2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

음소 모델의 Back-Off 기법을 이용한 어휘독립 음성인식기의 성능개선

구동욱°, 최준기 오영환
한국과학기술원 전자전산학과 전산학전공

Performance Improvement of Vocabulary Independent Speech Recognizer using Back-Off Method on Subword Model

Dong-Ook Koo°, Joon Ki choi, Yung-Hwan Oh
Division of Computer Science
Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology
e-mail : {dokoo, jkchoi}@bulsai.kaist.ac.kr , yhoh@cs.kaist.ac.kr

요약

어휘독립 음성인식이란 음향학적 모델 훈련에 사용하지 않은 어휘들을 인식하는 것이다. 단어모델을 이용한 어휘독립 음성인식 시스템은 발음표기로 변환된 인식대상어휘에 대하여 문맥 종속형 부단어(context dependent subword) 단위로 훈련된 모델을 연결하여 단어 모델을 만들고 이 단어 모델로 인식을 수행한다. 이러한 시스템의 경우 훈련과정에서 나타나지 않는 문맥 종속형 부단어가 인식대상어휘에서 나타나게 되고, 따라서 정확한 단어모델을 구성할 수 없다는 문제점이 있다. 본 논문에서는 문맥 종속형 부단어 구분의 계층화를 통한 back-off 선택 방법을 이용하여 새롭게 나타난 문맥 종속형 부단어 대신 연결될 부단어 모델을 찾아내는 방법을 제안한다. 제안된 선택 방법은 새롭게 나타난 문맥 종속형 부단어를 포함하는 상위의 부단어를 찾아내는 방법이다. 실험 결과 10단어 세트에서 97.5% 50단어 세트에서 90.16% 100 단어 세트에서 82.08%의 인식률을 얻었다.

1. 서론

어휘독립 음성인식이란 미리 정해지지 않았거나 수시로 변하는 인식대상음성을 인식하는 것이다. 따라서 부단어(subword) 단위로 음성을 모델링하고 이러한 부단어 단위의 모델들을 연결하여 인식대상어휘를 모델링하여 음성을 인식하는 것이다[1]. 일반적으로 부단어 단위는 음소 단위를 널리 사용한다. 음성인식의 성능을 높이기 위하여 단어내 혹은 단어간의 조음현상을 고려한 문맥 종속형 부단어(context dependent subword) 단위로 음소를 모델링 한다. 이러한 문맥 종속형 부단어로 많이 쓰이는 것이 biphone 또는 triphone 모델이다. 어휘독립 음성인식 시스템에서 음향학적 문맥 종속형 부단어 모델 훈련을 위한 음성 데이터베이스는 가능한 모든 문맥 종속형 부단어를 포함해야 하고 충분한 음운현상을 포함해야 한다. 이러한 음성 데이터베이스를 수집하는 것은 많은 시간이 들고 데이터베이스의 크기가 커지기 때문에 현실적으로 어려운 일이다.

적합한 음성 데이터베이스 수집의 어려움 때문에 음향학적 모델 훈련 과정에서 나타나지 않은 음소 모델이 인식대상어휘에서 나타나는 unseen model 문제가 발생하게 된다[2][3]. 이러한 문제의 해결을 위하여 음운학적으로 비슷한 성질의 triphone들을 하나의 triphone으로 군집화한 generalized triphone을 사용하는 방법이 제안되었다[3][4][5]. 그러나 데이터의 부족현상이 근본적으로 해결된 것이 아니므로 unseen model 문제가 완전히 해결되는 것은 아니다.

본 논문에서는 back-off 선택 방법을 이용하여 새롭게 나타난 문맥 종속형 부단어 대신 연결될 부단어 모델을 찾아내는 방법을 제안한다. back-off 방법이란 monophone, biphone, triphnone 단위로 각각의 음소 모델을 훈련하고, 요구되는 triphone 모델이 훈련된 모델들 중에 없을 경우 그 triphone을 포함하는 biphone을, biphone이 없을 경우 그 biphone을 포함하는 mono-phone을 대신 사용하는 방법이다.

본 논문의 구성은 다음과 같다. 2장에서는 어휘독립 음성인식 시스템의 구성에 대하여 설명하고, 3장에서는 본 논문에서 제안한 back-off 방법에 대하여 설명한다. 그리고 4장에서 제안된 방법에 의한 실험과 결과에 대해서 논의한 후, 5장에서 결론과 앞으로 할 일에 관해 언급한다.

