# Robust 음성 인식기의 Front-End 설계를 위한 고찰

° 이 우 형* 정 현 열* Richard M. Stern**
* 영남대학교 전자공학과
**The robotics Institute, Carnegie Mellon University

# A Consideration for Front-End Design of Robust Speech Recognizer

° Woo-Hyoung Lee* Hyun-Yeol Chung* Richard M. Stern**
* Dept. of Electronic Eng. Yeungnam University
**The robotics Institute, Carnegie Mellon University

## 1. 서론

최근 음성 인식 기술의 발전으로 실생활의 여러 곳에 응용될 수 있는 화자 독립 음성 인식기가 속속 출현하고 있다. 이에 따라 잡음환경에 강한 음성 인식기의 필요성이 강조되고 있으나 현재까지의 음성 인식기에서의 문제점은 training 시와 testing때에 다른 마이크를 사용하거나 다른 음향적 환경이 달라지면 인식기의 성능이 크게 저하되는 문제가 있다 [1]. 따라서 모터 팬, 문소리, 다른사람들의 대화 등에 의한 부가 잡음(additive noise)로 인한 음성 정보의 훼손뿐만 아니라 실내의 반향(reverberation), 마이크의 변동에 의한 스펙트럼 왜곡, 화자의 성도의 변동에 의한 스펙트럼 왜곡등과 같은 선형 filtering의 효과로 인한 음성 정보의 훼손을 감소시키는 방법이 요구된다. 나아가서 원거리 전화선로를 통한 음성의 인식이나 자동차 내부, 또는 공장 내부와 같은 잡음이 많은 환경에서도 인식이 가능하도록 하여야 한다 [2] [3].

이를 위하여 음성 인식 시스템의 전처리 단계를 둠으로서 음성에 포함된 잡음을 제거하여 잡음으로 인한 음성 인식 시스템의 성능저하를 최소화할 수 있다.

본 논문에서는 음향적, 환경적인 변화에 강한 자동 음성 인식 시스템을 개발하기 위하여 현재까지 알려진 음성 인식 전처리 단계에서 많이 이용되는 켑스트럼영역에서의 켑스트럼 고역 통과 필터링에 의한 RASTA, CMN 처리, 켑스트럼 remapping에 의한 CDCN처리에 대해 검토하고, 이를 화자 독립 연속 음성 인식 실험을 통해 그 유효성을 비교, 검토한다.

## 2. 전처리 방법

### 2.1 RASTA 처리

Hermansky등은 신호 스펙트럼의 각 성분안에 갖이 일정한 요소를 간단하고 효과적으로 억압할 수 있는 RASTA(RelAtive SpecTrAl)처리 방법을 제안하였다[4]. 이 방법은 이전의 LPC에 기초를 한 스펙트럼 추정법 대신에 영 주파수에서 날카로운 영 스펙트럼을 가지는 필터에 의해 각 주파수 대역이 필터링되는 스펙트럼 추정법을 사용하는데 식 (1)과 같이 나타낸다.

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

이를 이용하면 정재적(stationary)인 부과잡음은 스펙트럼 영역에서 제거할 수 있고, 정재적(stationary)인 채널 왜곡(channel distortion)은 로그 스펙트럼 영역에서 제거할 수 있다.

즉 $x(m)$을 입력 음성 신호, 부과된 잡음을 $n(m)$이라고 할 때의 신호는 식 (2)와 같이 나타낼 수 있다.

$$y(m)=x(m)+n(m) \qquad (2)$$

정재적(stationary)인 잡음이 부과된 신호의 스펙트럼 $Y_i(\omega)$는 식 (3)와 같이 나타낼수 있다.

$$Y_i(\omega_k)=X_i(\omega_k)+N_i(\omega_k) \qquad (3)$$

i : 분석 윈도우 인덱스
k : 주파수 대역

잡음 $n(m)$이 정재적(stationary)이라면 주파수 응답 $N(\omega_k)$는 스펙트럼상의 낮은 영역에 집중되므로 이 필터를 이용하여 제거할 수 있고 또 정재적(stationary) 채널 왜곡이 존재하는 경우 식 (4)과 같이 로그 스펙트럼 영역에서 분리되므로 위에서 설명한 정재적(stationary)인 부가 잡음에 대한

처리와 마찬가지로 채널 왜곡을 줄일 수 있다.

$$y(m) = x(m) * h(m)$$
$$Y_t(\omega_k) = X_t(\omega_k)H_t(\omega_k) \quad (4)$$
$$Log(Y_t(\omega_k)) = Log(X_t(\omega_k)) + Log(H_t(\omega_k))$$

그러나 RASTA 처리를 통해서 음질이 개선 될 수 있으나 대역 필터링에 의해 음성 스펙트럼의 성분 또한 제거되는 위험이 있으므로 만일 켑스트럼 벡터의 c[0]성분이 있는 곳에서 이 처리가 이루어지면 그 문제는 심각하다고 할 수 있다. 또 다른 문제점은 부가잡음과 채널왜곡이 함께 존재하는 경우에는 이 처리를 통한 효과는 기대할 수 없다.

