논문 00-25-12T-14 한국통신학회논문지 '00-12 Vol.25 No.12T

기저막 특성을 이용한 음성 특징 추출 및 성능 분석

정회원 문 인 섭*

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

In-Seob Moon* *Regular Members*

요 약

본 논문에서는 음성인식 시스템의 성능 향상을 위해 청각구조의 기저막(Basilar Membrane) 특성을 묘사한 gammatone 대역통과 필터를 이용하여 추출된 음성 특징 파라메터 GFCC (Gammatone Filter Frequency Cepstrum Coefficients)를 제시한다. 또한 기존의 음성 특징 파라메터인 LPC Cepstrum 그리고 MFCC(Mel Frequency Cepstrum Coefficients)의 성능과 비교 분석하였다. 음성인식 모델을 생성하기 위해 널리 사용되는 이산 HMM 알고리즘을 기반으로 음성 인식 시스템을 구성하였으며, HMM의 상태수의 변화에 따른 인식률을 비교 하였다. 결과적으로 GFCC가 LPC 켑스트럼이나 MFCC보다 1~3% 높은 인식률을 보였다.

ABSTRACT

In this paper, new method for extracting speech feature parameter set, namely GFCC(Gammatone Filter Frequency Coefficients), which was extracted using a set of gammatone band pass filters-describing the basilar membrane of human auditory system-was proposed. We also compared the resulted performance of GFCC with that of traditional LPC cepstrum and MFCC(Mel Frequency Cepstrum Coefficients). Discrete HMM(Hidden Markov Model) algorithm based recognition system was composed for the examination and it was to compare the recognition rate using not only for the performance analysis between speech feature parameters but also for the change of state in HMM. Consequently newly suggested GFCC was proven to have better performance in recognition rate by 1~3% than existing LPC cepstrum and MFCC.

Ⅰ. 서 론

1970년대에 들어오면서 컴퓨터를 이용한 음성의 인식과 합성에 대한 연구가 디지털 신호처리 기술의 발달에 힘입어 활성화되었으며, 자신의 의사를 음성을 이용하여 자연스러운 방법으로 컴퓨터에 전달하려는 욕구가 발생되었다. 이것은 컴퓨터의 대용량화와 고속화에 힘입어 더욱 많은 계산량을 갖고 있는 알고리즘의 실현이 점점 가능하게 되었지만, 쉽게만 생각되었던 사람의 대화에 대한 인식률은 아직까지 그리 높지 않고, 다만 좁은 응용범위를 갖는 음성에 대해서는 어느 정도의 실용화 가능한 인식률을 나타내고 있다.

인간의 음성을 인식하기 어렵게 하는 원인은 크게 2가지를 들 수 있다.[1][8] 첫째로는 음성의 변이성을 들 수 있다. 의미를 갖는 소리를 예로 들었을 경우, 동일한 시간에 서로 다른 사람이 발성한 경우의 음성을 분석해보면 서로간에 많은 상이점이 발견되며, 또 동일한 사람이 서로 다른 시간에 발성한 음성 역시 많은 상이점을 갖는 것을 발견할 수 있다. 이러한 상이점들은 동일한 음성의 인식 시에 인식률을 낮추는 요인으로 작용하며 또한 서로 다른 음성에 대해서는 같은 음성인 것처럼 오 인식이 되도록 작용하기도 한다. 두 번째로는 음성의 분석을

* 조선이공대학 정보통신과(mis@mail.chosun-c.ac.kr)
논문번호 : T00026-0810, 접수일자 : 2000년 8월 10일

위해 현재 사용하고 있는 분석방법이 이러한 변이성을 잘 수용하지 못한다는 것이다. 완벽한 음성의 인식을 위해서는 첫 번째 요인인 음성의 변이성 같은 인식기 외적인 요인보다는 두 번째의 요인을 해결할 수 있는 새로운 알고리즘의 연구가 시급하다 할 수 있을 것이다. 음성인식 기술의 초기 단계에서는 필터뱅크(Filter Bank)와 LPC 분석을 통하여 파라메터를 추출하였다. 현재, 음성인식에 많이 사용되는 파라메터로는 LPC 계수로 부터 유도된 LPC 켑스트럼과 인간의 청각 특성을 이용한 멜 주파수 켑스트럼(Mel Frequency Cepstrun)이 있다.

