韓國 OA學會 論文誌
第7卷 第2號, 2002. 6.

2002-7-2-2-3

음성인식 시스템에서의 잡음 제거 개선에 관한 연구

이 창 윤* 이 영 훈**

Study of the Noise Processing to Technique Speech Recognition System

Chang-Yun, Lee* Young-Gun, Lee**

요 약

본 논문에서는 음성인식 시스템에서의 잡음 처리 기술로서 SNR 정규화와 RAS를 결합한 방법을 사용하여 여러 가지 잡음 처리 방법을 연구하여 인식 시스템의 성능을 개선하였다. 인식 시스템으로는 범용 DSP (TI사의 TMS320C31)가 내장된 모듈을 사용하였다. 실험에 사용된 인식 단어 샘플은 일반 사무 및 컴퓨터의 명령을 위한 60단어이며. 일반환경에서 잡음과 함께 가상의 여러 유색 잡음을 고려하여 샘플된 데이터를 시뮬레이션 하였다.

녹음된 데이터에 대한 컴퓨터 시뮬레이션 상에서 잡음 처리 방법으로 SNR정규화와 스펙트럼 차감법을 결합하여 실험한 경우 최고 94.61%의 높은 인식 성능을 보였다.

Abstract

Recognition system of noise processing technique. A method combining SNR normalization with RAS is considered as a noise processing and the performance of the speech recognition system can be improved using other noise processing technique.

Experiment of recognition system is the internal organs that using a general digital signal processor(TMS320C31). Recognition word set is composed of 60 command words for office environment and order of computer. Simulation is considered as a colored noise of general environment. The results of experiment showed that the recognition word set gives 94.61% of efficiency of recognition at maximum in case of the combination of SNR normalization and spectral subtraction.

* 한남대학교 대학원 전자공학과 졸업
** 한남대학교 전자통신공학부(전자공학전공)교수

논문접수 : 2002. 3. 29
심사완료 : 2002. 6. 4

Ⅰ. 서론

음성인식 시스템은 잡음이 없는 환경에서는 만족스러운 결과를 얻을 수 있지만. 주변 잡음이 많은 환경에서는 만족스런 인식성능을 기대하기 어렵다. 이러한 문제는 배경 잡음과 채널 효과가 존재하는 환경에서 사용될 때는 필연적으로 존재하게 되므로 배경 잡음과 채널의 영향을 받지 않는 강인한 음성인식 시스템을 요하게 된다.

본 논문에서는 잡음 처리방법으로 SNR 정규화와 RAS(Relative Autocorrelation Sequence)를 결합한 것을 제안하였으며. 또한 다른 방법으로서 스펙트럼 차감법(Spectral Subtraction)[1]을 사용하여 실험을 하였고 널의 효과를 제거하기 위해 켑스트랄 평균 차감법(Cepstral Mean Subtraction)[2]을 이용하였다. 인식 알즘은 HMM중 인식률이 높은 반 연속 HMM[3][4]을 이용였다.

본 논문의 구성은 Ⅰ장의 서론에 이어 Ⅱ장에서는 잡음처리 기술. Ⅲ장에서는 음성인식기술에 관하여 설명하며. Ⅳ장에서는 다양한 실험과 결과에 대해 설명하고, Ⅴ장에서는 결론을 통해 제안된 방법의 성능이 우수함을 입증한다.

Ⅱ. 잡음처리 기술

잡음에 강한 특징 벡터로는 멜 켑스트럼 계수와 루트 켑스트럼 계수가 있다. [5,6]

1. 멜 켑스트럼 계수

멜 켑스트럼은 DFT(Discrete Fourier Transform) 또는 FFT(Fast Fourier Transform) 크기를 멜과 주파수 사이의 대응관계에 따라 주파수 축에서 와핑(warping)하여 이의 대수 값을 역 DFT하여 8차에서 14차 정도의 계수를 구한다.

