자동차 주행 환경에서의 음성 인식 성능 및 음향 특성의 검토

이광현, 최대림, 김영일, 김봉완, 이용주*
원광대학교 음성정보기술산업지원센터, 원광대학교 전기전자 및 정보공학부*

A Study on Performance of Speech Recognition & Acoustic Parameter in Car Environment

Kwang-Hyun Lee, Dae-Lim Choi, Young-Il Kim, Bong-Wan Kim, Yong-Ju Lee*
SiTEC, *Dept.of Electric and Electronic Eng.,Wonkwang Univ.
{khlee, dlchoi, yikim, bwkim}@sitec.or.kr, yjlee@wonkwang.ac.kr

요약

주행 상태에서의 자동차 내부 음 환경은 다양한 소음 및 구조적 요인으로 인하여 음성에 대한 정상적인 전송 특성을 갖기 어렵다. 이는 음원으로부터 음성 입력 장치(Microphone)에 이르기까지의 채널 왜곡에 기인한 문제로써, 실제 주행 환경에서의 음성 인식 성능에 대해서도 심각한 악영향을 초래한다. 본 논문에서는 주행 소음의 크기에 따른 채널별 음성 왜곡 정도에 따른 명료도를 음성 전달 지수인 STI(Speech Transmission Index)를 통하여 분석하고 그 결과를 음성 인식률과 상호 비교하였다. 그리고 수음 패턴에 따른 명료도 척도와 음성 인식 성능과의 상관성을 검토하고, 이를 통해 단일 채널 환경에서 최적의 마이크로폰 위치에 대하여 고찰해 보았다.

실험 결과, 주행 중의 소음 환경에서도 음성의 명료도 척도와 인식률과의 관계는 높은 상관성이 얻어짐을 알 수 있었고, 각 채널 간의 성능 편차 패턴도 주행 환경에 따라 비슷한 양상을 보이는 것으로 나타났다.

1. 서론

정차한 상태와는 달리 주행 중의 자동차 내부 음 환경은 엔진 소음과 같은 내부 소음원이나 바람 잡음 및 노면 마찰음 등의 외부 소음원에 노출될 수 밖에 없는 것이 현실이다. 이러한 문제로 인하여 자동차 주행 환경에서는 신뢰성 있는 음성 인식 성능을 나타내기 어려우며, 안정적인 음성 인식 시스템의 상용화가 늦어지는 것도 사실이다.

자동차 주행 환경을 비롯한 실제 소음 환경에서의 음성 인식 성능 향상을 위한 접근으로는 음 환경을 개선하는 방법과 멀티 채널을 이용한 잡음 제거, 신호 영역 및 특징 벡터 영역에서의 전처리, 그리고 실제 환경에서의 대규모 어휘를 기반으로 구축한 데이터 베이스의 이용 등을 들 수 있다.

본 논문에서는 SiTEC(Speech Information Technology & Industry Promotion Center)에서 구축한 자동차 환경에서의 모의 음성 DB를 이용하여 주행 환경과 채널(마이크로폰) 위치에 따른 음성 명료도 파라메터를 비교하여 음성 인식률에 미치는 영향과 그 상관성을 검토해 보고자 한다.

2. 데이터 베이스

음성 인식 실험을 위한 데이터 베이스로는 2500cc의 자동차 주행 환경에서 모의 음성 데이터를 수집한 SimulCar01을 이용하였다[1]. 본 DB는 기존에 320명의 화자를 대상으로 1인당 417 단어를 발성하도록 하여 방음 스튜디오 환경에서 구축한 음성 데이터를 HATS (Head and Torso Simulator)의 의사 입을 통해 재생하고 이를 안정된 주행 상태에서 8채널의 마이크로폰으로 재녹음한 것이다. 재녹음은 저속 주행 환경(40~60km /h)과 고속 주행 환경(70~90km/h)으로 나누어 각각 160명분의 데이터를 수집한 것이고, 화자 구성에 있어서는 발성 SET과 성, 연령, 지역에 대하여 균등한 분포를 가지도록 한 것이다. 한편, 배포를 위한 DB의 또 다른 구성으로서 별도 40명의 화자를 이용하여 동일한 방법으로 5가지의 랜덤 주행 환경에서 구축한 평가용 데이터가 있다. SimulCar01을 포함한 기타 자동차 환경에서 구축한 다양한 DB의 사양은 SiTEC 홈페이지에서 자세히 소개하고 있다[2]. 본 논문의 실험은 바운더리 콘덴서 타입의 AKG C400-BL 마이크로 수집된 6채널 중 4채널만을 이용하였다.