### 2.2 CMN 처리

RASTA 처리와 마찬가지로 켑스트럼영역에서 고역 통과 필터링에 의한 보상법의 하나로 CMN(Cepstral Mean Normalization)이 있다. 이를 식 (5)에 나타낸다.

$$c_x[n] = c_t[n] - \frac{1}{N}\sum_{n=1}^{N} c_t[n] \quad (5)$$

### 2.3 CDCN 처리

Acero등은 부가잡음과 채널왜곡이 함께 존재할 때에 사용될 수 있는 CDCN(Codeword Dependent Cepstral Normalization)을 제안하였다[1].

CDCN은 크게 다음과 같은 두단계로 나누어 처리된다.

1) 잡음 섞인 입력 벡터 Z에 대하여 ML(Maximun Likelihood) 추정법을 이용하여 equalization 벡터 q와 잡음 벡터 n을 추정.

2) 두번째 단계는 MMSE(minimun Mean-Square Error) 추정법을 사용하여 주어진 Z, q, n에 대한 잡음이 제거된 입력 음성 벡터 x를 추정.

이하 CDCN 알고리즘에 대해 간단히 설명한다.

잡음에 의한 음성 신호의 열화를 그림 1에 보인다.

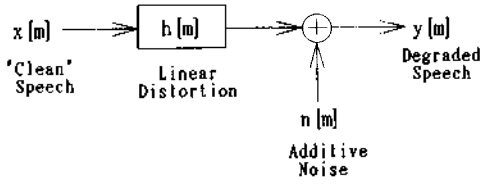


그림 1 잡음에 의한 음성 신호의 열화

이에 대한 전력 스펙트럼(power spectrum)은 식 (6)와 같다.

$$P_y(f) = P_x(f)|H(f)|^2 + P_n(f) \quad (6)$$

식 (6)의 켑스트럼 벡터를 구하면

$$\mathbf{y} = \mathbf{x} + \mathbf{q} + \mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) \quad (7)$$

여기서

$$\mathbf{q} = IDFT[\ln(|H(\omega)|^2)]$$
$$\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) = IDFT\{\ln(1 + \exp(DFT[\mathbf{n} - \mathbf{q} - \mathbf{x}]))\} \quad (8)$$

이다.

#### 2.3.1 ML 추정법

CDCN은 그림 2와 같이 입력 음성 프레임의 음향 공간(acoustic space)에, 환경적인 파라메타를 포함한 일반화 음향 공간(universal acoustic space)이 잘 정합될 수 있도록 equalization 벡터 q와 잡음 벡터 n를 계산한다.

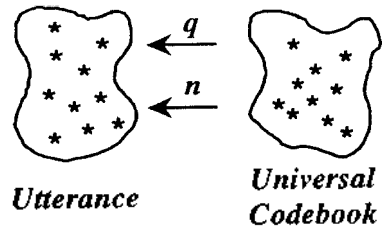


그림 2 CDCN

일반화 음향 공간(universal acoustic space)은 정규화된 무잡음 환경하의 입력 프레임의 분포로 정의되며 켑스트럼 벡터와 코드북으로써 표현되어 진다. 일반화 음향 공간(universal acoustic space)을 현재 환경의 음향 공간(acoustic space)으로 변환하는 파라메타 q, n를 추정하는 데는 식 (9)와 같은 ML(Maximun Likelihood) 추정법을 사용한다.

$$(\hat{\mathbf{n}}_{ML}, \hat{\mathbf{q}}_{ML}) = \arg\max p(Z|\mathbf{q}, \mathbf{n}) \quad (9)$$

각 프레임은 서로 독립적이라는 가정하에 식 (10)와 같이 표현할 수 있다.

$$\ln p(Z|\mathbf{q}, \mathbf{n}) = \sum_{i=0}^{N-1} \ln p(\mathbf{z}_i|\mathbf{q}, \mathbf{n}) \quad (10)$$

n에 대하여 미분하여 식 (10)의 최대치를 구한다.

$$\nabla_{\mathbf{n}} \ln p(Z|\mathbf{q},\mathbf{n}) = \sum_{i=0}^{N-1} \frac{\nabla_{\mathbf{n}} p(\mathbf{z}_i|\mathbf{q},\mathbf{n})}{p(\mathbf{z}_i|\mathbf{q},\mathbf{n})} = 0 \qquad (11)$$

이 과정은 q에 대해서도 적용할 수 있다. 이상의 ML의 해를 얻기 위해서는 잘 알려진 EM 알고리즘을 사용한다[5].

