DOI : 10.5392/JKCA.2011.11.5.138

한국어에서의 성인과 유아의 음성 인식 비교

Comparison of Adult and Child's Speech Recognition of Korean

유재권*, 이경미**
덕성여자대학교 전산정보통신학과 지능형 멀티미디어 연구실*, 덕성여자대학교 컴퓨터학과**

Jae-Kwon Yoo(chhjk2010@duksung.ac.kr)*, Kyoung-Mi Lee(kmlee@duksung.ac.kr)**

요약

현재 한국의 음성 데이터베이스 구축 현황을 살펴보면 유아에 맞춰진 음성 데이터베이스는 구축이 되지 않은 실정이다. 국외 연구를 분석한 결과, 다양한 언어를 기반으로 유아 대상의 음성 데이터베이스가 구축되어 있다. 이는 성인의 음성과 유아의 음성은 언어학적으로 차이가 있기 때문에 유아는 유아에 맞는 음성 데이터베이스가 필요하다. 한국어에서 성인과 유아의 음성 차이를 알아보기 위해, HMM을 이용하여 유아와 성인의 음성인식을 비교하였다. 유아와 성인의 음성인식 비교는 성별, 나이별, 성도 길이 정규화의 적용 유무에 따라 실험한다. 본 논문에서는 한국어에서 유아의 음성을 유아의 음성인식기로 인식했을 때가 성인의 음성인식기로 인식했을 때 보다 월등히 인식률이 높았으며, 성도 길이 정규화의 적용이 인식률 향상에 도움이 되고 있음을 보여준다.

■ 중심어 : | 음성인식 | 성인과 유아의 음성비교 | 히든마코프 | 성도 길이 정규화 |

Abstract

While most Korean speech databases are developed for adults' speech, not for children's speech, there are various children's speech databases based on other languages. Because there are wide differences between children's and adults' speech in acoustic and linguistic characteristics, the children's speech database needs to be developed. In this paper, to find the differences between them in Korean, we built speech recognizers using HMM and tested them according to gender, age, and the presence of VTLN(Vocal Tract Length Normalization). This paper shows the speech recognizer made by children's speech has a much higher recognition rate than that made by adults' speech and using VTLN helps to improve the recognition rate in Korean.

■ keyword : | Speech Recognition | Comparison of Adult and Children's Speech | HMM | VTLN |

I. 서 론

현재 한국의 음성 데이터베이스 구축 현황을 살펴보면 유아에 맞춰진 음성 데이터베이스는 구축이 되지 않은 실정이다. 한국의 공동 이용을 위한 음성 언어 자원의 구축 현황을 보면 초등학교에 재학 중인 학생 500명을 대상으로 숫자, 명령어와 제어어, 단위로 구성된 단어 중심의 데이터베이스가 구축되어 있다[1]. 하지만 언어 발달 상 만 3~5세는 유아에게 언어 발달의 결정적 시기라고 할 수 있다[16]. 이 무렵 형성된 언어 발달은

접수번호 : #110114-006
접수일자 : 2011년 01월 14일
심사완료일 : 2011년 03월 31일
교신저자 : 이경미, e-mail : kmlee@duksung.ac.kr

이후의 같은 노력으로는 보상되거나 대체될 수 없다. 언어 발달 중 말하기 능력은 언어의 습득과 사회생활에 밀접한 연관이 있다. 하지만 현재 교육 과정에서는 말하기 보다는 읽고 듣는 독해 학습 중심의 교육 위주이기 때문에 유아가 자신의 의견을 조리있게 전달하는 능력은 점점 약화되고 있다. 따라서 놀이를 통한 이런 교육 과정의 보완을 위해 유아에게 맞춰진 음성 인터페이스 개발이 필요하지만 현재 한국의 음성 DB에 만 3~5세의 음성 DB에 관한 자료는 거의 없는 실정이다.

유아 음성의 음향적 특징은 성인과는 다르다[2]. 유아 음성은 음의 높낮이, 성도, 포먼트와 같은 음성의 음향학적 상관성에 대해 나이에 의존하는 체계적인 구조를 따른다[3]. 유아의 음성 특징은 아이가 성장함에 따라 해부학적, 생리학적 변화로 인해 급속도로 발달하게 된다. 따라서 기존에 성인 대상으로 구축되어진 음성 DB를 이용하여 유아의 음성을 인식할 경우 인식률 저하가 일어날 수 있다. 물론 국내 연구에는 성인과 유아의 음성을 비교한 연구 자료는 없다. 하지만 여러 국외 논문에서 성인과 유아의 음성 비교가 연구되고 있기 때문에 그 차이를 알 수 있고, 한국어에서도 비슷한 결과가 나올 수 있다는 것이 예측가능하다. 하지만 실제 한국어에서는 유아와 성인의 음성이 어느 정도 차이가 있는지 실험을 통해 증명하고 이 실험 결과를 통해 만 3~5세 유아에 맞는 음성 DB 구축의 필요성을 강조하고자 한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 성인과 유아 음성 비교를 다양한 언어에서 연구한 결과를 기술한다. 한국어가 아닌 다른 언어에서 언어학적 비교와 인식률 측면의 비교를 통하여 성인과 유아 음성의 차이점을 분석한다. Ⅲ장은 한국어에서 성인과 유아 음성 비교를 음성인식기를 이용한 비교를 한다. 인식기를 이용한 비교는 전체, 성별, 나이에 관해 분석하며 각각의 단어에 대한 비교를 위해 혼돈표를 사용하여 실험 결과를 분석한다. 그 다음 VTLN을 적용한 음성인식기의 인식률을 분석하여 인식률 향상이 있는지 관찰한다. 또한 성인과 유아의 음성을 통합한 음성인식기에서 성인과 유아 각각의 음성 인식률을 측정하여 결과를 비교한다. Ⅳ장에서는 연구의 전반적인 내용을 정리하며, 향후 연구 방향에 대해 기술하고 결론을 맺는다.

