차량환경에서 DMB용 음성명령어기 사용을 위한 음성개선방법

백승권*, 한민수* 남승현**, 강경옥***,
* 한국정보통신대학교 공학부 멀티미디어 그룹
** 배재대학교 전자공학과
*** 한국전자통신연구원 무선방송 연구부
skbeack@icu.ac.kr

Speech Enhancement for DMB Voice commander in Car environment

Seung Kwon Beack, Minsoo Hahn, Seung Hyon Nam, Kyung Ook Kang
Multimedia Group, Information and Communications University
Electronic Engineering, PAICHAI
Radio & Broadcasting Research Laboratory ETRI

요약

본 논문에서는 차량용 음성명령어기의 사용을 위한 전처리 과정으로 음성개선 방법을 다룬다. 특히 DMB 사용환경에서 보다 주위 소음에 자유롭고 단말 조작에 있어 안정성을 보장하기 위하여 일반적 단일 마이크로 폰으로 처리되는 잡음뿐만 아니라 음성명령어를 제외한 오디오 신호 등 비정적 통계적 특성을 갖는 소음들도 제거 될 수 있도록 음성개선 방법을 제안한다. 우리는 2개의 마이크로폰을 가지고 BSS 알고리즘을 적용하여 비정적 신호들을 분리하고, 분리된 신호에 대하여 Kalman Filter를 이용하여 시간상 단구간 정적 잡음을 제거한다. 본 논문의 인식 실험 결과를 통하여 공간적, 시간적 음성개선 방법이 순차적으로 적용될 때, 실제 차량 환경에서 음성 개선 알고리즘으로 적용될 수 있음을 보였다.

1. 서론

향후 수년 내에 차량에서의 DMB(Digital Multimedia Broadcasting) 서비스가 가능할 것이다. 이때 주행중인 운전자의 안전과 편의를 보장하기 위하여 DMB단말 서비스 요청수단으로 음성명령어기 사용이 요구된다. 일반적으로 차량 내에서 신뢰도 높은 음성명령어기 사용을 위하여 음성개선 방법이 전처리 과정으로 요구된다. DMB용 음성명령어기에서 잡음은 주행으로부터 야기되는 잡음뿐만 아니라, 동승자나 라디오로부터 발생하는 음성명령어가 아닌 음성 및 오디오신호도 잡음으로 고려되어야 한다. 전자의 경우 시간상 음성과 상호 상관관계가 적은 관계로 단일 마이크로폰을 이용한 제거가 가능하나 후자는 음성명령어와 시간적 상관관계가 무시될 수 없으므로 단일 마이크로폰을 이용한 제거가 용이하지 않다. 우리는 2개의 마이크로폰을 이용한 제거 방법을 고려한다. 먼저 음성과 시간적 상관성이 큰 신호는 2개의 마이크로폰으로부터 공간적 독립성을 이용하여 분리한 후, 이차적으로 음성신호로 기대되는 채널에 단일마이크로폰 잡음제거 기술을 적용한다. 전자의 공간상 신호분리 방법으로는 Lucas parra의 BSS(Blind Source Separation)방법을 적용하고[1] 후자의 시간 축 상의 잡음제거방법은 Kalman Filter를 이용하였다.[2] 실험에 사용된 데이터는 실제환경에서 녹취한 차량 잡음과 음악 및 음성 신호를 인식 대상 음성신호와 인위적으로 혼합하여 열화시켰다. 잡음의 공간상 혼합은 실제환경을 반영하기 위해 복적분 혼합회로(convolutive mixture)로 모델을 사용하였다.

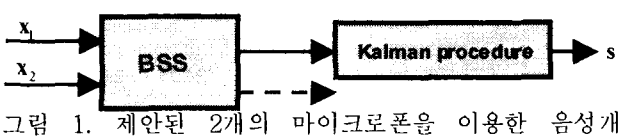

그림 1. 제안된 2개의 마이크로폰을 이용한 음성개선 방법[$\mathbf{x}_i$: 열화된 음성 신호, $\mathbf{s}$: 복원된 원 음성신호]

우리는 각각의 알고리즘이 열화된 음성에 적용 전/후 음성인식 성능을 측정하여 알고리즘의 신뢰도를 측정하려 한다. 본 논문의 구성은 시간적 잡음제거 기술과 공간적 잡음 제거 알고리즘에 대하여 간략히 설명하고 각각의 인식성능 및 제안된 전처리 과정에 의한 인식성능을 측정하고 이를 비교 분석한다.

