한국음향학회지 제23권 제2호 pp. 173~179 (2004)

음소 질의어 집합 생성 알고리즘

Phonetic Question Set Generation Algorithm

김 성 아*, 육 동 석*, 권 오 일**
(Sung-a Kim*, Dongsuk Yook*, Ohil Kwon**)

*고려대학교 컴퓨터학과 음성정보처리 연구실, **현대 오토넷 주식회사

(접수일자: 2003년 7월 1일; 수정일자: 2003년 10월 13일; 채택일자: 2003년 10월 28일)

음소 질의어 집합은 문맥 속에서 비슷한 조음 효과를 보이는 음소들을 분류해 놓은 것으로서, 음성 인식 시스템 학습 시 결정트리를 기반으로 HMM (hidden Markov model)의 상태들을 클러스터링할 때 사용된다. 현재까지의 음소 질의어 집합은 대부분 음성학자나 언어학자들에 의해 수작업으로 제시되어 왔는데, 이러한 지식 기반 음소 질의어들은 언어 또는 유사음소 단위 (PLU: phone like unit)에 종속될 뿐 아니라 생성된 클러스터 내의 동질성을 저하시킬 수 있다는 단점이 있다. 본 논문에서는 이와 같은 문제점들을 해결하기 위해 음성 데이터를 사용하여 측정한 음소들 사이의 유사도를 기반으로 언어나 유사음소단위에 상관없이 자동으로 음소 질의어 집합을 생성하는 알고리즘을 제안한다. 실험결과, 제안한 방법으로 생성된 음소 질의어들을 사용한 인식기의 에러율이 약 14.3% 감소하여 데이터 기반의 음소 질의어 집합이 상태 클러스터링에 효율적임을 관측하였다.

핵심용어: 음소 질의어 집합, 결정트리, 상태 클러스터링, 대용량 어휘 연속 음성인식, 문맥 종속 음향 모델
투고분야: 음성처리 분야 (2.5)

Due to the insufficiency of training data in large vocabulary continuous speech recognition, similar context dependent phones can be clustered by decision trees to share the data. When the decision trees are built and used to predict unseen triphones, a phonetic question set is required. The phonetic question set, which contains categories of the phones with similar co-articulation effects, is usually generated by phonetic or linguistic experts. This knowledge-based approach for generating phonetic question set, however, may reduce the homogeneity of the clusters. Moreover, the experts must adjust the question sets whenever the language or the PLU (phone-like unit) of a recognition system is changed. Therefore, we propose a data-driven method to automatically generate phonetic question set. Since the proposed method generates the phone categories using speech data distribution, it is not dependent on the language or the PLU, and may enhance the homogeneity of the clusters. In large vocabulary speech recognition experiments, the proposed algorithm has been found to reduce the error rate by 14.3%.

Keywords: *Phonetic question set, Decision tree, State clustering, Large vocabulary continuous speech recognition, Context dependent acoustic model*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

현재 우수한 성능을 보이고 있는 HMM (hidden Markov models) 기반 대용량 어휘 연속 음성인식에서는 음성의 전후 문맥을 고려하는 문맥 종속 음소 (context dependent phone)를 음향 모델의 단위로 사용한다[1]. 이는 같은 음소도 문맥에 따라 조음 효과로 인해 다르게 발음이 되는 연속 음성의 특성을 반영하기 위한 것으로, 일반적으로 좌, 우의 음소 하나씩을 고려하는 트라이폰 (triphone)을 사용한다. 그런데 트라이폰은 하나의 음소에 대해 서로 다른 전후 문맥을 갖는 모든 경우를 고려해야 하므로 그 수가 너무 많아질 수 있다. 예를 들어 음소의 수가 50개라면 트라이폰은 50^3개가 존재하는데 이를 모두 학습시킬 수 있는 데이터를 확보하는 것은 현실적으로 매우 어렵다. 결국 충분한 학습 데이터를 갖지 못하거나 혹은 전혀 학습되지 못하는 모델들이 발생하게 되고, 이러한 문제를 해결하기 위해 유사한 모델들끼리 클러스

책임저자: 육동석 (yook@voice.korea.ac.kr)
136-701 서울시 성북구 안암동 5-1
고려대학교 컴퓨터학과 음성정보처리 연구실
(전화: 02-3290-3202)

