마이크로폰의 종류에 따른 음성인식성능의 검토

김연화*, 이광현**, 정영조**, 김봉완**, 이용주*
원광대학교 전기전자 및 정보공학부*, 원광대학교 SiTEC**

The Validation of Speech Recognition Performance according to Microphones

Yoen-Whon Kim*, Kwang-Hyun Lee**, Young-Jo Jung**, Bong-Wan Kim**,Yong-Ju Lee**
Div. of Electric and Electronic Eng., Wonkwang Univ.*, SiTEC,Wonkwang Univ.**
E-mail : kimyw@wonkwang.ac.kr, {khlee, notpp, bwkim, yjlee}@sitec.or.kr

Abstract

Speech recognition performance depends on various factors. One of the factors is the characteristic of a microphone which is used when speech data is collected. Thus, in the present experiment speech databases for tests are created through varying types of microphones. Then, acoustic models are built based on these databases, and each of the acoustic models is assessed by the data to determine recognition performance depending on various microphones.

1. 서론

음성 인식 성능은 여러 가지 요인들에 의해서 달라질 수 있다. 그 요인들 중의 하나로 음성데이터를 수집 및 인식에 사용된 마이크로폰의 특성을 들 수 있다. 음성인식에 따른 훈련용 음성DB 및 테스트용 음성DB에 사용되는 마이크로폰의 차이에 따른 영향을 고찰 하기 위하여 공통환경의 시험용 음성 DB를 구축하였다. 이에 따른 마이크로폰 종류별 음성 데이터의 수집과정과 이 DB를 바탕으로 실제 인식 실험한 결과를 통해 사용된 마이크로폰의 종류에 따른 영향을 검토하였다.

2. 마이크로폰종류별 시험용 음성DB의 설계 및 구축

2.1 DB 수집 개요

음성 인식 시스템의 구축 과정에서 우선적으로 당면하는 문제는 데이터 수집 및 시스템의 입력 단을 구성하는 마이크로폰(Microphone)의 선택일 것이다. 음성 수집 및 음성 인식에서의 마이크로폰 성능은 수집된 데이터의 품질을 결정지을 뿐만 아니라 인식 시스템의 성능을 좌우할 수 있기 때문이다.

마이크로폰의 선정은 최종 어플리케이션에 대한 발성 환경이나 시스템의 특성에 따라 결정되어지는 것이 일반적이지만, 특징에 따른 마이크로폰의 종류가 무수히 많을 뿐만 아니라 목적에 적합한 유닛(Microphone Unit)을 설계하거나 선택하는 것도 쉬운 일 만은 아니다.

따라서, 형태 및 특징에 따라 분류되는 다수의 마이크로폰을 대상으로 공통 환경에 대하여 음성 데이터를 수집하였고, 이를 바탕으로 각 마이크로폰에 따른 인식 성능을 비교하였다.

본 데이터베이스 구축에는 다양한 특성을 갖는 8종의 마이크로폰 모델이 이용되었으며, 기존 방음부스에서 수집된 70명분의 PBW 452 단어를 방음부스에서 HATS(Head and Torso Simulator)를 통해 재생한 음향 신호에 대하여 각 마이크로폰이 수음(Sound Pick-up)하도록 하는 방식을 취하였다.

2.2 데이터 수집 시스템

마이크로폰 종류에 따른 시험용 음성 DB를 수집하기 위하여 그림 1과 같은 녹음 시스템을 구성하였다. 구축된 70명의 화자 발성은 Hand-held Type과 Boundary Type의 경우 20cm 거리에서 0° 방향으로 하였으며, Headset의으로부터 얻어진 PBW452 데이터는 Master PC에서 재생하도록 하였고, 재생 신호는 증폭기를 통해 방음 부스 내의 HATS로 입력된 후 Mouth Simulator를 통해 공간상에 방출되어 설치된 마이크로폰에 의하여 수음되도록 하였으며, 마이크로폰에서 얻어진 신호는 프리 앰프를 거쳐 Slave PC로 녹음되도록 하였다.

