2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

음소 군집화 기법을 이용한 어휘독립음성인식의 음소모델링

구동욱°, 최준기, 윤영선, 오영환
한국과학기술원 전자전산학과 전산학전공

Subword Modeling of Vocabulary Independent Speech Recognition Using Phoneme Clustering

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

요약

어휘독립 고립단어인식은 미리 훈련된 부단어(subword) 단위의 음향모델을 이용하여 수시로 변하는 인식대상어휘를 인식하는 것이다. 본 논문에서는 소용량 음성 데이터베이스를 이용하여 어휘독립음성인식 시스템을 구성하였다. 소용량 음성 데이터베이스에서 미관측 문맥 종속형 부단어에 대한 처리에 효과적인 백오프 기법을 이용한 음소 군집화 방법으로 문턱값을 변화시키며 인식실험을 수행하였다. 그리고 훈련용 데이터의 부족으로 인하여 문맥 종속형 부단어 모델이 훈련용 데이터 베이스로 편중되는 문제를 deleted interpolation 방법을 이용하여 문맥 종속형 부단어 모델과 문맥 독립형 부단어 모델을 병합함으로써 해결하였다. 그 결과 음성인식의 성능이 향상되었다.

1. 서론

어휘독립 음성인식이란 미리 정해지지 않았거나 수시로 변하는 인식대상음성을 인식하는 것이다. 즉 부단어(subword) 단위로 음성을 모델링하고 이러한 부단어 단위의 모델들을 연결하여 인식대상어휘를 모델링 하여 음성을 인식하는 것이다[1]. 일반적으로 부단어 단위는 음소 단위를 널리 사용한다. 음성인식의 성능을 높이기 위하여 조음현상을 고려한 문맥 종속형 부단어(context dependent subword) 단위로 음소를 모델링 한다. 이는 각각의 음소에 대한 음향학적 해상도를 높게 하기 때문에 음성인식의 성능이 향상된다. 어휘독립 음성인식 시스템에서 문맥 종속형 부단어 모델 훈련을 위한 음성 데이터베이스는 가능한 모든 문맥 종속형 부단어를 포함해야 하고 충분한 음운현상을 포함해야 한다. 그러므로 어휘독립음성인식을 위해서는 대용량의 음성 데이터베이스가 요구된다. 그러나 이러한 대용량의 음성 데이터베이스를 수집하는 것은 많은 시간과 비용이 들게 되고, 이러한 요인 때문에 시스템의 개발이 상대적으로 어려워 진다.

어휘독립 음성인식시스템의 개발에 소용량의 음성 데이터베이스를 사용하게 되면 음성 데이터베이스 수집에 드는 시간과 비용을 절약할 수 있으므로 개발이 보다 용이하게 된다. 그러나 이러한 경우 훈련을 위한 음성 데이터베이스는 모든 종류의 문맥 종속형 음소를 포함할 수 없고 또한 각각의 문맥 종속형 음소의 출현 빈도수도 음향학적 모델 훈련을 위한 최소한의 빈도수 보다 작게 되는 경우가 빈번하게 발생한다. 따라서 음향학적 모델 훈련 과정에서 모델링이 불가능한 음소 모델이 인식대상어휘에서 나타나는 미관측 모델(unseen model) 문제가 발생하게 된다[2][3][8]. 이러한 문제를 해결하기 위하여 비슷한 문맥정보를 가지는 음소들을 하나의 음소로 군집화하는 방법이 연구되어왔다[3][5][8]. 그러나 군집화 방법을 사용하여도 미관측 모델에 대한 문제가 완전히 해결되는 것은 아니다. 소용량의 음성 데이터베이스를 사용하는 경우에는 데이터의 문맥정보 뿐만 아니라 훈련에 필요한 데이터의 양이 많이 부족하기 때문이다. 그러므로 훈련된 문맥 종속형 음소 모델에 포함된 문맥정보가 쉽게 관측 문맥쪽으로 편향되게 되고 이로 인하여 음성인식의 성능이 오히려 저하되는 현상이 발생하게 된다.

