음성인식을 위한 알고리즘에 관한 연구

김선철*, 이정우*, 조규옥*, 박재균*, 오용택*
한국기술교육대학교 정보기술공학부 전기공학과*

A study on the algorithm for speech recognition

Sun Chul Kim*, Jung Woo Lee*, Kyu Ok Cho*, Jae Gyun Park*, Yong Taek Oh*
*Korea University of Technology and education

Abstract - 음성인식 시스템을 설계함에 있어서는 대표적으로 사람의 성도 특성을 모방한 LPC(Linear Predict Cording)방식과 청각 특성을 고려한 MFCC(Mel-Frequency Cepstral Coefficients)방식이 있다. 본 논문에서는 MFCC를 통해 특징파라미터를 추출하고 해당 영역에서의 수행된 작업을 매틀랩 알고리즘을 이용하여 그래프로 시현하였다. MFCC 방식의 추출과정은 최초의 음성신호로부터 전처리과정을 통해 아날로그 신호를 디지털 신호로 변환하고, 잡음부분을 최소화하며, 음성 부분을 강조한다. 이 신호는 다시 Windowing을 통해 음성의 불연속을 제거해 주고, FFT를 통해 시간의 영역을 주파수의 영역으로 변환한다. 이 변환된 신호는 Filter Bank를 거쳐 다수의 복잡한 신호를 몇 개의 간단한 신호로 간소화 할 수 있으며, 마지막으로 Mel-cepstrum을 통해 최종적으로 특징 파라미터를 얻고자 하였다.

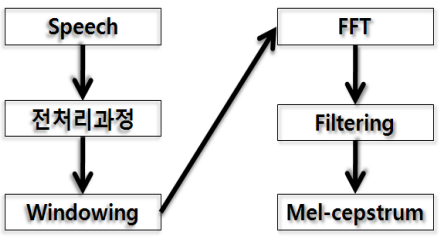
1. 서 론

오늘날 사람은 컴퓨터나 기계의 사용이 생활화되면서 기존의 아날로그(기계식, 터치식)식의 양자간 통신시스템에서 드러나는 불편함을 해소하기 위해 디지털(전자식, RF식)식의 양자간 통신시스템을 사용하여 컴퓨터 또는 기계와 사람간의 보다 신속하고 편리한 정보전달을 취할 수 있는 음성인식 신호처리 기술이 부각되고 있다. 현재 음성 인식 분야의 연구가 활발히 진행 중에 있으며, 이를 이용한 Door-Lock, Remote- Controller, Car-Navigation등의 제품들이 시중에 선보이고 있다. 이런 음성인식 신호처리 알고리즘에서 음성의 특징 파라미터 값을 얼마나 정확하게 추출할 수 있는 것이 가장 핵심이며, 정확한 음성 특징 파라미터는 인식률의 향상을 가져온다. 음성 신호의 특징 파라미터를 추출해내는 대표적인 방법에는 LPC방식과 MFCC방식이 있는데, LPC방식은 음성 신호의 초기 신호 일부분을 가지고 다음 신호들을 미리 예측하여 전체 음성의 특징 파라미터를 추출해 내는 방식이다. LPC를 기반으로 한 음성인식 시스템은 잡음이 없는 환경에서는 좋은 성능을 보이지만, 잡음 환경 하에서는 그 인식 성능이 급격히 낮아진다. 그러나 인간의 경우에는 매우 시끄러운 환경에서도 다른 사람의 말을 알아들을 수 있다.[1] 이러한 인간의 청각특성을 고려한 MFCC방식을 본 논문에서는 사용하였다.

<표 1> LPC와 MFCC의 장단점[2]

	장점	단점
LPC	·많은 데이터를 처리하지 않으므로 속도가 MFCC에 비해 빠르다.	·잡음 환경 하에서 인식률이 MFCC방식에 비해 떨어진다.
MFCC	·잡음 환경 하에서 인식률이 LPC방식에 비해 정확하다.	·많은 데이터를 처리하므로 속도가 LPC에 비해 느리다.

본 논문의 구성은 실제의 음성을 받아들여 2.1장에서는 전처리 과정(Pre-emphasis, Sampling, Quantization, Coding,), 2.2장에서는 Windowing, 2.3장에서는 FFT, 2.4장에서는 Filtering, 2.5장에서는 Mel-cepstrum(log, DCT)의 MFCC를 하기위한 일련의 과정을 매틀랩 알고리즘을 통해 그래프로 제시하고 분석하는 것으로 하였다.

