논문 2004-41SP-6-31

PCA-optimized 필터뱅크 기반의 MFCC 특징파라미터 추출 및 한국어 4연숫자 전화음성에 대한 인식실험

(Extraction of MFCC feature parameters based on the PCA-optimized filter bank and Korean connected 4-digit telephone speech recognition)

정 성 윤*, 김 민 성*, 손 종 목**, 배 건 성**

(Sungyun Jung, Minsung Kim, Jongmok Son, and Keunsung Bae)

요 약

음성신호의 스펙트럼으로부터 MFCC를 추출할 때, 일반적으로 필터뱅크의 처리과정에서 삼각형 형태의 필터를 사용한다. 그러나 더 나은 인식성능을 위해, 훈련 음성데이터의 스펙트럼에 PCA를 적용하여 필터뱅크의 필터형태를 최적화하는 PCA-optimized 필터뱅크 방법이 Lee et al. 에 의해 제안되었다. 본 논문에서는 대용량의 4연숫자 전화음성 DB를 사용하여 PCA-optimized 필터뱅크 기반의 MFCC 특징파라미터를 추출하고 인식실험을 수행한 후, 기존의 삼각형 형태의 필터를 사용하는 MFCC와 각 대역별 로그에너지로 가중시켜서 얻어지는 MFCC와의 인식성능을 비교하였다. 실험결과, PCA-optimized 필터뱅크 기반의 MFCC 특징파라미터가 기존의 삼각형 형태의 필터뱅크 기반 MFCC에 비해 조금 향상된 인식률을 나타내었지만, 각 대역별 로그에너지로 가중치를 주어 얻어지는 MFCC보다는 인식률이 떨어졌다.

Abstract

In general, triangular shape filters are used in the filter bank when we extract MFCC feature parameters from the spectrum of the speech signal. A different approach, which uses specific filter shapes in the filter bank that are optimized to the spectrum of training speech data, is proposed by Lee et al. to improve the recognition rate. A principal component analysis method is used to get the optimized filter coefficients. Using a large amount of 4-digit telephone speech database, in this paper, we get the MFCCs based on the PCA-optimized filter bank and compare the recognition performance with conventional MFCCs and direct weighted filter bank based MFCCs. Experimental results have shown that the MFCC based on the PCA-optimized filter bank give slight improvement in recognition rate compared to the conventional MFCCs but fail to achieve better performance than the MFCCs based on the direct weighted filter bank analysis. Experimental results are discussed with our findings.

Keywords : PCA-optimized 필터뱅크, 4연숫자 전화음성인식, MFCC특징파라미터 추출

I. 서 론

전화음성의 인식률은 전화망 환경에서 수반되는 신호의 왜곡 및 잡음으로 인해 일반 마이크 음성의 인식률에 비해 만족스럽지 못한 수준이며, 특히, 한국어 연속숫자음의 경우 다양한 조음효과로 인해 인식에 어려움이 많다. 따라서 유/무선 전화망 환경에서 연속숫자음의 인식성능을 향상시키기 위한 연구가 국내에서도 지속적으로 수행되어 왔는데, 특히 배경잡음에 강인한 특징파라미터의 추출 기법 및 채널보상 기법에 대한 연구가 지속적으로 진행되고 있다[1][2]. 음성인식에 사용되는 특징파라미터로는 MFCC(Mel Frequency Cepstral Coefficient)가 가장 보편적으로 사용되는데, 이것은 MFCC가 다른 특징파라미터에 비해 계산량도 적고 수월하게 얻을 수 있으면서도 비교적 좋은 인식성능을 보이기 때문이다. 따라서, MFCC를 기반으로 인식률을 향

* 학생회원, ** 정회원, 경북대학교
(Department of Electronic Engineering, Kyungpook National University)
접수일자: 2003년12월4일, 수정완료일: 2004년9월24일