2.3.2 MMSE 추정법

잡음이 제거된 음성 벡터 x에 대한 MMSE(Minimun Mean-Square Error) 추정법은 식 (12)과 같다.

$$\hat{\mathbf{x}}_{MMSE} = \frac{\sum_{k=0}^{K-1} p_k \int \mathbf{x} p(\mathbf{z}|\mathbf{x},\mathbf{n},\mathbf{q},k) p(\mathbf{x}|k) d\mathbf{x}}{\sum_{k=0}^{K-1} p_k \int p(\mathbf{z}|\mathbf{x},\mathbf{n},\mathbf{q},k) p(\mathbf{x}|k) d\mathbf{x}} \qquad (12)$$

식 (12)에서 공분산 행렬과 Γ의 대소 구별에 의한 근사화를 하면 식 (13)와 같이 나타낼 수 있다.

$$\hat{\mathbf{x}}_{MMSE} = f_0 \mathbf{c}_0 + \sum_{k=0}^{K-1} f_k \hat{\mathbf{x}}_k \qquad (13)$$

여기서 $\hat{\mathbf{x}}_k = \mathbf{z} - \mathbf{q} - \mathbf{r}_k$ (14)

이고 $f_k$는

$$f_k = \frac{\frac{p_k}{|\mathbf{C}_k|^{1/2}} \exp(-d_k/2)}{\frac{p_0}{|\Gamma|^{1/2}} \exp(-d_0/2) + \sum_{i=1}^{K-1} \frac{p_i}{|\mathbf{C}_k|^{1/2}} \exp(-d_i/2)} \qquad (15)$$

와 같이 나타난다. 여기서

$$d_0 = (\hat{\mathbf{x}}_0 - \mathbf{c}_0)\Gamma^{-1}(\hat{\mathbf{x}}_k - \mathbf{c}_k)$$
$$d_k = (\hat{\mathbf{x}}_k - \mathbf{c}_k)\mathbf{C}_k^{-1}(\hat{\mathbf{x}}_k - \mathbf{c}_k)$$

이다.

## 3. 인식기의 구성

SPHINX system[6]은 그림 3과 같이 크게 Training 5부분, Recognition 3부분으로 구분할 수 있으며 각 부분의 기능을 간략히 설명한다.

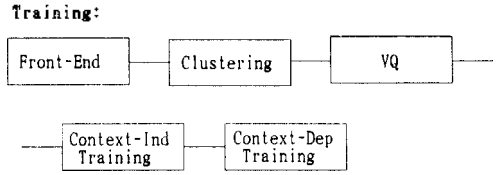


그림 3 SPHINX 시스템의 개요

3.1 Training부

음성신호는 16kHz로 sampling되고, Front-End부에서 0.97의 factor로 pre-emphasis하며 20ms(320sample)의 Hamming windowing하여 처리된다. 10ms씩 shift되며, 각 frame별로 14차 LPC계수를 구한 후 bilinear transform을 사용하여 12차의 LPC cepstral 계수를 구한다. 또한 clustering부에서 음성신호의 동적 특성을 반영하기 위하여 LPC cepstrum 계수로부터 1차 및 2차의 미분 계수 및 power 성분을 구하여 사용한다. hierarchical clustering 알고리즘을 사용하여 256개의 codebook이 만들어지고, Euclidean 거리를 측정하여 양자화함으로써 입력 음성신호의 대부분의 정보를 유지하면서 많은 양의 정보를 줄인다. 이때 VQ 알고리즘은 Linde-Buzo-Gray 알고리즘[7]을 약간 변형한 것을 사용한다.

본 system에서 사용하는 연속 음성 인식 기술은 Hidden Markov Model(HMM)이다. 먼저 Training의 흐름을 설명하면, 레이블된 TIMIT 문장으로부터 추출된 48개의 context-independent phone model을 AN4 데이터베이스를 이용하여 training한 후 context-dependent phone 모델을 생성한다. 이를 다시 forward-backward 알고리즘을 이용하여 interpolated context-dependent phone 모델을 생성하여 인식용으로 사용한다.

