훈련데이터 기반의 temporal filter를 적용한 한국어 4연숫자 전화음성의 인식실험

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

Recognition experiment of Korean connected digit telephone speech using the temporal filter based on training speech data

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**

yunij@mir.knu.ac.kr

Abstract

In this paper, data-driven temporal filter methods[1] are investigated for robust feature extraction. A principal component analysis technique is applied to the time trajectories of feature sequences of training speech data to get appropriate temporal filters. We did recognition experiments on the Korean connected digit telephone speech database released by SITEC, with data-driven temporal filters. Experimental results are discussed with our findings.

I. 서론

유/무선 전화음성의 인식은 전화망 환경에서 수반되는 신호의 왜곡 및 잡음으로 인해 마이크 음성에 비해 인식성능이 저하되는 문제점이 있다. 이러한 문제점을 극복하기위해, CMN(Cepstral Mean Normalization), MRTCN(Modified Real Time Cepstral Normalization), RASTA(Relative Spectra) 등과 같은 기법을 사용하여 채널왜곡 및 배경잡음에 강인한 파라미터를 추출하고자하는 연구가 이어져왔다[2,3,4]. 이러한 기법들은 채널왜곡이나 잡음을 제거하기 위해 음성 특징파라미터 시퀀스에 HPF(High Pass Filter) 나 BPF(Band Pass Filter)의 필터링을 수행한 것이며, 이는 인식태스크에 독립적이다. 특정 인식태스크의 특성을 고려해 줄 수 있다면 보다 효율적인 보상을 수행하는 게 가능 할 수 있다. 따라서, 적용할 인식태스크에 최적인 필터계수를 구하기 위해, PCA(Principal Component Analysis), LDA(Linear Discriminant Analysis) 그리고 MCE(Minimum Classification Error)와 같은 데이터 기반의 접근방법들이 연구되고 있다[5,6].

본 논문에서는, 한국어 4연숫자 전화음성의 인식성능 개선을 위한 특징파라미터의 연구를 위해, PCA를 적용하여 훈련 DB의 특징파라미터 time trajectories로부터 temporal filter 들을 구한 후, 이를 사용하여 인식성능이 향상된 특징파라미터를 구한 [1]의 논문을 검토하였다. SITEC (Speech Information Technology & Industry Promotion Center)의 4연숫자 전화음성 DB에 [1]에서 제안한 방법을 적용한 인식실험을 수행하여, 데이터 기반의 temporal 필터링 방법에 의한 한국어 4연숫자 전화음성의 인식성능을 알아보았다.

본 논문[1]의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 [1]에서 제안한 데이터 기반의 temporal filters에 대해 기술한다. 그리고, 3장에서 인식실험 및 결과를 검토한 후, 4장에서 결론을 맺는다.

II. PCA를 적용한 temporal filters

K 차원의 특징벡터 $x(n)$이 그림 1과 같이 프레임

1) 본 논문은 한국전자통신연구원 네트워크기술연구소 음성정보연구센터의 연구비 지원으로 수행되었습니다.

시간축에 따라 순차적으로 나열되어 있다면, $x(n)$은 식 (1)과 같이 표현 될 수 있고, $x(n)$의 k번째 time trajectory 는 $[x(1,k)\,x(2,k).....x(N,k)]$로 표현되고, $y_k(n)=x(n,k)$로 표현된다. 여기에서 n n은 time index 이고, k 는 특징벡터의 차수에 대한 index이다.

$$x(n)=[x(n,1),....,x(n,k),....x(n,K)]^T,$$
$$n=1,2,...N \qquad k=1,2,...K \qquad (1)$$

데이터기반의 temporal filter는 k번째 time trajectory $y_k(n)$을 필터링하는 L 샘플 FIR(Finite Impulse Response) filter $W_k(z)$이다. 이를 위해 먼저, k번째 time trajectories에 대해, L개의 특징파라미터를 취하여 식(1)과 같이 L차원의 벡터 $z_k(n)$을 얻는다.