터를 구축하여 정보를 공유하는 방법들이 제시되어 왔다[2,3]. 클러스터들은 대개 문맥 종속 음소 HMM의 상태들로 구성되며, 이를 위해서 결정트리 기반의 top-down 방식이 사용된다. Top-down 클러스터링은 음소 질의어에 따라 데이터를 양분하였을 때 우도확률 (likelihood)이 높아지거나[2] 엔트로피 (entropy)가 낮아지는[3] 범주를 채택하는 과정을 반복하며 결정트리를 구축하고, 트리의 각 리프 노드에 속해있는 HMM 상태들이 하나의 클러스터를 이룬다. 이와 같이 클러스터로 묶여진 유사한 상태들은 대표 분포를 공유함으로써 학습 데이터의 부족 문제를 해결할 수 있다. 또한, 음소 질의어 집합을 이용하여 결정트리를 구축하면 학습되지 않은 모델이 필요한 경우 해당 중심 음소의 각 상태에 대해 트리의 루트 노드에서부터 주어진 음소 질의어에 따라 가지를 선택하는 과정을 반복하여 리프 노드에 도달하면 유사 분포의 파라미터를 할당받을 수 있다. 이와 같이 결정트리 기반의 top-down 클러스터링은 상태들의 클러스터를 구축함과 동시에 unseen 트라이폰과 같은 학습되지 않은 문맥 종속 음소들의 모델을 생성할 수 있다는 장점이 있어 단순히 유사한 분포들을 두 개씩 묶어나가는 bottom-up 방법에 비해 높은 인식 성능을 보인다[3].

Top-down 방법은 문맥 종속 음소들을 클러스터링하는 결정트리를 구축하고 학습되지 않은 문맥 종속 음소를 생성하기 위해 음소 질의어 집합을 필요로 한다. 음소 질의어 집합은 유사한 특성을 보이는 음소들을 분류해 놓은 것으로서, 현재까지는 대부분 음성학자나 언어학자들에 의해 제공되어왔다. 그러나 전문가들의 지식 기반으로 생성된 음소 질의어 집합은 실제 학습 데이터의 정보는 고려하지 않으므로 제시된 질의어 집합의 각 범주들이 반드시 유사도가 높은 분포를 갖는 데이터들로 이루어져 있다는 것을 기대하기 어렵다는 문제점이 있다. 만약 문맥내의 특성이 유사하지 않은 음소들이 묶여있는 범주가 결정트리 구축에 사용된다면 이는 생성된 클러스터의 동질성을 떨어뜨려 음성인식기의 성능을 저하시킬 수 있으므로 이에 대한 해결책이 필요하다. 게다가 음소 질의어는 비슷한 음소들의 분류이기보다는 문맥 내에서 유사한 영향을 주는 음소들의 범주이어야 하는데, 기존의 음소 질의어들은 단순히 유사한 조음 위치나 언어적 특징을 가지는 음소들을 묶어놓은 것에 불과하다. 뿐만 아니라, 기존 방식의 음소 질의어 집합은 전문가들에 의해 수작업으로 제공되기 때문에 언어와 유사음소단위 (PLU: phone like unit)에 종속되어 그 효율성도 떨어진다. 음성인식기가 사용하는 유사음소단위가 변경되거나 인식하고자 하는 언어가 바뀌면 새로운 유사음소단위나 언어에 대한 음소 질의어 집합이 요구되는데 이 때마다 해당 언어 전문가의 추가적인 도움이 필요하다. 따라서 본 논문에서는 기존의 지식 기반 음소 질의어 집합이 갖는 이와 같은 문제점들을 해결하기 위해 음소 질의어 집합을 데이터 기반으로 자동 생성하는 알고리즘을 제안한다.

제안한 음소 질의어 집합 생성 알고리즘은 문맥 독립 음소 (context independent phone) 모델간의 거리를 Bhattacharyya distance를 이용하여 측정하고, 이를 바탕으로 각 음소에 대해 유사한 음소 모델들을 찾아낸 뒤 이들을 계층적으로 묶어나간다. 이는 실제 데이터간의 거리를 측정하여 음소 질의어 집합을 생성하기 때문에 생성된 각 범주에 속하는 음소의 동질성을 높일 수 있다. 또한 단순히 음소의 유사도만을 고려하는 음소 질의어들이 아니라 문맥 속에서 유사한 영향을 주는 음소들의 범주를 생성하기 위해서는 음소 모델 전체보다는 하나의 음소가 다른 음소와 연결되는 부분의 정보가 중요하다. 따라서 본 논문의 알고리즘에서는 발음된 음소의 시작부분 소리와 끝부분 소리만을 고려하기 위해 음소 모델간의 유사도 뿐 아니라 HMM의 첫 번째 상태 또는 마지막 상태만이 유사한 음소들도 하나의 범주로 생성되도록 하였다.