3. 주행 환경에 따른 채널 명료도 실험

3.1 Speech Transmission Index

공간 전송계에서 음성의 전달 품질을 예측할 수 있는 방법으로 H.J.M. Steeneken에 의해 제안된 STI(Speech Transmission Index)가 널리 알려져 이용되고 있다[3]. 음성 명료도의 척도로 사용되는 STI는 방사된 변조 신호가 공간을 전반하는 과정에서 실내의 반사음이나 잔향, 잡음에 의해 외견상의 변조도가 열화되어 수신되는 신호를 측정하여 평가할 수 있는 물리 지표이다. 이 파라미터는 MTF(Modulation Transfer Function)에 의해 구해질 수 있는데, MTF는 그림 1과 같이 랜덤 신호를 정현파로 100% 변조하여 전송한 후 변조 신호에 유입되는 잡음 및 잔향음 성분 등이 반영된 수신 신호의 열화 정도를 측정하는 것이다.

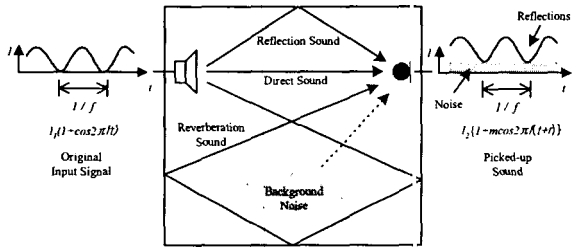


그림 1. MTF에 의한 잡음 및 잔향음장의 품질 평가

측정된 STI 결과는 0~1의 값을 가지며, 일반적으로 대형 공간에서 0.75 이상이면 Excellent하다고 평가하지만, 규모가 작은 자동차 실내에서는 큰 의미를 갖지 못한다. 한편, STI는 V.M.A. Peutz에 의해 제안된 자음 명료도 손실도(%ALcons, Articulation Loss of consonants)와도 잘 대응된다고 알려져 있다.

3.2 임펄스 응답을 이용한 음성 전달 품질 평가

주행 중인 자동차 실내 음장에는 음성 신호를 저해하는 엔진 소리를 비롯한 노면 마찰 소음, 주변 교통 소음 등의 여러 가지 요인이 발생하게 된다. 이러한 소음 요인과 흡음률이 낮은 차 유리 등으로부터의 반사음 요인 등을 고려하여 임펄스 응답에 의한 방법으로 MTF-STI를 측정하였다.

측정은 그림 2에 나타낸바와 같이 4채널의 마이크로폰을 통하여 SIA Soft의 AcousticTools v4[4]에 의해서 이루어졌고, 측정 음원으로는 의사 랜덤 신호인 MLS(Maximum Mean Sequence)를 이용하여 해당 주행 속도 상황에서 HATS의 의사 입을 통하여 방사하였다. 이 때 주행 소음으로 인한 잡음 변조가 클 것으로 판단하여 MLS length=64K, Sampling Rate=44.1kHz, Average=5로 설정하여 1,486ms의 Time Constant 응답을 도출하였다. 그림 3은 고속 주행에서 APL 채널을 통해 얻어진 임펄스 응답을 MLSSA Tool[5]로 구한 것이다.

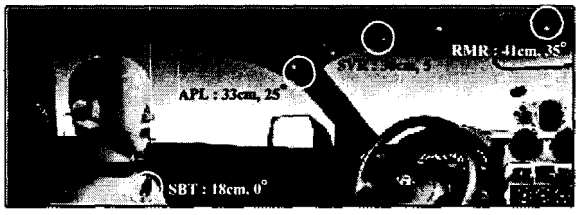


그림 2. 측정 마이크의 위치와 거리 및 각도
(APL: A-Pillar, SVR: Sun-Visor, RMR: Room Mirror, SBT: Safety Belt)

