음질 개선을 통한 음성의 인식

Speech Recognition through Speech Enhancement

조 준 희*, 이 기 성**
(Junhee Cho and Keeseong Lee)

* 홍익대학교 전기정보제어공학과(전화:(02)325-7514, 팩스:(02)320-1110, E-mail : cch1022@hitel.net)
** 홍익대학교 전기정보제어공학과(전화:(02)320-1670, 팩스:(02)320-1110, E-mail : leeksyh@yahoo.com)

Abstract : The human being uses speech signals to exchange information. When background noise is present, speech recognizers experience performance degradations. Speech recognition through speech enhancement in the noisy environment was studied. Histogram method as a reliable noise estimation approach for spectral subtraction was introduced using MFCC method. The experiment results show the effectiveness of the proposed algorithm.

Keywords : Speech Recognition, MFCC(Mel Frequency Cepstrum Coefficient), Histogram, Speech Enhancement, Spectral Subtraction

Ⅰ. 서론

인간은 오랜 시간동안 문자, 음성, 기호 등 다양 방법들을 사용해 정보를 교환해 왔다. 이 중에서 음성은 간편하며, 보편적인 정보 교환의 수단이다. 그래서 많은 연구가들이 이런 음성의 특징들을 이용해 인간과 기계가 상호 커뮤니케이션을 할 수 있도록 만들었으나 현재의 음성 인식 기술은 잡음이 섞인 환경에서는 인식률이 낮다. 때문에 많은 연구자들이 잡음 환경에서 음성인식기의 성능을 향상시키기 위해 많은 노력을 기울이고 있다[1,2].

음질 개선이란 잡음 환경 속에서 손상된 음성 신호를 인식하기 위한 전처리 단계로서, 손상된 음성 신호의 파형 또는 파라미터를 복구하는 방법이다. 대표적인 방법으로 스펙트럼 차감법, comb필터, 베이시안 추정법[10] 등이 있다. 특히 스펙트럼 차감법[6]은 다른 필터를 이용해 처리하는 방법보다 속도 면에서 우수할 뿐만 아니라 이론적으로 간단하고 널리 사용되는 방법이다. 그러나 스펙트럼 차감법은 음성과 비 음성 구간을 나눠 잡음 스펙트럼을 구하기가 어렵고 신뢰도가 떨어진다.

히스토그램을 통한 스펙트럼 차감법은 스펙트럼 상에서 음성과 잡음의 히스토그램을 통해서 어떤 임계값을 구해 음성의 영역과 잡음의 영역을 분리하여 잡음을 제거는 방법이다. 따라서 각 프레임 별로 존재하는 잡음과 음성의 스펙트럼을 히스토그램을 통해서 분리한다면, 시간에 따라 변하는 잡음을 제거할 수 있어 신뢰성 있고 좋은 음성 특징을 구할 수 있다.

인식률을 향상시키기 위한 방법으로 히스토그램을 통한 스펙트럼 차감법을 사용하였고, 특징 추출방법으로 MFCC(Mel Frequency Cepstrum Coefficient) 방법을 사용했다.

Ⅱ. 본론

그림 1은 음성이 인식되는 과정을 4가지 단계로 나눠 나타낸 것이다. 첫번째 Pre Processing과정은 음질 개선하여 잡음을 제거하거나 더 좋은 특징을 추출할 수 있게 도와주는 과정이다. 두번째 Feature Extraction과정은 특징을 추출하는 부분으로, LPC (Liner Prediction Coding)방법, MFCC방법 등이 있다. 세번째 Classification과정은 신경망(Neural Network), HMM (Hidden Markov Model), DTW (Dynamic Time Wapping) 등을 이용해 인식하는 과정을 말한다. 마지막 Post Processing 과정[5]은 후처리 과정이라고 하며, 단어사전, 문법, 의미 등을 고려하여 인식률을 높이는 과정을 말한다.

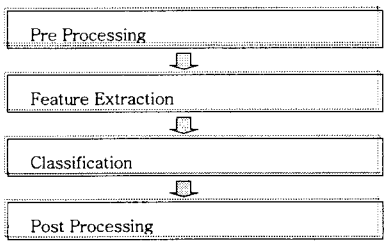

그림 1. 음성인식 과정

2.1 스펙트럼 차감법(Spectrum Subtraction)

스펙트럼 차감법[6]은 전처리 단계에서 음질을 개선하기위해 잡음을 최소화하는 방법이다. 이 방법은 주변 잡음에 의해 손상된 음성 스펙트럼에서 잡음 스펙트럼의 크기 성분만을 제거해 음성 스펙트럼을 구하는 방법이다. 이는 주변 잡음이 음성에 산술적으로 더해지면 잡음과 음성 신호사이에 상관관계가 없다는 가정

과 음성을 인지하는 청각의 특성은 위상 정보보다는 크기 정보에 더 많은 영향을 받는다는 점을 이용한 것이다.

