말소리와 음성과학 제1권 제2호 (2009)
pp. 9~14

잡음음성인식을 위한 음성개선 방식들의 성능 비교

Performance Comparison of the Speech Enhancement Methods for Noisy Speech Recognition

정 용 주[1)]
Chung, Yongjoo

ABSTRACT

Speech enhancement methods can be generally classified into a few categories and they have been usually compared with each other in terms of speech quality. For the successful use of speech enhancement methods in speech recognition systems, performance comparisons in terms of speech recognition accuracy are necessary. In this paper, we compared the speech recognition performance of some of the representative speech enhancement algorithms which are popularly cited in the literature and used widely. We also compared the performance of speech enhancement methods with other noise robust speech recognition methods like PMC to verify the usefulness of speech enhancement approaches in noise robust speech recognition systems.

Keywords: Speech enhancement, noisy speech recognition, speech recognition systems

1. 서 론

잡음 환경의 음성인식을 위하여 지금까지 많은 연구가 진행되어 왔으며 그 결과물들은 실제 인식시스템에 성공적으로 사용되며 많은 발전을 이루어 왔다. 그중에서도 특히 음성개선(Speech Enhancement) 방식은 잡음 환경 음성인식에서 인식성능이 심하게 저하되는 것을 막기 위하여 가장 보편적으로 사용되어 왔다. 여러 가지 방식의 음성개선 알고리듬들이 존재하고 이에 따른 음성인식 결과들이 제시되었지만 이들 중에서 실제의 다양한 잡음 환경 하에서 어떤 알고리듬이 최상의 인식성능을 나타내는 가에 대해서는 다소 명확하지 않은 것이 사실이다.

주파수 차감법(Spectral Subtraction)은 음성 개선 방식 중에서 가장 먼저 제안된 알고리듬이라 할 수 있다[1] [2]. 주파수 차감법은 구현방식이 간단하다는 장점으로 인하여 전통적으로 많이 사용되어 왔으나 단순한 스펙트럼의 차감이 가져오는 음성신호의 왜곡현상 등으로 인하여 인식성능의 향상이 그리 만족스럽지 않으나 최근에는 다양한 변형된 방식이 제안되어 많은 성능 향상을 보임을 알 수 있다[3].

음성향상 알고리듬 중에서 현재까지 일반적으로 많이 사용되는 방식중의 하나는 Wiener 필터방식이다[4]. Wiener 필터는음성신호의 스펙트럼에 대한 MMSE(Minimum Mean Square Error) 추정 기반의 최적필터라는 점에서 다소 직관적인 주파수 차감법에 비해서는 개선된 방식으로 평가되나 선형필터라는 제약조건이 따르고 있다. 한편 Wiener 필터와 같은 MMSE 추정방식을 따르지만 음성신호와 잡음신호의 스펙트럼에 대한 사전 확률분포를 가정하고 통계모델에 근거하여 음성신호의 스펙트럼의 크기를 추정하는 MMSE-STSA(Short-Time Spectral Amplitude) 추정방식은 Wiener 필터에 비해서 다소 향상된 음성개선 성능을 보이는 것으로 알려져 있다[5]. MMSE-STSA 방식은 Wiener 필터방식에 비해서 선형필터라는 제약조건이 없으며 음성신호 존재 여부에 대한 확률을 음성개선에 이용한다는 차이점이 있다. 비교적 최근에는 기존의 음성개선 방식들과는 조금 다른 관점에서 진행된 연구결과가 소개되었는데 이들을 부공간(Subspace) 기반 방식이라 한다[6]. 이들 방식에서는 잡음음성신호의 벡터공간이 음성신호의 공간과 잡음신호의 공간으로 분리(decomposition) 될 수 있다는 가정하에 KLT(Karhunen-Loeve Transformation)을 잡음음성신호에 적용하여 원래의 깨끗한 음성신호를 분리해낸다.