본 논문에서는 실제 청각 구조의 내이(Inner ear)와 와우각(Cochlear) 내에 있는 기저막 특성을 묘사한 Gammatone 대역통과 필터를 이용하여 추출된 음성특징 파라메터 GFCC(gammatone filter frequency cepstrum coefficients)를 제안하고, 기존의 대표적인 음성특징 파라메터인 LPC-켑스트럼과 멜 켑스트럼의 성능과 비교 분석한다. 본 논문의 구성은 2장에서는 멜 주파수 켑스트럼 방법에 대해 설명하고, 3장에서 본 논문에서 제안된 음성특징 파라메터 GFCC의 추출법을 설명한다. 4장에서는 음성특징파라메터 성능분석을 위해 구축된 음성인식 시스템에 대해 설명하고, 5장에서는 실험 및 결과를 보이고, 6장에서 결론을 맺는다.

Ⅱ. Mel Frequency Cepstrum

멜 주파수 켑스트럼 계수(MFCC)는 현재 음성인식에서 널리 사용되는 파라미터이다. 멜 단위(mel scale)는 Stevens과 Volkman(1940)에 의해 연구 되어왔고, O'Shaughnessy는 멜 단위의 식을 다음과 같이 정의하였다.

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

멜 주파수 켑스트럼 계수를 구하는 블록도는 그림 1.과 같다[3].

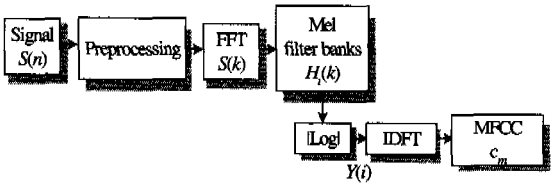


그림 1. MFCC 특징 파라미터 추출 블록도

MFCC는 그림 2.와 같은 Critical band 에너지를 이용하여 켑스트럼 계수를 구하는 것으로 i번째 임계 대역 필터의 출력 $Y(i)$는 식 (2)와 같다.

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

여기서 M은 필터의 개수를 나타낸다.

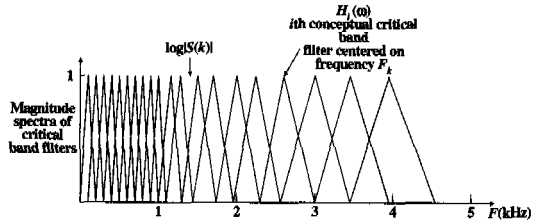


그림 2. 멜(mel) 단위 임계 대역 필터

식 (2)는 식 (3)과 같이 중심 주파수 k_i의 작은 영역(small range)으로 다시 나타낼 수 있다.

$$\hat{Y} = \begin{cases} Y(i), & k=k_i \\ 0, & \text{other } k \in [0,\ N-1] \end{cases} \tag{3}$$

마지막으로 IDFT를 변환을 하여 멜 주파수 켑스트럼 계수 c_m을 구한다.

$$c_m = \frac{1}{M}\sum_{k=0}^{M-1} \hat{Y}(k) e^{j\pi(2\pi/M)m}, \quad m=1,\cdots,n \tag{4}$$

n은 MFCC의 차수를 나타낸다. 그러나 $\hat{Y}(k)$가 실수이고 대칭이 되므로 식 (4)는 DCT(discrete cosine transform)함수로 지수 함수를 대신할 수가 있다. 따라서 식 (5)과 같이 된다.

$$c_m = \frac{1}{M}\sum_{k=0}^{M-1} \hat{Y}(k)\cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \tag{5}$$

여기서 M을 필터 수, m은 필터의 차수를 나타낸다.

