1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

어휘독립 환경에서의 가변어휘 음성인식에 관한 연구

황병한, 김형순
부산대학교 전자공학과

A Study on the Variable Vocabulary Speech Recognition in the Vocabulary-Independent Environments

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

요약

본 논문은 어휘독립(Vocabulary-Independent) 환경에서 별도의 훈련과정 없이 인식대상 어휘를 추가 및 변경할 수 있는 가변어휘(Variable Vocabulary) 음성인식에 관한 연구를 다룬다. 가변어휘 인식은 처음에 대용량 음성 데이터베이스(DB)로 음소모델을 훈련하고 인식대상 어휘가 결정되면 발음사전에 의거하여 음소모델을 연결함으로써 별도의 훈련과정 없이 인식대상 어휘를 변경 및 추가할 수 있다.

문맥 종속형(Context-Dependent) 음소 모델인 tri-phone을 사용하여 인식실험을 하였고, 인식성능의 비교를 위해 어휘종속 모델을 별도로 구성하여 인식실험을 하였다. Unseen triphone 문제와 훈련 DB의 부족으로 인한 모델 파라메터의 신뢰성 저하를 방지하기 위해 state-tying 방법 중 음성학적 지식에 기반을 둔 tree-based clustering(TBC) 기법[1]을 도입하였다.

Mel Frequency Cepstrum Coefficient(MFCC)와 대수에너지에 기반을 둔 3 가지 음성특징 벡터를 사용하여 인식 실험을 병행하였고, 연속 확률분포를 가지는 Hidden Markov Model (HMM) 기반의 고립단어 인식시스템을 구현하였다.

인식 실험에는 22 개 부서명 DB[3]를 사용하였다. 실험결과 어휘독립 환경에서 최고 98.4%의 인식률이 얻어졌으며, 어휘종속 환경에서의 인식률 99.7%에 근접한 성능을 보였다.

I. 서론

현재 상용 중인 대다수 음성인식 시스템은 인식대상 어휘들이 미리 정해져 있는 어휘종속 환경에서의 음성인식 시스템이다. 즉 사전에 결정된 인식대상 어휘를 대상으로 음성 DB를 구축하고, 이 음성 데이터베이스를 토대로 단어모델을 훈련한다. 이와 같은 방식의 음성인식 시스템은 선정한 인식대상 어휘들에 대해서는 높은 인식성능을 보이지만 인식대상 어휘를 변경하거나 추가가 필요할 경우, 새로운 어휘에 대해 별도로 음성 DB를 수집하고 처음부터 모델을 다시 훈련해야 하는 문제점이 발생한다.

여기서 사용되는 가변어휘 인식기술은 이러한 문제

점을 해결하고 화자독립 및 어휘독립 환경에서의 음성인식을 수행함으로써 다양한 응용분야의 각종 명령어를 음성인식에 의해 대체함으로써 사용자에게 편리한 입력 인터페이스를 제공해 줄 수 있다.

본 논문에서는 이러한 가변어휘 인식시스템에 사용될 최적의 음성특징 벡터의 선정 및 Gaussian Mixture의 수를 결정하기 위해 다양한 인식실험을 병행하여 인식성능의 변화를 살펴보았다.

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.

II. 가변어휘 인식시스템

2.1 인식시스템의 구성

인식대상 어휘가 결정되면 발음사전을 구성하고 미리 구성한 음소모델을 발음사전에 의거하여 음소모델을 연결하여 인식 네트워크를 형성하여 인식대상 어휘가 추가 및 변경된 경우에 적용될 인식시스템을 결성한다. 훈련 및 인식에 사용되는 음성특징벡터는 입력된 음성신호를 16bit, 16kHz로 샘플링한 후 10 msec 마다 20 msec 길이의 프레임 단위로 추출하여

- 13차 (12차 MFCC 및 대수에너지)
- 24차 (12차 MFCC 및 12차 delta MFCC)
- 25차(12차 MFCC, 12차 delta MFCC ,delta 에너지)

를 각각 사용하여 인식실험을 병행하였다.

