음성 인식을 위한 환경 파라미터 보상에 관한 연구

A Study on Environment Parameter Compensation Method for Speech Recognition

이호웅

(동원대학 정보통신과 교수)

홍미정

((주)디알텍 연구원)

Key Words : 음성인식, CMN, VTS

목 차

Ⅰ. 서 론

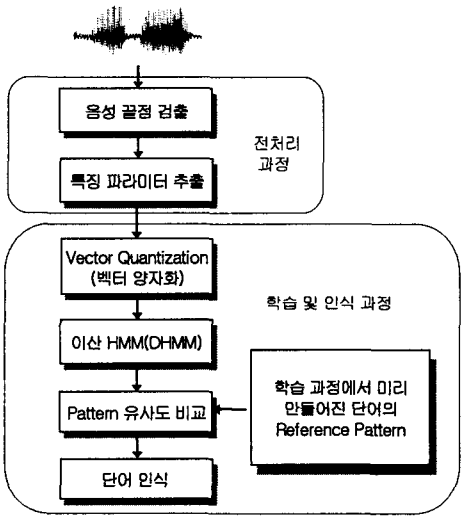
최근 몇 년간 Environmental Robustness 분야는 음성 인식 연구 분야 중에 가장 주목받고 있으며, 많은 연구소에서 음성의 음향학적인 특징, 음성 특징들의 필터링에 기반을 둔 접근과 다른 알고리즘들로 인식 시스템의 정확성을 증대시키기 위해 연구되고 있다[1][2].

환경 보상(Environment Compensation)에 접근하는 가장 성공적인 방식으로는 특징 벡터 영역에서의 변환이나 청각 모델을 이용한 인식에 비하여 재학습이 필요하지 않고, stereo Database가 없는 상황에서도 인식 대상의 단어를 이용하여 환경 잡음을 추정하여 모델을 보상하는 모델 파라미터 변환 방식이다[3][4].

본 논문에서는 각 단어마다 환경 잡음을 추정한 후 주어진 모델 파라미터를 몇 개의 보상(Compensation) 알고리즘을 적용시키는 방식에 대하여 연구한다. 특히 음성 인식 시스템이 환경에 강인하도록 하기 위한 방법 중 잡음에 강한 특징을 기반으로 한 방법의 CMN(Cepstral mean normalization)과 채널잡음에만 국한된 CMN과는 다르게 부가 잡음과 채널 왜곡을 동시에 감소시키는 최신 기법으로 Carnegie Mellon University에서 Moreno가 제안한 모델에 기반을 둔 보상 방법의 VTS(Vector Taylor Series)알고리즘[3] 을 비교하고 Moreno의 실험 환경과 다른 실험 환경에서 직접 실험하여 음성 인식 시스템의 인식 결과를 얻고자 한다.

Ⅱ. 음성 인식 시스템의 이론적 배경

고립 단어 음성 인식 시스템은 기본적으로 <그림 1>과 같이 몇 개의 블록으로 연결되어 구성된다. 음성 검출(Endpoint Detection)에서는 신호 속에서 음성을 검출하며, 특징 추출(Feature Extraction)에서는 검출된 음성으로부터 그 음성의 성질을 잘 표현해 주는 특징 벡터를 추출한다. 유사도 비교(Distance Measure) 부분에서는 이미 저장되어 있는 기준 모델과 검출된 임의의 음성 신호간의 유사 도를 측정하고 마지막에 인식단어를 결정한다.



<그림 1> 고립단어 음성 인식 시스템

1. 음성 전처리

1) 음성 끝점 검출(End-point Detection)

음성 입력 파형이 들어오면 우선 끝점 검출 과정, 즉 입력된 음성 신호로부터 묵음과 음성을 구분해야 한다. 인식 시스

템의 성능을 크게 좌우하는 필수적인 요소로서 정확한 끝점 검출이 필요하다. 일반적으로 끝점 검출에 사용되는 파라미터로는 단 구간 대수 에너지(Log Energy)와 영교차율(Zero Crossing Rate) 등이 있다[5].