2. 어휘독립 음성인식 시스템

단어 모델을 이용한 어휘독립음성인식 시스템은 부단어 단위로 음향학적 모델을 훈련하고, 인식대상어휘의 발음 표기에 따라 해당되는 부단어 모델을 연결하여 단어 모델을 만들고 이렇게 만들어진 단어 모델을 통하여 음성을 인식하는 시스템이다. 그림 1은 어휘독립 음성인식 시스템의 블럭도이다.

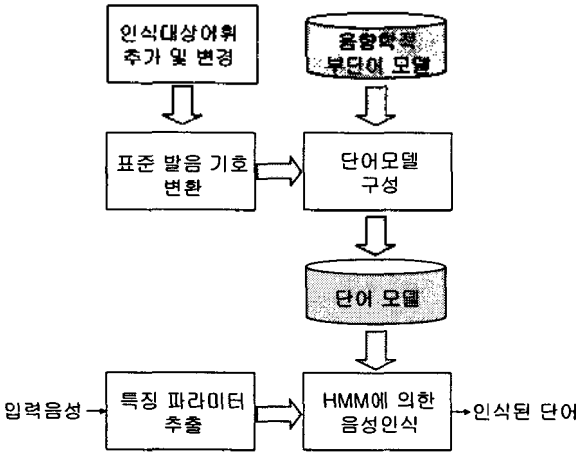

그림 1 어휘독립 음성인식 시스템

2.1 음소 모델링

음소 모델은 훈련용 음성 데이터가 충분한 경우 그 음소의 좌우 음운 현상을 고려한 문맥 종속형 음소 모델을 이용하는 것이 문맥 독립형 음소 모델을 이용하는 것보다 성능이 우수한 것으로 알려져 있다[7]. 그러나 문맥 종속형 음소모델을 사용할 경우 그 모델의 수가 너무 많아지는 문제점이 발생한다. 한국어의 경우 최소 38개의 음소로 나눈다면 모든 종류의 triphone을 모델링 하려면 그 수가 38*38*38 = 54872 이고 음성학적으로 불가능한 것을 제외한다면 약 25000 이다[2]. 이처럼 많은 수의 triphone에 대하여 모델 파라미터를 추정하기 위해서는 방대한 양의 훈련용 음성 데이터 베이스가 있어야 한다. 그러나, 이러한 음성 데이터베이스를 수집하는 일은 현실적으로 매우 어려운 일이다. 따라서 음향학적 모델 훈련 과정에서 나타나지 않은 음소 모델이 인식대상어휘에서 나타나는 unseen model 문제가 발생하게 된다.

이러한 문제의 해결을 위하여 음운학적 지식에 기반한 트리 기반 군집화 방법이 제안되었다[4]. 이 방법은 음운학적으로 비슷한 성질을 나타내는 문맥을 하나의 문맥으로 군집화 하는 방법으로 이진 결정트리를 이용하는 방법이다. 즉, 하나의 음소에 대한 데이터를 하나의 집합으로 만들고 그 집합을 음운학적인 질문에 대하여 두 개로 나눈다. 계속하여 재귀적으로 집합을 두 개로 나누어 가는 방식을 말한다. 부분집합으로의 분할은 각 집합에 해당하는 노드에서 문맥에 대한 가부를 묻는 질문과 그 질문에 대한 평가함수 그리고 분할을 언제 멈추어야 하는지에 관한 기준이 있어야 한다. 그림 2는 음소결정트리에 의한 음소 'ㅏ' 의 군집화를 나타낸다.

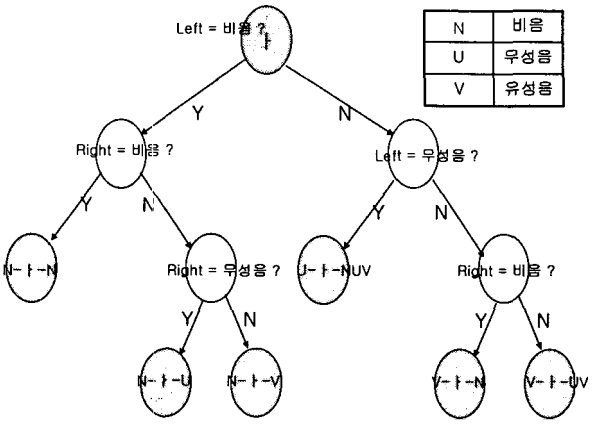

그림 2 음소결정트리에 의한 음소 'ㅏ'의 군집화

2.2 단어 모델링

인식대상어휘를 추가하거나 변경할 때, 표준발음 전사기를 통하여 인식대상어휘의 표준 발음기호열을 만들어 낸다. 각각의 어휘에 대한 발음기호열로 발음사전을 만들고 발음사전을 기반으로 해당되는 음소모델을 연결하여 단어모델을 만든다.

