# 음성의 시간변이와 상태분할을 고려한 강건한 문맥의존 음향모델에 관한 연구

오세진°, 김광동*, 노덕규* 정현열**
한국천문연구원°*, 영남대학교**
{sjoh°, kdkim, dgroh}@trao.re.kr, hychung@yu.ac.kr

# A study on the robust context-dependent acoustic models by considering the state splitting and the time variant of speech

Sejin Oh° Kwangdong Kim* Duk-Gyoo Roh* Hyunyeol Chung**
Korea Astronomy Observatory°*, Yeungnam University**

**요 약**

일반적으로 음성은 시간함수로 표현되며 음성인식에서 표준모델을 모델링하는 것은 매우 중요한 문제이다. 음절, 단어, 연속음성을 발성할 때 자음과 모음에 따라 발성시간에 차이가 있으며 이를 잘 모델링하는 것 또한 음성인식에서는 중요한 문제라고 할 수 있다. 따라서 본 연구에서는 강건한 음향모델을 학습하기 위해 시간의 변화와 상태분할과정에서의 모델의 변화를 고려하여 다양한 구조의 초기모델을 작성하였다. 각 초기모델에 의한 HM-Net 문맥의존 음향모델은 음소결정트리 기반 SSS 알고리즘(PDT-SSS)을 이용하였다. PDT-SSS 알고리즘은 미지의 문맥정보를 해결하기 위해 문맥방향과 시간방향으로 목표 상태수에 도달할 때까지 상태분할을 수행하여 모델을 작성하는 방법이다. 음성의 시간변이를 고려한 강건한 문맥의존 음향모델을 작성하기 위해 설정한 각 모델의 구조에 대한 유효성을 확인하기 위해 국어공학센터의 452 단어를 대상으로 음소와 단어인식 실험을 수행한 결과, 음소인식의 경우 상태수 2000개에서 2상태 구조의 모델에 비해 4상태 구조가 약 11.4% 향상된 인식성능과 39.2초의 인식시간을 단축할 수 있었다. 또한 단어인식의 경우 상태수 2000개에서 1상태 구조의 모델에 비해 4상태 구조가 약 5% 향상된 인식성능과 4상태 구조에서 한 단어를 인식하는데 평균 0.8초가 소요되었다. 따라서 강건한 문맥의존 음향모델을 작성하기 위해 수행한 초기모델의 구조에 관한 연구가 향후 음성인식 시스템을 구축하는데 유효함을 확인할 수 있었다.

## 1. 서 론

음성인식에서 초기모델의 구조를 선택하는 것은 인식 시스템의 성능을 결정하는데 매우 중요한 부분이다. 현재 음성인식에서는 확장성과 실용성을 고려하여 다양한 모델을 이용하고 있는데 그 중 가장 많이 사용하는 단위는 유사음소단위(Phoneme Likely Units; PLUs)이다[1]. 이와 함께 음성인식에서 널리 사용되고 있는 HMM(Hidden Markov Model)의 확률분포에는 이산분포형과 연속분포형이 있다[1]. 과거의 음성인식에서는 이산분포형을 많이 사용하였으나 모델을 학습하는데 많은 시간이 소요되어 현재는 연속분포형을 많이 사용하고 있다. 그리고 이들의 장점만을 고려한 semi-연속분포형을 이용하기도 한다[1][2]. 또한 다양한 문맥정보와 상태수를 고려하여 모델의 구조를 자동으로 결정할 수 있으며 HMM을 개량한 HM-Net(Hidden Markov Network)이 사용되고 있다[6].

유사음소단위를 이용한 모델의 구조는 가장 단순한 형태는 모델학습에 이를 그대로 하나의 모델을 구성하는 방법이 있으며 이를 monophone이라고 한다. 그리고 음소의 문맥정보를 고려하여 하나의 음소를 기준으로 앞/뒤 음소 하나에만 영향을 고려한 diphone을 생각할 수 있다. 또한 하나의 음소를 기준으로 앞과 뒤 모두를 고려한 triphone을 생각할 수 있다. 현재 대부분의 대어휘 연속음성인식에서는 monophone보다 다양한 문맥정보를 고려한 triphone 음향모델을 많이 사용하고 있다[1][3][4][5].