본 논문에서는 소용량 음성 데이터베이스를 이용하여 어휘 독립음성인식 시스템을 구성하였다. 문맥 종속형 음소 모델을 백오프 군집화 방법[8]에 의하여 군집화 하였고, 데이터의 부족으로 인하여 음소모델이 훈련용 데

이터에 편중되어 훈련되는 현상을 deleted interpo-lation 방법을 사용하여 문맥 독립형 음소모델과 문맥 종속형 음소모델을 병합함으로써 해결하였다[9].

본 논문의 구성은 다음과 같다. 2장에서는 어휘독립 음성인식 시스템의 기본 구성에 대하여 설명하고, 3장에서는 음소모델의 군집화와 군집화된 음소군의 최적화와 deleted interpolation 방법에 의한 모델 평활화(smoothing)에 대하여 설명한다. 그리고 4장에서 실험과 결과에 대해서 논의한 후, 5장에서 결론과 앞으로 할 일에 관해 언급한다.

2. 어휘독립음성인식 시스템

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

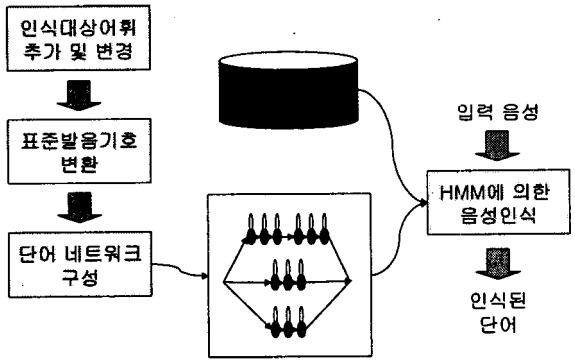

그림 1 어휘독립 음성인식 시스템

2.1 문맥 종속형 음소 모델링

어휘독립음성인식을 위해서는 음성신호를 부단어 단위로 모델링 하여야 한다. 부단어 단위로 널리 사용되고 있는 것이 음소 단위이다. 음향학적 모델의 해상도를 높여서 음성인식의 성능을 향상하고자 음소 각각의 좌우 문맥을 고려한 문맥 종속형 음소모델이 사용되고 있다. 문맥 종속형 음소 모델에는 음소 좌측 또는 우측의 문맥만을 고려한 바이폰 모델과 좌측과 우측 문맥을 모두 고려한 트라이폰 모델이 있으며 또한 그 이상의 문맥을 고려하는 모델들도 있다. 고려되는 문맥이 길어지면 길어질수록 반영되는 문맥정보가 많아 지게 되고 이러한 모델을 사용한 음성인식의 성능 또한 높아지게 된다. 일반적으로 문맥 종속형 음소모델로는 트라이폰 모델이 널리 사용되고 있으며 그 성능 또한 우수한 것으로 알려지고 있다[7].

트라이폰을 음성인식의 기본 단위로 사용할 경우 훈련용 음성 데이터베이스가 해당언어에서 존재 가능한 모든 트라이폰을 포함하고 있어야 한다. 그러나 가능한 모든 트라이폰을 모두 포함한 음성 데이터베이스의 구축은 현실적으로 불가능하다. 따라서 트라이폰 단위의 음소모델링을 위해서는 음향학적으로 또는 음운학적으로 비슷한 트라이폰들을 하나의 트라이폰으로 묶어주는 군집화 작업이 필요하다.

특히 소용량의 음성 데이터베이스를 이용한 어휘독립 음성인식 시스템에서는 훈련하고자 하는 트라이폰 모델에 해당하는 음성 데이터의 수가 모델 훈련에 필요한 최소한의 데이터 수 보다 적게 출현하는 경우가 빈번하다. 이 경우 트라이폰 모델의 훈련이 불가능하게 되고 반드시 트라이폰의 군집화 작업을 수행하여야 한다.

트라이폰의 군집화에는 음소결정트리가 이용된다[5][8]. 음소결정트리를 이용한 트라이폰의 군집화 작업은 트리의 내부 노드에서 음소 각각의 좌우 문맥에 대한 질문을 하고 질문에 대하여 만족하는 음소 군과 만족하지 않는 음소 군으로 나누어 가는 작업이다. 이러한 일련의 작업을 마치면 트리의 단말노드에는 훈련 가능한 수의 군집화된 트라이폰들이 모이게 되고 이를 훈련하여 트라이폰 모델이 만들어진다.

