

히스토그램 처리를 이용한 음성인식시스템의 환경잡음 전처리기법

A Front-End Processing for Environmental Noise Reduction Using Histogram Processing in Speech Recognition System

김 광 수*, 정 현 열**

(Kwang-Soo Kim*, Hyun-Yeul Chung**)

Abstract

In this paper, to reduce additive noise and channel distortion simultaneously contained in input speech of speech recognition system a new method that uses estimated power by histogram processing technique as a noise level of the input for noise interval is proposed. In the evaluation tests, the effectiveness of this method was verified by improving about 15% of recognition accuracy in 45 Korean words recognition experiments.

I. 서 론

최근 음성 인식 기술의 발전으로 음성 인식 시스템의 실용화가 점차 늘어남에 따라 잡음 환경에 강한 음성 인식기의 필요성이 강조되고 있다.

이는 기존의 음성인식시스템은 잡음이 거의 없는 환경에서 실험되어왔기 때문에, 실제 환경에 적용하였을 경우 인식오류가 매우 증가하여 실시스템으로서의 기능을 발휘할 수 없었기 때문이다.

잡음환경에 강한 음성 인식기의 구현을 위해서는, 기존의 음성 인식 시스템에 전처리 단

계를 두어 음성에 포함된 환경잡음을 제거함으로써 이로 인한 음성 인식 시스템의 성능 저하를 최소화 할 수 있다.

음성 인식기의 성능을 저하시키는 음향적, 환경적 변화의 요인에는 음성의 변화에 비하여 서서히 변화하는 요인과 음성 데이터의 녹음시의 환경에 의한 요인등이 있다.

특히 훈련과 평가시의 부가 잡음 레벨의 변화는 음성 인식 시스템의 성능을 크게 저하시키는 요인이 되며 마이크등의 녹음 장비의 변동에 의한 스펙트럼 왜곡과 같은 채널 왜곡 또한 인식 시스템의 성능을 저하시키는 요인이 된다.

이를 위한 기존의 연구들은 대부분 부가잡음, 채널왜곡을 위한 처리를 분리해 처리하고 있다.

* 영남대학교 전자공학과 석사과정

** 영남대학교 전기전자공학과 교수

따라서 본 논문에서는 음성 인식기의 성능을 저하시키는 요인중 부가 잡음과 마이크의 변동에 의한 채널 왜곡을 동시에 감소시키는 방법으로 일반적으로 알려진 SS, CMN, RASTA등을 이용한 전처리방법의 단어인식시스템에 대한 유효성을 검토한 후, 이 위에 Histogram에 의한 추정법을 적용하는 새로운 잡음처리방법을 도입하여 그 유효성을 확인하고자 한다.

II. 환경잡음의 전처리

2.1 기존의 전처리방법

2.1.1 Spectral Subtraction

비음성구간에 존재하는 부가적인 잡음의 평균치를 전체음성 스펙트럼값에 대해서 모두 차감하는 간단한 방법이다.

이는 음성신호의 초기목음구간에 부가적인 잡음이 존재한다는 가정에 따른다.

2.1.2 Cepstral Mean Normalization

아래와 같이 cepstral 영역에서의 전체평균값을 차감함으로써 필터링과 유사한 효과를 가지게하는 방법이다.

$$c_{\hat{x}}[n] = c_x[n] - \frac{1}{N} \sum_{n=1}^N c_x[n]$$

2.1.3 RASTA

음성스펙트럼의 각 성분 내에서 음성에 비해 느리게 변화하는 부분을 아래의 전달함수를 가지는 고역통과필터를 거쳐 억압되도록 하는 방법이다.

$$H(z) = \frac{z^4(0.2+0.1z^{-1}-0.1z^{-3}-0.2z^{-4})}{(1-0.98z^{-1})}$$

2.2 도입한 방법

2.2.1 CDCN

음성 신호의 환경 잡음에 의한 열화를 부가 잡음과 채널 왜곡으로 모델링한 것을 Fig.1에 보인다.

