잡음에 강한 음성 인식에서 SNR 기준 함수를 사용한 가우시안 함수 변형 및 결정에 관한 연구

A Study on Variation and Determination of Gaussian Function Using SNR Criteria Function for Robust Speech Recognition

전 선 도*, 강 철 호*

(Sun Do June*, Chul Ho Kang*)

요 약

잡음에 강한 음성 인식 시스템을 위하여 주파수 차감법을 사용할 경우 음성 신호마저 차감하여 신호를 더욱 부식시키는 경우가 존재한다. 본 연구에서는 이러한 경우를 위해서 프레임마다 추정 잡음과 차감 신호의 SNR(Signal to Noise Ratio) 함수로부터 반연속 HMM(Hidden Markov Model)의 가우시안 함수를 변형 및 결정하는 방법을 제안한다. 이 방법의 타당성을 위해 프레임마다 추정 잡음의 오류 정도가 추정 잡음의 크기와 관계함을 신호 파형 형태로써 보였으며, 이러한 이유에서 SNR을 기준으로 가우시안 함수를 변형 및 결정하게 된다. 실험에서 80km/h 이상의 속도로 달리는 차량내에서 배경 잡음과 음성이 혼합되었을 때의 음성 인식율을 평가하였다. 그 결과 주파수 차감한 경우와 차감하지 않은 경우에 비해 본 논문에서 제안한 SNR에 의한 가우시안 결정 방법이 더욱 향상된 인식율을 보였다.

ABSTRACT

In case of spectral subtraction for noise robust speech recognition system, this method often makes loss of speech signal. In this study, we propose a method that variation and determination of Gaussian function at semi-continuous HMM(Hidden Markov Model) is made on the basis of SNR criteria function, in which SNR means signal to noise ratio between estimation noise and subtracted signal per frame. For proving effectiveness of this method, we show the estimation error to be related with the magnitude of estimated noise through signal waveform. For this reason, Gaussian function is varied and determined by SNR. When we test recognition rate by computer simulation under the noise environment of driving car over the speed of 80km/h, the proposed Gaussian decision method by SNR turns out to get more improved recognition rate compared with the frequency subtracted and non-subtracted cases.

I. 서 론

잡음에 강한 음성 인식 시스템을 구현하기 위해서 먼저 배경 잡음을 제거한 이후에 인식에 적용을 한다. 그러나 잡음 제거 시스템에서 잡음을 제거한다고 하여도 인식율의 증가를 가지고 오는 것은 아니다. 왜냐하면 잡음 제거하는 과정에서 잡음뿐만이 아니라 음성까지 제거시켜 오히려 음성 인식의 저하를 가져오는 경우가 발생한다. 이러한 이유에서 잡음 제거 후 잡음 보상을 위한 알고리즘에 관한 여러 방법이 연구되고 있다. 이러한 방법으로는 주파수 차감 이후 HMM의 평균 및 분산 등의 파라미터를 가변 시키는 방법[1][2], HMM의 모델을 학습 시에 잡음에 대한 정보를 미리 모델화 하는 방법[3] 등이 있다. 본 연구에서는 잡음에 강한 음성 인식 시스템을 위하여 인식 전에 주파수 차감법을 이용하여 잡음 제거를 한다. 그 이후 추정한 잡음과 차감된 추정 음성

*광운대학교 전자통신공학과
접수일자 : 1999년 8월 31일

신호간의 SNR을 기준으로 차감된 음성의 가우시안 분포 및 혼합된 음성의 가우시안 분포를 변형 및 결정하면서 반연속 HMM에 적용하는 방법이다. 이러한 방법은 잡음 제거방법인 차감법이 오히려 음성 인식에 저하를 가지고 오는 단점을 보상해주는 방법이 된다. 실험에서 차량내의 잡음 정도가 극심한 배경 잡음과 34개의 고립 단어를 혼합하여 6명 화자에 대한 화자 종속 실험을 하였다. 실험 결과에서는 제안하는 방법이 차감하거나 차감하지 않은 경우보다 높은 인식율을 보임으로써 이 방법이 주파수 차감한 신호의 가우시안 함수와 차감하지 않은 신호의 가우시안 함수를 효과적으로 변형 및 결정하는 방법임을 알 수 있다. 결국 이러한 방법은 음성 정보의 손실을 보완해 주는 것이다.