본 논문에서는 영교차율을 사용하여 음성 신호의 끝점 검출을 하였다. 영교차율은 한 프레임 내에서 음성 파형의 영점과 교차하는 횟수를 말하며 화자의 성량에 대해서는 독립적이다. 음성 생성 모델에 의하면 유성음은 음성의 에너지가 스펙트럼의 3kHz정도의 낮은 주파수에 밀접해 있으나 무성음은 에너지가 높은 주파수에서 주로 발견된다. 이때 높은 주파수는 많은 영교차율을 갖게 되며, 낮은 주파수에는 적은 영교차율을 갖게 되므로 영교차율과 주파수에 의한 에너지 분포는 밀접한 관계가 있다. 일반적으로 음성 신호에서의 에너지 크기는 무성음 부분 보다 유성음 부분에서 크게 나타나며 에너지 E_n 은 다음과 같이 주어진다.

$$E_n = \sum_{m=n}^{n+N-1} x^2(m) \qquad (1)$$

여기서 N은 한 프레임의 길이를, n은 프레임의 색인(index)을 나타낸다.

2) 음성 특징 파라미터 추출

입력 신호에서 얻어지는 신호에는 여러 가지 원인에 의해 여러 주파수 성분이 포함되어 있으며 원하지 않는 잡음 들이 포함 될 수 있다. 따라서 음성 신호로부터 다음 단계의 처리에 유용하게 사용되도록 음성을 표현하는 파라미터를 추출하는 것이 필요하다.

따라서 음성 전처리에 의해서 프레임 별로 Pre-emphasis를 거쳐 Window가 취해지고, LPC(Linear Prediction Coefficient), MFCC (Mel Frequency Cepstral Coefficient), PLP(Perceptual Linear Predictive)에 의한 켑스트럼에 의하여 음성의 파라미터를 얻는다[6].

2. VQ(Vector Quantization) 분석

시간 t에서의 입력 특징 벡터를 X_t 라 하고 코드워드를 $\overline{X_t}$ 라 할 때 이러한 코드워드로 구성된 코드벡터가 최적이 되려면 모든 입력 벡터와 코드워드간의 평균 왜곡 D가 최소가 되어야 한다.

$$D = \frac{1}{T}\sum_{t=1}^{T} d\left(X_t, \overline{X}\right) \qquad (2)$$

여기서 $d(\cdot)$ 는 왜곡을 나타내는 함수로서 일반적으로 유클리디안(Euclidean) 거리에 의해 식 (3)로 표현된다.

$$d\left(X_t, \overline{X}\right) = \sum_{k=1}^{K} \left| x_{tk} - \overline{x_k} \right| \qquad (3)$$

X_t : 원 패턴 $\quad \overline{X_t}$: codeword $\quad k$: 차원

이렇게 새로운 양자화 벡터를 생성 및 Update 하는 방법에 의해 k-means, LBG(Liden Buzo Gray) 등의 방법이 있다.

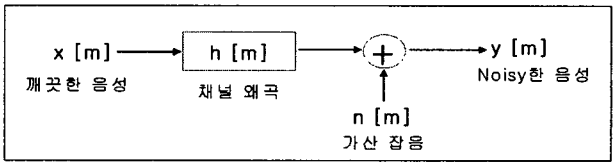
III. VTS(Vector Taylor Series) 근사화 방법

1. VTS방법의 이론적 배경

순수 음성(Clean speech)을 나타내는 vector x 가 주위 환경으로부터 영향을 받아 새로운 vector y를 만들었다고 가정하자. 이때 vector y는 noisy한 음성을 나타내고 식(4)로 표현할 수 있다.

$$y = x + g(x, a_1, a_2, \cdots) \qquad (4)$$

여기에서 함수$g(\cdot)$는 환경함수(Environmental Function)이고, a_1,a_2는 환경을 나타내는 파라미터들(vectors, scalars, matrices, ...)이다. 본 논문에서는 환경은 그림 2에 표현되는 것처럼 모델링 된다고 가정한다[3][7].