열화된 음성의 전력 스펙트럼 $Y(\omega)$ 는 식 (1)과 같이 표현할 수 있다.

$$Y(\omega) = X(\omega)|H(\omega)|^2 + N(\omega) \quad (1)$$

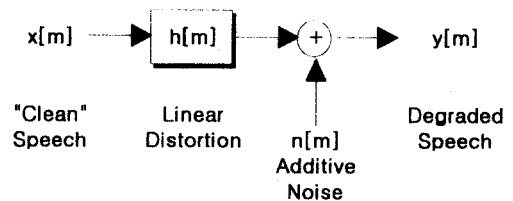


Fig 1. Model of a degradation

켈스트럼(cepstrum) 영역에서 입력 음성 벡터를 식 (2), 잡음 벡터를 식 (3), 열화된 음성 벡터를 식 (4), 채널 equalization 벡터를 식 (5)와 같이 정의한다.

$$\mathbf{x} = \text{IDFT} \{ \ln X(\omega) \} \quad (2)$$

$$\mathbf{n} = \text{IDFT} \{ \ln N(\omega) \} \quad (3)$$

$$\mathbf{y} = \text{IDFT} \{ \ln Y(\omega) \} \quad (4)$$

$$\mathbf{q} = \text{IDFT} \{ \ln |H(\omega)|^2 \} \quad (5)$$

따라서 켈스트럼 벡터를 구하면 식 (6)과 같이된다.

$$\mathbf{y} = \mathbf{x} + \mathbf{q} + \mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) \quad (6)$$

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

여기서 교정 벡터 $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ 와 $\mathbf{s}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ 는 아래의 식과 같다.

$$\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q}) = \text{IDFT} \{ \ln(1 + e^{\text{DFT}(\mathbf{n} - \mathbf{q} \cdot \mathbf{x})}) \} \quad (8)$$

$$\mathbf{s}(\mathbf{x}, \mathbf{n}, \mathbf{q}) = \text{IDFT} \{ \ln(1 + e^{\text{DFT}(\mathbf{x} + \mathbf{q} \cdot \mathbf{n})}) \} \quad (9)$$

높은 SNR(Signal to Noise Ratio)에서는 아래와 같이 간략화될 수 있다.

$$y = x + q \quad (10)$$

또한 낮은 SNR인 경우에는 아래와 같이 간략화 될 수 있다.

$$y = n \quad (11)$$

2.2.1.1 캡스트림 재정합에 의한 정규화

잡음 벡터 n 과 채널 equalization 벡터 q 는 입력 음성의 음향적 공간과 환경적 파라미터를 가진 일반화 음향 공간을 재정합함으로써 얻어질 수 있다.

열화된 관측 벡터로부터 환경 파라미터 n, q 를 찾고 이를 이용하여 입력 음성 벡터 x 를 추정하기 위하여 CDCN은 크게 다음과 같은 두단계로 나누어 처리된다.

1) ML(Maximum Likelihood) 추정법을 이용하여 잡음 벡터 n 과 equalization 벡터 q 를 추정한다.

2) MMSE(Minimum Mean Square Error)추정법을 사용하여 주어진 Z 에 대하여 잡음이 제거된 음성 벡터 x 를 추정한다.

2.2.1.2 캡스트림 벡터의 MMSE 추정

무잡음 캡스트림 벡터 x 의 추정은 환경 파라미터인 잡음 벡터 n 과 채널 equalization 벡터 q 를 알고 있다는 가정 아래 이루어진다.

관측벡터 z, n 과 q 가 주어진 경우 무잡음 캡스트림 벡터 x 에 대한 MMSE (Minimum Mean-Square Error) 추정은 식 (12)와 같은 형태를 가진다.

$$\hat{x}_{MMSE} = E \{ x | z, n, q \} = \frac{\sum_{k=0}^{K-1} P |k| \int x p(z | x, n, q, k) p(x | k) dx}{\sum_{k=0}^{K-1} P |k| \int p(z | x, n, q, k) p(x | k) dx} \quad (12)$$

$p(z | x, n, q, k)$ 와 $p(x | k)$ 가 가우시안 분포라는 가정하에서 가우시안 분해하면 식 (12)는 식 (13)과 같은 형태로 나타낼 수 있다.

