산업기술연구소 논문집, 3, 301-318, (1997), 서경대학교
Journal of Institute of Industrial Technology, 3, 301-318, (1997),
SeoKyeong University

잡음 환경하에서의 음성인식에 관한 연구

강 선 미*

요 약

음성인식기의 실용화에 따른 가장 큰 문제점은 다양한 잡음 환경하에서 그 성능이 급격히 떨어지는데 있다. 이는 음성인식기의 인식환경과 훈련환경과의 차이에 의해 야기되므로 음성을 오염시키는 환경을 그 원인에 따라서 적절하게 모델링하여 주거나 환경에 크게 영향을 받지 않는 강한 특징추출 및 유사도 계산 방법을 모색해야 한다. 본 논문에서는 배경잡음과 채널왜곡으로 오염된 환경하에서 음성인식의 성능을 향상시키기 위한 여러 가지 방법들에 대해서 소개하고 그 성능을 비교하였다.

Ⅰ. 서 론

최근 음성인식기술의 발달과 컴퓨팅 처리속도의 급격한 향상에 따라서 음성인식기의 실용화 작업이 활발히 진행되고 있다. 현재 음성인식기의 실용화 작업에서 우선적으로 고려되어야 할 조건은 주변 환경, 사용자 및 작업에 영향을 받지 않아야 하며 돌발적으로 발생되는 음성에 대한 처리가 가능해야 한다는 것이다. 본 논문에서는 실용화 단계에서 치명적인 요건이 되는 잡음 환경하에서도 강인한 음성인식기술(Environmental Robustness)에 관하여 다루고자 한다.

대다수의 음성인식기는 실험실 환경하에서는 만족할 만한 인식률을 유지하지만 실제 상황에서는 매우 낮은 인식결과를 얻게 된다. 이러한 현상은 인식이 수행되는 환경에 따라 음성인식기의 훈련환경과 인식환경이 서로 다르기 때문에 나타난다. 훈련환경과 인식환경과의 차이를 야기시키는 요인으로는 배경잡음, 마이크로폰, 통신채널, 화자의 발성형태 등을 들 수 있다.

본 논문에서는 인식환경을 훈련환경과 유사하게 만들어 줌으로써 잡음 환경하에서도 음성인식기가 그 성능을 유지하도록 하는 음성신호 전처리방법들을 살펴보고 그 성능들을 비교해보고자 한다. 2장에서는 음성인식의 개요에 대해서, 3장에서는 음성을 오염시키는 환경을 모델링하는 방법에 대하여 기술하였다. 4장에서는 음성인식을 환경에 강하도록 하는 방법들에 대하여 살펴본다. 마지막 5장에서는 결론을 맺는다.

* 서경대학교 컴퓨터과학과 전임강사

Ⅱ. 음성인식의 개요

일반적인 음성인식방법은 훈련단계와 훈련된 참조패턴을 이용하여 인식하는 두단계로 이루어진다. 훈련단계는 음성인식기가 사용되는 응용분야에 적용될 어휘를 구성하는 음성신호들을 나타내 줄 참조패턴을 학습하는 단계이다. 각 패턴은 몇 개의 음성예들을 가지고 학습되고, 어떤 템플릿의 형태나 패턴의 통계적 특성을 나타내는 모델의 형태로 저장된다. 인식단계는 임의의 음성패턴이 입력되었을 때 훈련단계에서 만들어진 참조패턴과 비교하여 입력패턴을 확인하는 과정이다. 훈련 및 인식단계에서는 다음과 같은 과정이 필요하다.

1. 특징추출(Feature Extraction)

특징추출이란 음성신호로부터 그 신호의 특징을 나타내 주는 파라미터를 추출하는 과정이다. 음성신호의 주파수 영역에서의 특징을 나타내는 power spectrum과 음성생성의 관점에서 살펴본 Vocal tract를 모델링하는 AR(AutoRegressive)모델이 음성인식에 사용된다. 음성의 power spectrum을 나타내기 위한 신호분석방법으로서는 FFT나 Filter-bank방법[1]이 주로 사용되고, AR모델을 위해서는 LPC분석방법[2]이 가장 많이 사용된다. 생물학적 분석에 의하면 귀가 음성을 인식하는 방법은 저주파영역은 상세하게 고주파영역은 개략적으로 분석하는데 이러한 분석결과를 이용하여 Bark-scale[3]이나 mel-scale[4]같은 log-scale로 주파수를 분석하여 사용하기도 한다. 음성신호는 nonstationary한 신호이기 때문에 음성신호의 분석은 프레임단위로 수행된다. 현재 음성인식알고리즘에서 가장 많이 사용되는 신호분석방법은 Cepstral Analysis[5]인데 이 방법은 여기(excitation)정보와 vocal tract의 모양정보를 분리해서 음성인식에 보다 중요한 vocal tract 정보만을 추출하는 가장 간단한 방법이다.

2. 패턴분류(Pattern Classification)

임의의 입력 음성으로 참조패턴을 학습시키기도 하며 또한 이미 학습된 참조패턴과의 유사도를 측정하는 단계이다. 패턴분류를 위해서 사용되는 알고리즘으로는 크게 DTW (Dynamic Time Warping)[6], HMM(Hidden Markov Model)[7], Neural Network[8] 등을 들 수 있다.

3. 결정(Decision)

입력음성과 참조패턴의 유사도에 따라 입력음성에 가장 가까운 참조패턴을 선택하는 단계이다. 이때 인식기의 성능에 따라서 거절 기능을 둘 수 있으며, 거절은 가장 높은 점수를 얻은 후보가 일정값 이하를 나타낼 때 인식 대상이 없다고 판단하는 것을 말한다.

Ⅲ. 잡음환경의 모델링

훈련환경과 인식환경이 차이가 날 때 음성인식기의 성능은 저하된다. 두 환경사이의 차이를 야기시키는 요인은 그림 1과 같이 크게 두가지로 분류할 수 있는데 기계소리나 다른 화자의 음성등과 같은 부가적인 잡음(additive noise)과 방안에서의 반향, 마이크로폰, 통신채널등의 선형필터링(linear filtering)을 들 수 있다.