2. 시간성 잡음 제거 알고리즘

우리는 시간성 잡음을 시간 샘플상에서 음성신호 $s(t)$ 와 상관관계가 적은 잡음신호 $w(t)$ 로 정의한다. 이는 음성신호 $s(t)$ 에 대한 자기상관함수를 $R_s(\tau)=E\{s(t)s(t+\tau)\}$ 라 할 때, 잡음에 의해 열화된 $\hat{s}(t)$ ($s'(t)=s(t)+n(t)$)의 자기 상관함수 $R_{\hat{s}}(\tau)$ 는 $R_{\hat{s}}(\tau)\cong R_s(\tau)$ 로써 표현 가능하다. 만일에 시간상에서 $w(t)$ 가 백색화(whitening) 특성을 지닌다면, 이론적으로 $R_w(\tau)=R_s(\tau)$ 이 된다. 그러므로 시간성 잡음은 백색화가 가능할 때 시간 축 상에서 성공적으로 제거가 가능하다. 이러한 백색화가 가능하기 위해서 시간성 잡음은 분석구간 내에서 정적인 통계적 특성(stationary statistic characteristic)을 지녀야 한다. 이러한 조건하에서 시간성 잡음은 단일 마이크로폰으로 입력된 열화음성으로부터 제거가 가능하다.

차량에서 순수하게 발생하는 잡음의 대부분이 단구간 내에서 정적인 통계적 특성을 갖는다. 이는 차량 자체내에서 발생하는 엔진 소음이나 진동, 에어컨 소음등과 유체의 흐름에 의한 소음-바람과 창문, 지면과 타이어-등이 대표적이라 할 수 있다. 물론 이것이 완벽한 백색화를 통하여 음성으로부터의 제거가 가능한 것은 아닐지라도, 음성보다 장구간에서 정적인 통계특성을 유지한다는 것으로부터 훌륭하게 제거 될 수 있다. 우리는 차량내에서 성공적인 시간성 잡음 제거를 위하여 단구간(200msec) 잡음정보를 기반으로 백색화를 통한

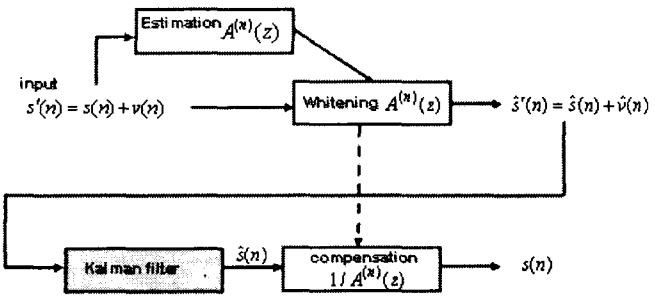

그림 2. Kalman Filter를 이용한 시간성 잡음 제거 알고리즘 구성도

Kalman Filter를 적용한다[2]. Kalman Filter는 백색 가산성 잡음을 효과적으로 억제할 수 있는 알고리즘이다. 그림 2.는 본 논문에서 적용한 Kalman Filter를 이용한 시간성 잡음 제거 알고리즘이다. 모든 신호는 디지털화된 것으로 가정한다. 먼저 열화된 음성신호 $s'(n)$ 은 음성신호 $s(n)$ 이 존재하지 않는 무발성 구간(초기 200msec)으로부터 예측된 $v(n)$ 에 대한 AR(Auto-Regression) 모델 $A^{(n)}(z)$ 에 의하여 백색화된 신호 $\hat{s}'(n)$ 를 생성한다. $\hat{s}'(n)$ 에 포함된 $\hat{v}(n)$ 는 가산성 백색잡음으로 가정할 수 있으며 이는 Kalman Filter를 통해 제거 된다. Kalman Filter의 출력신호 $\hat{s}(n)$ 은 $A^{(n)}(z)$ 에 의한 스펙트럼 왜곡이 있으므로 $1/A^{(n)}(z)$ 에 의하여 이를 보상해 줌으로써 $s(n)$ 을 최종적으로 출력한다.

