한국음향학회지 제22권 제5호 pp. 388~402 (2003)

음성인식에서 문맥의존 음향모델의 성능향상을 위한 유사음소단위에 관한 연구

A Study on Phoneme Likely Units to Improve the Performance of Context-dependent Acoustic Models in Speech Recognition

임 영 춘*, 오 세 진**, 김 광 동**, 노 덕 규**, 송 민 규**, 정 현 열***
(Young-Choon Lim*, Se-Jin Oh**, Kwang-Dong Kim**, Duk-Gyoo Roh**, Min-Gyu Song**, Hyun-Yeol Chung***)

*주식회사 자모바, **한국천문연구원 KVN 사업본부, ***영남대학교 전자정보공학부

(접수일자: 2003년 2월 13일; 채택일자: 2003년 4월 10일)

본 논문에서는 음소결정트리 기반 HM-Net 문맥의존 음향모델링에 적합한 유사음소단위를 재정의하고 그 유효성을 확인하기 위해 한국어에 대해 단어인식, 4연속 숫자음 인식, 연속음성인식, 태스크 독립 단어인식 실험을 각각 수행하였다. 48개의 유사음소단위의 경우 자음 /ㅂ/, /ㄷ/, /ㄱ/에 대해 음절, 단어 또는 문장에서 위치하는 자리에 따라 초성, 중성, 종성으로 구분하고, 자음 /ㄹ/, /ㅈ/, /ㅎ/에 대해서는 초성, 종성으로 구분하는 문제점이 있다. 따라서 본 논문에서는 문맥의존 음향모델을 효율적으로 작성하기 위해 48개의 유사음소단위의 초성, 중성, 종성으로 나눈 부분을 하나의 음소로 통일하여 39개의 유사음소단위로 새롭게 정의하였다. 새롭게 정의한 39 유사음소단위를 이용하여 인식실험을 수행한 결과, 문맥독립 음소모델을 이용한 단어인식실험은 기존의 48 유사음소단위가 재정의한 39 유사음소단위에 비해 평균 7.06% 향상된 인식성능을 보였으나, 화자독립 단어인식실험에서는 39 유사음소단위가 평균 0.61% 향상된 인식성능을 보였다. 또한 연음현상이 많은 4연속 숫자음 인식실험의 경우에서도 재정의한 39 유사음소단위가 평균 6.55% 향상된 인식률을 보였다. 그리고 연속음성 인식실험에서도 재정의한 39 유사음소단위가 48 유사음소단위에 비해 평균 15.08%향상된 인식률을 보였다. 마지막으로 미지의 문맥요소에 대한 태스크 독립 단어인식실험에서는 48, 39 유사음소단위 모두 전반적으로 낮은 인식률을 보였으나, 39 유사음소단위가 48 유사음소단위에 비해 평균 1.17% 더 향상된 성능을 보였다. 따라서 이상의 인식실험 결과를 바탕으로 본 논문에서 재정의한 39 유사음소단위가 기존의 48 유사음소와 비교하여 문맥의존 음향모델을 구성할 때보다 유효함을 확인할 수 있었다.

핵심용어: 48, 39 유사음소단위, HM-Net, PDT-SSS 알고리즘, 문맥의존 음향모델
투고분야: 음성처리 분야 (2,5)

In this paper, we carried out the word, 4 continuous digits, continuous, and task-independent word recognition experiments to verify the effectiveness of the re-defined phoneme-likely units (PLUs) for the phonetic decision tree based HM-Net (Hidden Markov Network) context-dependent (CD) acoustic modeling in Korean appropriately. In case of the 48 PLUs, the phonemes /ㅂ/, /ㄷ/, /ㄱ/ are separated by initial sound, medial vowel, final consonant, and the consonants /ㄹ/, /ㅈ/, /ㅎ/ are also separated by initial sound, final consonant according to the position of syllable, word, and sentence, respectively. In this paper, therefore, we re-define the 39 PLUs by unifying the one phoneme in the separated initial sound, medial vowel, and final consonant of the 48 PLUs to construct the CD acoustic models effectively. Through the experimental results using the re-defined 39 PLUs, in word recognition experiments with the context-independent (CI) acoustic models, the 48 PLUs has an average of 7.06% higher recognition accuracy than the 39 PLUs used. But in the speaker-independent word recognition experiments with the CD acoustic models, the 39 PLUs has an average of 0.61% better recognition accuracy than the 48 PLUs used. In the 4 continuous digits recognition experiments with the liaison phenomena, the 39 PLUs has also an average of 6.55% higher recognition

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

accuracy. And then, in continuous speech recognition experiments, the 39 PLUs has an average of 15.08% better recognition accuracy than the 48 PLUs used too. Finally, though the 48, 39 PLUs have the lower recognition accuracy, the 39 PLUs has an average of 1.17% higher recognition characteristic than the 48 PLUs used in the task-independent word recognition experiments according to the unknown contextual factor. Through the above experiments, we verified the effectiveness of the re-defined 39 PLUs compared to the 48 PLUs to construct the CD acoustic models in this paper.

Keywords: *48, 39 phoneme likely units, HM-Net (Hidden Markov Network), PDT-SSS algorithm, Context dependent acoustic models*

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

I. 서 론

1960년대 이후로 널리 연구되고 많이 사용되는 HMM Hidden Markov Model)은 시간적, 공간적인 특징을 잘 활영하여 통계적 방법으로 음성인식 등의 다양한 분야에서 널리 사용되고 있다[1,2]. 음성인식에서 HMM으로 음소단위를 모델링할 때 음소를 구성하는 방법에 따라 문맥독립 (CI: Context-Independent) 음소모델링과 문맥의존 (CD: Context-Dependent) 음소모델링으로 나눌 수 있다. 문맥독립 음소모델은 대부분 n상태 m출력의 단순한 구조로 음소를 독립적으로 모델링하기 때문에 이웃하는 음소에 의한 변이음의 정보를 모두 수용하기에는 부족하다[3].

이와 반대로 문맥의존 음소모델은 모델 자체의 수는 많지만 이웃 음소에 대한 변이음을 고려한 모델로서 강건한 음향모델을 생성하는 방법으로 많은 연구가 진행되고 있다[3,4]. 특히 선행음소와 중심음소 또는 중심음소와 후행음소의 결합으로 구성한 다이폰 (diphone)이나 선행음소와 중심음소, 그리고 후행음소를 결합한 트라이폰 (triphone) 구조의 음향모델이 많이 사용되고 있다. 하지만 문맥의존 음소모델은 하나의 중심 음소를 기준으로 선행 및 후행음소에 따라서 수천 가지의 서로 다른 음소가 생성되기 때문에 통계적인 방법인 HMM 기반에서 강건한 음향모델을 작성하기 위해서는 다양한 문맥요소가 포함되어 있는 충분한 학습데이터가 있어야 한다[5,6]. 그 이유는 문맥요소가 부족한 음성데이터로는 다양한 변이음을 효과적으로 학습할 수 없을 뿐만 아니라 미지의 문맥요소가 많이 발생하기 때문이다. 비록 미지의 문맥을 고려한 모델이 문맥독립 음소모델로 대치될 수 있다고 하지만 인식성능에는 그리 많은 영향을 미치지 못한다[7,8].

그러나 지금까지 수행되어 온 연구의 경우 충분한 음성데이터를 확보하기 어려운 실정이기 때문에 한정된 음성데이터 내에서 신뢰성 있는 문맥의존 음소모델을 작성하기 위해 상태결합 (state tying)이나 상태 클러스터링 (state clustering) 등의 연구가 진행되고 있다[9-12].

최근 새로운 방식의 문맥의존 음소모델을 작성할 수 있는 방법으로 HMM과 유사한 방법으로 HM-Net (Hidden Markov Network)에 대한 연구가 활발히 진행되고 있으며 이에 대한 유효성이 입증되고 있다[8-10]. HM-Net은 연속적인 상태분할에 기반한 HMM의 확장 구조로서 부족한 학습 데이터로 강건한 문맥의존 음소모델을 작성하는 방법으로 국내의 경우 연구가 미흡한 실정이다. 따라서 한국어에 대한 다양한 실험이 수행되어야 할 필요성이 있으며, 특히 HM-Net은 변이음을 효과적으로 모델링할 수 있는 방법으로 소개되고 있기 때문에 기존의 문맥독립 음소모델 형태의 HMM 변이음들은 HM-Net 음향모델의 강건함을 저하시킬 수 있기 때문에 HM-Net 음향모델에 효과적인 유사음소단위를 다시 정의할 필요가 있다.

