소음 환경에서 body-conducted 신호를 이용한 음성인식 성능 비교

최대림*, 이광현*, 이용주**, 김종교***
*원광대학교 음성정보기술산업지원센터, **원광대학교 전기전자 및 정보공학부
***전북대학교 전자정보공학부

Performance Comparison of Speech Recognition Using Body-conducted Signals in Noisy Environment.

Dae-Lim Choi*, Kwang-Hyun Lee*, Yong-Ju Lee**, Chong-Kyo Kim***
*Speech Information Technology & Industry Promotion Center(SiTEC)
**Dept. of Electrical, Electronics and Information Eng. Wonkwang University
***Dept. of Electronics and Information Eng. Chonbuk National University
E-mail: {dlchoi, khlee}@sitec.or.kr, yjlee@wonkwang.ac.kr, ckkim@chonbuk.ac.kr

요약

본 논문에서는 음성정보기술산업지원센터(SiTEC)에서 현재 배포중인 고소음 환경 음성 DB를 이용하여 air-conducted 음성과 body-conducted 음성의 인식 성능을 비교 실험하였다. 소음 환경에서 일반적인 마이크로폰으로부터 수집된 air-conducted 음성은 잡음의 영향을 받기 쉬우며 이는 인식률을 저하시킨다. 반면에 진동 픽업 마이크로폰에서 수집된 body-conducted 음성은 소음에 보다 강인한 특성을 보인다. 이러한 특성에 근거하여 소음 환경에서 일반 다이나믹 마이크로폰 음성에 음질 개선 방법과 채널 보상 방법을 적용한 인식 결과와 3종류의 진동 픽업 마이크로폰에서 수집된 음성과의 인식 성능을 비교 분석하여 body-conducted 음성 인식 시스템의 활용 가능성을 살펴보았다.

1. 서론

최근 음성인식 기술은 자동차 주행 소음 상황에서와 같은 다양한 환경에서 응용되고 있다. 그러나 음성인식 시스템은 실험실 환경이나 조용한 일반 사무실 환경에서는 비교적 높은 성능을 나타내지만 소음 환경에서는 인식률이 현저히 저하된다. 특히 군사작전 현장, 건설현장, 공장, 항공, 선박, 공연장, 스포츠 경기장과 같은 고소음 환경에서는 SNR이 매우 낮아 음성 인식 성능을 기대할 수 없으며 따라서 관련 연구는 그다지 주목을 받지 못했다.

일반적으로 음성 데이터 수집에 사용되는 마이크로폰은 공기 중에 방사된 음향 신호를 픽업하기 때문에 외부 잡음에 영향을 받기 쉬우며 이는 인식 성능의 저하를 야기한다. 그러나 진동 픽업 마이크로폰은 두개골이나 성대를 통해 전달되는 진동을 고감도의 센서 소자로 집음하므로 주변 잡음의 영향을 좀처럼 받지 않아 고소음 환경에서 보다 강인한 인식을 기대할 수 있다.

따라서 본 논문에서는 동일한 조건에서 일반적인 마이크로폰으로 수집한 air-conducted 음성 신호와 새로운 입력원인 진동 픽업 마이크로폰으로부터 수집된 body-conducted 음성 신호를 비교 분석하고, 소음 환경에서 body-conducted 신호를 이용한 인식 시스템의 활용 가능성을 살펴보고자 한다. 이를 위해 다이나믹 마이크로폰과 headgear, ear, neck 타입의 진동 픽업 마이크로폰을 사용하여 무소음 환경과 70dBA, 90dBA(SPL)의 화이트노이즈를 재생한 소음환경에서 음성 신호를 수집하고 인식 실험을 수행하였다. 또한 소음의 영향으로 인한 훈련 환경과 인식환경의 불일치를 줄이기 위해 음질 개선

방법인 주파수 차감법(spectral subtraction), 위너 필터링(Wiener filtering) 방법, 채널 왜곡의 영향을 보상하는 기법인 CMN(cepstral mean normalization)을 적용하여 성능을 비교 분석하였다.

