1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

휴대 멀티미디어 단말용 음성인식 시스템 개발

김승희, 송명규, 정용원, 하태호, 김형순
부산대학교 전자공학과

Development of Continuous Speech Recognition System for Multimedia Mobile Terminal Applications

Seunghi Kim, Myung Gyu Song, Yong Won Jeong, Tae Ho Ha, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
cygnus@hyowon.cc.pusan.ac.kr

요 약

본 논문에서는 한국전자통신연구원의 Handy Combi 응용 도메인을 대상으로 한 화자독립 연속음성인식 시스템 개발에 관하여 기술한다. 불특정화자가 자연스럽게 발음한 연속음성을 인식하는 기술은 핸인식 등과 더불어 멀티모달 인터페이스의 핵심 요소로서, 이동 환경에서 사용자의 다양한 요구사항을 처리하는 지능형 에이전트의 구현을 위해 필수적으로 개발되어야 하는 기술이다. 본 논문에서는 연속확률분포를 가지는 Hidden Markov Model(HMM) 기반의 연속음성인식 시스템을 구현하였다. 개발된 시스템은 음성특징벡터로 MFCC 를 사용하였으며, 음소 모델의 강인한 훈련을 위해 음성학적 지식에 기반을 둔 tree-based clustering 방식을 도입하였다. 인식단계에서는 인식속도를 개선시키기 위해 beam-search 기법을 적용하였다. 인식 실험 결과, 99.7%의 어절 인식률과 98.8%의 문장 인식률을 얻었으며, 최종적인 문장의 이해도는 99% 이상이었다.

1. 서 론

음성인식 기술은 음성합성 기술과 함께 인간의 가장 편리한 의사전달 수단인 음성을 통해 인간이 컴퓨터와 대화를 할 수 있도록 해주는 도구로서 정보화의 진전과 더불어 그 필요성이 더욱 증대되고 있다. 특히 임의의 화자가 자연스럽게 발음한 연속음성을 인식하는 화자독립 연속음성인식 기술은 사용자에게 편리한 입력 인터페이스를 제공해 줄 수 있어서 매우 유용하다. 사람이 문장구조 등에 제한없이 자연스럽게 발음한 연속음성에 대한 인식기술은 아직 성능면에서 뒤떨어지고 있으나, 제한된 도메인에서의 음성인식 기술은 어느 정도 실용화가 이루어지고 있다.

본 논문에서는 휴대 멀티미디어 단말에 적용할 수 있는 불특정화자 연속음성인식 시스템 개발에 관하여 기술하고 있다. 이동 환경에서 사용자의 다양한 요구사항을 처리해야 하는 지능형 에이전트를 구현하기 위해서는 화자에 상관없이 자연스럽게 발음한 연속음성을 인식하는 것이 반드시 필요하다. 이를 위해 본 논문에서는 HMM 을 이용하여 Handy Combi 응용도메인을 대상으로 한 음성인식 시스템을 구현하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 장에서는 HTK 기반의 연속음성인식 시스템에 대해 기술하며, 3 장에서 인식실험과정 및 결과를 다룬 후, 4 장에서 결론을 맺는다.

2. 연속음성인식 시스템의 구성

2.1 Task Domain

본 논문에서 대상으로 삼은 task domain 은 Handy Combi 응용 도메인(PDA domain)으로서, 모두 20 개의 문장생성규칙과 394 개의 vocabulary size 를 가지고 있다. 이에 대한 언어모델의 perplexity 는 10.3 이다. 인식 대상 문장의 전형적인 예는 다음과 같다.

이 지역에 있는 식당들을 보여 주세요
대전역에서 5 킬로 내에 있는 호텔을 보여줘
여기서부터 에트리까지의 거리가 얼마지?
선택된 메뉴를 삭제해줘
해외출장 신청서의 현재상태를 알려 주십시오.

2.2 시스템 개요 및 특징

휴대 멀티미디어 단말용 연속음성인식 시스템은 주어진 도메인 내에서 자연스럽게 발음된 문장을 인식하도록 구현되었다. 또한 문맥종속 음소모델, 즉 triphone 모델을 사용하였기 때문에 새로운 어휘 및 문장형태를 추가할 경우에도 단지 발음사전과 태스크 문법만 수정하면 된다. 다만, 본 시스템에서의 문맥종속 음소모델은 제한된 도메인을 대상으로 한 음성 데이터베이스에 의해 훈련되었기 때문에, 도메인에 포함되어 있지 않은 문맥종속 음소모델이 필요한 어휘에 대해서는 성능 저하가 불가피하다.

