잡음 필터를 이용한 음성 인식 시스템의 성능향상에 관한 연구

이 양 교, 김 학 진, 김 순 협
광운대학교 컴퓨터공학과

A Study on the Improvement of Speech Recognition System using Noise Filtering.

Yang-Gyo Lee, Hack-Jin Kim, Soon-Hyob Kim
Kwangwoon Univ.

Abstract

본 논문에서는 HMM알고리즘을 이용한 중규모급, 화자독립, 연결음성시스템에서의 인식성능 향상을 위해, 단어 인식기가 가지고 있는 고려사항들 중에 잡음(Noise)에 강한 모델을 위해 동작환경에 따른 적절한 필터를 구성하고 이차적으로 특징 파라미터를 개선하여 Noise를 보상하는 방법을 적용하였다. 인식기의 성능에 큰 영향을 미치는 요인중 하나인 전처리 기능의 평가로 성능향상의 요인을 찾아 음질개선을 위한 보다나은 잡음보상 방법을 제시하고자 하였다.

1. 서론

음성신호에서의 전처리 과정에 잡음 제거기법이 적용되지 않은 경우 음성신호 구간내의 잡음에 대한 처리가 수행되지 않은 채 훈련과정의 특징벡터와 비교하게 되므로 특징벡터열의 패턴매치과정에서 낮은 확률값을 가지게 된다. 또한 SN비가 충분치 못한 경우, 즉 SN비가 매우 낮아 충분한 에너지를 구하지 못하는 경우에 인식에 효과적으로 사용할 수 없고 마스킹 효과에 따라 잡음성분에 묻힌 음성구간을 찾아내지 못하여 인식성능을 떨어뜨리는 문제를 가지고 있었다. 이에 따라 본 논문에서는 스펙트럼 차감법과 J-RASTA기법을 이용, MFCC특징벡터를 추출, 사용하여 인식성능을 개선한 방법에 대해 제안하고자 한다.

2. 본론

(1)스펙트럼 차감법

음향학적 특징중에는 노이즈 마스킹효과가 있는데 이의 응용중 한가지가 바로 스펙트럼 차감법이다. 마스킹 효과에 의해 일정 문턱치 이하에 묻혀있는 음성이 청각적으로 구분이 불가능해 질 수 있음을 역으로 이용한 방법이다.

차감법을 사용하게 될 때 얻을 수 있는 장점은 크게 두가지로 요약될 수 있다.

첫째, 스펙트럼 차감법에 의해 남겨지는 환경 종속적인 잔여성분은 마스킹 될 수 있다는 점과,

두 번째, 낮은 음압의 음성발성이 인식기로 하여금 잡음성분으로 인해 구별이 불가능 할 수 있는 마스크되는 음성부분에 대해 훈련과정을 생략할 수 있다는 점이다.

본 논문에서 사용한 스펙트럼 차감법은 아래와 같이 구해질 수 있다.

$$Y_{ss(\omega)} = \max(Y(\omega) - \alpha\hat{N}(\omega), \beta Y(\omega))$$

$\hat{N}(\omega)$: estimated noise.

α : over-estimation factor.

β : flooring factor.

여기서 입력 음성신호의 초기 비음성 구간을 이용해서 잡음레벨을 추정하고, 그 레벨을 마스킹하는 정도의 target 잡음 레벨을 정하여 정규화를 적용하는 방식을 이용하였다. 입력음성신호의 잡음레벨로부터 target 잡음 레벨의 선정은 초기 몇 개의 입력 frame의 각 필터뱅크 결과들로부터 잡음레벨의 평균, 표준편차를 구해 그 잡음레벨의 평균으로 스펙트럼 차감법을 수행한다. target 잡음 레벨의

