용한 학습 음성 데이터가 특정화자의 한정된 음성 데이터이므로 대어휘 연속음성인식을 위한 음향모델을 작성하기 위해서는 많은 어려움이 따르며, 모든 화자의 학습 음성 데이터를 이용한 HM-Net 구조의 결정에도 어려움이 많은 문제점이 존재한다.

따라서 본 논문에서는 대어휘 연속음성인식을 위한 보다 정밀한 문맥의존 음향모델을 작성하기 위해 HM-Net 구조결정 알고리즘[13,16]을 적용한다. HM-Net 구조결정 알고리즘은 음소결정트리에 기반한 상태 클러스터링으로부터 상태위치에 대한 독립성을 배제하고 문맥방향과 시간방향의 상태분할을 수행하여 정밀한 문맥의존 HM-Net 음향모델의 구조를 결정할 수 있다. 즉, 미리 작성해 둔 문맥의존 음향모델의 각 상태위치마다 음소 질의어 집합에 의해 음소결정트리를 생성하고, PDT-SSS (Phonetic Decision Tree-based SSS) 알고리즘[16]에 의해 문맥의존 음향모델의 상태열을 다시 학습하는 방법이다. 본 논문에서 적용한 HM-Net 구조결정 알고리즘의 유효성을 확인하기 위해 국어공학센터 (center for Korean Language Engineering; KLE)의 452단어와 YNU (YeungNam University)의 200문장을 대상으로 음소, 단어 및 연속음성 인식 실험을 수행하고자 한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 일반적인 HM-Net 모델의 기본구조와 SSS 알고리즘에 대해서 설명한다. 그리고 Ⅲ장에서는 음소결정트리 기반 상태 클러스터링과 상태대응확률을 이용한 HM-Net 모델의 생성을 위한 구조결정 알고리즘에 대해서 소개하고, Ⅳ장에서는 인식실험 및 고찰에 대해서 기술한 후, 마지막으로 Ⅴ장에서 결론을 맺는다.

## Ⅱ. HM-Net과 SSS 알고리즘

### 2.1. HM-Net (Hidden Markov Network)

HM-Net은 SSS 알고리즘에 의해 HMM의 각 상태를 임의의 노드로 설정하여 네트워크로 연결한 구조로 표현되며 문맥의존 HMM의 각 상태를 서로 공유하게 된다. 각 상태는 상태번호, 가능한 문맥 클래스, 선행상태와 후행

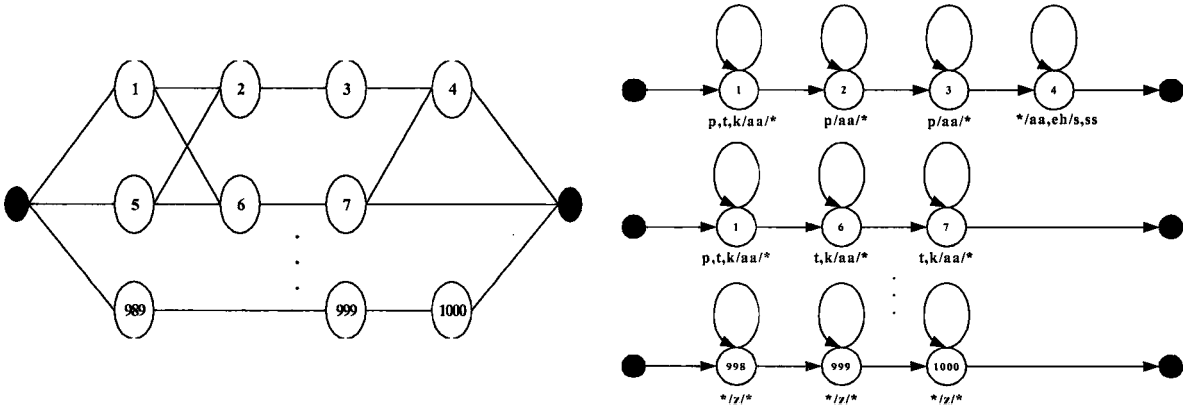


그림 1. HM-Net 모델의 예
Fig. 1. An example of HM-Net models.

상태 리스트, 자기천이 확률과 상태천이 확률, 그리고 출력확률 분포 파라미터 등의 정보를 가지고 있다. HM-Net에서는 문맥 정보가 주어질 경우 이 문맥을 만족하는 상태를 선행상태와 후행상태 리스트의 제약 조건 내에서 서로 연결하여 이 문맥에 대한 모델을 하나로 결정할 수 있다. 이 모델은 자기 루프와 이웃하는 상태로의 천이만을 허용하는 left-to-right 형 HMM과 동일하며 일반적인 HMM과 마찬가지로 Baum-Welch 알고리즘[14,15]에 의해 파라미터를 추정할 수 있다.

그림 1에 나타낸 HM-Net의 예에서 한 개의 음소를 전후해서 의존하는 triphone 모델의 경우 각 상태는 처리할 수 있는 문맥 클래스로서 선행/중심/후행 음소의 집합을 가지게 된다 (단, "*"는 모든 음소의 집합을 나타냄). 그림 1에서 첫 번째 경로는 2개의 triphone (p/aa/s, p/aa/ss)을 나타내는데 "p/aa/s"는 선행문맥 "p"와 후행문맥 "s"가 주어진 경우의 "aa"의 음향모델을 각각 나타낸다. 두 번째 경로는 96 개의 triphone (t/aa/*, k/aa/*)을 나타내는데, 여기서 "*"는 48개의 한국어 유사음소단위 중에서 임의의 하나를 의미한다.

### 2.2. SSS 알고리즘

SSS 알고리즘[10]을 전체적으로 간략히 설명하면 다음과 같다. 먼저 유사음소단위 (PLUs)를 기본단위로 하여 전체 모델을 연결한 구조의 초기모델로서 각각의 모델은 하나의 상태와 그 상태를 시단에서 종단까지 연결하여 전체 학습 음성 데이터로부터 생성한다. 상태의 분할은 경로 분할을 동반하는 문맥방향과 경로 분할을 동반하지 않는 시간방향이 있는데, 출력확률의 우도에 따라 한 방향으로만 수행된다. 문맥방향으로 분할할 때는 경로 분할에 동반된 각각의 경로에 할당된 문맥 클래스도 같이 분할된다. 따라서 문맥 클래스의 분할에 포함된 모든 상태 중에서 학습 데이터에 대한 누적 우도 확률이 가장 큰 쪽의 상태를 분할하도록 선택하게 된다. 시간 방향으로의 상태분할에서도 누적 우도확률이 높은 쪽의 상태를 분할하도록 선택하게 된다. 이러한 상태분할을 반복하여 HM-Net의 구조를 결정하게 된다. 하지만 상태를 분할하는 각 시점에서 가장 높은 우도확률을 가지는 상태를 분할하기 위해 분할 가능성이 있는 모든 상태에 대해서도 상태를 분할해야 하며, 상태분할 후에 추정된 파라미터의 누적 우도확률을 모두 구하여야 하는데, 이는 상당히 많은 계산과 메모리를 필요로 하는 작업이다. 위의 설명과 같이 SSS 알고리즘에 의한 HM-Net의 생성은 기본적으로 작은 상태수를 가지는 HM-Net으로부터 상태분할을 반복하여 보다 정밀한 HM-Net으로 성장시키는 작업을 수행한다.

