# 한국어 대어휘 음성DB를 이용한 HM-Net 음성인식 시스템의 성능평가

오 세 진, 김 광 동, 노 덕 규, 송 민 규, 김 범 국*, 황 철 준*, 정 현 열**
한국천문연구원, *대구과학대학, **영남대학교
전화 : 042-865-3280 / 핸드폰 : 011-543-0971

# Performance Evaluation of HM-Net Speech Recognition System using Korea Large Vocabulary Speech DB

Se-Jin Oh, Kwang-Dong Kim, Duk-Gyoo Roh, Min-Gyoo Song, Bum-Koog Kim*,
Chul-Jun Hwang*, Hyun-Yeol Chung**
Korea Astronomy Observatory, *Taegu Sicence College, **Yeungnam University
E-mail : sjoh@trao.re.kr

## Abstract

본 논문에서는 한국전자통신연구원에서 제공된 대어휘 음성DB를 이용하여 HM-Net(Hidden Markov Network) 음성인식 시스템의 성능평가를 수행하였다. 음향모델 작성은 음성인식에서 널리 사용되고 있는 통계적인 모델링 방법인 HMM(Hidden Markov Model)을 개량한 HM-Net을 도입하였다. HM-Net은 PDT-SSS 알고리즘에 의해 문맥방향과 시간방향의 상태분할을 수행하여 생성되는데, 특히 문맥방향 상태분할의 경우 학습 음성데이터에 출현하지 않는 문맥정보를 효과적으로 표현하기 위해 음소결정트리를 채용하고 있으며, 시간방향 상태분할의 경우 학습 음성데이터에서 각 음소별 지속시간 정보를 효과적으로 표현하기 위한 상태분할을 수행한다. 이러한 상태분할을 수행하여 파라미터를 공유하게 되며 최적인 모델 네트워크를 작성하게 된다. 대어휘 음성데이터를 이용하여 음향모델을 작성하고 인식실험을 수행한 결과, 100명의 100단어와 60문장에 대해 평균 97.5%, 96.7%의 인식률을 보였다.

## I. 서론

본 논문은 강건한 음성인식 시스템을 개발하기 위한 기초적인 연구로서 HMM[7]의 개량형인 HM-Net[1]을 도입하여 대어휘 음성데이터에 대해 성능평가를 수행하고자 한다. HM-Net 모델의 작성은 음소결정트리와 자동상태분할을 채용한 PDT-SSS(Phonetic Decision Tree-based Successive State Splitting) 알고리즘[4][5][6]을 이용한다. 이를 위하여 우선 소규모의 음성데이터에 대한 HM-Net 인식 시스템의 성능을 고찰하기 위해 국어공학센터(KLE)에 제공된 452단어를 대상으로 HM-Net 문맥의존 음향모델의 작성과 인식성능을 고찰한 후, 한국전자통신연구원(ETRI)에서 대어휘 음성데이터에 대해서도 고찰하고자 한다. 대어휘에 대한 성능평가 실험의 경우 HM-Net 모델 작성에는 ETRI에서 제공된 1000명(남녀 각 500명)의 음성 DB에서 400명(남녀 각 200명)의 음성 데이터를 이용하였다. 작성한 HM-Net 음향모델의 성능평가는 학습에 참가하지 않은 나머지 100명(남녀 각 50명)이 발성한 100단어와 60문장을 대상으로 One-Pass Viterbi 빔 탐색 알고리즘[7]을 이용하여 음소 및 단어, 연속음성 인식실험을 수행한 후 결과를 고찰하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 HM-Net 음성인식 시스템과 주요 알고리즘에 대해 설명하고, 3장에서는 시스템 평가에 사용된 음성데이터에 대해 간략히 설명한다. 4장에서는 인식 시스템을 이용한 인식실험 및 결과에 대해 고찰한 후, 마지막으로 5장에서 결론을 맺는다.

