훈련음성 데이터에 적응시킨 필터뱅크 기반의 MFCC 특징파라미터를 이용한 전화음성 연속숫자음의 인식성능 향상에 관한 연구

정 성 윤, 김 민 성, 손 종 목, 배 건 성, 강 점 자 *
경북대학교 전자공학과, 한국전자통신연구원 *

A study on the recognition performance of connected digit telephone speech for MFCC feature parameters obtained from the filter bank adapted to training speech database

Sung Yun Jung, Min Sung Kim, Jong Mok Son, Keun Sung Bae, Jeom Ja Kang *
School of the Electronic & Electrical Engineering, Kyungpook National University,
Electronics Telecommunications Research Institute *
E-mail : yunij@mir.knu.ac.kr

Abstract

In general, triangular shape filters are used in the filter bank when we get the MFCCs from the spectrum of speech signal. In [1], a new feature extraction approach is proposed, which uses specific filter shapes in the filter bank that are obtained from the spectrum of training speech data. In this approach, principal component analysis technique is applied to the spectrum of the training data to get the filter coefficients. In this paper, we carry out speech recognition experiments, using the new approach given in [1], for a large amount of telephone speech data, that is, the telephone speech database of Korean connected digit released by SITEC. Experimental results are discussed with our findings.

1. 서론

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 연속숫자음 음성인식에서는 일반적으로 MFCC(Mel Frequency 의 인식성능을 향상시키기 위한 연구는 국내에서도 지속적으로 수행되어 왔다[2,3].

Cepstral Coefficient)가 비교적 좋은 특징파라미터로 사용되고 있는데, MFCC를 기반으로 더 좋은 인식률을 얻을 수 있는 특징파라미터를 추출하기 위한 연구도 꾸준히 진행되고 있다[4,5]. MFCC를 추출하는 과정은 크게 필터뱅크 처리과정과 log-spectra 영역에서 DCT(Discrete Cosine Transform)를 이용하여 cepstral 영역으로 변환하는 두 가지 처리과정으로 나눌 수 있다. 일반적으로 필터뱅크 처리과정에서는 음성신호의 스펙트럼에다가 삼각형 창함수를 적용하여 대역별 에너지를 계산하는데, 이러한 과정에서 원 음성신호의 스펙트럼 특성이 충분히 고려되지 못하여 다소의 정보손실이 발생한다고 볼 수 있다. 이러한 정보손실을 최

소화 하고, 좀 더 변별력 있는 음성특징을 표현하기 위해 필터뱅크의 형태를 훈련 음성데이터의 스펙트럼에 PCA(Principal Component Analysis)를 적용하여 얻은 후, 이를 이용하여 MFCC를 구하는 방법이 [1]에서 제안된 바 있다. 따라서, 본 연구에서는 대용량 전화음성 DB(Database)인 SITEC(Speech Information Technology & Industry Promotion Center)의 4연숫자음 전화음성에 [1]에서 제안된 방법을 적용하고, 인식실험을 수행하여 기존의 MFCC 특징파라미터 및 보상기법 등을 적용한 결과와 인식성능을 비교, 검토하고자 하였다.

SITEC의 4연숫자음 전화음성 DB를 사용한 인식실험에서, PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터를 사용한 경우가 기존의 심각형 모양의 필터뱅크 기반의 MFCC에 비해 약 0.3%의 인식률 증가를 보였으며, 여기에 채널 보상기법인 CMN(Cepstral Mean Normalization)을 적용할 때에는 MFCC에 CMN을 보상기법으로 적용한 경우에 비해 약 1.0%의 인식률 증가를 나타내었다. 다시 말해, 훈련음성에 적응된 필터뱅크 형태를 사용하여 MFCC를 구함으로써 기존의 방법에 비해 약간의 인식률 증가를 얻을 수 있었지만 [1]에서 주장한 것처럼 뛰어난 성능 향상은 볼 수 없었다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 [1]에서 제안한 PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터를 추출하는 방법에 대해 기술한다. 그리고, 3장에서 인식실험 및 결과를 검토한 후, 4장에서 결론을 맺는다.

II. PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터 추출

PCA의 개념은, 데이터 셀(set) 내에 있는 변이를 가능한 한 많이 유지하면서, 상관관계를 갖는 많은 변수들로 이루어진 데이터 셀의 차원을 감소시키는 것이다 [6]. PCA에 대해 간략히 설명하면 다음과 같다. 만약 x가 $N\times 1$의 랜덤 벡터라면, PCA의 목적은 식 (1)의 각 w_i 와 x의 곱이 최대변이를 갖도록, $N\times 1$의 orthonormal 벡터들의 셀 $w_i|1\le i\le k, k\le N$을 구하는 것이다.