## III. 상태 클러스터링에 의한 HM-Net의 생성

### 3.1. 음소결정트리 기반 상태 클러스터링

HM-Net의 구조결정은 음소결정트리 기반 상태분할[6]과 많은 관계가 있다. 음소결정트리는 그림 2에 나타낸 것과 같이 뿌리 (根)를 음소환경에 독립적인 단위로 하는 2진 트리로서 트리의 뿌리에서 잎 (葉) 방향으로 음소환경의 분할이 수행되는 계층적인 구조를 가지고 있다.

음소결정트리 기반 상태분할은 트리의 뿌리에서 음소 질의어에 의한 노드 분할을 반복하여 음소결정트리를 성장시키는 방법이다. 트리의 성장은 음소환경독립 HMM의 상태위치에 따라 수행된다. 기본적인 아이디어는 다음과 같다. 우선 분할할 모든 상태를 트리의 뿌리에 위치시키고, 분할하기 위한 기준에 따라 가장 좋은 음소 질의어에 의해 상태분할을 반복 수행한 후, 분할이 종료되는 각 잎의 상태에 하나의 상태를 공유시키는 방법이다. 음소결정트리 상태분할의 장점은 음소 질의어에 의한 yes와 no의 분할에서 학습 음성 데이터에 출현하지 않는 음소환경을 포함하는 모든 환경에 대해 상태열을 구한다는 것이다.

이 방법에 의해 생성된 상태공유 음향모델은 음소환경독립 HMM으로부터 문맥방향의 분할로만 생성된 HM-Net으로 생각할 수 있다. 그러나 음소결정트리는 상태위치마다 독립으로 생성되며, 이 사이 학습 샘플과 상태의 대응관계가 변화하지 않는다고 가정한다. 따라서 분할할 때마다 학습을 수행하는 SSS 알고리즘과 비교하여 학습 샘플에 대한 전체 모델의 공유관계를 구할 수 있다.

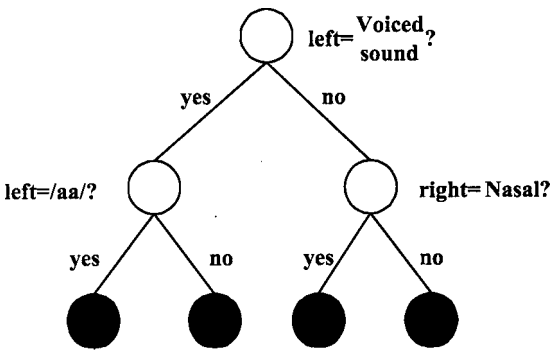


그림 2. 음소결정트리
Fig. 2. Phonetic decision tree.

SSS 알고리즘으로 상태분할을 수행할 때마다 학습 데이터를 이용하여 분포 파라미터를 재학습하는 방법도 고려할 수 있다. 그러나 상태분할에서는 각 상태의 분포 파라미터를 음소환경의존 HMM 파라미터의 대표 값 (중심)을 구한 후 상태분할을 먼저 수행하고 학습을 하기 때문에 다음 상태분할에서 다시 분포 파라미터를 구한다면 학습의 효과가 떨어지게 된다. 또한 상태분할을 수행하고 재학습할 때의 우도를 분할할 상태와 음소 질의어를 선택기준으로 이용한다면 상태분할을 상태수와 음소 질의어의 수를 곱한 회수의 재학습을 수행해야 하는 문제점이 있다.

### 3.2. 음소결정트리와 SSS 알고리즘과의 관계

SSS 알고리즘의 상태 분할과정은 2진 트리로 볼 수 있다. 즉, SSS 알고리즘은 상태단위의 트리구조 모델링으로 간주할 수 있기 때문에, 상태의 위치별 음소결정트리를 적용하는 방법과 공통점이 많다[4,5]. 예를 들어, SSS 알고리즘에서 문맥방향으로의 상태분할은 음소결정트리에서 노드 분할에 해당하고 HM-Net의 상태는 음소결정트리의 잎에 해당한다. 그러나 두 가지 방법의 기본적인 차이점은 SSS 알고리즘이 상태분할과 파라미터 추정을 반복하여 단계적으로 모델을 정밀화하는 방법인데 반해서, 상태의 위치별 음소결정트리는 이미 정밀하게 학습된 모델을 분할하는 방법이다. 두 알고리즘 중에서 SSS 알고리즘이 상태의 분할과 파라미터 추정에 많은 계산이 필요하지만, 보다 정밀한 상태공유 모델을 얻을 수 있다.

또한 문맥 클래스의 분할 방법에도 차이점이 있다. SSS 알고리즘은 모든 문맥의 분할에서 학습 데이터에 대해 우도확률이 최대가 되는 것을 기준으로 다음의 문맥분할을 위해 선택한다. 하지만, 모든 문맥의 조합이 선택되도록 조사하는 것이 어렵기 때문에 근사 계산을 도입한다. 이와 반대로 음소결정트리는 적용한 질문의 범위 내에서 우도확률이 최대가 되는 문맥을 분할하기 위해 선택한다. 따라서 모든 문맥의 조합이 선택되도록 조사할 필요가 없으며, yes와 no 형식의 분할에 의해 출현하지 않는 문맥을 반드시 어느 쪽의 클래스에 할당할 수가 있다. 따라서 문맥 클래스의 분할에 대해서는 음소결정트리가 SSS 알고리즘의 문맥방향 분할보다 효율적임을 알 수 있다. 그러나 음소결정트리는 질의어의 집합을 미리 준비할 필요가 있다. 이러한 연구로 과거에 발성학적인 음소의 유사성에 기반한 질의어가 많이 이용되고 있으며 적절한 보정을 통하여 출현하지 않는 문맥 클래스를 모델링하는데 효과를 나타내고 있다[4-7].

