論文2002-39TE-2-9

인지적 청각 특성을 이용한 고립 단어 전화 음성 인식

(Isolated-Word Speech Recognition in Telephone Environment Using Perceptual Auditory Characteristic)

崔 炯 箕 *, 朴 基 榮 **, 金 種 圭 ***

(Hyung-Ki Choi, Ki-Young Park, and Chong-Kyo Kim)

요 약

본 논문에서는, 음성 인식률 향상을 위하여 청각 특성을 기반으로 한 GFCC(gammatone filter frequency cepstrum coefficients) 파라미터를 음성 특징 파라미터로 제안한다. 그리고 전화망을 통해 얻은 고립단어를 대상으로 인식실험을 수행하였다. 성능비교를 위하여 MFCC(mel frequency cepstrum coefficients)와 LPCC(linear predictive cepstrum coefficient)를 사용하여 인식 실험을 하였다. 또한, 각 파라미터에 대하여 전화망의 채널 왜곡 보상기법으로 CMS(cepstral mean subtraction)를 도입한 방법과 적용시키지 않은 방법으로 인식실험을 하였다. 실험 결과로서, GFCC를 사용하여 인식을 수행한 방법이 다른 파라미터를 사용한 방법에 비해 향상된 결과를 얻었다.

Abstract

In this paper, we propose GFCC(gammatone filter frequency cepstrum coefficient) parameter which was based on the auditory characteristic for accomplishing better speech recognition rate. And it is performed the experiment of speech recognition for isolated word acquired from telephone network. For the purpose of comparing GFCC parameter with other parameter, the experiment of speech recognition are carried out using MFCC and LPCC parameter. Also, for each parameter, we are implemented CMS(cepstral mean subtraction)which was applied or not in order to compensate channel distortion in telephone network. Accordingly, we found that the recognition rate using GFCC parameter is better than other parameter in the experimental result.

* 正會員, 全北大學校 電子工學科
(Dept. of Electronic Eng. Chonbuk National University)
** 正會員, 全州工業大學, 情報通信科
(Dept. of Information & Communication Eng. Jeonju Technical College)
*** 正會員, 全北大學校 電氣電子制御工學部, 教授
(Dept. of Electrical & Control Eng. Chonbuk National University)
接受日字:2002年1月16日, 수정완료일:2002年4月23日

Ⅰ. 서 론

음성 인식 기술의 응용에서 전화를 통한 음성정보 서비스의 경우, 사용자의 환경에서 발생하는 잡음의 형태는 주변잡음, 채널에 의한 신호왜곡, 화자의 발음변화 등 다양하며, 이 응용 분야에서 중요한 음성 인식률의 저하를 발생시킨다. 따라서, 인식률 향상 및 시스템 실제 적용을 위해서 본 논문에서 인지적 청각 특성을 갖는 gammatone filter를 이용하여 음성 특징 파라미터를 추출하였다.

음성 특징 파라미터 추출하는 방법으로 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]

제안한 실험 방법으로 인식률 향상을 위해서 실제 청각 구조의 내이(inner ear)의 와우각(co chlear)에 있는 기저막(basilar membrane)의 특성이 대역 통과 필터 열을 형성한다는 것을 이용하여 파라미터 추출을 수행하였고, 전화 음성에 대한 인식 실험을 하였다.[2-5]

전화 음성 인식을 위해 채널 왜곡 보상 기법으로 스펙트럼 차감법(spectral subtraction), CMS(Cepstral Mean Subtraction), RASTA(RelAtive SpecTrAl)방법이 있지만 CMS방법이 간편하면서도 높은 인식률을 보이기 때문에 이 방법을 이용하여 실험하였다.[7]

본 논문의 구성은 2장에서는 채널 왜곡 보상 방법으로 CMS방법과 음성 특징 추출 방법에 대해 알아보고, 인지적 청각 특성인 기저막의 대역 통과 필터에 대한 파라미터 추출 방법을 다루며, 3장은 실험 및 결과를 보이고, 4장에서 결론을 맺는다.

Ⅱ. 채널 왜곡 보상 및 음성 특징 추출 방법

1. 전화망 채널 보상 방법

전화망의 경우에는 주변잡음, 채널왜곡, 마이크로폰왜곡 등으로 인식 성능이 저하된다. 본 논문에서는 주변 잡음이 거의 존재하지 않는 조용한 환경에서 휴대폰과 유선전화를 이용하여 음성을 수집하였다. 따라서 주변 잡음을 무시 할 수가 있다. 이러한 채널 잡음은 스펙트럼 영역이나 켑스트럼 영역에서 필터링이나 가중함수를 사용하여 왜곡을 보상해줄 수 있다. 본 논문에서 채널 보상 방법으로 CMS(Cepstral Mean Subtraction)가 가장 널리 이용되고 우수하기 때문에 이 방법을 적용하여 성능을 분석하였다.