앞에서 설명된 바와 같이 잡음음성신호의 음질개선을 위한

1) 계명대학교 yjjung@kmu.ac.kr, 교신저자
본 연구는 2009년도 계명대학교 비사연구기금으로 이루어졌음.

접수일자: 2009년 4월 30일
수정일자: 2009년 6월 4일
게재결정: 2009년 6월 20일

기존의 방식들은 주파수 차감법을 활용한 방식, Wiener 필터방식, MMSE-STSA로 대표되는 통계모델 기반 방식 그리고 KLT를 이용한 부공간 방식 등의 순서로 제안되어 왔으며 또한 각각의 방식들은 보다 나은 성능을 위하여 개선된 알고리듬들과 변형들이 존재하고 있다. 본 연구에서는 이러한 음질개선 방식들 중에서 가장 대표적이라고 알려진 주요방식들에서 대해서 잡음환경의 음성인식성능을 비교 검토하고자 한다. 또한 이러한 음질개선 방식들이 최근 많이 소개되고 있는 모델보상 방식 및 특징보상방식과 비교하여 어느 정도의 성능을 나타내는지에 대해서도 검토할 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 음질개선 방식 중에서 중요한 몇 가지 방법에 대하여 간략히 소개하며 3장에서는 인식실험에 사용된 음성데이터 및 인식시스템에 대해서 소개하며 4장에서는 음질개선방식을 이용한 잡음환경 음성인식시스템의 성능에 대한 실험결과를 소개하며 5장에서 결론을 기술한다.

2. 음성개선 방식의 개요

2.1 주파수 차감법 (Spectral Subtraction Methods)

$s(t)$ 와 $d(t)$는 각각 음성신호와 부가잡음신호를 나타낸다고 하면, 오염된 잡음음성신호 $x(t)$는 일반적으로 다음과 같이 표현될 수 있다.

$$x(t) = s(t) + d(t) \tag{1}$$

한편 식(1)의 각 신호들의 k번째 주파수 성분들은 각각 $X_k = R_k e^{j\nu_k}$, $S_k = A_k e^{j\alpha_k}$ 그리고 D_k로 표시된다고 가정하면 각 신호들의 전력(power) 스펙트럼은 다음과 같은 관계식을 가진다고 가정 된다.

$$R_k^2 = A_k^2 + |D_k|^2 \tag{2}$$

따라서 식(2)에 근거하여 음성신호의 전력스펙트럼의 추정치 $\widehat{A_k^2}$ 는 다음과 같이 구해진다.

$$\widehat{A_k^2} = R_k^2 - \widehat{D_k^2} \tag{3}$$

여기서 $\widehat{D_k^2}$ 는 잡음신호의 전력스펙트럼의 추정치이며 이는 음성신호가 없는 구간에서 추정된다. 위의 식(3)에서 결정된 음성신호의 전력스펙트럼은 잡음신호의 전력스펙트럼 추정치의 부정확성으로 인해서 추정의 오류를 가져오는데 특히 musical 잡음이라는 새로운 잡음신호가 발생하는 바람직하지 않은 결과를 가져오는 문제점이 있다. 이를 개선하기 위하여 잡음 스펙트럼을 추정치보다 과도하게 차감해주며 또한 음성신호의 스펙트럼의 최소값을 보장하는 아래식과 같은 비선형 주파수 차감법이 제시되었다[2].

$$\widehat{A_k^2} = \begin{cases} R_k^2 - \alpha \widehat{D_k^2} & \text{if } R_k^2 \geq (\alpha+\beta)\widehat{D_k^2} \\ \beta \widehat{D_k^2} & else \end{cases} \tag{4}$$

식(4)에서 $\alpha \geq 1$이며 $0 < \beta \ll 1$ 이다. 위의 비선형주파수 차감법에 대한 여러 가지 변형들이 존재하는데, Lockwood 등은 비선형 주파수 차감법의 과차감 요소인 $\alpha(k)$ 값을 주파수성분 k에 따라서 다르게 선정하는 방식을 제안하였으며[7], Kamath 등은 Lockwood 방식과 근본적으로 유사하나 주파수 성분별로의 SNR (Signal to Noise Ratio) 값이 변동이 심한 점을 감안하여 과차감 요소인 α_i($i = 1, \cdots, B$, B = 주파수대역의 개수)를 각 주파수 성분이 아닌 주파수대역별로 다르게 선정하는 방식(Multi-band)을 제안하여 우수한 성능을 보임을 확인하였다[3].

