음소결정트리 상태분할을 이용한 한국어 연속음성인식에 관한 연구

오세진*, 황철준**, 김범국**, 정호열*, 정현열*
*영남대학교 전자정보공학부
**대구과학대학 정보전자통신계열

A Study on the Korean Continuous Speech Recognition using Phonetic Decision Tree-based State Splitting

Se-Jin Oh*, Cheol-Jun Hwang**, Bum-Koog Kim**, Ho-Youl Jung*, Hyun-Yeol Chung*
*School of Electrical Eng., & Computer Science, Yeungnam University
**Informational Electronics & Communications Div., Taegu Science College

요 약

본 연구에서는 연속음성인식 시스템의 성능개선을 위한 기초 연구로서 음소결정트리 상태분할과 한국어 음성학적 지식을 이용하여 문맥의존 음향모델의 작성방법을 검토하고, 한국어 연속음성인식에 적용을 소개한다. 음소결정트리 상태분할 알고리즘은 각 노드에서 한국어 음성학적 지식으로 구성된 음소 질의어 집합에 따라 2진 트리로 SSS(Successive State Splitting) 알고리즘에 의해 상태분할 하는 방법으로서 상태분할 후 각 상태를 네트워크로 연결한 구조를 HM-Net(Hidden Markov Network)이라 하며 문맥의존 음향모델로 표현된다. 작성한 문맥의존 음향모델의 유효성을 확인하기 위해 본 연구실의 항공편 예약 문장(YNU200)에 대해 연속음성인식 실험을 수행하였다. 인식실험결과, 문맥의존 음향모델에 대한 화자독립 연속음성인식률이 기존의 단일 HMM 모델보다 평균적으로 1-pass의 경우 9.9%, 2-pass의 경우 4.1% 향상된 인식률을 보였다. 따라서, 문맥의존 음향모델을 작성하는데 음소결정트리 상태분할과 한국어 음성학적 지식이 유효함을 확인하였다.

I. 서 론

HMM(Hidden Markov Model)은 단위 음성을 몇 개의 상태열로 정의하고 각 상태는 각 음성 세그멘트의 특징벡터가 발생하는 확률적 분포로 정의하는 방법으로서 높은 인식성능과 강건함 등과 같은 이유로 음성인식에서 널리 사용되고 있다. 음성인식의 인식단위로서 확장성, 훈련성, 대용량성 등을 고려하여 유사음소단위(PLUs: Phone-Likely Units)가 많이 이용되고 있다[1]. 이 경우, 유사음소는 조음결합 등의 영향으로 인해 하나의 모델로 표현하는데는 한계가 있다. 또한 모델의 파라미터 수가 많아져서 학습 데이터가 부족할 경우 각 모델의 학습 데이터 수가 균일하지 못한 경우가 발생하게 된다.

최근에는 이러한 문제점을 해결하기 위해 단일음소의 음향학적 특성을 변화시키는 환경요인으로 선행음소와 후행음소까지 고려한 문맥의존 모델로서 이음(allophone)을 인식 단위로 하는 방법이 소개되고 있으며 유효성이 확인되고 있다[2]. 문맥의존 모델은 음소의 평균 특징만을 가지며, 전후의 음소에 의존한 조음결합과 이음현상을 잘 표현할 수 있기 때문에 보다 정확한 음성인식을 가능하게 한다. 또한 문맥의존 모델의 유사한 상태의 확률분포를 공유하는 방법이 제안되고 있다[3]. 그러나 이음을 인식단위로 할 경우 단일음소와 비교하여 다음과 같은 두 가지 문제점을 해결할 필요가 있다. 첫째, 음소환경에 의존하는 단위를 이용할 경우 총 모델의 수를 고려한 환경요인의 종류에 따라 모델의 수가 지수 함수적으로 증가하게 된다. 만약 충분한 학습 데이터가 있다면 문제가 없지만, 불충분할 경우 각 모델에 대응하는 학습 데이터의 부족으로 모델 파라미터를 정확하게 추정할 수 없게 된다. 둘째, 유한한 학습 데이터를 사용하기 때문에 학습 데이터 중에 출현하지 않는 문맥정보가 존재할 수 있다.