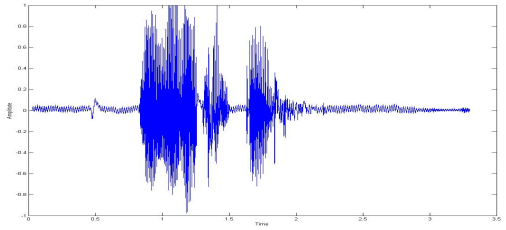


<그림 1> MFCC 특징파라미터 추출 과정

2. 본 론

2.1 전처리 과정

전처리 과정은 음성신호의 특징 파라미터를 추출하기 위한 최초의 과정으로서, 입력된 음성 신호를 음향학적 파라미터(A/D)로 변환하며 음성 신호의 특징을 강조하는 작업을 수행한다. Pre-emphasis, Sampling, Coding, Quantization 과정이 이에 해당된다. <그림 2>는 최초에 녹음하여 입력한 '열려라 참깨'라는 음성신호이며, 신호의 길이는 3.3초이다.

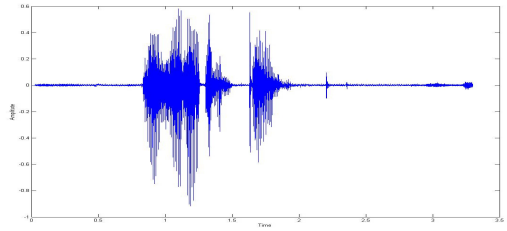


<그림 2> 입력 음성 신호

2.1.1 Pre-emphasis

입력된 음성신호는 고대역 통과 특성을 갖는 Pre- emphasis 필터를 거친다. 이 필터를 사용하는 이유는 첫째, 인간의 외이 / 중이의 주파수 특성을 모델링하기 위하여 고대역 필터링을 한다. 이는 입술에서의 방사에 의하여 20 dB/decade로 감쇄되는 것을 보상하게 되어 음성으로부터 성도 특성만을 얻게 된다. 둘째, 청각시스템이 1 kHz이상의 스펙트럼 영역에 대하여 민감하다는 사실을 어느 정도 보상하게 된다.[4] <그림 3>은 입력 음성신호를 Pre-emphasis 필터를 통과한 신호의 파형으로서, 낮은 주파수의 값은 감소되고 높은 주파수의 값이 강조된 것을 볼 수 있다. Pre-emphasis 필터는 다음 식으로 나타낼 수 있다.

$$H(z) = 1 - z_0 z^{-1}, z_0 \approx 1$$



<그림 3> Pre-emphasis 후의 신호 파형

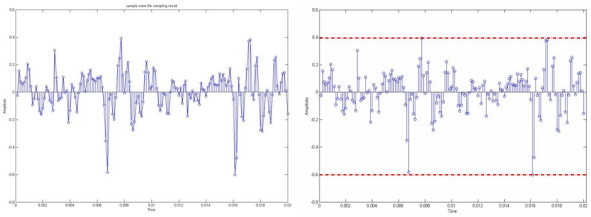
2.1.2 신호의 디지털화

Pre-emphasis 필터를 거친 신호는 Sampling, Quantization(양자화), Coding(부호화)를 거쳐 디지털 신호가 된다. Sampling이란, 연속적으로 변화하는 신호를 Discrete Time에서의 아날로그 샘플값을 얻는 것이다. 인간의 음성은 약 4kHz의 대역폭을 가지고 있으므로 Nyquist sampling 이론에 따라 8kHz로 샘플링 하였다.

Fs = 8000(Hz), Ts = 1/Fs = 1/8000(s)

<그림 4>-(a)는 Sampling한 신호에서 길이가 20ms인 구간을 잘라내었을 때, 샘플값이 160개가 있는 것을 보여주고 있다.(샘플갯수 = 8000 * 0.02s = 160)

Coding(부호화)이란 Sampling된 이산적인 음성 신호를 Quantization(양자화)하고 코드를 할당하여 디지털화 하는 것이다. 음성신호가 샘플당 a[bit]로 부호가 할당된다면, 레벨수는 2^a 보다 작거나 같은 값으로 결정할 수 있다.[3] <그림 4>-(b)는 신호를 양자화 하기 위해 샘플값을 이산적으로 표현하고 있다. 이 각 샘플값에 부호를 할당하면 최종적으로 디지털 신호가 된다.