2.2 Wiener 필터(Wiener filter)

앞에서 언급된 주파수 차감법과 그 변형들은 주로 직관적이며 경험적인 관점에 근거하며 제안된 방식으로서 최적화된 기준에 의하여 유도된 방식들이 아니다. 반면에 Wiener 필터 기법은 MSE(Mean Square Error)를 최소화하는 선형 FIR 필터의 구현이라는 최적기준에 근거하여 제안된 방식이다.

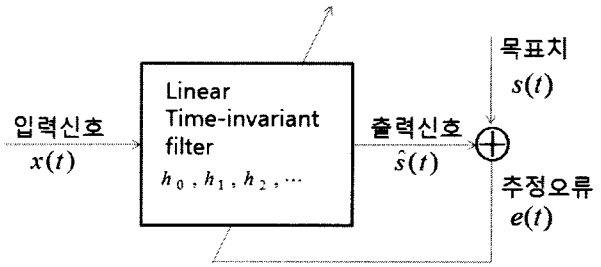

그림 1. 선형 최적 필터의 블록다이어그램
Figure 1. Block diagram of the linear optimal filter

선형필터 이론에 근거하여 <그림1>에서의 필터링 오차 $e(t)$의 자승을 최소화하는 Wiener 필터에 의하여 추정된 음성신호의 전력스펙트럼 값은 다음과 같다.

$$\widehat{A_k^2} = \left(1 - \frac{\widehat{D_k^2}}{R_k^2}\right)^2 R_k^2 \tag{5}$$

2.3 통계모델 기반 방법

통계적 모델 기반 기법이란 잡음과 음성신호 스펙트럼에 대한 통계적 모델을 가정하여 다양한 최적화 기준에 근거하여 음성스펙트럼을 추정하는 방식을 말한다. 특히 이 중에서도

MMSE(Minimum Mean Square Error) 기반의 음성신호 크기 스펙트럼을 추정하는 방식이 매우 효과적인 것으로 알려져 있다[5]. 아래 식은 음성신호 $s(t)$의 크기 스펙트럼에 관한 MMSE 추정치를 나타낸다.

$$\hat{A}_k = E\{A_k | X_k\} \tag{6}$$

A_k에 대한 MMSE 추정치를 나타내는 식(6)에서 잡음음성스펙트럼 X_k에 대한 조건 확률밀도 함수 $P(X_k|S_k)$와 음성스펙트럼 S_k에 사전 확률밀도함수 $P(S_k)$를 가정함으로서 아래와 같은 결과를 얻을 수 있다.

$$\hat{A}_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} exp(-\frac{v_k}{2}) \cdot \left[(1+v_k) I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right] R_k \tag{7}$$

$\Gamma(.)$는 감마 함수를 나타내며 $I_0(\bullet), I_1(\bullet)$는 각각 0차와 1차의 베셀함수를 의미한다. 한편 식(7)에서 v_k와 γ_k는 다음과 같이 정의된다.

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \tag{8}$$

$$\xi_k = \frac{\lambda_s(k)}{\lambda_D(k)} = \frac{E\{A_k^2\}}{\widehat{D_k^2}}, \gamma_k = \frac{R_k^2}{\widehat{D_k^2}}$$

여기서 ξ_k와 γ_k는 각각 사전(prior) SNR 및 사후(posteriori) SNR 이라 칭해지며 사전 SNR은 일반적으로 decision-directed 추정방식으로 얻어진다[5].