1.1 CMS(Cepstral Mean Subtraction)

전화망을 통한 음성신호는 채널 필터링의 영향에 의해 선형왜곡이 일어난다. CMN(Cepstral Mean Normalization)의 일종인 CMS는 모델 훈련과정과 테스트 과정에서 조금씩 변하는 채널의 왜곡성분을 켑스트럼 영역에서 제거하는 방법이다. 전체 구간에 대해 켑스트럼의 평균을 구하고, 이를 차감하여 전화선 채널에서 발생하는 선형왜곡을 보상한다. 다음 식 (1)과 같이 구해진다.

$$c_t = c_t - E[c_t]$$
$$E[c_t] = \frac{1}{T}\sum_{t=1}^{T} c_t \tag{1}$$

여기서 $E[c_t]$는 채널 켑스트럼 평균값, c_t는 t번째 프레임의 켑스트럼, T는 전체 프레임의 수이다. 순수한 음성의 켑스트럼에서 장구간 평균이 0이라고 가정하면, 채널 켑스트럼의 추정치는 필터링된 음성의 켑스트럼을 구할 수 있다.

2. 음성 특징 추출 방법

2.1 MFCC(Mel frequency cepstral coefficient)

멜 주파수 켑스트럼 계수 (MFCC)는 현재 음성 인식에서 널리 사용되는 파라미터 추출 방법이다.

$$\text{mel frequency} = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{2}$$

본 논문에서 gammatone filter와 비교하기 위해 식 (2)로 멜 단위로 변환한 임계 대역(critical bandwidth) 삼각 필터들을 사용하여 파라미터를 구한다.

멜 주파수 켑스트럼 계수를 구하는 블록도는 그림 1과 같다.[3]

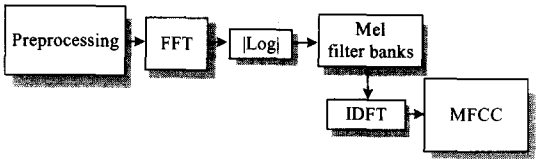


그림 1. MFCC 특징 파라미터의 추출 블록도
Fig. 1. Extraction of MFCC feature parameter.

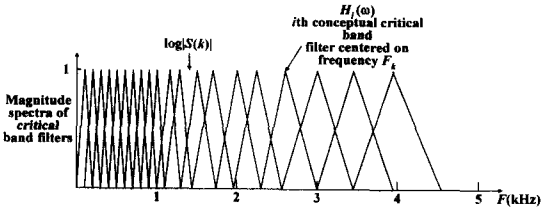


그림 2. mel 크기의 임계대역 필터
Fig. 2. Filter of mel scaled critical bandwidth.

각 멜 단위 임계 대역을 갖는 필터를 그림 2에 보였다. 다음 식(3)으로 MFCC 파라미터를 구한다.

$$c_m = \frac{1}{M}\sum_{k=0}^{M-1} \tilde{Y}(k)\cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \qquad (3)$$

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

2.2 기저막 특성의 대역 통과 필터

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 대역통과 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]

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

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

$$g(t) = \frac{at^{n-1}\cos(2\pi f_c t)}{e^{2\pi Bt}} \qquad (4)$$

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

$$ERB = \left[\left(\frac{f}{Q}\right)^{order} + B_n^{order}\right]^{\frac{1}{order}} \qquad (5)$$

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 (6)와 같다.[8]

$$f_{c_i} = -QB_n + (f_x + QB_n)e^{i(-\log(f_x+QBn)+\log(f_l+QB_n))/M} \quad i = 0,\cdots,M-1 \qquad (6)$$

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

표 1. 변수 값
Table. 1. Value of variable.

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

그림 3은 본 논문에서 사용한 40개의 gammatone 필터를 보인 것이다.[6]

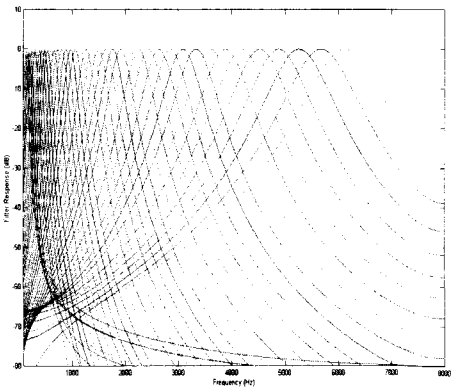


그림 3. ERB 대역폭을 갖는 gammatone 필터
Fig. 3. Gammatone filter bank with ERB bandwidth.