트리의 단말노드에서만 훈련하도록 트리를 구성할 경우 모든 단말노드를 훈련 시킬 음성 데이터가 부족하기 때문에 음성 데이터에 존재하는 문맥정보를 전부 반영할 수 없는 문제점이 있다. 이러한 문제점을 보완하고 각 음소의 문맥정보를 최대한 활용하기 위하여 위하여 백오프(back-off) 기법을 이용한 군집화 방법이 제안 되었다[8]. 백오프 기법을 이용한 군집화 방법은 트리의 단말노드뿐만 아니라 내부 노드에서도 모델 훈련 이 가능하도록 구성 되므로 문맥정보를 최대한 이용할 수 있고, 또한 음향학적 모델의 해상도도 최대한 높일 수 있다는 장점이 있다.

2.2 단어 Lexicon의 구성

인식대상어휘를 추가하거나 변경할 때, 표준 발음 표기 변환기를 통하여 인식대상어휘의 표준 발음기호열을 만들어낸다. 각각의 어휘에 대한 발음기호열로 발음사전을 만들고 발음사전을 기반으로 해당되는 음소모델을 연결하여 단어 네트워크를 구성한다. 단어 네트워크의 구성은 각 단어의 발음 기호에 해당하는 음소모델을 직렬로 연결하고 좌우 무음구간을 첨가하여 구성한다.

3. 음소군의 군집화와 평활화

3.1 백오프 기법을 이용한 음소 군집화

백오프 기법을 이용한 음소 군집화 방법은 결정트리를 이용한 음소 군집화 작업을 기반으로 한다. 기존의 결정트리를 이용하는 방법에서는 모델의 훈련이 단말노드에서만 일어나는 반면 백오프 기법을 이용한 방법에서는 트리의 내부노드에서도 모델이 훈련된다. 이 방법을 이용할 경우 각 음소의 문맥정보를 최대한 활용할 수 있고 또한 미관측 문맥 종속형 음소가 인식단계에서 요구될 경우 이를 처리하기가 용이하다.

그림 2는 백오프 기법을 이용한 음소결정트리의 생성 알고리즘을 나타내고 있다. 음소집합의 분할에 대한 제한을 가하지 않고 음소집합의 분할이 완료된 후에 훈련 가능한 노드와 그렇지 않은 노드를 단말 노드에서부터 부모노드로 거슬러 올라가면서 판단하게 된다.

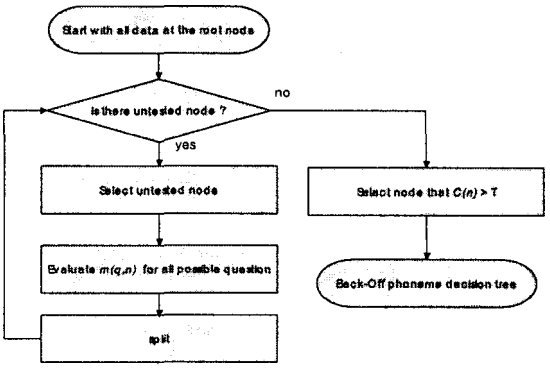

그림 2 백오프 음소결정트리 생성 알고리즘

그림 3은 음소 'ㅇ'에 대한 백오프 기법을 이용한 음소 결정트리의 예이다. 그림에서 회색노드가 훈련 되는 노드이다. 흰색 노드는 훈련되지 않는 노드이고 검은색 노드는 훈련 데이터의 양이 문턱값보다 작으므로 훈련되지 않고 버려 지는 노드이다. 모든 내부노드에서 분할이 허용되므로 가능한 한 많이 분할되고 단말노드에서부터 거슬러 올라가면서 훈련 가능 여부를 판단한다. 그림에서 점선 표시가 있는 노드가 내부노드이면서 훈련되는 노드이다.

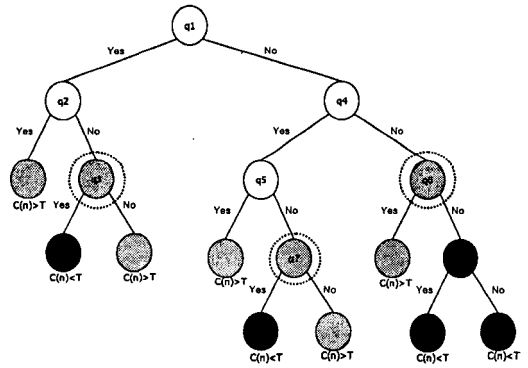

그림 3 음소 'ㅇ'의 음소결정트리

3.2 음소군의 최적화

백오프 기법을 이용한 음소 군집화 방법을 사용할 경우 단말노드와 함께 내부노드에서도 모델이 만들어지게 되고 모델의 수가 상대적으로 많아지게 된다. 모델의 개수는 음성인식의 성능향상에 매우 중요한 요소로서 작용하기 때문에 적절한 모델의 개수가 고려되어야 한다.