따라서 본 논문에서는 HM-Net 문맥의존 음소모델을 효과적으로 작성하기 위해 기존에 정의한 48개의 유사음소단위를 새롭게 정의하고자 한다. 48개의 유사음소단위의 경우 자음 /ㅂ/, /ㄷ/, /ㄱ/에 대해 이 음소들이 음절, 단어 또는 문장에서 위치하는 자리에 따라 초성, 중성, 종성으로 구분하고 있으며, 자음 /ㄹ/, /ㅈ/, /ㅎ/에 대해서는 초성, 종성으로 구분하고 있다. 따라서 단일구조 형태의 문맥독립 음소모델로서 48개의 유사음소단위를 사용할 경우 인식성능이 우수한 결과를 보이고 있다[13,14]. 다양한 문맥정보를 포함한 대량의 음성데이터를 이용할 경우에는 문제가 없지만, 일반적으로 인식 시스템을 구성할 때 인식대상에 따라 음성 데이터가 달라지기 때문에 다양한 문맥정보를 포함한 대량의 음성데이터를 사용하지 못하는 경우가 발생한다. 따라서 이 경우에는 모델의 학습에 사용한 음성데이터에는 문맥정보가 포함되지만 인식에 사용한 음성데이터에는 그 문맥정보가 포함되지 않아 인식성능이 떨어지는 문제점이 있다. 따라서 본 논문에서는 특정 태스크 또는 중·소규모 음성인식 시스템에 사용할 문맥의존 음향모델을 효율적으로 작성하기 위

해 48개의 유사음소단위의 초성, 중성, 종성부분을 하나의 음소로 통일한 형태로 39개의 유사음소단위로 새롭게 정의하였다.

이렇게 정의한 유사음소단위를 본 논문에서 도입한 HM-Net 문맥의존 음향모델링 방법에 적용하여 한국전자통신연구원 (ETRI)의 445 음성 데이터, 국어공학센터 (KLE)의 452 음성 데이터, 그리고 한국과학기술원 (KAIST)의 무역상담 음성 데이터를 대상으로 숫자음 인식, 단어인식, 연속음성인식 실험을 각각 수행한 후 그 유효성을 확인하고자 한다.

본 논문의 구성은 다음과 같다. II장에서는 HM-Net과 PDT-SSS 알고리즘에 대해 간략히 기술한다. III장에서는 기존에 사용한 48개의 유사음소단위의 문제점과 본 논문에서 재정의한 유사음소단위의 장점과 재정의 필요성에 대해 기술한다. IV장에서는 재정의한 유사음소단위의 유효성을 확인하기 위해 단어, 연속 숫자음, 연속음성 인식 실험을 수행하고 그 결과에 대해 고찰한 후, 마지막으로 V장에서 본 논문의 결론을 맺는다.

II. HM-Net과 PDT-SSS

2.1. HM-Net (Hidden Markov Network)

HM-Net[7]은 SSS (Successive State Splitting) 알고리즘[7]에 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 후행상태 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 자기루프와 이웃하는 상태로의 천이만을 허용하는 left-to-right 형 HMM과 동일하며 일반적인 HMM과 마찬가지로 Baum-Welch 알고리즘[12]에 의해 파라미터를 추정할 수 있다.

HM-Net 구조결정에 사용되는 SSS 알고리즘은 모든 문맥을 나타내는 1 상태의 초기모델로부터 문맥방향과 시간방향으로 상태분할 후 자동적으로 HM-Net의 구조를 결정하는 알고리즘이다. SSS 알고리즘을 전체적으로 간략히 설명하면 다음과 같다. 우선 유사음소단위 (PLUs: Phone Likely Units)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.

2.2. PDT-SSS 알고리즘

본 논문에서는 한국어 음성학적 지식의 음소 질의어에 의한 음소결정트리 (PDT: Phonetic Decision Tree)와 SSS 알고리즘의 장점을 결합한 PDT-SSS (Phonetic Decision Tree-based SSS) 알고리즘[10]을 도입하였다. PDT-SSS는 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.

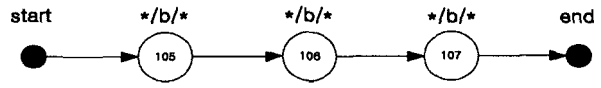
5) 학습 샘플의 우도에 근거하여 문맥방향과 시간방향에서 최적의 HM-Net을 선택한다.
6) Baum-Welch 알고리즘에 의해 HM-Nets의 상태를 재학습한다.
7) 미리 정의한 상태수에 도달할 때까지 단계 3부터 반복한다.

단계 3에서 분할될 상태의 선택은 식 (1)에 의해 계산되어진다.

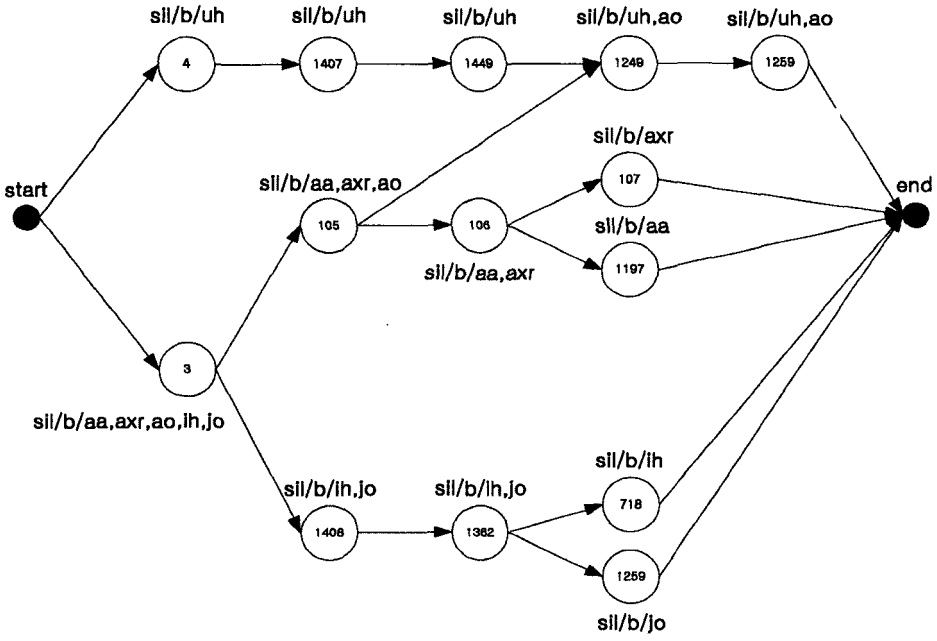
$$d_i = n_i \sum_{p=1}^{P} \frac{\sigma_{ip}^2}{\sigma_{Tp}^2} \tag{1}$$

여기서, σ_{ip}^2, σ_{Tp}^2는 상태 i의 분포 분산과 모든 샘플의 분산 (정규화 계수)을 나타내고, n_i는 상태 i의 추정에 이용한 음소 샘플의 수를, P는 특징 벡터의 차원 수를 각각 나타낸다.

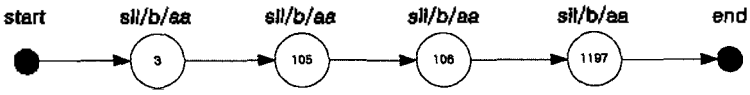
그림 1에 본 논문에서 /b/ 중심음소를 기준으로 선행음소 /sil/인 경우에 대해 HM-Net의 예를 나타내었다. 그림 1의 (a)는 /b/ 음소의 HM-Net 초기모델을 의미하며, (b)는 최종적으로 생성된 /b/ 음소의 HM-Net 구조를 나타낸다. 그리고 (c)는 생성된 HM-Net 문맥의존 음소모델의 하나인 sil/b/aa 모델을 일반적인 HMM과 동일한 구조를 가지고 있는 4상태 4출력의 모델을 나타내고 있다. 여기서 105, 106, 107의 초기모델 상태에서 HM-Net 모델 학습 후 3, 1197 상태를 다른 모델과 공유하는 형태를 보이고 있다. 이 모델의 각 상태는 자기천이와 후속상태로 천이가 가능하며 left-to-right 구조로 모델링되어 있다.



(a) /b/ 음소의 초기 HM-Net 모델
(a) Initial HM-Net model of /b/ phoneme



(b) /b/ 음소의 최종 HM-Net 모델
(b) Final HM-Net model of /b/ phoneme



(c) /b/ 음소의 일반적인 HMM 모델
(c) General HMM model of /b/ phoneme

그림 1. /b/ 모델의 예
Fig. 1. An example of /b/ model.

따라서 일반적인 left-to-right HMM 모델과 동일한 구조로서 Baum-Welch 알고리즘에 의한 재 추정과 비터비 디코딩 (Viterbi decoding)[1,2,12] 등을 수행할 수 있다.

III. 유사음소단위의 고찰

음성인식에 관한 연구 중 우선적으로 수행되어야 할 분야는 인식단위에 관한 연구이다. 일반적으로 기본 인식단위의 선정에 따라 인식시스템에서 인식성능의 차이를 보이는 것으로 알려져 있다[6]. 따라서 본 논문에서는 한국어와 새로운 문맥의존 음소모델 작성법인 HM-Net 음향모델링에 유효한 유사음소단위에 대해 고찰하고자 한다.