Ⅱ. 배경 연구

지금까지 연구된 성인과 유아의 음성 비교를 분석하면 언어학적인 비교와 인식기를 이용한 비교로 나눌 수 있다. 언어학적인 비교란 유아가 성장함에 따라 해부학적, 생리학적 변화로 인해 성인의 음성과 언어학적 차이가 있다는 것을 의미한다. 인식기를 이용한 비교란 성인의 음성인식기에서 성인의 음성과 유아의 음성 인식률을 비교하고, 유아의 음성인식기에서 성인의 음성과 유아의 음성 인식률을 비교한 결과를 분석하는 것이다.

2.1 언어학적 비교

성인과 유아의 음성 비교를 위해 네 가지 측면의 언어학적 비교를 할 수 있다. 성인과 유아 음성의 언어학적 비교는 M. Gerosa 등[4]에 의해 소개된 내용을 다룬다. 네 가지 요소는 단음 지속구간(Phone duration), 화자 내 변이(Intra-speaker variability), 음향 공간의 특징(Characterization of the acoustic space), 성도길이 변화의 영향(Effects of vocal tract length variations)이 있다.

단음 지속구간은 주로 미국, 이탈리아에서 이루어진 연구로 해당 언어의 음성 코퍼스를 사용하여 언어상의 단음 지속구간을 분석했다. 두 개의 코퍼스는 다른 목적으로 디자인되었으나, 유아가 성장함에 따라 단음 지속구간이 더 짧아진다는 동일한 결과가 나왔다. 단음 지속 구간이 나이에 따라 변하는 사실은 유아에게 성인의 음성으로 훈련된 음성인식기로 음성 인식률을 측정하게 되었을 때 저조한 인식률을 보이는 이유 중 하나라고 할 수 있다.

화자내 변이는 음성을 제어할 수 있는 측정치이다. 이는 화자가 발언한 모든 단음을 통해 계산된 평균적인 시간과 스펙트럼을 비교하여 계산된다. 이전 연구에서는 나이가 증가함에 따라 스펙트럼과 시간이 줄어드는 것을 연구 분석 결과를 통해 알 수 있었다[9].

음향 공간에서의 특징은 음소 모델들의 관찰 밀도의 분산 정도를 측정하여 관찰한다. 이를 위해 가우시안 밀도(Gaussian density)를 이용하게 되는데, 각 음소에

대해 모델링을 한 후, 음향 특징 공간에서 가우시안 밀도의 분산 정도에 따라 판단하게 된다. [그림 1]에서 Bhattacharyya 거리는 이탈리아 언어를 사용한 음성 분석 결과를 나이그룹과 성별에 대한 수치를 나타낸 것이다[4]. Bhattacharyya 거리는 훈련 데이터들로부터 음소들 간의 통합에 대한 정보를 얻는 방법이다. [그림 1]에서 알 수 있듯이 음소 분포 사이에 평균 Bhattacharyya 거리는 성별과 상관없이 나이에 따라 증가하며 이것은 음소의 분포가 나이가 더 많은 그룹의 음향 공간에서 덜 겹쳐진다는 것을 의미한다. 이는 나이가 증가할수록 화자 내 변이의 감소가 영향을 미친다는 것을 알 수 있다. 이 분석은 나이와 관련이 있으며 성별로 나누지 않는 것은 중요한 요소가 아니기 때문이다.

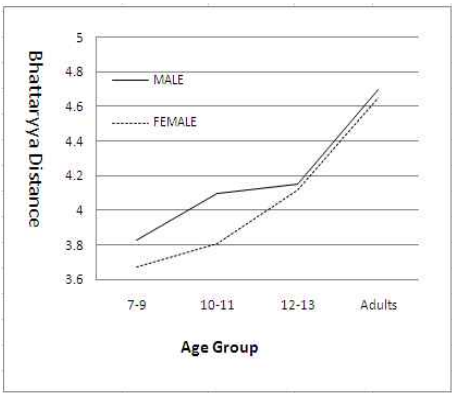

그림 1. 이탈리안 음성 코퍼스를 사용한 나이그룹과 성별로 나타낸 Bhattacharyya 거리[4]