20개의 삼각 대역 통과 필터를 이용하여 임계대역 필터를 통과한 로그 에너지 출력을 Xk라 하면 M 개의 켑스트럼 계수는 다음과 같이 나타내어진다.

$$C_n = \frac{1}{20}\sum_{k=1}^{20} X_k \cos[n(k-\frac{1}{2})\frac{\pi}{20}] \quad (1)$$

2. RPS(Root Power Sum)

스펙트럼의 기울기에 의한 거리 측정 d_{RPS}는 다음과 같이 표현된다.

$$d_{RPS} = \int_{-\pi}^{\pi} \left| \frac{\omega}{\partial\omega} \ln \frac{\sigma^2}{|A(e^{jw})|^2} - \frac{\omega}{\partial\omega} \ln \frac{\sigma'^2}{|A'(e^{jw})|^2} \right|^2 \frac{d\omega}{2\pi} \quad (2)$$

여기서 , $\sigma/A(e^{jw})$와 $\sigma'/A'(e^{jw})$는 선형 예측 방법에 의해 얻어진 두 모델 스펙트럼이다.

3. 잡음 제거

1) 스펙트럼 차감법

잡음신호 n(k)와 음성신호 s(k)에 더해졌을 때 잡음섞인 음성신호 x(k)는

$x(k) = s(k) + n(k)$ (3)이다.

식을 퓨리에 변환하면 다음과 같다.

$X(\omega) = S(\omega) + N(\omega)$ (4)이고,

$$X(\omega) = \sum_{k=0}^{N-1} x(k)\, e^{-j\omega k} \quad (5)$$

$$x(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)\, e^{j\omega k} d\omega \quad (6)$$ 이다.

이때, 스펙트럼 차감 필터 $H(\omega)$는 미리 구한 잡음의 스펙트럼 $N(\omega)$의 평균을 사용한다.

$$H(\omega) = \frac{|X(\omega)| - a\mu(\omega)}{|X(\omega)|} \quad (7)$$

여기서

$$\mu(\omega) = E[|N(\omega)|] = \frac{1}{M}\sum_{i=0}^{M-1} |N_i(\omega)| \quad (8)$$

이다.

이러한 과정의 결과인 잡음 제거된 신호는 다음과 같다.

$$S'(\omega) = [|X(\omega) - \mu(\omega)|]\, e^{j\theta x(\omega)} \quad (9)$$

2) SNR 정규화

음성 신호의 스펙트럼이 멜 밴드를 통과한 후 각 밴드들은 각 밴드의 동적 범위에 의해 적응적으로 마스킹 상수가 더해져 각 밴드의 SNR 정규화가 이루어진다. 측정된 순시 SNR보다 작은 경우에는 마스킹 상수가 시상수에 의해 감소 되고, 순시 SNR의 값이 표적 SNR보다 큰 경우에는 일정한 크기에 의해 마스킹 상수가 증가된다. 여기서, 순시 SNR은 각 밴드의 최대와 최소의 비에 의해 구해진다.[7]

3) RAS (Relative Autocorrelation Sequence)

채널의 왜곡과 부가적 잡음에 오염된 음성신호를 식으로 표현하면 다음과 같다.

$$y(m,n)=x(m,n)\otimes h(n)+w(m,n) \qquad (10)$$

여기서 m은 프레임을 나타내고, n은 한 프레임에서의 이산시간 이다. x는 오염되지 않은 원신호, y는 잡음 신호, h는 채널의 임펄스 응답, w는 부가잡음을 의미한다.

부가잡음이 정적이고 각 음성신호들이 상관관계가 없다고 가정을 하면 다음 식으로 나타낼 수 있다.

$$r_{yy}=r_{xx}(m,k)\otimes h(k)\otimes h(-k)+r_{ww}(k) \qquad (11)$$

여기서 k는 프레임에서의 자기 상관 순열 인덱스를 나타낸다.