3.1. 48개의 유사음소단위

음소 (phoneme)란 화자가 서로 다른 소리로 인식하지 않는 즉 /아/와 /어/ 같이 확실히 구분되는 음들을 말한다. 그리고 이음 (allophone)은 하나의 음소가 문맥구조에 따라 서로 다르게 발음되는 경우를 말한다[15,16]. 우리 국어의 음소는 표 1에 나타낸 것과 같이 초성자음으로 19개의 음소와, 중성모음으로 21의 음소, 종성자음으로 7개의 대표 자음을 사용한다.

일반적으로 유사음소단위는 음성인식에 사용되는 최소 인식단위로 많이 사용되며 기본적인 음소에 변이음을 포함하고 있다. 음향학적 및 음성학적 유사성이 큰 경우에는 음소와 유사음소단위는 동일하게 취급될 수 있지만 그렇지 않을 경우 큰 차이가 있다. 한국어와 영어를 대상으로 한 인식시스템의 경우 약 50여개 정도가 사용되고 있다. 표 2에 나타낸 기존에 정의한 48개의 유사음소단위는 기본 음소에 음성학적인 변이음을 추가하여 구성한 것이다[13,14].

3.2. 문맥독립과 문맥의존 음소모델

문맥독립 음소모델은 유사음소단위를 단독으로 사용한 경우이고 문맥의존 음소모델은 선행 및 후행하는 음소에 따라서 서로 다른 음소를 사용하는 것으로 정의할 수 있다. /ㅏ/ 음소에 대해 예를 들면, 문맥독립 음소모델은 /ㅏ/ 음소 하나로 모든 문맥을 표현한다. 문맥의존 음소모델은 표 1에 나타낸 것과 같이 총 47개의 음소가 있으며 총 47 × 47 = 2,209개의 서로 다른 문맥정보를 가진다. 문맥독립 음소모델의 경우 선행 및 후행하는 음소에 따라 /ㅏ/ 음소는 음향학적 특징이 달라지게 되는데 2,209개의 문맥환경을 하나의 대표음소 /aa/로 표현되므로 음향학적 분해능이 낮아진다. 그러나 문맥의존 음소모델은 선행 및 후행하는 음소의 결합에 따라 모두 다른 음소로 표현되므로 음향학적 분해능도 높아지게 되고 여러 가지 문법규칙에 강건한 모델을 구성할 수 있게 된다. 그러나

표 1. 한국어 음소
Table 1. Korean phoneme.

Initial sound consonant	/ㄱ/, /ㄲ/, /ㄴ/, /ㄷ/, /ㄸ/, /ㄹ/, /ㅁ/, /ㅂ/, /ㅃ/, /ㅅ/, /ㅆ/, /ㅇ/, /ㅈ/, /ㅊ/, /ㅋ/, /ㅌ/, /ㅍ/, /ㅎ/
Middle vowel	/ㅏ/, /ㅓ/, /ㅗ/, /ㅜ/, /ㅡ/, /ㅣ/, /ㅐ/, /ㅔ/, /ㅑ/, /ㅕ/, /ㅛ/, /ㅠ/, /ㅘ/, /ㅝ/, /ㅚ/, /ㅖ/, /ㅙ/, /ㅟ/, /ㅞ/, /ㅒ/, /ㅢ/
Final consonant	/ㄱ/, /ㄴ/, /ㄷ/, /ㄹ/, /ㅁ/, /ㅂ/, /ㅇ/

표 2. 48개 유사음소단위의 정의
Table 2. The definition of 48 phoneme likely units.

Vowel	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,얘/	Wi /의/	
Consonant	b~ /ㅂ/	d~ /ㄷ/	g~ /ㄱ/	z~ /ㅈ/	hh~ /ㅎ/
	bb /ㅃ/	dd /ㄸ/	gg /ㄲ/	zz /ㅉ/	ss /ㅆ/
	s /ㅅ/	p /ㅍ/	t /ㅌ/	k /ㅋ/	ch /ㅊ/
	r /ㄹ/	n /ㄴ/	m /ㅁ/		
First syllable	b /ㅂ/	d /ㄷ/	g /ㄱ/	z /ㅈ/	hh /ㅎ/
Final consonant	b1 /ㅂ/	d1 /ㄷ/	g1 /ㄱ/	l /ㄹ/	ng /ㅇ/
Silence	sil				

만약 학습데이터가 /ㅏ/ 음소에 대해 100개가 존재한다고 가정하면, 문맥독립 음소모델에서는 100개의 데이터를 모두 사용하여 /aa/ 음소를 학습하지만 문맥의존 음소모델에서는 해당 문맥에 맞게 100개의 데이터를 나누어 학습하므로 음향학적 분해능은 높아지지만 신뢰성 있는 모델을 학습하기에는 학습데이터가 충분하지 못한 문제가 발생한다.

3.3. 48 유사음소단위의 문제점

3.3.1. 첫음절 변이음

표 3은 48 유사음소단위에 포함된 자음의 변이음을 나타낸 것이다. 표 3에서 /g~/, /d~/, /b~/, /z~/, /hh~/ 변이음은 선행음소가 묵음 이외의 음소에 의해 영향을 받을 경우를 나타내고, /g/, /d/, /b/, /z/, /hh/ 변이음은 선행하는 문맥환경이 묵음인 경우를 나타낸다. 그리고 /g1/, /d1/, /b1/ 변이음은 종성에 올 경우의 음소를 나타낸다. 그림 2는 "가위바위보"라는 실제 음성 데이터에 48 유사음소단위를 사용하여 레이블링과 트랜스크립션한 예를 나타낸 것이다. 여기서 변이음에 해당하는 자음을 표 4에 나타내었다.

표 4에서 /ㄱ/음소는 선행하는 음소가 묵음인 경우이고 /ㅂ/음소는 선행하는 음소가 묵음 이외의 음소인 경우

표 3. 48 유사음소단위에 포함된 자음의 변이음
Table 3. Allophone of consonants including 48 PLUs.

Phoneme	Allophone
/ㄱ/	/g/, /g~/, /g1/
/ㄷ/	/d/, /d~/, /d1/
/ㄹ/	/r/, /l/
/ㅂ/	/b/, /b~/, /b1/
/ㅈ/	/z/, /z~/
/ㅎ/	/hh/, /hh~/

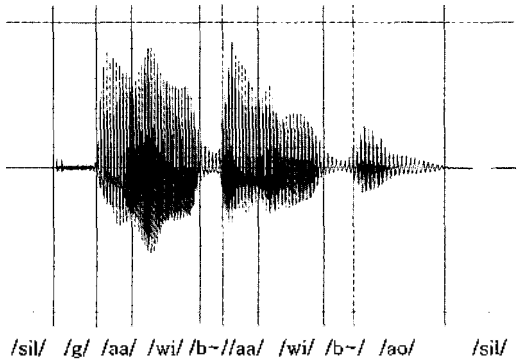


그림 2. "가위바위보"의 음성
Fig. 2. Speech wave form of "가위바위보".

표 4. "가위바위보"의 자음
Table 4. The consonant of "가위바위보".

/silence/ - /ㄱ/	/sil/ - /g/	first syllabel (precede phone=silence)
/ㄱ/ - /ㅂ/	/wi/ - /b~/	phone (/ㄱ/) except the precede phone is silence
/ㄱ/ - /ㅂ/	/wi/ - /b~/	

에 대해 48 유사음소단위에서는 서로 다른 변이음으로 정의하여 사용한다. 이는 유사음소단위로 문맥독립 음소모델을 구성할 때 문맥의존 요소를 일부 도입하여 구성한 것이다. 변이음은 음향학적으로 서로 다른 음가를 가지므로 단독으로 표기를 할 때는 구분하여 사용해야 하지만 문맥의존 음소모델에서의 중심음소는 선행 및 후행하는 음소에 따라 음향학적 특징을 받게 된다. 따라서 중심음소에는 변이음이 올 필요가 없다. 표 2를 바탕으로 문맥의존 음소모델을 구성하면 다음과 같다.

/sil/ - {/g/, /d/, /b/, /z/, /hh/} + {/모음/} (가)
{*} - {/g~/, /d~/, /b~/, /z~/, /hh~/} + {/모음/} (나)

여기서, '*' 은 묵음 (/sil/)을 제외하고 자음과 모음 결합 방법에 의한 모든 모음과 종성자음을 의미한다. (가)와 (나)의 경우, 후행하는 음소로는 모음밖에 올 수 없으며 선행하는 음소에 따라 서로 다른 문맥의존 음소모델이 구성된다. 묵음도 '*' 부분에 올 수 있다면 중심음소를 구분할 필요가 없게 된다.

그림 3의 (b)와 같이 선행음소가 /sil/ 경우의 음소 문맥환경을 따로 고려할 필요없이 그림 3의 (c)와 같이 중심음소가 /g~/, /d~/, /b~/, /z~/, /hh~/ 인 음소 문맥환경에 포함시킬 경우에도 경로에 의해 서로 다른 문맥의존 모델이 구성된다. 따라서 48 유사음소단위의 첫음절에 해당되는 변이음은 문맥의존 모델의 구조에 의해 그 필요성이 없어지게 된다.