그림 3은 훈련된 음소 모델을 연결하여 만들어진 단어모델의 예이다. 그림에서와 같이 각 음소간의 천이확률은 마지막 상태에서 종료되는 천이확률을 사용한다.

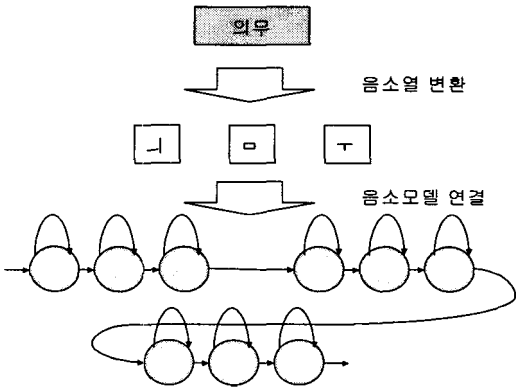

그림 3 음소모델을 연결하여 구성된 단어모델

3. Back-off 기법

2장에서와 같이 음소결정트리에 의해 군집화 된 generalized triphone을 사용할 경우 훈련용 음성데이터에서 나타나지 않는 triphone이 인식대상어휘에서 나타난다면 결정트리에서 지정해주는 triphone으로 사상된다. 그러나 결정트리는 음운학적인 질문들로 구성되어있기 때문에 실제 데이터에서 나타나는 음향학적 특성들을 반영하고 있지 않다. 따라서 결정 트리에만 의존하여 군집화 된 generalized triphone을 이용하는 것에는 문제가 있다.

본 논문에서는 이러한 문제를 해결하기 위하여 back-off 방법을 제안한다. back-off 방법이란 unseen model을 대신할 상위의 모델을 찾는 방법이다. 이 방

법을 적용하기 위하여 음향학적 음소모델을 triphone, biphone, monophone 단위로 각각 훈련한다. 단어 모델을 만들기 위하여 먼저 triphone 모델을 찾아서 연결하고, unseen model에 대하여 적합한 biphone 또는 mono phone을 찾아서 연결한다. ,

3.1 문맥 종속형 음소의 군집화

제안한 방법에서 사용하는 이진트리의 구성은 다음과 같다. 이진 질문들의 집합을 Q라 하고 n을 트리에서의 노드 그리고 m(q,n)을 노드 n에서의 분할에 대한 평가 함수라 하고 단말노드에서 데이터의 개수를 c(n)이라 한다. 각각의 노드에서 m(q,n)을 계산하고 m(q,n)이 가장 큰 값을 나타내는 노드에서 데이터집합이 분할된다. 이때 종료 조건 없이 모든 분할을 허용한다. 트리의 구성이 끝나고 단말노드에서 데이터집합 원소의 개수에 의하여 훈련할 음소모델을 결정한다. 표 2는 본 논문에서 제안된 back-off 방법을 적용하기 위하여 고안된 음소결정트리의 구성이다.

1. 루트노드에서 모든 데이터를 포함한다.

2. 검사하지 않은 노드가 없을 때까지 다음을 계속한다.
   1.검사하지 않은 노드를 선택한다.
   2.m(q,n)을 가능한 모든 q에 대하여 계산한다.
   3.m(q,n)의 값이 가장 큰 값에서 두개의 자식노드로 분할한다.

3. 단말노드에서 c(n)이 한계값 보다 작으면 그 노드를 훈련에서 제외한다.

표 1 back-off 방법을 위한 음소결정트리의 구성

그림 3은 제안된 방법의 음소결정트리에 의해 음소 'ㅏ'에 대한 triphone을 군집화한 예이다. 그림 3에서 보는 바와 같이 흑색으로 표시된 단말노드들은 c(n)의 값이 한계값 보다 작기 때문에 음향학적 모델 훈련에서 제외 시킨다. 기존방법에서는 음소결정트리 구성도중에 단말노드가 결정되기 때문에 단말노드에 포함된 데이터들 상호간에 변이가 크다. 그러나 제안된 방법에서는 이러한 단점이 극복된다.

