문장 거부를 위한 음소기반 인식 네트워크에서의 필러 모델 비율과 단어 검출률의 성능비교

*김형태° *이병혁 **하진영
*강원대학교 컴퓨터정보통신공학과
**강원대학교 전기전자정보통신공학부
{ds2swd° , iamdmania , jyha}@kangwon.ac.kr

Performance Comparison of Filler Models and Word Spotting Ratio for Sentence Rejection in Phoneme-based Recognition Networks

*Hyung-Tai Kim° *Byung-Hyuk Lee **Jin-Young Ha
*Dept. of Computer and Information Communication Engineering, Kangwon National Univ.
**Dept. of Electrical and Computer Engineering, Kangwon National Univ.

요 약

음성인식 시스템에서 입력된 음성 데이터에 대해 비인식 대상을 거부하는 기능은 신뢰도 보장 측면에 있어서 상당히 중요하며, 신뢰도를 높이기 위해서는 단순한 인식기능 외에 부적절한 입력 패턴의 거부 기능이 필요하다. 본 논문에서는 이러한 신뢰성 문제를 해결하기 위하여 음소기반 인식 네트워크에서 필러 모델 방법과 단어 검출률 방법을 사용하여 실험하였고, 문장의 단어 수에 따른 두 방법의 문장 거부 성능을 FAR과 FRR의 평균을 최소화 하는 값을 각각 구함으로써 비교 · 분석 하였다. 그 결과 필러모델 방법이 좀 더 나은 거부 성능을 보였고, 단어 검출률을 이용하는 방법이 인식 네트워크를 전부 거치지 않아도 되므로 실행속도와 메모리 절약에서 효과적이었다.

1. 서 론

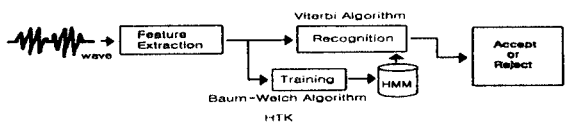
음성 인식 기술이 발전함에 따라 좀 더 자연스럽고, 편리한 인터페이스 방식의 음성인식 시스템이 등장하고 있으나 시스템 제작 시 정해 놓은 음성 문장 데이터외의 다른 데이터들이 입력되었을 때는 이를 처리하기 힘든 단점을 갖는다[1][2]. 따라서 음성인식 시스템의 신뢰도를 높이기 위하여 단순 인식 대상 이외의 다른 단어나 문장을 발성하였을 경우, 이를 무조건 인식하려 하지 않고 거부하여 사용자에게 제대로 된 문장 음성을 재입력하게 함으로써 시스템의 신뢰도를 높일 수 있는 거부기능이 필요하다[3].

본 논문에서는 비인식 대상 문장거부 기능을 구현하기 위한 방법으로 음소기반의 인식 네트워크에서 비인식 대상 단어들을 별도의 필러모델(filler model)로 만들어 이를 인식대상 단어와 병렬연결 처리함으로써 거부기능을 구현할 수 있는 필러 모델(filler model)방법[4]과, 문장 내 인식된 단어 비율과 인식에서 누락된 단어의 비율로 문장 거부를 판단할 수 있는 단어 검출률에 의한 방법을 사용하여 실험하였다. 그리고 단어 수에 따른 문장의 FAR(False Acceptance Rate : 제시된 문장 이외의 다른 문장이 입력되었을 때 거부하지 못하는 오류)과 FRR(False Rejection Rate : 제시된 문장이 입력되었음에도 이를 거부하는 오류)의 평균을 최소화 하는 값을 각각 구함으로써 두 방법의 성능을 비교 · 분석 하였다[5][6].

2. 문장 거부 네트워크의 구성

2.1 시스템의 구성 및 구현

시스템은 음성인식 분야에서 우수한 성능을 보여 많이 사용하고 있는 HMM(Hidden Markov Model)을 채택하였고, 음소 모델의 훈련 및 인식 실험은 HTK(Hidden Markov Model Toolkit)를 사용하여 <그림 1>과 같이 전반적인 비인식 대상 문장거부 기능을 수행하였다.