3.3.2. 초성과 종성

본 논문에서는 문맥의존 음소모델의 특징 및 음소결정 트리 기반 상태 클러스터링 방법을 고려하여 초성과 종성을 하나의 유사음소로 설정하는 방법에 대해 고찰하고자 한다.

초성과 종성은 그 음가에 있어서 서로 다른 음소로 구분되기 때문에 문맥독립 음소모델에 있어서는 구분하는 것이 당연하다. 문맥의존 음소모델의 경우 중심음소를 기준으로 선행 및 후행음소의 결합을 통해 서로 다른 음소로 표기되기 때문에 하나의 기본음소로 나타내는 것을

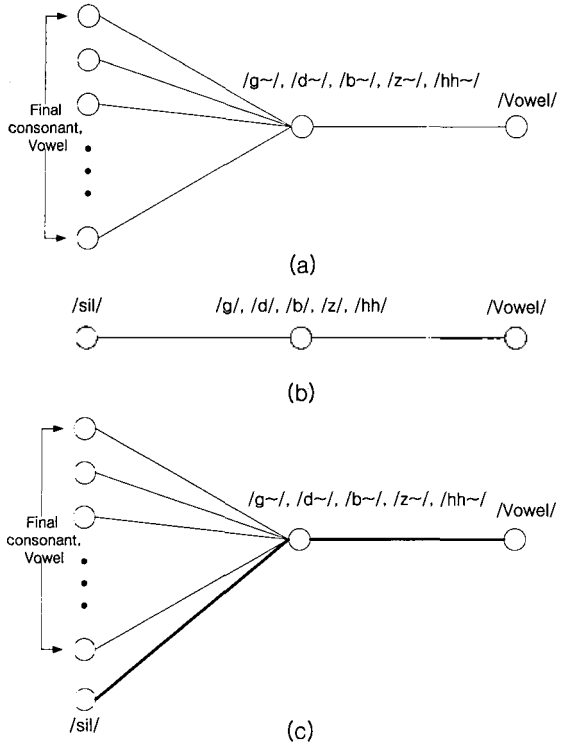


그림 3. 다른 문맥환경의 결합
Fig. 3. Incorporation of different context environment.

살펴보면 다음과 같다. 한국어에 대해 초성 자음을 C (Consonant), 중성 모음을 V (Vowel), 종성 자음을 C' 라고 할 때의 경우를 표 5에 나타내었다.

표 5의 음영부분은 한국어의 초성, 중성, 종성의 결합 방법에 의해 나타날 수 없는 경우로서 여기서는 고려할 필요가 없다. 표 5에서 중심음소가 모음인 (5), (6), (7)의

표 5. 한국어 음소에 의한 문맥의존 모델의 모든 가능성
Table 5. All possible case of CD model (triphone) by Korean phoneme.

Ex.	Base	Case1	Case2	Case3	Case4
(1)	C/C/V	C'/C/V	C/C'/V	C/C/C'	C/C/V
(2)	V/C/C	V/C'/C	V/C/C'	V/C'/C	V/C/C
(3)	V/C/V	V/C'/V	V/C/V		
(4)	C/C/C				
(5)	C/V/V	C'/V/V	C/V/V		
(6)	V/V/C	V/V/C'	V/V/C		
(7)	C/V/C	C'/V/C	C/V/C'	C'/V/C'	C/V/C
(8)	V/V/V	V/V/V			

경우를 살펴보면 다음과 같다. 먼저 (5), (6)의 경우 초성과 종성에 의해 2가지의 문맥의존 음소모델이 생성될 수 있고, (7)의 경우 초성과 종성에 의해 서로 다른 4가지의 문맥의존 음소모델이 생성될 수 있다.

본 논문에서 사용되는 음소결정트리의 질의어는 [14]에서 정의한 한국어 음운규칙을 이용하였다. [14]에서는 한국어 종성에 오는 자음을 따로 구분하지 않고 초성음소와 동일하게 구분하고 있다. 따라서 [14]에서 제시한 한국어 음운규칙을 바탕으로 표 5의 (7)의 경우에 대해 음소결정트리를 구성한 것을 그림 4에 나타내었다. 그림 4에서 중심음소가 같은 모음인 C'/V/C, C/V/C', C'/V/C', C/V/C를 루트노드에 위치시킨 후 음소결정트리의 상태분할을 수행하면 최종 노드인 잎 노드에서 초성과 종성이 같은 클래스에 속하게 되고 HM-Net은 같은 상태를 공유하게 된다. 결국 초성과 종성을 구분하더라도 한국어 자음의

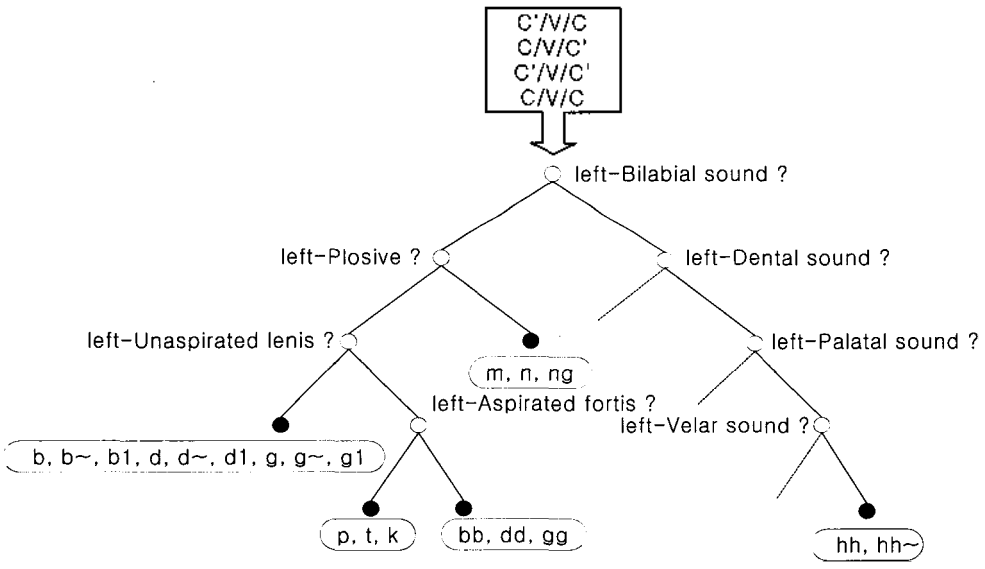


그림 4. 표 5의 (7)에 대한 음소결정트리의 예
Fig. 4. An example of PDT by (7) of table 5.

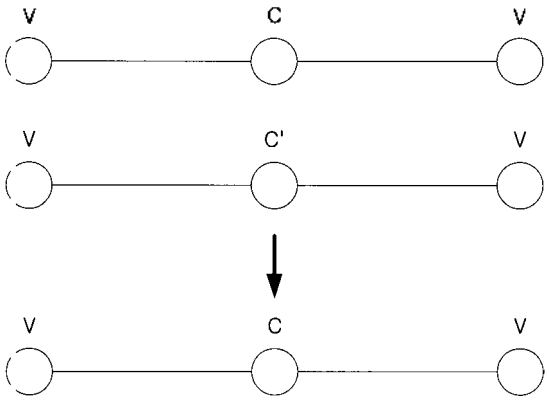


그림 5. VCV와 VC'V의 조합
Fig. 5. Combination of VCV and VC'V.

분류상 초성과 종성을 분리할 수 없고 같은 질의어의 범주에 속하게 된다. 따라서 음소결정트리 기반 상태 클러스터링에 의해 같은 상태를 공유하기 때문에 같은 파라미터를 사용하게 된다. 표 5의 (5), (6), (7), (8)의 경우 중심음소가 동일한 모음이기 때문에 음소결정트리와 질의어에 의해 자동으로 상태를 공유하게 된다. 따라서 선행 및 후행음소에 대해 초성과 종성을 구분할 필요가 없게 된다.

표 5의 (1), (2), (3)의 경우는 중심음소가 자음인 경우를 나타낸다. 표 5에서 (1), (2)의 경우는 다른 계열의 음소가 위치하므로 그 구조에 의해 자동적으로 초성과 종성을 분류할 수 있게 된다. 그림 5에 나타낸 것과 같이 표 5의 (3)의 경우는 V/C'/V, V/C/V와 같은 중심음소에 초성과 종성이 같다고 하면 선행 및 후행에 위치하는 음소가 동일하기 때문에 두 가지를 구분할 수 없다. 하지만 이 경우는 한국어의연음현상에 해당되는 것으로 V/C'/V는 모음/종성/모음 유형이고, V/C/V형은 모음/초성/모음 유형으로 자연스러운 발성의 경우 V/C'/V형이 연음현상에 의해 V/C/V형이 된다. 따라서 이와 같은 두 가지의 문맥을 하나의 문맥으로 볼 수 있다. 이러한 연음현상은 조음현상과 더불어 오인식을 유발하는 대표적인 문제점으로 특히 연속 숫자음에서 많이 나타나는 현상이다.

