가변어휘 음성인식기 구현에 관한 연구

황병한, 김형순
부산대학교 전자공학과

A Study on the Implementation of Vocabulary-Independent Korean Speech Recognizer

Byounghan Hwang, Hyung Soon Kim,
Dept. of Electronics Eng., Pusan National University,
E-mail : {bhealge, kimhs}@hyowon.cc.pusan.ac.kr

요 약

본 논문에서는 사용자가 별도의 훈련과정 없이 인식대상 어휘를 추가 및 변경이 가능한 가변어휘(vocabulary-independent) 인식시스템에 관하여 기술한다. 가변어휘 음성인식에서는 미리 구성된 음소모델을 토대로 인식대상 어휘가 결정되면 발음사전에 의거하여 이들 어휘에 해당하는 음소모델을 연결함으로써 단어모델을 만든다.

사용된 음소모델은 현재 음소의 앞뒤의 음소 context를 고려한 문맥종속형(Context-Dependent) 음소모델인 triphone을 사용하였고, 연속확률분포를 가지는 Hidden Markov Model(HMM) 기반의 고립단어인식 시스템을 구현하였다. 비교를 위해 문맥 독립형(Context Independent) 음소모델인 monophone으로 인식실험을 병행하였다.

개발된 시스템은 음성특징벡터로 MFCC(Mel Frequency Cepstrum Coefficient)를 사용하였으며, test 환경에서 나타나지 않은 unseen triphone 문제를 해결하기 위하여 state-tying 방법 중 음성학적 지식에 기반을 둔 tree-based clustering 기법을 도입하였다. 음소모델 훈련에는 ETRI에서 구축한 POW(Phonetically Optimized Words) 음성 데이터베이스(DB)[1]를 사용하였고, 어휘독립 인식 실험에는 POW DB와 관련 없는 22개의 부서명을 50명이 발음한 총 1,100개의 고립단어 부서DB[2]를 사용하였다. 인식실험 결과 문맥독립형 음소모델(monophone)이 88.6%를 보인데 비해 문맥종속형 음소모델(triphone)은 96.2%의 더 나은 성능을 보였다.

1. 서 론

음성은 사람에게 있어서 가장 자연스러운 의사전달 수단이다. 음성인식 기술은 음성합성 기술과 함께 음성을 통해 인간이 컴퓨터와 대화를 할 수 있도록 해주는 도구로서 정보화와 진전과 더불어 그 필요성이 더욱 증대되고 있다.

현재 사용중인 대다수의 음성인식 시스템은 인식할 대상어휘를 미리 선정하고, 이 어휘들에 대해서 음성 DB를 수집한다. 그리고 이 음성 DB를 사용하여 인식할 단어 또는 음소의 모델을 훈련한다. 이와 같은 방식의 음성인식 시스템은 선정한 어휘들에 대해서는 높은 인식성능을 보이지만, 인식대상 어휘를 변경 또는 추가할 필요가 있을 때는 새로운 어휘에 대해 음성 DB를 별도로 수집하여 처음부터 다시 모델을 훈련해야 하는 문제점이 발생한다.

본 논문에서 개발된 가변어휘 인식기술은 특별한 훈련과정을 거치지 않은 사람이 발음한 음성을 인식하는 화자독립 음성인식 기술로서, 인식대상 어휘의 변경 및 추가 필요시 나타나는 문제점을 해결하고, 다양한 응용분야의 각종 명령어를 음성인식에 의해 수행함으로써 사용자에게 편리한 입력 인터페이스를 제공해 줄 수 있어서 매우 유용하다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 가변어휘 인식시스템에 대해 기술하고, 3장에서는 훈련 및 인식에 사용한 음성 DB를, 4장에서는 실험 및 결과를 다룬 후, 5장에서 결론을 맺는다.

2. 가변어휘 인식시스템

2.1 인식시스템의 구성

훈련 및 인식의 통계적 자료로 사용되는 음성특징 벡터는 입력된 음성신호를 16bit, 16kHz로 샘플링한 후 10 msec 마다 20 msec 길이의 프레임 단위로 추출한 24차의 벡터(12차 MFCC 및 12차 delta MFCC로 구성)를 사용하였다.

음소 모델은 통계적 자료에 기반한 모델로, 세 개의 상태(state)를 가지는 left-to-right HMM(Hidden Markov Model)으로 구성하였다. 각 상태는 단일 Gaussian 확률 밀도 함수를 가지고 관측 벡터의 발생 확률을 계산하도록 하였다. 각 음소모델은 상용 음성인식 tool인 HTK(Hidden Markov Model Tool Kit)[3]를 사용하여 모델링하였다.