성도길이 변화에 대한 영향은 음성의 음향학적 이론 예측에 의해 성도 모양과 형태에 의한 것임을 알 수 있다[6]. 이런 관련성을 이유로 음성의 포먼트 패턴과 성도길이에 대한 연구가 진행 되었다. 유아가 성장하면서 포먼트 주파수가 감소하는 반면에 성도 길이는 꾸준히 길어지고 있다는 사실을 알 수 있다[8]. 화자 사이에 변이의 원인은 억양과 감정과 같은 외부적인 요인과 성도 길이와 모양과 같은 화자의 차이에 의해 발생하는 내재적인 요인이 있으며, 이 중 내재적인 요인을 감소시킬 수 있는 화자의 성도 길이에 따른 특징벡터의 변이를 최소화하는 VTLN이 널리 사용되고 있다. 화자의 성도 길이에 따라 음성의 스펙트럼은 주파수 축으로 확대되거나 축소되어 나타난다. 유아는 성인보다 성도의 길이가 더 짧기 때문에 성인음성인식기로 음성인식을 할 경우 인식률 저하가 나타나는 원인이 된다. 따라서 본 논문에는 음성 인식률 상승을 위해 VTLN을 적용하여 실험한다.

2.2 음성 인식기를 이용한 비교

성인과 유아의 음성은 인식기를 이용한 비교를 할 수 있다. 성인 음성인식기와 유아 음성인식기를 구축하여 각각의 인식기에 대해 성인과 유아의 음성으로 실험한다.

2.2.1 이탈리아 언어

이탈리아의 경우 성인 144명, 7-13세 대상 87명의 화자를 이용하여 훈련과 실험을 했다[2]. 이 실험은 HMM을 사용하였고, VTLN을 적용시켜 실험을 진행했다. 실험에 대한 결과는 [표 1]과 같다[2]. 이 실험에서 훈련은 성인과 유아 음성을 각각 사용하며, 실험은 유아의 음성만을 이용하였고, VTLN의 적용 유무에 따라 인식률을 측정했다. 성인의 음성을 사용하여 훈련시킨 후 유아의 음성으로 실험할 경우의 인식률은 60.32%이며 VTLN을 사용하면 68.33%로 가장 좋은 인식률이 나왔다. VTLN 사용 시 인식률이 향상하는 것을 볼 수 있지만(68.33%), 유아의 음성으로 VTLN을 적용하지 않고 실험했을 때(77.3%)보다 인식률이 8.97% 낮은 것을 볼 수 있다.

표 1. 이탈리아의 음성인식기의 비교 결과[2]

Training	VTLN		UA(%)
	Training	Test	
Adults	no	no	60.32
	no	yes	67.65
	yes	yes	68.33
Children	no	no	77.3
	no	yes	77.24
	yes	yes	78.51

2.2.2 스웨덴 언어

스웨덴의 경우 성인 200명, 4-8세 대상 200명의 음성

을 수집하여 훈련과 실험을 했다[12]. 실제 실험에서 훈련에 참여한 유아의 음성은 140명이다. 이 실험은 HTK를 사용하고, 특징추출의 방법으로 MFCC를 이용했다. 실험에 대한 결과는 [그림 2]와 같다[12].

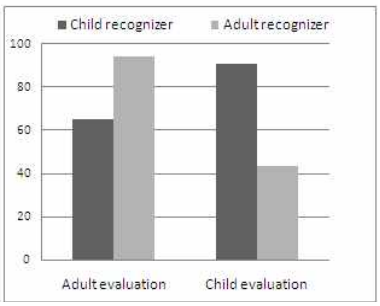

그림 2. 스웨덴의 음성인식기의 비교 결과[12]

이 실험은 유아의 음성인식기와 성인의 음성인식기를 구축하여 각각의 음성인식기에 성인과 유아의 음성을 인식시켜본 결과를 나타낸 것이다. [그림 2]와 같이 성인 음성인식기에 유아의 음성을 실험한 결과는 유아의 음성인식기에 유아의 음성을 실험했을 때(90.56%)보다 낮은 45.31%임을 알 수 있다.

2.2.3 미국 영어

미국 영어의 경우에 음성 인식률 비교를 위해 숫자열을 이용하여 훈련 데이터로 각각 성인 3026명, 10~17세 화자 대상 1234명이 참여했다[11]. 실험 데이터는 6~17세의 화자 501명이 참여하였고, 훈련된 음성인식기는 각각 성인과 10~17세의 화자의 음성을 사용했다. [그림 3]은 성인의 음성으로, [그림 4]는 10~17세 화자의 음성으로 훈련시킨 음성인식기에서 실험한 인식률을 보여주고 있다[11].

[그림 3]에서 알 수 있듯이 성인의 음성으로 훈련한 인식기는 10세 이하에서 저조한 인식률을 보인다. 하지만 [그림 4]에서는 [그림 3]보다 높은 인식률을 보인다. 이는 10~17세의 화자의 음성으로 훈련한 것으로 실험 데이터의 화자 나이와 비슷하기 때문이다. 따라서 나이에 따른 음성 인식은 나이에 맞는 음성 데이터를 사용하는 것이 효과적이라는 것을 알 수 있다.