한편 식(7)은 음성신호의 존재여부에 불확실성을 고려하면 다음과 같이 유도된다.

$$\widetilde{A}_k = \frac{\Lambda(X_k, q_k)}{1+\Lambda(X_k, q_k)}\hat{A}_k \tag{9}$$

$$\Lambda(X_k, q_k) = \mu_k \frac{p(X_k|H_k^1)}{p(X_k|H_k^0)}, \ \mu_k = \frac{1-q_k}{q_k} \tag{10}$$

여기서 q_k는 k번째 주파수 성분에서 음성신호가 존재할 확률을 나타내며 H_k^0와 H_k^1는 각각 k번째 주파수 성분에서 음성신호의 부존재와 존재 가설을 나타낸다. 한편, $\Lambda(X_k, q_k)$는 식(7)을 유도할 때 사용되었던 음성 및 잡음 음성의 스펙트럼의 주파수 성분들에 대한 Gaussian 통계모델을 사용함으로서 다음과 같이 얻을 수 있다.

$$\Lambda(X_k, q_k) = \mu_k \frac{\exp(v_k)}{1+\xi_k}, \ \xi_k = \frac{E\{A_k^2\}}{\lambda_d(k)} \tag{11}$$

2.4 부공간 방법(Subspace based Methods)

잡음음성신호의 공분산 행렬의 eigen decomposition을 이용하여 Ephraim 등은 잡음음성신호의 벡터 공간이 음성신호와 잡음신호의 공간으로 분리됨을 이용하여 음성향상이 가능함을 보였다[6]. 분리된 음성신호 공간 성분은 적절한 이득함수를 통하여 원래의 음성과 가깝도록 수정되고 잡음신호 공간 성분은 제거함으로서 음성개선을 이룰 수 있었다. 이 방법에서는 음성개선 결과의 잔류잡음이 특정 임계치보다 작다는 조건하에서 음성신호의 왜곡을 최소화 하는 최적 추정 알고리듬이 제안되었다.

식(1)과 같은 잡음음성신호에 대한 발생 모델이 주어져 있다고 가정하고 $\hat{s} = Hx$는 음성신호 s에 대한 선형 추정 값 이라고 가정하면 추정오차는 다음과 같다.

$$\varepsilon = \hat{s} - s = (H - I)s + Hn = \varepsilon_s + \varepsilon_n \tag{12}$$

여기서 ε_s는 음성신호의 왜곡을 나타내고 ε_n는 잔류 잡음신호를 나타내며 H는 $K \times K$ 추정 행렬식을 나타내며 K는 s의 차원(dimension)을 나타낸다.

$$\overline{\varepsilon_s^2} = trE[\varepsilon_s \varepsilon_s^T] \tag{13}$$

$$\overline{\varepsilon_n^2} = trE[\varepsilon_n \varepsilon_n^T] = \sigma_n^2 tr\{HH^T\} \tag{14}$$

$\overline{\varepsilon_s^2}$는 음성신호의 왜곡의 에너지 값이며 $\overline{\varepsilon_n^2}$는 잔류잡음신호의 에너지값을 나타낸다. 여기서 σ_n^2은 대각 공분산 행렬을 가지는 잡음신호 n의 공분산 값이다.

Ephraim 등이 제안한 부공간 방식의 음성개선방식은 잔류잡음신호의 에너지값 $\overline{\varepsilon_n^2}$의 최대 허용치에 제한을 두고 음성신호의 왜곡 에너지값 $\overline{\varepsilon_s^2}$을 최소화 하는 선형 추정행렬식 H를 구하는 것이며 아래와 같은 식으로 나타낼 수 있다.

$$\min_H \ \overline{\varepsilon_s^2} \tag{15}$$

단, $\frac{1}{K}\overline{\varepsilon_n^2} \le \alpha\sigma_n^2$ 이며 $0 \le \alpha \le 1$ 이다.