모델의 수를 고려하기 위하여 본 논문에서는 백오프 기법을 이용한 음소결정트리를 구성하고 단말노드에서의 해당하는 훈련용 음성데이타의 개수를 문턱값으로 삼아 문턱값을 변화시키면서 실험하였다. 문턱값이 커지면 트리의 깊이가 얕아지므로 많은 문맥정보를 포함할 수 없다. 반면에 데이터의 양이 많아지므로 보다 안정된 모델을 훈련할 수 있다. 반면에 문턱값이 작아지면 보다 세분화된 음소모델을 얻을 수 있고 보다 많은 문맥 정보를 이용할 수 있는 반면 데이터의 양이 적어지므로 훈련데이터에 편중된 모델이 얻어지게 된다.

3.3 모델의 평활화

문맥 종속형 음소모델을 훈련시키려면 충분히 많은 양의 음성 데이터가 존재하여야 한다. 소용량의 음성 데이터베이스를 사용하는 경우 음소 군집화에 의하여 군집화하여도 음성 데이터의 양이 적기 때문에 다양한 음운 현상을 충분히 포함하는 모델을 만들어 내는 것이 불가능하다. 따라서 훈련된 모델은 훈련용 데이터 내부에 존재하는 음소에 편중되어 훈련되는 현상을 나타낸다. 이러한 현상은 미관측 음소 모델이 인식과정에서 요구되는 경우 음성인식의 성능을 저하시키는 요인으로 작용한다.

훈련된 음소 모델이 훈련용 데이터에 편중되는 경우 모델을 평활화함으로서 문제를 해결 할 수 있다. 본 논문에서는 문맥 독립형 음소 모델과 문맥 종속형 음소 모델을 deleted interpolation 방법을 사용하여 병합 함으로써 모델이 훈련 데이터로 편중되는 문제를 해결하였다. deleted interpolation 이란 두 모델을 병합하여 하나의 다른 모델을 만들어내는 방법으로 다음 식과 같다.

$$\tilde{\lambda} = \varepsilon\lambda + (1-\varepsilon)\lambda'$$

문맥 독립형 음소모델 λ 와 문맥 종속형 음소모델 λ' 이 서로 병합되어 $\tilde{\lambda}$ 가 생성되고 이때 ε 은 0~1 사이의 상수이다. 이렇게 병합된 모델은 병합전의 문맥 독립형 음소 모델에 비하여 안정적이며, 특히 미관측 문맥 종속형 음소 모델에 보다 잘 적용된다.

4. 실험 및 결과

인식 실험을 위한 음성 데이터베이스는 한국전자통신연구소에서 만든 한국어 고립단어 445단어 음성 데이터베이스 이다. 어휘독립을 위하여 445개의 단어를 훈련과 인식을 위한 단어집합으로 나누었다. 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].

기본적인 인식시스템으로는 HTK Toolkit 을 사용하였다. 사용된 특징 파라미터로는 12차 MFCC와 12차 델타-MFCC 그리고 12차 델타-델타-MFCC를 사용하였다.

4.1 군집화 방법에 관한 실험

결정트리를 이용한 음소 군집화 방법으로 단말노드에서만 모델 훈련이 가능하도록 트리를 구성하는 기존의 방법과 백오프 기법을 이용한 방법을 비교 실험 하였다. 그림 4에서 음소모델 훈련을 위한 문턱값에 따른 인식률에 초점을 두어 살펴보면 문턱값이 작을 수록 백오프 기법을 이용한 방법이 기존의 방법보다 우수 함을 알 수 있다. 이는 소용량 음성 데이터베이스를 이용한 어휘독립 음성인식의 경우 상대적으로 많은 문맥정보를 포함하는 백오프 방법이 기존의 방법보다 인식률의 향상에 도움이 된다는 것을 의미한다.