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

이때, 벡터 셀 w_i는 가장 큰 k 개의 고유치들에 해당하는 x의 covariance matrix의 고유벡터들이다. 이러한 PCA의 개념은 MFCC를 추출할 때, 필터뱅크의 필터형태를 훈련음성 데이터에 따라 처리하여 최적의 필터모양을 구하는 문제에 적용될 수 있다. 즉, 필터뱅크의 각 필터는 차원감소의 과정으로 볼 수 있고, 각 필터의 주파수 대역에 있는 신호요소들은 필터와 가중되어져서 하나의 값으로 표현될 수 있다. 이러한 PCA를 적용하여 얻은 필터뱅크 기반의 MFCC 특징파라미터를 추출하는 과정은 다음과 같다.

먼저, 필터뱅크의 k 번째 필터의 주파수 대역에 속한 신호요소들의 수가 m_k 라면, k 번째 주파수 대역에 속한 m_k의 신호성분을 표현하는 랜덤 변수들을 $x_k(n),\ n=1,2,\ldots,m_k$로 표현할 수 있고, x_k를 벡터표현으로 식 (2)와 같이 다시 정의할 수 있다.

$$x_k=[x_k(1), x_k(2), \ldots\ldots\ldots\ldots, x_k(m_k)]^T \tag{2}$$

PCA를 적용한 필터뱅크 추출과정은, 그림 1과 같이 한 프레임의 음성신호의 스펙트럼에서, 먼저 19개의 필터뱅크를 정해놓고, 각 필터뱅크에 해당하는 spectral 요소를 벡터 $x_k,\ k=1,2,..19$ 로 하여 훈련음성 DB로부터 해당 필터들의 spectral component vectors를 구한다. 그리고, 훈련음성 DB에 대해 모든 벡터들을 구하고난 뒤, 식 (3)과 같이 이들의 covariance matrix를 계산한다. 식 (3)을 사용하여, 각 필터에 대한 covariance matrix를 구하고 나면, 식 (4)에서의 $cov(x_k)$의 고유치가 가장 큰 고유벡터를 구하여 해당 필터의 계수 값으로 정의한다.

$$cov(x_k) = E[(x_k-\mu_k)(x_k-\mu_k)^T] \tag{3}$$

$$cov(x_k) = F_k D_k F_k^{-1} \tag{4}$$

이때, F_k 의 열 벡터들이 $cov(x_k)$ 의 고유벡터들이고, D_k 에서 가장 큰 대각성분에 해당하는 F_k의 열 벡터가 k번째 필터의 계수 w_k 가 된다.

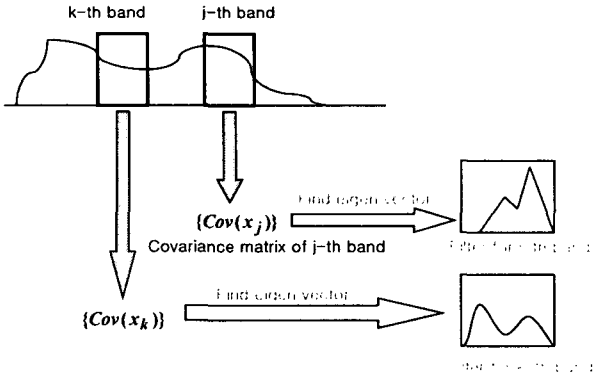

그림 14. PCA를 적용한 필터뱅크의 필터계수를 구하는 과정

본 논문에서는 PCA를 적용한 필터뱅크를 구하기 위해, SITEC 전화음성 DB중 훈련용 58388개의 전화음성 DB를 사용하였다. 모든 훈련음성 DB에 Mel scale로 19개의 필터뱅크를 적용하였고, 이웃 밴드간의 중심주파수를 각 밴드의 경계로 설정하였다. 그림 2에 훈련용 DB로부터 얻은 19개의 PCA 적용 필터뱅크의 필터형태를 첫 번째 필터부터 순서대로 나타내었다. 그림에서 가로축은 512-포인트 FFT(Fast Fourier transform) 스펙트럼에서의 주파수 샘플수를 나타낸 것이다. 필터뱅크의 필터형태가 서로 다른 모양을 하고 있고, 모든 필터가 삼각형의 모양을 나타내지 않음을 그림 2에서 볼 수 있다.