본 연구에서는 강건한 문맥의존 음향모델을 작성하기 위한 기초적인 연구로서 음성의 시간변이와 상태분할을 고려하여 HM-Net의 초기모델 구조와 인식성능의 관계에 관한 연구를 수행하였다. 다양한 형태의 초기모델 구조에 대해 국어공학센터(KLE)의 남성 38명이 1회 발성한 452단어를 이용하여 단어인식 실험을 수행하였다.

## 2. 상태분할 알고리즘

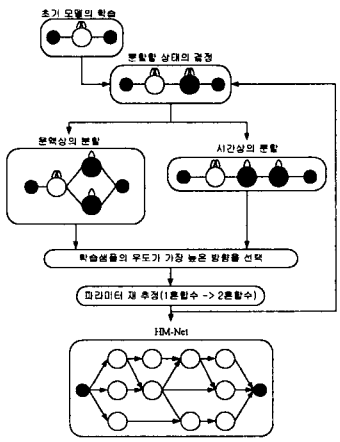



그림 1. SSS 알고리즘의 구성도

본 연구에서는 한국어 음성학적 지식의 음소질의어에 의한 음소결정트리와 SSS 알고리즘의 장점을 결합한 PDT-SSS (Phonetic Decision Tree-based SSS) 알고리즘[6][7][8]을 이용하였다. PDT-SSS 알고리즘은 SSS 알고리즘의 문맥방향 상태분할에 음소결정트리를 결합한 것으로 HM-Net에서 새로운 상태의 모델 파라미터 공유와 학습 데이터에 출현하지 않는 미지의 문맥에 대한 학습을 수행할 수 있도록 구성되어 있다.

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 설정한 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.

## 3. 모델 구조 설정

본 연구에서는 강건한 문맥의존 음향모델을 작성하기 위해 음성의 시간변화에 따른 다양한 형태의 모델구조를 설정하였다. 음성인식에서 널리 사용되고 있는 HMM의 상태는 음성의 시간에 따른 구조를 적절히 나타내기 때문에 널리 사용된다. 하지만 HMM에서 모델의 구조는 monophone의 경우 HMM의 상태를 음성의 길이 따라 다양한 구조를 설정하는 것이 아니고 특정한 상태수로 고정하여 사용한다. 또한 triphone의 경우에도 일정한 상태수를 가진 모델을 학습한 후 다양한 문맥환경에 따라 data-driven 방법과 결정트리 방법을 이용하여 모델을 학습한다. 하지만 일반적으로 사용되는 모델의 구조는 초기모델에 따라 결정되며 음성의 시간변화는 적절하게 고려하지 못한다. 하지만 하나의 음성은 일반적으로 자음과 모음의 조합에 의해 발성되는데 이는 발성에 따라 서로 다른 시간변화를 가진다. 따라서 음성의 시간변화를 적절히 고려하기 위해서는 음소(자음과 모음)에 따라 다양한 상태수를 가져야 한다. 따라서 이를 위해 본 연구에서는 다양한 문맥정보와 시간변이를 잘 표현할 수 있으며 모델의 구조를 자동적으로 결정하는 상태분할 알고리즘을 이용하였다.

그림 2에 음성의 시간변이에 따른 모델구조의 설정을 나타내었다. 그림 2의 (a)는 1 상태, (b)는 2 상태, (c)는 3 상태, (d)는 4 상태, 마지막으로 (e)는 5 상태에 음성의 자음 또는 모음이 지속할 수 있는 구조를 나타낸다. 이렇게 설정한 각 모델에 대해 음성의 시간변이와 문맥정보를 고려하여 PDT-SSS 알고리즘의 상태분할을 수행한 후 자음과 모음의 시간에 대해 적절한 음소의 길이를 가진 모델구조를 자동으로 결정하게 된다.