잡음 신호 n(k)와 음성 신호 s(k)는 서로 독립적이다, 따라서 잡음 신호 n(k)가 음성 신호 s(k)에 더하면, 손상된 음성 신호 x(k)는 다음과 같이 나타낼 수 있다.

$$x(k) = s(k) + n(k) \tag{1}$$

여기서 윈도우를 취하고 단시간 푸리에 변환을 하면

$$X(w) = S(w) + N(w) \tag{2}$$

음성신호의 크기는 추정치 잡음 스펙트럼의 평균을 사용하여 원래의 음성신호를 구한다.

$$|S(w)| = |X(w)| - \mu(w)$$
$$\mu(w) = \frac{1}{M}\sum_{i=1}^{M} |N_i(w)| \tag{3}$$

2.2 음성 특징 추출

그림 2는 MFCC를 사용해 음성의 특징을 추출하는 과정을 나타낸 것이다.

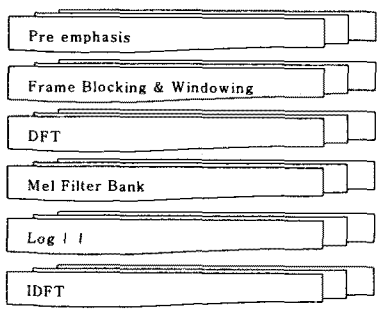

그림 2. MFCC Processing

음성을 20~30msec정도의 짧은 시간간격으로 나눠 음성의 특징이 변하지 않게 안정적으로 만들어 준다. 이를 프레임(Frame)이라고 이를 바탕으로 음성의 시간적 특성을 고려하여 10~20msec로 중첩된 프레임 간격으로 만드는 과정을 프레임 블로킹(Frame Blocking)이라고 한다.

FFT를 바탕으로 한 스펙트럼 성분은 서로 다른 파형의 스펙트럼 차이를 잘 나타낸다. FFT를 이용한 방법 중에서 인간의 청각 특성을 나타내는 MFCC 특징 추출방법은 1Khz정도까지의 저주파 영역의 신호에서는 선형 스케일에 따라 반응하지만, 고주파 영역의 신호에서는 로그 스케일의 특성을 가진다. 이런 특성을 고려한 것이 멜 스케일(mel scale)[5]이라 한다. 일반 선형 스케일을 멜 스케일로의 변환은 다음과 같다.

$$mel\ frequency = 2595\ \log_{10}\left(1 + \frac{frequency}{700.0}\right) \tag{4}$$

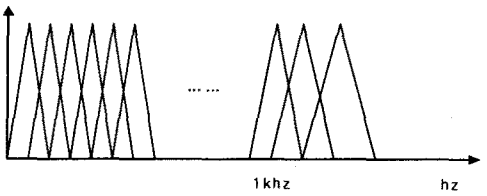

그림 3 Mel scaled Triangular Filter Bank

2.3 히스토그램을 이용한 스펙트럼 차감법

전처리 단계에서 사용되는 스펙트럼 차감법은 속도면에서 우수할 뿐만 아니라 이론적으로 간단하고 널리 사용되는 방법이다. 그러나 스펙트럼 차감법은 잡음환경에서 음성과 비 음성 구간의 잡음을 검출하는데 어려움이 있을 뿐만 아니라 신뢰도가 떨어진다. 물론 임의로 처음 부분의 아무런 음성이 검출되지 않는 부분을 잡음 스펙트럼으로 하여 계산 줄 수도 있다. 그러나 이 부분에서 신뢰성이 떨어지며, 정확하지 않다.

이런 부분을 개선하기 위해서 히스토그램을 통한 스펙트럼 차감법은 사용한다면 이 문제를 해결할 수 있다. 입력된 음성신호에서 잡음 스펙트럼의 추정치를 이용하여, 잡음을 제거하고 음성신호를 구하는 것이 아니라, 스펙트럼 상에서 음성신호의 스펙트럼과 잡음의 스펙트럼이 서로 독립적이라는 가정에서 스펙트럼 상의 음성과 잡음의 히스토그램을 통해서 음성의 영역과 잡음의 영역을 분리하여 잡음을 제거는 방법이다. 따라서 보다 신뢰성 있고 좋은 음질의 음성을 구할 수 있을 뿐만 아니라 각 프레임 별로 존재하는 잡음과 음성의 스펙트럼을 히스토그램을 통해 분리하여 사용하기 때문에 시간에 따라 변화하는 잡음 역시 제거하여 사용할 수 있다는 장점이 있다.