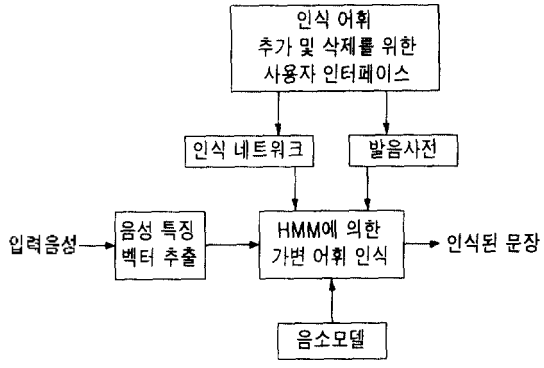

그림 2-1 가변어휘 인식시스템의 구성

2.2 문맥 독립형 및 문맥 종속형 음소모델[4]

음소 모델은 일반적으로 다양한 한국어 음운현상을 반영하는 40개 전후의 기본 유사음소 집합을 선정한다. 음소 모델을 구성할 때, 현재 음소의 앞뒤의 음소 context를 고려하지 않는 경우를 문맥 독립형(Context-Independent) 음소모델이라 하고, 앞뒤의 음소 context를 고려한 경우를 문맥 종속형(Context-Dependent) 음소모델이라 한다. 문맥 독립형 음소모델의 경우, 적은 수의 음소모델로 구현함으로써 인식 소요 시간이 짧고 구현하기에 쉬운 반면에, 문맥 종속형 음소모델에 비해 인식성능이 떨어지는 단점을 갖고 있다.

문맥 종속형 음소 모델은 문맥 독립형 모델에 비해 단어 내의 음운 현상을 효과적으로 반영하는 장점이 있지만, 그 모델의 수가 너무 많아지는 문제점이 발생한다. 앞뒤의 음소 정보를 포함하는 triphone의 경우, 예를 들어 기본 유사음소집합의 개수를 40개로 가정하면 40*40*40 = 64,000개의 triphone이 가능하고, 한국어의 경우 음성학적으로 발생할 수 없는 조합을 모두 제외하더라도 대략 20,000여 개의 triphone이 나타난다. 이와 같이 많은 수의 triphone들에 대해서 신뢰도 높은 모델 파라메터를 추정하기 위해서는 방대한 훈련용 데이터베이스가 필요하게 된다. 그리고, 이러한 대용량 훈련용 DB는 모든 한국어 음운현상을 포함하며 효율적으로 구성되어져야 한다. 그러나, 모든 한국어 음운 현상을 포함하는 DB를 구축하는 일은 현실적으로 어려운 문제이므로, 음운현상을 충분히 포함하면서도 효율적인 크기를 갖는 음성 DB 구성 방법에 관한 연구도 많이 진행되었다[5]. 이러한 방법에 의해 다양한 음소 context가 고려되도록 제작된 대용량 음성DB에서도 발견되지 않는 triphone (unseen triphone)이 존재할 수 있다. 가변어휘인식의 경우, 이러한 문제점의 해결은 대용량 음성 DB제작이나 선정 외의 다른 해결방법이 필수적으로 요구된다.

본 과제에서는 첫 번째로 음성 DB제작 및 선정의 문제점을 해결하기 위해 POW 3848 DB를 이용하였다[1]. 인식성능을 향상을 위해 이 음성 DB를 이용하여 문맥 종속형 음소모델(triphone)을 훈련하였다. 두 번째로, 문맥 종속형 음소모델을 사용시에 나타나는 unseen triphone 문제점을 해결하기 위해 tree-based clustering기법을 도입하였다. 이 방법은 훈련시 유사한 통계적 특성을 갖는 음소 모델의 state들을 하나의 그룹으로 묶음으로서 전체 모델의 파라메터 수를 줄여 상대적으로 신뢰도 높이는 State tying방법의 하나이다.

2.3 Tree-based Clustering

State tying에는 data-driven clustering 방법과 tree-based clustering 방법이 있다[3][6][7]. Data-driven clustering 방법은 훈련용 데이터에 포함된 triphone 모델

반 state tying하므로 대단위 어휘 인식 또는 가변 어휘 인식 시스템 구현할 때는 unseen triphone 문제점이 나타날 수 있다. 이 문제를 해결하기 위해 또 다른 state tying방법인 tree-based clustering방법을 적용하였다. Tree-based clustering에서는 먼저 동일한 음소에 해당하는 모든 triphone 모델의 상태들을 함께 모은 다음, 이를 두 개의 부분집합으로 나누고, 그 각각의 부분집합을 다시 두 개의 부분집합으로 나누어 가는 일련의 과정을 통해 트리를 구성한다. 부분집합으로의 분할은 각 집합에 해당하는 트리의 노드에서 문맥에 대한 binary question과 그 binary question에 대한 평가 함수를 필요로 하며 분할이 언제 멈추어야 하는지에 대한 기준도 설정해야 한다. 부분집합으로 분리했을 때 평가함수를 통한 관측 확률 값의 증가가 미리 정의한 임계값보다 작아지는 시점에서 분할을 멈추게 된다. 결국 최종적인 부분집합 내의 state들이 tying된다.