$$z_k(n)=[y_k(n)\,y_k(n+1).....y_k(n+L-1)]^T,$$
$$n=1,2,.....N-L+1 \qquad (2)$$

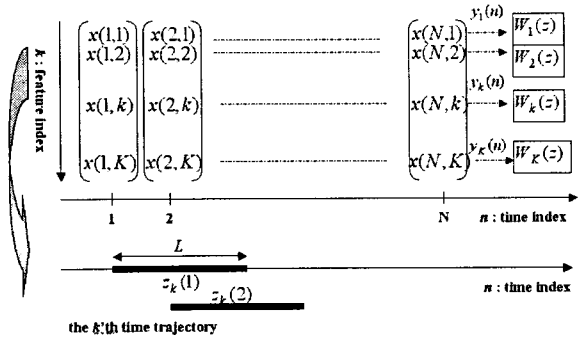

그림 1. 특징파라미터의 time trajectory 표현

본 논문에서의 인식실험에 사용된 데이터 기반의 temproral filter는 PCA에 의한 고유벡터의 적용방법에 따라 모두 3가지로 구분될 수 있는데, [1]에서 제안한 SETF(Single-eigenvector temporal filters)와 METF(Multi-eigenvector temporal filters) 외에 본 논문에서 제안한 PETE(Proposed eigenvector temporal filters)이다.

1. SETF(Single-eigenvector temporal filters)

L 차원의 벡터인 $z_k(n)$을 랜덤벡터 z_k의 샘플들로 본다면, z_k의 평균벡터와 covariance matrix는 식 (3),(4)와 같이 계산될 수 있다.

$$\mu_{z_k}=\frac{1}{N-L+1}\sum_{n=1}^{N-L+1} z_k(n) \qquad (3)$$

$$\Sigma_{z_k}=\frac{1}{N-L+1}\sum_{n=1}^{N-L+1}(z_k(n)-\mu_{z_k})(z_k(n)-\mu_{z_k})^T \qquad (4)$$

그리고, PCA의 계산 과정에 따라, covariance matrix 의 가장 큰 고유치에 해당하는 고유벡터 ϕ_k를 L샘플의 filter 계수로 선택한다. 이것은 L차원의 랜덤벡터 z_k가 최대의 변이를 갖는 1차원의 랜덤변수로 맵핑 된 것이다.

2. METF(Multi-eigenvector temporal filters)

PCA에 따라, covariance matrix Σ_{z_k} 의 가장 큰 고유치부터 순차적으로 L개 $\lambda_{i,k}, i=1,2,...L$ 에 해당하는 L개의 고유벡터들을 $\phi_{i,k}, i=1,2,...L$ 라 하면, $y_{i,k}, i=1,2,...L$ 는 $\phi_{i,k}$상에 랜덤벡터 z_k를 프로젝션 한 것을 나타내는 랜덤변수들이 된다. SETF는 가장 큰 고유치에 해당하는 고유벡터 하나만을 필터계수로 사용하였는데, 이것은 z_k의 가장 중요한 1차원의 표현으로 생각할 수 있다. 그러나, 여전히 다른 고유벡터들도 음성인식의 성능향상에 도움을 줄 수 있는 정보를 가지고 있다고 볼 수도 있다. 따라서, 이러한 관점에서 새로운 multi-eigenvector temporal filters를 식 (5)와 같이 고유치에 의한 가중치를 포함하여 정의한다.

$$w_k=\frac{\overline{w_k}}{|\overline{w_k}|}=\frac{1}{\sqrt{\sum_{i=1}^{M}\lambda_{i,k}^2}}\overline{w_k} \qquad (5)$$
$$\overline{w_k}=\sum_{i=1}^{M}\lambda_{i,k}\phi_{i,k}$$

여기에서, w_k는 time trajectory k에 대한 새로운 L차의 filter 계수이고, 합은 크기 순으로 M개 $(1<M\le L)$의 고유치에 해당하는 M개의 고유벡터들에 대해 수행된다. 따라서, temporal filters에 대한 최종출력은 식(6)과 같이 새로운 랜덤변수의 샘플들로 표현된다.

$$v_k=w_k^T z_k \qquad (6)$$

3. PETF(Proposed eigenvector temporal filters)

METF는 M개의 고유벡터들에 해당고유치의 가중을 취한 것인데 반해, PETF는 가장 큰 M개의 고유치에 해당하는 고유벡터 M개 각각을 filter로 사용한다.