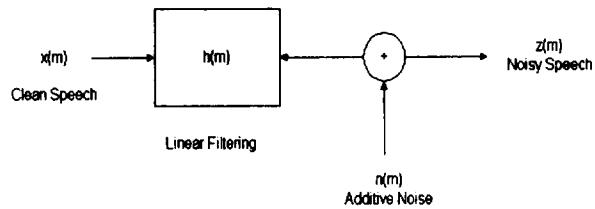

그림 28. 잡음환경의 모델링.

$$z(m) = h(m) * x(m) + n(m) \quad (1)$$

음성, $x(m)$은 먼저 convolutional한 선형필터링, $h(m)$에 의해서 왜곡되고 부가적인 잡음, $n(m)$에 의해서 오염된다고 가정한다. 식(1)을 power spectral density로 나타내면

$$P_z(f) = P_x(f)\,|H(f)|^2 + P_n(f) \quad (2)$$

이다. 식(2)를 음성인식의 특징벡터로서 가장 많이 쓰이는 켑스트럼으로 나타내면

$$z = x + q + r(x, n, q) \quad (3)$$

이다. 여기서 x, n, z, q는 각각 $x(m)$, $n(m)$, $z(m)$, $h(m)$의 켑스트럼 벡터이고 $r(x,n,q) = IDFT(\ln(1+e^{DFT(n-q-x)}))$이다. q는 선형필터링의 영향을 나타내고 $r(x,n,q)$는 부가적인 잡음의 영향을 나타낸다. 즉 켑스트럼영역에서는 선형필터링에 의한 영향과 부가적인 잡음에 의한 영향 둘 다 음성에 부가적으로 나타내어진다.

Ⅳ. 잡음환경에 강한 음성인식

음성인식기의 성능을 저하시키는 환경을 그림1과 같이 모델링하면 음성인식기가 환경에 강인하도록 하기 위해서는 부가적인 잡음과 선형필터링 요인을 제거해야 한다. 환경에 강한 음성인식기를 구현하는 방법은 음성강화(Speech Enhancement), 잡음에 강한 특징추출(Robust Feature Extraction), 잡음에 강한 거리측정(Robust Distance Measure), 모델에 기반을 둔 보상방법(Model-based Compensation) 을 들 수 있다.

1. 음성강화(Speech Enhancement)

음성강화란 배경잡음으로 오염된 음성에서 부가잡음을 제거하고 음성의 질이나 명료도(Intelligibility)를 향상시키는 알고리즘을 말한다. 음성강화 알고리즘은 배경잡음의 형태, 음성과의 관계(additive or convolutional), 채널의 수에 따라 달라질 수 있는데 대부분의 음성강화 알고리즘은 다음과 같은 가정을 전제로 한다.

① 음성과 배경잡음은 서로 상관이 없다(uncorrelated).
② 배경잡음은 음성에 비해 stationary하다.
③ 배경잡음은 음성에 부가적이다.

음성강화 알고리즘에는 스펙트럼 크기의 예측에 의한 방법(Spectral Subtraction [9][10], MMSE Spectral Estimator[11][12]), Iterative Wiener filtering[13], Adaptive Noise Cancelling[14], Microphone Array[15]등이 있다. 이 중에서 일반적으로 가장 많이 사용되는 방법인 스펙트럼 크기 예측방법을 소개하겠다.

1.1 스펙트랄 공제법(Spectral Subtraction)[9][10]

스펙트랄 공제법은 음성신호에 포함된 부가잡음을 제거하는 알고리즘으로서 오염음성의 크기 스펙트럼(magnitude spectrum)이나 파워 스펙트럼에서 잡음의 크기나 파워 스펙트럼을 제함으로써 순수음성의 크기나 파워 스펙트럼을 예측한다. 이때 위상(phase)은 음성의 질에 상대적으로 중요하지 않기 때문에 순수음성의 위상 스펙트럼(phase spectrum)은 오염음성의 위상스펙트럼을 그대로 사용한다. 만일 잡음 $n(m)$이 음성 $x(m)$에 부가적이라면 오염음성 $z(m)$은 식(4)와 같이 표현할 수 있다.

$$z(m) = x(m) + n(m) \tag{4}$$

이때 순수음성의 크기스펙트럼의 예상치, $|X'(\omega)|$는

$$|X'(\omega)| = |Z(\omega)| - |N'(\omega)| \tag{5}$$

가 된다. $|Z(\omega)|$는 오염음성의 스펙트럼 크기이고, $|N'(\omega)|$는 부가잡음의 스펙트럼 크기의 예상치로서 비음성구간 동안에 예측된다. 그러므로 순수음성의 예상치, $X'(\omega)$는

$$X'(\omega) = |X'(\omega)| \angle Z(\omega) \tag{6}$$

이 된다. 스펙트랄 공제법이 가지는 단점은 순수음성의 크기 스펙트럼이 음의 값을 가진 경우 최소값으로 처리(flooring)하기 때문에 musical noise을 야기시킨다. 그리고 부가잡음을 제거시켜 SNR을 증가시키지만 광대역 부가잡음의 경우는 음성의 질이나 명료도를 향상시키지 못한다.

1.2 MMSE(Minimum Mean Square Error) Spectral Estimator[11][12]

MMSE spectral estimator에서는 음성과 잡음의 Fourier expansion 계수들을 각각 통계적으로 독립적이고 평균이 0인 Gaussian 확률변수로 가정한다. X_k, D_k, Z_k를

각각 $x(m), d(m), z(m)$의 k번째 스펙트럴 요소라고 하면 예를 들어 Z_k는

$$Z_k = \frac{1}{T}\int_0^T z(m)\ \exp(-j\frac{2\pi}{T}km)dm \tag{7}$$

이다. 여기서 T는 프레임의 길이이다.

오염음성의 크기 스펙트럼이 주어질 때 순수음성의 크기 스펙트럼,$|X_k|$를 MMSE의 관점에서 예측한다.

$$|X'_k| = E[|X_k||Y_k|] \tag{8}$$

위와 같은 MMSE방법은 스펙트럼의 오차를 줄이기 위해서 전력 스펙트럼영역이 아니라 인식과 더 관련이 있는 log-spectral영역에서 수행되기도 한다.[12]

2. 잡음에 강한 특징추출(Robust Feature Extraction)

잡음에 강한 특징추출방법에서 대표적인 것으로 MFCC(Mel-Frequency Cepstral Coefficient), PLP(Perceptual Linear Prediction), SMC(Short-time Modified Coherence)를 들 수 있다. 음성의 동적특성이 정적특성에 비해서 잡음에 의한 영향을 덜 받기 때문에 정적특성과 더불어 동적특성도 고려된다. 동적특성으로서는 연속프레임간의 변화률을 나타내는 1차 동적특성과 가속을 나타내는 2차 동적특성이 주로 사용된다[19]. 훈련환경과 인식환경의 차이를 야기시키는 부가잡음과 선형필터링의 효과를 제거하기 위해서 켑스트럼 영역에서 수행되는 켑스트럼 보상방법[20][21]도 사용한다.