차량의 시간성 잡음중에서 엔진의 진동 소음등과 같이 장구간 안정된 소음이 존재한다. 이러한 소음은 실제 백색화도 용이하지만 High Pass Filter(HPF)의 적용으로 쉽게 제거될 수 있다. 이는 이러한 소음의 스펙트럼 구조가 음성의 스펙트럼 구조와 중복되지 않기 때문에 가능하다 우리는 GSM-ERF의 2차 HPF $H(z)$ 을 Kalman filter 의 전처리 과정으로써 적용한다.

$$H(z) = \frac{0.9273 - 1.8545z^{-1} + 0.9273z^{-2}}{1 - 1.9059z^{-1} + 0.9114z^{-2}} \qquad (1)$$

3. 공간성 잡음 제거 기술

차량에서 발생하는 소음은 앞서 정의한 시간성 잡음 외에 음성과 상관관계가 큰 소음들이 있다. 실제적으로 차량용 음성명령어기가 올바르게 동작하기 위해서는 음성 명령어를 제외한 모든 신호를 소음으로 정의할 수 있다. 예를 들어 차량 오디오에서 발생하는 음악 또는 음성 신호나 다른 동승자에 의한 음성은 음성명령어기의 잡음이다. 이러한 소음은 시간에 따라 그 특성이 변화되며 음성과의 상관관계가 존재하므로, 즉 $R_s(\tau) \neq R_{\hat{s}}(\tau)$ 이므로 시간성 잡음 기술로써 제거 되기 힘들다. 그러므로 이러한 소음들은 단일 마이크폰을 이용하여 제거하려 할 때, 음성의 스펙트럼 구조를 파괴시킴으로써 음성인식이 성공적으로 수행될 수 없다. 따라서 2개 이상의 마이크로폰을 이용한 제거 기술이 필요하다. 우리는 이러한 소음들을 공간상에서 독립적 위치에서 발생하는 소음으로 간주하고 공간성 잡음으로 정의한다. 이는 공간상에서 발생하여 혼합된 신호는 공간적으로 상관관계가 없음을 의미한다. 이러한 가정을 전제로, 공간 잡음은 공간 잡음 제거 기술로써 제거될 수 있다. 공간 Filtering 기술로써 크게 Beamforming과 BSS(Blind Source Separation) 기술이 있다. 본 논문에서는 BSS 기술을 공간잡음 제거 기술로써 사용 가능성을 검증하고자 한다.

3.1 BSS

BSS 기술의 최종 목표는 혼합된 신호로부터 원신호를 추출하는 것으로 혼합환경 및 원신호에 대한 정보 없이 혼합신호로부터 원신호를 추출해야 하므로 Blind라는 수식어를 동반한다. 이러한 상황에서 BSS 는 "상호 신호들은 서로 독립이다" 라는 사실을 이용하여 문제의 해에 접근한다. BSS의 해는 혼합회로를 추정하여 분리회로를 구하는 것이고 분리회로를 통해 원신호를 복원하는 것을 목적으로 한다. 분리회로를 예측하는 데 있어서 크게 세가지 모델이 있다. 첫째로 순시적 혼합회로에 대한 분리회로로, 하나의 혼합신호는 각각의 원신호가 임의의 상수값으로 확장 축소되어 더해져서 만들어 진다. 둘째로 지연 순시적 혼합회로에 대한 분리회로로써, 혼합신호는 임의의 상수값으로 확장 축소되어 혼합되지만 일정한 지연뒤에 혼합이 이루어진다. 셋째로 복적분 혼합회로에 대한 분리회로로, 하나의 혼합신호는 각각의 원신호가 일정시간의 응답을 갖는 임의의 시스템을 통과하여 합해지는 것을 말한다.