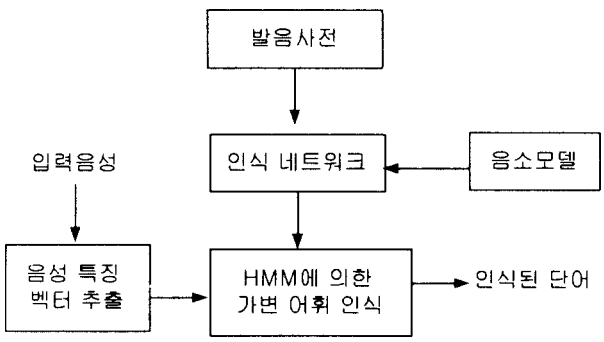

그림 1. 가변어휘 인식시스템의 구성

음소 모델은 통계적 자료에 기반한 모델로, 세 개의 상태(state)를 가지는 left-to-right HMM으로 구성하였다. 각 상태는 Gaussian 확률 밀도 함수를 가지고 관측 벡터의 발생 확률을 계산하도록 하였다. 각 음소모델은 상용 음성인식 tool인 HTK(Hidden Markov Model Tool Kit)[4]를 사용하여 모델링 하였다.

2.2 음소 모델링

음소 모델은 기본단위로 음소, 유사음소(PLU, phoneme-like unit), 변이음 등이 사용될 수 있다. 본 논문에서는 유사음소를 기본단위로 정하였고, 한국어 기본 음소단위에 유성음화, 불파음화, [r]의 [l]되기 등 3가지 규칙을 적용하였으며[5], '네'와 '내', '녜'와 '냬', 그리고 '니', '게', '새'를 각각 하나로 묶이서 45개의 유사음소 집합을 정하였다. 이 유사음소 집합에 묵음을 포함하여 총 46개를 기본 유사음소 집합으로 선정하였다.

앞뒤의 음소 정보를 포함하는 triphone의 경우, 예를 들어 기본 유사음소집합의 개수를 40개로 가정하면 46*46*46 = 97,336개의 triphone이 가능하다. 한국어의 경우 음성학적으로 발생되는 경우만 고려하더라도 대략 30,000여 개의 triphone이 나타나기 때문에 신뢰할 수 있는 모델 파라메터의 추정을 위해 방대한 훈련용 데이터베이스가 필요하게 된다. 그러나 모든 한국어 음운현상을 포함하는 음성 DB를 구축하는 일은 현실적으로 어려운 문제이므로, 음운현상을 충분히 포함하면서도 효율적인 크기를 갖는 음성 DB 구성 방법에 관한 연구도 많이 선행되었다[6]. 이와 같이 구축된 대용량 음성 DB에서도 인식과정에서만 나타나는 unseen triphone이 존재할 수 있다.

본 논문에서는 한국어 음운현상이 잘 반영된 대용량 음성 DB인 ETRI의 POW 3848 DB를 이용하였다[2]. 신뢰성 있는 모델 파라메터 추정 및 unseen triphone 문제를 해결하기 위해 TBC 기법[3]을 도입하였다. 이 방법은 훈련 시 유사한 통계적 특성을 갖는 음소 모델의 state들을 하나의 그룹으로 묶음으로서 전체 모델의 파라메터 수를 줄여 상대적으로 신뢰도를 높이는 State

tying 방법의 하나이다.