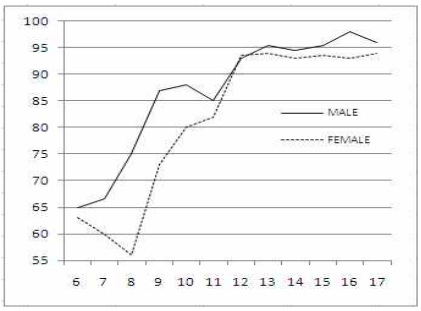

그림 3. 훈련 데이터 : 성인
실험 데이터 : 6~17세 [11]
(가로 : 나이 , 세로 : 인식률)

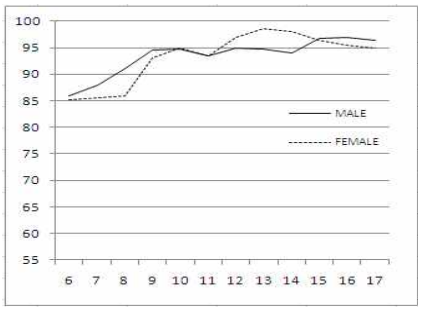

그림 4. 훈련 데이터 : 10~17세
실험 데이터 : 6~17세 [11]
(가로 : 나이 , 세로 : 인식률)

2.2.4 소론

이전의 연구들에서 인식기 측면에서 성인과 유아의 음성 비교 시 성인의 음성인식기에 VTLN, 주파수 변형, 화자 적응 등 여러 방법을 사용하여 유아의 음성 인식률 향상을 위한 방법을 연구하고 있다. 이러한 연구들을 통하여 인식률을 높이고 있지만, 가장 최적의 방법은 유아의 음성 데이터를 이용하여 유아에 맞는 음성인식기를 구축하는 것이다.

현재 한국의 음성 데이터베이스 구축 현황을 살펴보면 만3~5세 대상의 음성 데이터베이스는 구축이 되지 않은 실정이다. 이에 따라 한국어를 대상으로 성인과 유아의 음성 비교 연구 결과를 찾을 수 없다. 본 연구는 유아에 맞는 음성 데이터베이스의 필요성을 성인과 유아의 음성 비교를 통해 언급한다.

Ⅲ. 실험 및 결과

본 논문에서는 한국어에서 유아와 성인의 음성 인식을 비교하기 위하여 Cambridge대학이 개발한 HTK를 사용한다. HTK는 HMM 기반의 음성 인식기를 구현하는 사실상의 표준 도구로써, 세계 대부분의 연구기관과 학교에서 사용되고 있다. 연구에 사용되는 두 개의 데이터베이스는 유아, 성인으로 구성되어 있다. 성인의 음성데이터베이스는 ETRI에서 제공하는 숫자열 단어를 사용하며 인원은 남자 20명, 여자 20명으로 총 40명의 화자로 이루어져 있다. 성인의 음성 데이터베이스는 숫자, 사칙연산, 명령어로 이루어져 있다. 그 중 성인과 유아의 음성 비교를 위해 숫자열을 사용하며 숫자열은 고유어와 한자어 0~9까지의 총 20단어를 사용한다. 유아 음성데이터베이스는 앞서 언급한 것과 같이 만 3~5세 대상이 없기 때문에 본 연구를 위해 직접 데이터를 수집하였다.

3.1 유아 음성 데이터 수집

유아 음성데이터베이스는 만 3~5세 대상이며 수집 단어는 성인의 음성데이터베이스 단어와 동일한 숫자열이고, 인원은 남자 28명, 여자 32명으로 구성하여 총 60명이 참여했다. 유아의 참여 분포도는 [표 2]와 같다.

표 2. 참여 유아의 분포

나이(만)	3	4	5	
남	13	10	5	28
여	16	6	10	32
합계	29	16	15	60

유아 음성 데이터 수집을 위한 녹음은 유치원의 조용한 빈 공간에서 주변 잡음이 정제되지 않은 일반 잡음 환경에서 진행한다. 비교 실험으로 사용된 ETRI의 성인 음성 DB는 잡음이 정제된 곳에서 녹음이 진행되었지만, 유아의 음성을 녹음 할 수 있는 유치원과 어린이집 환경의 특성 상 잡음이 완전히 배제될 수는 없다. 녹음 진행 시간이 오래 지속되는 경우 입술을 만지거나 손으로 마이크를 만지는 등 산만해지는 경향 때문에 10분 이내에 이루어졌다. 각각의 단어들은 설치된 노트북의 스크린에 보여 지고 유아는 선생님들의 지도를 받아 녹음을 하기 시작한다. 같은 단어를 적게는 1번 많게는 5번씩 평균 3번의 발언을 한다. 녹음을 시행 시 아이들의 성격을 파악하여 지루함을 많이 느끼는 아이는 한 단어 당 세 번미만으로 녹음하며, 지도를 잘 따르는 아이들은 한 단어 당 세 번 이상 녹음 할 수 있도록 지도한다. 또한 녹음된 데이터를 정리하는 과정에서 유아의 소리가 너무 작거나 돌발 행동으로 인해 심한 잡음이 섞여 있는 경우 데이터를 삭제하였다.