## II. HM-Net 음성인식 시스템

그림 1에 HM-Net 인식 시스템의 전체 구성을 나타내었다. 이하에 인식 시스템의 핵심인 HM-Net과 주요 알고리즘에 대해 간략히 설명한다.

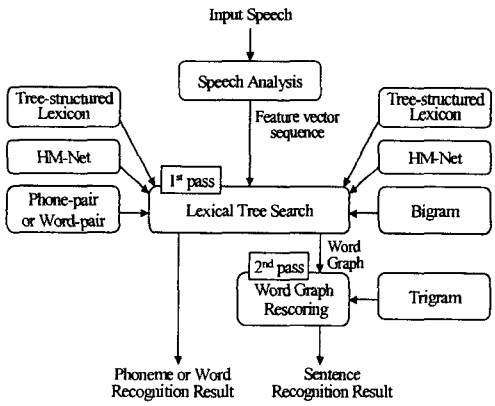


그림 1. 음성인식 시스템의 전체 구성도

### 2.1 Hidden Markov Network

SSS 알고리즘[1]에 의해 작성한 HM-Net은 여러 개의 상태를 연결한 네트워크로 표현되며, HM-Net의 각 상태는 상태번호, 허용할 수 있는 문맥 클래스, 선행음소와 후행음소 리스트, 자기천이확률과 후행상태로의 천이확률 그리고 출력확률분포 파라미터와 같은 정보를 포함한다. HM-Net에서는 문맥정보가 주어지면, 이 문맥을 허용할 수 있는 상태를 선행상태와 후행상태 리스트의 제약 내에서 연결하여 이 문맥에 대한 모델을 결정할 수 있다. 이 모델은 자기천이와 다음 상태로의 천이만을 고려한 left-to-right 모델로 간주할 수 있으므로 일반적인 HMM과 같이 Baum-Welch 알고리즘[7]에 의해 파라미터를 추정할 수 있다.

### 2.2 SSS 알고리즘

SSS 알고리즘[1]은 모든 문맥을 나타내는 1상태의 초기모델로부터 문맥방향과 시간방향으로 상태분할 후 자동적으로 HM-Net의 구조를 결정하는 알고리즘이다.

전체적으로 간략히 설명하면 다음과 같다. 우선 유사음소단위(PLUs)를 기본단위로 모든 모델을 연결한 네트워크 구조의 초기모델로서 각각의 모델은 하나의 상태와 그 상태를 시단에서 종단까지 결합하여 전체 학습 데이터로부터 작성한다. 상태의 분할은 경로분할을 동반하는 문맥방향과 경로분할을 동반하지 않는 시간방향이 있는데, 출력확률의 우도에 따라 한 방향으로만 수행된다. 문맥방향으로 분할할 때는 경로분할에 동반된 각각의 경로에 할당된 문맥 클래스도 동시에 분할된다. 따라서 문맥 클래스의 분할에 포함된 모든 상태 중에서 학습 데이터에 대한 누적우도 확률이 가장 큰 쪽의 상태를 분할하도록 선택된다. 시간방향으로의 상태분할에서도 누적우도 확률이 높은 쪽 상태를 분할하도록 선택된다. 이상의 상태분할을 반복하여 HM-Net의 구조가 결정된다.

### 2.3. PDT-SSS 알고리즘

PDT-SSS[4]는 SSS 알고리즘의 문맥방향 상태분할에 음소결정트리를 결합한 것으로 HM-Net에서 새로운 상태의 모델 파라미터 공유와 학습데이터에 출현하지 않는 미지의 문맥에 대한 학습을 수행할 수 있도록 구성되어 있다. 여기서 음소결정트리[6]는 2진 트리로서 각 노드는 음소질의어로 구성되어 있다. 각 음소모델의 공유 파라미터는 각 트리의 잎(leaf) 노드와 연관되고, 문맥의존 모델은 음소 질의어에 의해 트리의 뿌리(root) 노드에서 잎 노드까지 조사하여 임의의 문맥에 할당되어진다. PDT-SSS의 특징은 허용할 수 있는 문맥 클래스는 음소 질의어에 따른 결정트리에 의해 분할된다는 것이다. 또한, 하나의 상태가 분할될 때, 두 개의 혼합수는 새로운 상태와 관련된 것이 아니고 새로운 상태에 대한 단일 가우스 분포는 학습 샘플로부터 계산된다. 따라서, PDT-SSS 알고리즘이 적절한 문맥 클래스의 분할과 임의의 문맥을 표현할 수 있기 때문에 보다 정확한 HM-Net을 작성할 수 있게 된다. PDT-SSS 알고리즘의 주요 내용은 다음과 같다.

