반향제거기를 갖는 자동차 실내 환경에서의 음성인식[1)]

박철호* 허원철* 배건성**
* 경북대학교 대학원 전자공학과 석사과정
** 경북대학교 전자전기공학부 교수

Robust speech recognition in car environment with echo canceller

Chul Ho Park*, Won Chul Heo*, Keun Sung Bae**
Department of Electronics, Kyungpook National University

chilo@mir.knu.ac.kr

Abstract

The performance of speech recognition in car environment is severely degraded when there is music or news coming from a radio or a CD player. Since reference signals are available from the audio unit in the car, it is possible to remove them with an adaptive filter. In this paper, we present experimental results of speech recognition in car environment using the echo canceller. For this, we generate test speech signals by adding music or news to the car noisy speech from Aurora2 DB. The HTK-based continuous HMM system is constructed for a recognition system. In addition, the MMSE-STSA method is used to the output of the echo canceller to remove the residual noise more.

Ⅰ. 서론

오늘날 정보통신기술이 발전함으로써 차량 운전자는 GPS (Global Positioning System)를 통한 도로안내 및 관련 서비스, 디지털 방송을 통한 교통정보 등과 같은 다양한 형태의 서비스를 텔레매틱스(telematics) 시스템을 통해 제공 받을 수 있다[1]. 사용자와 차량용 정보시스템 간의 안전한 접속 기술 측면에서 볼 때, 기기의 조작을 음성으로 하고 서비스도 음성으로 받을 수 있도록 하는 게 바람직하다. 따라서 자동차 환경에서의 강인한 음성인식을 위한 기술 개발이 꾸준히 진행되고 있다. 그러나 대부분의 자동차 환경을 고려한 음성인식 연구는 주로 도로 주행 시에 자동차 실내 환경에서 발생되는 잡음을 대상으로 전처리 과정을 통한 잡음제거 또는 자동차 잡음에 강인한 특징파라미터 추출 등에 주안점을 두고 있다[2,3]. 그러나 실제 자동차 환경에서는 주행 중에 발생하는 배경잡음뿐만 아니라 자동차 오디오 시스템의 출력신호가 실내에서 여러 경로로 반사되어 마이크로 함께 입력됨으로써 인식성능을 크게 저하시키게 된다. 따라서 텔레매틱스 시스템의 음성인터페이스로 음성인식 기술을 이용하고자 할 때 자동차 실내 환경에서의 엔진 및 도로주행 잡음과 더불어 오디오 출력신호에 의한 반향잡음을 제거해주어야 한다.

본 논문에서는 자동차 환경에서의 텔레매틱스 시스템을 위한 음성인식 전처리 단으로서, 반향제거기와 barge-in 기능을 갖는 음성인터페이스[4]를 이용하여 자동차 잡음음성에 대한 인식실험을 수행하였다. 반향이 제거된 신호에 잔존하는 잡음성분을 줄이기 위해서 MMSE-STSA(Minimum Mean Square Error-Short Time Spectral Amplitude) 기반의 음성개선 기법을 후처리 과정으로 적용하였다. 기본 (baseline) 음성인식기로는 Aurora2-HTK를 사용하였으며, 실험 데이터로는 ETSI (European Telecommunications Standards Institute)에서 배포한 Aurora2 DB에서 인식기의 훈련은 clean condition으로 수행하였고, 테스트 집합 A의 자동차 잡음(car noise)에 배경음악 및 뉴스를 첨가하여 인식실험을 수행하였다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 본 연구에서 사용한 반향제거기 및 음성개선 과정을 포함

1) 본 연구는 한국과학재단의 핵심기초연구 (R01-2003-000-10242-0) 지원으로 수행되었습니다.

하는 음성인터페이스와 실험조건에 대해 간략하게 서술한다. 3장에서는 인식실험 방법 및 실험결과를 제시하고, 마지막으로 4장에서 결론을 맺는다.