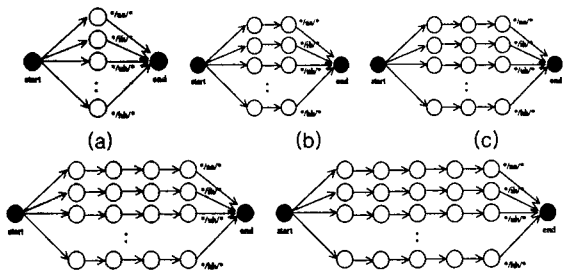



그림 2. 음성의 시간 변이에 따른 모델 구조 설정

## 4. 인식실험 및 결과

표 1. 음성데이터 및 분석조건

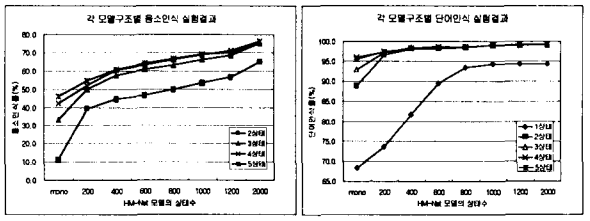
| 실 험 | 음소인식 | 단어인식 |
|---|---|---|
| 데이터 | KLE 452 단어 | |
| 학 습 | 35명 1회 발성 | |
| 평 가 | 3명 1회 발성 | |
| 인식형태 | 태스크 종속 화자독립 | |
| 주파수/분해능 | 16kHz/16bits | |
| 프레임 길이 | 25ms | |
| 프레임 주기 | 10ms | |
| 분석창 | Hamming Windows | |
| 특징 파라미터 | 12차 LPC-MEL cepstrum + delta power + 1, 2차의 회귀계수 = 39차원 | |

본 연구에서는 음성의 시간변이와 상태분할을 고려하여 강건한 문맥의존 음향모델의 작성하기 위해 다양한 구조의 초기모델을 설정하였다. 또한 초기설정 모델이 상태분할에 의해 어떤 형태로 구조변경이 되는가에 대해서도 조사하였다. 이렇게 설정한 구조의 인식성능에 대한 변화를 관찰하기 위해 국어공학센터(KLE)에서 채록한 452 단어를 대상으로 음소와 단어인식 실험을 수행하였다. 표 1에 사용한 음성데이터와 분석조건을 나타내었다. 모델의 학습에는 국어공학센터의 남성 35명이 1회 발성한 452단어 총 15,820단어를 사용하였으며, 평가에는 나머지 3명이 1회 발성한 452단어를 사용하였다. 모델의 학습은 HMM의 개량형으로 문맥의존 음향모델 학습시에 출현하지 않는 미지의 문맥에 대해서도 효과적으로 모델을 학습할 수 있는 PDT-SSS 알고리즘을 이용하였다. PDT-SSS 알고리즘에서 미지의 문맥을 고려하여 문맥방향의 상태분할에서 사용된 음소결

정트리의 질의어는 한국어 문법에서 조사한 162개(좌81,우81)를 사용하였다[8]. 모든 음성데이터는 16kHz, 16bits로 양자화하고 25ms의 Hamming Window를 곱해 10ms 단위로 이동하면서 LPC-MEL 분석[1]을 통해 39차의 특징 파라미터를 추출하여 사용하였다.

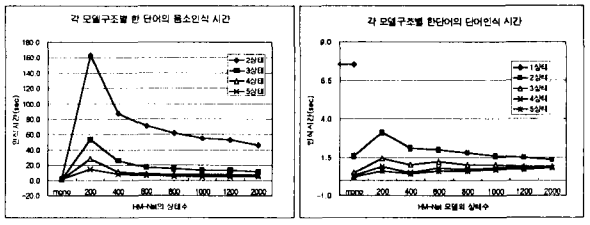
음성인식 실험은 다음과 같이 5가지의 종류의 초기모델을 학습한 모델을 이용하여 수행하였다. 5가지의 초기모델의 구조는 left-to-right HMM 구조[1]를 가지며 음성을 발성할 때 각 음소가 HMM의 1, 2, 3, 4, 5개의 상태에 각각 지속한다고 가정하였다. 이렇게 설정한 각 초기모델은 PDT-SSS 알고리즘에 의해 임의로 설정한 상태수가 도달할 때까지 각 음소의 길이에 적절한 문맥방향과 시간방향으로 상태분할을 자동적으로 수행한 후 최종 인식에 사용될 문맥의존 음향모델을 작성하였다.

음성인식 알고리즘은 Phone-pair/Word-pair 문법으로 작성한 언어모델과 트리구조의 단어사전에 의한 One-Pass Viterbi 탐색 알고리즘을 이용하였다[1]. 음소인식과 단어인식실험 결과를 그림 3의 (a), (b)에 각각 나타내었다. 또한 그림 4의 (a), (b)에는 각 정의한 모델의 구조에 따른 인식시간을 Sun Enterprise 450(400MHz dual CPU, 1Gbyte Memory)에서 측정한 결과를 나타내었다. 여기서 인식시간은 음성의 발성이 끝난 후 한 단어의 인식시간만 계산한 것을 나타낸다.