1) 한국어 음성학적 지식에 의한 음소 질의어 집합을 작성한다.
2) Baum-Welch 알고리즘으로 초기 HM-Net을 학습한다.(각 상태는 단일 가우스 분포)
3) SSS 알고리즘과 같이 식(1)에 의해 최적 분포를 가지는 상태를 선택한다.
4) 문맥방향과 시간방향으로 분할할 상태를 선택한다.
   - 각 음소 질의어에 대해 문맥방향으로 분할할 때,
     i) 질의어에 대해 허용할 수 있는 문맥 클래스의 분할과 두 개의 단일 가우스 분포를 추정한다.(각 가우스 분포는 yes 또는 no에 해당)
     ii) 새로운 상태에 각 문맥 클래스와 각 가우스 분포를 할당한다.
   - 각 음소 질의어에 대해 시간방향으로 분할할 때,
     i) Baum-Welch 재추정에 의해 두 개의 단일 가우스 분포를 추정한다.
     ii) 새로운 상태에 각 가우스 분포를 할당하고 문맥 클래스를 복사한다.
5) 학습 샘플의 우도에 근거하여 문맥방향과 시간방향에서 최적의 HM-Net을 선택한다.
6) Baum-Welch 알고리즘에 의해 HM-Nets의 상태를 재학습한다.
7) 미리 정의한 상태수에 도달할 때까지 단계 3부터 반복한다.

단계 3에서 분할될 상태의 선택은 식(1)에 의해 계산되어진다.

$$d_i = n_i \sum_{p=1}^{P} \frac{\sigma_{ip}^2}{\sigma_{Tp}^2} \qquad (1)$$

여기서, $\sigma_{ip}^2, \sigma_{Tp}^2$는 상태 $i$의 분포 분산과 모든 샘플의 분산(정규화 계수)을 나타내고, $n_i$ 는 상태 $i$의 추정에 이용한 음소 샘플의 수를, $P$는 특징 벡터의 차원 수를 각각 나타낸다.

## III. 음성데이터 및 분석조건

KLE에서 제공된 452단어는 방음부스에서 채록되었으며, PBW(Phoneme Balanced Word)로 구성되어 있다. 발성화자는 남성 38명과 여성 32명이 각각 2회씩 발성된 것으로 구성되어 있다.

ETRI에서 제공된 대어휘 음성DB는 성별, 연령별, 지역별로 분포된 1000명의 화자로 구성되어 있다. 남녀 화자의 성별은 50:50이며, 최소 SNR이 25dB이상인 조용한 사무실환경에서 수집되었다. 녹음장치는 일반 컴퓨터의 PC 마이크, PC 헤드셋, VoIP 통신망이 사용되었으며, VoIP 통신망은 PC 헤드셋으로 수집한 DB를 파일로 통신망에 전송하여 수집되었다. 총 발성내용은 10,000 단어, 10,000 숫자음, 100,000 문장으로 구성되어 있다. 단어는 주식회사명, 지명, 인명, 상호명, 제품명, PC 명령어, PDA 명령어, 그 외 일반명사로 구성되어 있고, 숫자의 경우 번호독식(connected digit) 방식과 봉독식(natural number) 방식에 대해 수집

되었다. 문장의 발성목록은 방송뉴스 대본에서 추출하였으며, 낭독체 50,000문장, 준낭독체 50,000문장으로 구성되어 있다.