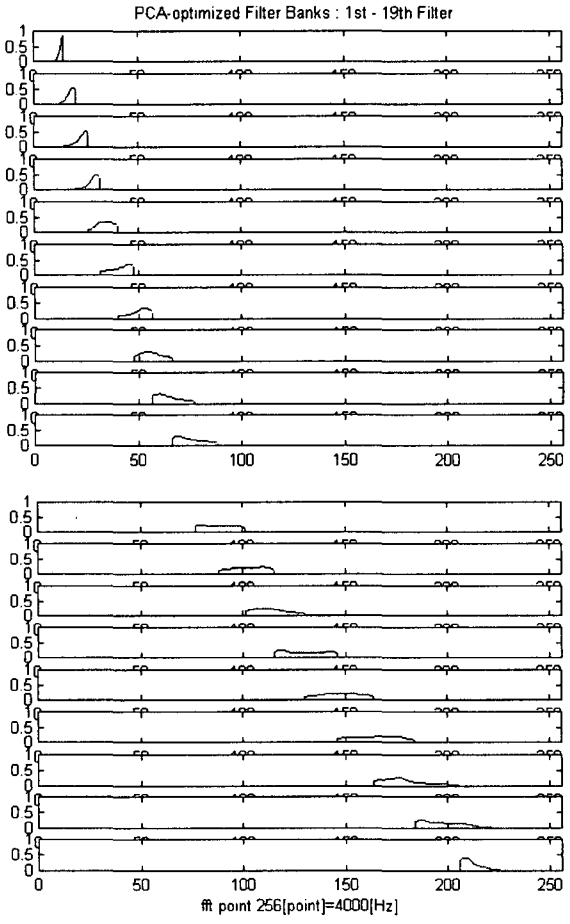

그림 2. SITEC의 훈련용 전화음성 DB로부터 추출한 19개의 PCA적용 필터뱅크 형태

위에서 구한 PCA 적용 필터뱅크를 사용한 MFCC 추출과정은 삼각형 모양의 필터를 사용하는 부분을 제외하고는 기존의 MFCC 추출 과정과 동일하다. 본 논문에서는 20ms 의 해밍 창함수 구간에 10ms 씩 중첩 이동하면서 1차의 에너지, 12차의 멜켑스트럼을 포함하는 13차의 MFCC 특징파라미터를 추출하였다.

III. 인식실험 및 결과

이 장에서는, 2장에서 구한 PCA 적용 필터뱅크기반의 MFCC 특징파라미터를 사용하여 인식실험한 내용과 그 결과를 기술한다. 먼저 인식실험에 사용한 전화음성 DB를 간략히 소개하고, PCA를 적용한 필터뱅크 기반의 MFCC의 인식결과를 기존의 MFCC 특징파라미터 및 보상기법을 적용한 인식결과와 비교, 검토하였다.

3.1 SITEC 전화음성 DB

음성정보기술산업지원센터(SITEC)에서 제작된 한국어 4연숫자음 전화음성 DB는 총 2000명 화자의 음성으로 이루어져 있는데, 유선전화, 무선전화, cellular, PCS 전화음성이 모두 포함되어 있다[7]. 녹음 환경은 연구실과 사무실, 가정집 환경으로 이루어져 있고, 모든 전화음성은 8kHz 샘플링에 16bits/sample linear PCM 형태로 파일에 저장되어 있다. 전화음성 파일은 각 화자별로 폴더에 저장되어 있으며, 각 폴더명 및 음성 파일은 일정한 규칙을 가지고 있다. SITEC 전화음성 DB에서는 훈련용 데이터로 1800명 화자의 58388개의 4연숫자음 데이터가 설정되어 있고, 테스트용 데이터로 200명 화자의 6468개의 4연숫자음 데이터가 설정되어 있다. 또한, 1620 종류의 4연숫자음은 50등분되어 화자당 32개의 4연숫자음으로 구성되어 저장 되어있고, 테스트용 데이터에는 1620 종류의 4연숫자음이 모두 포함되어 있으며, 훈련에 나타나지 않은 4연속자음은 포함되지 않았다. 그리고, 숫자음 중 "륙"과 "육"은 서로 다른 단어로 구분되어 레이블링 되어있다.