Ⅱ. 주파수 차감법 및 반연속 HMM

2.1 주파수 차감법

잡음 섞인 음성 신호의 모델은 다음과 같다.

$$s(n)+no(n)=x(n) \tag{1}$$

여기서 $s(n)$는 음성 신호의 모델이고, $no(n)$는 배경 잡음을 의미한다. 결국 입력되는 음성 신호는 잡음과 혼합된 $x(n)$신호이다.

본 시스템은 전력 스펙트럼에 의하여 잡음을 추정한다. 다음 식은 전력 스펙트럼 변환식이다.

$$X_i(k)=\left|\sum_{n=0}^{N} x_i(n) \quad e^{\left(-\frac{j2\pi kn}{N+1}\right)}\right|^2 \tag{2}$$

$$N_i(k)=\left|\sum_{n=0}^{N} no_i(n) \quad e^{\left(-\frac{j2\pi kn}{N+1}\right)}\right|^2 \tag{3}$$

윗 식은 DFT(Discrete Fourier Transform)으로서 실제 구현에서는 FFT(Fast Fourier Transform) 으로 구현했으며, N은 256이다.

잡음 추정 알고리즘에 의하여 임이의 i프레임에 대하여 잡음 전력 스펙트럼 $\overline{\mu_i}(k)$을 추정한다. 이렇게 추정한 잡음 스펙트럼으로 음성 신호를 추출할 수 있다. 다음 식은 차감식을 보여주고 있다.

$$\overline{S_i(k)}=X_i(k)-\overline{\mu_i}(k) \tag{4}$$

위 식에서 입력 전력 스펙트럼 $X_i(k)$에서 추정 잡음 전력 스펙트럼 $\overline{\mu_i}(k)$를 차감 하여 잡음 제거된 음성 전력 스펙트럼 $\overline{S_i(k)}$를 얻는다. 다음 식은 가중치 주파수 차감법에 의한 전 프레임의 추정 잡음 전력 스펙트럼에 의하여 현재 프레임의 잡음을 추정하는 식이다[4].

$$\overline{\mu_i}(k)=\beta((1-a) \quad X_i(k)+a\overline{\mu_{i-1}}(k)) \tag{5}$$

위 식은 i프레임의 입력 전력 스펙트럼 $X_i(k)$와 i-1 프레임의 추정한 잡음 전력 스펙트럼 $\overline{\mu_{i-1}}(k)$에 의하여 i 프레임의 잡음 추정 값 $\overline{\mu_i}(k)$을 결정한다. 이 때 a파라미터는 잡음을 추정할 때 $X_i(k)$와 $\overline{\mu_{i-1}}(k)$의 비율을 결정하며, β파라미터는 추정 잡음의 크기를 결정한다.

2.2 반연속 HMM

음성 인식을 하는 가운데 이산 HMM(discrete HMM)은 계산은 빠르나 코드북의 크기가 커야만 확률 분포를 잘 나타낼 수 있고, 연속 HMM(continuous HMM)은 계산량이 많다는 단점이 있다. 이런 점을 고려해서 이산 HMM보다 작은 코드북을 사용하고 연속 HMM보다 적은 계산량이 필요하도록 두 경우를 결합한 것을 반연속 HMM (Semi-Continuous HMM : SCHMM)이라 한다[5]. 이 경우는 백터 양자화의 크기 L인 코드북에서 각 코드워드에 해당하는 D차 평균값 μ와 공분산 행열의 주대각선 성분 $\sum$가 D개 주어지게 된다. 각 코드워드마다 공분산 행렬값이 주어지므로 일반적인 벡터 양자화의 경우와는 달리 유클리디안 (Euclidean) 거리 대신 마할라노비스(Mahalanobis) 거리를 사용하게 된다[6].