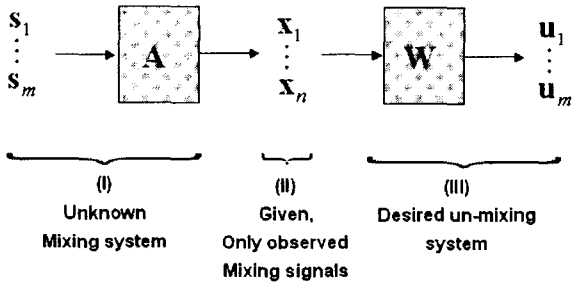

그림 3. BSS의 Global system [$\mathbf{A}$: 혼합회로, $\mathbf{W}$: 분리회로, $\mathbf{s}_m$: 원신호 벡터열, $\mathbf{x}_n$: 혼합신호 벡터열, $\mathbf{u}_m$: 분리신호 벡터열]

일반적으로 실제환경에서의 BSS는 복적분 혼합회로를 갖는다[1][3]. 이는 혼합신호 $\mathbf{x}$ 가 $\mathbf{A}*\mathbf{s}^{\mathrm{T}}$ 로부터 발생하는 것을 의미 하며, $\mathbf{A}$ 를 시간 lag p 를 가지는 혼합회로 $\mathbf{A}_p$ 로 표현할때, $\Re^{m\times n}$ 의 공간차원을 가지는 $\mathbf{A}_p = (A_{ji,p})_{mn}$ 으로 정의한다. 이에 의한 m 개의 혼합신호 $\mathbf{x}(k) = [x_1(k)\cdots x_m(k)]^T$ 는,

$$x_j(k) = \sum_{i=1}^{n}\sum_{p=-\infty}^{\infty} A_{ji,p}s_i(k-p)$$

따라서 분리회로 $\mathbf{W}$ 를 통해 분리신호 $\mathbf{u}(k) = [u_1(k)\cdots u_j(k)\cdots u_m(k)]^T$ 를 얻을 수 있다.

$$u_j(k) = \sum_{j=1}^{m}\sum_{p=-\infty}^{\infty} w_{ij,p}x_j(k-p) \qquad (2)$$

복적분 분리회로를 구하기 위해서, 각 원 신호에 대하여 공간적으로 독립하다는 가정뿐만 아니라, 시간적으로도 독립해야 한다는 가정이 고려될 때 분리회로의 해를 구할

수 있다. 그러나 복적분 회로의 역 모델은 장시간 시간응답 길이를 갖는 FIR(Finite Impulse Response) 시스템을 요구하므로 수렴을 위해 많은 훈련을 필요하며, 특히 시간적으로 독립하다는 가정으로 인하여 분리신호는 백색화 현상을 초래한다.

본 논문에서는 Lucas parra의 BSS알고리즘을 적용한다[1]. 이는 음성 및 오디오 신호가 시간적으로 비정적인 통계적 특성을 이용하여 분리회로의 해를 구함으로써, 실제환경에서의 역 분리회로 출력에 대한 백색화 현상에 대하여 좀더 자유롭다. Lucas의 분리회로는 최소위상 FIR(Finite Impulse Response) 필터로써만 훈련이 가능하는 단점이 있다. 만일에 $\mathbf{A}$ 가 비최소위상 FIR일 경우, $\mathbf{W}$ 에 대하여 비순시적(anti-causal) 부분을 야기 시킨다. 그러므로 순시적 또는 최소위상 FIR $\mathbf{W}$ 는 $\mathbf{A}$ 에 대한 올바른 역 시스템이 아님이 분명하다. 실제환경에서 $\mathbf{A}$ 는 일반적으로 비최소위상 시스템이다. 그럼에도 불구하고 우리는 Lucas의 최소위상 FIR $\mathbf{W}$ 를 적용하려 한다. 이는 차량 공간에서의 $\mathbf{A}$ 는 비최소위상이라 할지라도 협소한 공간으로 인한 그 반향시간이 짧다. 그러므로 순시적 FIR 모델로써 $\mathbf{W}$ 를 근사화 하여도 성공적인 분리를 기대할 수 있다. 그러므로 비순시적 부분을 훈련하기 위해 소요되는 연산을 제거할 수 있다. [1]로부터 시간성 잡음에 의한 왜곡을 고려하지 않을 때, 손실함수는 다음과 같이 정의한다.

$$E(\omega,k') = \mathbf{W}(\omega)[\bar{R}_x(\omega,k')]\mathbf{W}^{H}(\omega) - \Lambda_x(\omega,k') \qquad (3)$$