3.4. 유사음소단위의 재정의

본 논문에서는 3.3절에서 기술한 48 유사음소단위의 문제점을 바탕으로 변이음 9개 (첫음절: /b/, /d/, /g/, /hh/, /z/, 종성: /b1/, /d1/, /g1/, /l/)를 제거하고 /b~/, /d~/, /g~/, /z~/, /hh~/ 유사음소를 /b/, /d/, /g/, /z/, /hh/ 로 재정의 한다. 표 6에 48 유사음소단위와 39 유사음소단위를 비교하여 나타내었다. 그리고 표 7에 본 논문에서 재정의한 39 유사음소단위를 나타내었다.

표 6. 48, 39 유사음소단위의 비교
Table 6. Comparison of 48, 39 PLUs.

48 PLUs	Remarks	39 PLUs	Remarks
g	First syllable initial sound	g	Initial sound, Final consonant
g~	Initial sound		
g1	Final consonant		
d	First syllable initial sound	d	Initial sound, Final consonant
d~	Initial sound		
d1	Final consonant		
b	First syllable initial sound	b	Initial sound, Final consonant
b~	Initial sound		
b1	Final consonant		
z	First syllable initial sound	z	Initial sound
z~	Initial sound		
hh	First syllable initial sound	hh	Initial sound
hh~	Initial sound		
r	Initial sound	r	Initial sound
l	Final consonant		

표 7. 재정의한 39 유사음소단위
Table 7. The redefinition of 39 PLUs.

Vowel	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,얘/	Wi /의/	
Consonant	b /ㅂ/	d /ㄷ/	g /ㄱ/	z /ㅈ/	hh /ㅎ/
	bb /ㅃ/	dd /ㄸ/	gg /ㄲ/	zz /ㅉ/	ss /ㅆ/
	s /ㅅ/	p /ㅍ/	t /ㅌ/	k /ㅋ/	ch /ㅊ/
	r /ㄹ/	n /ㄴ/	m /ㅁ/		
Final consonant	ng /ㅇ/				
Silence	sil				

3.4.1. 재정의한 유사음소단위의 사용 예

표 8에 국어공학센터(KLE)의 452 단어음성의 발성 리스트에서 나타날 수 있는 48 유사음소단위와 39 유사음소단위의 문맥의존 음소모델 /ㅂ/의 예를 나타내었다. 48 유사음소단위에서 /b~/ 와 /b1/ 음소를 모두 /b/ 음소로 대체시키면 표 8의 중간 부분의 음소가 나온다. 이 때 음영부분의 문맥의존 음소를 살펴보면 48 유사음소단위에서 /aa-b~+eh/, /aa-b1+eh/ 음소를 39 유사음소단위에서는 /aa-b+eh/ 음소로 표현되어 하나의 문맥으로 나타나는 것을 알 수 있다. 그리고 /ao-b~+aa/, /ao-b1+aa/ 음소도 /ao-b+aa/ 로 표현된다. 여기서 /ao-b1+aa/는

표 8. 48, 39 유사음소단위에 의한 문맥의존 모델 /ㅂ/의 예
Table 8. An example of CD model /ㅂ/ by 48, 39 PLUs.

48 PLUs		39 PLUs
aa-b~+ao	aa-b+ao	aa-b+ao
aa-b1+b~	aa-b+b	aa-b+b
aa-b~+eh	aa-b+eh	aa-b+eh
aa-b1+eh	aa-b+eh	
aa-b1+g~	aa-b+g~	aa-b+g
aa-b1+hh~	aa-b+hh~	aa-b+hh
...		
ae-b~+axr	ae-b+axr	ae-b+axr
ae-b~+jv	ae-b+jv	ae-b+jv
ao-b~+aa	ao-b+aa	ao-b+aa
ao-b1+aa	ao-b+aa	
ao-b1+g~	ao-b+g~	ao-b+g
ao-b~+ih	ao-b+ih	ao-b+ih
...		
sil-b+aa	sil-b+aa	sil-b+aa
sil-b+ae	sil-b+ae	sil-b+ae
sil-b+ao	sil-b+ao	sil-b+ao
sil-b+axr	sil-b+axr	sil-b+axr
sil-b+eh	sil-b+eh	sil-b+eh
sil-b+ih	sil-b+ih	sil-b+ih
sil-b+jv	sil-b+jv	sil-b+jv
sil-b+uh	sil-b+uh	sil-b+uh
...		
81 kinds		79 kinds

/모음-자음(종성)+모음/인 VCV형으로 연음현상을 나타낸다. /aa-b~+ao/와 /aa-b1+b~/의 경우는 /aa-b+ao/와 /aa-b+b/로 표현되는데 후행음소에 의해 서로 다른 문맥을 나타낸다. 또한 48 유사음소단위의 첫음절에 해당되는 문맥들은 표 8에 나타낸 것처럼 음소의 재정의 후에도 표기상으로는 아무런 변화가 없는 것을 알 수 있다.

3.4.2. 재정의한 유사음소단위의 장점

재정의한 39개의 유사음소단위는 문맥의존형 모델의 선행 및 후행음소의 결합과 음소결정트리의 상태분할에서 다음의 두 가지 장점이 있다.

첫 번째, 문맥의존 모델을 구성할 때 학습 데이터의 부족에 대한 문제점을 어느 정도 해결할 수 있다. KLE 452 단어 리스트에서 48 유사음소단위와 재정의한 39 유사음소단위에 의한 음소출현 빈도를 표 9에 나타내었다.

표 9에서 48 유사음소단위의 중심음소 /b/는 8개의 학습데이터를, /b~/는 37개의 학습데이터를, /b1/은 36개의 학습데이터를 이용하여 상태분할과 학습을 수행하는 것을 알 수 있다. 하지만 재정의한 39 유사음소단위는 48

표 9. 48, 39 유사음소단위에 의한 KLE 452 단어리스트에서 음소 출현 빈도
Table 9. The frequency of phoneme occurrence in the KLE 452 word list by 48, 39 PLUs.

48 PLUs		39 PLUs	
b	8	b	79
b~	37		
b1	36		
d	10	d	113
d~	54		
d1	53		
g	15	g	138
g~	65		
g1	61		
hh	15	hh	71
hh~	56		
z	10	z	71
z~	61		
r	49	r	101
l	55		

유사음소단위와는 달리 모든 음소가 하나의 /b/음소로 대표되기 때문에 총 79개의 학습데이터를 사용하여 문맥방향 상태분할과 학습을 수행하게 된다. 따라서 제한된 학습데이터를 이용할 경우 보다 많은 학습데이터를 확보하게 되어 좀더 강건한 모델을 학습할 수 있게 된다.

두 번째, 연속음성인식에서의 단어사전을 구성할 때 48 유사음소단위의 문제점을 해결할 수 있다. 연속음성인식에서 48, 39 유사음소단위로 구성된 단어사전의 예를 표 10에 나타내었다. 연속음성인식에 사용되는 단어들은 문장의 어느 위치에나 올 수 있는 구조를 가지고 있다. 예를 들어 "가격" 이란 단어로 만들 수 있는 문장을 고려해 보면 다음과 같다.

표 10. 48, 39 유사음소단위에 의한 단어사전
Table 10. Word lexicon by 48, 39 PLUs.

Word	Phoneme sequence of 48 PLUs
가	g aa
가격	g aa g~ jv g1
가격대	g aa g~ jv g1 d~ ae
가격상승	g aa g~ jv g1 s aa ng s U ng
...	...
Word	Phoneme sequence of 39 PLUs
가	g aa
가격	g aa g jv g
가격대	g aa g jv g d ae
가격상승	g aa g jv g s aa ng s U ng
...	...

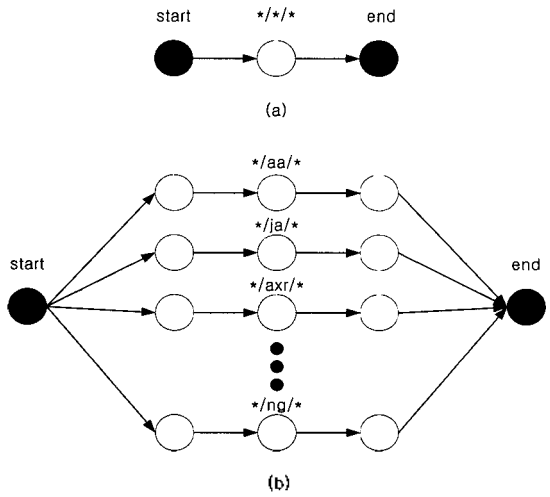


그림 6. HM-Net 모델의 초기구조
Fig. 6. The initial structure of HM-Net model.