이 때, Mouth Simulator와 마이크로폰 간의 거리 및 각도 설정 마이크는 5cm의 거리에서 수평 15° 각도를 취하였다.

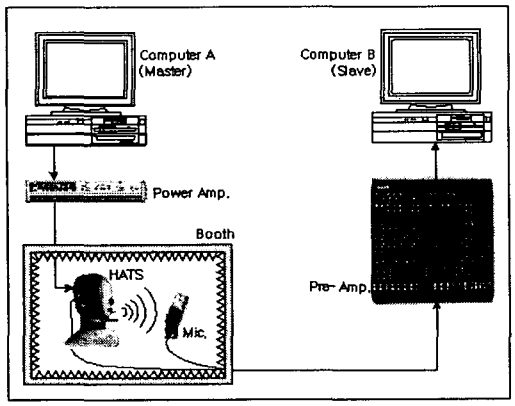

그림 1. 데이터 수집 시스템 구성도

방음 부스의 실내음향 파라메터와 주요 사용 장비는 표 1과 같으며, 실험에 이용된 마이크로폰의 모델 및 특성을 표 2에 나타내었다.

표 1. 방음 부스 및 장비의 제원

Booth /Equipment	Parameter / Model
Recording Booth	Volume : 1.5m³(1×1×1.5) RT : 0.1s (@500Hz) NC : 15 Background Noise : 20dB(A)
Equipment	Master PC(for Playback) : CPU 1.8GHz Slaver PC(for Recording) : CPU 1.4GHz HATS : Model 4128 (B&K) Audio Mixer : MXB1002 (Behringer) Power Amp. : 50W/ch, 0.05 %THD

표 2. 마이크로폰의 모델과 대표적 특성

Model	Mount/Unit Type	Directivity	Frequency Response
PBW452 DB 수집용 마이크로폰			
HMD 224X (SENNHEISER)	Head-worn Dynamic	Super-Cardioid	50~12,000Hz
실험에 이용된 마이크로폰(8종)			
280-Pro (SENNHEISER)	Head-worn Dynamic	Super-Cardioid	50~13500Hz
e825-s (SENNHEISER)	Hand-held Dynamic	Cardioid	80~15,000Hz
SM-10A (SHURE)	Head-worn Dynamic	Cardioid	50~15,000Hz
16A (SHURE)	Desktop Condenser	Cardioid	50~15,000Hz
C400-BL (AKG)	Boundary Condenser	Cardioid	150~15,000Hz
C414B-ULS (AKG)	Hand-held Condenser	Cardioid	20~20,000Hz
ANC-700 (ANDREA)	Head-worn ECM	Omni-directional	100~10,000Hz
M@B40 (SENNHEISER)	Head-worn ECM	Cardioid	40~12,500Hz

2.3 데이터 수집 절차

마이크로폰 종류에 따른 시험용 음성 DB를 구축하기 위해 사용된 재생 목록은 HMD 224X(SENNHEISER) 마이크로폰으로 수집된 Clean Speech 환경에서의 PBW 452 어절 70명분의 DB를 활용하였다.

기존에 구축된 이 DB는 모든 화자 발성에 대하여 단어 단위로 세그먼트 되어진 형태를 가졌고, 따라서 Mouth Simulator(HATS)를 통한 재생과 녹음 과정의 간소화를 위해 화자 성별로 분절된 각 발성 데이터를 연속 발성의 형태로 변환하여 최종적으로 2개(남성, 여성)의 재생 데이터만을 이용하였다.

한편, 연속 발성을 녹음한 후에는 Auto-Segmentation을 수행하기 위하여 분절된 원본 데이터를 통한 Log 파일을 사전에 작성해 두었다.

남성 화자 39명분과 여성 화자 31명분으로 구성된 2개의 재생 데이터는 각각 재생용 PC(Master PC)와 시뮬레이터를 통해 재생되었고, 재생된 음향 정보는 다시

같은 음장(Sound Field)을 통하여 각 시험용 마이크로폰 모델에 각각 입사되어 프리 앰프를 거친 후 녹음용 PC(Slave PC)로 입력되었다.