상 시킬 수 있는 특징파라미터를 추출하기 위한 연구도 꾸준히 진행되고 있다[3][4]. MFCC를 추출하는 과정은 크게 음성신호의 스펙트럼을 필터뱅크로 처리하는 과정과 로그 스펙트럼(log-spectra) 영역에서 DCT(Discrete Cosine Transform)를 이용하여 캡스트럼(cepstrum) 영역으로 변환하는 두 가지 과정으로 나눌 수 있다. 일반적으로 필터뱅크 처리 과정에서는 음성신호의 로그 스펙트럼에 멜(mel) 대역별로 삼각형 형태의 창함수(window function)를 적용하여 에너지를 계산하는데, 이러한 과정에서 원 음성신호의 스펙트럼 특성이 충분히 고려되지 못하여 다소의 정보손실이 발생한다고 볼 수 있다. 이러한 정보 손실을 최소화 하고, 좀 더 변별력 있는 음성특징을 표현하기 위해 필터뱅크의 필터형태를 훈련 음성데이터의 스펙트럼에 PCA(Principal Component Analysis)를 적용하여 얻은 후, 이를 이용하여 대역별 에너지를 계산하고 MFCC를 구하는 방법이 Lee et al.[5] 에 의해 제안된 바 있다. 따라서, 본 연구에서는 대용량 전화음성 DB(Data Base)인 SITEC (Speech Information Technology & Industry Promo -tion Center)의 4연숫자음 전화음성에 Lee et al.가 제안한 방법을 적용하여 MFCC 특징파라미터를 추출하고 인식실험을 수행하여, 기존의 삼각형 형태의 필터를 사용하는 MFCC 와 각 대역별 로그에너지로 가중시켜서 얻어지는 DWFBA(Direct Weighted Filter Bank Analysis)기반의 MFCC[6]와의 인식성능을 비교하였다. 또한 각 특징파라미터에 대해 CMN(Cepstrum Mean Normalization) 및 MRTCN(Modified Real Time Ceps -trum Normalization)의 채널보상 기법을 적용하여 얻어지는 인식성능도 함께 비교하였다.

본 논문의 구성은 다음과 같다. Ⅰ장의 서론에 이어 Ⅱ장에서는 PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터를 추출하는 방법에 대해 기술한다. 그리고, Ⅲ장에서 인식실험 환경 및 실험결과를 제시하고, Ⅳ장에서 결론을 맺는다.

Ⅱ. PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터 추출

PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터 추출과정[5]은 크게 2가지 과정으로 나눌 수 있다. 훈련데이터로부터 PCA-optimized 필터뱅크의 필터계수를 구하는 과정과 구한 필터계수를 사용하여 필터뱅크 처리를 통한 MFCC 특징파라미터 추출과정이다. 본 장에서는 각각의 처리과정에 대해 설명한다.

1. PCA-optimized 필터뱅크의 필터계수 추출

PCA의 목적은, 데이터 집합 내에 있는 변이를 가능한 한 많이 유지하면서, 상관관계를 갖는 많은 변수들로 이루어진 데이터 집합의 차원을 감소시키는 것이다[7]. 즉, $N\times 1$의 랜덤벡터 x가 있을 때, 식 (1)에서 각 w_i와 x의 곱이 최대변이를 갖도록, $N\times 1$의 orthonormal 벡터들의 집합, $\{w_i | 1 \le i \le k, k \le N\}$을 구하는 것이다.

$$y_i = w_i^T x \tag{1}$$

이때, 벡터집합 $\{w_i\}$는 k개의 고유치들에 해당하는 x의 covariance matrix의 고유벡터들에 해당한다. 이러한 PCA의 개념은 MFCC를 추출할 때 사용되는 필터뱅크의 필터형태를 훈련음성 DB에 적응된 최적의 필터모양을 구하는 문제에 적용될 수 있다. 즉, 필터뱅크의 각 필터는 차원 감소의 과정으로 볼 수 있고, 각 필터의 주파수 대역에 있는 신호 요소들은 필터와 가중되어져서 하나의 값으로 표현될 수 있다.