2.1 MFCC(Mel-Frequency Cepstral Coefficient)[16]

켑스트럼이 LP(Linear Parameter)보다 음성인식기의 특징으로 우수하며, 인간의 청각특성을 고려한 Mel-켑스트럼이 LPC 켑스트럼 보다 잡음이 없는 환경에서 뿐만 아니라 잡음 환경하에서도 성능이 우수하다. 이는 인간의 청각기관이 스펙트럼을 비선형적인 주파수 스케일(log-scale)로 분석하기 때문이다. MFCC는 이러한 청각기관을 모델링하여 음성신호의 스펙트럼을 mel-scale상태에서 동일한 간격을 갖는 필터뱅크로 분석한다. Mel은 tone의 인지된 피치 또는 주파수를 측정하는 단위이다. Mel-주파수와 물리적인 주파수는 식(9)에 의해서 얻어진다.

$$F_{mel} = 2595\ \log_{10}(1+\frac{f}{700}) \tag{9}$$

MFCC를 구하는 과정은 그림2과 같이 먼저 음성신호의 고주파수영역을 강조하기 위해서 Preemphasis를 수행한 다음, 프레임단위 해석을 위해 윈도우를 취하는데 보통 프레임의 시작과 끝에서의 연속성을 유지하기 위해 Hamming Window를 사용한다. 각 프레임마다 DFT를 취해 전력스펙트럼을 얻은 후 mel-scale의 주파수 스케일로 필터뱅크를 취하여 얻은 값에 로그를 취하고 DCT(Discrete Cosine Transform)를 사용

하여 MFCC를 구한다.

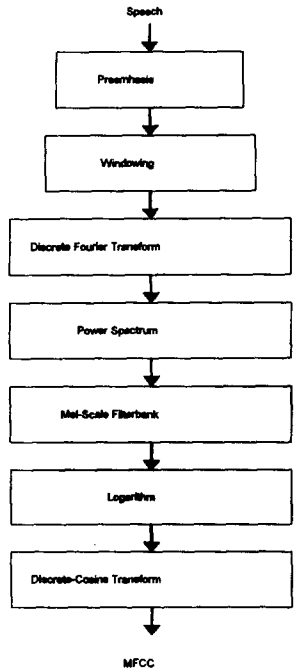

그림 2. MFCC 특징추출과정

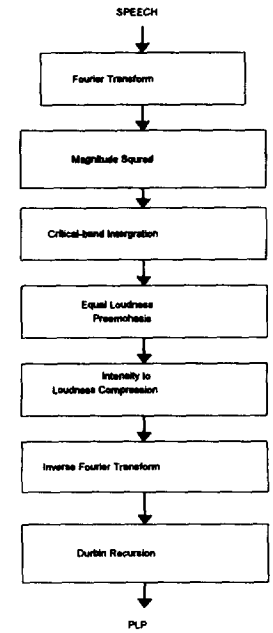

그림 3. PLP 특징추출과정

2.2 PLP(Perceptual Linear Prediction)[17]

PLP 특징추출방법은 자기상관 선형예측방법을 사용하는 all-pole 함수로 청각기관을 모델링하는 방법이다. PLP를 구하는 과정은 그림3과 같은데 Critical-band intergration, Equal-loudness preemphasis, Intensity-loudness compression등의 세가지 방법을 사용하여 인간의 청각기관을 모델링한다는 점에서 일반적인 LP방법과는 차이가 있다. PLP방법에 의해서 얻어지는 특징은 LP방법에 의해서 얻어지는 특징보다 유연한(smooth) 특성을 갖고 LP보다 더 낮은 차원을 갖는다. PLP방법의 단점은 Fourier Transform이후에 Critical band intergration을 수행하기 때문에 저주파수영역에서는 좋은 주파수 resolution을 고주파수에서는 좋은 시간 resolution을 갖는 인간의 청각시스템의 특성을 모델링하지 못하고 일정한 주파수 resolution을 갖는다.

2.3 SMC(Short-time Modified Coherence)[18]

All-pole 파라미터는 신호가 잡음에 의해 변형될 수 있기 때문에 신호로부터 직접 얻는 것보다는 자기상관(autocorrelation)함수로부터 얻는 것이 잡음에 더 강하다는 주

장에서 비롯되었다. 자기상관영역에서는 두 개의 인접한 프레임사이에서의 결합력(Coherence) 때문에 잡음에 강한 특성이 얻어진다. 이렇게 자기상관함수로부터 All-pole파라미터를 얻는 방법을 SMC라고 부른다. SMC는 일반적인 LP방법보다 프레임의 길이와 피치주기사이의 상호영향을 줄여주기 때문에 근접 프레임간의 변이성을 줄여주는 특징이 있다. 그리고 SMC방법은 부가적인 백색잡음에 강한 것으로 나타났다.

2.4 켑스트럼 보상(Cepstral Compensation)

켑스트럼 보상이란 그림1과 같은 잡음환경모델을 가정하고 오염음성의 켑스트럼, z가 주어질 때 부가잡음의 효과인 $r(x,n,q)$와 선형필터링의 효과인 q를 예측한 다음, 켑스트럼 영역에서 적절한 역연산을 가하여 순수음성(clean speech)의 켑스트럼, x을 예측하는 방법으로서 크게 cepstral highpass filtering과 empirical cepstral compensation방법이 있다.

(1) Cepstral highpass filtering

선형필터링의 영향을 제거하는데 효과적인 방법으로서 선형필터링의 영향이 음성에 비해 천천히 변하다는 가정하에 오염음성의 켑스트럼에 highpass필터링을 가함으로써 선형필터링의 영향을 제거한다. 즉 훈련환경과 인식환경에서의 켑스트럼계수의 평균값을 0으로 함으로써 두 환경사이의 평균 켑스트럼을 같게 만든다. RASTA(RelAtiveSpecTrAl)[20][21]와 CMN(Cepstral Mean Normalization)[22]등이 대표적인 방법이다.

RASTA는 음성의 켑스트럼 또는 로그 스펙트럼에 다음과 같은 전달함수를 갖는 대역통과필터를 가한다.

$$H(z) = \frac{z^4(0.2+0.1\,z^{-1}-0.1\,z^{-3}-0.2\,z^{-4})}{1-0.94\,z^{-1}} \qquad (10)$$

이 전달함수는 0.26Hz의 low cut-off frequency를 갖고 12.8Hz에서부터 전달함수의 기울기가 감소하기 시작하여 28.9Hz와 50Hz에서 0이 되는 모양을 갖는데, 전달함수의 highpass부분은 음성에 비해 천천히 변하는 선형필터링의 영향을 제거하고 lowpass부분은 프레임간의 빠른 변화성분을 제거한다.