Ⅱ. 반향제거기를 갖는 음성인터페이스

자동차 환경에서의 텔레매틱스 시스템을 고려한 음성인터페이스의 블록도는 그림 1과 같다. 반향제거기는 구조가 간단하고 계산량이 적은 NLMS(Normalized Least Mean Square) 알고리즘을 갖는 적응필터를 이용하여 구현하였으며, 동시통화(DT: Double-Talk) 검출기를 사용하여 운전자의 음성신호의 존재 여부를 판단함으로써 오디오 신호 또는 음성안내 메시지 등의 출력을 줄여주는 barge-in 기능을 포함하고 있다. 특히, 보조 적응필터를 사용하여 DT 검출 성능을 향상시키고, 배경잡음의 크기에 따른 적응필터 갱신 구간을 설정해 줌으로써 적응필터의 안정성을 향상시키고자 하였다[4].

반향제거기를 거친 후 반향이 제거된 음성신호에서 남아있는 잡음성분을 줄이기 위해 후처리 과정으로 MMSE-STSA 기반의 음성개선 기법을 적용하였다. 이 기법은 Y. Ephraim이 제안한 잡음제거 기법으로, 음성과 잡음의 스펙트럼이 통계적으로 서로 독립적인 가우시안(Gaussian) 분포를 가진다고 가정하고, 각 스펙트럼 빈(bin)에 대해 추정된 신호 및 잡음 전력을 이용하여 이득함수를 구하여 곱해줌으로써 잡음성분을 줄여준다[5]. 따라서 이득함수의 추정 정확도에 따라 음성개선 성능이 달라지는데, 이득함수는 사전 신호대잡음비(a priori SNR)와 사후 신호대잡음비(a posteriori SNR)의 함수이기 때문에 정확한 사전 신호대잡음비와 사후 신호대잡음비를 구할 필요가 있다. 사전 신호대잡음비와 사후 신호대잡음비를 구하기 위해서는 각 스펙트럼 빈에 대한 음성신호 및 잡음신호의 전력을 추정해야 하는데, 음성이 존재하는 구간에서는 잡음전력을 추정하기가 어려우므로 각 프레임마다 문턱치 기반의 VAD(Voice Activity Detection)를 사용하여 음성의 존재 유무를 결정하여 묵음 구간에서만 잡음전력을 추정하는 단순한 방법이 많이 사용되는데 이를 hard-decision 방법이라고 한다. 본 논문에서 MMSE-STSA 기법에서 신호전력 추정에 사용되는 파라미터 α 값은 0.90으로, 잡음전력 추정에 사용되는 파라미터 β 값은 0.4로 설정하였다[5].

자동차 오디오 출력에 의한 음악 및 음성을 포함하는 반향 신호를 생성하기 위하여 그림 2와 같이 가로×세로×높이를 각각 1.5 × 2.0 × 1m로 자동차 실내 환경을 가정하고 image method[6]를 사용하여 룸 임펄스 응답을 구하였다. 그림 2에서와 같이 마이크 위치, 4개의 실내 스피커, 차량 실내 환경을 고려하여 50ms의 잔향시간을 설정하고 마이크로폰은 등방성 빔 패턴을 가지며 한쪽 면에 붙어 있어 후면에서 반사되는 반사파는 없다고 가정하였다. 음악 및 음성을 포함하는 오디오 출력 신호에 위에서 구한 룸 임펄스 응답을 적용하여 반향음을 생성하고, 인식에 사용될 자동차 잡음음성에 반향음을 더하여 인식실험에 사용될 테스트 데이터를 생성하였다.