여기서 ω 는 블록단위로 DFT를 수행한 것을 나타내며, k' 는 전체 K 블록의 블록 인덱스이다. Lucas는 2차 통계적 특성을 이용하여 공간상의 자기상관함수 $\bar{R}_x(\omega,k')$ 에 대한 off-diagonal 항이 zero가 될 때 분리회로의 해를 가지며 해를 구하기 위한 방정식의 수를 K 개로 한것이다. 이때 분리회로의 계수는 다음과 같이 정의한다.

$$\hat{\mathbf{W}}, \Lambda_x = \underset{\substack{\mathbf{W},\Lambda_x \\ W^{H}(\tau)=0,\ \tau>Q\le T \\ W^{H}_{11}(\omega)=1}}{\arg\min} \sum_{\omega=1}^{T}\sum_{k=1}^{K} \|E(\omega,k)\|^2 \qquad (4)$$

Lucas알고리즘의 가장 큰 단점은 주파수영역 각 bin에서 permutation이 발생할 수 있다는 것이다. 이는 $\mathbf{W}\cdot\mathbf{A}$ 가 scaled permutation 행렬을 해로 가질 수 있기 때문이다. 이러한 문제를 해결하기 위하여 시간영역에서 $W(\tau) = 0, \tau > Q \le T$ 의 제약을 둔다. 일반적으로 Q 는 $T/8$로 설정한다.

4. Simulation

4.1 시간성 잡음제거 성능 측정

본 논문에서의 실험은 실제 차량환경을 고려하여 인위적인 혼합에 의한 열화된 음성신호를 생성하였다. 먼저 시간성 잡음 제거 기술의 성능측정을 위하여 실제 차량주행환경에서 발생한 시간적 잡음을 녹취 후, 음성신호와 단순히 더하여 혼합하였다. 실험을 위한 음성신호는 PBW(Phonetically Balanced Word) 셋 중

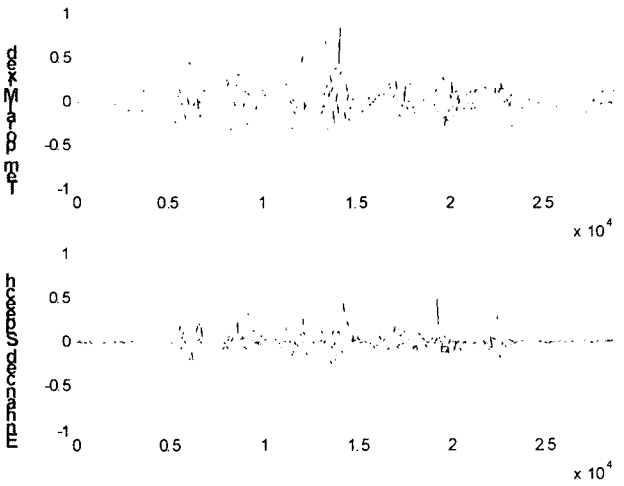

그림 4. Kalman Filter를 이용한 시간성 잡음 제거의 예.

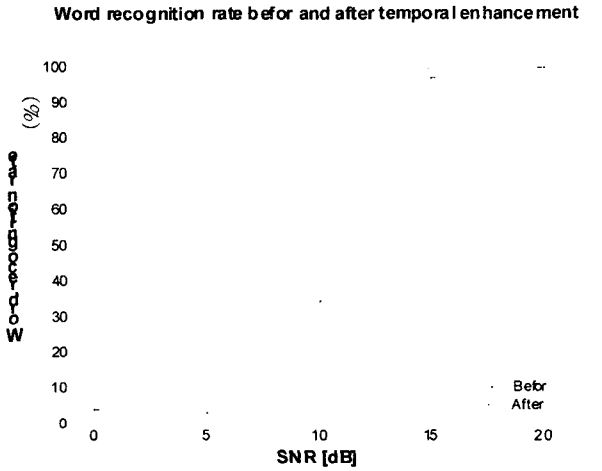

그림 5. SNR 별 시간성 잡음 제거 전/후 인식 결과.