본 논문의 II장에서는 기존 지식 기반 음소 질의어 집합이 결정트리 구축에 사용되었을 때 음성인식기의 성능을 저하시킬 수 있는 문제점들을 분석하고 그에 대한 해결방안과 제안하는 음소 질의어 집합 생성 알고리즘에 대해 설명한다. III장에서는 자동으로 생성된 음소 질의어 집합의 효율성을 검증하기 위한 음성인식 실험 결과를 분석하고, 마지막으로 IV장에서 결론을 맺는다.

II. 음소 질의어 집합 자동 생성

2.1. 지식 기반 음소 질의어 집합의 문제점

음소 질의어 집합은 문맥 내에서 특정 음소에 유사한 영향을 미칠 수 있는 음소들의 묶음을 모아놓은 것이다. 그런데 기존의 음소 질의어 집합은 언어학적 지식으로 음소들을 분류하기 때문에 비슷한 소리임에도 불구하고 같은 범주에 속하지 않는 경우가 발생할 수 있다. 그림 1은 각 음소들 사이의 유사도를 문맥 독립 음소 HMM들간의 Bhattacharyya distance (2.2절 참조) 값을 이용하여 표시한 것이다. 유사도가 높을수록 점의 크기가 크게 표현되었다. 예를 들어, 'ㅣ(i)'와 'ㅢ(eui)' 같은 음소들은 유사한 분포를 갖는다. 그러나 언어학 지식에 기초한

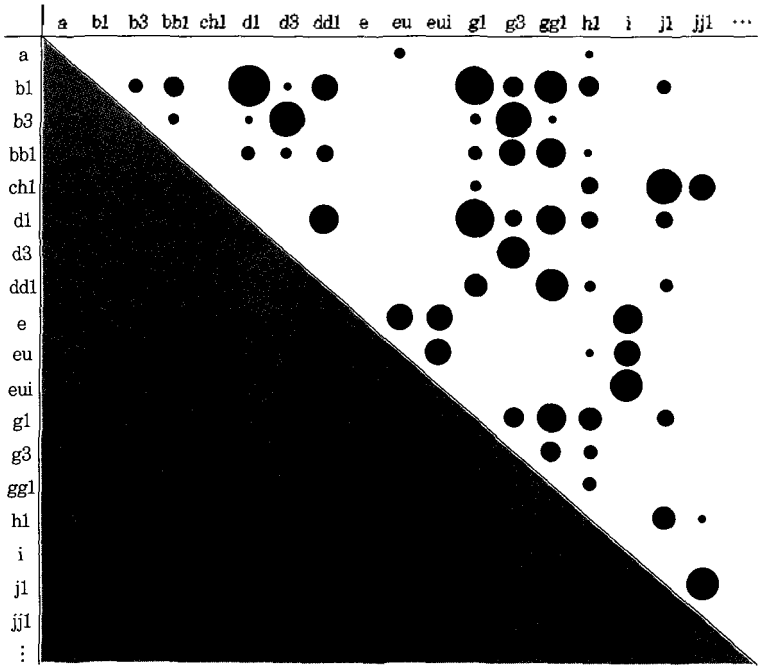

그림 1. 음소 모델 유사도
Fig. 1. Similarity between monophone HMMs.

음소 질의어 집합은 단모음과 이중모음이 다르게 분류되므로 두 개의 음소가 하나의 범주에 속하지 않는다. Resource Management linguistic question set[4]과 같은 영어 음소 질의어 집합 역시 'y'와 같은 음소를 모음으로 분류하지 않아서 'ih', 'iy', 'y'는 유사한 분포를 가짐에도 하나의 질의어 범주에 속하지 않는다. 이와 같이 기존의 음소 질의어 집합에서는 실제 데이터들이 비슷한 분포를 가져도 언어학적 기준에 따라서는 같은 범주에 묶이지 않는 경우가 발생하는데, 결정트리 기반 클러스터링을 위한 음소 질의어 집합에서는 음소들이 언어학적으로 어떤 집합에 속하는지에 상관없이 분포가 유사하면 하나의 그룹으로 묶어주는 것이 필요하다.