(a) 음소인식 (b) 단어인식

그림 3. 음소인식과 단어인식 실험결과



(a) 음소인식 시간 (b) 단어인식 시간

그림 4. 인식시간

그림 3의 음소인식과 단어인식 실험결과에서 모두 단일음소(monophone)보다는 문맥의존 음향모델(HM-Net)의 상태수가 증가함에 따라 인식성능이 향상되는 것을 볼 수 있다. 또한 그림 4의 인식시간에서도 음소인식과 단어인식 실험 모두에서 상태수가 증가할수록 인식시간이 감소하는 것을 볼 수 있다. 특히, 그림 3의 음소인식 실험결과에서 음성의 시간변이를 고려한 모델의 구조에 따라 인식성능이 변화하는 것을 확인할 수 있다. 마찬가지로 단어인식 실험결과에서도 음소인식 실험과 유사한 결과를 보임을 알 수 있다. 그림 3의 실험결과에서 초기모델의 상태수가 4개인 모델의 구조가 음소인식실험의 경우 상태수 2000개에서 평균 76.4%의 인식성능 보였으며 단어인식 실험의 경우 상태수 2000개에서 평균 99.3%의 인식성능을 보였다. 또한 단어인식의 경우 인식이 1초 이내에 이루어짐을 확인할 수 있었다.

전체적으로 위의 실험결과로부터 음성의 시간변이를 고려하여 HM-Net에 의한 강건한 문맥의존 음향모델을 작성하기 위해서는 적절한 모델의 구조를 설정하여야 한다. 이상의 실험결과로부터 음소의 시간변이를 고려하여 PDT-SSS 알고리즘에 의한 강건한 HM-Net 문맥의존 음향모델을 작성할 때 3개 또는 4개의 구조를 가지는 초기모델의 구조가 적절하다고 판단되며 본 연구에서 수행한 모델구조 설정이 유효함을 확인할 수 있었다.

## 5. 결 론

본 연구에서는 음성의 시간변이에 따른 적절한 모델의 설정과 강건한 문맥의존 음향모델을 학습하기 위해 다양한 구조의 초기모델을 설정하고 이를 음소결정트리 기반 SSS 알고리즘을 이용하여 각 초기모델에 대해 HM-Net 문맥의존 음향모델을 작성하였다. 작성한 각 모델의 유효성을 확인하기 위해 국어공학센터의 452 단어를 대상으로 음소 및 단어인식 실험을 수행한 결과, 상태수 2000개에서 음소인식의 경우 평균 76.4%의 인식성능을 보였으며 단어인식의 경우 평균 99.3%의 인식성능을 보였다. 또한 한 단어를 인식하는데 초기모델의 구조에 따라 인식속도가 향상됨을 확인할 수 있었다. 따라서 이상의 결과로부터 음성의 시간변이를 고려한 HM-Net 문맥의존 음향모델을 작성하기 위해서는 초기모델의 상태가 3개 또는 4개를 가진 구조가 적당하다고 판단된다. 이상의 실험으로부터 본 연구에서 수행한 모델구조설정에 관한 연구가 강건한 문맥의존 음향모델을 작성하는데 유효함을 확인할 수 있었다.

## 6. 참고문헌


[1] L.R. Rabiner and B.H. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, 1993.

[2] 박준영, "한국어 단어인식을 위한 최적의 연속 HMM모델에 관한 연구," 석사학위논문, 영남대학교, 1995.

[3] 김유진, 김회린, 정재호, "인식단위로서의 한국어 음절에 관한 연구," 한국음향학회지, 제16권 제3호, pp. 64-72, 1997.

[4] 김호경, 구명완, "기본음소 설정을 위한 음소인식률 이용 방안 연구," 제15회 음성통신 및 신호처리 워크숍 논문집, pp. 328-331, 1998.

[5] 이승훈, 김회린, "가변어휘 음성인식기의 응향모델 개선 및 성능분석," 한국음향학회지, 제18권 제8호, pp. 3-8, 1999.

[6] J. Takamia and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP'92, pp. 573-576, 1992.

[7] T. Hori, M. Katoh, A. Ito and M. Kohda, "A Study on a State Clustering-Based Topology Design Method for HM-Nets," Trans. of IEICE, Vol. J81-D-II, No. 10, pp. 2239-2248, 1998.

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