이를 해결하기 위해, 전자의 경우 비슷한 파라미터를 가지는 HMM의 상태와 출력확률분포를 하나로 하여 상향(bottom-up)으로 공유를 하는 방법이 제안되었고[3], 모든 음소모델에 대응하는 상태공유를 자동으로 결정하는 SSS 알고리즘에 의해 작은 상태에서 보다 정확한 문맥의존 모델인 HM-Net을 자동적으로 생성할 수 방법이 제안되고 유효성이 입증되었다[4]. 후자의 경우 음소결정트리(Phonetic Decision Tree) 방법이 소개되었다[5]. 이 방법은 2진 트리 구조의 하향(top-down) 분할에 의해

음소간의 유사성에 기반한 음성학적 질의어를 통해 yes와 no의 분할을 수행한 후 학습 데이터의 문맥에 존재하지 않는 미지의 문맥을 작성하게 된다. SSS 알고리즘과 음소결정트리에 의한 상태분할의 차이점은 전자는 상태의 음향학적 분포확률 크기에 따라 분할할 상태를 결정하는 것이고, 후자는 질의어와 같은 특정 문맥정보에 따라 분할할 상태를 결정하는 것이다.

따라서, 본 연구에서는 한국어에 대해 미지의 문맥정보를 보다 정확하게 표현할 수 있는 문맥의존 음향모델을 작성하기 위해, SSS 알고리즘과 음소결정트리의 장점을 결합한 PDT-SSS(Phonetic Decision Tree based Successive State Splitting) 알고리즘을 이용하여, 본 연구실의 남성 4인이 발성한 항공편 예약 관련 200문장을 대상으로 연속음성인식 실험을 통해 알고리즘의 유효성을 확인하고자 한다.

II. SSS 알고리즘과 음소결정트리

2.1 SSS 알고리즘

SSS 알고리즘[4,8]은 모든 문맥을 나타내는 1상태의 초기모델로부터 문맥방향과 시간방향으로 상태분할 후 자동적으로 HM-Net[4,8]의 구조를 결정하는 알고리즘이다. SSS 알고리즘으로 HM-Net을 작성하는 단계를 그림 1에 나타내었다. 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.

2.2 음소결정트리

음소결정트리[5]는 음소의 음향적 변동을 파악하는 것으로, 미지 음소환경의 음향적 특성을 예측하는 방법이다. 음소결정트리는 뿌리(root)를 음소환경에 독립한 2진 트리로 나타내고 뿌리에서 잎 방향으로 문맥클래스의 분할을 수행한다(그림 2). 이 트리는 뿌리에서 잎 방향으로 진행함에 따라 음소환경의 의존도가 강한 단위를 나타내는 계층적 구조를 가지며, 일반적으로 잎 부분에 모델을 대응시키게 된다. 트리의 각 노드에서는 경험적으로 음소유사성에 기인한 질의어를 할당하여 yes와 no에 의해 문맥클래스를 두 개로 분할한다. 음소환경과 음소군에 따라서 각 질의어를 구성한다. 이러한 음소환경을 트리의 뿌리 노드에서 질의어를 찾아 반드시 잎에 대응시키기 위해, 미지의 음소환경에서 음향학적으로 가장 유사한 잎의 노드로 분류된다고 할 수 있다. 이를 위해, 출현하지 않는 음소환경을 음소환경 독립모델 등으로 대치할 필요도 있다.

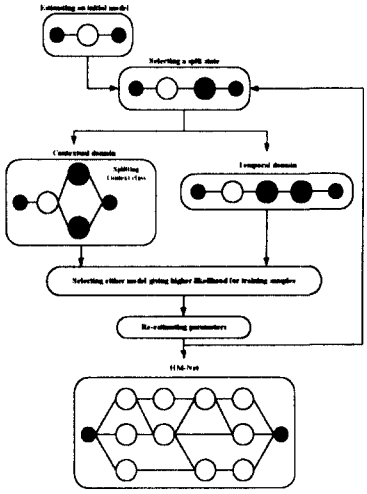

그림 1. SSS 알고리즘의 구성도

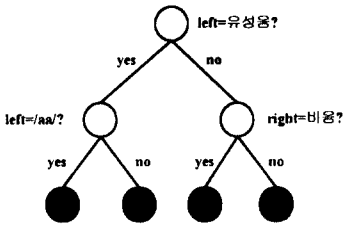

그림 2. 음소결정트리

III. 한국어 음성학적 지식

한국어에는 다른 언어와는 달리 많은 문법과 음운규칙이 있다. 본 연구에서는 한국어에 적합한 문맥의존 음향모델을 작성하기 위해 결정트리 기반 SSS 알고리즘의 상태분할에서 음소 질의어 집합의 구성에 한국어 음성학적 지식[7]을 이용하였다. 본 연구에서 적용한 음성학적 규칙은 크게 모음, 자음, 유성음, 비음, 유음, 반모음과 묵음으로 나눈다. 이 중에서 모음은 혀의 위치, 입의 크기, 혀의 높이, 좁힘점위치, 좁힘점간극 등과 같이 크게 5부