음성은 Cool Edit Pro 2.1을 사용하여 16KHz 샘플링 비율, Mono 채널, 16bit로 설정한다. 이는 추후 HTK 를 사용하여 시험 평가를 할 때 기본 설정이기 때문이다. 마이크는 코리아 디지탈의 MBL센서 KDS-1012를 사용하여 녹음을 시행했다.

3.2 시험 음성인식 과정

수집된 유아 음성 DB를 활용한 음성인식 시험평가는 HTK를 사용한다. HTK는 FFT 와 LPC 를 모두 지원한다. 이 실험에서는 FFT 기반 log spectra로 유도된 MFCCs 를 사용한다. 20개의 채널을 사용하여 추출한 12개의 켑스트럼 계수를 추출한다. MFCC는 39차 MFCC를 이용하여 특징추출을 한다. 음성 모델을 훈련하는 과정에서 HMM을 사용하며 본 논문에서는 20개의 숫자열을 이용하여 시험 음성 평가가 이루어지므로 20개의 훈련된 HMM이 생성된다.

훈련은 만 3~5세 60명 중 63%인 38명의 음성으로 구성하며 시험은 37%인 22명의 음성으로 구성한다. 훈련과 시험에 사용하는 음성에 대한 유아의 분포는 [표 3]과 [표 4]를 참고한다.

표 3. 훈련에 참여한 유아 분포

나이(만)	3	4	5	
남	8	6	3	17
여	11	3	7	21
합계	19	9	10	38

표 4. 실험에 참여한 유아 분포

나이(만)	3	4	5	
남	5	4	2	11
여	5	3	3	11
합계	10	7	5	22

3.3 시험 음성인식 결과 및 분석

HTK를 이용한 시험 음성 평가 결과는 [그림 5]와 같다. 시험 음성 평가는 성인 음성인식기에서 성인의 음성을 평가할 경우 96.5%, 유아의 음성을 평가할 경우 47.18%의 인식률이 나왔다. 유아 음성 음식기에서 유아의 음성을 평가할 경우 88.02%가 나오며 성인의 음성을 평가할 경우 45.31%의 결과가 나왔다.

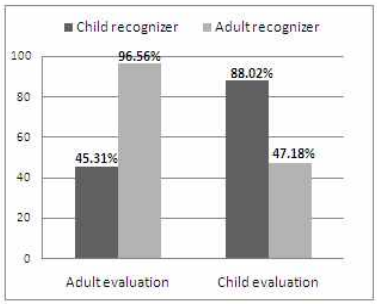

그림 5. 시험 음성 평가 결과

시험 음성 평가에서 알 수 있듯이 성인은 성인의 음성인식기를, 유아는 유아의 음성인식기를 사용할 때 인식률이 높음을 알 수 있다. 반면에 성인의 음성인식기에 유아의 음성을 시험하거나 유아의 음성인식기에 성인의 음성을 시험할 경우 인식률이 50%가 되지 않음을 확인한다.

3.3.1 나이에 따른 인식률 분석

[그림 6]는 나이에 따라 유아 음성인식기와 성인 음성인식기를 비교한 결과이다. 이 그래프에서 보면 알 수 있듯이 만 3~5세는 유아의 음성인식기에서 큰 차이 없이 전체 인식률과 비슷한 인식률을 보이며 고루 분포되어 있는 것을 볼 수 있다. 반면에 성인 음성인식기 측면에서 비교하자면, 만 3세의 인식률이 가장 낮으며, 만 4, 5세의 인식률은 그에 비해 높은 것을 알 수 있다. 이는 나이가 높아질수록 유아의 음성이 성인의 음성과 같이 변한다는 이전 연구결과를 통해 알 수 있다[5].

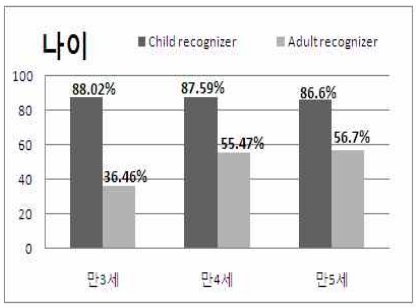

그림 6. 나이에 따른 인식률

3.3.2 성별에 따른 인식률 분석

[그림 7]은 성별에 따라 유아 음성인식기와 성인 음성인식기를 비교한 결과이다. 유아 음성인식기에서 또래의 여자가 남자보다 인식률이 높은 것을 볼 수 있다. 반면 성인의 음성인식기에서 실험 시 남아가 여아의 인식률이 더 높았지만 중요한 결과는 아니다. 왜냐하면 유아의 언어 발달은 연령의 증가에 따라 발달하는 것이기 때문에 성별에 따른 차이는 개인차로 해석할 수 있다[10]. 결론적으로 남, 여 성별에 상관없이 유아 음성인식기에서는 유아의 음성으로 훈련하여 실험한 것이 인식률이 높다는 것이 중요하다.