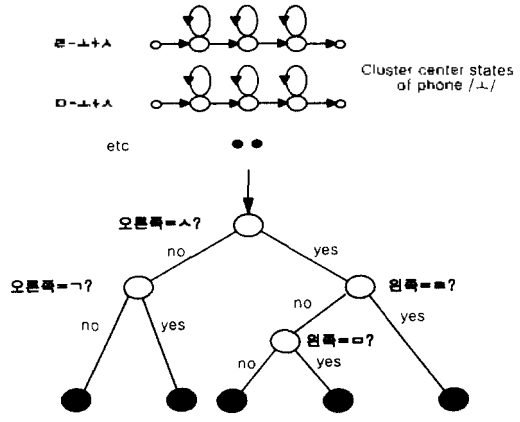

그림 2-2. Decision tree-based state tying 과정

3. 음성 데이타베이스

가변어휘 인식시스템을 구축하기 위해 사용된 데이터는 한국전자통신연구원(ETRI)에서 설계 및 수집한 대용량 음성 DB인 3,848개 어휘를 가지는 POW 3848 DB를 훈련에 사용하였다[1]. 3,848개 어휘를 8명이 481개씩 나누어 발성한 것을 1 개의 set으로 하고, 남성 5 set (40명)과 여성 5 set (40명)으로 모두 합하면 총 10 set (80명, 약 38,480개 단어)이 된다. 이 중 남성만 훈련에 사용하였고, 5 set 중 수작업으로 음소 경계가 표시된 3 set으로 모델을 초기화한 다음 전체 5set으로 최종 모델을 훈련하였다.

인식 실험에는 어휘독립성을 위해 POW DB와 관계없는 22개의 부서명을 50명이 발음한 부서명 DB를 사용하였다.

4. 실험 및 결과

기본 유사음소 집합을 묵음을 포함하여 46개로 선정한 후, 훈련용 DB에서 만들어진 9,908개(27,294개 states)의 triphone을 tree-based clustering 방법을 사용하여 state를 tying함으로써 실제 사용되는 파라메터의 수를 줄였다. Clustering 과정에서 부분집합의 분할을 결정하는 임계값을 변화해가면서 인식률을 평가하였다. 이 때 인식 실험용 DB에서 발견된 triphone의 개수는 128개 (384개 states)이며 이중 7개(21개 states)는 unseen triphone이다. 즉, unseen triphone이 차지하는 비율이 5%인 부서명 DB[2]를 사용하여 인식 실험을 하였다.

문맥 독립형 음소모델의 경우, 46개 monophone을 이용하여 88.6%의 인식률을 얻었고, 문맥 종속형 음소모델의 경우, 9,908개의 triphone을 state-tying하여 인식 실험한 경우는 95 ~ 96%의 높은 인식율을 보였다. 임계값에 따른 인식 결과는 표 4-1과 같다.

표 4-1 임계값에 따른 triphone 모델 인식율

문맥 종속형 음소모델 (27,294 states)		
임계값	인식율(%)	state # (tied ratio)
50	96.09	4207 (15.4%)
100	96.18	3292 (12.1%)
150	96.09	2539 (9.3%)
200	96.18	2044 (7.5%)
250	95.27	1729 (6.3%)
300	95.73	1518 (5.6%)
350	95.73	1351 (4.9%)

5. 결 론

본 논문에서는 다양한 응용분야의 각종 명령어를 변경 및 추가가 가능한 화자독립 가변어휘 고립음성 인식시스템을 구현하였다. 개발된 시스템은 POW DB을

이용하여 triphone model을 훈련하였다. 부서 DB를 이용하여 어휘독립 음성인식 실험을 하였다. 음성특징 파라메터는 잡음환경등에 강인한 것으로 알려진 MFCC를 사용하였고, 섬세한 음소 모델링이 가능한 연속확률분포 HMM을 기반으로 하여, 특히 음소모델의 강인한 훈련 및 test 환경의 unseen triphone문제를 해결하기 위해 tree-based clustering 기법을 도입하였다. 화자독립 가변어휘 인식결과, 95 ~ 96%의 높은 성능을 나타내었다.

앞으로, 인식성능 개선 및 가변어휘 연속음성 인식 시스템 구현을 위해 DB구축 및 추가 실험이 앞으로의 과제이다.

참 고 문 헌

[1] Yeonja Lim and Youngjik Lee, " Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," ICASSP95, In Proc. pp.89~91.

[2] 이영직 외, "ETRI의 음성 데이터베이스 구축 현황," 제12회 음성통신 및 신호처리 워크샵 논문집, pp.265~267, 1995년 6월.

[3] S. Young, " HTK: Hidden Markov Model toolkit V2.0," Eng. Dept., Speech Group, Cambridege, Univ., Cambridege UK, Tech., Rep., 1992.

[4] L. Villarrubia, L.H. Gomez, J.M. Elvira, H.C. Torrecill," Context Dependent Units for Vocaburary -Independent Spanish Recognition," In Proc. ICASSP96, pp.451~454.

[5] 임연자, 이영직, " Large scale word recognizer를 위한 음성 database POW," 12회 음성통신 및 신호처리 워크샵 논문집, pp.291~294, 1995년 6월.

[6] L. R. Bahl, P. V. de Souza, et al.," Decision trees for phonological rules in continuous speech." In Proc. ICASSP97, pp.185~188.

[7] H. J. Nock, M. J. F. Gales, S.J. Young, " A comparative study of methods for phonetic decision tree state clustering," In Proc. EUROSPEECH97, vol.1, pp.111~114.