PCA를 적용한 필터뱅크의 필터 추출 과정은, 그림 1과 같이 음성신호의 스펙트럼에서, 먼저 K개의 필터뱅

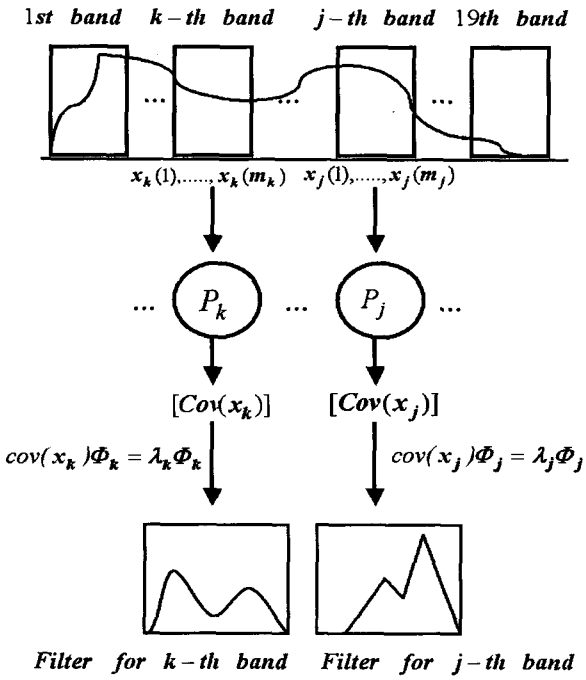


그림 1. PCA-optimized 필터뱅크의 필터계수를 구하는 과정

Fig. 1. The process of finding PCA-optimized filter bank coefficients.

크를 정해놓고, 각 필터뱅크에 속하는 스펙트럼 값들을 벡터 $x_j, j=1,2,...K$ 로 정의한다. 이때, 필터뱅크의 j번째 필터의 주파수 대역에 속한 스펙트럼 요소들의 수가 m_j라 한다면, j번째 주파수 대역에 속한 m_j개의 신호성분들을 식 (2)와 같이 벡터 형태로 정의할 수 있다.

$$x_j = [x_j(1), x_j(2), ..., x_j(m_j)]^T \quad (2)$$

훈련음성 DB의 모든 데이터에 대해, 각 필터뱅크에 해당하는 스펙트럼 값들을 벡터 형태로 수집하여, 각 필터뱅크에 해당하는 Pool, $P_j, j=1,2,...,K$ 에 수집한다. 훈련음성 DB에 대한 모든 벡터들을 구한 후, 식 (3)과 같이, 각 Pool에 속한 벡터들의 covariance matrix, $cov(x_j), j=1,2,...,K$를 계산한다. 식 (3)을 사용하여 각 필터뱅크에 해당하는 covariance matrix를 구하고 나면, 식 (4)와 같이 고유치 문제를 적용하여 고유치, λ_j와 해당 고유벡터, Φ_j를 구할 수 있다. 이때, μ_j는 j번째 Pool에 속한 x_j들의 평균벡터이다. 최종적으로, 각 필터뱅크의 covariance matrix의 λ_j에서 가장 큰 대각성분에 해당하는 Φ_j의 열벡터가 해당 필터의 계수 값으로 결정된다.

$$cov(x_j) = E[(x_j - \mu_j)(x_j - \mu_j)^T] \quad (3)$$

$$cov(x_j)\Phi_j = \lambda_j \Phi_j \quad (4)$$

2. PCA-optimized 필터뱅크 기반의 MFCC추출

PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터의 추출 과정은 그림 2와 같다. 전처리 및 해밍 윈도우된 음성프레임을 STFT(Short Time Fourier Transform)을 통해 전력스펙트럼을 구한 후, 필터뱅크

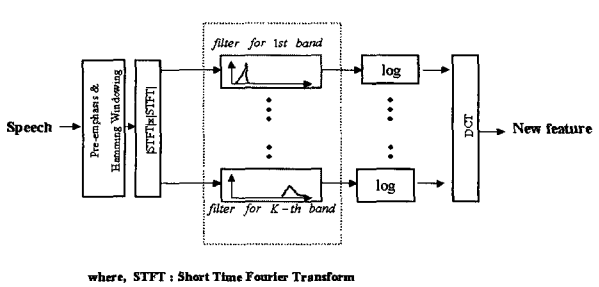


그림 2. PCA-optimized 필터뱅크 기반의 MFCC 추출 과정

Fig. 2. The process of extracting MFCC based on the PCA-optimized filter bank.

처리과정에서 삼각형 형태의 필터를 사용하지 않고, 앞 절에서 구한 각 필터뱅크의 필터계수를 사용한다. 따라서, PCA-optimized 필터뱅크 기반의 MFCC 추출은 삼각형 형태의 필터대신 PCA를 적용하여 구한 필터를 사용한다는 점을 제외하고는 기존의 MFCC 추출과정과 동일하다.

