기저막 특성을 이용한 새로운 음성 특징 추출 및 성능 분석

이 철 희, 신 유 식, 정 성 환, 김 종 교
전북대학교 전자정보공학부

Performance of analysis and extraction of speech feature using characteristics of basilar membrane

Chul-Hee Lee, You-Shik Shin, Sung-Hwan Chung, Chong-Kyo Kim
Division of Electronics and Information Eng., Chonbuk National University
E-mail : kissmeda@shinbiro.com

Abstract

본 논문에서는 음성 인식률 향상을 위한 여러 가지 방법들 중에서 음성특징 파라미터 추출 방법에 관한 한 가지 방법을 제시하였다. 본 논문에서는 청각 특성을 기반으로 한 MFCC(mel frequency cepstrum coefficients)와 성능 향상을 위한 방법으로 GFCC (gammatone filter frequency cepstrum coefficients)를 제시하고 음성 인식을 수행하여 성능을 분석하였다. MFCC에서 일반적으로 사용하는 임계 대역 필터로 삼각 필터(triangular filter) 대신 청각 구조의 기저막(basilar membrane)특성을 묘사한 gammatone 대역 통과 필터를 이용하여 특징 파라미터를 추출하였다. DTW 알고리즘으로 인식률을 분석한 결과 삼각 대역 필터를 이용한 것보다 gammatone 대역 통과 필터를 이용한 추출법이 약 2~3%의 성능 향상을 보였다.

1. 서 론

일상 생활에서 인간은 개개인의 의사 소통의 수단으로 음성을 이용하고 있다. 따라서 인간과 컴퓨터의 접촉이 일상화된 현시점에서 간단하고 편리한 정보 전달 기술의 한 수단으로 음성을 이용하려는 움직임이 커지고 있다. 음성 인식 기술에 대한 관심이 증대되면서 여러 분야에 접목하려는 움직임이 일어나고 있다. 이에 상응하여 인식률의 향상이 필수적이다. 인식률의 향상을 위한 방법으로 여러 가지가 있지만, 음성 특징 파라미터 추출 방법도 중요한 부분이다. 파라미터를 추출하는 방법으로 음성 인식 기술의 초기 단계에서는 필터 뱅크(filter bank)와 LPC 분석을 통하여 파라미터를 추출하였다. 현재는 음성 인식에 많이 사용되는 파라미터 추출 방법으로 LPC 계수로부터 유도된 LPC 켑스트럼과 인간의 청각 특성을 이용한 멜 주파수 켑스트럼(mel frequency cepstrum)계수를 이용한다. 또한, 청각 모델링(auditory modeling)을 이용한 파라미터 추출 방법 등이 있다.[1][3]

본 논문에서는 멜 주파수 켑스트럼이 멜(mel) 단위 임계 대역(critical bandwidth)을 갖는 필터를 사용하여 인간의 청각 특성을 적용했다는 이론에 접목해서 논문에서는 인식률 향상을 위한 방법으로 실제 청각 구조의 내이(inner ear)의 와우각(cochlear) 내에 있는 기저막(basilar membrane)의 특성이 대역 통과 필터 열을 형성한다는 것을 이용하여 파라미터 추출을 하여 인식 실험을 하였다.[2][3][4][5] DTW(dynamic time warping) 알고리즘으로 실험을 수행하여 음성 데이터의 각 차수에 따른 인식률을 비교하였다.

본 논문의 구성은 2장에서는 멜 주파수 켑스트럼 방법, 3장에서 기저막의 특성과 사용된 대역 통과 필터에

대한 파라미터 추출 방법을 다루며, 4장에서는 실험 및 결과를 보이고, 5장에서 결론을 맺는다.

2. 멜 주파수 켑스트럼(mel frequency cepstrum)

멜 주파수 켑스트럼 계수 (MFCC)는 현재 음성 인식에서 널리 사용되는 파라미터 추출 방법이다. 멜 단위(mel scale)는 Stevens과 Volkman(1940)에 의해 연구되어왔고, O'Shaughnessy는 멜 단위의 식을 다음과 같이 정의하였다.

$$\text{mel frequency} = 2595 \log_{10}\left(1+\frac{f}{700}\right) \quad (1)$$

본 논문에서도 식 (1)로 멜 단위로 변환한 임계 대역(critical bandwidth) 삼각 필터들을 사용하여 파라미터를 구한다. 삼각 필터뿐만 아니라 사각 필터와 Gaussian 분포 모양의 임계 대역 필터를 사용하여 파라미터를 구할 수 있다.