이 경우 관측 확률 파라미터 $b_j(l)$은 상태 j에서 코드북 l번째 가우시안 함수 $g(\cdot)$의 상대적인 크기와 내적 곱으로 표현된다. 그래서 상태 j에서 관찰값 o_t를 발견할 확률은 다음 식과 같다.

$$p_j(o_t)=\sum_{l=1}^{L} b_j(l)g(o_t \setminus \mu_l, \sum_l) \tag{6}$$

이와 같은 반연속 HMM을 음성 인식에 사용하기 위해서는 음성의 학습 데이터를 반연속 HMM의 파라미터로써 상태 i 에서의 초기파라미터 π_i, 상태 i 에서 상태 j로의 천이 파라미터 a_{ij} , 상태 i 에서 코드북 인덱스 l 에 의한 관측 파라미터 $b_j(l)$를 재추정 하는 과정이 필요하다. 또한 이 파라미터가 추정되면서 관찰열에서 가우시안 함수의 평균 μ와 분산 $\sum$이 추정되어야 하는데 대표적인 방법으로는 반복적으로 최대화시키는 EM(Expectation Maximization) 알고리듬이 있다[7]. 인식 과정에서는 반연속 HMM 파라미터들로부터 하나의 관찰열에 대응되는 가장 적합한 상태열을 찾는 방법으로 비터비 (Viterbi)알고리듬을 사용한다. 다음은 Baum-Welch 알고리듬과 EM 알고리듬에 의한 추정식이다[5].

$$\pi_i=\gamma_1(i) \tag{7}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\gamma_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{8}$$

$$b_j(l) = \frac{\sum_{t=1}^{T-1}\zeta_t(j,l)}{\sum_{t=1}^{T-1}\gamma_t(j)} \tag{9}$$

$$\mu_l = \frac{\sum_{t=1}^{T-1}\zeta_t(l)o_t}{\sum_{t=1}^{T-1}\zeta_t(l)} \tag{10}$$

$$\sum_l = \frac{\sum_{t=1}^{T-1}\zeta_t(l)(o_t-\mu_l)(o_t-\mu_l)^t}{\sum_{t=1}^{T-1}\zeta_t(l)} \tag{11}$$

이 때, 식 (7),(8),(9),(10),(11)의 파라미터를 결정하는 중간 변수 $\gamma_t(i,j),\gamma_t(i),\zeta_t(i,l),\zeta_t(l)$는 다음식에 의해 구한다.

$$\gamma_t(i,j) = \sum_{k=1}^{L}\chi_t(i,j,l), \quad 1 \le t \le T-1 \tag{12}$$

$$\gamma_t(i) = \sum_j \gamma_t(i,j), \quad 1 \le t \le T-1 \tag{13}$$

$$\zeta_t(i,l) = \sum_j \chi_{t-1}(j,i,l) \quad \text{if} 1 \le t \le T \tag{14}$$

$$\zeta_t(l) = \sum_i \begin{cases} \dfrac{\pi_i b_i(l) f(o_1 \mid v_l)\beta_1(i)}{f(O \mid \lambda)} & \text{if} \quad t=1 \\ \zeta_t(i,l), & 1 \le t \le T \end{cases} \tag{15}$$

그리고 위 식(12),(13),(14),(15)의 중간 파라미터들은 식 (16)의 $\chi_t(i,j,l)$에 의해서 결정된다.

$$\chi_t(i,j,l) = \frac{\alpha_t(i)a_{ij}b_j(l)f(o_{t+1} \mid v_l)\beta_{t+1}(j)}{f(O \mid \lambda)}, \quad 1 \le t \le T-1 \tag{16}$$

그리고 식 (10), (11)에서 구한 평균μ_l과 분산SUM $\sum_l$을 이용해 식 (6)의 가우시안 함수 $g(o_t \setminus \mu_l, \sum$를 결정할 수 있다.