지식 기반 음소 질의어 집합에는 위와 같이 실제로는 유사하지만 언어학적 의미가 달라서 하나의 범주로 묶이지 못한 경우가 존재할 뿐 아니라 언어학적으로는 같은 범주에 속하지만 데이터의 분포는 전혀 유사하지 않은 질의어들도 있다. 예를 들어, 언어학자들은 표 1에서와 같이 'ㅉ(jj1)', 'ㅆ(ss1)', 'ㅃ(bb1)', 'ㄸ(dd1)', 'ㄲ(gg1)'과 같은 음소들을 경음 (fortis)이라는 하나의 언어학적 범주로 묶는데 실제 이러한 음소들은 서로 다른 소리를 내며 'ㅉ(jj1)'과 'ㅃ(bb1)', 'ㅉ(jj1)'과 'ㄲ(gg1)' 등은 유사도가 낮음을 그림 1에서 알 수 있다. 이와 같이 하나의 음소 질의어 범주에 속하는 데이터들이 이질적일 경우 그것을 기반으로 구축된 결정트리가 생성해내는 클러스터 역시 동질성이 떨어진다. 하나의 클러스터로 묶여진 HMM 상태들의 동질성이 떨어지면 서로 다른 분포를 갖는 데이터들이 하나의 분포를 공유하게 되므로 모델링 에러가 발생하여 음성인식기의 성능이 저하된다.

표 1. 지식 기반 음소 질의어 집합의 예
Table 1. An example of knowledge-based phonetic question set.

Categories	음소
Stop	b1 d1 g1 kh1 ph1 th1 ⋯
Vowel	e i eu u o v a
Vowel-front	e i
Vowel-central	eu a o
Vowel-back	u v
Fortis	jj1 ss1 bb1 dd1 gg1
⋮	⋮

또한, 결정트리 기반 상태 클러스터링은 중심 음소가 같은 모델들 중 유사한 전후 문맥을 갖는 트라이폰들의 상태들을 하나의 그룹으로 묶는 것이 목적이기 때문에 음소 질의어는 그들 자체가 비슷한 분포를 갖는 음소 모델들로 이루어지기 보다는 다른 하나의 음소에 비슷한 영향을 주는 음소들로 구성되어야 한다. 'ㅏ(a)', 'ㅘ(wa)', 'ㅑ(ya)', 세 모음의 소리는 서로 다른 입 모양으로 시작되는 발음들로서 전체 모델들은 서로 다르지만 발음 마지막의 소리는 세 음소 모두 유사하다. 따라서 이러한 발음들이 특정 중심 음소의 왼쪽 문맥에 위치한다면 해당 음

소에 유사한 영향을 줄 수 있어 하나의 음소 질의어가 되는 것이 바람직하다. 그러나 기존의 방식은 음소 전체가 유사한 것들만을 묶어 음소 질의어들을 생성하므로 발음의 첫 부분이나 끝 부분만이 유사하여 같은 영향을 줄 수 있는 경우는 고려하지 않는다.

이와 같이 기존의 음소 질의어 집합은 수작업으로 제공되기 때문에 발생하는 언어 종속, 유사음소단위 종속 문제 뿐 아니라 제시된 음소 분류 범주들의 타당성 면에서도 문제점들을 갖는다. 본 논문에서는 언어나 유사음소단위에 상관없이 자동으로 음소 질의어 집합을 생성하면서, 기존의 음소 질의어 집합이 갖는 위의 단점들을 해결할 수 있는 알고리즘을 제시한다.

2.2. 음소 질의어 집합 생성 알고리즘

본 논문에서는 결정트리 기반의 HMM 상태 클러스터링의 성능 향상을 위해, 결정트리의 각 노드에서 데이터를 양분하는 규칙 (rules)으로 사용되는 음소 질의어들을 데이터 기반으로 자동 생성하는 알고리즘을 제안한다. 제안한 방법으로 생성되는 음소 질의어 집합은 위에서 지적한 지식 기반 음소 질의어 집합이 가지는 단점들을 해결함으로써 결정트리가 보다 동질성 높은 클러스터들을 생성할 수 있게 한다.