2.4 Mel Scaled Gaussian Filter Bank

Mel scale에 따른 필터 뱅크를 설계 시에 Triangular Filter Bank[8] 모양을 설계 하는 것이 아니라 좀더 삼각형에서 완화된 모습의 가우시안 모양의 필터 뱅크를 만들어 사용했다. 기존의 삼각형 모양의 필터 뱅크를 사용하여 추출된 특징들은 중심 주파수에 집중된다. 또한 1Khz 이하의 선형적인 부분에서는 특징이 잘 추출되나 1Khz 이상인 부분에서는 오히려 균일한 필터 뱅크를 촘촘히 설계 한 것이 더 우수하다. 따라서 삼각형 필터뱅크 보다 가우시안 모양의 필터를 멜 스케일에 따라서 분포시켜, 1Khz 이상의 필터들에서 좀더 좋은 특징 들을 얻을 수 있도록 만들었다.

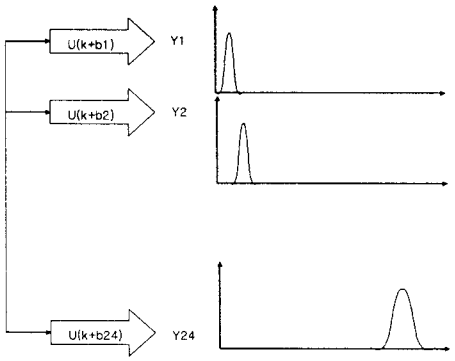

그림 4 Filter Bank의 특성

III. 실험

본 실험은 일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 영의 10가지 숫자 음을 실험하였으며, 기본적으로 11khz, 16비트로 5초간 sampling한 wave 파일을 이용하였다. 실험 대상자는 성인 남성 화자 9명을 대상으로 실험하였고, 사용한 컴퓨터는 AMD 750, Ram 512의 컴퓨터를 이용하였다.

다음 그림은 숫자 음 영의 음성을 스펙트로그램 상태에서 살펴 본 것이다.

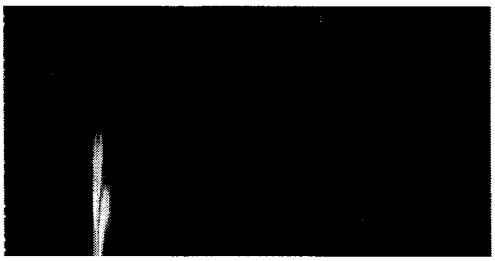

그림 5 숫자 "영" 음의 스펙트로그램

그림 5의 스펙트로그램을 살펴보면, 음성과 잡음의 스펙트로그램을 히스토그램 통해서 나타낸 것이다. 잡음의 영역과 음성의 영역을 확인하고, 임계 값을 구한 후에 서로 분리하여 잡음을 제거하고 음성의 부분만을 구한다.

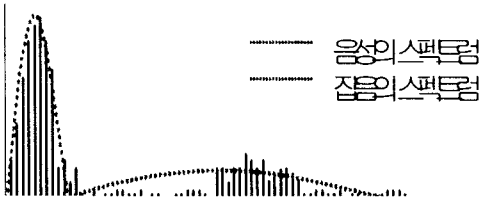

그림 6. 음성과 잡음의 히스토그램

다음은 개선된 음질을 이용하여 MFCC 방법을 사용해 음성의 특징을 추출하는 과정이다. MFCC 특징 추출 과정을 순서대로 나타냈다.

다음은 영이란 음성의 wave 파형을 그린 것으로 11khz, 16비트로 5초간 녹음한 음성이다.

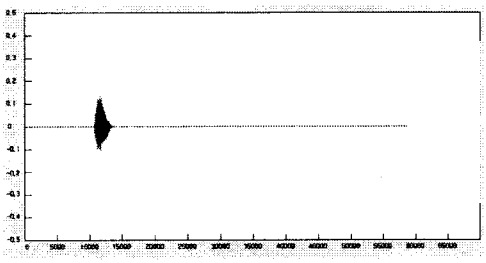

그림 7 a Sample Speech Signal

그림 8은 음성의 특징 추출 과정에서 Pre emphasis 과정을 거친 후에 256샘플을 기본으로 하는 Frame을 만들고 서로 128샘플씩 중첩시켜 Frame blocking 한 모습을 그린 것이다.