선정은 표준편차만을 이용해, 그 표준편차의 일정 비율에 해당하는 범위를 마스킹하는 정도를 이용하며, 그 레벨로 변형된 SNR 정규화를 통해 특징 파라미터를 추출한다. 이 이론은 스펙트럼 차감에 의해 현재 입력 신호의 잡음레벨의 평균값이 제거되고 남아있는 잡음의 fluctuation이 추정된 표준편차와 관련있다는 생각에 근거한다. 사용된 DB는 전화망을 통한 음성DB를 사용하였다. 음성 신호 발생 이전에 포함되는 채널 왜곡에 따른 열화잡음이 존재한다는 전제가 선행되었다. 따라서 채널 왜곡에 따른 hiss noise, 열화잡음등은 스펙트럼 차감법을 쓰기위해 각기 다른 장소와 전화망으로 30회 측정하여 DAT를 통한 Clean Speech에 test 데이터인 잡음DB를 구성하고 여기서 동일 채널 잡음을 이용하여 실험하였다.

(2)J-RASTA처리

잡음처리 기법중 SS(Spectral Subtraction)는 음성신호보다 느린 변화성분을 가진 신호를 미세하게 처리하기에는 다소 부적합한 면이 있으므로 이를 보강하기 위해 RASTA, J-RASTA기법을 사용하는데 본 논문에서는 RASTA기법을 확장하여 SS + J-RASTA 기법을 순차적으로 적용함으로써 채널 왜곡과 부가잡음 모두에 대해 적응적인 필터링 기법을 사용 하였다. RASTA처리는 음성스펙트럼의 각 성분내에서 음성에 비해 느리게 변화하는 부분을 필터링을 통해 억제하는 방법이다. 이는 훈련시 사용된 음성과 다른, 채널왜곡에 따른 열화의 결과 성분에 대한 필터링 기법이다. 실험에 적용해본 결과로는 예상되었던 결과로서 약간의 음성신호에 자체에 대한 필터링으로 원 신호에 대한 신호오류를 가짐을 알수 있었고, log영역에서의 처리이므로 전제했던 채널왜곡에는 상당한 효과를 가질 수 있으나, correlation이 낮은 부가잡음의 처리에 약점을 내포하게 됨을 알 수 있었다.

이에따라 본 논문에서는 이를 확장 적용한 J-RASTA기법을 사용하였으며, 더불어 MFCC12차 특징벡터를 이용하였다.

$$y(t) = h(t) * (x(t) + d(t))$$

$x(t)$: pure speech signal,

$d(t)$: additive noise,

$h(t)$: convolutional noise

대수크기의 스펙트럴 도메인에서, 아래와 같이 컨벌루션되는 잡음을 선형적으로 분리가 가능하다.

$$\log Y(\omega) = \log H(\omega) + \log(X(\omega) + D(\omega))$$

음성신호에 비해 채널에 의한 왜곡은 천천히 변하므로, 첫단계로 HP를 통해 $\log H(\omega)$를 걸러낼 수 있다. log-RASTA처리에서, 아래와 같은 BP(band-pass filter)가 로그 대수크기 스펙트럼에 적용된다.

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

그러면, 결과로 $\log(X(\omega) + D(\omega))$를 얻을 수 있는데, 이는 부가잡음 $D(\omega)$에 영향을 받은 것이다.

파워 스펙트럼 도메인상에서, 선형적으로 채널에 의한 컨벌루션 잡음이 부가된 음성신호로 분리가 가능하다.

$$Y(\omega) = H(\omega)X(\omega) + H(\omega)D(\omega)$$

로그-RASTA처리에서 동일 방법을 이용해 비교적 천천히 또는 급격히 변하는 부가잡음을 감소시킬 수 있다. 그러나 결과는 필터 $H(z)$를 벗어난 컨벌루션 잡음과 부가적인 잡음에 영향을 받는다. 여기서 두가지 사항이 고려되는데, 첫째는 부가잡음과 채널에 의한 컨벌루션 잡음 모두가 동시에 줄여질 수 있어야 한다는 것이며, 두 번째로 $H(z)$를 통한 결과 신호가 위 식과 같다는 전제가 되어야 한다.