<그림 1> 비인식 대상 문장 거부 기능 수행 흐름도

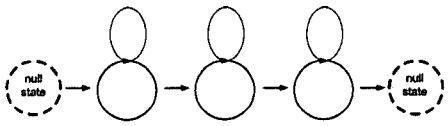
비인식 대상 문장거부 기능을 구현하는 방법으로 거부를 위한 적절한 후처리 과정을 통하여 인식 결과를 확인 하는 방법[2]이 있으나, 본 논문에서는 별도의 후처리 과정 없이 대상 문장의 단어단위 필러 모델을 구축하는 방법과 문장내의 단어를 선택적으로 인식하는 방법만을 사용하여 각각을 구현하였

※ 이 논문은 2005년도 강원대학교 두뇌 한국 21 사업에 의하여 지원되었음.

다.

2.2 음소 단위 인식 네트워크의 구성

IPA(International Phonetic Alphabet: 국제 음성 기호) 발음 표기에 의하면 /d/ /ŏu/ /n/ /t/, /m/ /ɪ/ /s/, /ð/ /ə/, /b/ /ʌ/ /s/로 나타 낼 수 있으며, 이때 각 음소는 <그림 2>와 같은 HMM 모델의 구조를 갖는다.



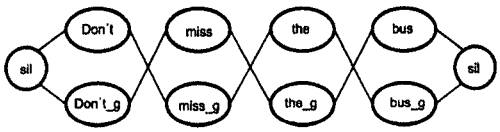
<그림 2> 음소모델의 구조

2.3 단어별 모델 인식 네트워크

본 논문에서는 "Don't miss the bus"라는 문장으로 필러 모델 인식 네트워크와 단어 선택 검출 모델 인식 네트워크를 구성 하였다.

2.3.1 필러 모델

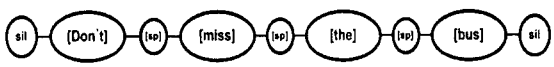
문장의 각 단어별로 사용한 필러 모델은 <그림 3>과 같은 구조이다. Don't_g, miss_g, the_g, bus_g는 각각 Don't, miss, the, bus의 필러 모델이며, 인식 대상단어와 단어별 필러모델을 병렬로 연결하여 전체 인식 결과 중 필러모델이 차지하는 비율을 문턱값(threshold)으로 설정하여 입력 발화 데이터에 대해서 이 문턱값을 넘는 필러모델 비율을 보이면 거부하고, 그렇지 않으면 받아들이는 방법을 사용하였다.



<그림 3> 문장의 단어별 필러 모델 인식 네트워크

2.3.2 단어 선택 검출 모델

문장의 단어 선택 검출률을 이용한 문장 거부모델은 <그림 4>와 같은 구조이며, 각 문장내의 모든 단어에 대하여 표준 발음의 인식 처리를 각각 선택적으로 하게끔 이들 음소를 직렬 연결함으로써 비인식 대상 문장에 대한 거부기능을 수행하는 인식 네트워크를 구성할 수 있다. 따라서 전체 문장을 구성하고 있는 단어를 선택적으로 인식하여 전체 단어 중 인식에서 누락된 단어의 비율에 따라 문턱값이 결정된다.



<그림 4> 문장의 단어 선택 검출 모델

3. 실험 및 결과 분석

3.1 실험환경 및 데이터베이스

비인식 대상 문장 거부 기능을 수행하기 위하여 HTK V.3.2.1을 사용하여 음향 모델 훈련과 인식 실험을 수행하였다. 본 실험에서 사용한 영어 음성 데이터 베이스는 언어교육을 위한 영어 발음 교정용 음향 모델 생성을 목적으로, PC 환경에서 영어를 모국어로 사용하는 성인 400명이 문장을 발음한 영어 음성 DB를 사용하였다. 음성 데이터는 16KHz, 16bit, Mono, linear PCM으로 녹음되었으며, 남자 200명, 여자 200명이 각각 발음한, 총 4120개의 영어 문장을 사용하였다.

실험에 사용된 사전은 4.58MB의 크기를 갖는 표준 발음사전이며 CMU 사전을 근간으로 하여 만들었으며, sp, sil 외 108개의 음소 모델을 사용하였다. 또한 가우시안 믹스쳐(Gaussian Mixture) 7개의 Continuous density HMM을 사용하였다.

3.2 실험 결과

본 논문에서 비인식 대상 문장 거부 기능을 실현하는 기본적인 방법으로 인식네트워크를 거친 입력 문장의 결과에 필러 모델을 사용할 경우 전체 인식된 단어 중 포함 되어있는 필러모델의 비율을 조사하고, 단어 검출률을 이용 할 경우 인식된 단어 모델의 비율과 인식에서 누락된 모델이 어떤 비율로 포함되어 있는가를 조사하는 방법을 택하였다.