본 논문에서는 KLE의 경우 남성 38명과 여성 32명의 1회 발성을 사용하였으며, ETRI의 경우 PC 헤드셋 마이크로 채록한 500명의 발성을 학습과 평가에 각각 사용하였다.

모든 음성 데이터는 16kHz, 16bits로 샘플링과 프리엠퍼시스 필터를 통과한 후 25ms의 해밍 윈도우를 곱하여 10ms씩 이동하면서 분석하였다. 이를 통해 음성 특징 파라미터는 12차 LPC-멜 켑스트럼 계수[7]와 정규화된 대수 에너지에 1차 및 2차의 차분 성분을 포함하여 총 26차와 39차의 특징 파라미터를 구하였다.

## IV. 인식실험 및 고찰

KLE에서 제공된 단어음성 데이터와 ETRI에서 제공된 대어휘 음성DB를 이용하여 HM-Net 음성인식 시스템의 유효성을 확인하기 위해 음소 및 단어, 연속음성 인식실험을 각각 수행하였다.

### 4.1 음향모델 작성

KLE에서 제공한 452단어에 대한 기본적인 실험의 음향모델 작성은 남성 35명이 1회 발성한 452단어를 이용하였으며, 나머지 3명은 평가에 사용하였다. PDT-SSS 알고리즘의 문맥방향 상태분할을 위해 162개(문맥의 좌, 우)의 음소 질의어를 한국어 음성학적 지식에 근거하여 작성하였다. 초기 HM-Net의 구조는 48개의 유사음소단위를 병렬로 연결하여 141개의 상태를 가지도록 구성하였다. HM-Net 모델은 26차와 39차의 특징 파라미터를 사용하여 혼합수 4를 가지며 200에서 1,200상태까지는 200상태씩 증가시켰으며, 상태수 2,000인 HM-Net도 학습하였다.

ETRI에서 제공한 대어휘 음성데이터를 이용한 음향모델 작성은 남녀 각 200명이 발성한 280발성을 이용하였으며, 나머지 남녀 각 25명의 100단어를 평가에 사용하였다. 남녀 각 200명이 발성한 데이터에서 학습에 사용된 총 단어는 5,287단어이고 평가에 사용된 총 단어 중 학습에 포함된 단어 수는 607단어로서 약 11.5%의 단어가 중복 사용되었다. ETRI 데이터의 경우 상태분할을 위해 152개의 음소 질의어를 준비하였으며, 초기 HM-Net의 구조는 43개의 유사음소단위를 병렬로 연결하여 126개의 상태를 가지도록 구성하였다. HM-Net 모델은 39차의 특징 파라미터를 사용하여 혼합수 1, 2, 4, 6, 8개를 가는 상태수 1000, 1500, 2000인 모델을 각각 학습하였다.

### 4.2 KLE 452 음소/단어인식 실험

KLE 단어음성 데이터에 대해 HM-Net 음성인식 시스템 기본적인 성능평가를 위해 남성 3명이 발성한 452단어의 첫 번째 발성에 대해 phone-pair/word-pair 문법을 가진 One-Pass Viterbi 빔 탐색[7]을 이용하여 음소 및 단어인식실험을 수행하였다.

표 1. HM-Net 상태수 및 차원별 음소/단어인식률

| 차원 | 실험 | HM-Net 상태수 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 200 | 400 | 600 | 800 | 1000 | 1200 | 2000 |
| 26차 | 음소 | 46.7 | 53.6 | 56.0 | 57.6 | 59.8 | 61.8 | 68.6 |
| | 단어 | 96.2 | 97.9 | 98.8 | 99.0 | 99.2 | 98.9 | 99.6 |
| 39차 | 음소 | 50.0 | 57.7 | 60.7 | 63.2 | 66.1 | 68.6 | 75.2 |
| | 단어 | 97.2 | 98.4 | 98.5 | 98.7 | 99.0 | 99.0 | 99.2 |

