제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

음성인식용 아동음성데이터베이스의 구축 및 음향모델의 검토

김연화*, 이용주*
*원광대학교 전기전자및정보공학부

e-mail:kimyw@mail.wonkwang.ac.kr, yjlee@wonkwang.ac.kr

The Construction of a Children Speech Database for Speech Recognition and The Validation of Acoustic Models

Yoen-Whoa Kim*, Yong-Ju Lee*
*Dept. of Electrical Electronic And Information Engineering, Wonkwang University

요 약

최근 아동음성을 이용한 응용분야가 활기를 띄고 있다. 따라서 아동음성DB의 구축이 시급히 필요하게 되었다. 이러한 요구에 따라 아동음성을 여러 응용분야에 적용하기 위한 한 방법으로, 아동음성DB를 구축하였고, 이를 이용한 음향모델을 작성하였다. 아동음성의 효율적인 인식을 위한 음향모델을 고찰하기 위하여 연령대별로 음향모델을 만들고, 이를 이용하여 훈련 및 평가용 데이터로 인식실험한 결과를 비교검토한다.

1. 서론

음성데이터베이스란 컴퓨터가 읽을 수 있는 형태로, 그 데이터를 재 사용할 수 있도록 annotation과 documentation이 충분하게 갖추어져 있는 음성녹음의 모든 집합으로 정의될 수가 있다.

음성DB는 데이터베이스의 내용정의, 녹음, 후처리등의 세단계로 개발이 이루어지며, 구축하는데에는 많은 시간과 예산, 그리고 전문적인 지식이 필요하며, 다양한 환경에서 다양한 발성화자들을 대상으로 수집된다.

이렇게 수집된 음성DB는 연구목적, 기술적응용등 사람의 말소리를 대상으로 한 연구분야에 쓰이게 되는데, 본 연구에서는 특히 아동음성DB를 대상으로 여러 종류의 음향모델을 만들어서 비교검토하였다.

2. 아동음성DB의 설계 및 구축

본 연구에서는 원광대학교 음성정보기술산업지원센터에서 작성한 KIDS01 DB를 이용하였으며, KIDS01 DB에 대한 대략적인 내용과 구조는 다음과 같다.

2.1 개요

KIDS01 DB는 각 지역별 초등학교 1학년에서 6학년까지의 남녀어린이 500명을 대상으로 수집한 음성인식용 단어에 대한 음성데이터베이스이다.

2.2 발성목록

발성목록의 내용은 다음과 같이 구성되어 있다.

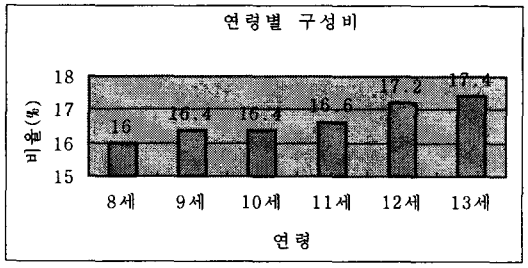
(1) PC용 명령어, 학습관련 지시어등 간단한 명령어 및 지시어 400종
(2) 단독 숫자 및 단위 41종
(3) 앞, 뒤의 다양한 숫자 환경을 포함한 4연 숫자 340종
(4) 다양한 음운환경을 포함한 PBW452어절

총 1233어휘로 구성된 발성목록의 내용을 한 어린이가 모두 발성하기에는 양이 너무 많아서, 1인당 100 내지 101 어휘를 발성할 수 있도록 20개의 세트로 재구성하였으며, 사용빈도가 많은 단독 숫자 및

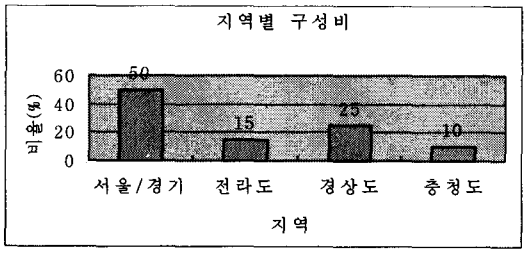
단위 41종은 모든 화자가 발성하도록 하였고, 나머지 목록내용은 세트에 골고루 분포하도록 구성하였다.