Ⅲ. 실험 및 결과

1. 4연숫자 전화음성 DB

음성정보기술산업지원센터(SITEC)에서 제작된 한국어 4연숫자 전화음성 DB는 총 2000명 화자의 음성으로 이루어져 있는데, 유선전화, 무선전화, cellular, PCS 전화음성이 모두 포함되어 있다[8]. 녹음 환경은 연구실과 사무실, 가정집 환경으로 이루어져 있고, 모든 전화음성은 8kHz 샘플링에 16bits/sample, 선형 PCM 형태로 파일에 저장되어 있다. 전화음성 파일은 각 화자별로 폴더에 저장되어 있는데, 각 폴더명 및 음성 파일은 일정한 규칙을 가지고 있다. SITEC 전화음성 DB에서는 훈련용 데이터로 1800명 화자의 58388개의 4연숫자음 데이터가 설정되어 있고, 테스트용 데이터로는 200명 화자의 6468개의 4연숫자음 데이터가 설정되어 있다. 4연숫자음은 총 1620 종류인데, 50등분 되어 각 화자당 32개의 4연숫자음으로 구성되어 저장되어 있고, 테스트용 데이터에는 1620 종류의 4연숫자음이 모두 포함되어 있다.

2. PCA-optimized 필터뱅크의 필터 형태

본 논문에서는 PCA를 적용한 필터뱅크의 필터 계수를 구하기 위해, SITEC 전화음성 DB중 훈련용 58388개의 전화음성 DB를 사용하였다. 모든 훈련음성 DB에 멜 대역별로 19개의 필터뱅크를 적용하였고, 이웃 대역간의 중심주파수를 각 대역의 경계로 설정하였다. 그림 3에 훈련용 DB로부터 얻은 19개의 PCA 적용 필터뱅크의 필터 형태를 첫 번째 필터부터 순서대로 나타내었다. 그림에서 가로축은 FFT(Fast Fourier Transform)의 주파수 샘플 수를 4000Hz까지를 나타낸 것인데, 필터뱅크의 모든 필터가 일반적으로 많이 사용되는 삼각형의 형태를 나타내고 있지 않으며 각 대역별로 서로 다른 모양을 하고 있음을 볼 수 있다.

3. 인식실험 및 결과

4연숫자 전화음성 인식기는 HTK(Hidden Markov

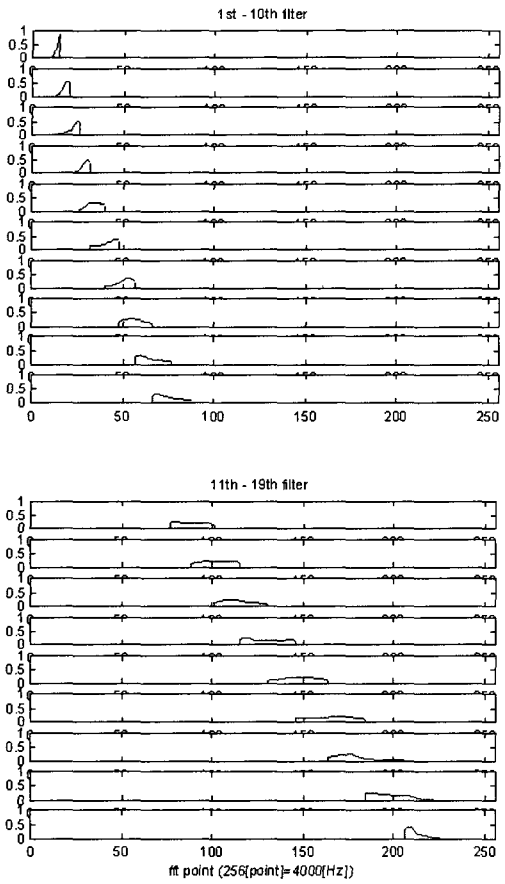


그림 3. PCA-optimized 필터뱅크의 19개 필터의 형태
Fig. 3. The shape of 19 filters in PCA-optimized filter bank.