RASTA 필터는 훈련환경과 인식환경이 비슷할 때 인식률을 향상시키지만 묵음구간에서 음성구간으로 천이할 때 천천히 변하는 주파수 특성을 감소시키지 못하므로 두 환경이 비슷할때는 인식률이 저하되는 현상이 있다. 또한 선형필터링의 효과는 제거하지만 부가적인 잡음을 제거하지 못한다. 그래서 선형필터링의 효과와 부가적인 잡음의 효과를 동시에 제거하기 위해서 RASTA의 변형형태인 Lin-Log RASTA[21]라는 방법이 제안되었다. 부가적인 잡음을 제거하기 위해서 스펙트럼상에서 highpass

filter를 수행하는 방법과 유사하게 스펙트럼 영역에서 부가적인 배경잡음과 동시에 로그 스펙트럼상에서 부가적인 채널왜곡을 필터링한다. 이를 위해서 Lin-Log RASTA에서는 RASTA처리에 로그변환을 $y=\ln(1+Jx)$변환으로 대치한다. 여기서 J는 신호에 의존하는 상수이고, $J \ll 1$이면 선형변환으로, $J \gg 1$이면 로그변환으로 근사화된다.

CMN는 선형필터링의 영향을 줄이기 위한 방법으로서 순수음성의 켑스트럼벡터의 평균은 0이거나 상수이다 라는 가정하에 한 발성(utterance) 전체의 켑스트럼벡터의 평균을 구한 다음, 이 평균을 각 켑스트럼계수에서 제하는 방법이다. 만일 그림1에서 배경잡음에 의한 왜곡은 없고 채널왜곡만 존재한다면, $n(m)=0$이 된다. $n(m)=0$로 하고 식(1)의 켑스트럼을 구하면 식(11)과 같다.

$$C_z(n) = C_x(n) + C_h(n) \tag{11}$$

여기서 $C_z(n)$, $C_x(n)$, $C_h(n)$은 오염음성 $z(m)$, 순수음성 $x(m)$, 선형필터링 $h(m)$의 켑스트럼이다. 선형필터링의 켑스트럼의 예측값, C_h'는 한 발성 전체의 켑스트럼의 평균을 구함으로써 식(12)에서 얻을 수 있다.

$$C_h' = \frac{1}{L}\sum_{i=0}^{L} C_z[i] \tag{12}$$

이 선형필터링의 예측값을 오염음성의 켑스트럼에서 제함으로써 순수음성 켑스트럼의 예측값을 식(13)에서 구할 수 있다.

$$C_s'[n] = C_z[n] - C_h' \tag{13}$$

CMN방법 역시 RASTA와 마찬가지로 훈련환경과 인식환경이 서로 차이가 날 때 인식률의 향상을 기대할 수 있다.

(2) Empirical cepstral compensation

Empirical cepstral compensation방법은 훈련환경과 인식환경의 차이를 보상하기 위해서 두 환경에서 동시에 기록된 음성 DB(stereo-recorded data)를 필요로 한다. 먼저 훈련환경과 인식환경의 음성 켑스트럼 벡터를 구한 다음, 두 환경사이의 평균 켑스트럼의 차이를 계산함으로써 보상벡터(compensation vector), v를 구한다.

$$v = \bar{x} - \bar{z} \tag{14}$$

이 보상벡터를 가지고 입력음성의 켑스트럼에 부가적인 보상을 가하여 순수음성의 켑스트럼의 예측값, x'을 구한다.

$$x' = z + v \tag{15}$$

이 때 보상벡터가 신호의 SNR, VQ코드북의 index, 음소, 인식환경등의 의존여부에

따라 여러가지 보상방법이 있다.

이 보상방법은 선형필터링의 영향뿐만아니라 부가잡음도 동시에 제거한다. SNR이 큰 경우 부가잡음의 효과인 $r(x, n, q)$가 0이므로 이 보상벡터는 주로 선형필터링의 영향을 보상하고, 반면에 SNR이 작은 경우 부가잡음의 효과가 크므로 주로 부가잡음의 영향을 보상한다. 중간정도의 SNR의 경우에는 부가잡음과 선형필터링의 영향을 동시에 보상한다. 환경에 의존하는 보상방법과 인식기를 훈련시킬 때 훈련환경에 대한 사전지식을 필요로 하지 않는 환경에 의존하지 않는 보상방법으로 나눌 수 있다. 환경에 의존하는 보상방법으로는 SDCN, FCDCN, PDCN, RATZ가 있고, 환경에 의존하지 않는 보상방법으로는 BSDCN, MFCDCN, IFCDCN, IPDCN, Blind RATZ가 있다.

SDCN(SNR Dependent Cepstral Normalization)[22][23]에서는 보상벡터, $w(SNR)$가 음성의 SNR, 즉 $z[0]-x[0]$에만 의존한다. 순수음성 켑스트럼의 예측값, x'는 식(16)과 같이 얻어진다.

$$x' = z - w(SNR) \tag{16}$$

음성의 각 프레임은 SNR값에 따라 다시 나누어지고 주어진 SNR에 해당하는 보상벡터는 훈련환경과 인식환경의 켑스트럼 차이에 의해서 예측된다. 임의로 새로운 인식음성이 입력되었을 때 입력음성의 각 프레임의 SNR을 계산한 다음, 그 SNR에 따라 보상벡터가 선택되고 이 보상벡터를 이용해 입력음성을 보상한다.

FCDCN(Fixed Codeword-Dependent Cepstral Normalization)[22][24]에서는 보상벡터, $r(k, SNR)$는 음성의 SNR 뿐만아니라 VQ의 코드북에 의존한다. 즉 순수음성의 예측값, x'는 다음과 같이 얻어진다.

$$x' = z + r(k, SNR) \tag{17}$$

여기서 k는 VQ코드북의 index이다. 적절한 코드북은 VQ단계에서 식(17)을 최소화하도록 선택된다.

$$\| z + r(k, SNR) - c(k) \|^2 \tag{18}$$

여기서 $c(k)$는 훈련음성 DB로 구성된 VQ코드북을 나타낸다.

