1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

부가 잡음 환경에서의 음성인식을 위한 비선형 변환을 이용한 캡스트럼 정규화 기법

*석용호, *최승호, *이황수

*서울 동대문구 청량리동 207-43 한국과학기술원 정보 및 통신공학과

Cepstral Normalization using Non-Linear Transform for Speech Recognition in Additive Noise Environments

*Yong Ho Suk, *Seung Ho Choi, *Hwang Soo Lee

*Depart. of Inform. and Comm. , KAIST

{bubble, shchoi, hslee}@spectra.kaist.ac.kr

요 약

본 연구에서는 입력 음성 특징 파라매타를 선형 및 비선형 변환함으로써 음성 특징의 1 차, 2 차 및 고차 통계치를 정규화하였다. 이러한 정규화 기법을 통해서 부가잡음 환경에서의 음성인식 성능향상을 얻을수 있었다.

1. 서론

잡음 섞인 음성 신호의 특징 벡터의 통계적 성질은 잡음이 섞이지 않은 음성 신호의 특징 벡터와 다르며 이러한 통계적 성질은 잡음의 종류, 신호 대 잡음비(SNR) 등의 잡음 조건에 따라 달라진다. 잡음 섞인 음성의 특징 벡터의 평균 벡터와 분산 행렬은 잡음 섞이지 않은 음성 특징의 평균 벡터와 분산 행렬에 비해 차이가 있으며 특히 부가 잡음의 경우 캡스트럼 계수의 각 분산이 줄어든다.[1]

음성 인식 시스템의 학습 환경과 인식 환경과의 불일치를 줄이기 위해 많은 보정 방법이 연구되었다. 이러한 방법은 parallel model combination (PCM)[2]과 같이 패턴 정합 단계에서 수행될 수 있으며 RATZ[3], 통계정 정합 알고리즘[4] 에서와 같이 음성 특징 단계에서 수행될 수도 있다. 그러나 RATZ[3]나 통계적 정합 알고리즘[4]의 경우 많은 양의 적응 데이터가 필요하며 또한 잡음 환경이 변할 경우 그에 따른 보정이 필요하다는 단점이 있다.

본 논문에서는 발성음 단위의 선형 및 비선형 변환을 이용한 캡스트럼 정규화 기법을 제안하고 이를 이용한 부가 잡음 환경에서의 음성인식 실험을 수행하였다. 1 차 및 2 차 통계치인 평균과 분산을 각 발성음 별로 선형 변환을 통해서 정규화한다. 또한 3 차 통계치 정규화를 비선형 변환에 의해 수행한다. 이러한 정규화 기법은 부가 잡음환경에서의 음성 특징의 선형 및 비선형 왜곡을 감소시킨다. 백색 가우시안 부가 잡음 환경하에서 인식실험 결과 상당한 인식율 향상을 얻을 수 있었다.

2. 음성 특징 정규화

음성 특징 정규화는 음성 특징을 잘 정의되고 정규화된 환경으로 변환시키는 방법이다. 이때 학습 환경과 인식 환경의 차이가 이러한 정규화를 통해서 감소되도록 정규화시킨다. 본 논문

에서는 발성음 단위로 음성의 캡스트럼 특징 벡터의 1차 및 2차 통계치인 평균과 분산을 각 발성음별로 선형 변환을 통해서 정규화한다.
또한 3차 통계치의 정규화는 비선형 변환에 의해 수행한다. 그 과정은 그림 (1)과 같다.

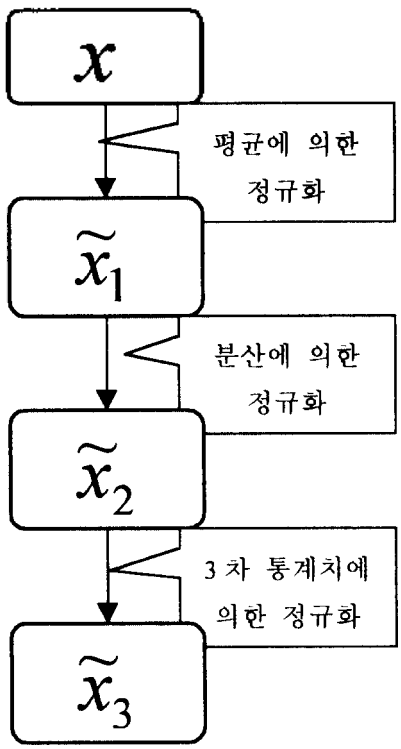

그림 1. 음성 특징의 정규화

3. 1차 및 2차 정규화와 선형 변환

음성 특징의 정규화 과정중 첫번째 과정은 평균의 정규화이다. 즉 (식 1)과 같이 발성음 별로 캡스트럼 특징 벡터의 평균을 구한 뒤 이를 빼준다.