2. 데이터 수집 및 분석

2.1 고소음 환경 음성 DB의 구축

일반 다이나믹 마이크로폰 1종과 진동 픽업 마이크로폰을 3종을 이용하여 무소음 환경과 소음 환경에서 음성 데이터베이스를 구축하였다. 음성 데이터 수집에 이용되는 마이크로폰의 종류는 그림 1과 같으며 방음 부스에서 멀티트랙 레코더를 이용하여 4채널 동시 녹음하였다. 소음환경은 무소음 환경과 다이나믹 마이크 위치에서 음압 레벨이 각각 70dBA, 90dBA가 되도록 스피커를 통하여 화이트 노이즈를 재생한 소음 환경으로 구성된다. 무소음 환경에서는 40명분의 데이터를 수집하고 70dBA와 90dBA의 소음 환경에서는 각각 20명씩 데이터를 수집하였다. 한편 소음 정도에 따른 화자의 롬바드 효과를 분석할 수 있도록 10명의 화자는 3가지 소음 환경에서 재차 발성하도록 하였다[1].

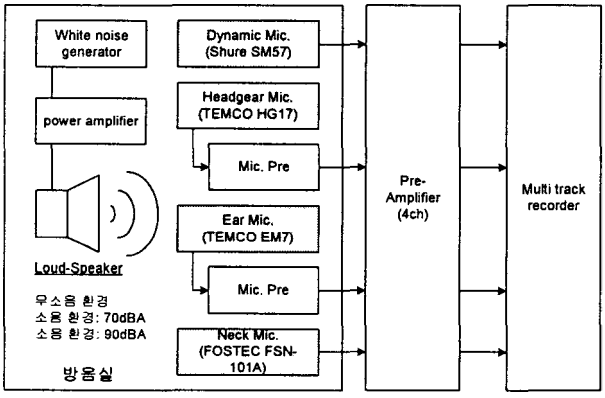

그림 1. 마이크로폰의 종류 및 수집 시스템의 구성

2.2 Body-conducted(Bone-conducted) 음성 분석

본 논문에서 사용된 body-conducted 음성 신호는 3가지의 다른 위치(headgear, ear, neck)에서 픽업되며 수집에 사용된 진동 픽업 마이크로폰에 대한 소개는 다음과 같다. Headgear 타입의 HG17은 고출력의 골도 스피커를 두 귀에 배치하고 송신용의 골도 마이크로폰은 머리 위 중앙에 위치하고 있다. Ear 타입인 EM7의 보이스튜서 마이크로폰은 뼈를 통해 들어오는 성대의 진동, 즉 골도(bone-conduction)음을 외이도에서 픽업하여 신호를 전달하는 구조이며 외형상으로는 이어폰과 동일하다[2]. 또한 Neck 타입의 FSN-101A는 목 부분에 접촉 장착되어 발성이 이루어질 때 목 부분의 피부조직의 울림을 센싱하여 음성 파형을 얻게 된다. 일반적으로 body-conducted 음성은 air-conducted 음성의 파형과 매우 유사하지만 명료도가 낮은 편이고 재생할 때 음색이 다르게 나타난다.

그림 2는 남성 화자의 */그때 누가 그녀의 책상 앞으로 다가왔다./* 발성에 대한 dynamic, headgear, ear, neck 마이크로폰에서 동시에 수집된 음성의 스펙트로그램을 보여주는데, body-conducted 음성은 4kHz 이상의 대역에서는 감도가 떨어지는 주파수 특성을 보이며 F2 이하의 포만트 정보만을 얻을 수 있다. 이는 진동 픽업 마이크로폰의 특성과 body-conduction을 통한 전달 특성에서 대역 제한에 기인한 것이다. 한편 body-conducted 음성 신호에서 저주파 대역의 에너지의 강도는 headgear, ear, neck 마이크로폰의 순으로 감쇄되며 이는 인식 성능에 크게 영향을 미칠 것으로 판단된다. Body-conducted 음성에서 주관적인 청감 음질은 headgear 타입이 가장 우수하고 주변 소음에 보다 강인한 특성을 보이며, 같은 조건하에서 신호가 입력될 때에는 ear 타입이 가장 높은 출력을 나타낸다.