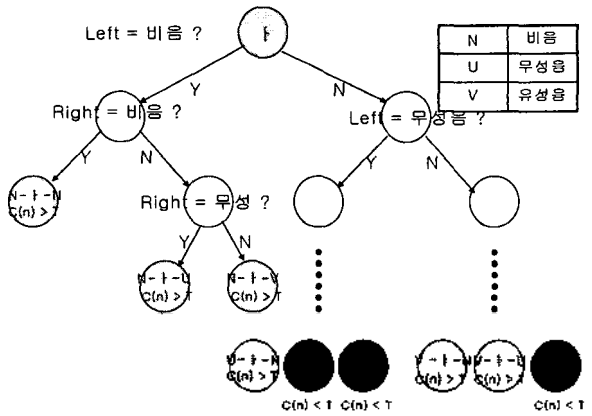

그림 3 제안된 방법의 음소 'ㅏ'에대한 음소결정트리

3.2 back-off 방법의 적용

본 논문에서 제안한 back-off 방법에서는 triphone 모델, biphone 모델, monophone 모델을 각각 훈련해야 한다. triphone 모델은 앞서 설명한 결정트리의 단말노드에 포함된 데이터집합의 사용하여 훈련하고, biphone 모델은 결정트리의 내부노드에 포함된 데이터집합을 이용하여 훈련한다. 그리고 monophone model은 결정트리의 루트노드에 포함된 데이터 집합을 이용하여 훈련한다.

위에서와 같이 각각의 음소모델을 훈련한 다음 단어모델을 만든다. 표준 발음 기호 변환기를 통하여 만들어진 발음기호열이 있을 때 우선 적으로 연결할 triphone 모델을 결정트리 단말노드에서 찾아내고 이를 연결한다. 만일 triphone 모델이 없다면 그 단말노드의 조상노드를 검색하여 적당한 biphone 모델을 찾아내고 이를 연결한다. 만일 요구하는 biphone 모델도 존재하지 않을 경우 루트노드에서 monophone 모델을 연결한다.

본 논문에서 제안한 back-off 방법을 사용하면 이미 존재하는 모델에 관해서는 정밀한 조음현상을 포함한 모델을 그대로 사용할 수 있다. 그리고 unseen model이 요구되는 경우에 있어서도 음향학적특징 분포를 따르는 모델을 사용하게 된다. 따라서 정밀한 조음형상이 유지되고 unseen model 문제도 합리적으로 처리할 수 있다.

4. 실험 및 결과

인식 실험을 위한 음성 데이터베이스는 한국전자통신연구소에서 만든 한국어 고립단어 445단어 음성 데이터베이스 이다. 이 음성 데이터베이스는 한 세트가 445개의 단어를 2번씩 발성한 것으로 구성되어 있고 총 14세트를 실험에 사용하였다. 훈련과정에서 8세트를 사용하였고, 인식과정에서는 6세트를 사용하였다.

어휘독립을 위하여 445개의 단어를 훈련과 인식을 위한 단어집합으로 나누어야 한다. 우선 445개의 단어들 중에서 인식을 위한 단어 100개를 골라내고 그 나머지 단어들 345개를 음향학적 음소 모델 훈련을 위하여 사용하였다. 인식을 위한 100단어를 선택하기 위하여 엔트로피를 최대화하는 반복 알고리즘을 사용하였다[6].

음소 훈련을 위하여 345단어 1세트를 수작업으로 음소단위로 분할하였고 나머지 7세트는 수작업으로 분할된 단어와의 동적 시간 정합 방법을 이용하여 분할하였다.

음향학적 모델로서 DHMM을 사용하였다. 각각의 음소당 3개의 상태를 부여하였다. 사용된 특징 파라미터로는 15차 LPC 멜켑스트럼과 15차 델타멜켑스트럼을 각각 256 코드워드로 벡터 양자화한 것을 사용하였다.

본 논문에서는 두 가지의 간단한 음소결정트리의 이진 질문 세트를 사용하였고, 각각에 대하여 같은 실험을 수행하였다.

4.1 파라미터특성을 고려한 문맥

triphone을 군집화 하는 음소결정트리의 구성은 다음과 같다.

Q = { left 비음, left-무성음, left-유성음,
right 비음, right-무성음, right-유성음 }
m(q,n) = 질문을 만족하는 데이터의 개수

위의 결정트리로 음소를 군집화하였고, 제안한 방법에 의한 음소 연결방법과 단지 결정트리 만을 사용한 경우를 비교 실험하였다. 10개의 단어부터 100개의 단어까지 어휘수를 증가시키면서 실험하였다. 그림 5는 인식률의 추이를 나타내고 있다.