표 1에 나타낸 것과 같이 음소/단어인식 모두 각 차원에 대해 상태수가 증가할수록 인식률이 향상되는 것을 볼 수 있다. 음소인식의 경우 26차원에 비해 39차원의 경우가 상태수 2000개인 모델에서 평균 6.6% 향상되었으나, 단어인식의 경우 상태수 2000개인 모델에서 두 차원에 대해 높은 인식률을 보이지만 39차원이 26차원에 비해 평균 0.4% 낮은 결과를 보였다. 이는 학습 데이터가 적은 경우 26차원이 보다 효과적인 것으로 생각된다. 즉, 적은 학습 데이터에 대해 상태수를 계속해서 증가시킬 경우 학습 데이터의 부족으로 인해 모델 학습이 제대로 수행되지 않은 원인도 된다.

### 4.3 ETRI/KLE 데이터에 대한 인식 실험

ETRI 음성 데이터로 작성한 각 HM-Net 모델에 대해 2099개의 인식단어 카테고리 내에서 ETRI의 남녀 각 25명이 발성한 100단어와 KLE의 남성 35명과 여성 32명의 452단어를 대상으로 4.2절과 동일한 인식방법으로 태스크종속 및 태스크 독립 단어인식 실험을 수행하였다. 또한 ETRI의 남녀 각 25명이 발성한 60문장에 대해 단어 bigram 문법[7]과 동일한 인식방법에 대해 연속음성인식 실험을 수행하였으며, 그 결과를 그림 2, 3, 4와 표 2에 각각 나타내었다.

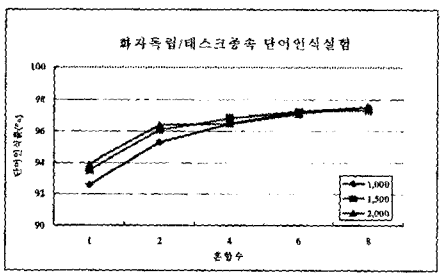


그림 2. 태스크 종속 ETRI 남녀 각 25명의 단어인식률

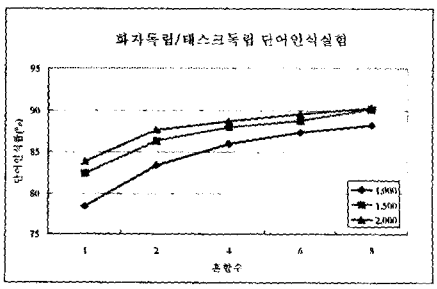


그림 3. 태스크 독립 KLE 452 남성 35명의 단어인식률

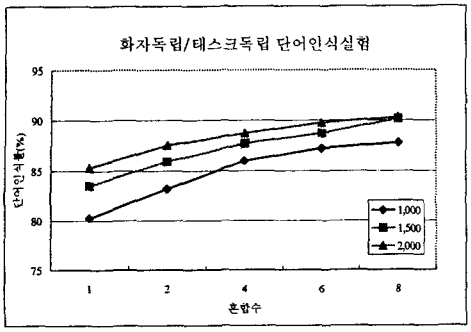


그림 4. 태스크 독립 KLE 452 여성 32명의 단어인식률

표 2. 남녀 각 25명의 60문장 인식률

| 상태 | 실험 | 혼 합 수(mixture) | |
|---|---|---|---|
| | | 4 | 8 |
| 1000 | 문장 | 95.5 | 96.7 |
| | 단어 | 97.7 | 98.3 |
| 1500 | 문장 | 95.5 | 96.7 |
| | 단어 | 97.6 | 98.3 |
| 2000 | 문장 | 94.9 | 96.6 |
| | 단어 | 97.0 | 98.2 |
| mono | 문장 | 94.5 | |
| | 단어 | 97.3 | |