결정트리 기반 상태 클러스터링은 음소 질의어들을 바탕으로 유사한 분포의 HMM 상태들을 클러스터링 하기 위한 것이고, 따라서 각 노드의 질의어들은 현재 노드에 있는 상태 중에서 유사한 분포를 갖는 것들을 추출할 수 있어야 한다. 이에 본 논문에서는 음소 데이터간의 거리를 수치적으로 측정한 뒤, 이를 바탕으로 유사한 음소들을 묶어 질의어 집합을 생성한다. 음소 데이터 간의 거리 측정을 위해서는 두 개의 정규 (Gaussian) 분포 사이의 거리를 측정하는 방법 중 하나인 Bhattacharyya distance를 이용한다. 두 개의 정규 분포 g와 h간의 Bhattacharyya distance는 다음과 같이 정의된다[5].

$$B(g,h)=\frac{1}{8}(M_g-M_h)^T\left[\frac{\Sigma_g+\Sigma_h}{2}\right]^{-1}(M_g-M_h)+\frac{1}{2}\log\frac{\left|\frac{\Sigma_g+\Sigma_h}{2}\right|}{\sqrt{|\Sigma_g||\Sigma_h|}} \quad (1)$$

여기서 M_g, Σ_g와 M_h, Σ_h는 각각 분포 g와 h의 평균 벡터와 공분산 행렬 (covariance matrix)을 나타내며 g와 h가 유사할수록 $B(g,h)$의 값은 작게 나타난다. Bhattacharyya distance는 그 계산이 간단하며 베이지안 (Bayesian) 분류 (classification) 에러 확률과 이론적 상계 (upper bound)가 같다는 장점이 있다[5]. 정규 분포로 모델링된 각 음소 HMM의 상태들 역시 수식 (1)을 이용하여 거리를 측정할 수 있고, Bhattacharyya distance값이 작은 데이터들을 묶으면 동질성이 큰 HMM들의 집합을 생성할 수 있다[6]. 제안한 방법에서는 위의 척도를 이용하여 음소 모델간의 유사도를 측정함으로써 그것이 어떤 언어의 어떤 종류의 음소인지에 상관없이 실제 소리의 특성이 비슷한 음소들을 찾아낸다.

본 논문의 목적은 결정트리 기반 상태 클러스터링을 위한 음소 질의어 집합을 생성하는 것이므로, 일반적인 데이터 클러스터링과 달리 고려해 주어야 할 문제들이 있다. 첫째, 하나의 음소 질의어에 속한 음소들은 그 자체가 유사한 것들의 클러스터이기 보다는 문맥 내에서 유사한 영향을 주는 음소들을 분류해 놓은 것이어야 한다. 2.1절에서 살펴보았듯이 'ㅏ(a)', 'ㅑ(ya)', 'ㅘ(wa)' 세 음소는 발음의 시작 부분 보다는 끝 부분이 유사하기 때문에 이러한 음소들은 특정한 트라이폰의 왼쪽 문맥에서 해당 중심 음소에 비슷한 영향을 준다. 이와 같이 연속 음성 속에서 하나의 음소에 문맥적 영향을 주는 것은 전에 오는 음소의 끝부분 발음이거나 후에 오는 음소의 앞부분 발음이다. 문맥 독립 음소 모델을 비교하면서도 음소 간의 유사도가 아닌 문맥적 영향을 비교하기 위해서는 특정 음소 모델의 상태들 중 앞의 일부나 뒤의 일부만의 유사도를 고려할 수 있다[7]. 따라서 본 논문의 알고리즘에서는 HMM 전체 상태의 평균거리를 비교하여 모델 전체가 유사한 음소들을 하나의 범주로 생성할 뿐 아니라, 문맥적 특성이 유사한 음소들을 하나의 그룹으로 묶기 위해 HMM의 첫 번째 상태만을 비교하여 오른쪽 문맥에 대한 질의어들을 생성하고 마지막 상태만을 비교하여 왼쪽 음소 질의어들을 생성한다.