이상의 이유로부터 SSS 알고리즘에 음소결정트리의 문맥 클래스 분할법을 도입한다면, 각각의 장점을 접목한 알고리즘이라 할 수 있으며, 보다 정밀한 모든 문맥 클래스를 표현할 수 있는 HM-Net을 생성할 수 있게 된다.

### 3.3. HM-Net 생성 알고리즘

SSS 알고리즘은 상태분할을 수행할 때 우도의 계산과 분할 후의 파라미터를 추정하는데 많은 계산량이 필요한데 이는 학습 샘플의 양에 크게 의존한다. 이에 반해, 상태 클러스터링 방법은 먼저 정밀한 문맥의존 음향모델을 학습한 후, 각 모델의 상태를 다시 분할하기 때문에 파라미터 추정에 필요한 계산량은 모델의 수에 의존하므로 SSS 알고리즘보다 많지 않은 장점이 있다.

기본적인 아이디어는 SSS 알고리즘의 구조를 가지면서 미리 작성해 둔 문맥의존 음향모델을 다시 상태를 분할하는 방법이다. 즉, 모델의 각 상태위치마다 음소결정트리를 생성하고, 학습 음성 데이터를 이용하여 SSS 알고리즘에 의해 문맥의존 음향모델의 상태열을 학습하는 알고리즘이라고 할 수 있다. 이 방법은 상태분할을 수행하는데 속도가 빠르기 때문에 SSS 알고리즘에 의해 분할할 상태를 선택하여 분할하는 동시에 전체 분할 가능한 상태에 대해 상태분할을 수행하고 우도가 최대가 되는 상태를 선택하게 된다.

전체적인 알고리즘을 설명하기 전에, HM-Net 구조결정 알고리즘의 기본원리가 되는 상태대응확률과 기본적인 파라미터 추정 방법에 대해 설명한다.

#### 3.3.1. 상태대응확률

음소결정트리에 기반한 상태분할에서, 문맥의존 음향모델은 모두 같은 상태수를 가지며, 상태의 위치마다 공유화를 수행한다[4,5]. 따라서 음소결정트리에서 임의의 상태가 잎의 어느 상태에 대응되는가가 명확해야 한다. 그러나 본 논문에서 적용한 방법에 있어서는 각 문맥의존 음향모델의 길이와 HM-Net 중에 포함된 상태열의 길이가 일치하지 않기 때문에 문맥의존 음향모델의 각 상태가 HM-Net의 어느 상태에 대응되는가가 명확하지 않다면 이를 확률적으로 대응시키는 방법을 고려할 수 있다. 따라서 상태대응확률의 도입으로 인해 상태위치마다 공유관계가 독립이라는 가정은 없어지게 된다. 이하에 상태대응확률의 계산방법과 기본적인 분포 파라미터의 추정방법을 소개한다.

문맥의존 음향모델 $m$의 $n$번째의 상태 $s_{mn}$의 학습에 이용된 샘플의 집합 $X_{mn}$이 HM-Net의 상태 $S$로부터 출력되는 대수우도는 식 (1)과 같이 근사화할 수 있다.

$$L(X_{mn}|S) \approx \int_{-\infty}^{\infty} P(x|s_{mn}) \log P(x|S) dx \times f(s_{mn})$$
$$= -\frac{1}{2} \sum_{k=1}^{K} \left\{ \log(2\pi\sigma_{S_k}^2 + \frac{(\mu_{mnk} - \mu_{S_k})^2 + \sigma_{mnk}^2}{\sigma_{S_k}^2} \right\} \times f(s_{mn}) \quad (1)$$

식 (1)에서 $P(s_{mn}|S) = e^{L(X_{mn}|S)}$을 $S$로부터 $s_{mn}$이 생성되는 확률로서 정의한다. 여기서 $\mu_{mnk}$, $\sigma_{mnk}^2$, $\mu_{S_k}$, $\sigma_{S_k}^2$는 상태 $s_{mn}$과 $S$의 분포 (무상관 정규분포)의 $k$번째 차원의 평균과 분산을, $K$는 특징 벡터의 차원 수를 각각 나타낸다. 또한 $f(s_{mn})$은 상태 $s_{mn}$에 할당된 학습 샘플의 총 프레임 수 (임의의 상태에 점유하는 시간)로서 문맥의존 음향모델의 학습에서 미리 구한다.

상태의 시작과 종료는 반드시 대응되어야 하는 제약조건 아래 문맥의존 음향모델 $m$의 상태열 $s_{mn}(1 \le n \le N_m)$이 모델 $m$을 만족하는 HM-Net의 상태열 $S_i(1 \le i \le H_m)$로부터 생성될 때의 전향확률과 후향확률은 아래 식에 의해 계산된다.

$$\alpha_m(1, i) = \begin{cases} P(s_{m1}|S_{ni}) & i = 1 \\ 0 & otherwise \end{cases} \quad (2)$$

$$\alpha_m(n, i) = \sum_{k=i-1, i} \alpha_m(n-1, k) P(s_{mn}|S_i) \quad (3)$$

$$\beta_m(N_m, i) = \begin{cases} 1.0 & i = H_m \\ 0 & otherwise \end{cases} \quad (4)$$

$$\beta_m(n, i) = \sum_{k=i, i+1} \beta_m(n+1, k) P(s_{m(n+1)}|S_k) \quad (5)$$

여기서, $N_m$은 문맥의존 음향모델 $m$의 상태수, $H_m$은 문맥의존 음향모델 $m$을 만족하는 HM-Net의 상태열의 길이(단, $H_m \le N_m$)를 각각 나타낸다. 상태천이 확률은 상태의 대응관계에 거의 영향을 주지 않는다고 생각할 수 있으므로 여기서는 고려하지 않는다.

위의 전향확률과 후향확률에 의해 $s_{mn}$이 $S_i$에 대응하는 확률은 식 (6)에 의해 구할 수 있다.

$$\gamma_{S_i}(s_{mn}) = \frac{\alpha_m(n, i)\beta_m(n, i)}{\sum_n \alpha_m(n, i)\beta_m(n, i)} \quad (6)$$

상태대응확률 $\gamma_{S_i}(s_{mn})$을 이용한다면, 최우추정법 (Maximum Likelihood Estimation; MLE)[14,15]에 의해 HM-Net의 상태 분포 파라미터를 식 (7)과 (8)을 이용하여 구할 수 있다.

$$\hat{\mu}_{S,k} = \frac{\sum_{m \in C(S), n=1} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn}) \mu_{mnk}}{\sum_{m \in C(S), n=1} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn})} \quad (7)$$

$$\hat{\sigma}^2_{S,k} = \frac{\sum_{m \in C(S), n=1} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn}) (\mu^2_{mnk} + \sigma^2_{mnk})}{\sum_{m \in C(S), n=1} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn})} - \hat{\mu}^2_{S,k} \quad (8)$$

여기서 $C(S)$는 상태 $S$를 만족하는 문맥 상태를 나타내고, 위의 재추정식에 의해 상태분할시의 분포 추정 및 분할 후의 파라미터 재추정이 가능하다.