그림 2의 태스크 종속 단어인식률의 경우, 학습 데이터의 양에 비해 모델의 상태분할 및 학습이 충분하지 않지만, 평균 97%를 상회하는 결과를 보이고 있다. 하지만 태스크 독립 단어인식의 경우 태스크 종속에 비해 인식률이 저조한데, 이는 HM-Net 모델을 학습할 때 많은 문맥환경과 발성리스트를 공유할 경우 높은 인식성능을 보이나 그렇지 못한 경우에는 다소 인식성능이 저하되는 것으로 생각된다. 즉, 태스크 독립실험의 경우, 모델의 상태분할과 학습에 사용한 음성 데이터의 문맥정보와 인식에 사용된 음성 데이터의 문맥정보가 서로 상이한 것과 포함되지 않은 것이 많이 있어 예상보다 저조한 인식률을 보이는 것으로 생각된다. 또한 사용된 음성 데이터의 문맥정보를 분석해 보면, ETRI 음성 데이터의 triphone 수는 16,380개로 문맥정보는 많지만, KLE 452단어(triphone 수: 2,164개)에서의 triphone을 만족할 만큼 충분히 표현하지 못하는 것으로 생각된다. 그리고 ETRI 데이터의 경우 데이터 양은 많지만 다양한 문맥정보(음소)가 KLE의 PBW와 비교하여 균형있게 분포하지 못하고, 문맥정보에 대한 음성 데이터가 적은 것도 인식률 저하의 원인으로 생각된다.

표 2의 연속음성인식 실험결과의 경우 태스크 종속으로 비교적 높은 인식률을 보이고 있다. 하지만 상태수와 혼합수가 증가함에도 인식률이 조금 저하되는 경향이 있는데 이는 앞에서 설명한 것과 같이 많은 문맥정보(음소)에 대한 음성 데이터의 양이 부족한 것에 기인한 것으로 생각된다. 이는 향후 다양한 문맥정보(음소)에 대해 많은 음성 데이터를 사용할 경우 해결될 것으로 생각된다.

## V. 결론

본 논문에서는 한국전자통신연구원에서 제공된 대어휘 음성DB를 이용하여 HM-Net(Hidden Markov Network) 음성인식의 음향모델 작성에 널리 사용되고 있는 통계적인 모델링 방법인 HMM(Hidden Markov Model)을 개량한 HM-Net을 도입하였다. HM-Net은 PDT-SSS 알고리즘에 의해 문맥방향과 시간방향의 상태분할을 수행하여 생성되는데, 특히 문맥방향 상태분할의 경우 학습 음성데이터에 출현하지 않는 문맥정보를 효과적으로 표현하기 위해 음소결정트리를 채용하고 있으며, 시간방향 상태분할의 경우 학습 음성데이터에서 각 음소별 지속시간 정보를 효과적으로 표현하기 위한 상태분할을 수행한다. 이러한 상태분할을 수행하여 파라미터를 공유하게 되며 최적인 모델 네트워크를 작성하게 된다. 대어휘 음성데이터를 이용하여 음향모델을 작성하고 인식실험을 수행한 결과, 100명의 100단어와 60문장에 대해 평균 97.5%, 96.7%의 인식률을 보였다.

※본 논문에서 사용된 음성 데이터는 한국전자통신연구원과 국어공학센터에서 제공되었습니다.

## 참고문헌

[1] J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. of ICASSP'92*, Vol. 1, pp. 573-576, 1992.

[2] 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.

[3] M. Ostendoft, and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, Vol. 11, pp. 17-41, 1997.

[4] T. Hori, *"A study on large vocabulary continuous speech recognition,"* Ph. D. thesis, Yamagata University, Japan, 1999.

[5] Se-Jin Oh, Cheol-Jun Hwang, Bum-Koog Kim, Hyun-Yeol Chung, and Akinori 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.

[6] 오세진, 황철준, 김범국, 정호열, 정현열, "결정트리 상태 클러스트링에 의한 HM-Net 구조결정 알고리즘을 이용한 음성인식에 관한 연구," *한국음향학회지*, 제21권 제2호, pp. 199-210, 2002.

[7] 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.