<그림 2> 부가 잡음과 채널 왜곡을 포함한 환경 모델

다음은 순수 음성(clean speech)에 잡음과 필터링이 미치는 영향을 Power spectral 영역에서 나타내고자 한다. 식은 (5)과 같다.

$$P_Y(\omega_k) = |H(\omega_k)|^2 P_X(\omega_k) + P_N(\omega_k) \qquad (5)$$

$P_Y(\omega_k)$: noisy한 음성 $y[m]$의 power spectra
$P_N(\omega_k)$: 잡음 $n[m]$의 power spectra
$P_X(\omega_k)$: clean 한 음성 $x[m]$의 power spectra
$|H(\omega_k)|$: 채널 $h[m]$의 power spectra
ω_k : 특정한 mel-spectral 대역

log-spectral 영역으로 변환하기 위해서 식(5) 양변에 식(6)과 같이 대수 연산자(logarithm operator)를 적용시킨다.

$$10log_{10}P_Y(\omega_k) = 10log_{10}(|H(w_k)|^2 P_Y(\omega_k) + P_N(\omega_k)) \qquad (6)$$

또한 noisy, clean한 음성과 잡음, 채널 왜곡을 다음과 같이 정의한다.

$$y[k] = 10log_{10}(P_Y(\omega_k)),\ x[k] = 10log_{10}(P_X(\omega_k))$$
$$n[k] = 10log_{10}(P_N(\omega_k)),\ h[k] = 10log_{10}(|H(w_k)|^2) \quad (7)$$

정의한 식 (7)식을 적용시켜 식(6)을 다시 정리해 보면,

$$10log_{10}(P_Y(\omega_k)) = 10log_{10}\left(10^{\frac{x[k]+h[k]}{10}} + 10^{\frac{n[k]}{10}}\right) \quad (8)$$

$$y[k] = x[k] + h[k] + 10log_{10}\left(1 + 10^{\frac{n[k]-x[k]-h[k]}{10}}\right) \quad (9)$$

앞에서 식(4)에서 보여준 형식에 맞춰서 식(9)을 다시 나타내면 다음과 같고 이 식은 VTS에 접근하는 첫 번째 가정이 된다.

$$y[k] = x[k] + g(x[k], h[k], n[k]) \quad (10)$$

벡터 형식으로는

$$y = x + g(g, h, n) \quad (11)$$

이 된다. 여기서

$$g(x[k], h[k], n[k]) = h[k] + 10log_{10}\left(1 + 10^{\frac{n[k]-x[k]-h[k]}{10}}\right) \quad (12)$$

이 되고, 환경 파라미터는 vector형식이다.

두 번째 가정은 clean 한 음성의 log-spectrum 랜덤 변수는 아래 식과 같이 gaussian 분포의 Mixture에 의해 나타낼 수 있다는 것이다.

$$P(x_t) = \sum_{k=0}^{K-1} P_k N_{x'}\left(\mu_{x,k}, \sum_{x,k}\right) \quad (13)$$

여기에서 K는 Mixture의 개수이다.

2. VTS 근사화

앞의 가정들을 이용하여 noisy 음성의 log-spectral vector의 확률 분포들을 계산하고자 한다. y의 확률 밀도 함수에 대한 해를 얻기 위해서는 확률 분포가 Gaussian 분포가 되도록 단순화 시킨다. 수학적인 문제를 단순화하기 위해 환경 벡터 함수 $g(x, a_1, a_2, \cdots)$를 VTS 근사 화에 의해 대치한다. 이 단순화는 환경함수 $g(\cdot)$가 해석적(Analytical)인 조건이 필요하다. 이러한 가정 하에 랜덤 변수 x 와 y사이의 관계는 식(14)와 같다.

$$y = x + g(x_0, a_1, a_2, \cdots) + g'(x_0, a_1, a_2, \cdots)(x - x_0) + \frac{1}{2}g''(x_0, a_1, a_2, \cdots)(x - x_0)(x - x_0) + \cdots \quad (14)$$