2.3 기저막 특성을 이용한 파라미터 추출

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

그림 4의 블록도와 같이 기저막 특성 필터인 gamma tone 필터를 적용하여 파라미터를 구하는 방법인 gamma tone 필터 주파수 켑스트럼 계수(GFCC : gammatone filter frequency cepstrum coefficients)를 제안한다. 따라서 멜 단위 임계 대역을 이용한 MFCC

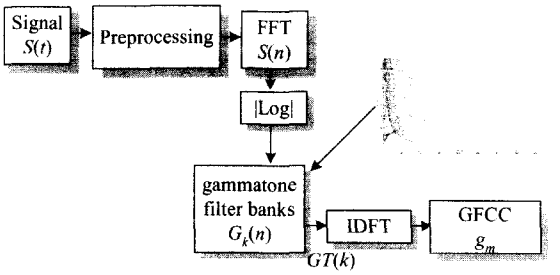


그림 4. GFCC 특징 파라미터의 추출 블록도
Fig. 4. Extraction of GFCC feature parameter.

의 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 (7)와 같다.

$$GT(k) = \sum_{n=0}^{N/2} \log|S(n)|G_k(n), \quad k = 1, \cdots, M \tag{7}$$

여기서 M은 필터의 개수를 나타낸다. 식 (7)는 식 (8)과 같이 중심 주파수 k_i의 작은 영역(small range)으로 다시 나타낼 수 있다.

$$\widehat{GT} = \begin{cases} GT(k), & k = k_i \\ 0, & \text{other } k \in [0,\ N-1] \end{cases} \tag{8}$$

마지막으로 IDFT를 변환을 하여 GFCC를 g_m을 구한다.

$$g_m = \frac{1}{M}\sum_{k=0}^{M-1} \widehat{GT}(k) e^{jk(2\pi/M)m}, \quad m = 1, \cdots, n \tag{9}$$

n은 GFCC의 차수를 나타낸다. 그러나 $\widehat{GT}(k)$가 실수이고 대칭이 되므로 식 (9)은 DCT(discrete cosine transform)함수로 지수 함수를 대신할 수가 있다. 따라서 식 (10)와 같이 된다.

$$g_m = \frac{1}{M}\sum_{k=0}^{M-1} \widehat{GT}(k) \cos\left\{m\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right\} \tag{10}$$

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

Ⅲ. 실험 및 결과

1. 실험 방법

본 논문에서는 음성 인식 실험을 위해 전화 음성을 이용하였다. 수집된 전화음성은 전라도, 충청도, 경상도 등 각 지역에서 유선전화와 무선(휴대폰)전화로 수집된 음성 데이터이다. 본 논문에 사용된 음성 데이터의 환경은 다음과 같다.

- 전화음성 데이터
  - 남성화자 : 28명
  - 여성화자 : 28명
  - 단어 수 : 10개
  - 발성회수 : 1회
    - 학습 데이터 : 남성화자 - 20명, 여성화자 - 20명
    - 시험 데이터 : 남성화자 - 8명, 여성화자 - 8명

- 유·무선에 따른 데이터

	무선전화 (휴대폰)	유선전화
남	16	12
여	19	9

- 학습·시험 데이터

		무선전화 (휴대폰)	유선전화
학습	남	12	8
	여	15	5
시험	남	4	4
	여	4	4

- 수집 환경
  - 주변잡음이 적은 조용한 장소에서 수집
  - 8kHz의 16 bit 데이터

실험은 각각의 LPCC, MFCC, GFCC의 12차의 파라미터를 추출하여 인식률을 비교 분석하였다. 그림 5는 "개인정보" 단어의 각 파라미터 추출의 한 예를 보인다. 전화 음성의 채널 왜곡을 보상해주기 위한 방법으로 성능이 좋은 CMS를 사용하여 인식률을 비교하였다.

인식 알고리즘으로 DHMM을 이용하였으며 코드북 크기와 상태는 다음과 같이 정하였다. 코드북 사이즈의 증가에 따라, 각각의 인식률은 증가하지만, 각각 파라미터간의 차이는 거의 변화를 보이지 않았기 때문에, 계산량을 감안해 코드북 64를 사용하였으며, 상태수의 변화에 따른 인식률은 8에서 10사이일 때 가장 높은 인식률을 보였다. 따라서, 본 논문에서는 파라미터의 비교를 목적으로 코드북 64와 상태수 8을 이용해 인식실험을 하였다.

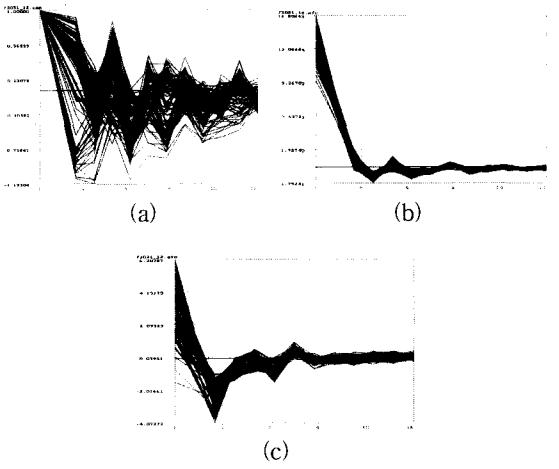


그림 5. LPCC(a), MFCC(b), GFCC(c) 특징 파라미터
Fig. 5. Figure of parameter LPCC(a), MFCC(b), GFCC(c).