$$\tilde{x}_{1,t} = x_t - E[x_t] \quad (1)$$

$$E[x_t] = \frac{1}{N}\sum_{t=1}^{N} x_t$$

이러한 1차 정규화의 결과는 평균이 0인 특징 벡터가 되며 Cepstral Mean Normalization (CMN)[5] 와 동일하다.
1차 정규화의 결과에 분산에 의한 2차 정규화를 수행시킬 수 있다. 즉 (식 2) 에서와 같이 각 특징벡터를 발성음 단위로 구한 공분산 행렬의 행렬 제곱근의 역행렬로 나눠주면 평균 0 벡터, 단위 공분산 행렬을 가지는 2차 정규화 벡터를 얻는다.

$$\tilde{x}_{2,t} = \sqrt{V^{-1}}\, x_{1,t} \quad (2)$$

$$V = E[x_{1,t} x_{1,t}^T]$$

이러한 1차 정규화와 2차 정규화를 종합하면 (식 3)과 같은 선형 정규화를 얻는다.

$$\tilde{x}_{2,t} = \sqrt{V^{-1}}(x_t - E[x_t]) \quad (3)$$

일반적으로 공분산행렬은 대각행렬로 근사화시킬 수 있으며 따라서 공분산 행렬의 행렬 제곱근과 이의 역행렬은 (식 4)와 같이 간략화 된다.

$$\sqrt{V} = diag(\sqrt{\sigma_{ii}^2}) \quad (4)$$

$$\sqrt{V^{-1}} = diag(1/\sqrt{\sigma_{ii}^2})$$

4. 3차 정규화와 비선형 변환

확률 변수 X 가 가우시안 분포를 따를 때 1차 통계치 $E(X)$와 2차 통계치 $E(X^2)$가 정해지면 고차 통계치 $E(X^n)$, (n >=3)는 일의적으로 결정된다. 따라서 가우시안 확률 변수의 고차 통계치와 주어진 확률 변수의 고차 통계치의 차이를 가우시안 분포에서 벗어나는 척도로 간주할 수 있다.[6] 따라서 본 논문은 음성 특징 벡터의 3차 통계치를 가우시안 확률 변수의 3차 통계치로 변환시키는 비선형 변환과 이러한 변환의 과

라메타 값을 구하는 방법을 제안한다.

1 차 통계음성 특징의 2 차 정규화 과정은 (식 3)의 선형 변환으로 나타낼 수 있다. 그러나 3 차 이상의 통계치를 정규화하기 위해서는 비선형 변환이 요구된다. 따라서 (식 5)와 같은 비선형 변환으로 나타내었다. 이때 캡스트럼 벡터의 각 차수간의 상호연관이 존재하지 않는다고 가정하였으며 따라서 각 차수별로 비선형 변환을 수행하였다.

$$\tilde{x}_{3,t} = a\tilde{x}_{2,t}^2 + b\tilde{x}_{2,t} + c \tag{5}$$

이때 평균, 분산, 3 차 통계치의 정규화에 따라 파라메터 a, b, c 를 구한다.

$$E[\tilde{x}_{3,t}] = 0 \tag{6}$$

$$\sigma_{\tilde{x}_3}^2 = E[\tilde{x}_{3,t}] = 0 \tag{7}$$

$$E[\tilde{x}_{3,t}^3] = 0 \tag{8}$$

여기서 3 차 정규화된 음성 특징의 평균값에 대한 관계와 분산값의 근사화[6] , 3 차 통계값의 정규화 과정에 따라 다음과 같은 식들 얻을 수 있다.

$$\begin{aligned} E[\tilde{x}_{3,t}] &= aE[\tilde{x}_{2,t}^2] + bE[\tilde{x}_{2,t}] + c \\ &= a\sigma_{\tilde{x}_2}^2 + b\cdot 0 + c \\ &= a + c = 0 \end{aligned} \tag{9}$$

$$\therefore \quad c = -a$$

$$\sigma_{\tilde{x}_3}^2 \approx \left|\frac{d\tilde{x}_3}{d\tilde{x}_2}\right|^2 \sigma_{\tilde{x}_2}^2 = b^2\sigma_{\tilde{x}_2}^2 \tag{10}$$

$$\therefore \quad b = 1$$

$$\begin{aligned} E[\tilde{x}_{3,t}^3] = {} & a^3\cdot\{E(\tilde{x}_{2,t}^6) - 3E(\tilde{x}_{2,t}^4) + 2\} \\ & + a^2\cdot\{3E(\tilde{x}_{2,t}^5) - 6E(\tilde{x}_{2,t}^3)\} \\ & + a\cdot\{3E(\tilde{x}_{2,t}^4) - 3\} \\ & + E(\tilde{x}_{2,t}^3) \\ & = 0 \end{aligned} \tag{11}$$

(식 11)은 a 에 대한 3 차 방정식이며 3 개의 근을 가진다[8]. 이러한 3 개의 근 중 절대값이 가장 작은 것을 선택한다. 이것은 비선형 성분을 최소화하기 위함이다.