임의의 남성화자가 1회 발성한 452개 단어를 일괄적으로 사용하였다. 그림 4.는 SNR 0dB에 대하여 Kalman Filter를 이용한 시간적 잡음제거 알고리즘을 적용한 결과 파형이다. 인식실험에 사용된 인식기는 HTK로 452단어에 대하여 55명(남:20, 여:25)이 1회 발성한 것을 훈련데이터로 사용한 것이다. MFCC 39차를 피처로 사용하였으며, triphone을 인식단위로 하였다. Mixture 수는 5개를 사용하였다. 그림 5.는 시간성 잡음의 SNR대비 잡음제거 전/후의 인식율을 나타낸다. 여기서 시간성 잡음만이 존재할 경우 열화된 음성은 Kalman Filtering 과정만으로 성공적으로 음성이 개선됨을 인식결과 알수 있다.

4.2 공간성 잡음제거 성능 측정

공간성 잡음에 의해 열화된 음성의 생성을 생성하기 위하여, 차량에 대한 혼합회로 $\mathbf{A} = (A_{ij})_{2,2}$ 를 다음과 같이 정의하였다.

$$A_{11} = 2.2 + z^{-1} - 0.75z^{-2} + 0.4z^{-3} + 0.3z^{-4} + 0.2z^{-5}$$

$$A_{12} = 1.8z^{-1} - 0.7z^{-2} + 0.45z^{-3} + 0.4z^{-4} + 0.1z^{-5}$$

$$A_{21} = 1.7z^{-1} - 0.45z^{-2} + 0.4z^{-3} + 0.2z^{-4} + 0.1z^{-5}$$

$$A_{22} = 1.9 + 0.8z^{-1} - 0.35z^{-2} + 0.3z^{-3} + 0.2z^{-4} + 0.1z^{-5}$$

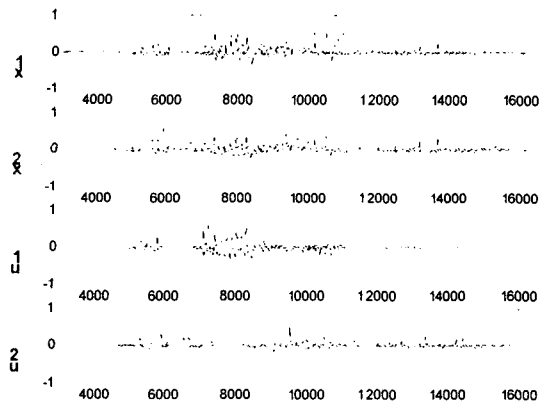

그림 6. 서로다른 음성신호의 혼합에서 음성 개선

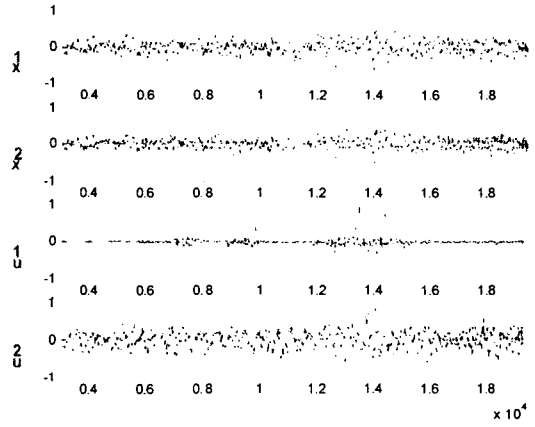

그림 7. 음성과 음악신호의 혼합에서 음성 개선

혼합 방법은 452개 각각의 단어를 $\mathbf{s}_1$ 로하여 이것을 인식대상어휘로 하고, 각각에 대하여 $\mathbf{s}_2$를 임의의 오디오 또는 다른 음성신호로 하여 $\mathbf{A}*[\mathbf{s}_1\ \mathbf{s}_2]^T$ 에 의하여 $[\mathbf{x}_1\ \mathbf{x}_2]^T$ 를 생성한다. $[\mathbf{x}_1\ \mathbf{x}_2]^T$ 에 대한 BSS 분리 신호 $[\mathbf{u}_1\ \mathbf{u}_2]^T$ 는 그림 6.과 7.에 나타내였다. 각각에 대한 인식 결과는 그림 8.과 같다. 여기서 공간성 잡음 제거 후 인식 결과는 제거 전과 큰 차이를 보임을 알 수 있다.