2.3 Tree-based Clustering(TBC)[1]

State tying 에는 data-driven clustering(DDC) 기법과 TBC 기법이 있다[1][4][7]. DDC 방법은 훈련용 데이터에 포함된 triphone 모델만 state tying 하므로 신뢰성 있는 모델 파라메터를 추정할 수는 있지만 unseen triphone 문제는 여전히 남아 있다. 이 문제를 해결하기 위해 또 다른 state tying 방법인 TBC 기법을 적용하였다. TBC 기법에서는 먼저 동일한 음소에 해당하는 모든 triphone 모델의 상태들을 함께 모은 다음, 음운론에 기반하여 미리 정해둔 binary question 들 중에서 분할 후의 누적 확률값을 최대로 하는 binary question 을 선택하여 분할한다. 나뉘어진 두 개의 부분집합을 다시 두 개의 부분집합으로 나누어 가는 일련의 과정을 통해 결정 트리를 구성한다. 분할되는 것을 노드라 하고, 노드의 분할 종료는 분할 후의 누적 확률값의 증가가 미리 정해둔 문턱값 보다 작은 경우에 행해진다. 이러한 과정을 거쳐 분할이 일어나 지 않는 경우에 최종적인 결정 트리가 형성되고 이를 기준으로 최종적인 부분집합 내의 state 들이 tying 된다.

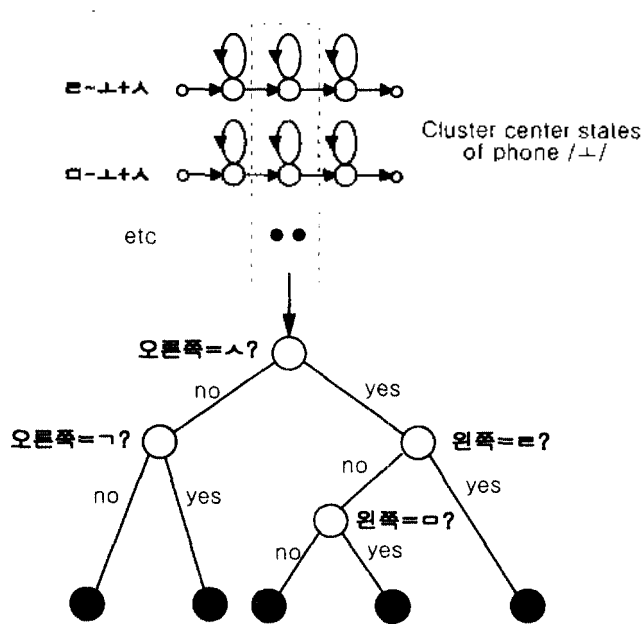

그림 2. Decision tree-based state tying 과정

III. 음성 데이터베이스

가변어휘 인식시스템을 구축하기 위해 사용된 음성 DB 는 한국 전자통신 연구원(ETRI)에서 설계 및 수집한 대용량 음성 DB 인 3,848 개 어휘를 가지는 POW 3848 DB 를 훈련에 사용하였다[2]. 3,848 개 어휘를 8 명이 481 개씩 나누어 발성한 것을 1 개의 set 으로 하고, 남성 5 set (40 명)과 여성 5 set (40 명)으로 모두 합하면 총 10 set (80 명, 약 38,480 개 단어)이 된다. 이 중 남성만 훈련에 사용하였고, 5 set 중 수작업으로 음소 경계가 표시된 3 set 으로 모델을 이용하여 초기 음소모델을 구성하여 모델 파라메터의 초기화 문제를 해결하였고 이 초기 모델을 전체 5 set 으로 다시 훈련하여 최종 모델을 구성하였다.

인식 실험에는 어휘독립성을 위해 POW DB 와 관계없는 22 개의 부서명을 50 명이 발음한 부서명 DB[3]를 사용하였다. 인식 성능을 비교하기 위해 어휘종속 실험을 별행하였다 어휘독립 실험에 사용된 22 개 부서명 DB 에서 50 명 남성 중 35 명을 어휘종속 음소모델 훈련에 사용하였고, 나머지 15 명으로 인식실험을 하였다.