5. 인식 실험 및 결과

이 논문에서 이용한 음성 DB 는 20 명의 남성화자가 100 단어를 2 회 반복 발성한 것을 8kHz, 16 bit 로 샘플링한 것이다. 그중 12 명의 음성 데이터로 학습하고 나머지 8 명의 음성 데이터로 인식 실험하였다. 부가 잡음은 20 dB, 10 dB, 5 dB, 0 dB 신호 대 잡음 비의 백색 가우시안 잡음을 이용하였다. 음성 특징 분석은 다음과 같다. 분석 구간은 30 msec frame 이며 20msec overlap 을 주었다. 전처리 과정은 (1- $0.98z^{-1}$)로 Preemphasis 한뒤 Hamming Window 를 부가시켰다. 그후 12 차의 MFCC 음성 특징 파라메터를 추출하였다. 이렇게 추출된 음성 특징을 1 차, 2 차 및 3 차 정규화 과정에 의해 정규화 시켰으며 각각의 정규화 에 의한 학습과 인식을 수행하였다.

학습과 인식에 이용된 HMM 은 Simple left to right Semi-contiuous HMM, codebook 크기는 128 이며 대각 공분산 행렬을 이용하였다. HMM 의 상태수는 모든 단어에 대해 10 개로 통일시켰다. 각 정규화 방법에 대한 결과는 그림 (2)와 같다.

6. 결론

본 연구에서는 입력 음성 특징 파라메타를 선형 또는 비선형 변환함으로써 음성 특징의 통계치를 정규화하였다. 음성 특징의 3 차 이상의 통계치를 정규화하기 위해서는 비선형 변환이 필수적이며 이러한 비선형 변환의 파라메타를 구하는 방법을 제안하였다. 이때 3 차 방정식의 근을 구해야 하므로 수치적 안정성에 문제가 있을 수 있다. 따라서 2 차 이하의 방정식으로 근사화하는 방법이 필요이러한 정규화 기법을 통해서 부가잡음 환경에서의 음성인식 성능향상을 얻을수 있었다.

ACKNOWLEDGEMENTS

본 연구는 과학재단의 수탁과제 연구지원에 의해 수행되었습니다. (과제번호 : 96-0102-15-01-3)

참고문헌

[1] A. Acero, Acoustical and environmental robustness in automatic speech recognition, Kluwer Academic

[2] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. On Speech and Audio Process., Vol. 4, No. 5, pp. 352-359, Sep. 1996.

[3] P. J. Moreno, B. Raj, E. Gouvea, and R. M. Stern, "Multivariate Gaussian based epstral normalization for robust speech recognition," in Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 137-140, May 1995.

[4] A. Sankar and C. H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. on Speech and Audio Process., vol. 4, no. 3, pp. 190-202, May 1996.

[5] Furui, S. "Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, ASSP-29, pp. 254-272, 1981.

[6] C. Nikias, A. P. Petropuou, Higher Order Spectra Analysis

[7] Athanasios Papoulis, "Probability, Random Variables, and Stochastic Processes", 3rd Ed., pp.112 – 113, McGraw-Hill

[8] Milton Abramowitz and Irene A. Stegun, "Handbook of Mathematical Functions With Formuls, Graphs, and Mathematical Tables", p.p. 17

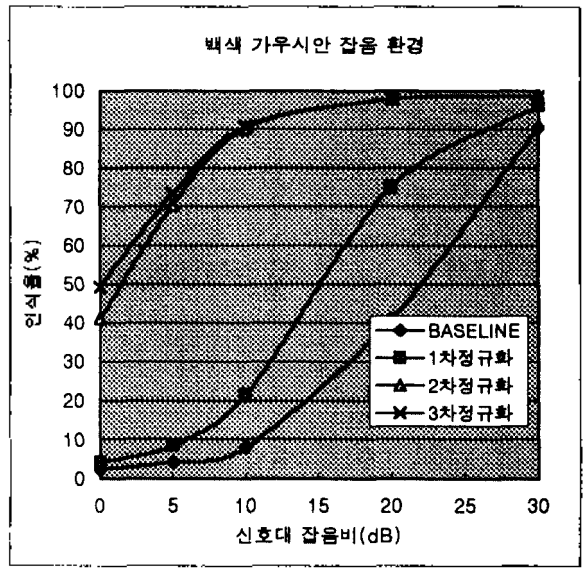

그림 2. 백색 잡음 환경에서의 인식 실험 결과