### 3.3.2. HM-Net 구조결정 알고리즘

본 절에서는 상태의 길이가 긴 문맥의존 음향모델의 작성과 상태분할에 의한 HM-Net의 구조결정 알고리즘 [13,16]을 소개하고자 한다. 전체 알고리즘의 구성을 그림 3에 나타내었으며 주요 알고리즘은 다음과 같다.

① 각 음소모델은 3상태, 단일 정규분포를 가지는 문맥독립 HMM 모델을 학습한다.

② 중심 음소가 일치하는 문맥독립 HMM의 파라미터를 문맥의존 HMM의 초기값으로 하고, 각 문맥에 대응하는 학습 음성 데이터를 이용하여 문맥의존 HMM 모델을 학습한다. 이때, 각 상태에 할당한 음소의 샘플 수와 총 프레임 수에 의해 각 상태의 평균 프레임 수를 식 (9)에 의해 계산한다.

$$\text{평균 프레임 수} = \frac{\text{총 프레임 수}}{\text{음소 샘플 수}} \quad (9)$$

③ 단계 ②에서 작성한 문맥의존 HMM 모델의 모든 상태를 대상으로 총 상태수가 미리 정한 임의의 수에 도달할 될 때까지 다음의 (a), (b)의 처리를 반복한다.

(a) 평균 프레임수가 가장 큰 하나의 상태를 선택하고, 시간방향으로 한 개를 복사한다. 단, 모델의 길이(상태수)에 대해 최대 길이와 최소 길이를 설정할 때, 최소 길이에 도달하지 않는 상태를 우선하고, 최대 길이에 도달하는 모델의 상태는 선택 대상으로부터 제외한다.

(b) 평균 프레임 수를 원본과 이것의 복사본에 대해 균등하게 분배한다.

④ 단계 ③에서 작성한 긴 문맥의존 HMM 모델을 단계 ②와 같은 방법으로 학습한다.

⑤ HM-Net의 초기모델의 구조를 정의하고, 각 상태의 분포 파라미터를 상태의 길이가 긴 문맥의존 HMM 모델의 분포 파라미터에 대해 식 (7)과 (8)을 이용하여 최우 추정한다. 초기모델의 구조는 임의로 정의하지만, 모든 음소는 1 상태의 HM-Net, 3 상태의 문맥독립 음향모델을 병렬로 접속한 HM-Net, 3 상태의 문

맥독립 음향모델의 1 상태와 3 상태마다 모든 음소를 공통으로 한 HM-Net 등을 고려한다.

⑥ 문맥상태를 분할하기 위해 음소 질의어를 이용하여 임의의 상태수가 될 때까지 다음의 (a), (b)를 반복한다.

(a) 분할할 수 있는 모든 상태에 대해 다음의 (i)와 (ii)의 문맥방향과 시간방향의 상태분할 처리를 수행한다.

(i) 문맥방향의 상태분할

우도의 비가 최대가 되는 음소 질의어에 대해 문맥방향으로 상태를 분할한다. yes측과 no측의 상태 분포 파라미터는 먼저 문맥상태를 분할하는 것으로 식 (7)과 (8)을 적용하여 구한다. (이때의 상태대응확률은 분할 전에 구한 값을 이용한다.) 우도 비는 분할 전과 분할 후의 대수우도의 기대값의 차로부터 구한다. HM-Net의 상태 $S$에 있어서 대수우도의 기대값은 식 (10)을 이용한다.

$$\begin{aligned} L(S) &= \int_{-\infty}^{\infty} P(x|S)\log P(x|S)dx \\ &\quad \times \sum_{m\in C(S),\, n=1} \sum_{n=1}^{N_m} \gamma_S(s_{mn}) f(s_{mn}) \\ &= -\frac{1}{2}\sum_{k=1} K(\log(2\pi\sigma_{Sk}^2)+1) \\ &\quad \times \sum_{m\in C(S),\, n=1} \sum_{n=1}^{N_m} \gamma_S(s_{mn}) f(s_{mn}) \end{aligned} \tag{10}$$

그리고 상태 $S$를 질의어 $q$에 의해 분할될 때의 우도 비는 식 (11)을 이용하여 계산한다.

$$L_G(S,q) = L(S_{q,yes}) + L(S_{q,no}) - L(S) \tag{11}$$

여기서, $S_{q,yes}$와 $S_{q,no}$는 상태 $S$를 분할 때의 yes 측과 no측의 상태를 각각 나타낸다.

(ii) 시간방향의 상태분할

시간방향으로 상태를 분할할 때, 분할할 상태의 분포 파라미터와 문맥상태를 새로운 상태에 복사한다. 그리고 분할하는데 영향을 주는 상태에 대해 분포 파라미터를 최우 추정 (식 (7), (8))한다. 분할에 영향을 주는 상태와는 분할 후 파라미터 추정에 의해 파라미터가 변화할 가능성이 있는 상태를 지정한다. 즉, 분할할 상태를 $S$, 상태의 집합 $\theta$내의 상태를 통과하는 모든 경로 상에 있는 상태의 집합을 $\zeta(\theta)$라고 하고, 분할에 영향을 주는 상태의 집합 $R(S)$는 $R(S) = \zeta(\zeta(\cdots\zeta(\{S\})\cdots))$로 표현할 수 있다.

상태 $S$의 시간방향 분할에 있어서 우도 비는 식 (12)와 같다.

$$L_G(S, temp.) = \sum_{s\in R(S)} \{L(s) - L_{old}(s)\} \tag{12}$$

여기서, $L_{old}(s)$는 재추정 전의 분포 파라미터에 대한 대수우도의 기대값을 나타낸다. 단, $S$의 복사본 $S'$에 대해서는 $L_{old}(S') = 0$이 된다.