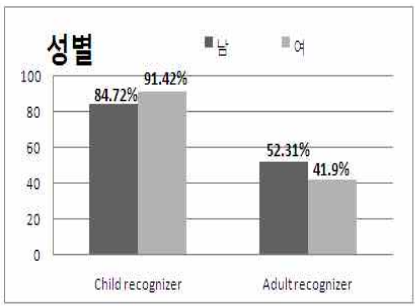

그림 7. 성별에 따른 인식률

3.3.3 유아의 음성 인식률 분석

인식기 측면에서 유아의 음성은 유아의 음성데이터베이스로 구축된 유아의 음성인식기를 사용할 때가 인식률이 높음을 알 수 있다. 자세한 분석을 위해 비교를 위해 선정된 숫자열 단어에 대해 혼동 행렬(confusion matrix)을 [표 5]를 통해 확인할 수 있다. [표 5]에서 유

아 음성인식기에 유아의 음성을 시험평가 할 경우 대부분의 단어가 인식이 잘 되고 있음을 보여준다. 하지만 특정 숫자열 '영'의 인식이 좋지 않음을 알 수 있다. 숫자열 '영'과 혼동되는 단어는 같은 음절의 단어인 '육'이다. '영'과 '육'의 'ㅕ'와 'ㅠ'의 혼동이다. 이는 반모음 /y/를 포함한 이중 모음이기 때문이다. 만 3~5세의 유아는 지속적인 언어발달 과정이기 때문에 분명하고 또렷한 발음을 하는 유아가 있는 반면 불분명한 발음을 하는 유아도 있다. '육' 이외에 '여섯'과 '여덟'은 다른 음절이지만 '영'과 혼돈되는 것을 관찰할 수 있다. 이는 유아와 어른의 음성 분할 지속구간의 차이에서 발생한 혼돈이다[13][14]. 평균적으로 아이의 말하는 지속구간의 비율은 어른보다 느리다. 또한 유아의 말하는 속도와 음성 영향은 높은 가변성으로 인해 음절의 수와 상관없이 혼동하는 모습을 볼 수 있는 것이다.

3.3.4 성도 길이 정규화

화자의 성도 길이에 따라서 음성의 스펙트럼은 주파수 축으로 확대되거나 축소되어 나타난다. 성도의 길이가 성인에 비해 유아의 경우 더 짧기 때문에 성인음성 인식기로 음성인식을 할 경우 인식률 저하가 나타난다. 앞서 실험 결과를 보면 유아는 유아의 음성인식기로 음성인식을 했을 때 인식률이 높은 것을 알 수 있다. 성도 길이의 차이를 줄이기 위해 VTLN을 사용한다[15]. VTLN 알고리즘은 주파수 축을 확대나 축소하여 정규화하는 방법으로 주파수 축의 척도를 변환하는 워핑 함수(warping function)를 사용한다.

HTK는 간단한 조각적 선형 주파수 워핑을 지원한다. 워핑 요소를 α라 칭한다. 선형으로 주파수축이 변하게 되면, 변환된 신호는 원래의 입력 신호와 다른 대역폭을 갖게 된다. 그러므로 주파수 범위의 경계를 주의하여야 한다. 만일, $\alpha < 1.0$ 이면 변환된 주파수 축을 축소하고 $\alpha > 1.0$ 이면 확대한다.

조각적 선형 주파수의 경우 형태를 결정하는 요소는 α뿐이다. 최적 α를 추정하기 위해서 $0.6 \leq \alpha \leq 1.4$의 범위 내에서 값을 조절하면서 추정한다. 이전 연구에서 유아의 음성에 적합한 α값을 추정하기 위한 연구가 진행되었으며, α값을 1.22로 워핑하는 것이 인식률 향상에 기여한다는 것은 이전 연구 결과를 통해 알 수 있었다[7]. 따라서 성인과 유아의 음성 비교를 위해 음성인

표 5. 유아 음성인식기에 유아의 음성을 인식할 경우의 혼동행렬(행 : 입력값 열 : 출력값)

	영	일	이	삼	사	오	육	칠	팔	구	공	하나	둘	셋	넷	다섯	여섯	일곱	여덟	아홉	인식률(%)
영	5						4										7		5		23.8
일		19						1						1							86.36
이			21																	1	95.65
삼	1			19																	95.24
사				1	17											1				1	80.95
오						20				1											90.9
육							20			2											86.95
칠								21			1										91.3
팔									20										1		95.45
구						1	1			19											82.6
공	1					1		1			18										81.82
하나									1			21									91.3
둘								1		1			19								86.36
셋								3						19							86.96
넷		1												1	20						95.45
다섯				1												19	1				86.36
여섯																1	17		4		78.26
일곱																		22			95.65
여덟																	2		19		86.36
아홉											1									20	95.45

식기의 인식률 향상을 위한 VTLN 적용 시 α값을 1.22로 실험하였다. 실험 결과는 [표 7]과 같다.