FCDCN의 보상벡터를 구하기 위해서는 먼저 SNR에 따라 음성의 각 프레임을 구분하고 인식환경의 각 SNR에 따라 켑스트럼을 VQ하여 나눈 다음, 각 SNR의 VQ index에 해당되는 보상벡터를 예측한다. 순수 음성켑스트럼, x의 분포를 가우시안 분포의 혼합으로 가정하고 식(19)에 EM(Expectation Maximization)을 사용하여 보상벡터를 예측하기 위해 FCDCN을 훈련시킨다.

$$p(x) = \sum_{k=0}^{K} P[k]\, N_x(c(k), \Sigma_k) \tag{19}$$

오염음성 켑스트럼, z의 분포는 다음과 같이 분산이 SNR과 VQ 코드북에 의존하는

가우시안 분포로 모델링한다.

$$p(z|k, r, SNR) = \frac{C}{\sigma(SNR)} \exp(-\frac{1}{2\sigma^2} ||z + r(k, SNR) - c(k)||^2) \quad (20)$$

FCDCN을 훈련시키는 EM 알고리즘은 다음과 같이 반복적으로 수행된다.

① $r(k, SNR)$과 $\sigma^2(SNR)$의 초기값을 가정한다.

② Estimatioo: $r(k, SNR)$, $\sigma^2(SNR)$와 $c(k)$가 주어질 때 식(19)의 파라미터를 예측한다.

③ Maximization: $r(k, SNR)$와 $\sigma^2(SNR)$을 예측한다.

④ 수렴이 될 때까지 ②, ③의 과정을 반복한다.

PDCN(Phone-Dependent Cepstral Normalization)[22][25]는 보상벡터가 SNR이나 VQ의 index가 아닌 음소에 의존한다. 보상벡터를 얻기 위해서 PDCN을 훈련시키는 과정은 다음과 같다. 먼저 훈련환경과 인식환경의 음성을 음소에 따라 구분하고 각 음소에 대해서 프레임별로 훈련환경과 인식환경의 음성의 켑스트럼 차이를 구한다. 이렇게 구한 켑스트럼 차이를 가지고 식(21)과 같이 보상벡터를 구한다.

$$c(p) = \frac{\sum_{u=1}^{A} \sum_{t=1}^{T} (x_t - z_t)\delta(f_t - p)}{\sum_{u=1}^{A} \sum_{t=1}^{T} \delta(f_t - p)} \quad (21)$$

여기서 A는 발성의 수를 나타내고 f_t 는 한 발성의 t번째 프레임의 음소, p는 음소의 인덱스이고 T_u 는 프레임의 길이이다. 이 때 순수음성 켑스트럼의 예상치, x_t' 는

$$x_t' = z_t + c(p) \quad (22)$$

이 되고 프레임단위로 구해진다.

RATZ(Multivariate-Gaussian-Based Cepstral Normalization)[26]에서는 음성 켑스트럼의 분포를 multivariate-Gaussian mixture로 모델링한다.

$$x = [x_0 ..... x_p]^T$$

$$p(x) = \sum_{k=0}^{M-1} P[k] N_x(\mu_{x,k}, \Sigma_{x,k}) \quad (23)$$

여기서 $P[k]$, $\mu_{x,k}$, $\Sigma_{x,k}$는 k번째 가우시안 혼합의 평균벡터, covariance matrix, 사전확률값들이다. 부가잡음과 선형필터링등의 환경에 의한 음성에 대한 영향을 통계적인 관점에서 보면 다음과 같이 순수음성 켑스트럼 평균의 이동과 분산의 확장 또는 축소로 나타낸다.

$$\mu_{z,k} = \mu_{xk} + r_k \quad (24)$$

$$\Sigma_{z,k} = \Sigma_{x,k} + R_k \quad (25)$$

여기서 $\mu_{z,k}$, $\sum_{z,k}$는 오염음성의 평균벡터와 covariance matrix이다. r_k, R_k는 평균의 이동과 분산의 축소 또는 확장을 나타내는 요소이다. 다음과 같은 과정을 통해서 r_k, R_k를 예측한 다음, 켑스트럼을 보상한다.

① 순수음성의 통계를 예측: covariance matrix를 diagonal이라고 가정하고 $P[k]$, $\mu_{z,k}$, $\sum_{z,k}$를 EM알고리즘을 사용하여 예측한다.

② 오염음성의 통계 (r_k, R_k)를 예측

$$r_k = \frac{\sum_{i=0}^{N-1}(z_i - x_i)\,P(k|x_i)}{\sum_{i=0}^{N-1}P(k|x_i)} \tag{26}$$

$$R_k = \frac{\sum_{i=0}^{N-1}(z_i - \mu_{z,k})(z_i - \mu_{z,k})^T\,P(k|x_i)}{\sum_{i=0}^{N-1}P(k|x_i)} \tag{27}$$

③ 오염음성의 보상: r_k, R_k를 사용해서 순수음성의 기대값이 최대화되도록 MMSE (Minimum Mean Square Estimation)을 수행해서 순수음성을 예측한다.

$$x' = z - \sum_{k=0}^{M-1} P(k|z)\,r_k \tag{28}$$

BSDCN(Blind SDCN)[27]은 각 SNR에서의 모든 데이타를 총괄함으로써 인식환경에 대한 사전지식을 피한다. 훈련환경과 인식환경에서의 SNR 사이의 대응은 두 환경 각각의 SNR의 histogram에 nonlinear warping을 사용함으로서 확립된다. 일단 훈련환경과 인식환경사이의 대응이 확립되면 보상벡터는 SDCN처럼 두 환경에서의 각 SNR에서의 평균 켑스트럼의 차이를 계산함으로써 얻어진다.

MFCDCN(Mutiple FCDCN)[25][28]는 FCDCN처럼 보상벡터가 신호의 SNR과 VQ코드북의 위치에 의존한다. 인식환경에 대한 사전지식의 필요를 피하기 위해 여러 환경에서 계산된 보상벡터를 구한 다음, 최소한의 VQ왜곡을 하는 보상환경을 선택한다. 선택된 환경에 대한 보상벡터를 사용하여 식(29)와 같이 프레임단위로 보상을 수행한다.

$$x' = z + r(k, SNR, e) \tag{29}$$

여기서 e는 먼저 선택된 환경을 나타내고 나머지는 FCDCN과 같다. 즉 MFCDCN은 FCDCN방법에다 환경을 결정하는 과정이 하나 첨가된 형태이다. 환경을 선택하는 방법에는 전체 발성에 대한 VQ의 평균왜곡을 최소화하는 환경을 선택하는 방법, 각각 훈련환경에 대한 VQ코드북을 만든 다음, 인식음성을 VQ코드북으로 VQ를 수행하여

최소의 VQ왜곡을 갖는 환경을 선택하는 방법으로 각각의 환경을 가우시안 분포의 혼합으로 가정하고 최상의 확률을 갖는 환경을 선택하는 방법등이 있다.