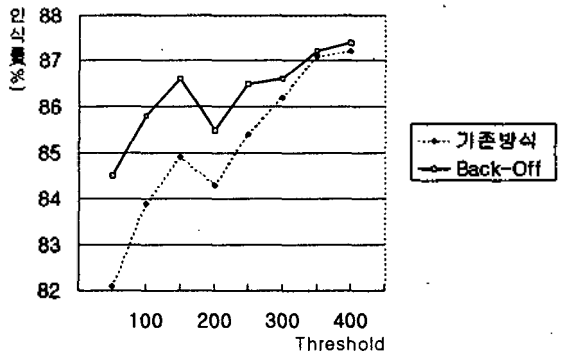

그림 4 음소군집화 방법에 대한 비교 실험

4.2 Deleted Interpolation에 의한 모델 평활화 실험

데이터의 양이 부족하므로 훈련된 문맥 종속형 음소 모델은 훈련용 데이터에 편중되는 경향이 있다. 따라서 미관측 음소 모델에 대하여 인식률이 저하되는 현상이 나타난다. 표 1에서 보는 바와 같이 미관측 모델이 없는 경우 트라이폰을 이용한 경우가 모노폰을 이용한 경우 보다 월등히 좋지만, 미관측 모델이 존재하는 경우 역전되는 경향이 보인다.

표 1 미관측 모델과 관측 모델의 비교 실험

	관측 모델을 요구	미관측 모델을 요구
모노폰	89.3	87.9
트라이폰	92.5	85.8

이러한 경향은 문맥정보가 인식률에 많은 영향을 미치기는 하지만 데이터의 부족 현상을 극복하기 힘들다는 의미로 해석 될 수 있다. 모노폰 모델과 트라이폰 모델을 deleted interpolation 방법으로 병합하여 사용함으로써 이러한 문제에 대처하였다. 그림 5에서 보듯이 deleted interpolation 의해 병합된 모델을 사용하는 경우 월등히 우수한 성능을 나타내고 있다.

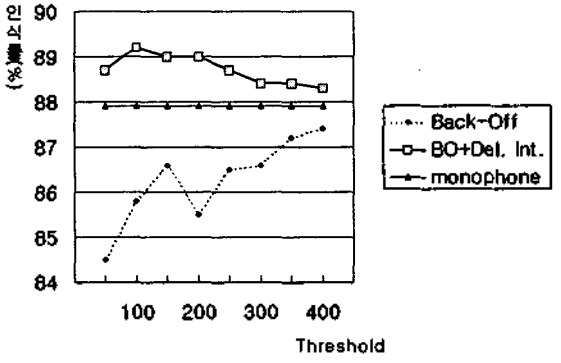

그림 5 Deleted Interpolation에 의한 병합된 모델과 병합되지 않은 모델 그리고 모노폰 모델의 비교 실험

5. 결론

본 논문에서는 소용량의 음성 데이터베이스를 사용하는 어휘독립음성인식 시스템에서 문맥 종속형 음소모델을 사용하는 경우 음소 군집화 방법으로 백오프 기법을 이용한 음소 군집화 작업이 유용함을 보였다. 그리고 백오프 기법을 이용할 경우 모델의 수와 인식기의 성능은 밀접한관계가 있으며 적절한 모델의 수를 결정하는 것이 중요함을 실험을 통하여 살펴 보았다. 데이터 양의 부족으로 인한 모델이 훈련 데이터로 편중되는 문제는 deleted interpolation 방법으로 모델을 평활화함으로써 해결할 수 있었다.

앞으로 소용량 음성 데이터베이스에서 유용한 백오프 기법의 음소 군집화 방법을 대용량 음성 데이터베이스에 적용해 보고자 한다. 그리고 엔트로피 개념의 매저를 사용하여 모델의 개수와 관련된 최적화된 음소 군집화를 시도해 보고자 한다.

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] 황병환, "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 음성인식", 박사 학위 논문, KAIST, 1999.

[8] 구동욱,최준기,오영환, "음소모델의 Back-Off 기법을 이용한 어휘독립 음성인식 시스템의 성능개선", 한국음향학회 하계학술발표대회 논문집, 2000.

[9] Rabiner, Juang, "Fundamentals of Speech Recognition", Prentice Hall, pp. 372~374,1993.