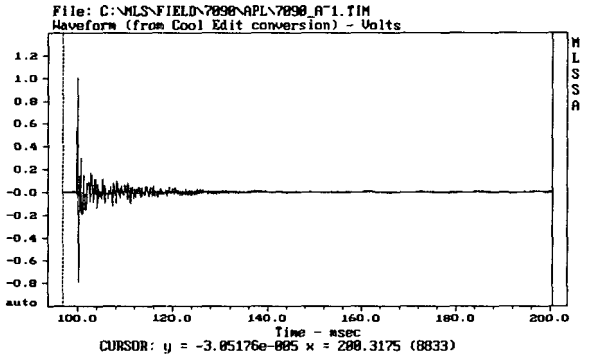


그림 3. 고속 주행 환경의 APL 채널의 임펄스 응답

3.3 주행 환경 및 채널에 따른 음성 명료도

주행 환경에서 STI의 측정을 통해 명료도를 평가하기 위하여 3가지 주행 환경에서의 채널별 임펄스 응답을 구하였다. 얻어진 응답 $h(t)$로부터 전송로의 MTF $m(\omega)$를 나타내는 데에는 식 (1)에 의해 계산되며, 전송로의 방해 잡음을 적용한 식을 (2)에 나타내었다.

$$m(\omega) = \frac{\int_0^\infty h^2(t)e^{-j\omega t}dt}{\int_0^\infty h^2(t)dt} \quad (1)$$

$$m(\omega) = \frac{\int_0^\infty h^2(t)e^{-j\omega t}dt}{\int_0^\infty h^2(t)dt}(1+10^{-(S/N)/10})^{-1} \quad (2)$$

여기에서 $S/N=10log(I_0/I_{noise})$이고, I_0와 I_{noise}는 각각 전송로의 평균 강도와 입력 신호의 평균 강도를 나타낸다. 중심주파수 125Hz-8kHz Octave band(k=1,2,...,7)에 대하여 1/3octave band로 구성된 63-12.5Hz 범위의 변조 주파수 $F_i(i=1,2,...,14)$를 평균하면 98종의 $m_{k,f}$가 얻어지고, 여기에 외견상의 SNR을 정규화하여 STI를 산출한다. 단, Overall STI의 산출은 언어 구조에 의한 고려로서 각 Octave band에 MLSSA 표준 가중 계수 W_k

가 적용되었다. 표 1과 그림 4는 주행 환경에 따른 4 채널의 STI 산출 결과를 나타낸 것이다.

표 1. 주행 환경에 따른 채널별 STI
(Low : 40~60km/h, High : 70~90km/h)

Environments	APL	RMR	SBT	SVR
Stop	0.924	0.953	0.943	0.933
Low	0.723	0.727	0.819	0.777
High	0.730	0.654	0.814	0.757

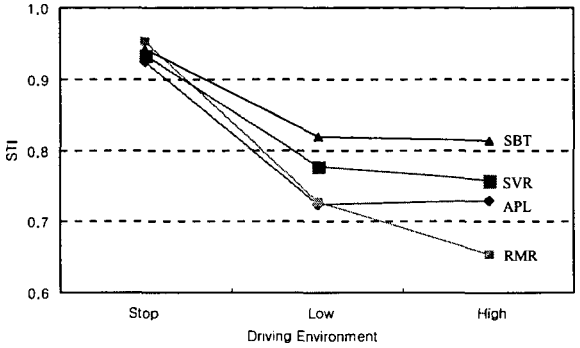


그림 4. 주행 속도별 각 채널의 명료도 변화

STI를 구한 결과 정차 상태에서는 각 채널의 명료도가 비슷한 수치를 기록한 반면, 주행 속도가 빨라질수록 보편적으로 명료도가 나빠지는 것을 알 수 있었다. 그러나 SBT 채널은 주행 환경에서 정차 상태와 큰 편차를 보이지 않을 뿐만 아니라 저속과 고속 환경에서 명료도 값이 거의 일정한 값을 기록하여, 속도가 빨라질수록 명료도가 크게 저하되는 RMR 채널과 대조가 되었다.

3.4 주행 환경에서의 파워 스펙트럼 분석

고속 주행 환경에서 채널간의 명료도 차이를 규명하고자 측정된 시간 도메인의 임펄스 응답을 주파수 도메인으로 변환하여 스펙트럼 분석하였다. 분석은 정상 상태일 때와의 비교를 위하여 정차 환경과 고속 주행 환경으로 선정하였으며, 채널은 대조적 결과를 나타낸 SBT와 RMR 채널 데이터를 사용하였다. FFT는 N_{FFT} =32K(sample), Hanning Window에 의해 이루어졌고, 그림 5에 파워 스펙트럼 분포를 나타내었다.