3.2 인식실험 및 결과

4연숫자음 인식기는 HTK(Hidden Markov Tool Kit)를 사용하여 구현하였다[8]. 음성신호는 20ms 의 분석 구간에 10ms 씩 중첩 이동하면서 특징파라미터를 추출하였다. 특징 파라미터로는 12차의 멜켑스트럼 및 이들의 차분, 차차분 그리고 차분 에너지 및 차차분 에너지를 포함한 총 38차를 사용하였으며, 음향모델은 트라이폰(triphone) HMM 모델을 사용하였는데, 육과 륙을 구분하여 모두 17개의 음소를 정의하였고, 5 states, 9 mixture의 연속 HMM 모델을 적용하였다. 다. 또한, 4연숫자음 인식의 특성을 고려하여, 언어모델은 FSN(Finate State Network)을 사용하였다.

인식실험에 사용된 음성데이터는 1620 종류의 4연숫자음에 대해 2000명의 화자(남자 980명, 여자1020명)가 발성한 64856개이다. 이 중 1800명분의 58388개의 숫자음성을 훈련에 사용하였고, 200명이 발성한 6468개의 숫자음성을 테스트에 사용하였다.

표 1은 특징파라미터 추출방법 및 보상기법 적용에 따른 인식실험의 결과를 보인 것이다. 본 논문에서의 실험결과와의 비교를 위해, 필터뱅크 출력에 적절한 가중치를 준 다음 DCT를 취해 MFCC를 구하는 방법인 DWFBA(Direct Weighted Filter Bank Analysis)를 이용한 인식결과[9]를 함께 나타내었다. 기존의 삼각형 형태의 필터뱅크를 기반으로 하는 MFCC 특징파라미터를 사용한 baseline 인식기를 기준으로 비교하면, PCA를 적용한 필터뱅크 기반의 MFCC 특징파라미터가 약 0.3%의 인식률 증가를 나타내었고, 여기에 보상기법인 CMN을 적용할 때에는 기본의 MFCC에 CMN을 보상기법으로 적용한 경우에 비해 약 1.0%의 인식률 증가를 나타내었다. 따라서, 필터형태를 훈련음성 DB에 적응시킨 필터뱅크 기반의 MFCC 특징추출이 기존의 MFCC에 비해 다소 나은 인식성능을 나타낸다고 할 수 있다. 그러나, DWFBA를 사용한 경우보다는 전체적으로 인식률이 낮았으며, [1]에서 주장한 것처럼 뛰어난 성능 향상은 볼 수 없었다.

표 1. 특징파라미터 및 보상기법에 따른 인식결과
(4연속숫자열 인식률/개별숫자 인식률)

특징파라미터	인식률(%)
MFCC	87.06 / 96.17
PCA	87.34 / 96.32
DWFBA	88.54 / 96.71
MFCC+CMN	88.19 / 96.48
PCA+CMN	89.10 / 96.82
DWFBA+CMN	90.28 / 97.19

Ⅳ. 결론

본 논문에서는, 기존의 심각형 모양의 필터뱅크 기반의 MFCC에 비해 훈련음성 DB로 부터 추출한 PCA 적용 최적 필터형태를 갖는 필터뱅크 기반의 MFCC 특징파라미터가 더 나은 인식성능을 나타낸다는 [1]의 논문을 검토하였다. 이를 위해 대용량 전화음성 DB인 SITEC 4연숫자음 전화음성에 [1]에서 제안된 방법을 적용하고, 인식실험을 수행하여 기존의 MFCC 특징파라미터 및 보상기법 등을 적용한 인식결과와 비교, 검토 하였다.

인식실험결과, PCA를 적용한 필터뱅크 기반의 MFCC의 특징파라미터가 기존의 MFCC에 비해 약 0.3%의 인식률 증가를 나타내었고, 여기에 보상기법인 CMN을 적용하면, 약 1.0%의 인식률 증가를 얻을 수 있었다. 하지만, 전체적으로 DWFBA를 사용한 경우보다는 인식률이 낮았으며, [1]에서 주장한 것처럼 뛰어난 성능 향상은 얻을 수 없었다.

본 연구는 한국전자통신연구원 네트워크기술연구소 음성정보연구센터의 연구비 지원으로 수행되었습니다.

참고문헌

[1] Shang Ming Lee, Shi Hau Fang, Jeih weih Hung, Lin Shan Lee, "Improved mfcc feature extraction by pca-optimized filter-bank for speech recognition,"*Automatic Speech Recognition and Understanding*, pp. 49-52, 2001

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

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

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

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

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

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

[8] Steve Young, Gunnar Evermann and D. Kershaw, *The HTK Book (HTK Version 3.1)*, Cambridge University Engineering Department

[9] 최종연구보고서, *전화망 환경에서의 연속숫자음 신호왜곡 연구*, 전자통신연구원, 2002