음성특징벡터로는 12 차 MFCC 및 12 차 delta MFCC 로 구성된 24 차의 벡터를 사용하였으며, 각 음소 모델은 세 개의 상태(state)를 가지는 left-to-right HMM 모델로 구성하였다. 각 상태는 단일 Gaussian 확률 밀도 함수를 가지고 관측 벡터의 발생 확률을 계산하도록 하였다. 각 음소는 상용 음성인식 tool 인 HTK 을 사용하여 모델링하였다[1].

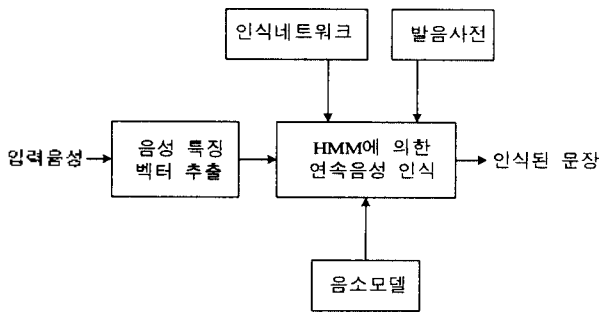

그림 2-1. 연속음성인식 시스템의 구성

2.3 Tree-based Clustering

본 논문에서는 연속음성인식에서의 자연스러운 발화를 고려하기 위해 한국어 표기를 자동으로 발음표기로 변환하는 프로그램을 이용하여 훈련용 문장 내에서 어절간의 연음 효과를 고려한 다음 단 음소를 생성하였고, 음소 모델링 중 문맥종속 음소 모델링에 기반을 둔 인식 실험을 수행하였다. Triphone 모델은 음소 모델보다 단어 내의 음운 현상을 효과적으로 반영하는 장점이 있지만 신빙성 있는 모델 파라메터를 추정하기 위해서는 방대한 훈련 데이터가 필요하게 되는 문제점이 있다. 이에 따라 본 논문에서는 훈련시 유사한 특성을 가지는 state 들을 하나의 그룹으로 묶어 state-tying 을 수행하였다.

State tying 에는 data-driven clustering 방법과 tree-based clustering 방법이 있다[1][2][3]. Data-driven clustering 방법은 훈련용 데이터에 포함된 triphone 모델만 state-tying 하므로 대단위 어휘 시스템에서는 훈련용 데이터에 포함되지 않은 triphone(unseen triphone)이 발생할 수 있다. 본 논문에서는 이 문제를 해결하기 위해 또 다른 state-tying 방법인 tree-based clustering 방법을 적용하였다.

Tree-based state tying 에서는 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 question 과 그 binary question 에 대한 평가 함수를 필요로 하며 분할이 언제 멈추어야 하는지에 대한 기준도 설정해야 한다. 부분집합으로 분리했을 때 평가함수를 통한 관측 확률 값의 증가가 미리 정의한 임계 값보다 작아지는 시점에서 분할을 멈추게 된다. 결국 최종적인 부분집합 내의 상태들이 tying 된다.

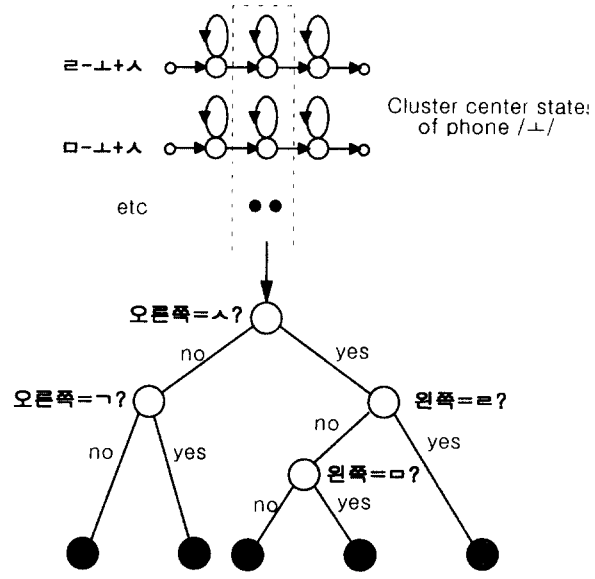

그림 2-2. Decision tree-based state tying 과정

그리고 전체 triphone 목록은 훈련용 데이터와 관계없이 만들고 이를 state-tying 하게 되므로 data-driven 에서 발생하는 unseen triphone 문제를 해결할 수가 있다.