주행 환경에서 채널간 명료도가 가장 심한 차이를 나타냈던 SBT와 RMR 채널에 대하여 파워 스펙트럼 분석한 결과, 주행시 RMR 채널의 경우 1kHz 이하의 대역에서 높은 파워 분포를 나타내는 것을 볼 수 있었다. 특히 음성 대역의 특징이 잘 반영되는 100~500Hz 구간에서 SBT 채널에 비해 다소 높은 차이를 나타내는 것으로 음성 인식에 있어서 SBT 위치보다 RMR 위치에 유입되는 소음의 영향력이 큰 것으로 판단된다. 한편, APL과 SVR 채널의 경우에는 상호간 차이가 거의 없는 것으로 분석되었다.

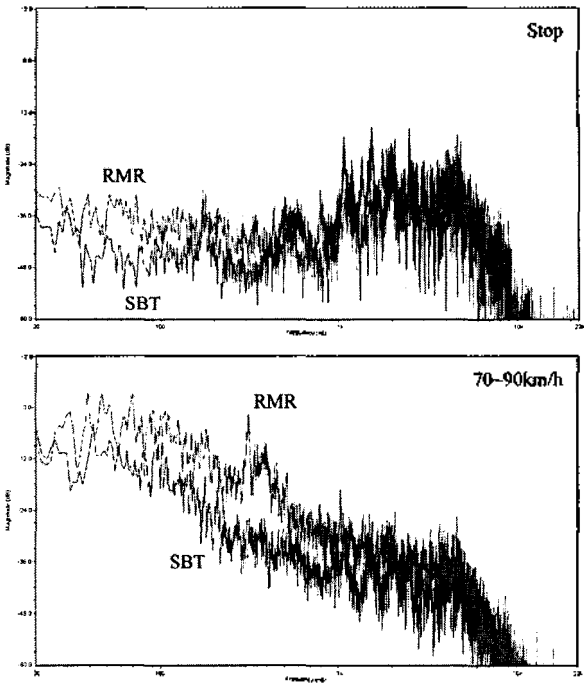


그림 5. 정차와 주행환경에서의 채널 파워 스펙트럼

4. 인식 실험 및 결과

4.1 실험 환경

본 논문의 실험은 주행 환경별, 채널별로 구분된 총 8종의 환경에서 각각 120명의 화자 데이터를 훈련하였고, 훈련에 참여하지 않은 40명의 화자 데이터를 이용하여 인식 실험을 수행하였다.

음향 모델 훈련 및 평가에 사용된 특징 벡터로는 C0를 포함한 MFCC 13차, delta, delta-delta를 이용하여 총 39차 음성특징벡터를 적용하였다. 음향 모델에 따른 자동차 환경별, 채널 별 인식 실험은 트라이폰 단위의 3 state의 left-to- right 방식의 연속 밀도 HMM을 기반으로 하였다.

4.2 실험 결과

실험은 4개의 채널(SBT, APL, SVR, RMR)별로 주행 환경을 저속(Low), 고속(High)으로 구분하여 진행하였다. 인식률 비교를 위해 각각 채널별, 주행 환경별로 독립적인 음향 모델을 작성하여 저속과 고속의 주행 환경에서 마이크의 위치에 따른 인식률의 차이를 비교하였으며 모노폰과 트라이폰의 Mixture 수에 따른 인식률의

변화는 그림 6과 같다.

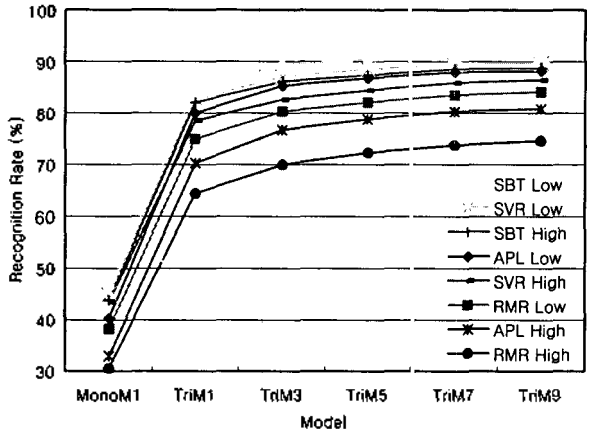
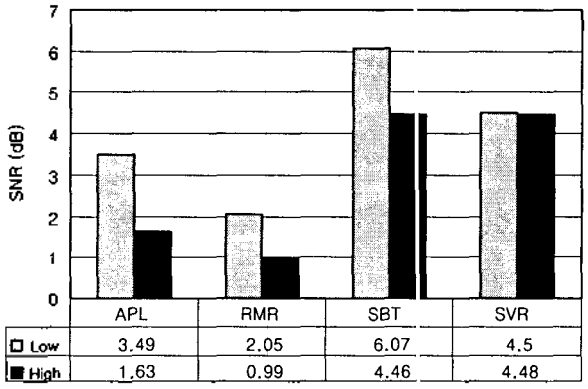


