분산음성인식을 위한 내장형 고속/경량 음소인식기 개발

김승희, 황규웅, 전형배, 정훈, 박준
한국전자통신연구원 음성/언어정보연구센터
e-mail : { seunghi, kyuwoong, hbjeon, hchung, junpark }@etri.re.kr

Development of Embedded Fast/Light Phoneme Recognizer for Distributed Speech Recognition

Seung Hi Kim, Kyuwoong Hwang, Hyunbae Jeon, Hoon Jeong, Jun Park
Speech/Language Information Research Center, ETRI

요 약

ETRI 음성/언어정보연구센터에서는 분산음성인식을 위해 메모리를 작게 사용하며 속도가 빠른 음소인식기를 개발 중이다. 음향 모델, 언어 모델, 탐색 네트워크 등 고정되어 있는 정보는 인식기를 수행하기 이전에 미리 binary 형태로 구축하여 ROM 형태로 저장함으로써 실제 사용해야 할 RAM 용량을 대폭 줄일 수 있었다. Tied state 에 기반한 triphone 모델에서는 unique HMM 만을 사용함으로써 인식시간 및 메모리 사용량을 대폭 줄일 수 있었다. Monophone 인식기의 경우 RAM 사용량이 179KB 였으며, triphone 인식기의 경우 435KB 의 RAM 사용량과 RTF(Real Time Factor) 0.02 를 확인하였다.

1. 서 론

음소인식기는 2-pass decoding 구조에서 acoustic decoder 로서 사용되기도 하고[1], 혹은 소형의 인식기가 필요한 내장형 시스템 응용 분야에서 사용될 수도 있다. 한국전자통신연구원에서는 2-pass 구조의 acoustic decoder 로 소형 단말기에서도 음성인식이 가능하도록 내장형 고속/경량의 음소 인식기를 개발 중이다[2].

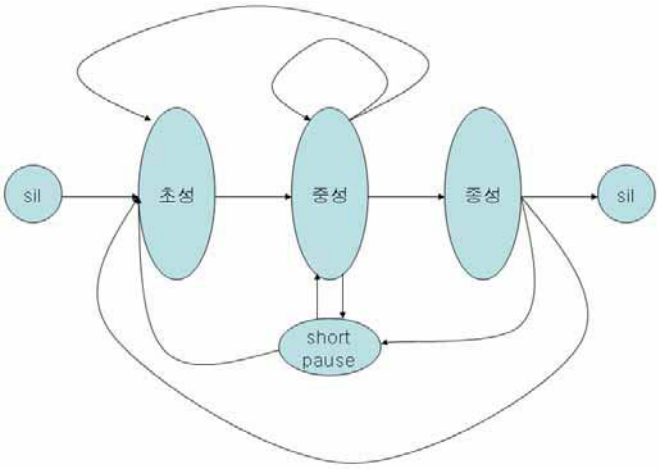
본 논문에서는 고속/경량 음소인식기의 개발 단계별 성능에 중점을 두고 기술한다. 서론에 이어 2 장에서는 음향/언어모델 훈련 및 범용의 인식 시스템을 이용한 기본 인식 실험에 대해 간략히 설명하고, 3 장에서 고속/경량 음소인식기에 대해 개발 단계별로 설명한다. 그리고, 4 장에서 결론을 맺는다.

2. 음향/언어 모델 훈련 및 기본 인식 실험

음소 모델을 훈련하기 위해서 POW 및 단어 19 만 발화와 4 만 발화의 문장을 사용하였다. 인식 실험을 위한 테스트 DB 로는 여행계획 분야의 대화체 문장 690 발화(약 50 분 분량)를 사용하였다.

매 10ms 마다 20ms 분량의 320 개의 음성 신호 sample 로부터 30 차의 MFCC 를 추출하였다. 그리고, HTK 를 사용하여 음소당 3 개의 state 를 가지는 음소 HMM 을 훈련하였으며, 문맥 종속 모델로는 decision-tree 기반의 tied-state triphone 을 사용하였다.

기본적으로 탐색 네크워크는 그림 1 과 같은 음절 FSN(Finite State Network)의 구조를 가진다.



(그림 1) 한국어 음절 FSN

기본 인식 실험을 위해 HTK 의 HVite 와 유사한 성능을 내는 센터 내부의 범용 인식기를 사용하였다.

<표 1> Monophone 모델을 사용한 음소인식 결과

Net-work	Mix.	Correct-ness (%)	Accuracy (%)	RTF	Mem (MB)
all	128	62.00	47.63	0.15	23.4
syllable	128	62.51	49.75	0.14	18.3
syllable	512	64.59	52.83	0.66	33.5

표 1 에서 Network 중 'syllable'은 그림 1 의 음절 FSN 을 사용한 경우이며, 'all'은 음절 FSN 을 사용하지 않고 모든 음소의 연결이 가능한 탐색 네트워크를

의미한다. 위 실험 결과는 beam pruning 을 적용한 결과이며 beam width 는 100.0 이다. Beam width 가 100.0 이상이 되면 인식률에 거의 변화가 없었다. 이후 모든 실험결과의 beam width 는 100.0 이다.

인식 실험에 사용된 컴퓨터는 2 개의 Xeon 3.0GHz CPU 를 사용한다. 언어모델은 음향 모델 훈련 DB 의 전사문을 사용하여 훈련하였다.