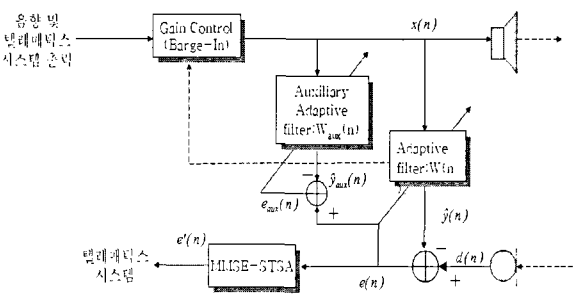

그림 1 음성인터페이스의 블록 다이어그램

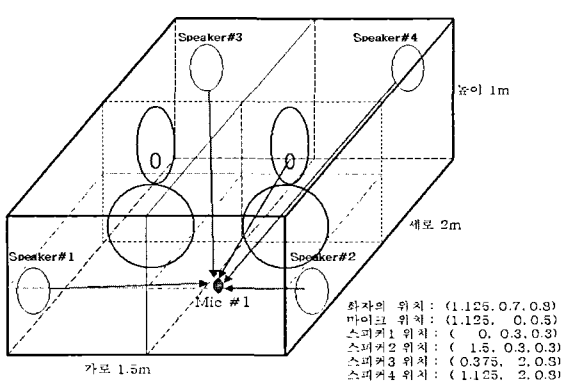

그림 2 자동차 실내의 시뮬레이션 환경

Ⅲ. 실험 및 검토

반향음을 생성하기 위한 오디오 출력은 약 30분 정도의 한국 음악을 사용하였으며, 샘플링주파수 8kHz에, 16bit로 양자화 된 총 14,312,466 샘플 수를 가진다. 인식실험에 사용되는 테스트 데이터는 반향음의 전력이 테스트 잡음음성 전력의 1/2, 1.0이 되는 2가지 경우에 대해서 반향음과 잡음음성을 더해주어 생성하였다. 그림 3(a)는 반향음이 포함된 상태에서의 잡음음성에 대한 인식실험 경우를 나타내며, 3(b)는 반향음이 포함된 잡음음성을 음성인터페이스를 통과시켜 반향음을 제거한 후의 인식실험 경우를 보인 것이다.

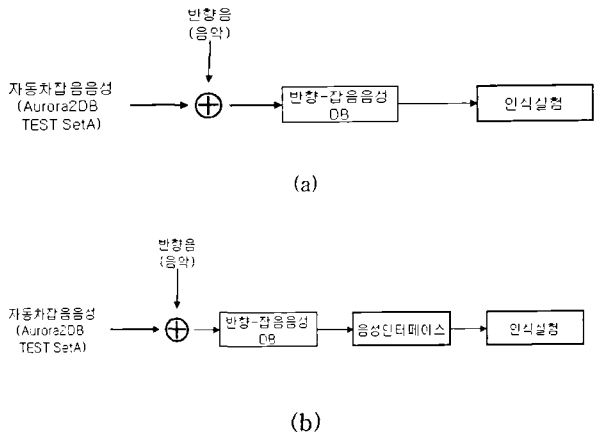

그림 3 (a) 반향음이 포함된 상태에서의 잡음음성 인식실험 경우
(b) 음성인터페이스로 반향음을 제거한 후의 인식실험 경우

인식실험은 Aurora2 DB에서 테스트 집합 A의 자동차 잡음에 대해, clean 음성을 포함하여 20dB부터 -5dB까지 7가지 잡음레벨에 대한 수행하였다. 기본(baseline) 음성인식기는 Aurora2-HTK를 사용하였는데, 이것은 CUED-HTK 음성인식시스템의 단어모델과 훈련절차를 따른 것으로 Aurora2 DB에 적합하도록 구성된 것이다. 단어모델은 one, two, three, four, five, six, seven, eight, nine, zero, oh의 11개로 정의되어 있고, 각 단어모델은 3 mixture, 16 state를 갖는 CHMM(Continuous Hideden Markov Model)으로 구성된다. 인식시스템에는 11개의 단어모델 외에 2개의 묵음 모델이 포함되어 있는데, 각각 3 state와 1 state CHMM으로 구성되어 있다. 특징파라미터는 23차 필터뱅크를 이용한 12차 MFCC와 1차 로그에너지, 그리고 각각의 delta 및 acceleration 을 포함한 총 39차로 구성된다. 분석 프레임의 크기는 25ms 이며, 10ms씩 이동시키면서 특징파라미터를 추출하였다.