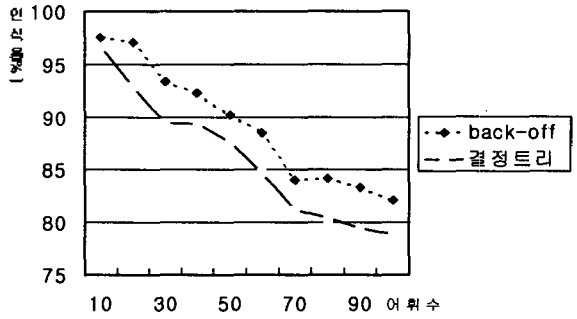

그림 5. 인식 실험 결과

그림에서 실선은 결정트리 만을 이용하여 실험한 결과이고 점선은 제안된 back-off 방법을 이용하여 실험한 결과 이다. 보는 바와 같이 제안된 방법을 적용하는 것이 기존의 결정트리 만을 이용한 방법보다 인식 성능이 더 높음을 알 수 있다.

4.2 음운론적특징을 고려한 문맥

triphone을 군집화 하는 음소결정트리의 구성은 다음과 같다.

Q = { left-전설모음, left-후설모음, left-혀끝소리,
left-입술소리, left-구개음, left-비음,
right-전설모음, right-후설모음,
right-혀끝소리, right-입술소리,
right-구개음, right-비음 }
m(q,n) = 질문을 만족하는 데이터의 개수

실험 1과 마찬가지로 10단어부터 100단어까지 인식 대상 어휘를 늘려가면서 실험하였다. 그림 6은 인식률의 추이를 나타내고 있다.

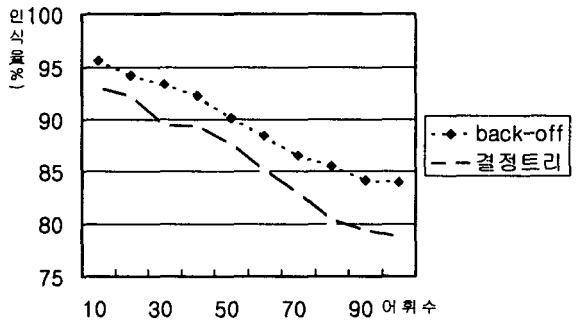

그림 6. 인식 실험 결과

위의 실험 결과를 살펴보면 두 경우 모두 제안된 방법에 의한 경우가 기존의 결정트리 만을 이용한 방법에 비하여 높은 성능을 나타내는 것을 알 수 있다. 이는 제안한 방법이 정밀한 조음현상을 유지와 음향학적특징 분포에 따른 unseen model 처리를 모두 고려하고 있기 때문이라고 판단된다.

5. 결론

본 논문에서는 어휘독립 음성인식 시스템에서 나타나는 unseen model 문제에 대처하는 방안으로 back-off 방법에 기반한 음소 선택방법을 제안한다. 이러한 방법으로 실험한 결과 기존의 음소결정트리에만 의존한 방법에 비하여 평균적으로 3.12%의 인식률 향상을 보였다. 따라서 어휘독립 음성인식 시스템의 성능향상에는 음운학적 음소 세분화와 함께 음향학적 특징의 반영이 중요한 요소로 작용함을 알 수 있었다. 현재는 이러한 음향학적 특징을 반영하는 보다 체계적이고 수학적인 방법에 관하여 연구가 진행중이다.

Reference

[1] Rafid A. Sukkar and Chin-Hui Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition," IEEE Trans. Speech and Audio processing Vol.4, No.6 pp. 420~429 Nov. 1996.

[2] 황병한, "한국어 가변어휘 인식을 위한 음소 모델링 방법에 관한 연구", 석사 학위 논문, 부산대학교 1999.

[3] Lynn C. Wood, David J. B. Pearce and Frederic Novello, "Improved Vocabulary-Independent Sub-Word HMM Modeling," Proc. Int. Conf. On Acoustics, Speech and Signal Processing, pp. 181~184, 1991.

[4] R. K. Moore, M. J. Russell, S. N. Downey and S. R. Browning, "A Comparison of Phoneme Decision Tree and Context Adaptive Phone Based Approaches To Vocabulary-Independent Speech Recognition,"Proc. Int. Conf. On Acoustics, Speech and Signal Processing, pp. I-541~I-544, 1994.

[5] L. R. Bahl, P. V. desouza, "Decision Tree for Phonological Rules in Continuous Speech," Proc. Int. Conf. On Acoustics, Speech and Signal Processing, pp. 185~188, 1991.

[6] 오영환, "음성인식을 위한 잡음 처리기술에 관한 연구", 전자통신연구소 중간보고서, 1995.

[7] 윤성진, "확률 발음 사전을 이용한 대어휘 연속 음성 인식", 박사 학위 논문, KAIST, 1999.