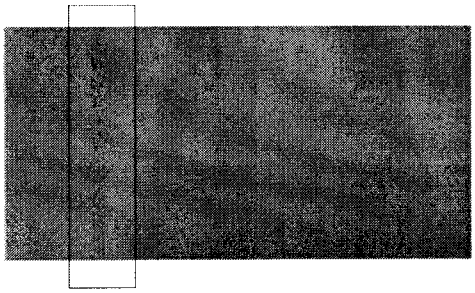

그림 8 Frame Blocking

Frame blocking 다음 Hamming windowing를 사용해 시작과 끝부분의 에러를 최소화 하고 256사이즈의 Short Time Fourier Transform(STFT)을 실행하였다. 여기서 프레임의 총 길이는 256으로 제한하여 사용했으며 프레임의 사이즈 역시 256으로 만든 후 STFT를 실행하였다.

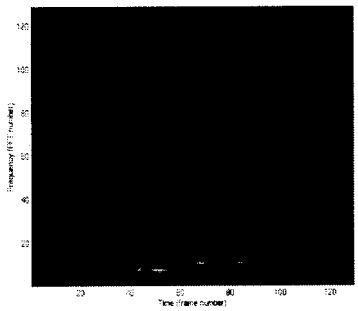

그림 9. 프레임 단위로 STFT를 실행

STFT를 실행한 후에 각각의 프레임에 24개의 Mel Scaled Gaussian Filter Bank를 사용해서 Mel Spectrum을 나타낸 것이다.

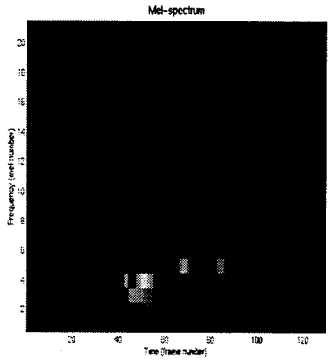

그림 10. Mel Spectrum

마지막으로, Mel spectrum에서 DCT를 실행하여 39차의 Mel Frequency Cepstrum Coefficient를 구한다.

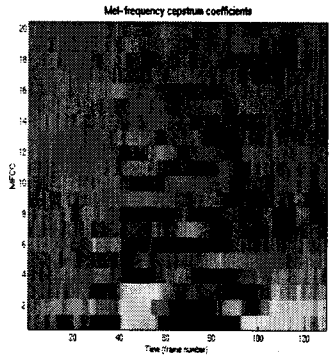

그림 11 MFCC

Ⅳ. 결과

히스토그램을 통한 스펙트럼 차감법을 이용해, 음질을 개선하기 전과 음질을 개선한 후의 상황을 알아보고, 각 숫자 음의 인식 빈도수를 통해 인식률을 비교해보았다. 매칭 방법으로는 고립단어 음성인식에 뛰어난 성능을 발휘하는 DTW를 사용했다.

표 1은 음질개선 전 및 후의 인식률을 나타내었다.

음성	전	후
일	84.44	90.00
이	88.88	91.11
삼	100.0	100.0
사	93.33	97.77
오	87.77	88.88
육	100.0	100.0
칠	93.33	97.77
팔	100.0	100.0
구	91.11	94.44
영	100.0	100.0

표 1. 음질개선 전 및 후의 인식률

음성의 특징이 잘 나타나는 영이나 팔과 같은 경우 음질 개선 전이나 후의 결과 비슷하였으나, 일이나 이처럼 구별하기 힘든 숫자의 경우 음질 개선 후의 인식률이 높아졌다.

참고문헌

[1] 이수영, "잡음에 둔감한 음성인식을 위한 청각모델", 한국과학기술원 뇌과학연구센터, 한국뇌학회지, Vol. 1, No. 2, pp. 157-172, Dec. 2001.

[2] 오영환, 음성언어정보처리, 홍릉과학출판사, 서울, 1998.

[3] Rabiner, L.R. and Sambur, M.R. "An Algorithm for Determining the Endpoints of Isolated Utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297-315, Feb. 1975.

[4] Rabiner, L,R and Juang B.H, *Fundamentals of Speech Recognition*, Prentice Hall PTR, 1996

[5] 이건상, 양성일, 권영현, 음성인식, 한양대학교 출판부, 2001.

[6] S.F. Boll and D.C. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaptive Cancellation," *IEEE. Trans. on ASSP*, Vol. 28, pp. 752-755, 1980.

[7] Hwang, Mei-Yuh, "Subphonetic Acoustic Model- ing for Speaker-Independent Continuous Speech Recognition," Tech Report No. CMS-CS-93-230, Carnegie Mellon University, Dec. 1993.

[8] 현동훈, "음성인식을 위한 멜켑스트럼의 최적화", 연세대학교 석사 논문, 1998.

[9] 이한구, "강인한 정합과정을 이용한 텍스트종속 화자인식에 관한 연구", 홍익대학교 석사 논문, 2003.

[10] J.-C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition,* Kluwer Academic Publishers, 1996.