$$\hat{x}_{MMSE} = \sum_{k=0}^{K-1} f[k] b[k] \quad (13)$$

이고, $f[k]$ 는 식 (14)와 같다.

$$f[k] = \frac{P |k| N_c(q+r | k| + c |k|, \Gamma + \Sigma_k)}{\sum_{l=0}^{K-1} P |l| N_c(q+r | l| + c |l|, \Gamma + \Sigma_l)} \quad (14)$$

2.2.1.3 ML을 이용한 환경 파라미터의 추정

Maximum Likelihood를 이용하면 잡음 벡터 n 과 equalization 벡터 q 에 대한 사전 정보(a priori Information)가 없는 경우 관측 벡터 z 로부터 n, q 를 추정할 수 있다. 즉,

$$(\hat{n}_{ML}, \hat{q}_{ML}) = \arg \max p(Z | q, n) \quad (15)$$

각 프레임은 독립이라는 가정 아래

$$\ln p(Z | n, q) = \sum_{i=0}^{N-1} \ln p(z_i | n, q) \quad (16)$$

이고 이를 최대로 하는 것은

$$\nabla_n \ln p(Z | n, q) = \sum_{i=0}^{N-1} \frac{\nabla_n \ln p(z_i | n, q)}{p(z_i | n, q)} = 0 \quad (17)$$

를 의미한다.

식(17)는 비선형 방정식이기 때문에 EM 알고리즘을 사용한다. 결과적으로 잡음 벡터 n 과 equalization 벡터 q 의 ML 추정치는 식 (18)와 (19)에 의해 구해진다.

$$\hat{n} = \frac{\sum_{i=0}^{N-1} f_i[0] z_i}{\sum_{i=0}^{N-1} f_i[0]} \quad (18)$$

$$\hat{q} = \frac{\sum_{k=0}^{K-1} \sum_{i=0}^{N-1} f_i[k] (z_i - c[k] - r[k])}{\sum_{k=0}^{K-1} \sum_{i=0}^{N-1} f_i[k]} \quad (19)$$

2.2.2 RASTA처리의 확장

RASTA처리는 훈련에 사용한 음성과는 다른 성질을 가지는, 채널왜곡에 의하여 열화된 음성의 침예하게 변화하는 고조파성분을 필터링하는 처리방법이다.

문제는 필터링시 원하지 않는 음성정보의 훼손이 발생할 수 있다는 것과 log 영역에서의 처리이기 때문에 선형적인 채널왜곡에는 상당한 효과를 기대할 수 있으나, correlation이 낮은 부가적인 잡음의 처리에는 약하다는 단점이 있다.

즉, 스펙트럼영역에서 부가적인 uncorrelated 잡음은 log영역에서의 필터링만으로는 그 제거 효과를 기대하기가 힘들기 때문이다.

따라서 부가잡음과 채널왜곡을 동시에 처리하기 위한 RASTA의 확장된 방법인 J-RASTA처리법을 사용한다.

2.2.2.1 잡음이 부가된 음성의 처리

잡음을 다루는 방법중 SS는 느리게 변화하는 성분(정재적인 convolutionalnoise)은 효과적으로 처리하기가 어렵다.

RASTA는 SS와는 달리 단순한 차감방법이 아닌 필터링에 의한 방법을 사용한다.

이러한 두 가지 방법을 분리하여 처리하는 것은 인식률의 저하가 발생할 수 있기 때문에, 동시에 처리하는 방안이 바람직하다.

2.2.2.2 J-RASTA

이 방법은 RASTA의 log영역 처리 대신에 아래와 같은 근사화된 영역에서 처리한다.

$$y = \ln(1 + Jx) \quad (20)$$

여기서 J 는 signal-dependent인 작은 값

을 갖는 상수이고, y 는 변환후의 출력, x 는 입력이다.

이 식을 사용한 warping 변환영역은 log영역이 아닌, $J < 1$ 일 때 linear-like, $J \geq 1$ 일 때는 log영역인 변환이다.