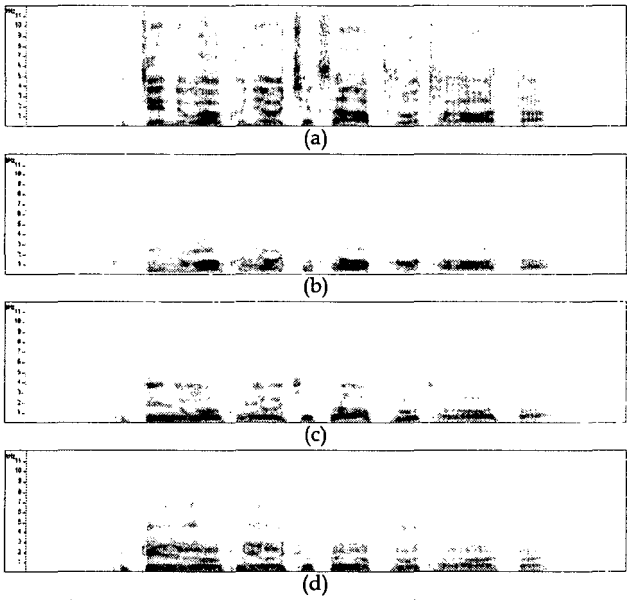

그림 2. /그때 누가 그녀의 책상 앞으로 다가왔다./ 음성에 대한 스펙트로그램:
(a)Dynamic Mic. (b)Headgear Mic. (c)Ear Mic. (d)Neck Mic.

3. 소음 환경에서 음질 개선

소음 환경에서 진동 픽업 마이크로폰의 특성을 살피

고 body-conducted 음성 신호의 소음에 대한 강인성을 파악하기 위해 다이나믹 마이크로폰에서 수집된 음성에서 잡음 제거 방법을 적용시킨 결과와 body-conducted 음성 신호를 이용한 인식 결과를 비교 실험하였다.

본 논문에서는 훈련 조건과 인식 조건 사이의 불일치(mismatch)를 줄이기 위해 주파수 차감법과 위너 필터링 방법을 적용하여 음질 개선을 시킨 후 얻어진 결과를 사용한다[3].

3.1 주파수 차감법

전처리 과정에서 잡음을 제거하는 기법 중에서 간단하고 구현이 용이하여 널리 쓰이는 주파수 차감법은 잡음이 섞인 음성(noisy speech)의 power spectrum $Y^2(m,f_k)$에서 잡음의 power spectrum $N^2(m,f_k)$을 제거함으로써 원음성(clean speech)의 power spectrum $S^2(m,f_k)$를 구하는 것으로서 식 (1)과 같이 표현할 수 있다. 여기에서 m은 power spectrum상에서 프레임의 값이고, f_k는 k번째 spectral component를 나타낸다.

$$S^2(m,f_k) = Y^2(m,f_k) - N^2(m,f_k) \tag{1}$$

잡음의 power spectrum을 모를 경우에는 식 (2)를 이용하여 추정된 잡음 power spectrum $\hat{N}(m,f_k)$을 이용하여 noisy speech에서 잡음을 제거한다.

$$H(m,f_k) = 1 - \left(\frac{\hat{N}(m,f_k)}{Y^2(m,f_k)} \right)^{1/2} \tag{2}$$

3.2 위너 필터링 방법

위너 필터는 음성 신호의 단구간 스펙트럼 크기(short time spectral amplitude)를 추정하여 음성 신호의 음질을 향상시키는 방법이다. 잡음 $n(k)$이 섞인 음성 신호 $y(k)$로부터 $x(k)$를 얻기 위한 위너 필터는 식 (3)과 같다.

$$\begin{aligned} G(w) &= \frac{P_{yn}(w)}{P_{yy}(w)} \\ &= \frac{P_{yy}(w) - P_{nn}(w)}{P_{yy}(w)} \\ &= \frac{|Y(w)|^2 - \hat{\mu}(w)^2}{|Y(w)|^2} = 1 - \frac{\hat{\mu}(w)^2}{|Y(w)|^2} \end{aligned} \tag{3}$$