IV. 실험 및 결과

기본 유사음소 집합을 묵음을 포함하여 46 개로 선정한 후, 훈련용 DB 에서 만들어진 9,908 개(27,294 개 states)의 triphone 을 TBC 방법을 사용하여 state 를 tying 함으로써 실제 사용되는 파라메터의 수를 줄였다. 인식실험은 2.1 절에서 정의한 3 가지 음성특징벡터를 가지고 상태 확률밀도함수를 나타내는 Gaussian 의 개수를 1 에서 6 개로 변화시키면서 진행하였다. 이 때 인식 실험용 DB 에서 발견된 triphone 의 개수는 128 개 (384 개 states)이며 이중 7 개(21 개 states)는 unseen triphone 이다. 즉, unseen triphone 이 차지하는 비율이 5%인 부서명 DB 를 사용하여 인식 실험을 하였다. 훈련용 음성 DB 에서 나타난 9,908 개의 triphone 을 state-

tying 하여 98.4%의 인식 성능을 보였다. 특히, 어휘종속 환경에서 99.7%의 인식성능에 근접한 성능을 보였다. 표 1 과 표 2 에서 MFCC 는 12 차 MFCC 를, D 는 delta MFCC 를, E 는 대수 에너지를, 그리고 dE 는 delta 에너지를 나타낸다.

표 1. 어휘독립 환경에서의 triphone 모델 인식율

(%)

Feature	Gaussian Mixture #					
	1	2	3	4	5	6
MFCC_E	97.0	98.4	98.2	98.3	97.6	98.0
MFCC_D	97.3	97.2	97.6	97.4	97.6	97.7
MFCC_D_dE	97.6	97.6	98.0	97.7	98.4	97.9

표 2. 어휘종속 환경에서의 triphone 모델 인식율

(%)

Feature	Gaussian Mixture #					
	1	2	3	4	5	6
MFCC_E	97.3	97.9	97.9	97.9	97.3	97.3
MFCC_D	98.9	99.7	99.4	99.4	99.4	99.4
MFCC_D_dE	99.1	99.4	99.7	99.1	99.7	99.4

V. 결론

본 논문에서는 어휘종속 인식시스템의 성능에 근접하는 화자독립 가변어휘 고립단어 인식시스템을 구현하였다 실험결과 음성특징 벡터나 Mixture 의 수에 따른 뚜렷한 경향성은 나타나지 않으나, 음성특징 벡터의 경우 어휘독립 환경에서는 델타 MFCC 에 비해 에너지나 델타에너지의 영향이 큰 것으로 보이고, 어휘종속 환경에서는 반대의 경향을 보였고, Mixture 의 수는 2~3 개가 적당한 것으로 보인다. 실험에 사용된 인식 시스템은 POW DB 을 이용하여 다양한 한국어 음운현상이 반영된 대용량 음성 DB 구축 문제를 해결하였고, 음운론적 지식에 기반을 둔 TBC 기법을 도입하여 unseen triphone 문제 및 모델 파라메터의 신뢰성을 높였다. 앞으로 대단위 가변어휘 인식기 구현을 위해 음성 DB 의 크기가 크고 unseen triphone 의 비율이 높은 평가용 DB 를 사용하여 추가로 인식실험을 할 예정이다.

참고문헌

[1] L. R. Bahl, P. V. de Souza, et al., "Decision trees for phonological rules in continuous speech," In Proc. ICASSP 97, pp.185~188.

[2] Yeonja Lim and Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," ICASSP 95, In Proc. pp.89~91.

[3] 이영직 외, "ETRI 의 음성 데이터베이스 구축 현황," 제 12 회 음성통신 및 신호처리 워크샵 논문집, pp.265~267, 1995 년 6 월.

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

[5] 유재원, "연속음성인식을 위한 음성 단위 발음사전 구성방법 연구," 위탁과제 최종 연구보고서, 한국전자통신연구원, 1995 년 1 월.

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

[7] H. J. Nock, M. J. F. Gales, S.J. Young, "A comparative study of methods for phonetic decision tree state clustering," In Proc. EUROSPEECH 97, vol.1, pp.111~114.