식(15)에 대한 해는 Kuhn-Tucker 알고리듬을 적용하여 얻을 수 있으며 아래와 같다.

$$H_{opt} = R_s(R_s + \mu\sigma_n^2 I)^{-1} \tag{16}$$

$$\alpha = \frac{1}{K} tr\{R_s^2(R_s + \mu\sigma_n^2 I)^{-2}\} \tag{17}$$

한편 음성신호 s의 공분산 행렬 R_s에 대한 eigen decomposition은 아래와 같이 이루어진다는 것이 알려져 있다.

$$R_s = U\Lambda_s U^T \qquad (18)$$
$$\Lambda_s = diag[\Lambda_s', 0I]$$
$$\Lambda_s' = diag(\lambda_s(1), \cdots, \lambda_s(M))$$

여기서 $U=[u_1,\cdots,u_K]$는 잡음음성신호 x의 공분산행렬인 R_x의 eigenvector $u_k \in R^K$들로 이루어진 직교정상행렬이며 $\lambda_s(k)$는 아래 식에서와 같이 구해진다.

$$\lambda_s(k) = \lambda_x(k), \; k = 1, \cdots M \qquad (19)$$

여기서 $\lambda_x(k)$는 R_x의 eigenvalue를 의미하며 $M = Rank(R_s)$이다.

식(18)을 식(16)에 대입하면 아래와 같이 최종적으로 H_{opt} 값을 얻을 수 있다.

$$H_{opt} = \sum_{m=1}^{M} g_\mu(m) u_m u_m^T$$
$$g_\mu(m) = \frac{\lambda_s(m)}{\lambda_s(m) + \mu\sigma_n^2} \qquad (20)$$

또한 식(17)과 (18)로부터 아래과 같이 μ 값을 얻을 수 있다.

$$\alpha = \frac{1}{K} tr\left\{\Lambda_s'^2 (\Lambda_s' + \mu\sigma_n^2 I)^{-2}\right\} \qquad (21)$$

3. 음성데이터 및 인식시스템

본 연구를 위하여 잡음음성인식을 위한 Aurora 2 데이터베이스를 사용하였다. Aurora 2 데이터베이스는 TI digit 데이터베이스에 인공적으로 부가잡음을 더해주고 채널왜곡을 인가한 것이다. 훈련은 Clean과 Multi 방식이 있으며 전자는 HMM의 훈련시에 깨끗한(clean) 음성데이터 만을 이용하고 후자는 깨끗한 음성과 함께 여러 가지 종류의 잡음음성을 이용하여 HMM을 훈련하는 방식이다. 인식을 위해서는 3가지 종류의 음성데이터 Set 이 사용된다. Set A는 인식잡음이 훈련시 사용된 잡음의 종류와 같은 경우(Subway, Babble, Car, Exhibition noises)이며 Set B는 인식잡음이 훈련 잡음과 다른 경우(Restaurant, Street, Airport, Train-Station noises) 이고 Set C는 앞의 부가 잡음 외에도 채널 왜곡이 인위적으로 조성된 경우이다.

음성특징을 위해서는 ETSI (European Telecommunications Standard Institute) Aurora 그룹에서 제안된 0차의 cepstral 계수를 포함한 13차의 MFCC(Mel-frequency Cepstral Coefficient) 를 사용하였고 delta 와 acceleration 계수를 추가하여 전체 39차의 특징벡터를 사용하였다.

숫자에 대한 HMM은 3개의 Gaussian 확률밀도함수 성분을 가지는 16개의 상태들로 이루어지며 묵음에 관한 HMM 모델은 6개의 Gaussian 성분을 가지는 3개의 상태로 이루어진다. 한편 1개의 상태를 가지는 짧은 묵음에 관한 HMM 모델도 구성되며 이는 묵음모델의 가운데 상태와 동일하게 모델링 된다[8].