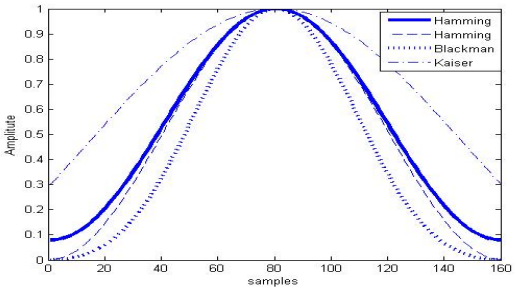
(a) Sampling (b) Quantization

<그림 4> 20ms 구간의 이산신호

2.2 Windowing

전처리 과정을 거친 후에 특징 파라미터를 추출하기 위해서 음성 신호를 20ms씩 구간을 블록화 하여 Framing한다. 20ms씩 Framing하는 이유는 인간의 음성은 20ms안에서 특징변화가 미미하기 때문이다. 구간을 나눈 후에 각 프레임의 주파수 특성을 알아내야 하는데 이때 프레임 양끝단의 불연속 지점은 주파수 영역으로 변환 시 원하지 않는 정보를 포함하게 된다. 그래서 이러한 시작과 끝 지점의 불연속 부분을 최소화하기 위해 각 프레임마다 Window계수를 곱하게 된다. 종류로는 Hanning, Hamming, Blackman, Kaiser 등이 있다. <그림 4>는 각 Window 함수들의 종류를 보여주고 있다. 본 논문에서는 가장 보편적으로 많이 사용되는 Hamming 계수를 사용하였다.[5]

Hamming Window : $W_H(n) = 0.54-0.46\cos(2\pi n/N-1)$



<그림 4> Window 함수 종류

2.3 FFT(Fast Fourier Transform)

특징 파라미터를 추출하기 위해서는 FFT알고리즘을 사용하여 시간영역의 신호를 주파수 영역으로 변화시켜 주파수 특성을 파악할 수 있다. 시간영역에서의 이산적이고 비주기적인 음성 신호는 퓨리에 변환을 통해 주파수 영역에서 주기적이고 연속적인 신호로 변환된다. 그러나 퓨리에 변환된 신호는 연속함수이므로 계산처리가 어려운 단점이 있다. 따라서 본 논문에서는 DFT(Discrete Fourier Transform) 즉, 시간영역에서의 초기의 이산신호를 스펙트럼의 샘플들로 표현하는 방법을 사용한다.

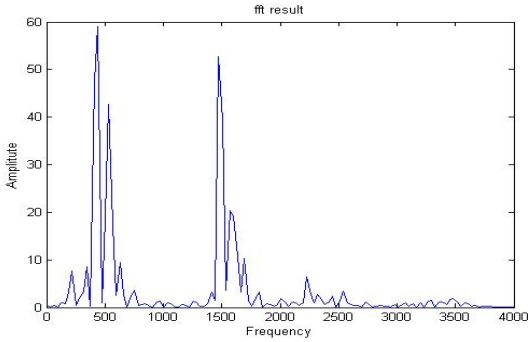
x(n)의 DFT식 : $X(2\pi k/N) = X(k)$일 때

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}, k = 0,1,2,...N-1$$

W_N의 대칭성(symmetry property과 주기성(periodicity property)을 이용하면 계산량을 대폭 줄일 수 있다. FFT알고리즘은 이러한 성질을 이용하여 DFT를 계산한다.[6] <그림 5>는 hamming window를 씌운 프레임을 시간축에서 주파수축으로 변환시킨 FFT 변환 파형이다.

FFT알고리즘을 이용한 DFT계산 : $W_N = e^{-J\angle\pi/N}$라고 놓으면

$$X(k) = \sum_{n=0}^{N-1} x(n)\, Wn^{kn}, 0 \le k \le N-1$$



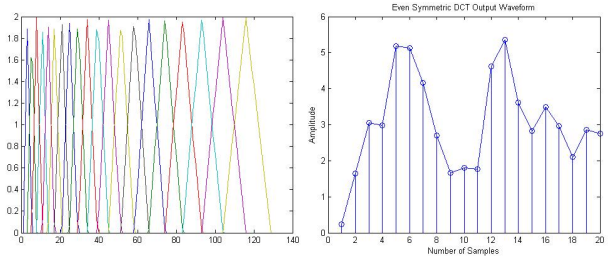
<그림 5> FFT 변환 파형

2.4 Filtering

FFT에서 얻어진 결과 값을 필터뱅크에 통과시키면, 주파수 대역을 여러 개의 필터뱅크로 나누고 각 뱅크에서의 에너지를 구한다. 필터뱅크의 모양 및 중심 주파수 설정방법은 귀의 청각적 특성에 맞게 결정 된다.[7]

$$\tilde{S}(l) = \sum_{k=0}^{N/2} S(k) M_l(k), \quad l = 0,1,....,L-1$$

필터뱅크 주파수 테이블에 의하면 8kHz로 샘플링된 음성신호는 FFT 작업 수행 후 필텅뱅크를 거쳐 20개의 에너지값으로 출력된다.