또한 음소 질의어 집합은 모든 음소들의 각 상태에 대해 결정트리를 만들 때 쓰이는 분류 범주의 후보들이기 때문에 다양한 음소들의 묶음이 생성되어야한다. Bhattacharyya distance에 따르면 초성 'ㅂ(b1)', 초성 'ㄷ(d1)', 초성 'ㄱ(g1)'은 유사한 분포를 갖는 데이터인데 초성 'ㅂ(b1)'과 'ㅍ(ph1)' 역시 유사한 분포를 갖는다. 그런데 'ㄷ(d1)'과 'ㅍ(ph1)', 'ㄱ(g1)'과 'ㅍ(ph1)'은 상대적으로 다른 분포를 가지므로 'ㅂ(b1)', 'ㄷ(d1)', 'ㄱ(g1)'의 범주와 'ㅂ(b1)', 'ㅍ(ph1)'의 범주가 따로 존재하는 것이 타당하다. 그러나 음소 질의어 집합을 단순히 전체 음소에 대한 bottom-up 이나 top-down 방식과 같이 각 음소들을 리프 노드로 갖는 이진 트리의 형태로 생성[7,8]한다면 위와 같은 두 개의 범주가 동시에 생성될 수

없다. 따라서 본 논문에서는 각 음소에 대해 타당한 문맥 정보들을 생성해 주기 위해서, 특정 음소에 대해 가까운 거리를 갖는 음소들을 찾은 뒤 해당 음소와 찾아진 유사 음소들의 쌍을 각각 병합하여 하나의 클러스터로 초기화하고 그것들을 기본 클러스터로 하여 계층적인 음소들의 묶음을 생성하였다. 이렇게 함으로써 하나의 음소에 대해 유사한 음소들이 만드는 범주에는 해당 음소가 항상 포함될 수 있다. 또한, 특정 거리 이내에 속하는 음소들만 병합해 나가기 때문에 상위 레벨로 올라가도 서로 상이한 음소들이 묶이는 경우가 발생하지 않아서 각 범주 내 데이터의 동질성이 보장된다. 이와 같이 자동 문맥 정보를 생성하는 알고리즘을 정리하면 표 2와 같다.

III. 실 험

자동 음소 질의어 집합은 문맥 독립 음소의 HMM을 이용하여 위에서 설명한 알고리즘으로 생성되었으며 표 3이 만들어진 범주들 중 초성 자음 'ㅂ(b1)'과 이중모음 'ㅘ(wa)'를 포함하는 것의 일부이다. 기본 bottom-up 알고리즘은 하나의 음소가 포함되는 범주가 한 집합의 진부분집합들로서만 생성되는 것에 비해 제안한 방법으로 생성된 음소 질의어 집합은 유사한 분포를 갖는 보다 다양한 범주들로 구성됨을 표에서 확인할 수 있다. 예를 들면, 2.1절에서 언급한 경음들의 집합은 생성하지 않으면서 2.2절에서 설명한 'ㅂ(b1)', 'ㄷ(d1)', 'ㄱ(g1)' 집합뿐만 아니라 'ㅂ(b1)'과 'ㅍ(ph1)'의 집합도 생성할 수 있었다. 또한 마지막 상태의 유사도만을 비교함으로써 'ㅘ(wa)', 'ㅏ(a)', 'ㅑ(ya)'와 같이 소리의 끝 부분이 유사한 음소들의 범주들 역시 생성함을 확인하였다.

이와 같은 음소 질의어 집합 생성 알고리즘의 성능 평가를 위해서 표 3과 같이 자동으로 생성된 음소 질의어 집합과 표 1의 기존 음소 질의어 집합을 각각 적용한 음성인식기의 고립 단어 인식 에러율을 측정하였다. 본 실험에서 사용한 지식 기반 한국어 음소 질의어 집합은 언어학자들에 의해 제공받은 것으로써 각 음소의 소리와 발음시의 입 모양, 혀의 위치, 스펙트로그램 (spectrogram) 등을 참조하여 수작업으로 생성된 것이다.

음향 모델의 학습과 인식 실험에는 본 연구실에서 자체 개발한 SLT (spoken language toolkit) version 1.0을 사용하였다. SLT 1.0은 음향 모델 학습 도구와 오프라인 및 온라인 적응 도구[9,10], 그리고 음성 인식 엔진 등으로 구성되어 있으며 병렬 처리가 가능하도록 설계된 음성인식 연구용 소프트웨어이다. 학습에 사용된 데이터는 현대 오토넷 CNS (car navigation system) 데이터베이스이며, 음성 특징 벡터로는 매 10 ms마다 멜 스케일 켑스트럼 계수에 에너지 값을 추가한 13차원 벡터와 그 1,2차 미분값이 더해진 총 39차원의 벡터를 추출하여 사용하였다. HMM은 세 개의 상태로 이루어진 left-to-right 구조 모델을 사용하였고 각 상태는 20개의 Gaussian들로 구성된다. HMM 상태 클러스터링을 위해서는 로그 우도확률 기반으로 결정트리를 구축하였다. 이와 같이 학습된 모

표 2. 음소 질의어 집합 생성 알고리즘
Table 2. Phonetic question set generation algorithm.