Tool Kit)를 사용하여 구현하였다[9]. 음성신호로부터 20 ms 의 분석 구간에 10 ms 씩 중첩 이동하면서 특징파라미터를 추출하였다. 음향모델은 트라이폰(triphone) HMM(Hidden Markov Model)을 사용하였는데, 육과 륙을 구분하여 모두 17개의 음소를 정의하였고, 5 states, 9 mixture의 연속 HMM 모델을 적용하였다. 또한, 4연숫자음 인식의 특성을 고려하여, 언어모델은 FSN(Finate State Network)을 사용하였다.

인식에 사용된 전체 특징파라미터의 구성은 12차의 멜켑스트럼 및 이들의 차분, 차차분 그리고 차분 에너지 및 차차분 에너지를 포함한 총 38차로 이루어져 있으며, 인식실험에 사용된 특징파라미터들은 기본 특징파라미터인 MFCC와 DWFBA, 그리고 PCA-optimized 필터뱅크 기반의 MFCC이다.

표 1은 특징파라미터 및 채널 보상기법 적용에 따른 인식실험의 결과를 보인 것이다. 표에서 MFCC는 기존의 삼각형 형태의 필터를 사용하여 특징파라미터를 구한 경우를 나타내고, MFCC_PCA는 PCA-optimized 필

표 1. 특징파라미터 및 채널 보상기법에 따른 인식실험 결과(4연숫자음 인식률 / 개별숫자 인식률)
Table 1. Recognition experiment results for various feature parameters and channel compensation techniques. (4-digit connected speech recognition rate / word recognition rate)

Feature Parameter	Recognition rate(%)		
	Without channel compensation	Channel compensation with CMN	Channel compensation with MRTCN
MFCC	87.06/96.17	88.19/96.48	88.65/96.70
MFCC_PCA	87.34/96.32	89.10/96.82	89.84/97.10
MFCC_DWFBA	88.54/96.71	90.28/97.19	90.88/97.36

터뱅크를 이용하여 구한 경우이며, MFCC_DWFBA는 필터뱅크 출력에 적절한 가중치를 준 다음 DCT를 취해 MFCC를 구한 경우[6]를 나타낸다. 기존의 삼각형 형태의 필터뱅크를 기반으로 하는 MFCC 특징파라미터를 사용한 인식기를 기준(baseline)으로 인식성능을 비교하면, MFCC_PCA는 0.28 %의 인식률 증가를 보였으며, 여기에 채널 보상기법으로 CMN을 적용할 때에는 MFCC에 CMN을 적용한 경우에 비해 0.91 %, MRTCN을 적용할 때에는 MFCC에 MRTCN을 적용한 경우에 비해 1.19 %의 인식률 증가를 나타내었다. 그러나 MFCC_DWFBA 보다는 채널보상 기법을 적용한 경우나 적용하지 않은 경우나 상관없이 대체적으로 약 1 % 정도 인식율이 낮았다. 따라서, 필터 형태를 훈련음성 DB에 적응시킨 필터뱅크 기반의 특징파라미터인 MFCC_PCA가 기존의 MFCC에 비해 다소 나은 인식성능을 나타낸다고 할 수 있으나 MFCC_DWFBA에 비해서는 전체적으로 인식률이 낮았다. 결국, Lee et al.이 주장한 방법이 잡음환경에서는 강인한 성능을 나타내었다 할지라도 전화망 환경에서는 뛰어난 성능 향상을 볼 수 없었다.

Ⅳ. 결 론

본 논문에서는 한국어 4연숫자 전화음성의 인식실험에서 훈련음성 DB로부터 얻어지는 PCA-optimized 필터 형태를 갖는 필터뱅크 기반의 MFCC 특징파라미터를 추출하여 기존의 삼각형 형태의 필터뱅크를 이용하는 MFCC 특징파라미터와 각 대역별 로그에너지로 가중시켜 구한 MFCC와의 인식성능을 비교하였다. SITEC의 전화음성 DB를 사용한 인식실험 결과, PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터가 기