가격이 얼마입니까 (다)
이 상품의 가격이 얼마입니까 (라)

(다) 문장에서 "가격이" 라는 단어는 어두, 즉 첫음절에 있기 때문에 묵음의 영향을 받게 되어 음소열의 구성은 /sil g aa g~ jv gl ih axr l m aa ih bl n ih gg aa sil/와 같이 된다. 하지만 (라) 문장의 "가격이" 라는 단어는 발성 중간에 나타나는 단어이므로 /sil ih s aa ng p uh m Vi g~ aa g~ jv gl ih axr l m aa ih bl n ih gg aa sil/와 같은 음소열의 구성을 가진다. (다)와 (라)의 문장으로 사전정보를 구성한다면, "가격" 이란 단어의 트랜스크립션 정보가 2가지 존재하는 것을 볼 수 있다. 즉, 모델을 학습할 때 어두에 오는 "가격" 과 어중에 오는 "가격"의 /ㄱ/ 음소가 서로 다른 음소 /g/, /g~/로 학습된다. 하지만 인식할 때 사전구성에서는 일반적으로 어두와 어중에 오는 단어를 구분하지 않고 단지 어두에 오는 경우만을 고려하여 작성하는 경우가 많다. 따라서 학습할 때 사용한 문맥정보와 인식할 때의 음소열에 대한 문맥정보가 일치하지 않은 경우가 많아 인식률의 저하를 초래하게 된다. 이에 대해 해당 단어에 대해서 여러 개의 사전정보를 고려할 수도 있지만 연속음성인식에서는 수천에서 수만 가지의 단어가 사용되기 때문에 여러 개로 사전정보를 구성한다면 인식성능은 조금 향상되지만 어휘 수의 증가로 인해 인식속도가 저하되는 원인이 된다. 하지만 39 유사음소단위에서는 첫음절 변이음을 모두 제외하였기 때문에 학습과 인식에 사용되는 음소열의 문맥정보가 일치하여 이러한 문제점을 해결할 수 있게 된다.

IV. 인식 실험 및 결과

본 논문에서 한국어에 적합한 문맥의존 음향모델을 작성하기 위해 도입한 PDT-SSS 알고리즘으로 작성한 HM-Net 음향모델과 재정의한 39 유사음소단위의 유효성을 확인하기 위해 여러 가지 다양한 태스크에 대해 음성인식 실험을 수행하였다. 본 논문에서 사용한 음성 데이터베이스를 표 11에 나타내었다. 인식실험에 사용된 음성 데이터는 단어인식의 경우 KLE 452 단어를, 숫자음 인식의 경우 KLE 4연속 숫자음을 각각 사용하였다. 연속음성 인식실험은 한국과학기술원 (KAIST)에서 작성한 무역 상담용 연속음성 데이터를 사용하였다. 또한 태스크 독립 단어인식 실험에는 한국전자통신원 (ETRI) 445 단어음성 데이터를 사용하였다. 표 12에 음성 데이터의 분석 조건을 나타내었다. 그림 6은 본 논문에서 사용한

표 11. 음성 데이터베이스
Table 11. Speech databases.

Exp.	Word recognition	Digit recognition
speech data	KLE 452	KLE 4 connected digit
Train	Male 35 spks. 1 utt.	Male 35 spks. 1~4 utt.
Test	Male 3 spks.1 utt.	Male 3 spks.1 utt.
Model	Monophone, HM-Net	HM-Net
Recog.	Speaker independent	Speaker independent
Exp.	Continuous speech recognition	Task independent
speech data	Trade related continuous speech	ETRI 445
Train	Male 90 spks. 1 utt.	KLE452 male 35 spks. 1 utt.
Test	Male 10 spks. 1 utt.	ETRI445 male 20 spks. 1 utt.
Model	HM-Net	HM-Net
Recog.	Speaker independent	Speaker independent

표 12. 음성 데이터의 분석조건
Table 12. Analysis conditions of speech data.

Sampling Frequency	16 kHz
Quantization	16 bit
Frame length	25 ms
Frame period	10 ms
Analysis window	Hamming Window
Feature parameters	12 orders LPC-MEL cepstrum + delta power + 1st, 2nd regression coefficient = 39 orders

초기 HM-Net 모델의 구조를 나타내었다. 본 논문에서는 문맥독립 음소모델에서 널리 사용하고 있는 3개 또는 4개의 상태를 고려하여 문맥의존 음소모델의 구조로 그림 6의 (b) 구조를 사용하였다.

4.1. 문맥독립 음소모델에 대한 단어인식실험

48 유사음소단위와 39 유사음소단위를 이용하여 작성한 문맥독립 음소모델에 대해 인식성능 차이를 비교검토하기 위해 KLE 452 단어를 대상으로 단어인식실험을 수행하였다. 표 11에 나타낸 것과 같이 모델의 학습에는 남성 35명이 1회 발성한 단어를 사용하였으며 문맥독립 음소모델로서 각 모델은 1, 4, 8개의 혼합수를 가진다. 그리고 인식에는 학습에 참가하지 않은 3명의 1회 발성을 사용하였다. 인식 알고리즘은 단어쌍 (Word-pair) 문법[1]을 사용하는 단일 경로 비터비 (One-Pass Viterbi) 알고리즘[1,2]을 사용하였다.

그림 7에 나타낸 인식실험 결과에서 48 유사음소단위가 39 유사음소단위에 비해 혼합수 1인 경우 13.43%, 4인 경우 4.65%, 8인 경우 3.14%의 향상된 인식률을 각각 보였다. 따라서 문맥독립 음소모델의 경우 앞에서 설명한 것과 같이 39 유사음소단위보다 48 유사음소단위가 인식시스템에 보다 효과적임을 실험을 통해 확인할 수 있었다.

4.2. 문맥의존 음향모델에 대한 단어인식실험

문맥의존 음소모델에 대해 48 유사음소단위와 39 유사음소단위의 성능을 평가하기 위해 KLE 452 단어를 대상으로 단어인식 실험을 수행하였다. 모델의 학습과 평가는 4.1절과 동일하다. 그리고 HM-Net 문맥의존 음소모델링 방법에서 초기 HM-Net 모델은 그림 6의 (b) 구조를 사용하였으며 초기모델의 상태수는 114개이다. HM-Net 문맥의존 음소모델은 상태수의 증가에 따른 인식률의 변화 정도를 파악하기 위해서 상태수를 200에서 3,000까지 200상태씩 증가시키면서 학습하였으며 혼합수는 각 모델마다 4개를 가진다. 인식 알고리즘은 4.1절과 동일하다. 그림 8에 인식실험 결과를 나타내었다.

그림 8에서 48 유사음소단위의 경우 상태수가 600개일 때 98.6%의 최고 단어인식률을 나타내었고 그 이후에는 학습에 포함된 문맥정보에 대해 상태분할을 수행할 때 사용된 음성데이터가 부족하여 학습이 제대로 되지 않아 인식성능이 감소하는 결과를 보이고 있다. 반면 39 유사음소단위는 600 상태 이후 48 유사음소단위에 비해 우수한 인식성능을 보이고 있다. 전체적으로 39 유사음소단위가 48 유사음소단위와 비교하여 평균 0.61%의 인식률 향상을 보였다. 따라서 39 유사음소단위가 문맥의존 HM-Net 음향모델에 보다 효과적임을 확인할 수 있었다.

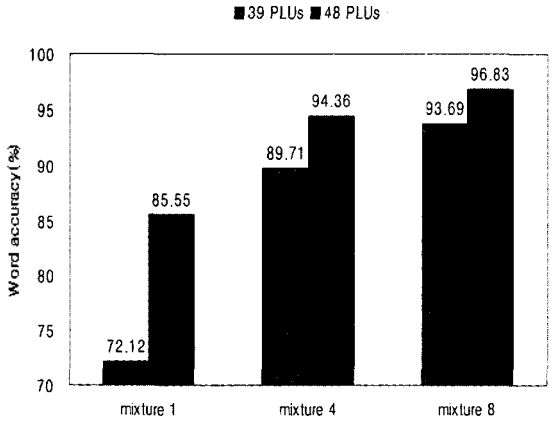


그림 7. 48, 39 유사음소단위의 문맥독립 모델에 대한 단어인식률의 비교
Fig. 7. Comparison of word recognition results by CI model of 48, 39 PLUs.

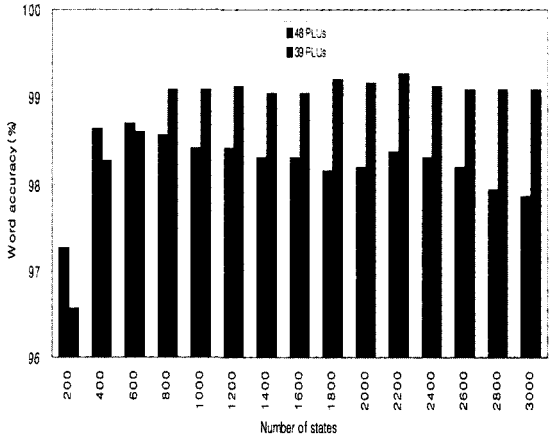


그림 8. 48, 39 유사음소단위에 대한 단어인식률 비교
Fig. 8. Comparison of word recognition results by 48, 39 PLUs.