인식기는 본 연구실에서 자체적으로 개발하였으며 Baum-Welch 기반의 훈련과정을 통해서 HMM 모델을 구성하고 Viterbi 알고리듬을 이용하여 인식결과를 얻도록 하였다[9].

4. 인식실험 및 고찰

본 연구에서는 7개의 대표적 음질개선 방식을 구현하고 Aurora 2 데이터베이스를 이용한 인식실험을 수행하였다. 각각의 방식에 대한 약어와 관련된 이론 및 출처는 아래와 같다.

1) SS: Berouti 등에 의해서 제안된 방식이며 과차감 요소를 이용한 주파수 차감법의 초기 방식이며 다른 많은 주파수 차감 기반 음성향상 기법의 근간이 된다[2].

2) MBAND: Kamath 등에 의해서 제안된 방식이며 1)의 SS 방식과의 주된 차이점은 주파수 대역별로 그리고 SNR 값에 따라 서로 다른 과차감 요소를 사용한다는 점이다[3].

3) WIENER: Prior SNR 값 추정을 이용한 Wiener 필터링 방식이다[4].

4) MMSE-STSA: Ephraim 등이 제안한 통계모델 기반의 대표적 추정 방식이다[5].

5) MMSE-STSA(SPU) : MMSE-STSA 방식에서 음성존재 확률 정보를 추가로 이용한 방법이다.

6) logMMSE-STSA: MMSE-STSA 방식에서는 음성신호의 스펙트럼 크기를 추정하는 것에 비해서 여기서는 스펙트럼 크기의 log 값을 추정 한다[10].

7) SUBSPACE : Ephraim이 제안한 부공간 방식에 기반한 음성 개선방법 이다[6].

위의 방식들 중에서 1)과 2)는 주파수 차감법에 기반하고 3)은 Wiener 필터 방식을 나타내며 4),5),6)는 통계모델 기반의 방식들이며 7)은 부공간 방식을 대표한다. <표1>에서는 위의 음성개선 알고리듬을 Aurora 2 데이터베이스에 적용한 인식결과를 나타낸다. HMM 모델은 Clean 방식으로 훈련되었으며 베이스라인은 어떠한 음성개선도 적용되지 않은 경우를 나타낸다. 향상률은 베이스라인의 단어인식율(Word accuracy)에 비해서 상대적으로 얼마만큼의 인식율의 향상이 있었는지를 퍼센트 단위로 나타낸 것이며 평균인식률은 [12]에서 사용된 것처럼(Set A 인식률+Set B 인식률)*0.4 + set C 인식률*0.2로 나타낸다.

<표1>의 결과로부터 우리는 잡음음성 환경에서 전통적으로 많이 사용된 주파수 차감법이나 Wiener 필터 방식에 비해서 통

계모델방식이나 부공간 방법의 성능이 우수함을 알 수 있으나 주파수 차감법의 경우에는 최근에 개선된 방식의 경우에는 성능이 매우 우수한 것을 알 수 있다. 특히, 전통적인 스펙트럼 차감법 SS 는 거의 성능의 향상을 이루지 못함을 알 수 있는데, 이는 잡음의 제거를 통하여 SNR의 향상은 이루었지만 musical 잡음 등에 의한 음성신호의 왜곡으로 인하여 전반적으로 인식성능의 향상을 이루지 못한 원인인 것으로 판단된다.