IFCDCN(Interpolated FCDCN)[25]은 MFCDCN에서 보상벡터들을 구하기 위해 사용했던 환경들이 인식환경과 유사하지 않을때는 MFCDCN에서 처럼 정확하지 못한 하나의 환경에서 구한 보상벡터를 사용하는 대신에 여러 환경에서 구한 보상벡터를 보간하여 사용하는 방법이다. IFCDCN에서 새로운 인식환경에 대한 보상벡터, $r(k, SNR)$ 는 다음과 같이 MFCDCN의 몇 개의 보상벡터들, $r(k, SNR, e)$을 선형보간함으로써 얻을 수 있다.

$$r(k, SNR) = \sum_{e=1}^{E} f_e \cdot r(k, SNR, e) \tag{30}$$

여기서 f_e 는 e번째 환경에 대한 가중치인데 VQ 왜곡에 의존하여 얻어진다.

$$f_e = \frac{\exp(D_e/(2\sigma^2))}{\sum_{j=1}^{E}\exp(D_j/(2\sigma^2))} \tag{31}$$

D_e 는 e번째 환경의 VQ 왜곡을 나타낸다.

IPDCN(Interpolated PDCN)[25][28]에서는 IFCDCN 처럼 미리 계산된 PDCN의 보상벡터들의 ensemble에 기초해서 새로운 인식환경에 대한 보상벡터는 그 인식환경과 가장 가까운 몇 개의 PDCN 보상벡터들을 선형보간함으로써 구한다. 보간하는 방법은 IFCDCN과 같은 방법으로 수행된다.

Blind RATZ[26][29]는 인식환경에 대한 사전지식 없이 수행하기 위해 RATZ의 오염음성의 통계를 예측하는 단계에서 $P(k|x_i)$대신에 $P(k|z_i)$를 사용하고 EM알고리즘을 사용하여 수렴될 때까지 반복적으로 수행한다.

$$r_k^{l+1} = \frac{\sum_{i=0}^{N-1} z_i P^l(k|z_i)}{\sum_{i=0}^{N-1} P^l(k|z_i)} - \mu_{x,k} \tag{32}$$

$$R_k^{l+1} = \frac{\sum_{i=0}^{N-1}(z_i - r_k^l - \mu_{x,k})(z_i - r_k^l - \mu_{x,k})^T P^l(k|z_i)}{\sum_{i=0}^{N-1} P^l(k|z_i)} - \Sigma_{x,k} \tag{33}$$

3. 잡음에 강인한 거리측정(Robust Distance Measure)

지금까지 음성인식을 위한 특징벡터로서 켑스트럼 벡터를 사용하는 것이 두드러지고 인식률을 향상시키기 위해 켑스트럼 계수에 가중치를 가해 거리측정을 하는

weighted cepstral distance measure방법이 널리 연구되어 왔다[30][31]. 켑스트럼에 가중치를 두는 것은 잡음이 없는 경우 음성의 인식률을 향상시킬뿐 아니라, 잡음 환경하에서도 음성인식기의 인식률을 향상시킨다. 한 예로서 Juang et al.[31]은 켑스트럼계수를 raised sine function을 가지고 가중치를 주는 bandpass lifering이라는 방법을 제안하였다. 큰 분산값을 갖는 하위차수의 켑스트럼계수와 고차의 켑스트럼계수의 영향을 줄이는 것은 인식시스템의 분별력을 향상시켜주는 결과를 가져온다.

Mansour et al.[32]은 단지 두 벡터의 거리(norm)보다는 두 벡터사이의 각을 고려하는 거리측정방법을 제안하였다. 이 방법은 부가적인 백색잡음이 켑스트럼 벡터들의 거리를 줄이고 각도편차가 부가적인 백색잡음에 아주 강인하다는 것을 보였다.

Soong et al.[33]는 frequency-weighted 거리측정방법을 제안하였다. SNR이 높은 경우에는 가중치를 가하지 않은 방법과 같은 성능을 유지하지만, SNR이 작아질수록 가중치를 가하지 않은 방법에 비해 우수한 성능이 나타난다. 그 이유는 비선형 스펙트럼 가중치를 주는것과 가중치 요소의 적응 조정(adaptive adjustment)라는 주파수-가중치 거리측정방법의 특징 때문이다. 이와 비슷한 방법들이 Nocerino et al.[34]과 Noda[35]에 의해서 제안되었다.

3.1 Bandpass Liftering을 사용한 방법

왜곡에 대한 측정을 speech spectral envelope가 윈도우 위치변이(window position fluctuations), 여기간섭(excitation interference), 측정잡음(measurement noise) 등에 의해 통계적으로 변화하는데 근거를 둔 방법이다. 바람직하지 않은 스펙트랄 측정변화가 적절한 신호처리 기술에 의해 부분적으로 제어될 수 있다. 특히, bandpass "liftering" 과정은 LPC based spectral measurement의 통계적 요소들의 변화를 줄여주므로 이러한 liftering과정을 음성인식에 응용한다.

3.2 Projection Operation을 기반으로 한 Distortion Measures의 집합을 사용한 방법

Linear prediction 계수로부터 유도되는 음성 켑스트럼에 초점을 맞춘 방법으로서 오염된 음성에 대한 일반적인 모델을 사용하여 부가적인 백색잡음이 LPC 켑스트럼벡터의 길이를 감소시킨다는 것을 보인다. 파라미터 히스토그램에 대한 관찰을 통해 잡음모델을 통한 분석적인 결과를 확신할 수 있고, 주어진 SNR에서 큰 값을 가지는 켑스트럼벡터의 길이 감소는 작은 길이를 가지는 켑스트럼벡터의 길이 감소보다 작다는 것을 알 수 있다. 그리고 낮은 차수의 계수가 높은 차수의 계수보다 더 많이 영향을 받으며, 더욱이 켑스트럼벡터의 orientation이 벡터길이보다 잡음교란(perturbation)에 덜 민감한 것이 관찰되었다. 그러므로 두 개의 켑스트럼벡터사이의 투영에 근거한 family of distortion measures가 제안된다. 새로운 측정법들은 bandpass cepstral distortion 측정시 같은 computational 효율을 가진다. 이러한 측정법들의 효율성에 대한 계산은 화자종속과 화자독립 고립어 인식작업에 실험되었다.