Ⅲ. 제안한 방법

3.1 차감법의 현상

주파수 차감법은 처음에 음성을 검출하기 전의 몇 프레임을 초기 추정 잡음값으로 본다. 그러나 이 음성 검출이 잘못된 경우에 이후 프레임의 추정한 잡음은 많은 오류를 가지게 된다. 또한 추정 방법에서 추정한 잡음에 음성 신호가 섞어 들어가게 된다. 다음 그림은 잡음으로 부식된 "의정부"라는 단어의 음성 신호와 이것의 주파수 차감을 위해 추정한 잡음과 차감된 음성 신호의 그림이다.

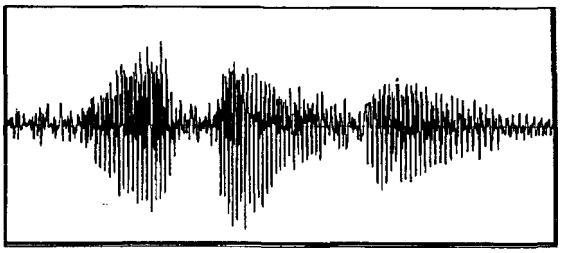

그림 1 . 잡음에 의해 부식된 "의정부" 단어
Fig. 1. Corrupted Word " Uijongbu" by Noise.

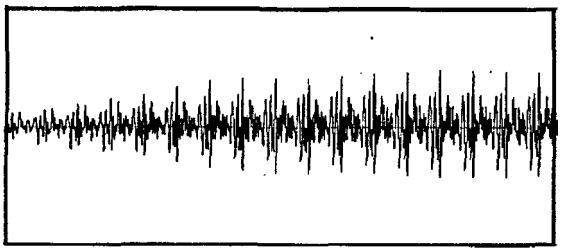

그림 2. 추정한 잡음 신호
Fig. 2. Estimated Noise Signal.

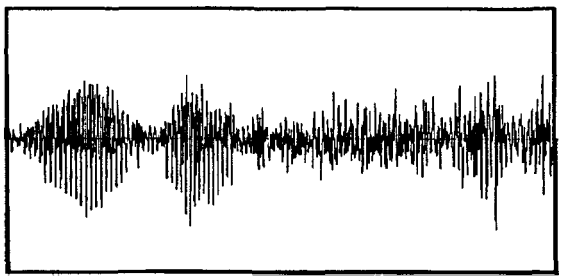

그림 3. 차감된 "의정부" 단어
Fig. 3. Subtracted Word "Uijongbu"

그림 1은 잡음에 의해 부식된 혼합 음성 신호이다. 그림 2는 추정한 잡음 신호를 나타내며 점점 추정 오류가 증가하고 있다. 이렇게 추정 잡음에 오류가 생긴 경우 차감한 음성 신호는 프레임이 진행될수록 잡음 오류로 인해 차감 하지 않은 경우보다 더 부식되는 결과를 초래한다.

그림 3에서 초기 프레임에서의 "의저～ " 까지는 차

감이 잘 되어진 음성 신호의 모양을 갖으나 이후 프레임 "-o 부"에서는 차감된 오류 잡음에 의해 더 큰 잡음으로 부식되는 모양을 나타낸다.

3.2 제안한 SNR 함수에 의한 가우시안 함수 변형 및 결정

전 절에서 보는 것과 같이 잡음에 의한 부식 정도가 어떤 프레임에서는 차감된 신호가 또 어떤 프레임에서는 처음에 혼합된 신호가 오히려 더 많은 음성 정보를 가지고 있는 경우가 있다. 이것은 그림 2에서와 같이 추정 잡음의 정도에 의해서 결정될 수 있다. 다음식은 제안한 SNR 평가 함수를 이용하여 가우시안 함수를 변형 및 결정하는 식이다.

$$SNR_i = 10\log\frac{\sum_{j}^{M}\overline{s_i^2(n-j)}}{\sum_{j}^{M}no_i^2(n-j)} \tag{17}$$

$$g(o_i \setminus \mu_l, \sum_l) = G(\overline{s_i(n)}), SNR > Threshold \tag{18}$$

$$g(o_i \setminus \mu_l, \sum_l) = aG(s_i(n)) + (1-a)G(\overline{s_i}(n)),\quad 0 \le SNR \le Threshold \tag{19}$$

$$g(o_i \setminus \mu_l, \sum_l) = G(s_i(n)),\quad SNR < 0 \tag{20}$$

식 (17)은 임의의 i 프레임에서 각 샘플에 대한 SNR 식이다. 이 SNR은 경계값 *Threshold* 를 기준으로 가우시안 함수를 결정한다. 그 경계값은 다음과 같다.