이득 함수를 살펴보면 특정 주파수 대역에서 SNR이 높으면 $H(w)$가 1에 가까워져서 필터는 관측 신호를 전부 통과 시키는 반면에 SNR이 낮으면 $H(w)$가 0에 가까워져서 관측 신호를 통과시키지 않게 된다. 따라서 위너 필터는 잡음 성분이 큰 주파수 영역은 통과시키지 않고 음성 신호 성분이 큰 주파수는 그냥 통과 시키는 특성을 갖고 있다.

4. 실험 및 결과

4.1 실험 환경

인식 실험을 위하여 29명의 화자가 PBW(Phonetically Balanced Words) 452단어를 발성한 데이터를 사용하였으며 한 화자는 226단어를 발성하였다. 인식 실험은 훈련 데이터 분량이 충분하지 않다고 판단되어 close test로 수행하였으며 9명(남성 5, 여성 4)에 한하여 테스트를 수행하였다.

음성신호는 20ms의 윈도우 구간에 10ms씩 중첩하여 분석하였고, 음향 모델 훈련 및 평가에 사용된 특징 벡터로는 C0를 포함한 MFCC 13차, delta, delta-delta를 이용하여 총 39차 음성특징벡터를 적용하였다. 인식에 사용된 HMM 모델은 트라이폰 단위의 1 gaussian mixture를 사용하여 3 state의 left-to-right 방식의 연속밀도 HMM을 기반으로 하였다.

4.2 Matched condition

다이나믹 마이크로폰과 진동 픽업 마이크로폰 종류에 따른 무소음 환경과 소음 환경에서의 인식률의 차이를 비교하기 위하여 훈련 과정에서 사용된 데이터와 인식 과정에서 테스트를 위한 데이터들 사이에 환경이 일치된 경우의 인식률을 조사하였고 그 결과는 표 1과 같다.

*DYN: Dynamic Mic., HGR: Headgear Mic.
EAR: Ear Mic. NEC: Neck Mic.

	DYN	HGR	EAR	NEC
clean	98.21	85.42	73.91	61.95
noise(70dBA)	98.08	85.70	72.44	62.52
noise(90dBA)	94.34	76.64	66.57	55.29

표 1. Matched condition에서의 인식률(%)

해당 마이크로폰과 실제 소음 환경의 음성 데이터를 이용하여 음향 모델을 작성하고 테스트를 수행하였기 때문에 소음 환경 변화에 따른 인식률의 하락은 크지 않다. 예상처럼 다이나믹 마이크로폰에서 수집된 air conducted 음성의 인식률이 가장 높으며, body-conducted 음성은 headgear, ear, neck 마이크로폰 순으로 인식 성능이 좋다.

4.3 Mismatched condition: clean *vs.* noisy

일반적으로 성능이 좋은 인식기를 만들기 위해서는 깨끗한 환경에서 구축된 음성을 이용하여 음향 모델로 작성하지만, 테스트 음성이 주변 잡음의 영향을 받을 경우 모델과 테스트 음성과의 환경 불일치로 인하여 성능이 저하된다. 따라서 무소음 환경의 음성을 이용하여 모델을 훈련하고 잡음 환경에서의 음성으로 테스트 하였을 때 결과를 표 2에 나타내었다.

	DYN	HGR	EAR	NEC
clean	98.21	85.42	73.91	61.95
noise(70dBA)	64.91	77.81	59.17	54.39
noise(90dBA)	1.54	43.01	8.18	22.39

표 2. Mismatched condition에서의 인식률(%)

각각의 마이크로폰에서 무소음 환경의 음성을 이용하여 음향 모델을 작성한 후, 70dBA와 90dBA의 소음 환경의 음성을 인식 할 경우에는 급격한 인식률의 저하를 보인다. Body-conducted 음성 신호는 air-conducted 음성 신호에 비해 전반적으로 소음 환경에서 강인한 특성을 보이며, 90dBA의 고소음 환경에서는 다이나믹 마이크로폰의 경우 배경 잡음이 그대로 반영되어 거의 인식이 되지 않는 반면, headgear 타입의 마이크로폰은 소음에 가장 강인한 특성을 보이며 headgear, nec, ear 마이크로폰 순으로 좋은 인식 성능을 나타내었다.