그림 2에 나타낸 환경 모델의 형태에 적용시킨 VTS 근사화는

$$y = x + g(x_0, h, n) + g'(x_0, h, n)(x - x_0) \frac{1}{2}g''(x_0, h, n)(x - x_0)(x - x_0) \quad (15)$$

이 된다. 식(15)에서 $g(x_0,h,n)$은 아래와 같이 확장 시킬 수 있다.

$$g(x_0, h, n) = h + 10log_{10}\left(i + 10^{\frac{n-x-h}{10}}\right) \quad (16)$$

또한 $g(\cdot)$는 vector x_0 에서 나타낸 것이고, $g'(x_0,n,h)$는 vector x_0 에서 벡터 함수 $g(\cdot)$를 vector x_0 에서 벡터 변수 x에 대하여 미분한 것으로 아래와 같다.

$$g'(x_0, h, n) = -diag\left[\left(1 + 10^{\frac{x_{0,i}+h_i-n_i}{10}}\right)^{-1}\right] \quad (17)$$

IV. 실험결과 및 고찰

1. 본 논문에서의 실험환경 및 방법

VTS를 적용시킨 고립 단어 인식 시스템의 성능 평가를 위해서 남성 17명이 30개의 고립 단어를 10번 발음한 Database를 가지고 인식 실험을 하였다. 인식에 선택한 단어는 <표 1> 에 기입하였다.

<표 1> 실험에 사용한 단어

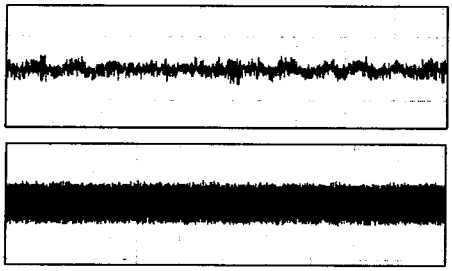
거울(1)	당선(6)	손님(11)	옛날(16)	임신(21)	첫날(26)
계단(2)	돌풍(7)	수업(12)	오디오(17)	입시(22)	파도(27)
구조(3)	명령(8)	약속(13)	외국(18)	장갑(23)	하늘(28)
군인(4)	비행기(9)	얼굴(14)	우표(19)	주차(24)	활용(29)
귀걸이(5)	성장(10)	얼음(15)	음식물(20)	주택(25)	흡연(30)

인식 알고리즘은 VQ의 K-means와 이산 HMM(DHMM)의 Baum-welch와 ML(Maximum Likelihood)방법을 사용하였고, 조용한 연구실 환경에서 녹음한 20대 후반의 남성 17명이 10번 씩 발음한 후 16,000Hz로써 sampling한 음성으로 학습하였다. 이때 VQ는 128개의 codeword를 사용하였고, HMM의 상태(state)는 5개로 하였다. 실험에 사용한 모든 음성 파라미터(parameter)는 13차로 하였고, 256 sample을 한 개의 프레임(frame)으로 하였다. NSR(Noise to Signal ratio)은 0dB, 2dB, 3dB의 범위로 하였고,

여기서 NSR은 임의적인 신호에 대한 잡음의 크기를 나타

낸 것으로 가정하였다.

그림 3은 실험에 사용한 배경잡음들로 길거리의 배경잡음과 컴퓨터로 만든 임의의 백색 잡음(2dB)을 보여주는 그림이다.



<그림 3> 길거리 잡음과 컴퓨터에서 만든 백색 잡음(2dB)

이러한 배경 잡음과 단어들을 각각 혼합하여 인식 실험을 하였는데, 임의의 백색잡음(0, 2, 3 dB) 와 길거리 잡음인 경우로 하여 4가지 배경 잡음 환경으로 실험 하였다.

2. 기본 인식 시스템의 실험결과

기본적인 인식 시스템에 거리잡음과 임의의 백색잡음(0, 2, 3 dB)이 섞인 음성 data를 입력으로 했을 때의 database 인식률은 <표 2>에 나타내었다.