[표 6]은 음성인식기를 이용해 성인과 유아의 음성인식률을 비교한 것이다. 훈련은 성인과 유아가 각각 포함되어 있고, 시험 평가는 유아의 음성을 사용했다. 이번 실험은 훈련과 시험 평가 둘 다 VTLN을 적용하여 인식률을 비교한 것이다. 앞서 연구 결과에서 성인 음성인식기로 유아의 음성을 실험할 경우 47.18%의 인식률이 나왔으며, 훈련과 실험에 VTLN을 적용 시 0.94% 인식률이 향상 되었지만 여전히 낮은 음성 인식률이 나온다. 유아 음성인식기로 유아의 음성을 인식 할 경우 88.02%의 인식률을 보이며 훈련과 시험 평가에 VTLN을 적용 시 2.35%의 인식률 향상으로 전체 인식률 90.37%의 음성 인식률이 나온다.

표 6. 한국어에서 VTLN을 적용한 성인과 유아의 음성인식기 비교 결과

학습	VTLN		인식률(%)
	학습	테스트	
성인	no	no	47.18
	yes	yes	48.12
유아	no	no	88.02
	yes	yes	90.37

성도의 길이 차이가 큰 경우 주파수 대역에서 주파수 축의 확대 비율이 비선형적이므로 이를 정규화하기 위해 HTK에서 지원되는 조각적 선형 워핑 함수를 사용하여 VTLN을 적용하였다. VTLN의 효용성을 검증하기 위해 실험을 진행하였고, 그 결과 음성 인식률이 향상됨을 알 수 있다.

3.3.5 성인의 음성과 유아의 음성을 포함한 음성 인식 실험

지금까지의 실험은 성인과 유아의 음성 데이터를 이용하여 각각의 음성인식기를 구현하여 실험했다. 이전 논문에서는 각각의 음성으로 성인과 유아의 음성의 차이를 비교했지만, 본 논문에서는 성인과 유아의 음성 데이터 모두를 활용한 음성인식기로 성인과 유아의 음성인식의 차이를 비교한다. 실험 과정은 3.2절에서 기술한 내용과 같으며 앞서 실험한 결과와 비교하기 위해 훈련 데이터와 실험 데이터는 위와 같은 데이터로 진행한다. 훈련과 실험에 참여한 유아와 성인의 분포는 [표 7][표 8]과 같다. 이 실험에서 VTLN의 워핑 요소는 유아에 맞게 α값을 1.22로 하여 실험한다. 이는 추후에 유아와 성인이 공동으로 사용할 수 있는 음성 인터페이스의 개발을 위함이다.

표 7. 훈련에 참여한 성인과 유아의 분포

나이(만)	3	4	5	성인	
남	8	6	3	12	29
여	11	3	7	12	33
합계	19	9	10	24	62

표 8. 실험에 참여한 성인과 유아의 분포

나이(만)	3	4	5	성인	
남	5	4	2	8	19
여	5	3	3	8	19
합계	10	7	5	16	38

HTK를 이용한 유아와 성인 각각의 데이터를 포함한 시험 음성 평가 결과는 [그림 8]과 같다. 시험 음성 평가 결과는 성인의 음성으로 평가할 경우 61.25%, 유아의 음성으로 평가할 경우 77.1%이다. 이 실험에서는 VTLN의 워핑 요소를 유아에 맞는 1.22를 사용했기 때문에 유아의 음성 인식률이 더 높았다. 이는 다시 말해 성인에 맞는 워핑 요소 사용 시 성인의 음성 실험 결과의 인식률이 유아보다 더 높은 인식률이 나오리라는 것을 기대할 수 있다.

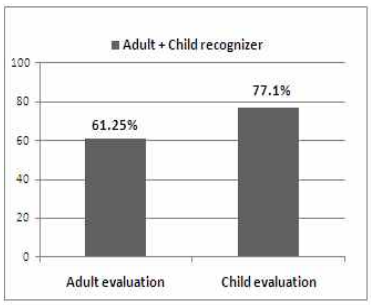

그림 8. 성인+유아 음성 인식기의 시험 음성 평가 결과

[그림 5]와 비교하여 성인의 경우 성인의 음성인식기를 사용하여 성인의 음성으로 실험했을 때(96.56%)보다 낮은 61.25%의 인식률이 나왔고, 유아의 경우 유아의 음성인식기를 사용하여 유아의 음성으로 실험했을 때(88.02%)보다 낮은 77.1%의 인식률이 나왔다. 따라서 이 실험은 성인, 유아 둘 다에게 좋은 영향을 주진 못했다. 이런 실험 결과를 통해 각각의 데이터를 통합하여 음성을 훈련 시켜도 성인과 유아의 음성 차이 때문에 높은 인식률을 기대하긴 힘들며 각각의 음성의 특징에 맞는 음성 DB가 필요하다는 것을 알 수 있다.