Ⅲ. 기저막(basilar membrane) 특성을 이용한 음성 특징 파라메터

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. 현재 basilar membrane에 대한 지식은 von Bekesy의 실험에 의해 밝혀졌으며, 그는 등골을 주기적으로 진동시키면서 달팽이관의 길이에 따라 basilar membrane의 변화율을 측정했다. basilar membrane는 비교적 넓은 대역 통과 필터로 나타나며, 등골에 가까울수록 고주파에 민감하고 끝으로 갈수록 저주파에 민감하다. 연속적인 점에서 각 필터는 대략 일정한 첨예도를 가지며, 이성질로 인해 저주파에 민감한 부분에서는 주파수 해상도가 높고 고주파에 민감한 부분에서는 시간에 대한 해상도가 높다. 바로 이러한 성질이 basilar membrane을 모델화 하는데 중요한 이론적 바탕이 되며 음성 특징 추출에 활용된다[2].

2. 기저막 특성의 대역 통과 필터

본 논문에서는 기저막 특성을 묘사하기 위해 주로 이용되는 4차 gammatone 필터를 사용하였으며, 임펄스 응답을 8차 recursive digital 필터로 구현하였다. 식(6)은 gammatone 필터를 나타낸 것이다.

$$g(t)=\frac{at^{n-1}\cos(2\pi f_c t)}{e^{2\pi Bt}} \tag{6}$$

여기서 f_c는 필터의 중심 주파수, B는 f_c에서의 대역폭을 나타낸다. 대역폭 결정에 있어서 ERB (equivalent rectangular bandwidth)를 사용하였다. ERB(식 7)는 청각 필터의 모양을 추정하는 과정에서 유도된 것으로 500Hz 이하에서도 주파수 감소에 따라 계속 감소하게 된다.

$$ERB=\left[\left(\frac{f}{Q}\right)^{order}+B_n^{order}\right]^{\frac{1}{order}} \tag{7}$$

여기서 Q는 필터의 quality factor이며 B_n은 최소 대역폭을 나타낸다. 각 변수의 값은 여러 실험을 통해서 다양하게 제안되었다. 본 논문에서 사용한 값은 Glasberg 와 Moore 변수 값을 이용하였다. 또한 중심주파수는 식 (8)과 같다.

$$f_{c_i}=-QB_n+(f_x+QB_n)e^{i(-\log(f_x+QBn)+\log(f_l+QB_n))/M},$$
$$i=0,\cdots,M-1 \tag{8}$$

여기에서 f_x는 최대 주파수, f_l은 최소 주파수이고 M은 필터 개수를 나타낸다. 표 1에 식(8)의 각 변수의 값들을 표시하였다. 그림 3.은 본 논문에서 사용한 gammatone 필터의 주파수 특성을 보인 것이다.

표 1. 변수 값

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

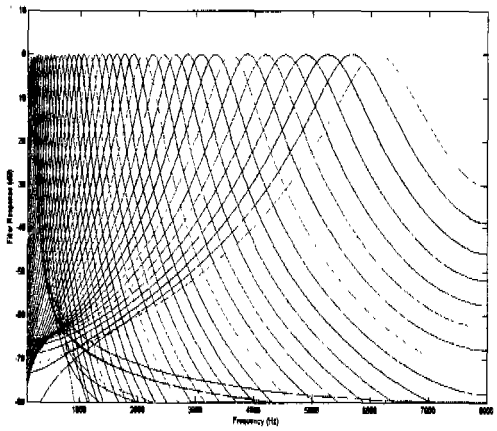


그림 3. ERB 대역폭을 갖는 gammatone 필터 주파수 특성

3. 기저막 특성을 이용한 파라미터 추출

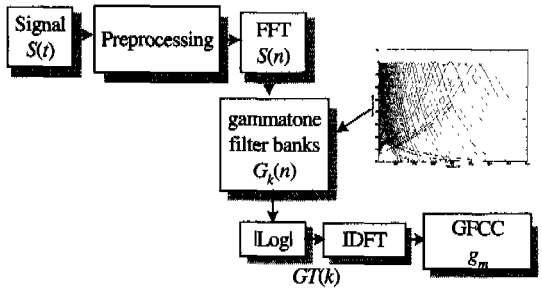


그림 4. 음성 특징 파라미터 추출 블록도

본 논문에서 제안한 음성 특징 파라미터 추출 방법은 그림 4.에 나타내었으며, 이는 MFCC를 구하는 방법과 유사하다. 그러나 기존의 MFCC의 파라미터 추출 방법에서 이용되는 멜 단위 임계 대역

(critical band)을 갖는 삼각 필터(triangular filter)를 대신하여 기저막 특성 필터인 gammatone 필터를 적용하였다. 따라서 멜 단위 임계 대역을 이용한 MFCC의 파라미터 보다 더 청각적인 특징을 가질 수 있게 된다.