표 1. 음질개선 방식의 인식성능비교(Word accuracy(%))
Table 1. Performance comparison between speech enhancement methods

	Set A	Set B	Set C	평균	향상률
베이스라인	59.3	55.1	67.5	59.3	-
SS	58.5	56.2	67.0	59.3	0
MBAND	72.3	69.2	78.3	72.3	21.9
WIENER	55.4	53.7	61.3	55.9	-5.7
MMSE_STSA	71.6	66.2	78.9	70.9	19.5
MMSE-STSA (SPU)	67.1	63.2	72.4	66.6	12.3
logMMSE-STSA	71.4	66.7	77.3	70.7	19.2
SUBSPACE	62.3	63.9	63.2	63.1	6.4

반면에 MBAND의 경우에는 과차감 요소를 주파수 대역과 SNR 값에 의존하여 조정하여 줌으로서 매우 우수한 인식성능을 얻을 수 있음을 알 수 있었다. MMSE-STSA로 대표되는 통계모델링 방식은 전반적으로 높은 인식성능을 보임을 알 수 있었다. 여기서는 음성신호와 잡음신호의 스펙트럼에 대한 확률분포를 음질개선에 활용함으로서 매우 높은 성능을 얻을 수 있었다고 판단된다. MMSE-STSA 방식은 MBAND 와 함께 가장 높은 인식성능을 보임을 알 수 있었다. 한편 logMMSE-STSA 방식은 음성신호의 크기 스펙트럼의 log 값을 추정하는 방식인데, 기존의 연구결과에서는 logMMSE-STSA의 음질개선 향상이 MMSE-STSA 보다 뛰어난 것으로 보고되었으나[10], 본 실험결과에 따르면 인식성능은 오히려 다소 저조한 것으로 보인다. 아마도 다소 주관적인 음질개선의 효과와 인식기를 통한 인식성능의 향상과는 여러 가지 면에서 차이점이 있기 때문인 것으로 생각된다. 한편, Ephraim 등은 MMSE-STSA 방식에서 음성신호의 존재에 대한 불확실성(SPU:Signal Presence Uncertainty)를 고려한 음질개선방식 MMSE-STSTA(SPU)를 제안하였다. 실험결과 MMSE-STSA(SPU)는 원래의 MMSE-STSA보다 저조한 인식성능을 보임을 알 수 있었다. 이는 결국 음성신호의 존재에 대한 불확실성에 관한 정확한 정보가 제공되지 않으면 음질개선에 오히려 악 영향을 미친다는 것을 의미한다. 이상과 같이 우리는 기존의 음질개선 방식이 인식성능에 미치는 효과를 살펴보았는데, 일반적으로 통계모델 기반의 방식이 우수한 성능을 보임을 알 수 있었으나, 주파수 차감법의 경우에는 과차감 요소를 적절히 설정함으로서 통계모델 방식과 비슷한 성능을 보임을 알 수 있었다. 또한 비교적 최근에 제안된 부공간 방식은 성능 에서는 통계모델 방식이나 향상된 주파수 차감법에 미치지는 못한 것으로 보인다.

<표2>에서는 앞에서 언급된 음질개선 방식 중에서 성능이 가장 우수한 MBAND 방식과 모델파라미터보상 방식과의 성능비교를 하였다. 모델파라미터보상방식으로는 PMC(Parallel model combination) 방식을 적용하였다.

표 2. 음질개선 방식과 모델파라미터보상 방식간의 인식성능비교(Word accuracy(%))
Table 2. Performance comparison between the speech enhancement method and the model parameter compensation method.

	Set A	Set B	Set C	평균	향상률
베이스라인	59.3	55.1	67.5	59.3	-
MBAND	72.3	69.2	78.3	72.3	21.9
PMC	79.3	81.2	78.0	79.8	34.5

<표2>의 결과를 보면 MBAND는 모델파라미터보상 방식인 PMC 방식에 비해서는 다소 저조한 인식성능을 보임을 알 수 있다. 잡음음성인식의 또 다른 방식인 특징보상 방식의 성능이 PMC 등의 모델보상 방식과 유사한 인식성능을 나타냄을 고려해 볼 때[11], 음성개선 방식은 특징보상 방식에 비해서도 저조한 성능을 나타내는 것으로 판단된다. 따라서 잡음음성인식의 3가지 방식 중에서 음질개선 방식은 다른 2가지 방식에 비해서는 다소 저조한 성능을 보이는 것으로 판단되므로 잡음음성인식을 위한 음성인식시스템의 구성시에는 음질개선 방식을 단독으로 사용하기 보다는 다른 보상방식과 함께 적용하는 접근 방식을 고려하는 것이 바람직 할 것으로 보인다.