식을 프레임에 대해 미분하여 정리하여 근사화 하면 식 12와 같이 표현된다.

$$\frac{\partial}{\partial m} r_{yy}(m,k)=\frac{1}{T_L}\sum_{t=-L}^{L} t\cdot r_{yy}(m+t,k)$$
$$0\le m\le M-1, 0\le k\le N-1 \qquad (12)$$

4. 채널영향 제거 방법

음성신호에 가해진 왜곡의 영향이 선형 필터로 표현된다면, 왜곡은 관찰된 신호를 역 필터링 함으로 제거될 수 있다. 켑스트럼 영역에서는 왜곡의 영향은 관찰된 신호의 켑스트럼에서 왜곡과 관련된 켑스트럼을 빼줌으로 제거될 수 있다. 채널 왜곡 특성이 음성 신호의 관찰 구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 켑스트럼의 추정치는 관찰된 신호의 켑스트럼의 평균으로 구해질 수 있다. [8,9,10]

Ⅲ. 음성인식 기술

1. 음성신호 분석

음성신호가 전체적으로는 시변 신호이지만 보통 20ms 구간에서는 정적인 성질을 가지고 있다고 가정하여 이 구간을 한 단위로 분석을 수행한다.

1) 켑스트랄 분석방법

켑스트럼 계수{Ck}는 선형 예측 계수{ak}와 최소자승 평균 오차 Emin을 식 13 에 대입하여 구할 수 있다.[10]

$$C_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}\log|S(\omega)|\,e^{j\omega k}d\omega \qquad (13)$$

$$C_0=\log E_{\min} \qquad (14)$$

$$C_k=-a_k-\sum_{i=1}^{k-1}\frac{i}{k}C_i\,C_{k-i}, 1\le k\le P \qquad (15)$$

$$C_k=-\sum_{i=1}^{P}\frac{i}{k}C_i\,C_{k-i}, k>P \qquad (16)$$

2. 거리 측정 방법

켑스트럼 계수를 사용하는 거리 측정 방법은 두 패턴의 대수 스펙트럼의 차이를 구하는 방법으로서, 테일러 급수전개에 따른 켑스트럼 계수{Ck}는 식 17 같이 정의된다.[11,12]

$$\ln A(z)=-\sum_{k=1}^{\infty}C_k\,z^{-k} \qquad (17)$$

스펙트럼 모델에 대하여 대수 스펙트럼 거리(log spectrum distance) d_{CEP} 를 정의하면

$$d_{CEP}=\int_{-\pi}^{\pi}\left|\ln\left(\frac{\sigma^2}{|A(e^{j\omega})|^2}\right)-\ln\left(\frac{\sigma'^2}{|A'(e^{j\omega})|^2}\right)\right|^2\frac{d\omega}{2\pi} \qquad (18)$$

이고 식 18 에 Parserval의 정리를 적용하면 다음과 같다.

$$d_{CEP}=\sum_{k=-\infty}^{\infty}(C_k-C'_k)^2$$

$$= (C_0 - C'_0)^2 + 2 \sum_{k=1}^{\infty} (C_k - C'_k)^2 \quad (19)$$

P차까지 근사화한 켑스트랄 거리측정 방법은 다음과 같다.

$$d_{CEP} = \sum_{k=1}^{P} (C_k - C'_k)^2 \quad (20)$$

3. HMM을 이용한 음성인식

HMM은 음성신호의 시간적 변화를 Markov 프로세스로 모델링 하며, 이중 확률과정으로 음성의 단구간 스펙트럼과 시간적 변화를 각각 모델링 하는 구조를 갖는다. 학습 데이터를 사용하여 모델 파라메터를 추정한 후, 미지의 입력 데이터가 어떤 모델에 해당되는가 하는 것을 확률적으로 판단하는 것이다.