화자의 음성 신호로부터 음성 특징 파라미터를 추출하기 위해 먼저 전처리 과정을 거친 후 신호의 각 프레임에 대해 N개의 FFT 성분을 구한다.

이 때 k번째 임계대역 필터의 출력 $GT(k)$는 식 (9)와 같다.

$$GT(k)=\sum_{n=0}^{N/2}\log|S(n)|G_k(n),\quad k=1,\cdots,M \tag{9}$$

여기서 M은 필터의 개수를 나타낸다.

식 (9)는 식 (10)과 같이 중심 주파수 k_i의 작은 영역(small range)으로 다시 나타낼 수 있다.

$$\widehat{GT}=\begin{cases}GT(k), & k=k_i\\ 0, & \text{other } k\in[0,\ N-1]\end{cases} \tag{10}$$

마지막으로 IDFT를 변환을 하여 GFCC를 g_m을 구한다.

$$g_m=\frac{1}{M}\sum_{k=0}^{M-1}\widehat{GT}(k)e^{j\pi(2\pi/M)m},\quad m=1,\cdots,n \tag{11}$$

n은 GFCC의 차수를 나타낸다. 그러나 $\widehat{GT}(k)$가 실수이고 대칭이 되므로 식 (11)은 DCT(discrete cosine transform)함수로 지수 함수를 대신할 수가 있다. 따라서 식 (12)와 같이 된다.

$$g_m=\frac{1}{M}\sum_{k=0}^{M-1}\widehat{GT}(k)\cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \tag{12}$$

식 (12)에서 보는 바와 같이 삼각 필터 대신에 gammatone 필터 $G_k(n)$를 삽입하여 파라미터 계수 g_m을 구하게 된다.

Ⅳ. 인식 시스템의 구성

1. DHMM

음성은 장소, 시간, 발성 자에 따라 다르며 발성 시간 등에서도 많은 차이를 보이는데 HMM은 이러한 변이성을 확률로 계산하며 그간 음성인식에서 주도적인 역할을 해온 DTW(Dynamic Time Warping)의 단점(계산량 과다, 연속음성인식에서의 어려움)을 보완하는 대안으로 많이 사용되어지고 있다[7][8]. HMM모델은 음성신호가 이전의 특징과 관계가 있는 확률 프로세스라고 가정하고서 도입된 방법으로서 성대가 10-30ms 정도의 짧은 시간 동안에는 그 특성이 변하지 않는 준 안정 프로세스(quasi-stationary process)라는 성질에 기초를 두고 있다[1][4]. 본 논문에서는 단순 left-to-right 모델을 적용한 DHMM(Discrete Hidden Markov Model)을 사용하였으며 기존의 상태 수를 고정시키는 방법과 인식어휘 내의 음소 개수에 따라 상태 수를 가변시키는 방법을 이용하였다. HMM을 구성하는 두 가지 요소로 상태천이 확률과 관측확률을 들 수 있는데 그림 5.에 상태천이 확률을 결정하기 위한 단순 left-right 모델의 예를 나타내었다.

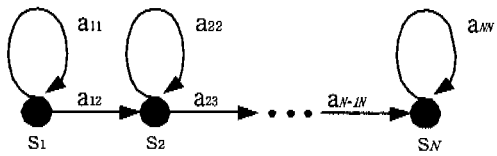


그림 5. 단순 left-right 모델

DHMM은 관측되는 심벌이 미리 정의된 코드 북에 포함되는 경우 동일한 코드 북의 중심과의 거리가 얼마인가에 관계없이 동일한 거리로 간주하여 계산되어지는 방식으로 자주 발생하지 않는 심벌의 경우 두 개의 코드 중간에 걸쳐지게 되고 이런 경우 잡음등 기타의 요인에 의해 종종 다른 코드로 인식되어 인식률을 저하시키는 요인으로 작용하기도 한다. 전처리 과정에서 구한 음성신호의 특징벡터를