5. 결 론

본 연구에서는 잡음음성인식을 위하여 전통적으로 많이 활용되었던 여러 가지 음질개선방식들에 대하여 성능비교를 수행하였고 그 결과 주파수 차감법의 개선된 기법인 MBAND 방식과 잡음신호에 대한 통계적 모델에 기반한 MMSE-STSA 방식의 성능이 가장 우수한 것을 확인 할 수 있었다. 그러나 단순한 음질개선 방식은 잡음음성인식의 또 다른 방식들인 모델파라미터보상방식이나 특징보상 방식에 비해서는 성능이 다소 저조한 것으로 판단된다. 따라서 음질개선 방식만을 적용하여 잡음에 강인한 음성인식시스템을 구성하는 것은 다소 한계가 있을 것으로 판단되며 최근에 많이 제안되고 있는 모델파라미터보상 방식이나 특징보상 방식등과의 연계를 통해서 보다 나은 잡음음성인식시스템을 구성할 수 있을 것으로 판단된다.

참 고 문 헌

[1] Ball, S. F. (1979). "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 27, pp. 113-120.

[2] Berouti, M., Schwartz, R., Makhoul, J. (1979). "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp.208-211.

[3] Kamath, S., Loizou, P. (2002). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 101-111.

[4] Scalart, P., Filho, J. (1996). "Speech enhancement based on a priori signal to noise estimation", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 629-632.

[5] Ephraim, Y., Malah, D. (1984). "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 32, pp. 1109-1121.

[6] Ephraim, Y., Van Trees, H.L. (1995). "A signal subspace approach for speech enhancement", *IEEE Trans. Speech, Audio Processing.*, Vol. 3, No. 4, pp. 251-266.

[7] Lockwood, P., Buddy, J. (1992). "Experiments with a nonlinear spectral subtractor(NSS), hidden Markov models and the projection, for robust speech recognition in cars", *Speech Communication.*, Vol. 11, No. 2-3, pp.215-228.

[8] Pearce, D., Hirsch, H. (2000). "The Aurora experimental framework for the performance evaluation of speech recognition systems under conditions", *Proc. ICSLP*, Vol. IV, pp. 29-32, Beijing, China.

[9] Kim, H. K., Chung, Y. J. (2006). "An implementation of the baseline recognizer using the segmental K-means algorithm for the noisy speech recognition using the Aurora DB", Malsori, No. 57, pp. 113-122.
(김희근 · 정용주, (2006). "Aurora DB를 이용한 잡음음성인식 실험을 위한 Segmental K-means 훈련방식의 기반인식기의 구현", 말소리, No. 57, pp. 113-122.)

[10] Ephraim Y., Malah, D. (1985). "Speech enhancement using a minimum mean square error log-spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 23, pp. 443-445.

[11] Kim, N. S. (2003). "Speech recognition under noisy environments", SK Telecommunications Review, pp. 650-661.
(김남수, (2003). "잡음환경에서의 음성인식 기술", SK Telecommunications Review, pp. 650-661.)

[12] ETSI draft standard doc. Speech Processing, (2000). Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, ETSI ES 202 108 V1.1.2 (2000-04), http://pda.etsi.org/pda/queryform.asp, April

• **정용주 (Chung, Yongjoo)**
계명대학교 전자공학과
대구시 달서구 신당동 1000번지
Tel: 053-580-5925 Fax: 053-580-5165
Email: yjjung@kmu.ac.kr
관심분야: 음성인식, 신호처리
1999~현재 전자공학과 부교수
Ph.D., Dept. of Electrical Engineering, KAIST, 1995