멜 주파수 켑스트럼 계수를 구하는 블록도는 그림 2.1과 같다.[3]

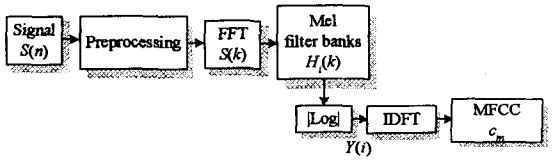

그림 2.1 MFCC 특징 파라미터 추출 블록도

화자의 음성 신호로부터 음성 특징 파라미터를 추출하기 위해 먼저 전처리 과정을 거친 후 신호의 각 프레임에 대해 N-point FFT를 하여 N개의 성분을 구한다. k번째 인덱스의 중심 주파수는 식 (2)와 같다.

$$f_k = k\frac{F_s}{N}, \quad 0 \le k \le N-1 \quad (2)$$

F_s는 샘플링 주파수이다. 이 때 i번째 임계대역 필터의 출력 $Y(i)$는 식 (3)과 같다.

$$Y(i) = \sum_{k=0}^{N/2} \log|S(k)|H_i(k), \quad i=0,\cdots,M \quad (3)$$

여기서 M은 필터의 개수를 나타낸다.

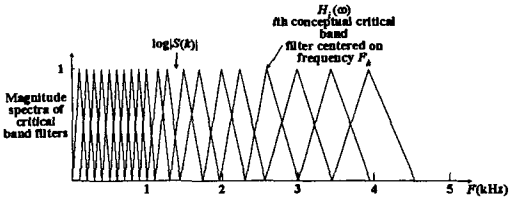

그림 2.2 멜(mel) 단위 임계 대역 필터

각 멜 단위 임계 대역을 갖는 필터를 그림 2.2에 보였다. 식 (3)은 식 (4)와 같이 중심 주파수 k_i의 작은 영역(small range)으로 다시 나타낼 수 있다.

$$\hat{Y} = \begin{cases} Y(i), & k=k_i \\ 0, & \text{other } k \in [0, N-1] \end{cases} \quad (4)$$

마지막으로 IDFT를 변환을 하여 멜 주파수 켑스트럼 계수 c_m을 구한다.

$$c_m = \frac{1}{M}\sum_{k=0}^{M-1} \hat{Y}(k) e^{jk(2\pi/M)m}, \quad m=1,\cdots,n \quad (5)$$

n은 MFCC의 차수를 나타낸다. 그러나 $\hat{Y}(k)$가 실수이고 대칭이 되므로 식 (5)는 DCT(discrete cosine transform)함수로 지수 함수를 대신할 수가 있다. 따라서 식 (6)과 같이 된다.

$$c_m = \frac{1}{M}\sum_{k=0}^{M-1} \hat{Y}(k)\cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \quad (6)$$

여기서 M을 필터 수, m은 필터의 차수를 나타낸다. 본 논문에서는 MFCC 값 각 차수 별로 인식 실험을 하여 파라미터를 추출하였다.

3. 기저막(basilar membrane) 특성을 이용한 음성의 특징 추출

3.1 기저막의 구조와 특징

귀의 구조는 크게 외이, 중이, 내이로 나눌 수 있다.

기저막은 내이의 와우각(cochlea) 내부에 위치한 길이가 30~35mm 정도이고 뒤로 갈수록 넓어지는 구조를 가지고 있다. 기저막은 서로 다른 주파수 성분을 전파하는 진행파로 묘사한 전송선 모델이 제안되어 대역 통과 필터로 나타낼 수 있게 되었다. 따라서 등골에 가까울수록 고주파에 민감하고 끝으로 갈수록 저주파에 민감하다. 이러한 성질로 인해 저주파에 민감한 부분에서는 주파수 해상도가 높고 고주파에 민감한 부분에서는 시간에 대한 해상도가 높다. 이 성질을 이용하여 기저막을 모델화하여 음성 특징을 추출하게 된다.[2]

3.2 기저막 특성의 대역 통과 필터

실제 특징 추출에 이용하는 것은 주파수 영역에서 대역 통과 필터의 특징을 가지고 저주파일수록 주파수 해상도가 좋은 것이 기저막의 일반적인 성질이다. 대역 통과 필터 설계시 필터의 모양과 각각의 대역폭과 중심 주파수를 고려해야 한다. 그러나 필터의 모양은 중요한 변수는 아니다.