인식실험의 성능 평가를 위한 측정치로 문장인식률(sentence correction), 단어인식률(word correction), 단어정확률(WA: Word Accuracy) 등이 있다. 일반적으로 음성인식의 성능 측정치로 단어인식률 보다는 삽입에러를 인식률에 포함시킨 단어정확률을 더 많이 사용하므로, 본 논문에서는 성능 비교를 위한 인식률로 단어정확률을 사용하였다. 단어정확률은 식(1)과 같이 정의 된다

$$WA = \frac{H-I}{N} \times 100 \qquad (1a)$$

$$H = N - D - S \qquad (1b)$$

여기서 H는 올바르게 인식된 단어 수, I는 삽입된 단어 수, D는 삭제된 단어 수, S는 대치된 단어 수, N은 전체 테스트 단어 수를 의미한다.

그림 4(a)는 SNR 10dB의 자동차 잡음음성에 같은 크기의 전력을 갖는 반향음을 더한 신호의 예를 보인 것인데, 4(b)는 반향제거기를 갖는 음성인터페이스를 이용하여 반향음을 제거한 결과를 보인 것이며, 4(c)는 반향음이 제거된 자동차 잡음음성에서 MMSE-STSA 음성개선 기법을 적용하여 추가로 잡음제거를 한 결과이다. 음성인식 실험은 반향음이 포함되지 않은 자동차 잡음음성에 대한 경우를 포함하여 4(a), 4(b), 4(c)의 세 가지 경우의 테스트 음성에 대하여 수행하였는데, 인식결과가 표 1에서 4에 주어져 있다. 표 1은 반향음이 포함되지 않은 경우의 인식률을 나타내며, 표 2에서 4는 반향음이 포함된 경우의 인식률인데, case A는 반향음의 전력이 잡음음성의 전력과 동일한 경우를 나타내고, case B는 반향음의 전력이 잡음음성 전력의 1/2이 되는 경우를 의미한다.

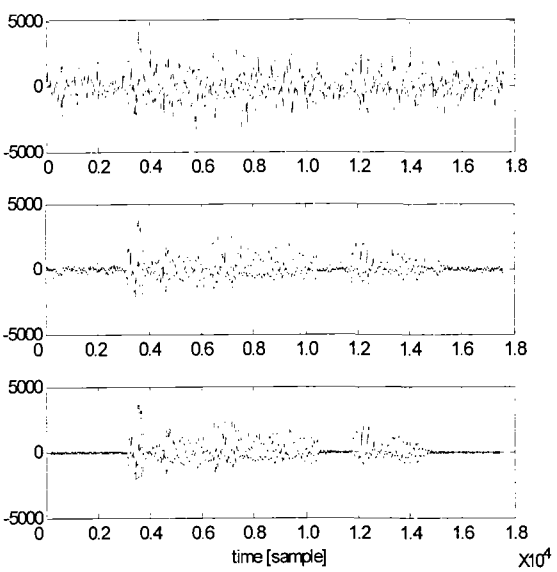

그림 4 (a) SNR 10dB의 잡음음성에 반향음이 부가된 음성신호
(b) 음성인터페이스를 이용하여 반향음이 제거된 음성신호
(c) 반향음 제거후 MMSE-STSA 음성개선 적용한 음성신호

표 1을 보면, SNR 15dB 까지는 90% 이상의 인식률이 얻어짐을 볼 수 있다. 그러나 표 2에서 반향음이 포함된 잡음음성에 대한 인식률은 clean 음성이나 SNR이 높은 경우의 음성에서도 전반적으로 40% 이하의 낮은 인식률을 보이며, 반향음이 첨가되지 않은 경우에 비해 거의 제대로 인식하지 못함을 알 수 있다. 표 3은 전처리 과정으로 반향제거기를 사용하여 반향음을 제거한 후의 인식 결과이다. 표 2의 경우와 비교하여 clean 음

성이나 SNR이 20dB 인 경우에는 인식률이 상당히 향상되었음을 볼 수 있다. 그러나 SNR이 15dB 이하에서는 반향음이 없는 경우에 비해 20~30% 정도 인식률이 저하되었다. 표 4는 반향음 제거한 후, 후처리 과정으로 MMSE-STSA 기법을 사용하여 잡음을 제거한 후의 음성에 대한 인식결과이다. SNR이 20dB 이상에서는 인식률이 약간 감소하였지만 10dB 내외에서는 인식률 향상을 가져오며 전체 평균인식률이 38.80%에서 53.32%로 향상되었음을 볼 수 있다.