2.3 발성화자

발성화자의 선정은 각 지역, 성별 및 연령을 고려하였다. 총 500명(남:250, 여:250)으로 이루어졌으며 연령별, 지역별 현황은 다음과 같다.



(그림 1) 연령별 구성비



(그림 2) 지역별 구성비

2.4 음성데이터의 녹음

음성데이터는 Microphone(Andrea ANC 750 headset)과 Sound Card(Sound Blaster PCI 128 Digital)를 이용하여 심한 소음이 배제된 실내공간에서 수집하였으며, 발성시 마이크의 위치는 발성화자의 입과 2cm의 거리를 유지했고, 모니터 요원이 화자의 발성오류단어를 확인한 후에 재발성 하도록 하였다. 재발성의 경우 무리하게 발성을 교정시키지 않고, 잠깐의 휴식을 주고 재발성하도록 하여 자연스러운 발성이 유지될 수 있도록 하였다. 녹음된 음성 데이터는 16kHz, 16Bit Windows Wave Format으로 저장하였다.

2.5 세그먼테이션

연속적으로 저장된 음성 데이터를 대상으로 자동 끝점 추출 알고리즘을 사용하여, 필요한 음성 데이터 구간을 찾아내고, 음성 데이터의 앞과 뒤에 일정한 길이의 무음구간 300msec을 확보하여 분절하였다.

2.6 전사

음성데이터에 대한 전사결과데이터는 여러 가지 규칙이 포함되어 이루어져 있다.

3. 아동음성인식을 위한 음향모델의 비교

3.1 음성자료의 준비 및 분석

본 연구를 위한 음성자료는 다음과 같이 준비하였다.

(1) 성별/학년별/지역별로 분류
(2) 총 486명의 데이터를 균형을 맞추어서 6개 집단으로 분류
(3) 각 집단의 데이터를 이용하여 각각의 모델을 작성
(4) 각 집단의 데이터를 이용하여 작성된 각각의 모델을 테스트
(5) 각 모델의 해당집단평가데이터는 closed test이고, 나머지집단평가데이터는 open test이다.

<표 1> 남자의 각 집단데이터

데이터 / 모델	M12Data	M34Data	M56Data
M12Mo	**closed**(81명)	open(81명)	open(81명)
M34Mo	open(81명)	**closed**(81명)	open(81명)
M56Mo	open(81명)	open(81명)	**closed**(81명)

<표 2> 여자의 각 집단데이터

데이터 / 모델	F12Data	F34Data	F56Data
F12Mo	**closed**(81명)	open(81명)	open(81명)
F34Mo	open(81명)	**closed**(81명)	open(81명)
F56Mo	open(81명)	open(81명)	**closed**(81명)

본 연구를 위한 음성데이터의 분석 및 추출된 특징 파라미터는 다음과 같다.

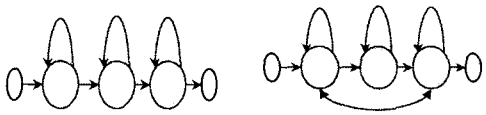
<표 3> 음성데이터의 분석조건 및 특징파라미터

음성데이터	
Sampling frequency	16kHz
Resolution	16bits
Pre-emphasis	$1-0.97z^{-1}$
Window	Hamming Windows(25msec)
Frame rate	10msec
Feature Parameters (39차)	MFCC(12)+E(1)+ΔMFCC(12) ΔE(1)+ΔΔMFCC(12)+ΔΔE(1)

3.2 음향모델의 작성과 인식시스템

HMM(Hidden Markov Model)기반 음성인식은 확률모델을 이용한 통계적 패턴인식방법으로서, 출력확률의 분포에 따라 크게 이산HMM, 반 연속 HMM, 연속HMM으로 분류한다.

본 연구에서는 CHMM(Continuous HMM)을 이용하여 음소모델을 작성하였으며,5state left-to-right 모델의 구조를 사용하였다.