(b) 모든 상태의 문맥방향과 시간방향의 분할 중 우도비

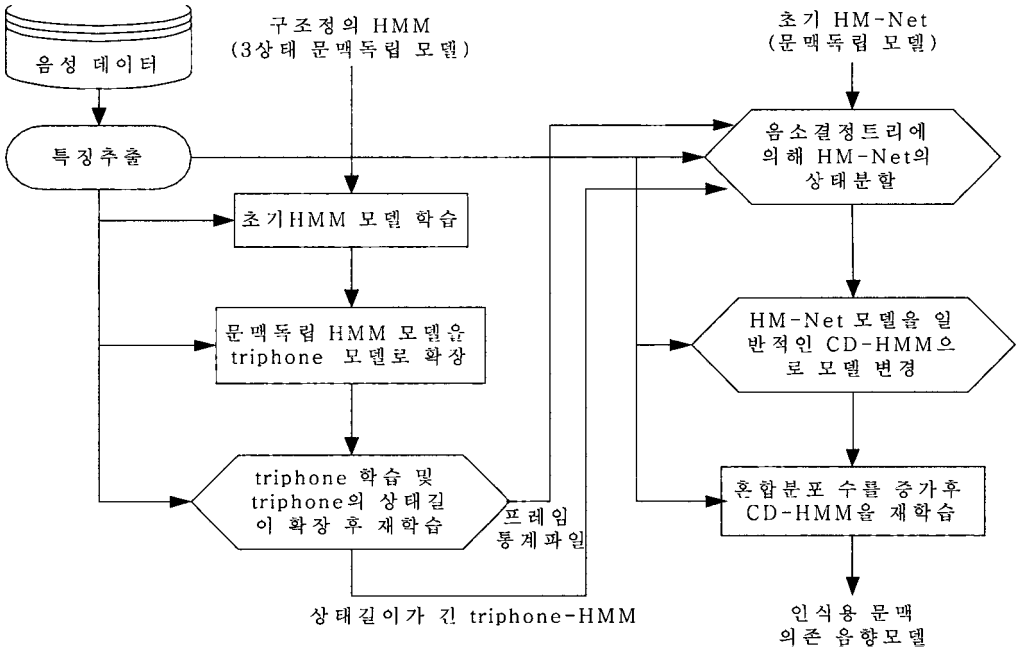


그림 3. HM-Net 구조결정 알고리즘의 전체 구성
Fig. 3. Overall diagram of HM-Net topology design algorithm.

가 최대가 되는 것을 선택하고 이 분할의 영향 및 상태에 대한 분포 파라미터를 최우 추정 (식 (7), (8)) 한다.

⑦ 각 상태의 혼합분포 수를 임의의 수로 증가시키고, 학습 음성 데이터를 이용하여 재학습 과정을 수행한다.

이상의 과정을 통해서 음성인식에서 사용되는 HM-Net triphone 음향모델을 완성한다.

문맥방향의 상태분할에서 yes측과 no측의 분포는 분할 전의 상태대응확률을 이용하여 구하는데, 여기서는 상태 분할의 전후에서 상태의 대응관계가 변화하지 않는다고 가정을 한다. 최적의 문맥방향의 분할을 선택하기 위해서 분할 처리는 최대 (질의어 수 X 상태수)의 회수만큼 처리가 필요하지만, 이 가정에 의해 계산량을 많이 줄일 수 있다.

한편, 시간방향의 상태분할의 경우에는 상태의 대응관계는 분할할 상태의 전후에서 크게 변한다고 생각되므로, 분할할 상태의 분포 파라미터를 새로운 상태에 복사하는 것으로 분할에 영향을 주는 모든 상태의 분포를 재추정한다. 이를 위해 분할할 필요한 계산량은 상당히 많지만, 최적의 상태분할을 선택하는데 필요한 분할처리는 (상태수) 회가 되며, 또한 전회의 분할에 있어서 선택된 분할 결과를 가지고 계산의 중복을 피한다면 그다지 문제가 되지 않는다.

# IV. 인식실험 및 고찰

## 4.1. 음성 데이터 및 분석조건

문맥의존 음향모델을 작성하기 위해 사용된 음성데이터는 국어공학센터 (KLE)의 단어음성과 본 연구실의 항공편 예약관련 200문장 (YNU200) 연속음성 데이터베이스를 사용하였다. 음향모델의 학습을 위해 452 단어를 35명이 1회 발성한 15,820단어와 200문장을 8명이 1회 발성한 1,600문장을 문맥의존 음향모델을 학습에 사용하였다. 평가용 음성 데이터는 음소 및 단어인식의 경우 학습에 참가하지 않은 국어공학센터 (KLE)의 3명이 1회 발성한 452단어를 사용하였고, 연속음성인식의 경우 학습에 참가지하지 않은 4명의 200문장을 사용하였다. 사용된 음성 데이터베이스를 표 1에 나타내었다.

표 2에 나타낸 것과 같이 모든 음성 데이터는 16 kHz의 샘플링과 16 bits로 양자화되었으며, $1-0.97z^{-1}$의 전달함수로 프리엠퍼시스하였으며, 25 ms의 해밍 윈도우를 곱하여 10 ms씩 이동하면서 분석하였다. 이를 통해

표 1. 음성 데이터 베이스
Table 1. Speech database.

| 실험 | 음소인식, 단어인식 | | 연속음성인식 | |
|---|---|---|---|---|
| 데이터명 | KLE452 | | YNU200 | |
| 발성형태 | 단어 | | 문장 | |
| 화자수 | 35명 | 3명 | 8명 | 4명 |
| 단어/문장수 | 452 | | 200 | |
| 발성회수 | 1 | | | |
| 사용단계 | 모델학습 | 인식 | 모델학습 | 인식 |
| 녹음환경 | 방 음 부 스 | | | |

표 2. 음성의 분석조건
Table 2. Analysis condition of speech.

| | |
|---|---|
| Sampling Frequency | 16kHz |
| Resolution | 16bit |
| Frame Length | 25ms |
| Frame Period | 10ms |
| Analysis Window | Hamming Window |
| Pre-emphasis | $1-0.97z^{-1}$ |
| Feature Parameter | 12차 LPC-MEL cepstrum + delta power + 1차와 2차의 회귀계수 (39차원) |
| Normalization | Cepstrum Mean Normalization |

음성 특징 파라미터는 12차 LPC-멜 켑스트럼 계수와 정규화된 대수 에너지에 1차 및 2차의 차분 성분을 포함하여 총 39차의 특징 파라미터를 구하였다.