J-RASTA처리는 아래와 같은 입력스펙트럼 맵에 의해 위의 식 모두를 고루 만족한다.

RASTA의 log영역 처리 대신에 아래와 같은 근사화된 영역에서 처리한다.

$$X^{\wedge}(\omega) = \ln(1 + JX(\omega))$$

J= signal dependent constant

x= input

이 식을 이용한 warpping 영역은 아래와 같다.

$$X^{\wedge}(\omega) = \begin{cases} J<1 : linear-like \\ J\geq : log-domain \end{cases}$$

역변환은 $x = (e^{y} - 1)/J$ 이므로,

근사화시킨 역변환을 사용하였다. 위의 J값은 SN비에 따라 특정 최적값이 존재하며, 최적값은 아래식에 의해 구할수 있다.

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

이 과정을 거치면 스펙트럼의 음성신호성분이 비선형성을 가진 log영역내에 존재하게 되며, 잡음신호성분은 선형영역내에 존재하게 된다.

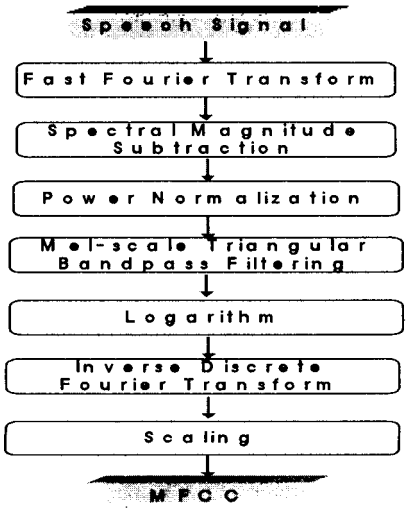
이때 만일 부가잡음이 지배적인 경우라면 J는 0에 가까울 것이고 그렇지 않은 경우라면 J값이 컨벌루션 잡음을 억제 하기위해 좀더 큰 값이어야 한다.

이 경우 $X^{\wedge}(\omega)$가 log-like하기 때문에 더 크게 조정될 때 컨벌루션된 잡음은 더 많이 줄게 되는 것이다. J-RASTA방법은 부가잡음과 컨벌루션 잡음에 대해 효과적으로 동작을 하지만, J-RASTA의 의의는 컨벌루션 잡음과 부가 잡음으로 인한 오류의 감소효과에 있다. 본 논문에서 현재까지의 실험결과로는 부가잡음과 컨벌루션잡음에 의한 영향이 매우 큰 경우에 두 가지 모두를 억제할 수 있는방안은 아직 이뤄지지 않았다. 따라서 두 잡음의 영향이 일정 제한된 영역임을 사전에 밝힌다.

(3)잡음에 강한 특징 추출

본 논문에서는 그간 연구에 사용되었던 PLP, LPC를 전처리 과정의 잡음 처리 기법을 포함하지 않은 채 특징벡터와 HMM 인식 알고리즘만을 비교대상으로 실험한 결과, 활발한 연구가 진행중인 MFCC계수에서 가장 나은 결과를 나타냈으며, MFCC + $\Delta\Delta$는 계산량이 많음에도 불구하고 성능에서 MFCC12차 + Δ의 경우와 별 차이를 보이지 않았다. 따라서 MFCC12차 + Δ를 사용하였다.

잡음에 강한 특징추출방법에 대표적인 것이 MFCC(Mel-Frequency Cepstral Coefficient)가 있다. 이는 본 연구에서 모델의 특징벡터를 뽑아내기 위해 사용한 HTK에서도 동일하게 MFCC가 효율적으로 쓰이는 점으로도 신뢰성이 인정된다. 이는 MFCC가 인간의 청각기관이 스펙트럼을 비선형적 주파수 스케일(log-scale)로 분석하는 것을 가장 잘 반영하여, 청각특성을 고려한 특징벡터로 최근까지 널리 쓰이고 있는 LPC에 비해 무잡음 환경에서나 잡음 환경에서 모두 그 성능이 우수하게 나타났기 때문이다. MFCC는 이러한 청각기관의 모델링으로 신호 스펙트럼을 mel-scale상의 동일간격을 갖는 필터뱅크로 분석하게된다.