3. 인식실험 결과 및 고찰

본 논문에서 사용한 음성 데이터베이스는 KOREAN SPEECH DB(ETRI Wonkwang SPEECH DB)[4]중 PDA 용 문장과 기타 단일 토큰 음성을 이용하였다. KOREAN SPEECH DB 에는 110 명 분의 데이터가 있으며, 훈련 데이터는 90 명 분의 PDA 9,000 문장과 단일 토큰 26,837 단어, 그리고 평가용 데이터는 PDA 2,000 문장으로 되어 있다. 훈련 데이터는 남성화자 45 명, 여성화자 45 명이 발음한 것이며, 평가용 데이터는 남성화자 10 명, 여성화자 10 명이 발음한 것이다. 음성 데이터는 16kHz 및 16bits 로 샘플링되었으며, 8kHz 로 down-sampling 시킨 데이터에 대한 인식실험도 함께 수행하였다. 그리고 각 음성 파일에 해당하는 한국어 transcription 정보가 제공된다. 각 utterance 의 앞뒤에는 묵음(silence)이 온다고 가정하였으며, 이 묵음 모델은 그림 3-1 과 같이 상태 천이를 하도록 했다. 이는 각 상태가 훈련용 데이터 내에 포함될 수 있는 impulsive 잡음에 대해 후방 천이(backward skip)를 허용함으로써 음소 모델로 강제 천이 하는 것을 방지할 수 있게 한다. 그리고 문장내의 각 단어 사이에 있을 수 있는 짧은 pause 를 표현하기 위해 묵음(silence) 모델의 가운데 상태와 동일한 상태를 가지는 short pause(sp)모델을 두었다.

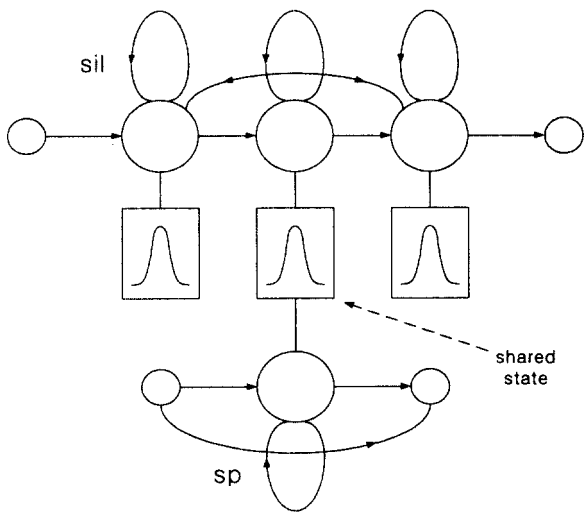

그림 3-1. 묵음 모델(silence model)

연속음성인식 시스템의 실제적인 구현을 위해서는 real-time 처리가 필수적이다. 본 논문에서는 인식률의 저하를 최소로 하면서 인식속도를 개선하기 위해 beam search 기법을 사용하였으며, 다양한 beam width 에 대해 인식률과 인식시간을 측정하였다.

먼저 표 3-1 에 beam search 를 적용하지 않았을 때의 인식결과가 나타나 있다. 남성화자의 경우 16kHz 샘플링 음성에 대해서 98.8%, 8kHz 에서는 97.1%의 문장인식률을 얻었으며, 여성화자의 경우에는 16kHz 에서 98.2%, 8kHz 에서는 91.3%의 문장인식률을 보인다. 남녀 통합모델을 사용했을 경우는 이보다 성능이 좀 떨어져서 16kHz 및 8kHz 샘플링에 대해 각각 96.4%, 88.6%의 인식률을 나타내었다. 여성모델과 남녀 통합모델의 경우 8kHz 샘플링에서는 인식률이 상당히 떨어졌다.

그리고 표 3-2 에서 보는 바와 같이 beam search 기법을 사용한 경우 인식률은 거의 감소하지 않고 상당한 속도의 개선을 얻을 수 있었다. 그림 3-2 를 보면 beam width 100 정도만 되어도 인식률은 거의 수렴하는 것을 알 수 있다. 반면 인식소요시간은 beam width 와 거의 비례하는 것으로 나타난다(그림 3-3). 참고로 file 당 평균 인식소요시간을 계산하여 보았다. UltraSparcII 167MHz 기종을 사용하여 평균 길이가 279frame(2.79 초)인 1000 개의 wave file 에 대해서 인식실험을 하였다. beam width 를 200 으로 한 경우에 file 당 평균 인식소요시간은 1.194 초로서 1 초당 0.428 초가 소요되는 계산량이다.