존의 삼각형 모양의 필터뱅크 기반의 MFCC에 비해 0.28%의 인식률 증가를 보였으며, 여기에 채널 보상기법인 CMN을 적용한 경우에는 0.91 %, MRTCN을 적용한 경우에는 1.19 %의 인식률 증가를 보였었다. 그러나 기존의 필터뱅크에 각 대역별 로그에너지로 가중치를 주어 얻어지는 MFCC 특징파라미터에 비해서는 전체적으로 1.04 %~1.2 % 인식률이 떨어졌었다. 따라서, 한국어 4연숫자 전화음성 인식을 위해 훈련 음성 DB에 적응된 필터뱅크 형태를 사용하여 MFCC를 구하는 기법이 기존의 MFCC에 비해 약간의 인식률 증가를 얻을 수 있겠지만 전화음성 연속숫자음의 인식성능 향상에 크게 기여하기는 어렵다고 생각된다.

참 고 문 헌

[1] 정성윤, 김민성, 손종목, 배건성, 김상훈, "채널보상 기법 및 특징파라미터추출 방법에 따른 연속숫자음 전화음성의 인식성능향상," 대한음성학회 정기총회 및 학술발표대회 논문집, 201-203쪽, 2002.

[2] 김성탁, 김상진, 정호영, 김회린, 한민수 "전화망 환경에서의 연속숫자음 인식 성능평가," 한국음향학회 논문집, 제 21권 1호, 253-256쪽, 2002.

[3] A.Biern, S.Katagiri, E.McDermott and B.H.Juang, "An application of discriminative feature extrac -tion to filter-bank based speech recognition," IEEE Transaction on Speech and Audio Processing, Vol.9, no.2, Feb. 2001.

[4] C. Benitez, L. Burget, H.Hermansky, P.Jain, and N.Morgan, "Robust ASR front-end spectral -based and discriminant features : experiments on the Aurora tasks," Proc. Eurospeech, 2001.

[5] S. M. Lee, S. H. Fang, J. Hung, and L. S. Lee, "Improved mfcc feature extraction by pca -optimized filter-bank for speech recognition," Automatic Speech Recognition and Under -standing, pp. 49-52, 2001.

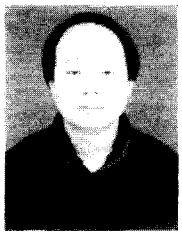
[6] 정성윤, 김민성, 손종목, 배건성, 김상훈, "한국어 연속숫자음 전화음성의 인식성능 개선," 대한전자공학회 추계학술대회 논문집, 제 25권 2호, 582 -585쪽, 2002.

[7] I. T. Jolliffe, Principal component analysis, Springer Verlag, 2002.

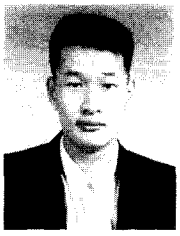
[8] http://www.sitec.or.kr/index.asp

[9] Steve Young, Gunnar Evermann and D. Kershaw, The HTK Book (HTK Version 3.0), Cambridge, 2000.

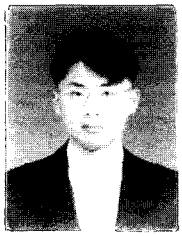
저 자 소 개



정 성 윤(학생회원)
1991년 2월 경북대학교 전자공학과 (공학사)
1994년 2월 영남대학교 전자공학과(공학석사)
2000년 3월~현재 경북대학교 전자공학과 박사과정.
<주관심분야: 디지털 신호처리, 음성신호처리, 음성인식>



김 민 성(학생회원)
2001년 2월 경북대학교 전자공학과 (공학사)
2003년 2월 경북대학교 전자공학과 (공학석사)
2003년 3월~현재 경북대학교 전자공학과 박사과정
<주관심분야 : 음성신호처리>



손 종 목(정회원)
1997년 2월 경북대학교 전자공학과(공학사)
1999년 2월 경북대학교 전자공학과(공학석사)
1999년 3월~현재 경북대학교 전자공학과 박사과정
<주관심분야: 디지털 신호처리, 음성신호처리, 음성인식 >

배 건 성(정회원)
1977년 2월 서울대학교 전자공학과(공학사)
1979년 2월 한국과학기술원 전기 및 전자공학과(공학석사)
1989년 5월 University of Florida (공학박사)
1979년 3월~현재 경북대학교 전자공학과 교수
<주관심분야: 음성분석 및 인식, 디지털 신호처리, 음성 부호화, 웨이브렛 분석>