Ⅳ. 결론

본 논문에서는 자동차 환경에서의 텔레매틱스 시스템을 위한 음성인식 전처리 단으로서, 반향제거기와 barge-in 기능을 가지며 음성인터페이스를 이용하여 오디오 출력신호가 포함된 자동차 잡음음성에 대한 인식실험을 하였다. 음성인터페이스로 반향음을 제거할 경우 SNR 10dB 이상에서 인식률이 많이 향상되었으며, 후처리 과정으로 MMST-STSA 기반의 잡음제거를 포함할 경우 전체적으로 높은 인식률 향상을 얻을 수 있었다.

테스트 데이터 생성을 위해 반향음을 첨가할 때 잡음음성 전체의 평균전력 크기에 비례하여 반향음을 더하였으므로, 잡음음성 신호의 크기가 작은 경우에는 상대적으로 반향음이 너무 커서 음성인터페이스가 반향음을 적절히 제거하지 못하는 경우가 많았다. 차후이러한 문제점과 실제 환경에서의 반향제거 성능을 향상시킬 수 있는 방법에 대한 연구를 수행할 계획이다.

표 1. 반향음이 포함되지 않은 잡음음성에 대한 인식결과(%)

clean	20dB	15dB	10dB	5dB	0dB	-5dB	Avg
98.96	97.41	90.04	67.01	34.09	14.46	9.39	60.60

표 2. 반향음이 포함된 잡음음성에 대한 인식결과(%)

반향음	case A	case B	Average
Clean	35.88	43.24	36.30
20 dB	36.86	43.81	36.87
15 dB	33.16	39.58	33.37
10 dB	25.71	30.54	26.08
5 dB	15.57	18.25	15.60
0 dB	9.48	10.50	9.66
-5dB	8.23	8.26	8.11
Average	24.16	28.54	24.32

표 3. 반향음을 제거한 후의 잡음음성에 대한 인식결과(%)

반향음	case A	case B	Average
Clean	92.10	94.12	91.96
20 dB	90.75	92.01	73.85
15 dB	48.14	60.10	50.24
10 dB	46.41	47.81	42.17
5 dB	23.98	7.19	17.02
0 dB	10.95	11.06	10.74
-5dB	8.68	8.44	8.81
Average	44.05	43.63	38.80

표 4. 반향음 제거 후 음성개선 한 음성에 대한 인식결과(%)
(MMSE-STSA, $\alpha = 0.90$, $\beta = 0.4$)

반향음	case A	case B	Average
Clean	87.83	91.17	87.70
20 dB	88.49	89.92	74.87
15 dB	59.56	72.03	61.58
10 dB	67.28	70.00	64.80
5 dB	49.96	53.86	49.24
0 dB	20.19	22.31	16.12
-5dB	7.25	7.78	7.94
Average	57.10	61.62	53.32

참고문헌

[1] Y, Zhao, "Telematics: safe and fun driving," *Intelligent Systems, IEEE Expert*, vol. 17 , no. 1 . pp.10-14, Jan.-Feb 2002 .

[2] T. Shinde, K. Takeda, F. Itakura, "Multiple regression of log spectra for in-car speech recognition," pp.797-800, ICSLP 2002.

[3] M. Matassoni, M. Omologo, A. Santarelli, P. Svaizer, "On the joint use of noise reduction and MLLR adaptation for in-car hands-free speech recognition," ICASSP 2002.

[4] 김준, 텔레매틱스 시스템을 위한 반향제거 및 Barge-in 기능을 갖는 음성인터페이스, 경북대학교 석사학위 논문, 2004.

[5] 박철호, "MMSE-STSA 기반의 음성개선 기법을 이용한 잡음음성의 인식성능 분석", 신호처리합동학술대회, 제18권, 1호, pp.2, 2005년 10월.

[6] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics", J. Acoustic. Soc. Am., Vol. 65, No.4, pp.943-950, 1979.