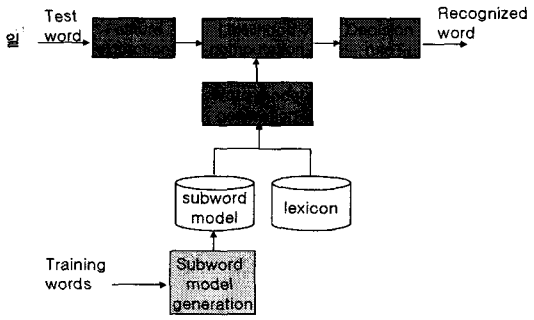


(가) 음소모델 (나) 묵음모델

(그림 3) 모델의 구조

또한 음향신호의 모델링을 위해서 문맥독립음소, 문맥종속음소등을 사용하였으며, 음소의 표기를 위해서는 42개의 PLU(Phoneme Likely Unit)를 사용하였다.

그리고 발음사전은 트랜스크립션을 이용하고 기본인식단위모델들을 연결하여 구성하였고, 워드네트워크는 SLF(Standard Lattice Format)을 이용하였으며, 음성인식시스템의 개략적인 내용은 다음 그림과 같다.



(그림 4) 인식시스템

3.3 인식실험 및 실험결과의 검토

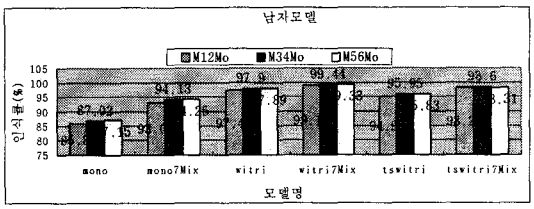
본 연구에서는 HTK3.1.1을 이용하여 인식실험을 하였으며, 여러 가지 모델을 작성하여 인식실험한 결과는 다음과 같다.

(1) Closed Test인 경우

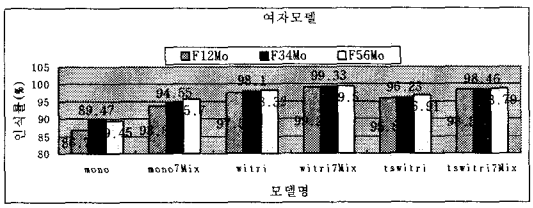
남/여 모두 word internal triphone 7mixture model에서 가장 높은 성능을 보였으며, 결과는 <표 4>와 같다.

<표 4> word internal triphone 7Mixture

각 집단의 모델	인식률
M12Mo	99.15%
M34Mo	99.44%
M56Mo	99.33%
F12Mo	99.24%
F34Mo	99.33%
F56Mo	99.50%



(그림 5) 남자모델의 Closed Test



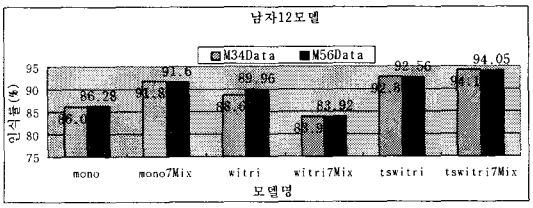
(그림 6) 여자모델의 Closed Test

(2) Open Test인 경우

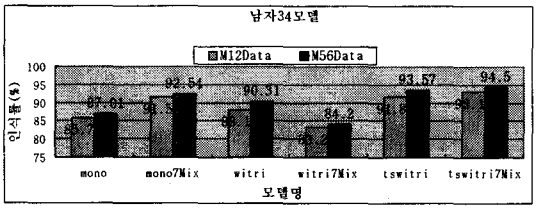
남/여 모두 tied state word internal triphone 7mixture model에서 가장 높은 성능을 보였으며, 남자음성모델의 인식실험결과는 <표 5>와 같고, 여자음성모델의 인식실험결과는 <표 6>과 같다.