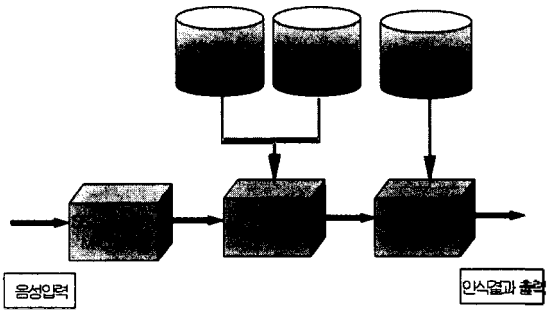


여기서 Mel-frequency와 물리적인 주파수는 아래와 같은 관계식으로 표현되며, MFCC의 추출은 위의 다이어그램과 같은 과정으로 추출한다.

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

(4)인식 시스템

본 논문의 test DB는 증권거래용의 전화채널을 이용한 증권거래 DB를 사용하였으며, 사전 훈련과정의 특징 추출의 효용성을 살피기 위해 핵심어 52개와 비핵심어 46개를 선정, 성인남성 30명을 대상으로 71문장을 구성하였고, Clean speech DB 50문장을 구성, 수작업에 의해 잡음을 감쇄하여, 훈련하였으며, test DB는, 전화 DB중 채널잡음으로 유추되는 잡음구간이 비교적 많이 분포되고, 부가잡음이 첨가된 DB중 5명 화자 DB를 랜덤 선정하여 테스트 하였다.



인식기는 현재 사용중인 단어 기반 인식기를 확장하여 연결어 인식기로 구성하였으며, 한 문장내에 핵심어가 최대 3회까지 포함될 수 있다. 기본적인 핵심어 인식 시스템은 핵심어 검출부분과 이의 검증부분으로 구성 하였다. 이중 핵심어 검출부에는 핵심어 모델과 필러모델을 구성하였으며, 이중 필러모델은 핵심어를 제외한 음성부나 비음성부로 구성하였다. 검출된 핵심어를 검증하는 부분은 핵심어의 Tri-Phone에 대한 Anti-model의 신뢰도 값을 결정, 사전에 정의된 임계값과 비교, 핵심어의 인정여부를 결정한다. 이때 반음소모델(Anti-model)은 전체 PLU(Phone Likely Unit)에서 특정 음소에 대한 반음소 모델을 구성한다고 할때, 해당 음소를 제외한 나머지 음소들로 구성된 음소배열이며, 이러한 음소열이 전체 PLU에 거쳐서 구성되게 된다. 이때 특정음소를 제외한 나머지 음소들의 Best Gaussian, 2nd Best Gaussian, 3nd Best Gaussian의 가중치, 평균, 분산을 취하여 적합한 모델을 구성하였다. 신뢰도 값은 다음과 같은 수식에 의해 구하였다. 사용된 PLU는 기존 사용중인 52개 PLU를 이용하였다.

$$s_i(O;\Theta) = \log\left[\frac{1}{M(i)}\sum_{q=1}^{M(i)}\exp\{f\cdot Lr_{(q)}(O_q;\Theta)\}\right]^{\frac{1}{f}}$$

여기서 신뢰도값이 임계값 τ_k이하인 경우 거절하게 된다. 특징벡터는 PLP, LPC, MFCC를 모두 실험하였으며 결과는 MFCC가 가장 우수한 성능을 보였다.