역변환은 $x = (e^y - 1)/J$ 가 되는데, y 값이 양수임을 보증하지 못하기 때문에, 근사화된 역변환을 사용한다.

$$x = e^y/J \quad (21)$$

위의 J 값은 SN비에 따라서 특정 최적값이 존재하며, 최적값은 아래식에 의해서 구하게 된다.

$$J = 1.0 / (C \cdot E_{\text{noise}}) \quad (22)$$

그러면, 처리된 스펙트럼들의 음성신호성분은 비선형성을 가지는 log영역에 존재하게 되고, 잡음성분은 선형영역에 존재하게 된다.

III. Histogram 기법과의 혼용

3.1 분포밀도함수에 의한 추정

기존의 잡음추정은 음성신호의 초기 비음성구간, 통산 묵음구간의 스펙트럼의 평균값을 추정된 잡음으로 가정한다.

이러한 방법은 음성/비음성 구간의 검출이 매우 신뢰성있게 수행되어야 한다는 것을 전제로 하는데, 실제로 이러한 검출은 매우 어려우므로, 더 효과적인 잡음추정이 요구된다.

따라서, 본 논문에서는 신뢰성있는 잡음의 추정을 위하여 신호스펙트럼의 분포밀도함수인

Histogram 처리기법을 사용한다.

기존의 방법과 달리, histogram 처리기법은 음성/비음성구간의 구분이 불필요한 장점이 있다.

Histogram 처리기법은 다음과 같은 관찰 결과에 바탕을 두고 있다.

진폭스펙트럼의 분포를 살펴보면, 높은 SN비에서는 잡음분포가 진폭스펙트럼이 낮은 쪽으로 분포하고, 낮은 SN비에서는 높은 쪽으로 분포하게 된다.

Histogram 처리에서는 이러한 사실을 바탕으로 각 주파수대역에 대해서 진폭스펙트럼의 분포밀도함수의 값이 최대가 되는 진폭스펙트럼을 해당 주파수대역에서의 잡음레벨이라고 판정한다.

따라서, SN비의 의존도가 적으며 시간적으로 변하는 잡음의 특성에도 적용가능하다.

3.2 잡음추정의 구현

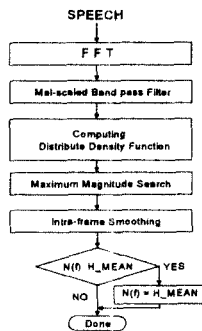


Fig 2. Histogram Processing

Histogram에 의한 처리방법을 Fig.2에 나타내었다.

프레임 단위로 입력음성을 FFT하여 주파수영역으로 변환, 청각특성을 고려한 mel-scale로 warping된 critical band pass

filter를 통과시켜서 각 band별로 출력을 얻는다.

각 band에 대해서 개별적인 histogram 처리를 수행하여 band별 잡음레벨을 추정하게 되며 그 결과를 합하여 잡음스펙트럼 특성을 구한다.

대부분의 음성프레임에 대해서는 잡음레벨이 잘 추정되지만, 일부 프레임에서 잘못된 추정값을 구할 수도 있으므로, 그림에서와 같이 현재의 잡음레벨을 이전프레임에서 구해진 잡음레벨로 smoothing하고, 지속적으로 잘못된 peak 점들이 형성되는 것을 해결하기 위해서, 평균값보다 큰 최대진폭스펙트럼값은 평균값으로 대체하였다.

3.3 전처리와의 결합

이러한 신뢰성 있는 잡음추정기법을 기존의 전처리기법과 결합시켜 사용함으로써 더 나은 성능향상을 볼 수 있을 것으로 생각된다.

IV. 인식기의 구성

본 실험에서 사용한 음성 인식기는 Fig. 3와 같이 크게 Training과 Recognition으로 구분할 수 있다.

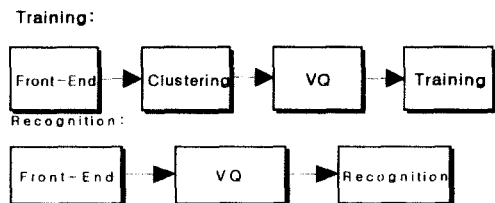


Figure 3. Block diagram of recognition system

Training 단계에서는 음성 신호를 Front-End부에 입력, 특징 파라미터로써 프레임별로 10차의 LPC mel 캡스트럼계수를 추출한다.

clustering부에서 256 코드워드를 가진 코드북(codebook)을 생성한다. VQ부에서는 Euclidean 거리를 이용하여 입력음성을 VQ한다. 이때 VQ 알고리즘은 LBG 알고리즘을 사용한다.