분으로 분류하였다. 그리고 자음은 조음자리, 조음방법 등과 같이 2부분으로 분류하고, 조음방법의 경우 파열음, 파찰음, 마찰음으로 다시 나누었다. 본 연구에는 음성학적 규칙을 문맥의 좌, 우를 포함하여 총 162부분으로 분류하였으며, 이를 이용하여 음소 질의어 집합을 작성하였다. 이렇게 작성한 음소 질의어 집합을 결정트리에 의한 상태분할에 사용하였다.

Ⅳ. PDT-SSS 알고리즘

본 연구에서는 한국어 음성학적 지식의 음소질의어에 의한 음소결정트리 상태분할과 SSS 알고리즘의 장점을 결합한 PDT-SSS(Phonetic Decision Tree-based SSS) 알고리즘[6]을 이용하였다. 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.

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

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

Ⅴ. 인식실험 및 고찰

음소결정트리 상태분할 알고리즘과 한국어 음성학적 지식을 이용하여 작성한 한국어 문맥의존 음향모델의 유효성을 확인하기 위해 연속음성인식 실험을 수행하였다. 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 학습하는데 사용하였으며, 학습에 참가하지 않은 4명의 200문장을 화자독립 연속음성인식 평가에 사용하였다.

모든 음성데이터는 16kHz의 샘플링과 16bits로 양자화되었으며, $1-0.97z^{-1}$의 전달함수로 프리엠퍼시스 하였으며, 25ms의 해밍 윈도우를 곱하여 10ms씩 이동하면서 분석하였다. 이를 통해 음성 특징 파라미터는 12차 LPC-멜 켑스트럼 계수와 정규화된 대수 에너지에 1차 및 2차의 차분 성분을 포함하여 총 39차의 특징 파라미터를 구하였다. 또한 PDT-SSS 알고리즘에 의한 문맥방향의 상태분할을 위해 162개(문맥의 좌, 우)의 음소 질의어 집합을 한국어 음성학적 지식에 근거하여 작성하였다. 초기 HM-Net의 구조는 48개의 유사음소단위를 병렬로 연결하여 141개의 상태를 가지도록 구성하였다. 모든 HM-Net은 혼합수 4를 가지며 200에서 1,200상태까지는 200상태씩 증가시켰으며, 상태수 2,000인 HM-Net도 학습하였다.

인식 알고리즘은 Multi-pass 탐색 알고리즘[3]으로서 1-pass 탐색의 경우, 단어 2-gram 언어모델을 이용하여 프레임 동기형 Viterbi beam 탐색을 수행한 후 단어 그래프를 출력한다. 2-pass 탐색의 경우 1-pass의 단어 그래프와 보다 정밀한 단어 3-gram을 이용하여 A* stack decoding 탐색을 수행한 후 인식결과를 출력한다.

그림 3에 상태수의 변화에 따른 화자독립 연속음성인식률을 나타내고, 그림 4에 인식 문장에 포함된 단어인식률을 각각 나타내었다.