$$Threshold = a_{th}\frac{\sum_{i}^{N}\sum_{j}^{M}s_i^2(n-j)}{\sum_{i}^{N}\sum_{j}^{M}no_i^2(n-j)} \tag{21}$$

이 식은 N개의 프레임과 각 프레임에서 M 샘플에 대하여 음성 $s(n)$과 잡음 $no(n)$의 SNR을 구한 후 실험 상수 a_{th}를 곱하여 경계값을 얻는다. 또 식 (18), (19), (20)에서 $G(\overline{s_i(n)})$는 차감된 신호에서의 가우시안 함수를 의미하고, $G(s_i(n))$는 원래의 혼합된 신호에서의 가우시안 함수를 의미한다.

식 (18), (19), (20)에 의하여 각 코드북 인덱스 l에 따라 식(6)의 가우시안 함수를 대응함으로써 HMM의 관측 확률 밀도 함수를 결정한다. 즉, 식 (18)은 잡음의 크기가 신호에 비해 상대적으로 작은 경우로써 차감한 음성의 가우시안 함수로 결정하는 식이다. 식 (19)은 차감된 음성의 가우시안 함수 $G(\overline{s_i(n)})$과 부식된 음성의 가우시안 함수 $G(s_i(n))$를 비율로 하여 변형시킨다. 식 (20)은 잡음의 크기가 큰 경우이므로 부식된 음성의 가우시안 함수로 결정하는 식이다. 이렇게 변형 및 결정된 가우시안 함수에 의해 식 (6)의 관측 확률 밀도 함수가 결정된다. 이러한 가우시안 함수의 결정 및 변형 정도는 그림 4와 같이 표현할 수 있다.

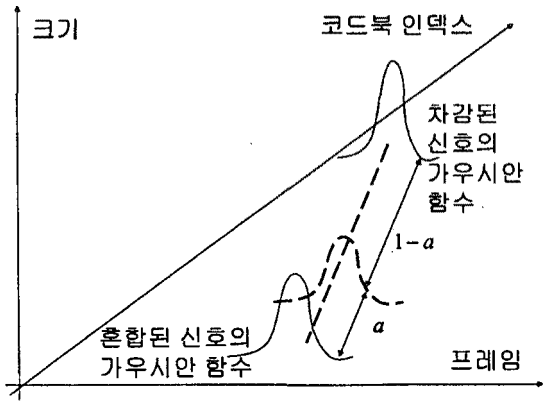

그림 4. 가우시안 함수 변형 및 결정
Fig. 4. Variation and Determination of Gaussian Function.

그림 4에서는 프레임열과 각 코드북 인덱스에서 차감된 신호와 부식된 신호와의 관계에 의해 가우시안 함수를 변형 및 결정하는 것을 보여 준다.

이러한 시스템을 설계하기 위해서는 차감 이후에 음성의 각 프레임마다 차감 신호와 추정 잡음 신호에 의한 SNR을 구한 후 가우시안을 결정하는 형태를 취한다. 다음 그림 5는 제안하는 잡음에 강한 음성 인식을 위한 전체적인 시스템을 보여준다.