3. 내장형 고속/경량 인식기

인식기에서 사용되는 메모리는 크게 2 가지 형태의 정보를 저장한다. 하나는 인식과정에 상관없이 고정된 정보로써 음향모델, 언어모델, 탐색네트워크 등이 여기에 포함된다. 이러한 정보들은 내장형 시스템에서 ROM 형태로 저장이 가능하다. 다른 하나는 인식과정 중에 수시로 변하는 정보로써 active path 들의 각종 정보들이 여기에 해당한다. 이러한 정보들은 RAM 에 저장된다. 이와 같이 두 가지 정보를 분리하여 다룸으로써 인식기 로딩 시간을 대폭 줄일 수 있었고 내장형 시스템에 탑재하는 경우 RAM 사용량을 대폭 줄일 수 있다.

우선 문맥 독립 음소모델을 사용하여 인식실험을 수행하였다. 음절 네트워크로는 그림 1 의 FSN 을 사용하였고 state 당 128 개의 Gaussian mixture 를 사용하지만 실제 관측확률 계산을 위해서는 입력된 특징벡터와 가장 가까운 하나의 Gaussian 만을 사용하였다.

<표 2> Monophone 모델을 사용한 음소인식 결과

Insertion penalty	Corr.(%)	Acc. (%)	RTF
0	62.56	45.21	0.1
25	57.84	53.04	0.09
45	53.24	50.91	0.08

<표 3> Monophone 모델을 사용할 때의 메모리 소요량

ROM	음향모델	4.2MB
	음소 천이 모델	6KB
	프로그램 코드	30KB
	계	4.2MB
RAM	고정	6KB
	HMM end information	173KB
	계	179KB

다음으로 triphone 모델을 사용하여 인식 실험을 수행하였다. 음절 네트워크로는 그림 1 의 음절 FSN 을 좌우 문맥 확장하여 사용하였고, state 당 16 개의 Gaussian mixture 를 사용하지만 관측확률 계산에는 입력된 특징벡터와 가장 가까운 하나의 Gaussian 만을 사용하였다.

<표 4> Triphone 모델을 사용한 음소인식 결과

Insertion penalty	Corr. (%)	Acc. (%)	RTF	RAM	ROM
40	62.56	45.21	0.2	7.1 MB	6.4 MB

Tied-state 기반의 triphone 모델의 경우에는 triphone 이름은 달라도 모델 파라메터 자체는 동일한 집합들이 존재한다. 위의 triphone 모델 실험에서 실질적으로 동일한 triphone 들을 각각의 대표 triphone (unique triphone)으로 묶게 되면 총 48627 개에서 2955 개의 triphone 으로 줄어든다. Triphone 으로 확장된 음절 네트워크에서 중복되는 triphone 들을 하나로 묶게 되면 네트워크의 크기가 줄어들며, 인식시간도 단축된다. 이렇게 탐색 공간을 구축하여 인식실험을 하였다. 이 경우 문맥에 정합한 탐색은 다소 흐트러질 수 있기 때문에 인식률이 하락할 수도 있다. 그러나 실험 결과에서는 인식률의 하락은 거의 없었다.

그리고 언어모델을 적용하여 실험을 하였다. 언어모델로는 monophone bigram 과 unique triphone 기반의 triphone bigram 을 사용하였다.

<표 5> Unique triphone 에 기반한 탐색공간에서의 인식 결과

LM	Pen.	Scale	Corr.(%)	Acc.(%)	RTF	RAM
-	40	-	64.13	57.83	0.03	760KB
Mono	5	30	68.69	61.52	0.03	570KB
Tri	15	20	69.53	63.9	0.02	435KB

LM 은 적용한 언어모델의 종류를 나타내며, Pen.은 phone insertion penalty, Scale 은 언어모델값의 가중치를 나타낸다. Insertion penalty 가 커질수록 삽입 오류는 줄어드는 반면 삭제 오류가 늘어나는데, 언어모델 값의 가중치를 크게 할 때도 비슷한 경향을 보인다.

참고로 위 실험들에서 내장형 음소인식기에서는 인식과정에서 transition probability 를 적용하지 않았다. 그 이유는, 고립 단어 혹은 연결 단어 인식 실험에서 transition penalty 가 인식률에 미치는 영향이 미미하였기 때문이다. 그러나 monophone 모델을 사용한 기본 실험에서 무시할 수 없는 성능 차이가 관측이 되어 이에 대한 추가실험 및 분석이 필요하다.

4. 결 론

본 논문에서는 고속/경량의 한국어 음소인식기를 개발하는 과정에서 각 단계별 인식기의 성능에 중점을 두어 기술하였다. 인식기에서 사용하는 정보들을 ROM 에 기록 가능한 정보와 RAM 에 기록해야 하는 정보를 분리함으로써 RAM 사용량을 대폭 줄일 수 있는 형태를 개발하였고, tied state triphone 모델을 사용하는 경우에 unique model 로만 구성된 탐색 네트워크를 구성함으로써 인식률의 하락 없이 메모리 사용량 및 인식 속도를 대폭 개선할 수 있었다.

참고문헌

[1] 김승희 외, “음소 인식 시스템의 인식 오류 분석,” 제 23 회 음성통신 및 신호처리 학술대회, 한국, 2006

[2] Kyuwoong Hwang, et al., “Distributed speech recognition for spoken query of internet by lexical access,” Interspeech’07, Antwerp, Belgium, 2007. Submitted.