<표 2> 4가지 잡음들 경우의 단어인식률

	0dB(N/S비)	2dB	3dB	거리 잡음
기본인식시스템 인식률	89.6%	85.1%	78%	88.2%

3. VTS를 적용한 인식 시스템의 실험결과

1) Moreno의 환경에서의 실험결과

첫 번 째 실험은 4개의 SNRs에서 Gaussian 백색 잡음인 섞인 5,000 단어(1993년에 평가된 WSJ set)를 사용하였다. 또한 256개의 Gaussian은 clean 음성 특징 벡터의 분포를 나타내었고, VTS 알고리즘은 40 차수의 log spectral 벡터를 사용하였다. 다음 <표 3>은 실험 결과를 나타낸다.

<표 3> VTS 알고리즘 실행 결과(Moreno)

SNR	0dB	5dB	10dB	20dB
0차 VTS	10%	33%	68%	**89%**

두 번째 실험은 달리는 자동차 안에서 녹음한 음성(1994년에 10평가 set)을 사용하였다. 인식 시스템은 37,000 개의 문장으로 구성된 WSJ에서 학습되었고 남자와 여자 모델은 약 10,000 senonic 클러스터로 구성되어 있다. 실험 결과는 <표 4>와 같다.

<표 4> 실제 데이터에서 VTS 알고리즘 실행 결과(Moreno)

SNR	15dB	20dB	25dB	30dB
0차 VTS	81%	85%	89%	**90%**

2) 본 논문에서의 실험 결과

본 실험에서는 인식에서 학습에 참가하지 않은 5명의 화자와 환경 잡음 4가지 인식실험을 수행한다. Baseline은 잡음처리를 하지 않은 기본 인식 시스템을 말하며, 화자는 A ~ E 로 나타내었다. <표 5>,<표 6>은 각각 환경 잡음 0dB, 2dB에서의 인식 율을 비교한 것이다.

<표 5> 환경 잡음(0dB)에서의 각각의 인식 율 비교

	Baseline	VTS	CMN
화자 A	256/300	265/300	260/300
화자 B	263/300	276/300	276/300
화자 C	248/300	258/300	247/300
화자 D	245/300	251/300	239/300
화자 E	264/300	270/300	272/300
합계	1276/1500	1320/1500	1294/1500
인식률	**85.1%**	**88%**	**86.3%**

<표 6> 환경 잡음(2dB)에서의 각각의 인식 율 비교

	Baseline	VTS	CMN
화자 A	269/300	274/300	271/300
화자 B	287/300	292/300	286/300
화자 C	260/300	266/300	258/300
화자 D	253/300	262/300	249/300
화자 E	275/300	280/300	283/300
합계	1344/1500	1374/1500	1347/1500
인식률	**89.6%**	**91.6%**	**89.8%**

<표 7> 환경 잡음(3dB)에서의 각각의 인식 율 비교

	Baseline	VTS	CMN
화자 A	231/300	249/300	247/300
화자 B	243/300	254/300	255/300
화자 C	233/300	239/300	230/300
화자 D	221/300	240/300	232/300
화자 E	242/300	263/300	244/300
합계	1170/1500	1245/1500	1208/1500
인식률	**78%**	**83%**	**80.5%**

<표 8> 환경 잡음(거리잡음)에서의 각각의 인식 율 비교

	Baseline	VTS	CMN
화자 A	272/300	261/300	250/300
화자 B	242/300	270/300	241/300
화자 C	261/300	288/300	260/300
화자 D	256/300	246/300	252/300
화자 E	277/300	276/300	279/300
합계	1308/1500	1341/1500	1282/1500
인식률	**87.2%**	**89.4%**	**85.5%**

<표 8>,<표 9>는 각각 환경 잡음이 3dB와 거리잡음 인 경우의 인식 율을 비교한 것이다. 인식 결과를 보면, 두 알고리즘(VTS, CMN)의 인식결과가 Baseline의 인식 결과보다 더욱 높은 인식률을 보이며, 특히 VTS 알고리즘을 적용시킨 인식 시스템의 인식률이 무 잡음 환경에서의 인식률에 대략적으로 접근 하게 된다. 즉, 위의 실험 결과를 통하여 다음과 같은 사실을 알 수 있었다. CMN은 채널 왜곡(channel

filtering)만을 모델하기 때문에 정확한 환경을 모델하지 못하고, 이러한 약점을 해결하려는 것이 VTS이다.