1) 반연속 HMM

반연속 HMM에서는 전체M 개의 가우시안 분포를 이용하여 관찰열의 전체 분포를 모델링 하며 측정 상태에서 관찰열 Ot의 확률은 이들 M개의 가우시안 분포 중에서 F개만을 사용하여 나타내게 된다. 반연속 HMM은 가우시안 분포의 수 M을 적절히 선택하면 연속HMM에서와 같이 분포를 잘 나타낼 수 있다.

2) HMM을 이용한 음성인식의 구현에서의 고려점

음성신호는 시간에 따라 변한다. 따라서 HMM의 상태간의 천이를 시간이 흘러가는 것처럼 한쪽 방향으로만 천이할 수 있도록 하면 음성 신호의 시간 천이 특성을 잘 모델링할 수 있다. 또한 HMM으로 전후방 알고리즘을 이용하여 전후방의 변수들을 구할 경우 0과 1사이의 확률값들을 계속 곱해야 하기 때문에 언더플로우(underflow)현상이 일어난다. 이러한 문제를 해결하기 위하여 정규화를

해주는데, 이것을 스케일링(scaling)이라 하며 전후방 변수를 스케일링값 Ct로 정규화 시키는 것을 말한다.[13]

$$C_t = \frac{1}{\sum_{i=1}^{N} a_t(i)} \quad (21)$$

Ⅳ. 실험 및 컴퓨터 시뮬레이션

사용된 특징 벡터는 12차 MFCC와 12차 델타 MFCC, 그리고 2차 델타 에너지를 결합한 것으로 구성되고, 가중함수는 RPS를 사용하였다. 잡음 제거 방법으로 SNR 정규화를 사용하였으며 채널의 제거 방법으로는 켑스트랄 평균 차감법을 사용하여 실험을 수행하였다.

제안된 방법을 실험하기 위해 사용된 잡음은 대부분의 에너지가 400Hz이하의 저주파 영역에 집중되어진 유색잡음으로 만들어졌고, 본 논문에서는 이러한 잡음을 제거하기 위하여 250Hz의 고역 통과 필터를 사용하였다. 이렇게 하여 잡음이 첨가된 음성신호에 거의 영향을 미치지 않으면서 상당부분의 잡음을 제거 할 수 있다.

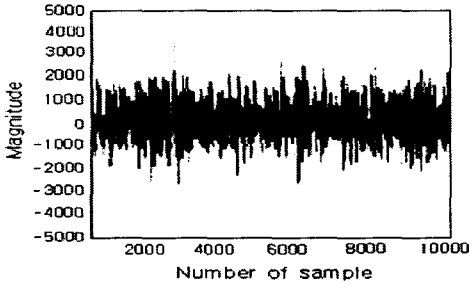

그림 2. 유색 잡음 음성 파형

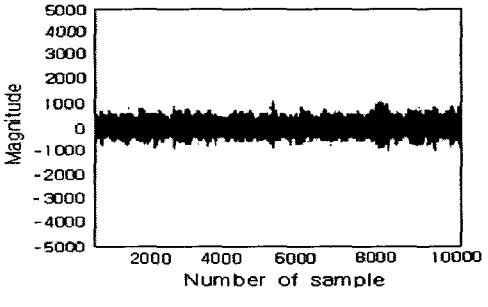

그림 3. 고역통과 필터링된 유색잡음

구현된 인식 시스템의 인식 성능을 알아보기 위하여 잡음환경에서 20명의 화자가 명령어 60단어를 3회씩 발음한 데이터를 가지고 컴퓨터 상에서 시뮬레이션 하였다.