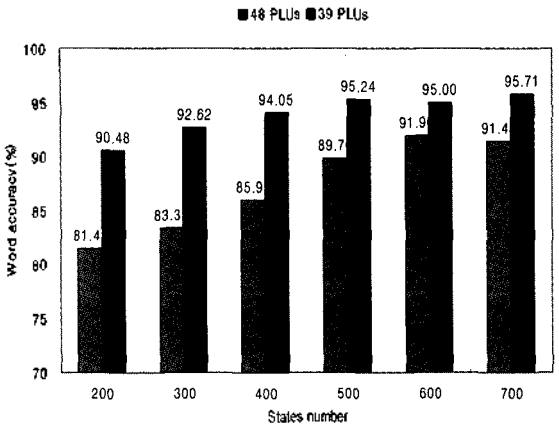


그림 9. 48, 39 유사음소단위에 대한 4연속 숫자음 인식률 비교
Fig. 9. Comparison of 4 continuous digit recognition results by 48, 39 PLUs.

4.3. 연속 숫자음 인식실험

본 절에서는 연속 숫자음에 대해 재정의한 39 유사음소단위의 유효성을 확인하기 인식실험을 수행하였다. 표 11에 나타낸 것과 같이 모델 학습에 사용한 음성데이터는 KLE에서 채록한 4연속 숫자음으로 35명 1회에서 4회까지 녹음한 음성데이터를 사용하였다. HM-Net 문맥의존 음소모델은 상태수를 200에서 700까지 100상태씩 증가시키면서 혼합수 4개를 가지도록 구성하였다. 숫자음 음성 데이터의 경우 데이터 량이 많지 않아 화자가 1회에서 4회까지 발성한 데이터를 사용하였으며 상태수에 대한 상태분할의 경우도 데이터 량이 적어 많은 상태수로 분할하지 않았다. 평가는 학습에 참가하지 않은 나머지 3명의 1회 발성을 사용하였으며, 인식 알고리즘은 4.1절과 동일하다. 그림 9에 인식실험 결과를 나타내었다.

그림 9에 나타낸 것과 같이 재정의한 39 유사음소단위가 48 유사음소단위에 비해 평균 6.55% 인식률 향상을 보이고 있다. 한국어 숫자음의 경우 국어문법 중 연음현상에 민감하게 반응을 보이고 있는데 실험을 통해 39 유사음소단위가 48 유사음소단위에 비해 연음현상이 많은 숫자음에서 더 유효함을 확인할 수 있다. KLE 4연속 숫자음의 경우 음성 데이터 량이 작기 때문에 748 상태에서 문맥방향과 시간방향의 우도가 무한대 값이 출력되어서 더 이상 상태분할을 수행할 수 없었다.

4.4. 연속음성인식 실험

본 논문에서 재정의한 39 유사음소단위가 연속음성에서도 유효함을 확인하기 위해 KAIST에서 작성한 무역 상담용 연속음성 데이터에 대해 연속음성인식 실험을 수행하였다. 표 11에 나타낸 것과 같이 HM-Net 문맥의존 음소모델의 학습에는 남성 90명이 1회 발성한 음성데이터를 사용하였으며, 평가에는 학습에 참가하지 않은 남성 10명이 1회 발성한 음성데이터를 사용하였다. 무역상담용 연속음성 데이터는 각 화자가 서로 다른 평균 95~100 문장을 각각 발성한 것으로 구성되어 있다. 무역 상담용 연속음성 데이터는 비교적 데이터 량이 많기 때문에 모델의 상태수는 각각 1,000, 2,000, 3,000개와 혼합수 4개를 가지는 문맥의존 HM-Net 음향모델을 작성하였다. 인식 알고리즘은 1-pass 탐색에서는 단어 2-gram 모델[12]을, 2-pass 탐색에서는 단어 3-gram 모델[12]을 사용하는 다중 경로 (Multi-Pass) 탐색 알고리즘[12,14]을 사용하였다. 그리고 실제 인식률의 정도를 확인하기 위해 본 논문에서는 형태소 분석을 사용하지 않은 단어 2-gram 모델만 사용하여 1-pass 탐색만 수행하였다.

그림 10과 11에 문장 인식률과 문장에 포함된 단어 인식

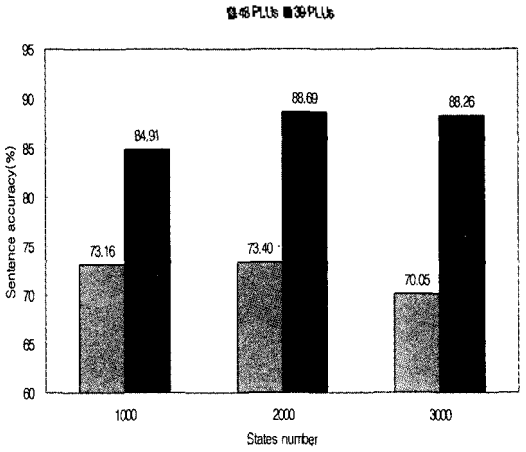


그림 10. 48, 39 유사음소단위의 문장 인식률의 비교
Fig. 10. Comparison of sentence accuracy by 48, 39 PLUs.

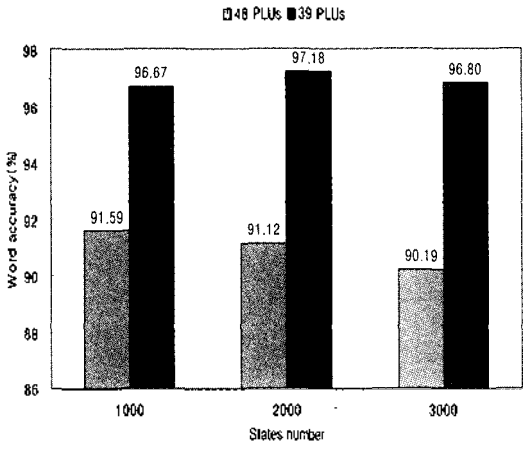


그림 11. 48, 39 유사음소단위의 문장에 포함된 단어 인식률의 비교
Fig. 11. Comparison of word accuracy by 48, 39 PLUs.

률을 각각 나타내었다. 그림 10에 나타낸 것과 같이 39 유사음소단위를 사용한 경우 각 상태수에 따라 문장 인식률이 48 유사음소단위에 비해 평균 11.75%, 15.29%, 18.21%의 향상된 인식률을 보였다. 또한 그림 11의 문장에 포함된 단어 인식률의 경우에도 각 상태수에 따라 48 유사음소단위보다 평균 5.08%, 6.06%, 6.61%의 향상된 인식률을 보였다.

4.2절의 단어인식 실험결과에 비하여 문장 내의 단어 인식 성능이 높은 이유는 사용된 유사음소단위와 언어모델의 영향으로서 단어사전의 유사음소단위 음소열에서 그만큼의 인식대상이 줄었으며, N-gram 언어모델의 경우 출현단어의 빈도가 제한적일지라도 인식대상에 포함될 경우 이웃한 단어의 상관관계로 인해 인식될 가능성이 높기 때문으로 생각된다.

실험결과에 의해 연음현상이나 조음현상이 가장 빈번히 나타나는 연속음성 데이터에서 39 유사음소단위가 48 유사음소단위에 비해 우수한 성능을 보임을 확인할 수 있었다. 이상의 실험결과로부터 본 논문에서 도입한 PDT-SSS 알고리즘에 의한 문맥의존 HM-Net 음향모델 작성법에 재정의한 39 유사음소단위가 유효함을 확인할 수 있었다.

4.5. 태스크 독립 단어인식 실험

SSS 알고리즘인 단점인 미지의 문맥에 대한 해결방법으로 도입한 PDT-SSS 알고리즘의 미지 문맥 모델링에 관한 성능 및 재정의한 유사음소의 유효성을 확인하기 위해 태스크 독립 단어인식실험을 수행하였다. HM-Net 음향모델은 4.2절의 단어인식실험에서 작성한 HM-Net 음향모델을 사용하였고 인식에는 문맥환경이 다른, 즉 미지의 문맥 요소를 포함하고 있는 ETRI 445 남성화자 20명이 1회 발성한 음성 데이터를 사용하였다. KLE 452 단어음성에 대한 단어리스트와 ETRI 445 단어음성에 대한 단어리스트에서 triphone 개수 및 조사한 문맥 유사도 결과를 표 13에 나타내었다.

표 13. KLE 452와 ETRI 445 단어의 문맥유사도
Table 13. Context likely rate of KLE 452 and ETRI 445.

	Number of triphone	Context likely rate
KLE PBW 452	2,117	33.8%
ETRI PBW 445	1,778	40.3%

두 음성 데이터의 단어 리스트를 조사한 결과 동일한 문맥환경의 개수는 717개로 나타났다. 따라서 KLE 452 단어음성은 ETRI 445 단어음성과 717/2117 = 33.8%의 문맥 유사도를 가지고, ETRI 452 단어음성은 KLE 452 단어음성과 717/1778 = 40.3%의 문맥 유사도를 가진다. 따라서 ETRI 445 단어음성의 59.7%는 KLE 452 단어음성에서 미지의 문맥요소에 해당된다. 이에 대한 인식결과를 그림 12에 나타내었다.