## 4.2. 음향모델의 구조 및 생성

본 논문에서 사용한 음소는 표 3에 나타낸 것과 같이 유사음소단위 (PLUs)로 묵음정보 (sil)를 포함하여 48개를

표 3. 48개의 유사음소단위
Table 3. The definition of 48 phoneme likely units.

| 구 분 | 음 소 정 의 | | | |
|---|---|---|---|---|
| 모 음 | aa /아/ | axr /어/ | ao /오/ | uh /우/ |
| | U /으/ | ih /이/ | ae /애/ | eh /에/ |
| | ja /야/ | jv /여/ | jo /요/ | ju /유/ |
| | wa /와/ | wv /워/ | wE /외/ | we /웨,왜/ |
| | wi /위/ | je /예/ | jE /얘/ | Wi /의/ |
| 자 음 | b~ /ㅂ/ | d~ /ㄷ/ | g~ /ㄱ/ | z~ /ㅈ/ |
| | bb /ㅃ/ | dd /ㄸ/ | gg /ㄲ/ | zz /ㅉ/ |
| | p /ㅍ/ | t /ㅌ/ | k /ㅋ/ | ch /ㅊ/ |
| | s /ㅅ/ | ss /ㅆ/ | hh~ /ㅎ/ | r /ㄹ/ |
| | n /ㄴ/ | m /ㅁ/ | ng /ㅇ/ | |
| 첫음절 | b /ㅂ/ | d /ㄷ/ | g /ㄱ/ | z /ㅈ/ |
| | hh /ㅎ/ | | | |
| 종 성 | b1 /ㅂ/ | d1 /ㄷ/ | g1 /ㄱ/ | l /ㄹ/ |
| 묵 음 | sil | | | |

사용하였다. 상태 클러스터링에 의한 음소환경 의존 HMM 모델은 선행음소와 후행음소를 고려한 triphone 모델로 표현된다. 먼저 3상태의 단일 가우스 분포의 triphone 모델을 학습한 후 상태의 평균 프레임 길이에 따라 총 상태 수를 1.5배에 도달하는 긴 triphone 모델을 작성한다. 이때 각 모델의 최대 길이는 6상태, 최소 길이는 4상태로 제한을 둔다.

HM-Net 초기모델의 구조는 각 음소는 3상태의 음소환경 독립 HMM을 병렬로 연결하여 141개 상태를 가지는 HM-Net 모델을 이용하였다. 문맥방향의 상태분할에서 사용되는 음소 질의어 집합은 표 4에 나타낸 한국어 변이음 분류표[17-19]를 이용하여 총 162 종류 (문맥의 좌,

표 4. 한국어 변이음 분류표
Table 4. Korean allophonic variations.

| 구 분 | | | | 음 소 |
|---|---|---|---|---|
| 유성음 | | | | aa ih uh ae eh ao ja jv jo ju je wa wv wE we wi Wi axr U m n ng r l |
| 모 음 | | | | aa ih uh ae eh ao ja jv jo ju je wa wv wE we wi Wi axr U |
| 모음 | 혀위치 | 전설 | 비원순 | ih ae eh |
| | | | 원순 | wE we wi |
| | | 중설 | 비원순 | aa axr |
| | | | 원순 | |
| | | 후설 | 비원순 | axr U |
| | | | 원순 | uh ao |
| | 입크기 | 협 (狹) | | ih uh jo ju |
| | | 반협 (半狹) | | ae eh ao jv je wv wE we wi Wi axr U |
| | | 광 (廣) | | aa ja wa |
| | 혀높이 | 고모음 | | ih uh wi U |
| | | 중고모음 | | eh ao wE we axr |
| | | 중저고음 | | ae |
| | | 저모음 | | aa axr |
| | 좁힘점위치 | 경구개음 | | ih ae eh ao wE we wi |
| | | 연구개음 | | uh U |
| | | 인두음 | | aa ao axr |
| | 좁힘점간극 | 폐모음 | | aa ih uh wi U |
| | | 반폐모음 | | eh wE we axr |
| | | 개모음 | | ae ao |
| 자음 | 조음자리 | 양순음 | | wa wi b b~ b1 bb p m |
| | | 치(조)음 | | d d~ d1 dd t n s ss r l |
| | | 경구개음 | | ja z z~ zz ch |
| | | 연구개음 | | wa wi Wi g g~ g1 gg k ng |
| | | 성문음 | | hh hh~ |
| | 조음방법 | 파열음 | 무기연음 | b b~ b1 d d~ d1 g g~ g1 |
| | | | 유기경음 | p t k |
| | | | 무기경음 | bb dd gg |
| | | 파찰음 | 무기연음 | z z~ |
| | | | 유기경음 | zz |
| | | | 무기경음 | ch |
| | | 마찰음 | 무기연음 | s hh hh~ |
| | | | 무기경음 | ss hh hh~ |
| | 비 음 | | | m n ng |
| | 유 음 | | | r l |
| | 반모음 | | | ja wa wi Wi |
| | 묵 음 | | | sil |

우)를 작성하였으며, 이를 각 음향모델의 환경요인 (선행음소, 후행음소)으로 이용하였다. 모든 HM-Net 음향모델은 4개의 혼합분포를 가지며 상태수가 200에서 1,200 상태까지는 200상태 단위로 증가시키면서 학습하였고, 상태수 2,000개인 HM-Net 모델도 학습하였다. 또한 비교를 위해 문맥독립 음향모델은 5상태 3출력분포의 혼합수 4개를 가지는 음향모델도 작성하였다.

본 논문에서 적용한 HM-Net 구조결정 알고리즘과 비교를 위해 HTK[20]에서도 단어인식 실험을 위해 동일한 음소 질의어 집합을 이용하여 상태공유에 의한 문맥의존 음향모델을 작성하였다.

### 4.3. 실험결과 및 고찰

결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘에 의해 작성한 문맥의존 음향모델의 유효성을 확인하기 위해 음소, 단어 및 연속음성인식 실험을 각각 수행하였다.

음성인식 방법은 음소 및 단어인식의 경우 One-Pass Viterbi 빔 탐색 알고리즘[14,15,21]으로서 음소인식의 경우 48개의 유사음소단위로 구성한 phone-pair 문법[14]을 사용하고, 단어인식의 경우 452단어로 구성한 word-pair 문법[14]을 사용하였다. 또한 연속음성인식의 경우 멀티 패스 탐색 알고리즘[22]으로서 1-pass 탐색의 경우, 단어 2-gram 언어모델을 이용하여 프레임 동기형 비터비 빔 탐색을 수행한 후 단어 그래프를 출력한다. 2-pass 탐색의 경우 1-pass의 단어 그래프와 보다 정밀한 단어 3-gram 언어모델을 이용하여 A* stack decoding 탐색을 수행한 후 그 결과에 대해 re-scoring에 의해 인식결과를 출력한다. 그림 4, 5, 6에 상태수의 변화에 따른 화자독립 음소인식률, 화자독립 단어인식률 그리고 화자독립 연속음성인식률을 각각 나타내었다.