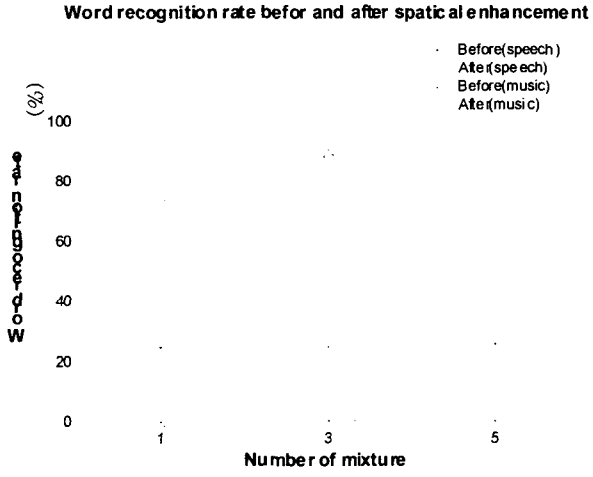

그림 8. Mixture 수에 따른 공간성 잡음제거 적용 전/후 인식율 결과

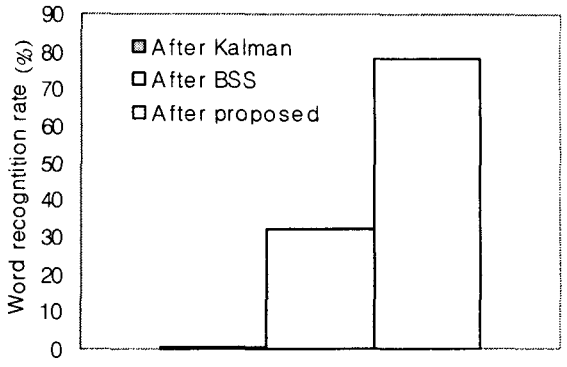

그림 9. 시간성+공간성 잡음에 대한 각 알고리즘 적용 전/후 인식성능

특히 그림 7.과 같이 구간 전반에 걸쳐 음악 신호와의 혼합된 혼합 신호는 인식율이 영에 가까웠으나 제거 후 최대 89.2%의 인식 성능을 보였다. 이러한 결과로부터 음성과 상관관계가 큰 신호일 지라도 인식 할 수 있도록 충분이 제거 될 수 있는 가능성을 입증한다.

4.3 제안된 알고리즘의 성능측정

마지막으로, 음성에 공간성 잡음과 시간성 잡음이 동시에 존재할 경우 각 알고리즘을 이용한 제거 후에 대하여 인식율을 측정하여 그림 9.에 나타내었다. 그림 9.의 결과를 통해 우리는 두가지 특성의 잡음이 음성내 존재할 때 하나의 제거방식으로 제거가 불가능 함을 알 수 있다. 또한 그림 1.에서 제안된 알고리즘의 구조로부터 시간성 잡음 제거는 공간성 잡음제거 필터보다 후처리 해야 함을 알 수 있다. 이는 공간성 잡음을 시간성 잡음제거로 처리할 경우 음성과의 상관관계로 인하여 음성의 스펙트럼 특성을 열화시킴으로서 인식이 될 수 없음을 그림 9.로부터 알 수 있다.

5. 결론

운전자의 안전과 편의를 보장할 수 있도록 DMB용 음성명령어기를 사용하기 위하여 차량 환경에서 시간성 잡음 뿐만 아니라 공간성 잡음도 제거 되어야 한다. 우리는 시간성, 공간성 잡음을 제거할 수 있는 가능성을 제안된 알고리즘을 가지고 실제환경을 연출한 인위적 열화 음성의 인식성능 측정으로부터 그 가능성을 검증하였다.

6. 참고문헌

[1] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320-327, May 2000.

[2] S.Jeong and M. Hahn, "Speech quality and recognition rate improvement in car noise environments," *Electronics Letters*, vol. 37, No 12, pp. 800-801, June, 2001.

[3] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129-1159, 1995.