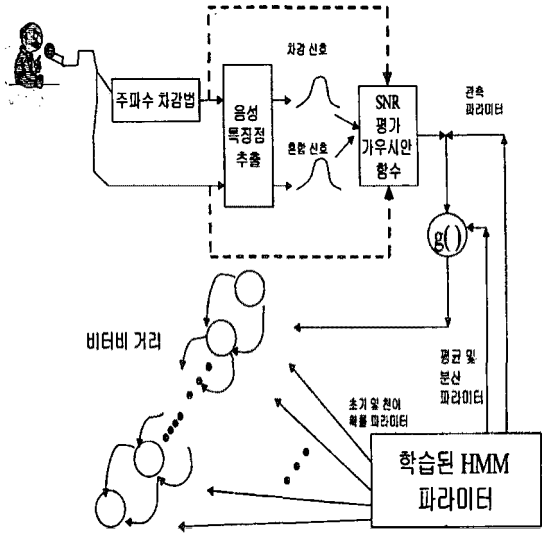

그림 5. 제안하는 전체 시스템
Fig. 5. Proposed total system.

Ⅳ. 실 험

동부 간선도로에서 창문을 어느 정도 열어놓은 상태와 많이 열어 놓은 상태에서 80 km/h 이상으로 달리는 차량내에서 받은 배경 잡음 데이터를 지역명 및 숫자음의 34 단어와 혼합하여 6 화자에 대해 화자 종속 인식 실험을 하였다. 표 1에서는 달리는 차량내에서 창문을 활짝 열어놓은 상태에서 PLP(Perceptual Linear Predictive) 켑스트럼[8]을 사용했을 때의 인식율을 보이며, 표 2에서는 표 1보다는 잡음 정도는 덜 한 상태로써 5분의 1정도 창문을 열어 놓은 환경에서 PLP 켑스트럼을 사용했을 때의 인식율을 보여준다. 또한 표 3에서는 표 2와 같은 잡음 배경 환경으로 RPS-PLP 켑스트럼[9]을 사용했을 때의 인식율을 보여준다. 또한 이 표들은 혼합된 신호, 차감한 신호 및 제안하는 가우시안 함수에 의한 인식율을 보이고 있다. 이때 반연속 HMM의 파리미터는 Baum-Welch 추정 알고리즘과 EM 알고리즘을 사용하여 10 명의 화자에 대해 학습하여 추출하였다. 그중 인식 테스트는 6명을 선택하여 실험하였다. 그리고 제안하는 함수에서 a 는 실험에 의해서 가장 높은 인식율을 보이는 값으로 0.01을 선택하였고, a_{th} 도 역시 실험에 의하여 0.125로 결정하였다.

표 1. 창문을 활짝 열어 놓고 80km/h 이상으로 달리는 차량 배경 잡음에서의 인식율 (PLP 켑스트럼)

Table 1. The recognition rates under the car noise environment over the speed of 80 km/h, of which window is opened widely(PLP Cepstrum).

화 자	혼합된 음성의 인식율	차감 이후의 인식율	가우시안 함수 변형 및 결정에 의한 인식율
화자 1	19/34	19/34	23/34
화자 2	28/34	29/34	31/34
화자 3	31/34	31/34	30/34
화자 4	16/34	21/34	23/34
화자 5	15/34	18/34	18/34
화자 6	20/34	23/34	24/34
전체	63 %	69 %	73 %

표 2. 창문을 작게 열어 놓고 80km/h 이상으로 달리는 차량 배경 잡음에서의 인식율 (PLP 켑스트럼)

Table 2. The recognition rates under the car noise environment over the speed of 80 km/h, of which window is opened narrowly (PLP Cepstrum).

화 자	혼합된 음성의 인식율	차감 이후의 인식율	가우시안 함수 변형 및 결정에 의한 인식율
화자 1	25/34	26/34	27/34
화자 2	29/34	31/34	31/34
화자 3	31/34	32/34	32/34
화자 4	18/34	24/34	24/34
화자 5	20/34	19/34	24/34
화자 6	26/34	29/34	29/34
전체	73 %	79 %	82 %

표 3. 창문을 작게 열어 놓고 80km/h 이상으로 달리는 차량 배경 잡음에서의 인식율(RPS-PLP 켑스트럼)

Table 3. The recognition rates under the car noise environment over the speed of 80 km/h, of which window is opened narrowly(RPS-PLP Cepstrum).