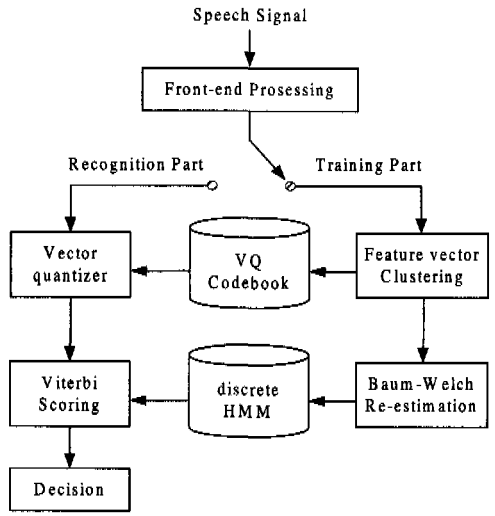


그림 6. DHMM의 구성도

벡터 양자화 과정을 거쳐서 코드 북의 가장 근접한 코드워드로 나타낸다. DHMM의 구성 도를 그림 6.에 나타내었다.

2. Viterbi Algorithm

Viterbi 알고리즘은 관측 시퀀스와 모델 λ가 주어졌을 때 최적의 상태경로 $Q = q_1, q_2, \cdots q_N$을 찾는 방법으로 Forward-Backward 알고리즘과 비교해 볼 때 시간 t마다 최적의 경로를 계산해 나가며 다음의 단계들을 통해 구해진다.

1 단계 : 초기화

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le n$$

$$\Psi_1(i) = 0$$

2 단계 : 루프

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \quad 2 \le t \le T \quad 1 \le i \le N$$

$$\Psi_t(j) = \arg\max_{1 \le i \le N} [\sigma_{t-1}(i) a_{ij}] \quad 2 \le t \le T \quad 1 \le j \le N$$

3 단계 : 종료

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \max_{1 \le i \le N} [\delta_T(i)]$$

4 단계 : 경로 역추적

$$q_t^* = \Psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots 1$$

Baum-Welch 알고리즘에 의해 학습이 이루어졌으며, 인식 과정에서는 입력음성의 특징 벡터 열과 학습 모델 사이에 Viterbi scoring 방식으로 최근거리의 학습 모델을 찾아서 인식단어로 결정하였다.

V. 실험 및 결과

본 논문에서는 음성 인식 실험을 위해 6명의 화자가 10개의 단어에 대해 20회씩 발성한 데이터 1200개와 10명이 2회씩 발성한 데이터 200개를 수집하였다. 그중 화자 3명이 각 20회씩 발성한 데이터 600개는 인식기의 훈련에 이용되었으며 나머지 13명분의 데이터 800개는 인식 테스트에 이용되었다. 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 12차의 LPC 켑스트럼, MFCC 그리고 GFCC를 추출하였다. 인식 실험은 4장에서 설명된 DHMM 알고리즘을 이용하였다. 인식결과는 그림 7.에 나타난 것처럼 GFCC(94.21%)가 LPC켑스트럼(92.56%), MFCC(93.37%)에 비해 높은 인식률을 보였다. 또한 HMM의 상태수를 고정(상태수 6, 8, 10) 시켰을 경우는 GFCC가 LPC 켑스트럼이나 MFCC 보다 높은 인식률을 보였고 HMM의 상태수를 가변(단어의 음소수)으로 하였을 경우 GFCC가 LPC 켑스트럼 보다는 높은 인식률을 보였으나, MFCC의 인식률과는 비슷한 인식 결과를 얻을 수 있었다.