재생과 녹음을 위한 툴은 Cool Edit Pro를 사용하였으며, 음향 및 신호 전송계에서의 시간 지연(Time Delay)은 매우 작을 것으로 판단되어 동기화 에러에 대한 부분은 무시하였다. 데이터 수집 절차를 나타낸 다이어그램은 그림 2와 같다.

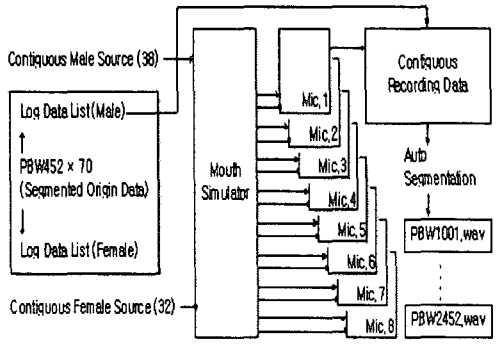

그림 2. 마이크로폰별 데이터 수집 절차

2.4 재생 목록 작성과 분할 정보 생성

여러 가지 마이크로폰에 대한 동일 환경에서의 데이터 수집을 위해 HATS로부터의 발성 데이터는 PBW 452 어절 DB를 이용하여 일반인 화자 70명에 의해 수집된 것으로서, 각 화자에 대하여 452어절을 2회씩 발성하도록 한 것이다. 발성 화자의 성별, 연령별 구성은 표 3과 같다.

표 3. PBW 452 DB의 화자 구성

연령 / 성	10대	20대	30대	40대	계
남성	3	27	6	3	39
여성	1	25	4	1	31

한편, 각 마이크로폰을 통해 녹음된 데이터를 기존에 구축된 어절 단위로 분할하기 위하여 원시 데이터 정보로부터 Log List를 작성하였으며, 이를 통해 그림 2에서와 같이 연속적으로 녹음된 데이터에 대하여 Auto-Segmentation을 수행하였다.

3. 마이크로폰의 종류별 음성인식실험

3.1 음성데이터의 준비

본 연구를 위한 음성데이터는 방음부스에서 SENNHEIZER HMD224X를 사용하여 녹음한 70명분의 Clean Speech pbw452어절과 이 어절을 음원으로 해서 수집한 재생음성 데이터를 이용하였다.

3.2 음성데이터의 분석

표 4. 음성데이터의 분석조건 및 특징파라미터

음성데이터	
Sampling frequency	16kHz
Resolution	16bits
Window	Hamming window
Window size	25ms
Window shift rate	10ms
Preemphasis factor	0.97
Number of filterbanks	26
Liftering order	22
Cepstrum order	12
Freature parameters (39차)	MFCC(12)+E(1)+△ MFCC(12)+△E(1)+△ △MFCC(12)+△△E(1)

3.3 음향모델의 작성과 음성인식시스템

CHMM(Continuous HMM)을 이용하여 초기 음소모델을 작성하였으며, 5 state left-to-right 구조를 사용하였고, 42개 PLU를 사용하여 Triphone모델로 구성하였다. 인식실험은 HTK(ver 3.1.1)를 사용하였다.

3.4 음성인식 실험

본 연구에서는 Clean Speech와 재생음성데이터를 이용하여 음향모델을 작성하고 각 음향모델을 Clean Speech와 재생음성데이터를 이용하여 평가하였다.

3.5 음성인식 실험의 결과

표 5. 마이크로폰종류별 인식결과 (*D: Dynamic, C:Condenser, EC:Electret Condenser)

Model \ TestData		D				C			EC	
		HMD224X	280Pro	E825	SM10A	16A	C400	C414	ANC700	M@400
D	HMD 224X	99.16	98.57	98.78	98.23	98.14	98.31	98.51	98.47	97.87
	280Pro	98.69	99.24	98.84	99.21	98.13	98.05	99.12	98.41	98.86
	E825	99.02	99.22	99.28	99.21	98.87	98.73	99.19	98.89	98.46
	SM10A	98.63	99.21	98.89	99.20	98.55	98.27	99.14	98.42	99.02
C	16A	98.38	98.85	98.95	98.73	99.09	98.70	99.04	99.03	97.73
	C400	98.80	98.95	98.98	98.87	98.95	99.12	98.85	99.00	98.41
	C414	98.22	99.05	98.62	98.84	98.27	97.83	99.15	98.50	98.52
EC	ANC700	98.86	99.05	99.08	98.94	99.05	98.73	99.08	99.15	98.03
	M@B400	97.16	98.98	96.74	98.84	96.01	97.25	98.85	96.65	99.08
평균 인식률		98.55	99.01	98.68	98.90	98.34	98.33	98.99	98.50	98.44