<표 5> tied state word internal triphone 7Mixture

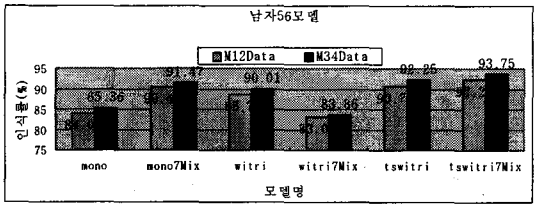
평가데이터 / 모델(tswitri7Mix)	M12Data	M34Data	M56Data
M12Mo	**98.25**	94.18	94.05
M34Mo	93.19	**98.60**	94.50
M56Mo	92.25	93.75	**98.31**



(그림 7) 남자12모델의 Open Test



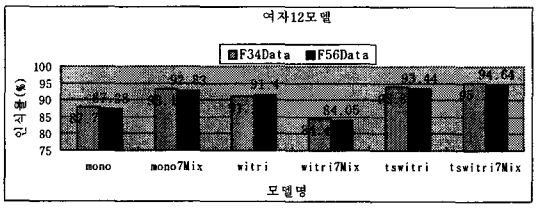
(그림 8) 남자34모델의 Open Test



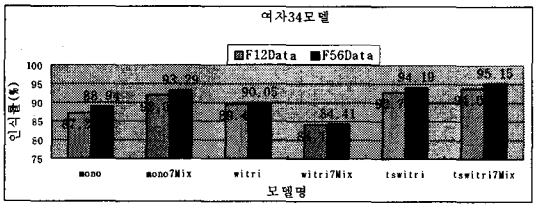
(그림 9) 남자56모델의 Open Test

<표 6> tied state word internal triphone 7Mixture

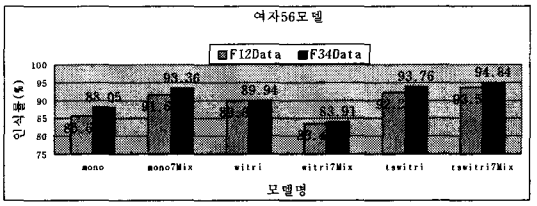
평가데이터 / 모델(tswitri7Mix)	F12Data	F34Data	F56Data
F12Mo	**98.51**	95.10	94.64
F34Mo	94.01	**98.46**	95.15
F56Mo	93.56	94.84	**98.79**



(그림 10) 여자12모델의 Open Test



(그림 11) 여자34모델의 Open Test



(그림 12) 여자56모델의 Open Test

3.3.1 결과의 검토

각 모델에 대해서 각 평가집단데이터로 평가한 화자별, 어휘별 인식률을 살펴보았다. 초등학교 어린이들의 발성습관상, 같은 단어라 할지라도 실제 발음되어 구현된 말소리는 물리적 또는 추상적으로 매우 다르게 나타나기도 하였다.

본 연구에서는 트랜스크립션을 이용하였고, 학습과정에서 나타난 데이터 부족현상으로 인하여 파라미터공유방법을 이용한 모델을 모델링하기도 하였다.

따라서 우리말의 음향음성학적 특징뿐만 아니라 음운학적 특징이, 실제 아동음성인식에서 어떠한 영향을 주고, 어느 정도의 영향을 주는지에 관한 연구가 앞으로 필요하다.

4. 결론

본 연구에서는 음성인식용 아동음성DB의 구축 및 음향모델의 검토를 위하여 아동음성DB에 대하여 살펴보고, 이를 이용한 음향모델을 만들고, 연령대별로 나누어 인식성능을 살펴보았다.

향후, 일반적으로 이용되는 방법들을 적용해 보면서, 특히 아동음성에서 많이 나타날 수 있는 여러 특징들을 고찰해 보고, 아동음성의 인식성능과의 관계를 살펴본다. 또한, 음향모델의 특징파라미터의 변경, 음향모델의 구조의 변경, 탐색 알고리즘의 변경 등에 따른 아동음성의 인식효율을 검토해 볼 필요가 있다.

참고문헌

[1] S. J. Young, The HTK Book, Cambridge, 1997
[2] 음성정보기술산업지원센터,KIDS01 DB,CD-ROM
[3] 김동화, 연속음성인식을 위한 향상된 결정트리 기반 상태공유 기법 연구, 박사학위 논문, 부산대학교, 1999