화 자	혼합된 음성의 인식율	차감 이후의 인식율	가우시안 함수 변형 및 결정에 의한 인식율
화자 1	27/34	29/34	29/34
화자 2	31/34	31/34	31/34
화자 3	30/34	28/34	28/34
화자 4	31/34	29/34	31/34
화자 5	28/34	30/34	33/34
화자 6	29/34	29/34	31/34
전체	86 %	86 %	89 %

Ⅴ. 고찰 및 결론

잡음에 강한 음성 인식 시스템을 설계시 인식 전에 잡음 제거를 선행한다. 그러나 이 잡음 제거기에서는 잡음뿐만 아니라 음성마저도 제거시켜 시스템의 성능에 오히려 잡음 제거기가 없을 때 보다 더 큰 저하를 가지고 올 수 있다. 본 연구는 잡음 제거기를 주파수 차감법을 사용하여 차감 후 반연속 HMM의 가우시안 함수를 차감 음성과 추정 잡음과의 관계로써 프래임 별 SNR 함수에 의해 결정하는 방법을 제안하였다. 이러한 방법은 차감 신호의 에너지와 추정 신호 에너지와의 비율에 의하여 가우시안 함수를 적응적으로 결정해 나가는 방법이다. 실험에 사용한 잡음은 달리는 차량내에 음성과 동일한 대역의 극심한 잡음 형태로써 음성과의 상관성이 많아 주파수 차감을 할 경우 음성 정보의 손실이 많은 경우이다. 표 1, 표 2 및 표 3에서의 실험 결과에서 차감된 신호가 오히려 원래 잡음에 혼합된 신호보다도 인식율이 떨어지는 경우를 볼 수 있다. 본 연구에서 제안한 방법을 적용하였을 때 전체적으로 혼합된 신호보다는 3~10%, 차감된 신호보다는 3~4% 정도의 높은 인식율을 갖음을 확인할 수 있다. 특히 RPS-PLP 켑스트럼을 사용한 경우 PLP 켑스트럼을 사용했을 때 보다 높은 인식률을 보이고 있다. 표 3에서 제안한 방법의 인식율이 낮은 경우가 한가지 있으나, 나머지 모든 화자들은 제안한 방법의 인식율이 높기 때문에 전반적으로 인식율이 증가하게 된다. 이것은 차감에 의해서 오히려 부식되는 부분을 SNR 기준으로 가우시안 함수를 결정 및 변형함으로써 손실된 음성 정보를 보상할 수 있음을 증명한 것이다.

참고 문헌

1. J. A. Nolazoco Flores, S. J. Young " Continuous Speech Recognition In Noise Using Spectral Subtraction and HMM Adaptation," ICASSP, Vol. 1, pp. 409-412 , 1994.
2. R. A., L. Chin-Hui, F. K. S., "Cepstral Channel Normalization Techniques for HMM-based Speaker Verification," ICSLP, pp.1835-1838. 1994.
3. Verga. A., Moore. R. "Hidden Markov Model Decomposition of Speech and Noise," ICASSP, pp. 845-848, 1990.
4. 전선도, 강철호, 김종찬, 김순협, "차량내 잡음 환경에서 적응적 경계값을 이용한 가중치 주파수 차감에 관한 연구," 한국음향학회지, 제17권, 제8호, pp. 73-77, 1998.
5. X. D. Haung, M. A. Jack, "Semi-Continuous Hidden Markov Models for Speech Signals," Computer Speech and Language, vol. 3, pp. 239-251, 1989.
6. X. D. Haung, Y. Ariki, M. A. Jack, "Hidden Markov Models for Speech Recognition," Edinburgh University Press, 1990.
7. B. H. Jaung, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chain," AT&T Technical Journal, Vol. 64, pp. 1235-1249, 1985.
8. H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis for Speech," J. Acoust. Soc. Am. pp. 1738-1752, 1990.
9. Jean-claude Junqua, H. Wakita. "A Comparative Study of Cepstral Lifters and Distance Measure for All Pole Models of Speech in Noise,"ICASSP, Vol.1, pp.476-479. 1989.

▲ 전 선 도 : 제 17권 8호 참조

▲ 강 철 호 : 제 17권 8호 참조