본 논문에서는 기저막 특성을 잘 묘사하는 gammatone 필터를 사용하였다. 이 필터는 기저막을 묘사하기 위해 많이 사용되는 필터로 차수 n이 4인 4차 gammatone 필터를 사용하였고 필터의 임펄스 응답으로 8차 recursive digital 필터로 구현할 수가 있다. 식 (7)은 gammatone 필터를 낸 것이다.

$$g(t)=\frac{at^{n-1}\cos(2\pi f_c t)}{e^{2\pi Bt}} \tag{7}$$

여기서 f_c는 필터의 중심 주파수, B는 f_c에서의 대역폭을 나타낸다. 대역폭 결정에 있어서 ERB (equivalent rectangular bandwidth)를 사용하였다.

$$ERB=\left[\left(\frac{f}{Q}\right)^{order}+B_n^{order}\right]^{\frac{1}{order}} \tag{8}$$

여기서 Q는 필터의 quality factor, B_n은 최소 대역폭을 나타낸다. 각 변수의 값은 여러 실험을 통해서 다양하게 제안되었다. 본 논문에서 사용한 값은 Glasberg와 Moore 변수 값을 이용하였다. 또한 중심주파수는 식 (9)와 같다.

$$f_{c_i}=-QB_n+(f_x+QB_n)e^{i(-\log(f_x+QB_n)+\log(f_l+QB_n))/M}$$
$$i=0,\cdots,M-1 \tag{9}$$

여기에서 f_x는 최대 주파수, f_l은 최소 주파수이고 M은 필터 개수를 나타낸다. 표 1에 각 변수의 값들을 표시하였다.

표 1. 변수 값

변수	값
Q	9.26449
B_n	24.7
$order$	1
f_x	6855
f_l	133
M	40

그림 3.1은 본 논문에서 사용한 40개의 gammatone 필터를 보인 것이다. 이들 필터는 Patterson 과 Holdworth 의 cochlea 설계 모델에서 기저막 특성의 ERB 대역폭을 갖는 gammatone 필터 뱅크로 구현된 필터 열들이다.[6]

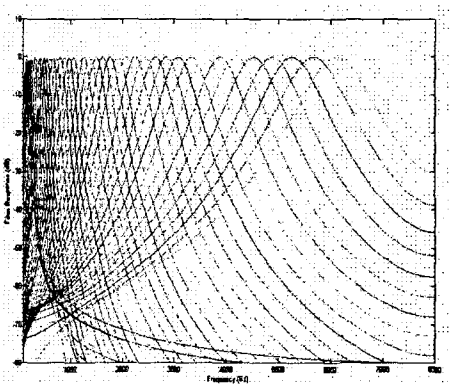

그림 3.1 ERB 대역폭을 갖는 gammatone 필터 뱅크

3.3 기저막 특성을 이용한 파라미터 추출

본 논문에서 제안 한 기저막 특성을 이용한 파라미터 추출 방법은 다음 블록도와 같다.

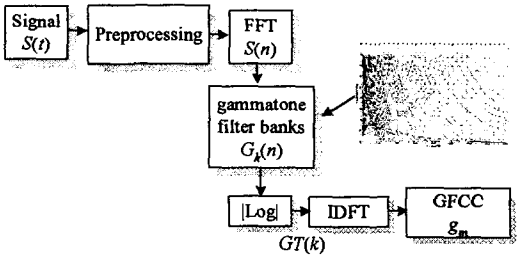

그림 3.2 특징 파라미터 추출 블록도

파라미터 계수를 구하는 방법은 그림 3.2의 블록도와 같이 MFCC를 구하는 방법과 유사하다. 그러나 기존의 MFCC의 파라미터 추출 방법에 있는 멜 단위 임계 대역(critical band)을 갖는 삼각 필터(triangular filter)를 대신하여 기저막 특성 필터인 gammatone 필터를 적용하여 파라미터를 구하는 방법인 gammatone 필터 주파수 켑스트럼 계수(GFCC : Gammatone filter Frequency Cepstrum Coefficients)를 제안한다. 따라서 멜 단위 임계 대역을 이용한 MFCC의 파라미터 보다 더 청각적인 특징을 가질 수 있게 된다.

$$GT(k)=\sum_{n=0}^{N/2}\log|S(n)|G_k(n),\quad k=1,\cdots,M \tag{10}$$

여기서 M은 필터 수, N은 FFT 크기이다.

$$g_m = \frac{1}{M}\sum_{k=0}^{M-1} GT(k)\cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \qquad (11)$$

식 (10)에서 보는 바와 같이 삼각 필터 대신에 gammatone 필터 $G_k(n)$를 삽입하여 파라미터 계수 g_m을 구하게 된다.