3. 실험결과 및 고찰

고립단어 인식기에서와 동일하게 PLP, LPC보다 MFCC가 더욱 우수한 성능을 보였다 이에 대한 결과는 http://www.yastalavista.ne.co.kr에 올릴 예정이며 음성 인식기에 MFCC에 전처리 단계에 제안된 방법을 적용하지 않은 경우와 전처리 단계에 제안된 방법으로 잡음에 대한 감쇄및 보상등의 과정을 거친 경우 실험결과를 나타내었다.

표 1-1 비교 실험 결과

항목 / mixture수	CA	FAI	FR	CR	FAO
GMM mixture 1	70.89	5.06	24.05	67.50	32.50
GMM mixture 2	69.51	8.54	21.95	77.50	22.50
GMM mixture 3	78.04	8.54	13.41	72.50	27.50

표 1-2 제안 실험 결과

항목 / mixture수	CA	FAI	FR	CR	FAO
GMM mixture 1	71.39	2.10	26.51	71. 48	28.52
GMM mixture 2	69.91	6.55	23.54	77.90	22.10
GMM mixture 3	79.63	4.21	16.16	79.95	20.05

표 3 Mixture3의 dB별 인식률 결과(CA+CR/2)

target SNR dB	Clean speech%	Noise speech%
0	79.81	79.79
6	80.98	79.99
12	84.76	84.53
18	88.39	88.39
24	88.94	88.90

4.결론

본 논문에서 제안된 Spectral Subtraction, J-RASTA, MFCC의 이용을 통하여 전처리과정을 거치지 않은 경우와 SS + J-RASTA를 거친 경우 그 성능에 있어 Mixture3일 때 가장 성능이 우수하게 나타났으며, 각 Mixture에 대해 비교적 고른 성능향상을 보였다. 주지할 점으로 FAI에 비해 FR의 증감폭이 큰 결과와 CR이 증가하는 결과를 가져온 것은 전처리 기능에 의한 것으로, 전처리과정에서 잡음에 대한 보상이 특징벡터열에 영향을 미쳤으며, 이 효과가 바르게 인식된 결과중에서의 오인식 부분이 현저히 감소한 모습으로 나타났다. 또한 Baseline 실험에서는 인식성능이 낮았는데 이는 SNR이 낮음으로 인해 음성신호가 잡음에 묻히는 경우로 추정되었다. 또 DAT에 PC환경을 이용한 본 실험과는 별도의 Clean DB에 전화DB의 잡음을 첨가하여 동일 실험을 한 결과는 target 잡음 레벨이 음성정보의 일부를 마스킹하는 것으로 추정되는 신호의 왜곡으로 인하여 잡음레벨을 높일수록 인식성능이 현저히 떨어졌다. 향후 연구 방향으로 현제 시스템의 성능문제중 시간복잡도를 고려한 속도 개선에 대해 진행하고, 음성신호의 한 특징인 harmonic sieving에 대해 연구를 진행하여 속도, 인식률 향상을 진행할 계획이다.

1. 김우성, 구명완 " 반음소 모델링을 이용한거절기능에 대한 연구" 한국음향학회지 18권 3호
2. 김광수, 정현열 "Histogram 처리와 Noise Threshold를 이용한 음성인식기의 환경잡음 처리 성능 향상", 1998년
3. 전선도, 강철호 "배경잡음에 강인한 음성 인식 시스템에 관한 연구" 1998년
4. 전선도, 강철호 "잡음에 강한 음성 인식에서 SNR 기준함수를 사용한 가우시안 함수 변형 및 결정에 관한 연구" 한국음향학회 제18권 7호
5. 정회인, 김형순 "신호대 잡음비의 정규화를이용한 잡음환경에서의 음성인식" 1998년
6. Abdallah, I., Montresor, S.&Bauding, M. "Speech signal detection in noisy environment using a local entropic criterion, in 'European Conference on Speech Communication & Technology" 1997년

본 연구는 한국과학재단 목적기초연구 R01-2000-00276 지원으로 수행되었음

This work was supported by grant No. R01-2000-00276 from the Korea Science & Engineering Foundation