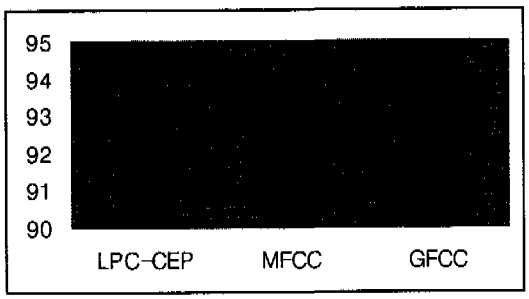


그림 7. 인식률 결과

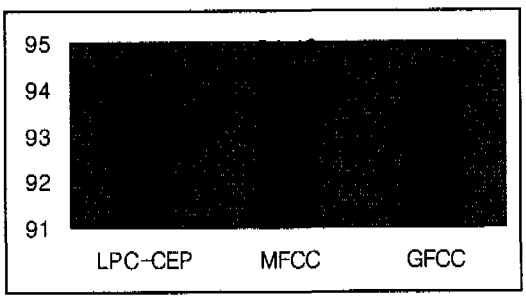


그림 8. DHMM 상태수가 가변일경우 인식률

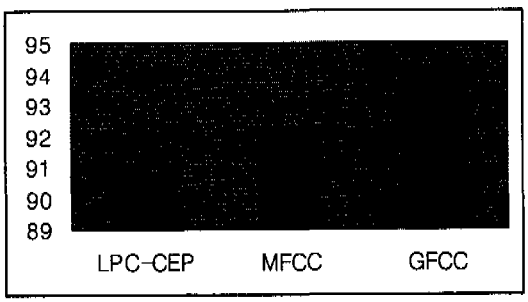


그림 9. DHMM 상태수가 6일경우 인식률

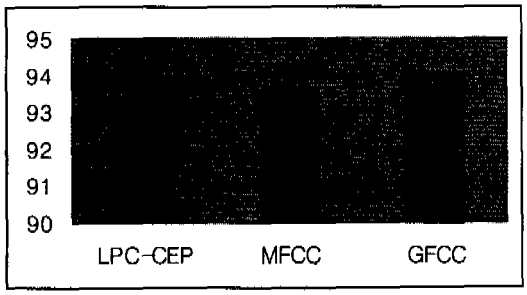


그림 10. DHMM 상태수가 8일 경우 인식률

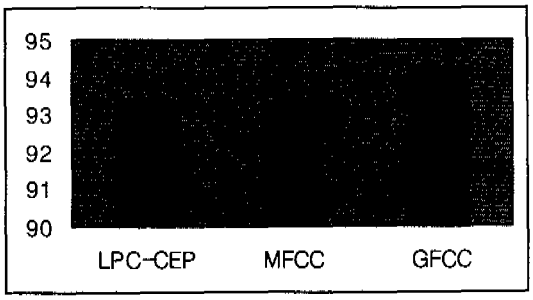


그림 11. DHMM 상태수가 10일 경우 인식률

Ⅵ. 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 파라메터 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 MFCC와 LPC 켑스트럼 보다 약 1~3% 정도 향상되었다. 그러나 DHMM의 상태수를 가변으로 하였을 경우 GFCC가 MFCC보다 낮은(0.03%) 인식률을 보여, gammatone 필터의 임펄수 응답 차수와 인식률과의 상관관계를 고려할 필요가 있으며, 또한GFCC는 임펄스 응답이 8차인 필터를 구현해야 하므로 계산량이 많은 단점이 있다. 앞으로 인식률을 더 향상하기 위한 보완과 여러 종류의 잡음과 잡음 지수에 따른 인식 실험을 수행 할 예정이다.

참 고 문 헌

[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.

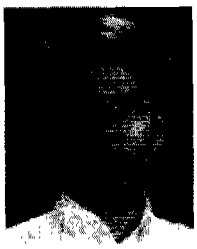
[5] Rivarol Vergin, Douglas O'Shaughnessy, "Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition," *IEEE Trans. on Speech & Audio Processing*, vol. 7, no. 5, pp. 512-532, 1999.

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

[7] X.D.Huang, Y.Ariki, M.A.Jack, "Hidden Markov Models for Speech Recognition,"Edinburgh University Press, 1990.

[8] 유창규, "전화망에서의 연속음성인식을 위한 빔 탐색 음소열 추출," 박사학위논문, 전북대학교, 1999.

문 인 섭(In-Seob Moon) 정회원



1992년 2월 : 전북대학교 공과대학 전자공학과 졸업(공학사)

1995년 2월 : 전북대학교 대학원 전자공학과 졸업 (공학석사)

1997년 2월 : 전북대학교 대학원 전자공학과 박사과정 수료

1997년 3월~1999년 3월 : 한려대학교 정보통신학과 전임강사

1999년 5월~현재 : 조선이공대학 정보통신과 전임강사

<주관심 분야> 음성인식, 음성코딩, 음성합성