- 모든 문맥 독립 음소 HMM p에 대해 다음을 반복한다.
  - p의 첫 번째 상태, 마지막 상태, 모든 상태들의 평균에 대해,
    1. 모델 p와의 Bhattacharyya distance 값이 기준치보다 작은 모노폰 (monophone)들을 찾는다.
    2. 위의 과정에서 찾아진 모노폰들과 p의 쌍들은 각각 하나의 분포로 병합되어 하나의 클러스터를 이루며 음소 질의어 집합에 추가 된다.
    3. 더 이상 병합할 클러스터가 존재하지 않거나, 집합에 새로이 추가된 클러스터가 없을 때까지 다음을 반복한다.
       1) 다른 것과 병합된 적 없는 클러스터들 중에서 거리가 가장 가까운 두 개를 현재 음소 질의어 집합에서 찾는다.
       2) 두 개의 클러스터를 병합하여 새로운 클러스터를 생성한다.
       3) 새로 생성된 클러스터에 속한 음소들 간의 거리 중 최대값이 기준치 이하이고, 음소 목록이 음소 질의어 집합 내에 존재하지 않으면 새로운 음소들의 묶음을 집합에 추가한다.

표 3. 자동 생성된 음소 질의어 집합
Table 3. Automatically generated phonetic question set.

Categories	음소
state1-b1-14	b1 ph1
state1-b1-15	b1 r1
state1-b1-16	b1 th1
state1-b1-17	b1 bb1 d3
state1-b1-18	b1 gg1 th1
state1-b1-19	b1 bb1 d3 kh1
state1-b1-20	b1 bb1 d3 dd1 kh1
state1-b1-21	b1 bb1 d3 dd1 gg1 kh1 ph1 th1
state1-b1-22	b1 b3 g3
state1-b1-23	b1 d1 g1
⋮	⋮
state3-wa-1	wa a ya
state3-wa-2	wa eu yu
state3-wa-3	wa v yo
⋮	⋮

표 4. 음소 질의어 집합에 따른 단어 인식 에러율
Table 4. Word recognition error rates for manual and automatic phonetic question set.

음소 질의어 집합	에러율 (%)
지식 기반 음소 질의어 집합	2.8
Bottom-up 방식으로 생성된 자동 음소 질의어 집합	2.6
제안한 알고리즘으로 생성된 자동 음소 질의어 집합	2.4

델을 이용하여 SITEC에서 배포한 PBW 데이터에 대해 단어 인식 실험을 수행한 결과, 자동으로 생성된 음소 질의어 집합을 이용한 인식기는 표 4와 같이 기존의 지식 기반 음소 질의어 집합을 이용한 인식기보다 상대적으로 약 14.3%의 에러율 감소를 보였다. 또한, 제안한 방식의 클러스터링이 아닌 단순히 모든 음소들을 리프노드에 두고 묶어 나가는 bottom-up 방식으로 질의어들을 생성하여 결정트리에 사용한 경우는 에러율이 2.6%로 본 논문에서 제안한 알고리즘보다는 성능이 약 8% 정도 떨어지지만 여전히 지식 기반 음소 질의어 집합에 비해서는 음성 인식기의 성능을 높임을 확인할 수 있었다.

자동 생성된 음소 질의어 집합을 이용하여 생성된 결정트리는 총 5207개의 질의어 노드수를 가지고 5331개의 상태 클러스터들을 생성하였으며, 기존의 지식기반 음소 질의어 집합은 총 질의어 노드수 5451개, 생성된 클러스터수 5575개로 유사한 크기의 결정트리를 생성하였다. 또한, 자동 음소 질의어 집합으로 구축된 결정트리에는 1400개의 고유한 질의어들이 사용되었는데 이 중에서 1031개가 모노폰의 첫 번째 상태 또는 마지막 상태의 유사도만을 바탕으로 생성된 것으로서, 소리의 앞부분과 뒷부분의 유사도를 바탕으로 생성된 질의어들이 실제로 많은 노드에서 선택됨을 알 수 있었다.