4. 실험 및 결과

본 논문에서는 음성 인식 실험을 위해 30명의 화자, 즉 남자 22명, 여자 8명이 3음절, 4음절로 구성된 단어 20개를 각 1회씩 발성한 데이터 총 600개의 데이터를 수집하였다. 녹음환경은 주변잡음이 존재하는 일반적인 실험실이며 IBM PC에서 마이크를 통한 음성신호를 사운드 카드로 8kHz, 16bit linear PCM으로 A/D 변환하여 수집하였다. 음성의 특징 파라미터 추출을 위한 분석 프레임의 크기는 25.625ms, 이동 프레임의 크기는 12.5ms로 하였다. 특징 파라미터 추출은 12차, 14차, 16차, 18차, 24차의 MFCC 와 GFCC를 추출하였다.

추출된 각 파라미터들 중 남자 2명, 여자 2명의 파라미터 값을 참조 패턴으로 하고 나머지 남자 20명, 여자 6명의 파라미터 데이터로 시험패턴으로 하여 인식 실험을 하였다. 인식 실험에는 DTW(dynamic time warping) 알고리즘을 이용하였다.

표 2와 표 3은 MFCC와 GFCC의 각 차수 별로 인식률을 나타낸 것이다.

표 2. 차수별 남녀 인식률 비교(%)

차수		12	14	16	18	24
GFCC	남	96.25	96.25	96.5	96.75	97.25
	여	92.5	90.83	92.5	93.33	93.33
MFCC	남	93.5	94.25	94.25	94.25	94.75
	여	92.5	90.83	91.67	90.83	90.8

표 3. 인식률의 결과 와 비교 (%)

차수	12	14	16	18	24
GFCC	95.38	95.00	95.58	95.96	96.34
MFCC	93.26	93.46	93.65	93.65	93.85

5. 결 론

음성 특징 파라미터 추출 방법들 중에 MFCC는 인간의 청각 특성을 반영한 추출법으로 많이 이용되고 있다. MFCC의 청각 특성은 1kHz이하에서는 선형적이고 이상에서는 지수적인 값을 갖는 멜 크기의 삼각 대역 필터를 이용하는데 그쳤다. 하지만 본 논문에서는 실제 청각 모델링(auditory modeling)에서 파라미터 추출에 이용하는 기저막(basilar membrane)의 특성인 gammatone 대역 통과 필터를 MFCC를 구하는 방법 중 필터 뱅크의 부분에 대치함으로써 새로운 파라미터 추출 방법인 GFCC를 제안하게 되어 실제 청각 모델링(auditory modeling)을 하지 않고 청각적 특성이 가장 가까운 특징을 가진 파라미터를 추출할 수 가 있다.

실험 결과, 인식률이 GFCC가 MFCC 보다 각 차수에서 약 2~3% 높음을 알 수 있고, 남자와 여자를 비교했을 때 여자의 인식률은 비슷했으나 남자의 인식률이 높음을 알 수 있었다.

GFCC가 임펄스 응답이 8차인 필터를 구현해야 하므로 계산 량이 많은 단점이 있다. 앞으로 인식률을 더 향상하기 위한 보완과 잡음 음성에 대한 인식 실험이 필요하다.

참고문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice- Hall, 1993.

[2] 정호영, 김도영, 은종관, 이수영,"청각구조를 이용한 잡음 음성의 인식 성능 향상", 한국음향학회지 제14권, 제5호, 1995.

[3] John R. Deller, Jr., John G. Proakis, John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.

[4] J. M. Kates, "A Time-Domain Digital Cochlear Model," *IEEE Trans. on Signal Processing*, vol. 39, no. 12, pp. 2573-2592, Dec. 1991.

[5] Rivarol Vergin, Douglas O'Shaughnessy, "Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition," *IEEE Trans. on Speech & Audio Processing*, vol. 7, no. 5, pp. 512-532, 1999.

[6] M. Slaney, "An Efficient Implementation of the Patterson-Holdworth Auditory Filter Bank," *Apple Computer Tech. Report #35*, 1993.