입력 음성에 대하여 네트워크에서 주어진 문장의 모든 단어가 모두 올바르게 인식이 되었다면 입력 음성은 주어진 문장을 발화한 것이라고 판단할 수 있고, 반대로 모두 인식 되지 않았다면 주어진 문장대신 다른 문장을 발화 하였거나 소음으로 판단하여 거부해야 할 것이다. 그러나 이것은 이상적인 결과이며 주어진 문장을 입력 하였음에도 이를 거부할 수도 있고, 다른 문장을 입력하였는데 정상 인식된 결과가 나올 수 있다. 따라서 적절한 문턱값을 설정하여 입력 발화 데이터에 대해 필러모델 비율이나, 인식 누락된 비율이 문턱값을 넘으면 거부하고 그렇지 않으면 받아들이는 방법을 사용해야 한다. 또한 문장거부 시스템의 문턱값을 높이면 FAR이 커지고, 문턱값을 낮추면 FRR이 커지기 때문에 FAR과 FRR의 평균치를 최소화 할 수 있는 문턱값을 선택하여, 목표 문장의 문장 내 단어수(2~6단어)에 따라 가변적인 길이 의존 문턱값을 사용하는 방법으로 실험 하였다.

본 논문에서는 과거의 비인식 대상 문장 거부를 위한 필러 모델 기반 인식 네트워크에 대한 연구[6]중 문장 내 인식한 필러 모델의 비율이 sp와 sil을 포함한 비율인 것을 확인하고, 오류를 수정하여 같은 문장 데이터로 재 실험을 하였으며, 이를 단어 검출률을 이용한 문장 거부 모델에도 적용하여 두 방법을 비교하였다.

실험결과를 얻기 위하여 먼저 실험에 사용될 각 문장의 단어수(2~6) 별로 문장을 분류하였고, 분류된 문장에 필러모델을 사용한 방법과 단어 검출률을 이용한 방법을 사용하여 각각의 필러모델의 수와 인식 누락 모델의 수의 백분율을 구하였으며,

이를 다시 <표 1>, <표 2>처럼 백분율 누적 값으로 나타내었다. <표 1>은 필러모델을 사용 하였을 때 문장 내 필러모델 검출수의 백분율 누적값이고, <표 2>는 단어 검출률 방법을 사용하였을 때 단어검출 누락수의 백분율 누적값이다.

<표 1> 단어수별 필러 모델 검출수의 백분율 누적값

필러 모델 비율(%) / 단어수	0	1~10	10~20	20~30	30~40	40~50	50~60	60~70	70~80	80~90	90~100	비고	문장수
2개	100	12.5	12.5	12.5	12.5	12.5	0	0	0	0	0	FRR	160
	20	20	20	20	20	55.62	55.62	55.62	55.62	55.62	100	FAR	
3개	100	31.94	31.94	31.94	31.94	6.24	6.24	6.24	0.7	0.7	0.7	FRR	1280
	2.1	2.1	2.1	2.1	29.06	29.06	29.06	73.82	73.82	73.82	100	FAR	
4개	100	49.92	49.92	49.92	16.75	16.75	4.07	4.07	4.07	0.48	0.48	FRR	1640
	0.36	0.36	0.36	7.31	7.31	37.55	37.55	37.55	82.55	82.55	100	FAR	
5개	100	62.06	62.06	27.06	27.06	7.59	7.59	1.97	1.97	0.62	0.62	FRR	960
	0	0	1.78	1.78	13.03	13.03	43.75	43.75	89.68	89.68	100	FAR	
6개	100	66.25	66.25	30	30	8.75	1.25	1.25	1.25	1.25	0	FRR	80
	0	0	0	0	3.75	17.5	17.5	66.25	66.25	100	100	FAR	