표 6. 모델별 동일 마이크로폰을 제외한 최고, 최저 인식률을 보인 마이크로폰

모델 훈련용 데이터		시험용 데이터	
마이크로폰 유형	모델	최고인식률을 보인 마이크	최하인식률을 보인 마이크
D	HMD224X 280Pro E825 SM10A	E825(D) SM10A(D) 280Pro(D) 280Pro(D)	M@B40(EC) C400(C) M@B40(EC) C400(C)
C	16A C400 C414	C414(C) ANC700(E.C) 280Pro(D)	M@B40(EC) M@B40(EC) C400(C)
EC	ANC700 M@B40	C414(C),E825(D) 280Pro(D)	M@B40(EC) 16A(C)

3.6 음성인식실험결과의 검토

• 일반적으로 알고 있는 바와 같이 학습 데이터와 인식 데이터를 수집하기 위한 마이크로폰의 종류가 동일할 경우 가장 좋은 인식 성능을 내고 있다.

• 다이나믹 마이크로폰으로 수집된 데이터를 통하여 학습된 모델을 사용할 경우 다이나믹 마이크로폰으로 수집된 테스트 데이터들에서 최고 성능을 보이며, EC나 C에서 최하 성능을 보인다.

• 그러나,C나 EC의 경우 이러한 경향성이 다소 모호하게 나타나고 있다.

• 최고 인식률을 보인 마이크로폰들
-280Pro(D):4회,E825(D):2회,C414(C):2회,SM10A(D):1회,ANC700(EC):1회

• 최하 인식률을 보인 마이크로폰들
- M@B40(EC):5회,C400(C):3회,16A(C):1회

• 좋은 인식 성능을 보이는 마이크로폰과, 그렇지 않은 마이크로폰이 구별되는 양상을 보이고 있다. 즉 최고 인식률을 보인 마이크로폰들은 한번도 최하 인식률을 보이지 않았으며, 최하 인식률을 보인 마이크로폰들은 한번도 최고 인식률을 나타내지 않았다.

• 마이크로폰에 따른 인식률에서, 마이크로폰 간 대칭성을 보이지 않고 있다. 즉 HMD240X로 학습된 모델의 경우 E825의 테스트 데이터를 잘 인식하는 것으로 나타났다. 그러나 E825로 학습된 모델의 경우 HMD224X보다 280Pro,테스트 데이터를 더 잘 인식하는 것으로 나타났다. 또한 280Pro로 학습된 모델의 경우 HMD224X나, E825가 아닌 SM10A 테스트 데이터를 더 잘 인식하고 있는 것으로 나타났다.

• 그러나 원본데이터를 HMD224X로만 받았다는 실험적인 한계가 있으며, 각 경우별 인식률이 확연하게 차이가 나지 않기 때문에 절대적인 데이터로 삼기에는 향후 연구가 더 필요하다.

4. 결론

본 논문에서는 음성인식에서 훈련 및 테스트에 사용되는 마이크로폰의 특성의 차이가 인식률에 미치는 영향을 검토하기 위해 구축한 시험용 음성DB의 설계 및 제작과정을 보고하였고, 이를 이용한 인식실험을 통하여 그 영향을 검토하였다. 제한 된 시험조건이나 마이크로폰의 특성에 따른 개략적인 경향은 살펴볼 수 있었다.

참고 문헌

[1] S.J. Young, The HTK Book, Cambridge, 1997
[2] 원광대학교 음성정보기술산업지원센터
VariMic01DB, CD-ROM, 2002