그림 4의 화자독립 음소인식률에서 문맥독립 음향모델 (monophone)에 대해 국어공학센터 (KLE)의 남성화자 3인 평균 32.3%를 나타내고 있다. 그리고 문맥의존 음향모델인 HM-Net triphone에 대해서는 상태수 200일 때 평균 47.3%, 상태수 2,000일 때 평균 71.5%를 나타내고 있다. 문맥독립 음향모델과 상태수 2,000일 때의 문맥의존 음향모델과 음소인식률을 비교하면 HM-Net 구조결정 알고리즘에 의해 작성한 HM-Net triphone을 이용한 경우가 평균 39.2%의 인식률 향상을 보였다. 또한 상태수 200과 상태수 2,000일 때의 문맥의존 음향모델의 인식률을 비교하면 평균 24.2%의 인식률 향상을 보이고 있다. 그리고 문맥의존 음향모델의 경우 상태수를 점진적으로 증가시킨 경우 인식률의 향상이 두드러짐을 알 수 있다.

그림 5의 단어인식률에서도 문맥독립 음향모델인 경우 국어공학센터 (KLE)의 남성화자 3인 평균 92.1%를 나타내고, 문맥의존 음향모델인 HM-Net triphone의 상태수가 200일 때 평균 96.4%, 상태수가 2,000일 때 평균 99.2%의 단어인식률을 나타내고 있다. 그림 5에서도 문맥독립 음향모델과 HM-Net triphone을 비교하면 상태수 2,000일 때의 HM-Net triphone을 이용한 경우가 평균 7.1%의 인식률 향상을 보이고 있다. 또한 HM-Net 구조결정 알고리즘의 유효성을 검토하기 위해 단어인식 실험에서 사

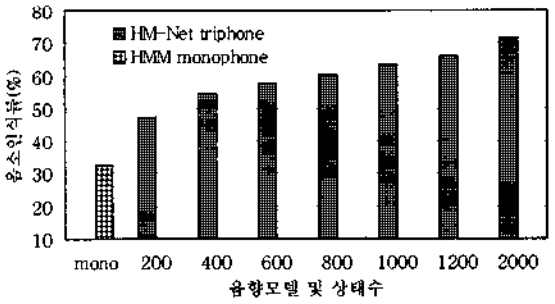


그림 4. 상태수에 따른 화자독립 음소인식률
Fig. 4. Speaker independent phoneme recognition accuracy according to the number of states variation.

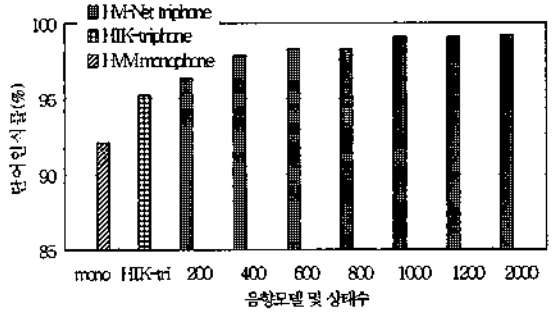


그림 5. 상태수에 따른 화자독립 단어인식률
Fig. 5. Speaker independent word recognition accuracy according to the number of states variation.

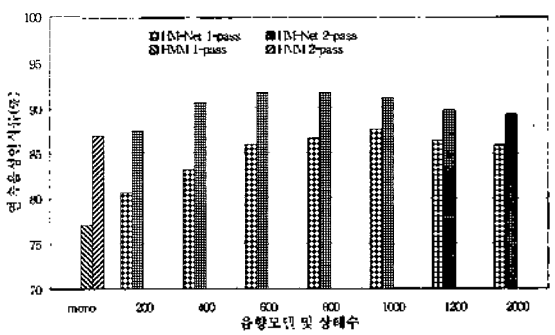


그림 6. 상태수에 따른 화자독립 연속음성인식률
Fig. 6. Speaker independent continuous speech recognition accuracy according to the number of states variation.

용한 동일한 학습 음성데이터와 음소 질의어 집합을 이용하여 HTK에 의해 상태공유 triphone 음향모델을 학습한 후 화자 3인에 대해 단어인식실험을 수행하였다. 그림 5에 나타낸 것과 같이 HM-Net 구조결정 알고리즘이 상태수를 변화시킨 각각의 HM-Net triphone 음향모델이 전체적으로 향상된 인식률을 보였으며, 특히 상태수가 2,000일 때의 음향모델은 HTK의 상태공유 triphone 음향모델보다 평균 4.0% 향상된 인식률을 얻었다.

그림 6의 연속음성인식 결과에서는 상태수 1,000일 때 HM-Net triphone 음향모델의 경우 1-pass 탐색의 인식률은 평균 87.5%로서 문맥독립 음향모델에 비해 평균 10.5%의 인식률을 향상을 보이고, 상태수 800일 때 HM-Net triphone 음향모델의 경우 2-pass 탐색의 인식률은 평균 91.6%로서 문맥독립 음향모델에 비해 평균 4.8%의 인식률을 향상을 보였다. 하지만 그림 6에서 HM-Net triphone 음향모델의 상태수가 1,000개 이상으로 증가함에 따라 인식률이 조금씩 감소하는 결과를 보이고 있는데 이는 학습에 참가한 음성 데이터의 부족으로 인해 HM-Net triphone 음향모델의 파라미터 추정이 제대로 수행되지 못해 정확한 HM-Net 음향모델이 생성되지 못한 것으로 생각된다. 이는 향후 음향모델을 작성하는데 많은 양의 음성 데이터를 사용할 경우 해결할 수 있을 것으로 기대된다. 이상의 결과로부터 본 논문에서 문맥의존 음향모델을 작성하기 위해 적용한 결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘의 유효성을 확인할 수 있었다.

## V. 결 론

본 논문에서는 한국어 음성인식에서 음향모델의 성능 개선을 위한 기초적 연구로서 결정트리 상태 클러스터링에 의한 HM-Net의 구조결정 알고리즘을 이용한 음성인식에 관한 연구를 수행하였다. 한국어는 다른 언어와 비교하여 많은 문법과 변이음이 존재하는데 국어 음성학에서 정의한 다양한 변이음을 조사하고, 음소결정트리를 위한 음소 질의어 집합을 작성하였다. 본 논문에서 적용한 HM-Net 구조결정 알고리즘은 SSS 알고리즘의 구조를 가지면서 미리 작성해 둔 문맥의존 음향모델의 상태를 다시 분할하는 방법으로서 모델의 각 상태위치마다 음소 질의어 집합에 의해 음소결정트리를 생성하고, PDT-SSS 알고리즘에 의해 문맥의존 음향모델을 학습하게 된다. 본 논문에서 적용한 알고리즘의 유효성을 확인하기 위해, 국어공학센터 (KLE)의 452단어와 항공편 예약에 관련된 YNU (Yeungnam University)200 문장을 대상으로 음성인식 실험을 수행한 결과 음소, 단어, 연속음성인식 실험에서 상태분할을 수행한 후 상태수의 변화에 따라 인식률이 점진적으로 향상됨을 확인하였다. 특히 상태수 2,000일 때 음소, 단어 인식률이 평균 71.5%, 99.2%를 각각 얻었으며, 연속음성인식률은 상태수 800일 때 평균 91.6%를 얻었다. 또한 HM-Net 구조결정 알고리즘과 비교를 위해 상태공유를 수행하는 HTK를 이용한 단어인식 실험을 수행한 결과, HTK를 이용한 경우보다 본 논문에서 적용한 방법이 평균 4.0%의 인식률 향상을 얻어, 결정트리 상태 클러스터링에 의한 HM-Net 구조결정 알고리즘의 유효성을 확인할 수 있었다.