표 3-1. 각 모델별 인식률 (beam width ∞)

	남 성		여 성		통 합	
	16kHz	8kHz	16kHz	8kHz	16kHz	8kHz
문 장 (%)	98.80	97.10	98.20	91.30	96.40	88.60
어 절 (%)	99.68	99.24	99.55	97.64	98 65	96 32

표 3-2. 각 모델별 인식률 (beam width 200)

	남 성		여 성		통 합	
	16kHz	8kHz	16kHz	8kHz	16kHz	8kHz
문 장 (%)	98.60	97.10	98.20	84.00	96.35	88.60
어 절 (%)	99.51	99.24	99.55	95.99	98.64	96.32

표 3-3. Beam Width 의 변화에 따른 인식성능 및 소요시간(16kHz sampling, 남성화자의 경우)

Beam Width	60	80	100	120	140	160
문 장 (%)	88.25	94.76	97.06	97.78	98.09	98.30
어 절 (%)	90.88	97.23	98.65	99.06	99.29	99.38
소요 시간	0.337	0.421	0.510	0.600	0.692	0.790
Beam Width	**180**	**200**	**220**	**240**	**260**	**∞**
문 장 (%)	98.5	98.6	98.7	98.7	98.8	98.8
어 절 (%)	99.48	99.51	99.58	99.63	99.68	99.68
소요 시간	0.889	1	1.119	1.246	1.382	33.029

* Beam width 200 인 경우의 소요시간 기준

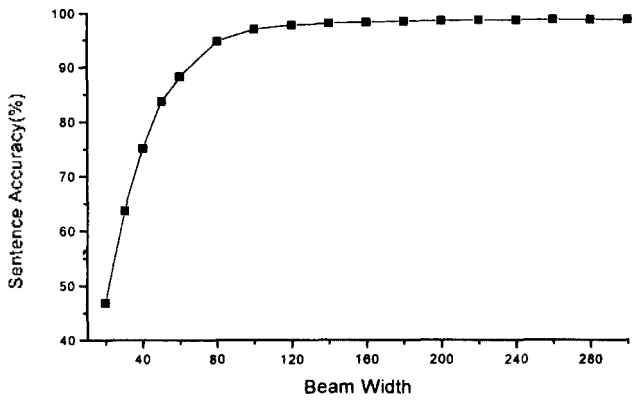

그림 3-2. Beam width 에 따른 인식률(남성, 16kHz)

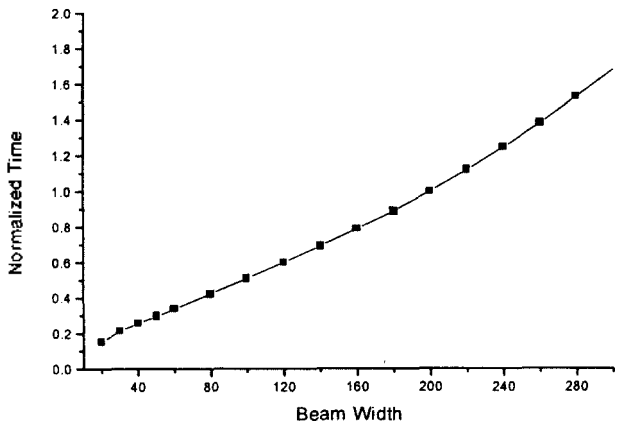

그림 3-3. Beam width 에 따른 인식소요시간(남성, 16kHz)
(Beam width 200 인 경우의 소요시간 기준)

4. 결 론

본 논문에서는 휴대 멀티미디어 단말에 적용할 수 있는 불특정화자 연속음성인식 시스템을 구현하였다. 개발된 시스템은 Handy Combi 응용 도메인을 대상으로 하였으며, 음성특징 파라메터로서 잡음환경등에 강인한 것으로 알려진 Mel-Frequency Cepstral Coefficients(MFCC)를 사용하였다. 또한 섬세한 음소 모델링이 가능한 연속확률분포 HMM을 기반으로 하였으며, 특히 음소모델의 강인한 훈련을 위해 음성학적 지식에 기반을 둔 tree-based clustering 방식을 도입하였다. 화자독립 연속음성인식 실험 결과, 최고 99.7%의 단어인식률과 98.8%의 문장인식률을 얻었으며, 최종적인 문장의 이해도는 99%이상으로서 우수한 결과를 얻었음을 확인하였다.

본 논문은 한국전자통신연구원 인공지능연구실에서 지원한 학·연 공동연구과제의 결과입니다.

참 고 문 헌

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

[2] L. R. Bahl, P. V. de Souza, *et al.*, "Decision trees for phonological rules in continuous speech," in Proc. ICASSP, pp.185-188, 1991.

[3] H. J. Nock, M. J. F. Gales, S. J. Young, "A comparative study of methods for phonetic decision-tree state clustering," Proc. EUROSPEECH, vol.1, pp.111-114, 1997.

[4] 이용주, 음성 데이타베이스 설계 및 제작, 용역결과 보고서, 한국전자통신연구원, 1998 년 5 월.