<표 2> 단어수별 단어검출 누락수의 백분율 누적값

단어검출 누락비율(%) / 단어수	0	1~10	10~20	20~30	30~40	40~50	50~60	60~70	70~80	80~90	90~100	비고	문장수
2개	100	21.25	21.25	21.25	21.25	21.25	1.25	1.25	1.25	1.25	1.25	FRR	160
	30.62	30.62	30.62	30.62	30.62	78.12	78.12	78.12	78.12	78.12	100	FAR	
3개	100	37.95	37.95	37.95	37.95	8.11	8.11	8.11	0.46	0.46	0.46	FRR	1280
	4.15	4.15	4.15	4.15	37.19	37.19	37.19	92.19	92.19	92.19	100	FAR	
4개	100	56.69	56.69	56.69	22.61	22.61	7.31	7.31	7.31	0.06	0.06	FRR	1640
	1.46	1.46	1.46	11.21	11.21	47.81	47.81	47.81	92.44	92.44	100	FAR	
5개	100	70.83	70.83	31.66	31.66	9.57	9.57	2.8	2.8	0.1	0.1	FRR	960
	0	0	1.66	1.66	14.88	14.88	48.42	48.42	95.32	95.32	100	FAR	
6개	100	82.5	82.5	43.75	43.75	13.75	3.75	3.75	0	0	0	FRR	80
	0	0	0	0	1.25	8.75	8.75	65	65	100	100	FAR	

실험결과 단어수가 증가하면 두 방법의 문턱값도 같이 증가하였으며, 단어수를 고려하여 최적의 적응 문턱값을 가변적으로 적용 하였을 때 <표 3>과 같이 FRR과 FAR의 평균값은 필러모델 방법의 경우 13.29%이고, 단어 검출률을 이용한 방법의 경우 17.25%로 필러모델을 이용한 방법이 성능 면에서 더 좋은 결과를 보였다. 또한 필러모델을 사용할 경우 입력문장이 인식 네트워크에서 각 단어별 필러모델을 거쳐야 하기 때문에 실행속도와 메모리 절약에 있어서 단어 검출률을 이용한 모델이 더 효과적이다.

<표 3> 단어수별 오류가 최소가 되는 최적의 평균값

단어수	필러 모델 방법		단어 검출률 방법		문장수
	문턱값	FAR과 FRR의 평균(%)	문턱값	FAR과 FRR의 평균(%)	
2	10	16.25	10	25.94	160
3	10	17.02	10	21.05	1280
4	40	12.03	40	16.91	1640
5	50	10.31	50	12.23	960
6	60	9.375	60	6.25	80
평 균		**13.29**		**17.25**	4,120

4. 결론 및 향후 과제

본 논문에서는 음성인식 시스템에서 비인식 대상 문장의 거부를 필러모델을 사용하는 방법과 단어 검출률을 이용한 방법을 통하여 구현하였다. 두 가지 방법 모두 주어진 문장 내 인식 단어들의 필러모델비율이나 인식에서 누락된 단어의 비율에 의존해 문장에 대한 거부를 판단한다. 또한 실험을 통해 이를 최적화 시킬 수 있는 문턱값과 문장을 구성하고 있는 단어수에 따른 문장의 거부 성능을 FRR과 FAR의 평균을 최소화 하는 값을 통하여 확인할 수 있었다. 그 결과 필러모델을 사용하는 방법이 더 나은 거부성능을 보였고, 단어 검출률을 이용하는 방법의 경우 인식 네트워크를 전부 거치지 않아도 되므로 실행속도와 메모리 절약에서 효과적이었다.

향후 연구에서는 이 두가지 방법을 통합한 새로운 방법을 연구가 필요하다.

5. 참고문헌

[1] 김무중, 김효숙, 김선주, 김병기, 하진영, 권철홍, "한국인을 위한 영어 발음 교정 시스템의 개발 및 성능 평가," *말소리*, 제46호, 대한음성학회, pp.87-102, 2003.

[2] 김무중, 김병기, 하진영 "음소기반 인식 네트워크에서의 비인식 대상 단어 거부 기능 성능 분석," 한국음향학회 하계학술발표, 제22권, 1(s) pp.85-88, *한국음향학회*, 2003.

[3] 김우성, 구명완 "반음소 모델링을 이용한 거절 기능에 대한 연구," 한국 음향학회지, 제 18권, 제 3호, pp.3-9, 1999.

[4] 김동화, 김형순, 김영호 "고립단어 인식 시스템에서의 거절기능 구현", *한국 음향 학회지*, 제 16권 제 6호, pp.106-109, 1997.

[5] RC.Rose "Discriminant wordspotting techniques for rejection nonvocabulary utterances in unconstrained speech," Proc. IEEE Conf, Acoustics, Speech, and Signal Processing, Vol.2, pp. 105-108, Mar. 1992.

[6] B.-H. Lee and J.-Y. Ha, "Length Dependent Threshold for Non-Recognition Sentence Rejection," The 4th Asia Pacific International Symposium on Information Technology, pp.462-465 Gold Coast, Austrailia, 2005.