그림 6. 주행 환경과 채널에 따른 인식률

또한 자동차 잡음 환경에서 수집된 음성의 SNR에 따른 인식률의 편차를 비교하기 위해 각각의 주행 환경과 채널에 따른 한 화자 음성 데이터의 평균 SNR 값을 조사하였으며 그 결과는 그림 7과 같다.



	APL	RMR	SBT	SVR
Low	3.49	2.05	6.07	4.5
High	1.63	0.99	4.46	4.48

그림 7. 실험 데이터의 주행 환경 및 채널별 SNR

그림 6은 음향모델과 인식 데이터들 사이의 채널별, 환경별 상황이 일치되었을 때의 결과이다. 그러나 훈련 과정에서 사용된 데이터와 인식과정에서 테스트를 위한 데이터의 불일치(Mismatch)가 발생했을 때 인식률의 저하를 비교하기 위해 음향모델과 인식 데이터의 주행환경이 반대일 경우와 저속과 고속의 데이터를 모두 사용하여 작성된 음향 모델에 따른 인식률의 차이를 표 2에 나타내었다.

5. 고찰 및 결론

표 2. 음향 모델에 따른 환경별, 채널별 인식률(%)

Acoustic Model(TriM1)		Test Environment	
		Low	High
APL	Low	82.17	71.84
	High	76.54	74.55
	Low+High	80.14	73.95
RMR	Low	77.69	65.44
	High	72.89	68.88
	Low+High	75.94	68.40
SBT	Low	86.65	81.34
	High	83.80	84.16
	Low+High	85.43	83.24
SVR	Low	85.14	80.29
	High	83.06	82.52
	Low+High	84.45	81.98

자동차 주행 환경에서 각 채널이 갖는 음향 전송 특성과 인식 실험을 수행하였다. 그 결과 음원으로부터의 거리에 따라 명료도(STI) 및 인식률이 변화함을 알 수 있었고, 변화 패턴 측면에서 두 척도가 높은 상관성이 얻어짐을 확인할 수 있었다.

명료도와 인식률 실험에서는 저속과 고속에서 모두 SBT - SVR - APL - RMR 순으로 높은 성능을 나타내었고, 특히 SBT와 SVR 채널에서의 인식률은 그 차이가 미미하며, 단일 채널 환경 하에서 유리한 위치임을 확인하였다. 또한 모든 채널에서 단순히 거리 요인에 의해서만 명료도와 인식률이 결정되는 것이 아니라 속도 증가에 따라 각 채널이 갖는 100~500Hz 대역의 주행 소음 역시 크게 작용함을 알 수 있었다. 단일 채널에서 저속 환경을 고속으로 속도를 높였을 때, SBT 채널의 경우 STI는 0.005, SNR은 0.02dB, 인식률에서도 2.5% 정도의 최소 하락폭을 나타내어 각 파라미터에서 0.07, 1.06dB, 10%의 큰 하락폭을 갖는 RMR 채널과 대조를 이루었다. 또한 음향 모델에 따른 인식률 실험에서는 저속과 고속 모두 일치(Match)된 음향 모델로 테스트할 경우에만 높은 인식률이 나타나는 것으로 분석되었다.

참 고 문 헌

1. 이광현, 김봉완, 이용주, "자동차 환경에서의 노이즈 DB 및 한국어 음성 DB 구축", 대한음성학회 말소리 제48호, 2003. 12

2. http://www.sitec.or.kr

3. Don Davis&Carolyn Davis, "Sound System Engineering", Howard W. Sams & Co., 1989

4. SIA-Smaart AcousticTools v.4 User Guide, SIA Software, 2001

5. Douglas D.Rife, "MLSSA Reference Manual", DRA Lab., 1987