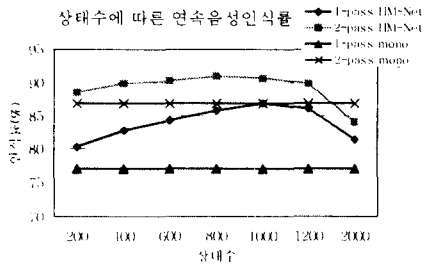

그림 3. 상태수에 따른 화자독립 연속음성인식률

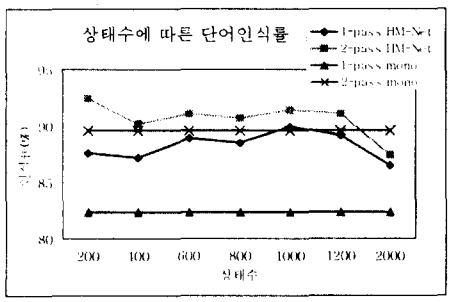

그림 4. 연속음성에 포함된 단어인식률

그림 3에서 상태수 1,000일 때 HM-Net triphone의 경우 1-pass의 인식률은 평균 86.9%로서 단일 HMM에 비해 평균 9.9%의 인식률을 향상을 보이고, 상태수 800일 때 HM-Net triphone의 경우 2-pass의 인식률은 평균 90.9%로서 단일 HMM에 비해 평균 4.1%의 인식률을 향상을 보였다. 또한 그림 4에서 인식대상인 연속음성에 포함된 798단어에 대한 인식률은 상태수 1,000일 때 HM-Net triphone의 경우 1-pass 인식률은 평균 89.9%로서 단일 HMM에 비해 평균 7.6%의 인식률 향상을 보이고, 상태수 200일 때 HM-Net triphone의 경우 2-pass의 인식률은 평균 92.4%로서 단일 HMM에 비해 평균 2.8%의 향상된 인식률을 구하였다.

그리고 상태수의 증가에 따라 연속음성인식률과 단어인식률이 감소하는 원인으로는 학습에 참가한 음성 데이터의 부족으로 인해 정확한 HM-Net이 생성되지 못한 것으로 생각된다. 이는 향후 음향모델을 작성하는데 많은 양의 음성 데이터를 사용할 경우 해결할 수 있을 것으로 기대된다. 이상의 결과로부터 본 연구에서 문맥의존 음향모델을 작성하기 위해 적용한 음소결정트리 상태분할과 한국어 음성학적 지식이 유효함을 확인할 수 있었다.

Ⅵ. 결 론

본 연구에서는 연속음성인식 시스템의 성능개선을 위한 기초 연구로서 음소결정트리 상태분할과 한국어 음성학적 지식을 이용하여 문맥의존 음향모델의 작성방법을 검토하였다. 작성한 문맥의존 음향모델의 유효성을 확인하기 위해 본 연구실의 항공편 예약 문장(YNU200)에 대해 연속음성인식 실험을 수행한 결과, 문맥의존 음향모델에 대한 화자독립 연속음성인식률이 기존의 단일 HMM 모델보다 평균적으로 1-pass의 경우 9.9%, 2-pass의 경우 4.1% 향상된 인식률을 보였으며, 단어인식률은 1-pass의 경우 평균 7.6%, 2-pass의 경우 평균 2.8%의 향상된 인식률을 보여 문맥의존 음향모델을 작성하는데 음소결정트리 상태분할과 한국어 음성학적 지식이 유효함을 확인하였다.

※본 연구는 BK21정보기술사업의 2001년도 연구비에 의해 수행되었음

참 고 문 헌

[1] 김범국, 정현열, "가변장 음소모델을 이용한 음소인식," 한국음향학회지, 제16권, 제8호, 1997.

[2] K.F. Lee, S. Hayamizu, H.W. Hon, C. Huang, J. Swartz, R. Weide, "Allophone Clustering for Continuous Speech Recognition," Proc. of ICASSP'90, pp. 749-752, 1990.

[3] S.J. Young, P.C. Woodland, "State Clustering in hidden Markov model-based Continuous Speech Recognition," Computer Speech and Language, Vol. 8, No. 4, pp. 369-383, 1994.

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

[5] L.R. Bahl, P.V.de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, "Dicision Trees for Phonological Rules in Continuous Speech," Proc. of ICASSP'91, pp. 185-188, 1991.

[6] S.J.Oh, C.J.Hwang, B.K.Kim, H.Y.Jung, H.Y.Chung, "A Study on Speech Recognition using New State Clustering Algorithm of HM-Net with Korean Phonological Rules," Proc. of IC-AI'2001, U.S.A, 2001. 6.

[7] 이호영, "국어음성학," 태학사, 1996.

[8] 오세진, 임영춘, 황철준, 김범국, 정현열, "Hidden Markov Network를 이용한 음향학적 음소모델 작성에 관한 검토," 2000년도 한국음향학회 학술발표대회 논문집, 제19권 제2(s)호, pp. 29-32, 2000. 11.