한편 채널 왜곡으로 인한 영향을 줄이기 위한 CMN 방법과 잡음으로 오염된 음성을 개선시키기 위한 주파수 차감법, 위너 필터링 방법 적용에 따른 인식 결과는 표 3과 같다.

Compensation	Env.	DYN	HGR	EAR	NEC
CMN	clean	98.56	87.00	83.28	64.19
	70dBA	91.77	80.28	72.93	57.69
	90dBA	23.21	70.43	59.85	37.43
SS	70dBA	93.33	37.29	11.69	31.98
	90dBA	21.72	24.46	5.15	23.20
SS+CMN	70dBA	94.37	40.20	26.06	34.52
	90dBA	65.11	23.33	14.64	26.94
WF	70dBA	95.71	32.40	17.97	38.58
	90dBA	8.10	52.63	8.83	35.24
WF+CMN	70dBA	96.90	55.89	59.81	43.69
	90dBA	69.72	67.72	52.33	43.56

표 3. 주파수 차감법(SS), 위너필터(WF), CMN을 적용하였을 경우 인식률(%) 비교

다이나믹 마이크로폰에서 수집된 음성은 소음 레벨이 커질수록 위너필터와 CMN을 결합시킨 방법에서 가장 좋은 인식 성능을 얻을 수 있었으며, 진동 픽업 마이크로폰은 음질 개선 방법보다는 채널 왜곡을 줄이기 위한 CMN 방법을 적용하였을 경우 보상 능력이 크며 보다 효과적이었음을 알 수 있다.

5. 결론

본 논문에서는 무소음과 소음 환경에서 일반적인 다이나믹 마이크로폰으로부터 수집된 air-conducted 음성과 새로운 입력원인 headgear, ear, neck 타입의 진동 픽업 마이크로폰에서 동시에 수집된 body-conducted 음성을 이용하여 인식 실험을 하였다. 70dBA, 90dBA의 소음 환경에서 air-conducted 음성은 환경 불일치로 인하여 인식 성능이 급격히 저하되지만, headgear 마이크로폰에서 수집된 body-conducted 음성은 비록 주파수 대역폭이 제한되어 있고 청감 명료도가 높지 않지만 외부 잡음에 영향을 덜 받으므로 보다 소음에 강인한 특성을 보였다.

음질 개선 방법과 채널 왜곡 보상 방법을 적용한 인식 실험의 경우, 90dBA의 고소음 환경에서 다이나믹 마이크로폰은 위너필터와 CMN을 결합하였을 때, headgear 마이크로폰은 CMN만을 적용하였을 때 각각 69.72%, 70.43%의 인식 성능을 보였다. 이처럼 보상 알고리즘을 달리 적용하였을 경우 air-conducted 음성과 headgear 타입의 body-conducted 음성의 최고 인식 성능의 차이가 거의 없음을 확인할 수 있다.

Body-conducted 음성의 경우에는 채널영역에서의 왜곡을 보상하는 방법이 인식 성능 향상에 상당히 효과적이었으므로 향후 잡음에 강인한 특징 추출 방법과 보다 적합한 채널 왜곡 보상 방법이 모색되어야 하며, SNR이 0dB 이하인 고소음 환경에서 body-conducted 음성에 대한 연구가 향후 과제로 남아 있다.

참고문헌

[1] http://www.sitec.or.kr

[2] 강성훈, 성만순, 김상훈, "소음 환경하에서 골도 마이크를 이용한 음성 인식 실험", 한국 음향학회 추계학술 대회 제 21권 제 2호, pp.131-134, 2002.

[3] Gillian M. Davis, "Noise Reduction in Speech Applications", CRC Press