실험결과를 살펴보면 새로운 측정법이 높은 SNR 조건에서는 degradation이 없지만, 잡음 환경하에서 clean reference template를 가지고 실험했을 때, 월등한 결과를 나타내었다.
SNR이 5dB 일 때, 새로운 측정법은 20dB 이상의 SNR에서의 bandpass cepstral 측정법과 동등한 인식률을 나타내었다.

4. 모델에 기반을 둔 보상방법(Model-based Compensation)

훈련환경과 인식환경사이의 차이를 통계적인 모델로 특징화하는 방법들을 말하는데 오염된 음성을 특징화하기 위해 음성으로부터 얻은 부가잡음에 대한 지식을 사용해서 순수음성으로 훈련된 음소모델의 평균이나 분산을 변환하는 방법인 HMM 분리(decomposition) [36]과 부가잡음 뿐만아니라 선형필터링의 영향도 제거하기 위해서 상기 방법을 확장한 PMC(Parallel Model Compensation)[37]방법등이 있다.

4.1 HMM decomposition[36]

HMM을 사용한 신호분리방법으로서 동시에 생성되는 신호들을 모델링하기 위해서 parallel HMM이 사용되고 혼합신호는 이들 HMM의 출력함수로 모델링된다. 음성과 잡음을 분리하는데 적용할 수 있다. 이 방법의 가장 큰 장점은 잡음은 분리된 HMM을 사용하여 모델링될 수 있기 때문에 다양한 형태의 잡음을 다룰수 있다는 점이다. 또한 혼합신호는 특정한 형태의 신호결합이라고 가정하지 않기 때문에 부가적인 잡음뿐만아니라 convolutional한 잡음도 다룰 수 있다. 그 외에도 spectral subtraction 잡음파워의 분산이 0이라고 가정하지 않아도 된다는 점을 들 수 있다.

4.2 PMC(Parallel Model Compensation)[37]

HMM을 기반으로 하는 음성인식시스템의 성능은 훈련환경과 인식환경 사이의 부정합(mismatching)이 증가함에 따라 급속히 저하된다. 따라서 이러한 부정합을 보상(compensation)하기 위한 모델 파라미터에 대한 수정이 요구되는데 그 중의 한가지 방법이 PMC이다. PMC는 훈련환경과 인식환경 사이에 부정합이 나타나지 않을 때 음성인식시스템이 최적의 성능을 보인다는 착상에서 출발했다. 특히 PMC는 간섭부가잡음(Interfering additive noise)이 있는 경우를 고려한다. 부가적인 잡음(additive noise)이 부정합성에 나타나는 영향을 알 수 있다면 새로운 테스트 환경에 정합(matching)시키기 위해서 훈련 데이터를 수정하거나 재훈련 시킬 수 있을 것이다. 하지만 이 작업은 환경이 바뀔 때 마다 전체 데이터베이스를 저장하고 수정해야하기 위한 많은 양의 계산과 충분한 시간을 필요로 하기 때문에 훈련 데이터를 다루기 쉬운 형태로 압축시킬 필요가 있다.

이를 위한 방법으로 훈련 데이터로부터 유도된 통계적 특성을 이용한다. 훈련 데이터로부터 얻은 모든 정보를 통계적 특성으로 표현하여 부정합성을 정확하게 모델링한다면 같은 환경하에서의 훈련과 인식에 있어서의 성능저하는 없게 된다. 이러한 통계

적 특성을 알고 있다고 가정한다면 실제 관찰(actual observation)대신 통계적 특성을 사용하기 위해서 표준 HMM 재추정 식(re-estimation formulae)를 수정 할 필요가 있다. 또한 최적화 기준을 선택하기 위해서 ML(Maximum Likelihood) 추정법이 사용된다. 잡음 보상된 음성모델의 ML 추정을 얻기 위해서는 오염음성의 평균과 공분산을 예측하는 것이 필요하다. 또한 훈련 환경과 인식환경 사이의 부정합 함수(mismatching function)을 정의하는 것이 필요한다. 이를 위해서 다음과 같은 가정을 한다. 음성과 잡음은 서로 독립적이다. 음성과 잡음은 선형 도메인에서 부가적이다. 즉, 스펙트럼 추정에 있어 충분한 스무딩(smoothing)이 존재하므로 음성과 잡음은 전력 스펙트럼 레벨에서 부가적일 것이다. 단일 Gaussian 또는 Gaussian mixture set은 로그 도메인에서 관찰벡터의 분포를 표현하기에 충분한 정보를 갖는다. 프레임 상태의 할당은 부가 잡음에 의해 변하지 않는다.

부정합 함수는 static, delta 파라미터 모두에 대해서 표현된다. Static 파라미터에 대해서는 산술 적분과 고도의 계산적 효율을 갖는 근사법이 사용되며, 이 모두가 잡음 환경에서의 훈련에 있어 상당한 인식 성능을 가져오는 것이 실험을 통해 입증된 바 있다. Delta 파라미터의 경우에는 ML 추정에 대한 효과적인 근사법이 유도되며 만족할 만한 결과의 평균 추정을 가져오는 것으로 알려져 있다.

V. 결론

본 논문에서는 배경잡음과 채널왜곡으로 오염된 환경 하에서 음성인식의 성능을 향상시키기 위한 여러 가지 방법들에 대해서 소개 및 성능 비교를 하였다. 상기에 소개된 방법들 외에도 인간의 청각시스템을 모방한 방법들[38][39]이 연구된 바 있다.

지금까지 잡음환경에 강한 음성인식기를 구현하기 위한 많은 방법들이 연구 발표되었지만, 대부분 stationary한 부가잡음과 선형필터링이라는 요소들만 제거하는데 제한되어 있다. 보다 향상된 성능을 위해서는 환경에 대한 보다 정확한 모델링이 필요하고 인간의 청각시스템이 잡음 환경 하에서도 비교적 정확히 음성을 인식해내는 것처럼 인간의 청각시스템을 모방하는 방법에 대한 연구가 계속되어져야 한다. 한편으로는 지금까지의 공학적인 어떤 연구와는 달리 인간의 기억을 기본으로 하는 직관에 의한 인식능력에 대한 연구 결과도 첨부되어져야 한다고 본다.

참고문헌

[1] Dautrich, B. A., Rabiner, L. R., and Martin T. B., "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," *IEEE Trans. on Acoustics, Speech, Signal Processing*, Vol. ASSP-31, No. 4, pp.793-807,

August 1983.