그림 12에 나타낸 것과 같이 4.2절의 태스크 의존 인식실험과 비교하여 인식률이 많이 저조함을 알 수 있다. 48 유사음소단위의 경우 상태수 600에서 89.98%의 최고 인식률을 나타내었고 재정의한 39 유사음소단위의 경우 상태수 800에서 91.58%의 최고 인식률을 나타내었다. 이는 학습 데이터로 사용된 KLE 452 단어 음성데이터가 미지의 문맥요소를 강건하게 학습하기에는 부족한 것으로 생각할 수 있다. 하지만 여기서도 재정의한 39 유사음소단위가 48 유사음소단위에 비해 평균 1.17% 향상된 인식률

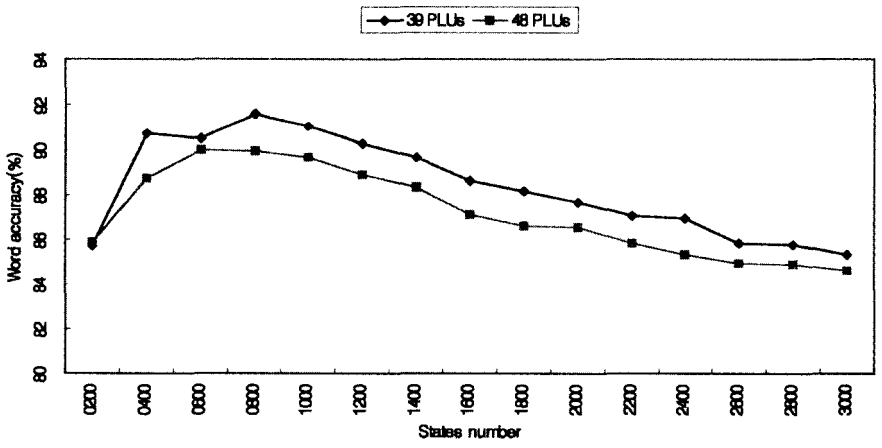


그림 12. 48, 39 유사음소단위의 태스크 독립 단어인식률의 비교
Fig. 12. Comparison of task independent recognition results by 48, 39 PLUs.

을 보였다. 따라서 미지의 문맥요소에 있어서도 재정의한 39 유사음소단위가 유효함을 확인할 수 있었다.

V. 결 론

본 논문에서는 음소결정트리 기반 HM-Net 문맥의존 음향모델링에 적합한 유사음소단위를 재정의하고 그 유효성을 확인하기 위해 한국어에 대해 단어인식, 4연속 숫자음인식, 연속음성인식, 태스크 독립 단어인식 실험을 각각 수행하였다.

48개의 유사음소단위의 경우 자음 /ㅂ/, /ㄷ/, /ㄱ/에 대해 이 음소들이 음절, 단어 또는 문장에서 위치하는 자리에 따라 초성, 중성, 종성으로 구분하고 있으며, 자음 /ㄹ/, /ㅈ/, /ㅎ/에 대해서는 초성, 종성으로 구분하고 있다. 다양한 문맥정보를 포함한 대량의 음성데이터를 이용할 경우에는 문제가 없지만, 그렇지 못한 경우에는 인식성능이 저하되는 문제점이 있다. 따라서 본 논문에서는 문맥의존 음향모델을 효율적으로 작성하기 위해 48개의 유사음소단위의 초성, 중성, 종성으로 나뉜 부분을 하나의 음소로 통일하여 39개의 유사음소단위로 새롭게 정의하였다.

새롭게 정의한 39 유사음소단위를 이용하여 인식실험을 수행한 결과, 문맥독립 음소모델을 이용한 단어인식 실험의 경우 기존의 48 유사음소단위가 재정의한 39 유사음소단위에 비해 평균 7.06% 향상된 인식성능을 보였으나, 화자독립 단어인식실험의 경우 39 유사음소단위가 평균 0.61% 향상된 인식성능을 보였다. 또한 연음현상이 많은 4연속 숫자음 인식실험의 경우에서도 재정의한 39 유사음소단위가 평균 6.55% 향상된 인식률을 보였다. 그리고 연속음성 인식실험에서도 재정의한 39 유사음소단위가 48 유사음소단위에 비해 평균 15.08%향상된 인식률을 보였다. 마지막으로 미지의 문맥요소에 대한 태스크 독립 단어인식실험에서는 48, 39 유사음소단위 모두 전반적으로 낮은 인식률을 보였으나, 39 유사음소단위가 48 유사음소단위에 비해 평균 1.17% 더 향상된 성능을 보였다.

따라서 이상의 인식실험 결과를 바탕으로 본 논문에서 재정의한 39 유사음소단위가 기존의 48 유사음소와 비교하여 문맥의존 음향모델을 구성할 때 보다 유효함을 확인할 수 있었다. 또한 본 논문에서 도입한 문맥의존 음향모델 작성법인 PDT-SSS 알고리즘의 유효성을 확인할 수 있었다. 향후에는 대량의 음성 데이터를 대상의 비교실험을 수행할 예정이다.

감사의 글

본 논문은 2000년도 한국과학재단 목적기초연구 (과제번호: R01-2000-000-00276-0) 지원으로 수행되었음.

참 고 문 헌

1. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc. 1993.
2. 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, ch. 3-4, 1988.
3. 박현상, 은종관, 박용규, 권오욱, "Diphone 단위의 hidden Markov model을 이용한 한국어 단어인식," 한국음향학회지, 13 (1), 14-23, 1994.
4. 이승훈, 김회린, "가변어휘 음성인식기의 음향모델 개선 및 성능분석," 한국음향학회지, 18 (8), 3-8, 1999.
5. 김유진, 김회린, 정재호, "인식 단위로서의 한국어 음절에 관한 연구," 한국음향학회지, 16 (3), 64-72, 1997.
6. 김효경, 구명완, "기본음소 설정을 위한 음소인식률 이용 방안 연구," 제15회 음성통신 및 신호처리 워크샵 논문집, 328-331, 1998.
7. J. Takami, and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. of ICASSP '92*, 1, 573-576, 1992.
8. M. Suzuki, S. Makino, A. Ito, H. Aso and H. Shimodaira, "A new HMnet construction algorithm requiring no contextual factors," *IEICE Trans. Info. & Syst.*, **E78-D** (6), 662-669, 1995.
9. Ostendoft and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, 11, 17-41, 1997.
10. 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*, 39-44, 2001.
11. K. Lee, S. Hayamizu, H. Hou, C. Huang, J. Swartz, and R. Weide, "Allophone clustering for continuous speech recognition," *Proc. of ICASSP '90*, 749-752, 1990.
12. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, 1997.
13. 김득수, 황철준, 정현열, "음성인식 기능을 가진 주소입력 시스템의 개발과 평가," 한국음향학회지, 18 (2), 3-10, 1999.
14. 오세진, 황철준, 김범국, 정호열, 정현열, "결정트리 상태 클러스트링에 의한 HM-Net 구조결정 알고리즘을 이용한 음성인식에 관한 연구," 한국음향학회지, 21 (2), 199-210, 2002.
15. 이호영, 국어음성학, 태학사, ch. 3-5, 1996.
16. 배주채, 국어음운론, 형설출판사, ch. 2-5, 15-76, 1995.

저자 약력

● 임 영 춘 (Young-Choon Lim)

2000년 2월: 제명대학교 물리학과 (이학사)
2002년 8월: 영남대학교 대학원 정보통신공학과 (공학석사)
2000년 8월~현재: 주식회사 자모바 연구원
※ 주관심분야: 음성분석 및 인식

● 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~2002년 12월: 대구과학대학 정보통신계열 전임강사
2002년 12월~현재: 한국천문연구원 KVN사업본부 그룹 선임연구원
※ 주관심분야: 디지털신호처리, 음성분석 및 인식, 상관기

● 김 광 동 (Kwang-Dong Kim)

1973년 2월: 영남대학교 전기공학과 (공학사)
1975년 5월~1982년 8월: 고미반도체(주) 기술과장
1982년 9월~1986년 7월: 대한통운(주) 전산실장
1986년 9월~1993년 4월: 제성전자(주) 기술부장
1993년 4월~현재: 한국천문연구원 책임연구원
※ 주관심분야: 디지털신호처리, 상관기, DSP 응용분야

● 노 덕 규 (Duk-Gyoo Roh)

1985년 2월: 서울대학교 천문학과 (이학사)
1994년 8월: 일본 동경대학 대학원 이학계연구과 천문학전공 (이학석사)
2002년 2월: 일본 동경대학 대학원 이학계연구과 천문학전공 (박사수료)
1984년 4월~현재: 한국천문연구원 KVN사업본부 선임연구원
※ 주관심분야: 디지털신호처리, DSP 응용 분야

● 송 민 규 (Min-Gyu Song)

2001년 2월: 강원대학교 전기공학과 (공학사)
2003년 2월: 강원대학교 대학원 전자공학과 (공학석사)
2002년 12월~현재: 한국천문연구원 KVN사업본부 그룹 연구원
※ 주관심분야: 디지털신호처리, DSP 응용분야

● 정 현 열 (Hyun-Yeol Chung)

한국음향학회지 제21권 4E호 참조