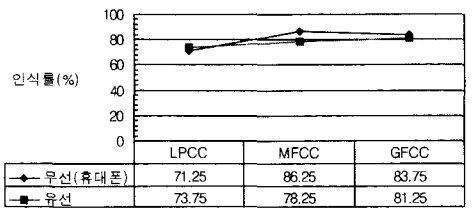
2. 실험 결과

실험은 채널 보상을 적용하지 않은 방법과 적용한 방법을 실험1)과 실험2)로 나누어 실험하였다. 무선(휴대폰)과 유선을 분류하여 인식률을 비교하여 보았다.

▸ 실험 1)

	LPCC	MFCC	GFCC
인식률(%)	72.5	82.5	82.5

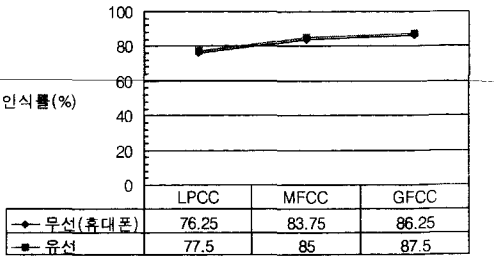
▸ 실험 1) :무선과 유선의 인식률 1



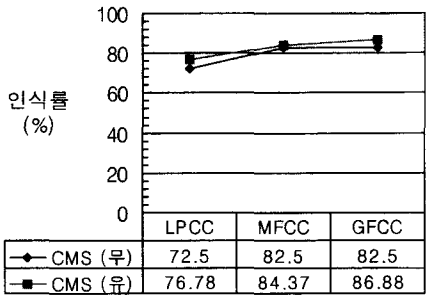
▸ 실험 2)

	LPCC	MFCC	GFCC
인식률(%)	76.78	84.37	86.88

▸ 실험 2) : 무선과 유선의 인식률 2



▸ 전체 인식률 비교



Ⅳ. 결 론

음성 특징 파라미터 추출 방법들 중에 MFCC는 인간의 청각 특성을 반영한 추출법으로 많이 이용되고 있다.

본 논문에서는 인지적 청각 특성을 위해 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 수가 있다.

전화 음성을 이용하기 때문에 많은 채널 왜곡에 따른 인식률의 저하가 있었다. 따라서, 채널 보상을 위해 CMS 방법을 적용한 결과 인식률의 향상이 있음을 알 수가 있었다.

실험 결과, 인식률이 GFCC가 MFCC 와 LPCC 비교했을 때 전화음성에서의 채널 왜곡 보상을 하기 전에

는 MFCC와 유사한 인식률을 보였고, LPCC보다는 우수한 인식률을 보였다. 채널 보상을 적용한 후 인식률 결과는 모든 파라미터의 인식률 값이 상승하였고, GFCC가 MFCC 보다 2.51% 향상됨을 알 수가 있다. 또한, 무선(휴대폰)과 유선 전화에 따른 인식률이 약간의 차이가 있는 결과로 각 매체에 따른 채널 왜곡이 다름을 알 수 있었다.

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

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.

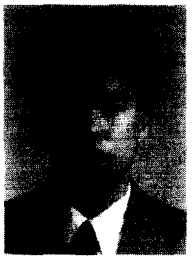
[5] Rivarol Vergin, Douglas O'Shaughnessy, "Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition," IEEE Trans. on Speech & Audio Processing, vol. 7, no. 5, pp. 512-532, 1999.

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

[7] 조태현, 김유진, 이재영, 정재호, "전화선 채널이 화자확인 시스템의 성능에 미치는 영향", 한국음향학회지 제18권, 제5호, 1999

[8] C. J. Moore and R. Glasberg, "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns," J. Acoust. Soc Am., vol. 74, pp. 750-753, Sep. 1983.

저 자 소 개



崔 炯 箕(正會員)

1995年 2月 : 전북대학교 전자공학과 졸업. 1997年 2月 : 전북대학교 대학원 전자공학과 공학 석사 학위 취득. 1999年 2月 : 전북대학교 대학원 전자공학과 박사과정 수료. <주관심분야 : 음성인식, 음성합성>

朴 基 榮(正會員) 第36卷 T編 第2號 參照

金 種 玟(正會員) 第36卷 T編 第2號 參照