[2] Markel, J. D. and Gray, A. H. Jr., *Linear Prediction of Speech,* Springer-Verlag,1976.

[3] Zwicker, E., and Terhardt, E., "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, 68(5), pp.1523-1525,1980.

[4] Stevens, S. S. and Volkman, J., "The relation of pitch to frequency", *American Journal of Psychology*, Vol. 53, pp.329, 1940.

[5] John R. Deller,Jr.,John G. Proakis and John H. L. Hansen, *Discrete-time Processing of Speech Signals*, Prentice-Hall, 1993.

[6] Harvey F. Silverman and David P. Morgan, "The Application of Dynamic Programming to conneted Speech Recognition," *IEEE ASSP Magazine*, pp.6-25, July 1990.

[7] Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications to Speech Recognition," *Proc. IEEE,* Vol. 77, No.2, pp.257-285, 1989.

[8] Waibel, A., Hanazawa, T., Hinon, G. E., Shikano, K., and Lang, K. J., "Phoneme Recognition using time-delay Neural Networks," *IEEE Trans. on Acoustics, Speech, Signal Processing,* Vol.37(3), pp.328-339, March 1989.

[9] Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, Signal Processing,* Vol. 27, pp.113-120, April 1979.

[10] Lockwood, P. and Boudy, J., "Experiments with a Non-linear Spectral Subtraction (NNS), hidden Markov models and the projection for robust speech recognition in cars," *EUROSPEECH,* pp.79-82, 1991.

[11] Ephraim, Y. and Malah, D., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, Signal Processing* ,vol. 32, pp.1109-1121, December 1984.

[12] Ephraim, Y. and Malah, D., "Speech Enhancement Using a Minimum Mean-Square log-spectral amplitude estimator,"*IEEE Trans. on Acoustics, Speech, Signal Processing* ,vol. 33, pp.443-445, 1985.

[13] Lim, J. S., and Oppenheim, A. V., "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech, Signal Processing* ,vol. 26, pp. 197-210, June 1978.

[14] Widrow, B., Grover, J. R., McCool, J. M., et al. "Adaptive Noise Cancelling: Principles and Applications," *Proc. IEEE,* vol. 63, pp.1692-1716, Dec. 1975.

[15] Flanagan, J. L., "Use of Acoustic filtering to control the beamwidth of steered micrphone arrays", *Journal of the Acoustical Society of America,* 78(2): 423-428, August, 1985.

[16] Davis, S. B. and Mermelstein. P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, Signal Processing* ,vol 28, No. 4, pp. 357-366, 1980.

[17] Hermansky, H., "Perceptual Linear Predictive(PLP) analysis of speech," *J. Acost. Soc. Am.*, 87(4), pp. 1738-1752, 1990.

[18] Mansour D. and Juang B.,"The short-time modified coherence representation and its application for noisy speech recognition", *IEEE Trans. on Acoustics, Speech, Signal Processing*, ASSP-37(6), pp.795-804, 1989.

[19] Furui S., "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoustics, Speech, Signal Processing*, ASSP-34, pp.52-59, 1986.

[20] Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, 2(4), pp. 578-589, 1994.

[21] Morgan, N. and Hermansky, H., "RASTA extensions: Robustness to additive and convolutional noise," *ETRW: Speech Processing in Adverse Conditions*, pp.115-118, 1992.

[22] Liu, F.-H., Stern R. M., Huang, X., and Acero, A., "Efficient cepstral normalization for robust speech recognition," Proc. ARPA speech and Natural Language Workshop, pp. 69-74, Princeton, March 1993.

[23] Acero A. and Stern R. M., "Environmental Robustness in Automatic Speech Recognition," *ICASSP,* 1990.

[24] Acero A. and Steran R. M., "Robust Speech Recognition by Normalization of the Acoustic Space," *ICASSP*, 1991.

[25] Liu, F.-H., Moreno, P. J., Stern, R. M., and Acero, A., "Signal Processing for Robust Speech Recognition," *Proc. of the spoken language Technology Workshop*, 1994.

[26] Moreno, P. J., Raj, B., Gouvea, E., and Stern, R. M., "Multivariate Gaussian-Based Cepstral Normalization for Robust Speech Recognition," *ICASSP*, vol 1, pp.137-140, 1995.

[27] Liu F.-H., Acero, A. and Stern, R. M., "Efficient joint compensation of speech for the effects of Additive Noise and Linear filtering," *ICASSP*, 1992.

[29] Moreno, P. J., "Speech Recognition in Noisy Environments", Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, April 1996.

[28] Liu, F.-H., "Environmental Adaption for Robust Speech Recognition," Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University,1994.

[30] Hanson, B. A., and Wakita, H., "Spectral Slope distance Measure with Linear Prediciton Analysis for Word Recognition in Noise," *IEEE Trans. on ASSP*, Vol. ASSP-35, No. 7, pp.968-973, July 1987.

[31] Juang, B. H., Rabiner, L., and Wilpon, J., "On the use of bandpass liftering in speech recognition," *ICASSP*, pp.765-768, 1986.

[32] Mansour, D., and Juang, B. "The short-time modified coherence representation and its application for noisy speech recognition," *IEEE Trans. on ASSP*, ASSP-37, No.6, pp.795-804, 1989.

[33] Soong, F. and Sondhi, M., "A frequency-weighted Itakura spectral distortion measure and its application to speech recognition," *ICASSP*, pp.625-628, 1987.

[34] Nocerino, N., Soong, F., Rabiner, L., and Klatt, D., "Comparative study of several distortion measures for speech recognition," *ICASSP*, pp. 25-28, 1985.

[35] Noda, H., "Frequency-warped spectral distance measures for speaker verification in noise," *ICASSP*, pp.576-579, 1988.

[36] Varga A. and Moore R.,"Hidden Markov Model decomposition of speech and noise," *ICASSP*, pp.845-848, 1990.

[37] Gales, M. and Young S. "HMM recognition in noise using parallel model combination" *EUROSPEECH*, pp.837-840, 1993.

[38] Oshima, Y., "Environmental Robustness in Speech Recognition using Physiologucally Motivated Signal Processing", Ph. D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1993.

[39] Seneff, S., " A joint Synchrony/Mean-Rate Model of Auditory Speech Processing,"*Journal of phonetics*, 16:55-76, January, 1988.

[39] Ghitza, O., "Auditory Neural feedback as a basis for speech processing," *ICASSP*, pp. 91-94., 1988.