## 감사의 글

본 연구는 한국과학재단 목적기초연구 (과제번호: R01-2000-00276)지원으로 수행되었음.

## 참 고 문 헌

1. S. J. Young, "A review of large vocabulary continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45-57, 1996.
2. P. Beyerlein, M. Ullrich, and P. Wilcox, "Modeling and decoding of cross-word context-dependent phones in the Philips large vocabulary continuous speech recognition system," *Proc. of Eurospeech'97*, pp. 1163-1166, 1997.
3. S. J. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech and Language* vol. 8, pp. 369-383, 1994.
4. M. Y. Hwang, X. Huang, and F. A. Alleva, "Predicting unseen triphones with senones," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 6, pp. 412-419, 1996.
5. S. J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *Proc. of ARPA Human Language Technology Workshop*, pp. 307-312, 1994.
6. L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision tree for phonological rules in continuous speech," *Proc. of ICASSP'91*, pp. 185-188, 1991.
7. S. Hayamizu, K. F Lee, and H. W. Hon, "Description of acoustic variations by tree-based phone modeling," *Proc. of ICSLP'90*, pp. 705-708, 1990.
8. K. F. Lee, H. W. Hon, C. Huang, J. Swartz, and R. Weide, "Allophone clustering for continuous speech recognition," *Proc. of ICASSP'90*, pp. 749-752, 1990.
9. S. Sagayama, and S. Honma, "Estimation of unknown context using a phoneme environment clustering algorithm," *Proc.*

*of ICSLP'90*, vol. 1, pp. 361-364, 1990.
10. J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. of ICASSP'92*, vol. 1, pp. 573-576, 1992.
11. M. Suzuki, S. Makino, A. Ito, H. Aso, and H. Shimodaira, "A new HMnet construction algorithm requiring no contextual factors," *IEICE Trans. Info. & Syst.*, vol. E78-D, no. 6, pp. 662-669, 1995.
12. M. Ostendolf, and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, pp. 17-41, 1997.
13. T. Hori, "*A study on large vocabulary continuous speech recognition*," Ph. D. Thesis, Yamagata University, Japan, 1999.
14. L. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc. 1993.
15. 中川聖一, "確率モデルによる音聲認識," 日本電子情報通信學會, 1988.
16. S. J. Oh, C. J. Hwang, B. K. Kim, H. Y. Chung, and A. Ito, "New state clustering of hidden Markov network with Korean phonological rules for speech recognition," *IEEE 4th workshop on Multimedia Signal Processing*, pp. 39-44, 2001.
17. 이호영, *국어음성학*, 태학사, chap. 3-5, 1996.
18. 배주채, *국어음운론*, 형설출판사, chap. 2-5, pp. 15-76, 1995.
19. 김상훈, 박준, 이영직, "골퍼스에 기반한 한국어 문장/음성변환 시스템," 한국음향학회지, 제20권 제3호, pp. 24-33, 2001.
20. S. Young, D. Kershaw, J. Odell, D. ollason, V. Valtchev, and P. Woodland, *The HTK Book*, 1997.
21. H. Ney, and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, pp. 64-83, 1999.
22. N. Deshmukh, A. Ganaparhiraju, and J. Picone, "Hierarchical Search for large vocabulary conversational speech recognition," *IEEE Signal Processing Magazine*, pp. 84-107, 9, 1999.
23. S. J. Oh, C. J. Hwang, B. K. Kim, H. Y. Jung, and H. Y. Chung, "A study on context-dependent acoustic modeling using the PDT-SSS Algorithm for Korean speech recognition," *Proc. of ICSP'01*, 2001.

## 저자 약력

**● 오 세 진 (Se-Jin Oh)**

1996년 2월: 영남대학교 전자공학과 (공학사)
1998년 2월: 영남대학교 대학원 전자공학과 (공학석사)
2002년 2월: 영남대학교 대학원 전자공학과 (공학박사)
2001년 9월~현재: 대구과학대학 디지털정보통신계열 전임강사
※ 주관심분야: 음성분석 및 인식, 언어처리

**● 황 철 준 (Chul-Joon Hwang)**

1996년 2월: 영남대학교 전자공학과 (공학사)
1998년 2월: 영남대학교 대학원 전자공학과 (공학석사)
2002년 2월: 영남대학교 대학원 전자공학과 (공학박사)
2000년 3월~현재: 대구과학대학 디지털정보통신계열 전임강사
※ 주관심분야: 음성분석 및 인식, 디지털 신호처리

**● 김 범 국 (Bum-Koog Kim)**

1990년 2월: 영남대학교 수학과 (이학사)
1992년 2월: 영남대학교 대학원 전자공학과 (공학석사)
1998년 2월: 영남대학교 대학원 전자공학과 (공학박사)
1997년 3월~현재: 대구과학대학 디지털정보통신계열 조교수
※ 주관심분야: 음성분석 및 인식, 언어처리, 멀티모달 시스템

**● 정 호 열 (Ho-Youl Jung)**

1988년 2월: 아주대학교 전자공학과 (공학사)
1990년 2월: 아주대학교 전자공학과 (공학석사)
1993년 2월: 아주대학교 전자공학과 (박사수료)
1998년: (프)리옹국립응용과학원 (INSA de Lyon) 전자공학전공 (공학박사)
1998년 4월~1998년 12월: (프)CREATIS 박사후 과정
1999년 3월~현재: 영남대학교 전자정보공학부 조교수
※ 주관심분야: 음성·영상 신호처리, 인공지능, 디지털 워터마킹 등

**● 정 현 열 (Hyun-Yeol Chung)**

현재: 영남대학교 전자정보공학부 교수
한국음향학회지 제20권 제7호 참조