Ⅳ. 결론

본 논문에서는 한국어에서 성인과 유아의 음성을 비교하기를 위해 인식기 측면의 연구를 하였다. 인식기 측면의 비교에서 정확한 분석 결과를 위하여 성인 음성인식기와 유아 음성인식기를 구현하여 각 인식기에 성인과 유아의 음성으로 실험하였다. 아래는 한국어에서 성인과 유아의 음성을 비교한 실험 결과이다.

- 유아의 음성 인식기로 유아의 음성을 실험할 때 특정 숫자열 '영'과 '육'의 단어가 혼동되었으며 이는 두 단어의 'ㅕ'와 'ㅠ'가 반모음 /y/를 포함한 이중 모음이기 때문이다.
- 음성 인식률 향상을 위한 VTLN의 사용이 유아의 음성 인식률을 향상시킴을 알 수 있었다.
- 성인과 유아의 음성을 통합하여 훈련한 인식기를 구현하여 실험한 결과는 성인의 음성인식기로 성인의 음성을 실험한 인식률과 유아의 음성인식기로 유아의 음성을 실험한 인식률보다 낮게 나왔다. 이 실험을 통해 음성인식기 구현 시 대상 화자에 맞는 음성 데이터의 수집이 중요하다는 사실을 알게 되었다.

현재 국내 음성데이터베이스 구축 현황을 살펴보면 미취학 아동인 만 3~5세 대상의 음성 데이터베이스는 구축이 되지 않은 실정이다. 유아에게 보다 즐거운 놀이와 교육을 통해 풍부한 경험을 주기위해 유아에게 맞춰진 콘텐츠 개발은 필수적이며, 유아에 맞는 인터페이스 연구도 중요하다. 이 연구를 바탕으로 향후 유아 음성데이터베이스를 구축하여 유아에게 맞는 인터페이스를 이용한 콘텐츠를 개발할 것이다.

부록. 약어표	
HMM	Hidden Markov Model
VTLN	Vocal Tract Length Normalization
HTK	Hidden markov model Tool Kit
MFCC	Mel Frequency Cepstral Coefficients
MBL	Microcomputer Based Laboratory
FFT	Fast Fourier transform
LPC	Linear Predictive Coding

참 고 문 헌

[1] 이용주, 김봉완, 김영일, 최대림, "한국의 공동이용을 위한 음성언어자원의 구축 및 보급현황", 한국어정보학회, 제10권, 제1호, pp.81-85, 2008.

[2] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," Proc. ICASSP, pp.137-140, 2003.

[3] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," IEEE Trans. on Speech and Audio Processing, Vol.10, No.2, pp.65-78, 2002.

[4] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," Speech Communication 49, pp.847-860, 2007.

[5] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in Proceedings of FONETIK, Stockholm, Sweden, 2004.

[6] H. Wakita, "Normalization of vowels by vocal tract length and its application to vowel

identification," IEEE Trans. on Acoustic. Speech and Signal Processing, 25, pp.183-192, 1977.

[7] S. öhgren, "Experiment with adaptation and vocal tract length normalization at automatic speech recognition of children's Speech," KTH, Stockholm, Sweden, 2007.

[8] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash and K. Johnson, "Formants of children women and men: The effect of vocal intensity variation," Journal of the acoustical society of america. Vol.106, No.3, pp.1532-1542, 1999.

[9] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in Proceedings of IEEE Multimedia Signal Processing Workshop, 2007.

[10] 장보경, 이연규, "유아의 연령과 성별에 따른 언어발달과 사회정서 발달의 차이", 한국 Montessori 교육학회, Vol.14, No.2, pp.61-77, 2009.

[11] G. Potamianos and S. Narayanan, "Robust recognition of children speech," IEEE Transaction on Speech and Audio Processing 11, pp.603-616, 2003.

[12] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," Speech Communication. Vol.49, No.12, pp.861-873, 2007.

[13] R. D. Kent and L. L. Forner, "Speech segment durations in sentence recitations by children and adults," Journal of Phonetics, 8, pp.157-168, 1980.

[14] S. Lee, A. Potamianos, and S. Naraynan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," Journal of the Acoustical Society of America, pp.1455-1468, 1999.

[15] H. John and H. Wendy, "Speech synthesis and recognition," Taylor & Francis, 2nd edition, 2001.

[16] J. Nicholas and A. Geers, "Effects of early auditory experience on the spoken language of deaf children at 3 years of age," Ear Hear, 27, pp.286-296, 2006.

저 자 소 개

유 재 권(Jae-Kwon Yoo) 준회원

- 2009년 2월 : 덕성여자대학교 인터넷정보공학과(공학사)
- 2011년 3월 ~ 현재 : 덕성여자대학교대학원 전산정보통신학과(석사과정)

<관심분야> : 음성인식, 유아 음성 DB

이 경 미(Kyoung-Mi Lee) 정회원

- 1993년 2월 : 덕성여자대학교 전산학과(이학사)
- 1996년 2월 : 연세대학교 전산과학과(이학석사)
- 2001년 12월 : 아이오와 주립대학교 전산학과(이학박사)
- 2003년 3월 ~ 현재 : 덕성여자대학교 컴퓨터학과 교수

<관심분야> : 영상처리, 패턴인식, 멀티미디어. HCI