Training부에서는 음성으로부터 추출된 49개의 음소 모델을 forward-backward 알고리즘을 이용하여 training한 후 이 음소 모델을 단어인식용으로 사용한다. 이때 음향 음소 모델은 Discrete Hidden Markov Model이다.

Recognition 단계에서는 training에서와 똑같은 처리를 거쳐 벡터 양자화된 음성 신호를 One-Pass HMM-based Viterbi beam search 알고리즘을 사용하여 인식한다.

V. 실험 및 고찰

5.1 음성데이터

본 논문에서 사용한 음성 데이터는 ETRI에서 구축한 445 단어음성데이터베이스와 611 단어음성데이터베이스를 이용하였다.

5.2 인식실험

4절에서 설명한 기본인식시스템을 사용하여 인식실험한 결과를 표로 나타낸다.

첫 번째 실험은 시스템의 기본성능에 대한 실험으로, training용 데이터로는 Headset 마이크인 Dynamic Shure 마이크를 통하여

녹음한 445단어중 15인의 남성 화자가 1회 발성한 45단어를 이용하였고, test용 데이터는 445 단어음성 데이터베이스중 training에 참석하지 않은 5인의 남성 화자가 2회 발성한 45단어와 Desktop 마이크인 AKG Condensor 마이크를 통하여 녹음한 611 단어중 3인의 남성 화자가 2회 발성한 45단어를 이용하였다.

기본인식결과를 Table 1에 나타내었다.

Table 1. Result of recognition(단위 : %)

	화자종속	화자독립A	화자독립B
마이크	Dynamic Shure mic.	AKG mic.	
BASE	94.3	85.9	85.1

두 번째 실험은 SNR을 달리한 부가잡음을 첨가한 경우에 대한 실험으로, test로 training에 참여하지 않은 5인의 남성화자가 2회 발성한 45단어를 이용하였다.

기본인식률과 SS처리, RASTA, CMN, CDCN, 그리고 J-RASTA처리를 수행한 경우와 hitogram에 의한 잡음추정을 도입한 경우에 대한 인식결과를 Table2에 나타내었다.

Table2. Result of recognition with Additive Noise

< 기존전처리를 사용한 경우 >

	10dB	15dB	20dB	25dB
기본 인식률	14.4	19.7	32.2	45.6
SS	48.5	55.3	64.4	69.3
CMN	15.6	26.5	33.4	44.5
RASTA	22.4	27.6	38.8	48.9
J-RASTA	50.2	55.3	63.3	67.8
CDCN	60.6	64.4	67.8	70.2

〈histogram처리와 결합한 경우〉

	10dB	15dB	20dB	25dB
SS	63.34	65.56	67.78	71.3
CMN	33.4	34.5	55.3	56.4
JRASTA	63.34	64.44	66.67	70.2

Table 2로부터 기본인식률에 비하여 SS처리의 경우 24~35%의 인식률 향상을 보였으며 CDCN과 J-RASTA처리후 SN비에 따라 25~45%의 인식률 향상을 보임을 알 수 있다.

또한, histogram에 의한 처리를 도입한 결과 더욱 향상된 성능을 보임을 알 수 있었다.

세 번째 실험은 마이크의 변동에 의한 채널 왜곡과 부가잡음이 함께 함유된 경우의 실험으로, test용 데이터는 Decktop 마이크인 AKG Condensor 마이크를 통하여 녹음한 611단어중 3인의 남성화자가 2회 발성한 45단어에 SNR을 달리 한 부가잡음을 첨가하였다.

기본 인식률과 전처리를 수행한 경우의 인식결과를 Table 3에 나타내었다.