특히 VTS의 인식률이 높은 경우는 그 만큼 단순히 잡음만의 환경 파라미터에 의해 모델화 한 것 보다 여러 다양한 환경 파라미터에 의한 좀 더 정확한 모델이 필요하다는 것으로 보인다.

<표 10>은 종합적으로 무 잡음 환경에서의 인식 시스템과 baseline, VTS와 CMN의 인식률을 정리해 보았다.

<표 10> 몇몇 noisy 음성들 경우에서의 인식률 정리

무 잡음에서의 인식률	N/S비	Baseline	VTS	CMN
92.9%	0dB	89.6%	**91.6%**	89.8%
	2dB	85.1%	88%	86.3%
	3dB	78%	83%	80.5%
	거리 잡음	88.2%	89.4%	85.5%

기존의 논문에서는 SNR비를 사용한 것에 비해 본 논문에서는 NSR비를 사용했기 때문에 가장 작은 값의 noisy 음성의 입력 경우, 인식률을 각각 살펴보면 기존의 논문에서는 89%, 90%(실제 데이터), 본 논문에서는 91.6%, 89.4%의 인식률을 나타내었다.

V. 결 론

본 논문에서는 환경에 강인한 voice portal system의 인식률 향상을 목표로 음성 인식 시스템의 성능을 저하시키는 요인 중 부가 잡음과 채널 왜곡을 동시에 감소시키는 Moreno가 제안한 최신 기법인 VTS(Vector Taylor Series)와 기존의 잡음 처리 방법 중 CMN(Cepstral Mean Normalization) 방법을 직접 비교 실험하여 인식률을 검토하였다.

그 결과 VTS 알고리즘은 보다 정확한 환경을 모델화하며, 보다 정확한 환경 파라미터를 예측할 수 있으므로 인식률도 CMN 방법보다 높게 나타나며, 무 잡음 환경에서의 음성 인식 성능에 가까워짐을 알 수 있었다.

그러나 기존의 Moreno의 실험 환경과 결과를 비교해 보면, 본 논문에서의 실험환경이 채널왜곡의 영향과 자동차나 길거리에서 직접 녹음 한 음성data가 부족했음을 알 수 있었다.

따라서 환경 잡음인 채널 왜곡의 영향을 좀 더 충분히 반영하고, 여러 가지 환경에서의 데이터 음성을 사용하여 직접 추가 실험을 한다면, 또한 Moreno의 연구에서도 해결하지 못한 계산 량의 감소를 위한 환경 함수의 수식을 더 단순화 시킬 수 있다면, 보다 좋은 인식 결과를 얻으리라고 기대한다.

참고문헌

1. B. Juang, "Speech Recognition in Adverse Environments," *Computer Speech and Language,* Vol.5, pp.275-294, 1991.
2. M. F. Gales, "Model-Based Techniques for Noise Robust Speech Recognition," *Ph.D dissertation, University of Cambridge*, Sept. 1995.
3. P. Moreno, "Speech Recognition in Environments," *Ph.D dissertation, Carnegie Mellon Univ. April 1996.*
4. A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph.D dissertation, CMU, Department of Eletrical and Computer Engineering, 1990.
5. 김광수, 정현열, "Histogram 처리와 Noise Threshold를 이용한 음성인식기의 환경잡음 처리 성능향상," 영남대학교 정보통신학과. 1998.
6. L. Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentic-Hall International, Inc., 1993.
7. A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm Adaptation Using Vector Taylor Series for Noisy Speech Recognition," *Processing of ICSLP, 2000.*