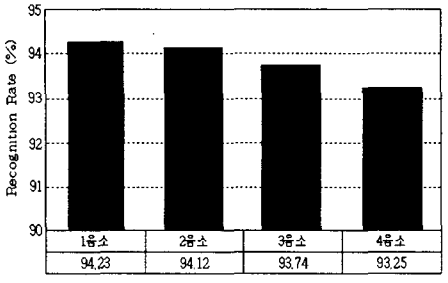

그림 4. 음소별 인식성능

V. 결론

본 논문에서는 음성 인식 시스템에서 잡음처리 방법으로 SNR 정규화와 RAS를 결합한 것을 제안하였으며, 여러 가지 잡음처리 방법을 연구하여 음성 인식 시스템의 성능을 향상 시켰다. 음성 인식 알고리즘으로는 반연속 HMM을 사용하였으며 명령어 60단어에 대한 인식을 수행하였다.

잡음처리 방법으로 스펙트럼 차감법과 SNR 정규화 방법을 사용하였으며, 채널의 영향을 제거하기 위하여 켑스트랄 평균 차감법을 이용하였다. 이러한 실험 결과 상태수를 5개로 고정한 경우 표적 SNR이 18dB일 때 최고 94.61%의 인식 성능을 보였다. 향상된 시스템은 기존의 채널의 영향만을 고려하는 시스템보다 2.11%의 인식 성능의 향상을 보였다.

앞으로의 연구과제는 보다 열악한 잡음환경에서의 적응력 강한 잡음제거 방안에 대한 연구이다.

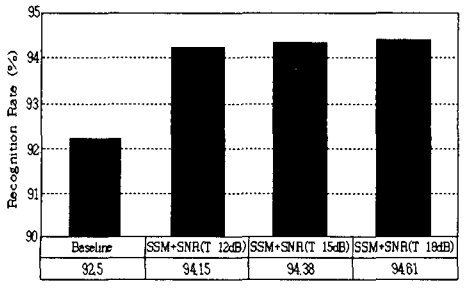

그림 5. 상태수 5개 일때의 인식 성능 비교

참고문헌

[1] P. Lockwood and J. Boudy, "Experiment with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars." in Proc. Eurospeech, pp. 79-82. 1991

[2] Jean-Claude Junqua and Jean-Paul Haton, Robustness in Automatic Speech Recognition Fundamentals and Applications, Kluwer Academic Publishers, 1996.

[3] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," Proc. IEEE, vol. 77, no. 2, pp. 257-285, 1989.

[4] X. D. Huang, Y. Ariki and M. A. Jack, Hidden Markov models for Speech Recognition, Edinburgh University Press, 1990

[5] S. F. Boll, "Suppression oacoustic noise in speech using spectral subtraction," IEEE Trans. Acoust, Speech, Signal Processing, vol. 27, no. 2, pp.113-120, Apr. 1979.

[6] S. V. Vaseghi, Advance Signal Processing and Digital Noise Reduction. New York: Wiley, 1996

[7] T. Claes and D. Van Compernolle, "SNR-normalization for robust speech recognition," in Proc. IEEE Int. Conf Acoust, Speech Signal Processing, vol. 1, pp. 331-334. May 1996.

[8] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," IEEE Trans,

Speech Audio Processing, vol. 2, no. 1, part. 1, pp. 97-102, Jan. 1994.

[9] N. Nocerino, F. K. Soong, L. R. Rabiner and D. H. Klatt, "Comparative Study of Several Distance Measures for Speech Recognition," Proc. ICASSP, vol. 1, pp. 25-28, Mar. 1985

[10] A. H. Gray and Jr. J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans, Acoust, Speech Audio Processing, vol. ASSP-24, no. 5, pp. 380-391, Oct. 1976.

[11] huzo Saito and Kazuo Nakata, Fundamental of Speech Signal Processing, Academic Press, 1985

[12] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication vol. 12, pp. 231-240, 1993

[13] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans, Acoust, Speech Signal Processing, vol. ASSP-26, no. 1, pp. 43-49. Feb 1978.

저 자 소 개

이 창 윤

2002.2 한남대학교 대학원졸업

이 영 훈

한남대학교 전자공학과 교수