3.2 인식부

인식 단계에서는 training에서와 똑같은 처리를 거쳐 벡터 양자화된 음성신호를 HMM-based viterbi beam search 알고리즘을 사용하여 인식한다.

## 4. 실험 및 고찰

본 논문에서 사용한 AN4 Database는 남자 53명, 여자 21으로부터 얻은 1018문장으로 학습 데이터를 구성하고, 인식 실험에는 학습에 참여하지 않은 남자 7명, 여자 3명의 화자가 발성한 140문장으로 구성된다. 문장은 문자열, 숫자, 방향어들로써 사무실 환경에서 녹음하였으며 마이크는 Senheiser HMD224 close-talking 마이크와 desk-top Crown PZM6fs 마이크를 사용한다.

학습 단계에서는 Senheiser HMD224 close-talking 마이크로부터의 데이터만을 사용하고 인식 실험에서는 두종류의 마이크를 사용한다.

위에서 설명한 몇가지 전처리 단계를 거친 경우에 대한 인식 실험 결과를 표 1에 나타낸다.

| parameter | CLSTK | CRPZM |
|---|---|---|
| Baseline | 87.8 | 41.7 |
| RASTA | 87.0 | 65.1 |
| CMN | 81.7 | 47.6 |
| CDCN | 87.0 | 77.2 |

표 1 세가지 전처리에 따른 인식율(%)

CLSTK인 경우 잡음을 거의 포함하지 않기 때문에 RASTA, CMN, CDCN 처리의 효과가 거의 나타나지 않으나, 사무실 잡음이 부가되고 마이크를 통한 스펙트럼 왜곡이 함유된 CRPZM 데이터를 사용한 경우 RASTA인 경우 23.4%, CMN은 약 6%, CDCN은 약 35% 인식율 개선을 가져옴을 알 수 있다.

## 5. 결론

본 논문에서는 음향적, 환경적인 변화에 강한 자동 음성 인식 시스템을 개발하기 위하여 현재까지 알려진 음성 인식 전처리 단계에서 많이 이용되는 켑스트럼영역에서의 켑스트럼 고역 통과 필터링에 의한 RASTA, CMN 처리, 켑스트럼 remapping에 의한 CDCN처리에 대해 검토했다.

이를 화자 독립 연속 음성 인식 실험을 통해 인식 결과를 비교, 검토한 결과 사무실 잡음의 부가, 마이크를 통한 스펙트럼 왜곡이 함유된 CRPZM 데이터를 사용한 경우 RASTA인 경우 23.4%, CMN은 약 6%, CDCN은 약 35% 인식율 개선을 가져옴을 알 수 있었다.

향후 음질 개선 효과가 뛰어난 CDCN 알고리즘을 한국어 연속 음성 인식 시스템에 이용하여 그 효과를 확인하고자 한다.

## 참고 문헌


1. A.Acero and R.M.Stern "Environmental robustness in the automatic speech recognition" Proc.ICASSP'90, PP. 849 - 852, 1990.
2. Pedro.j.Moreno "Speech recognition in the telephone environmental" Carnegie Mellon University December 1992.
3. Nobutoshi Hanai "Speech recognition in the automobile" Carnegie Mellon University May 1993.
4. H.Hermansky, N.Morgan, A.Bayya, P.Kohn "Compensation for the effect of the communication channel in auditory-like analysis of speech" Proc. EUROSPEECH '91, pp. 1367 - 1370, Genova, 1991.
5. Alejandro Acero "Automatic and environmental robustness in automatic speech recognition" Kluwer Academic Publisher 1993.
6. Kai-Fu Lee "Automatic speech recognition : The development of the SPHINX system" Kluwer Academic Publisher 1989.
7. Y.Linde, A.Buzo, R.M.Gray "An algorithm for vector quantization" IEEE Trans. on Communication, vol. com-28, No.1 Jan. 1980. pp. 84 - 95.
8. H.Hermansky "Perceptual linear predictive(PLP) analysis for speech" J. Acoust. Soc. Am., pp. 1738 - 1752, 1990.
9. H.Hermansky and N.Morgan "Toward handling the acoustic environment in spoken language processing" Proc. of Intl. Conf. on Spoken Lang. Processing.
10. H.Hermansky and N.Morgan "RelAtive SpecTrAl(RASTA) processing in the speech analysis" Proc. 12th Speech Research Symposium, Rutgers University, June 1992.
11. C.W.Therrien "Discrete random signals and statistical signal processing" Prentice-Hall International, Inc. 1992.