Table 3. Result of recognition with Channel Distortion and Additive Noise

〈 기존전처리를 사용한 경우 〉

	10dB	15dB	20dB	25dB
기본인식률	7.8	16.8	28.2	40.1
SS	21.1	35.6	46.2	51.3
CMN	19.9	24.5	32.3	45.6
RASTA	17.8	23.7	28.5	35.2
J-RASTA	35.2	42.3	53.4	58.7
CDCN	48.7	54.4	60.6	68

〈histogram처리와 결합한 경우〉

	10dB	15dB	20dB	25dB
SS	40.1	42.3	51.3	57.8
CMN	34.1	35.6	42.3	51.3
J-RASTA	50.1	55.6	58.7	67.8

Table 3으로부터 모든 전처리에서 인식률의 향상이 있음을 알 수 있으며 특히 CDCN과 J-RASTA는 처리후 기본인식률에 비해 SN비에 관계없이 15~30%정도의 인식률향상을 볼 수 있었으며 특히 SN비가 낮을수록 다른 전처리에 비해 인식률 향상이 뚜렷함을 알 수 있었다.

이 실험에서도 두 번째 실험에서와 같이 histogram 처리에 의한 잡음추정법이 인식률향상에 크게 기여함을 알 수 있었다.

VI. 결론

본 연구에서는 환경잡음에 강한 실용화를 위한 단어 인식시스템 개발을 목표로 음성 인식기의 성능을 저하시키는 요인중 부가 잡음과 마이크의 변동에 의한 채널 왜곡을 동시에 감소시키는 방법으로 기존의 방법과 쉐스트럼 재정합에 의한 CDCN처리방법과 RASTA를 확장한 J-RASTA를 도입하여 그 유효성을 비교 검토하였다.

또한, 기존 전처리방법에 사용되는 비음성 구간에서의 잡음추정 대신에 신뢰성이 높은 분포밀도함수를 이용한 histogram 처리방법을 사용하여 그 유효성을 확인하고자 하였다.

실험결과, 부가잡음이 첨가된 경우에 대한 실험에서는 일반적으로 알려진 SS, CMN, RASTA등과 비교하여, 전처리방법을 사용하

지 않은 경우의 기본인식률에 대해서 SN비에 따라 25% 이상의 인식률 향상을 볼 수 있었으며, 특히 CDCN처리와 J-RASTA를 사용한 경우 채널왜곡과 부가잡음이 함께 포함된 음성 에 대한 인식실험에서는 SN비에 관계없이 약 30% 정도의 인식률의 향상을 볼 수 있어 CDCN처리와 J-RASTA처리의 유효성을 확인할 수 있었다.

그리고, histogram 처리기법을 모든 전처리기법에 적용한 결과 인식성능의 향상에 크게 기여함을 보여 더욱 신뢰성있는 추정방법임을 알 수 있었다.

향후 현재까지 검토한 결과를 바탕으로 여러 가지 환경에서 채록한 음성자료를 이용한 실험을 수행한 후, 이 결과를 바탕으로 특히 부가잡음과 채널왜곡에 강건한 실용화를 위한 단어 인식시스템을 구축하고자한다.

SpecTrAl (RASTA) processing in the speech analysis" Proc. 12th Speech Research Symposium, Rutgers University, June 1992.

- [5] 이우형, 정현열 "환경잡음에 강한 음성인식기의 FRONT-END" 제13회 음성통신 및 신호처리 워크샵 논문집", pp. 356~360, 1996.
- [6] Alejandro Acero "Automatic and environmental robustness in automatic speech recognition" Kluwer Academic Publishers, 1989.
- [7] Y.Linde, A.Buzo, R.M.Gray "An algorithm for vector quantization" IEEE Trans. on Communication, vol. com-28, No. 1,Jan. pp. 84-95, 1980.

참 고 문 헌

- [1] J.E.Porter, S.F.Boll "Optimal estimaters for spectra restoration of noisy speech", Proc. CASSP84, pp.18A.2.1-4, 1984
- [2] H.Hermansky, N.Morgan, H.Hirsch "Recognition of Speech in Additive and Convolutional Noise based on RASTA Spectral Processing", Proc. ICASSP93, pp.II83 - 86, 1993
- [3] H. Hermansky "Perceptual linear predictive analysis for speech" J. Acoust. Soc. Am., pp. 1738-1752, 1990.
- [4] H.Hermansky, N.Morgan "RelAtive