(a) Filter Bank (b) Bank 에너지

<그림 6> Filter Bank를 통과한 20개의 Bank 에너지

2.5 Mel-cepstrum

Mel-cepstrum은 MFCC의 최종 과정으로서 Filter bank를 통과하여 추출한 뱅크 에너지를 LOG와 DCT 과정을 거쳐 필요한 최소한의 정보만을 얻는 과정이다. 우리의 귀가 소리의 크기에 대해 로그함수로 느끼기 때문에 필터뱅크 출력에너지 각각에 로그를 취하게 된다. 다음에 DCT를 하면 20개의 에너지값은 멜스케일로 된 주파수의 정현파 계수로 바뀌게 된다. 이중 13차까지의 mel 계수를 취하고 나머지는 버리게 된다.[8]

$$c(i) = \sqrt{\frac{2}{L}} \sum_{m=1}^{L} \log(\tilde{s}(m)) \cos\left[\frac{\pi i}{L}(m-0.5)\right], i = 0,1,....C-1$$

여기까지의 과정을 통해 20ms길이로 블럭화 된 음성신호에서 한 프레임당 13개의 특징벡터를 구하고자 한다

3. 결 론

음성인식 신호처리에서 가장 핵심이 되는 부분은 얼마나 정확하게 음성의 특징 파라미터를 추출하느냐에 달려 있다. 잘못된 음성 검출은 인식 시스템의 성능을 크게 저하시킨다. 따라서 본 논문에서는 처리속도가 LPC에 비해 다소 느리지만 주변 환경의 잡음에도 정확한 특징 파라미터를 추출할 수 있는 MFCC 알고리즘을 음성 인식 시스템에 적용시키고자 추출과정을 각 단계별로 확인하였다. 일반적인 주파수의 단위를 이 특징에 맞게 맵핑시키고, mel-cepstrum계수를 적용하여 불필요하게 중복되어 있는 음성정보를 없앤 후, 최소 정보만을 추출하였다. 향후 이 특징 파라미터를 이용하여 음성인식 분야에 유효하게 사용하는 일만 남았다. 새로운 특징 파라미터와 기존의 특징 파라미터 사이에서 유사도 비교 및 분석을 DTW(Dynamic Time Warping), 벡터 양자화, HMM(Hidden Markov Model)[9] 중의 적절한 알고리즘을 적용하여 음성 인식률이 얼마나 실효성이 있는지에 대해 연구 할 것이며, 이런 과정을 통해 DSP보드를 사용하여 높은 인식률을 가진 음성인식 Door-Lock 시스템을 구연할 계획이다.

[참 고 문 헌]

[1]김주곤, "MEL-LPC분석 방법을 이용한 한국어 음성인식 시스템의 성능향상", 정보통신연구소 논문집, 9권 제 1집, p.2~2.3, 2002

[2]신동성, "음성인식 시스템의 처리시간 단축에 관한 연구", 한국통신학회추계종합학술대회, 1호, p.5~8, 1999

[3]임제탁, 디지털신호처리, 희중당, p.1~19, 1995

[4]김현구, "인식점수의 궤환을 통한 음성 및 화자인식 시스템의 구현에 관한 연구", 과학기술원 석사학위논문, p.21~34, p.43~52, 2005

[5]우광방, 디지털 신호처리, 청문각, p.465~476, 2004

[6]우광방, 디지털 신호처리, 청문각, p.543~604, 2004

[7]Ben J. Shannon, "A Comparative Study of Filter Bank Spacing for Speech Recognition", MICROELECTRONIC ENGINEERING RESEARCH CONFERENCE, p.1~3, 2003

[8]Syed Ali Khayam, "The Discrete Cosine Transform(DCT): Theory and Application", Department of Electrical & Computer Engineering Michigan State University, 2003

[9]정익주, 실시간 디지털 신호처리, 생능출판사, p. 703~745, 2006