IV. 결 론

결정트리는 학습 데이터를 유사한 것끼리 클러스터로 묶거나 학습되지 않은 unseen 트라이폰의 모델을 생성할 때 음소 질의어 집합을 필요로 한다. 따라서 음소 질의어 집합은 유사한 분포를 가지는 트라이폰들을 구분할 수 있어야 하며, 이를 위해 하나의 중심 음소에 비슷한 영향을 줄 수 있는 왼쪽 음소나 오른쪽 음소들의 묶음들로 구성되어 있다. 기존의 음소 질의어 집합은 전문가들에 의해 제시되어 왔는데 이는 유사한 영향을 주는 음소들을 분류한 것이기 보다는 그 자체가 언어학적으로 같은 범주에 속하는 음소들의 묶음이어서 생성된 클러스터의 동질성을 보장하기 어렵다는 단점이 있었다. 따라서 본 논문에서는 이러한 문제를 해결하기 위해 음소 모델 간의 거리를 수치적으로 측정하고 유사한 것들을 묶어나감으로써 음소 질의어 집합을 생성하는 방법을 제안하였다. 제안한 방법은 데이터 기반으로 음소 문맥을 생성하기 때문에 인식기가 사용하는 언어나 유사음소단위를 고려할 필요가 없다는 장점이 있을 뿐 아니라, 자동으로 생성된 음소 질의어 집합이 수작업으로 제공된 음소 질의어 집합에 비해 음성인식기의 에러율을 감소시킴을 실험으로 확인하였다. 이는 실제 음소 데이터들의 분포가 반드시 언어학적 범주를 따르지는 않는다는 것을 의미하며, 이로 인해 언어학적 구분에 따른 음소 질의어 집합을 사용하는 것 보다는 데이터 기반의 음소 질의어 집합을 이용하여 클러스터링을 하는 것이 보다 효율적임을 알 수 있었다.

참 고 문 헌

1. K. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38** (4), 599-609, 1990.
2. S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *DARPA Human Language Technology Workshop*, 307-312, March 1994.
3. M. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," *IEEE Transactions on Speech and Audio Processing*, **4** (6), 412-419, November 1996.
4. J. Odell, "The use of context in large vocabulary speech recognition," PhD thesis, University of Cambridge, 1996.
5. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.
6. D. Yook, "Decision tree based clustering", *Lecture Notes in Computer Science*, **2412**, 487-492, August 2002.
7. K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, **2**, 805-809, May 1998.
8. R. Singh, B. Raj, and R. Stern, "Automatic clustering and generation of contextual questions for tied states in hidden Markov models", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, **1**, 117-120, May 1999.
9. D. Yook, "Unsupervised incremental online adaptation to unknown environment and speaker," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, 617-620, May 2002.
10. D. Yook, "Hidden Markov model and neural network hybrid", *Lecture Notes in Computer Science*, **2510**, 196-203, October 2002.

저자 약력

● 김 성 아 (Sung-a Kim)

2002년 8월: 고려대학교 컴퓨터교육과 (이학사)
2002년 9월~현재: 고려대학교 컴퓨터학과 석사과정
※ 주관심분야: 음성인식, 기계학습

● 육 동 석 (Dongsuk Yook)

1990년: 고려대학교 컴퓨터학과 (학사)
1993년: 고려대학교 컴퓨터학과 (석사)
1999년: 뉴저지 주립대학교 (Rutgers University) 컴퓨터학과 (박사)
1999년~2001년: IBM T.J. Watson Research Center, Senior S/W Engineer
2001년~현재: 고려대학교 컴퓨터학과 교수
※ 주관심분야: 음성인식, 화자인식, 기계학습

● 권 오 일 (Ohil Kwon)

1991년: 고려대학교 전자공학과 (학사)
1993년: 고려대학교 전자공학과 (석사)
1996년 8월: 고려대학교 전자공학과 (박사)
1996년 8월~2000년 3월: 현대전자산업주식회사 차장
2000년 3월~현재: 현대오토넷주식회사 차장
※ 주관심분야: 음성인식, 음성합성