4. Temporal filter의 주파수특성

본 논문에서는 데이터 기반의 temporal filter를 구하기 위해, SITEC의 4연숫자 전화음성 DB중 58292개의 훈련 DB에 PCA를 적용하였다. 훈련 DB의 특징파라미터는 MFCC에 MRTCN된 13차가 사용되었고, temporal filter의 길이 L을 7로, 고유벡터의 갯수 M을 3으로 설정하였다. 그림 2는 특징파라미터 13차에 대해 훈련 DB에서 구한 각 temporal filter의 주파수 응답을 모든 차수에 대해 나타낸 것이다.

SETF와 METF는 모두 LPF(Low Pass Filter)의 특성을 나타내고, PETF는 첫 번째 filter는 LPF를 나타내지만, 두 번째 및 세 번째 filter는 BPF의 특성을 나타낸다.

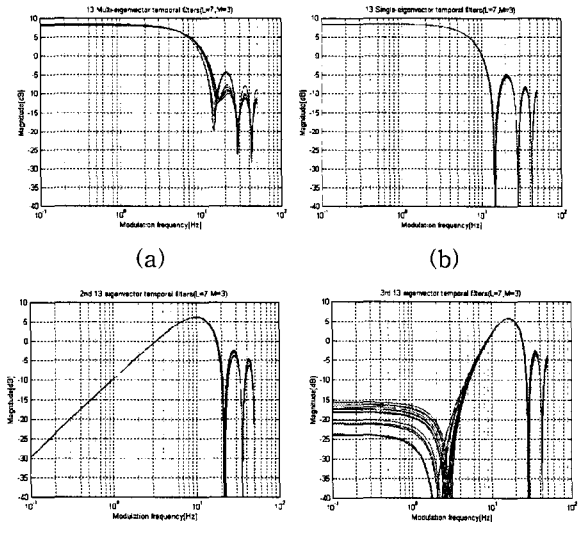

그림 2. Temporal filter의 주파수 응답 : (a)MFTE (b) SETF (c) PETF(M=2번째) (d) PETF(M=3번째)

III. 인식실험 및 결과

1. SITEC 4연숫자 전화음성 DB

음성정보기술산업지원센터(SITEC)에서 제작된 한국어 4연숫자음 전화음성 DB는 총 2000명 화자의 음성으로 이루어져 있으며, 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 58292개의 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.

2. 인식실험 및 결과

4연숫자 전화음성 인식기는 HTK(Hidden Markov Tool Kit)를 사용하여 구현하였다[8]. 음성신호는 20ms의 분석 구간에 10ms 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.

인식실험에 사용된 특징 파라미터는 temporal filters의 종류에 따라 SETF, METF, PETF의 3가지로 나뉜다. 각 temporal filters의 특징파라미터 구성은 다음과 같다. SETF와 METF는 temporal filters에 의해 필터링된 특징파라미터 13차와 이것의 차분 13차 및 차차분의 13차의 총 39차이고, PETF는 3개의 각 filters를 통해 필터링된 특징파라미터 13차씩의 총 39차로 이루어져 있다.

표 1은 특징파라미터에 따른 인식실험의 결과를 나타낸 것이다. temporal filter를 사용했을 때의 실험결과와 비교하기 위해, MFCC에 MRTCN를 적용한 인식결과를 함께 나타내었다. temporal filter를 적용한 실험들 중에서는 METF보다 SETF가 0.14%의 인식율 증가를 나타내었고, SETF보다 PETF가 0.51%의 인식율 증가를 나타내어 가장 높은 인식성능을 보였다. 그러나, PETF는 temporal filter를 적용하지 않은 MFCC+MRTCN보다 1.05%의 인식율 감소를 보였고, SETF와 METF는 각각 1.56%, 1.7%의 인식율 감소를 나타내었는데, 이는 [1]에서 제안한 temporal filter를

적용하여 구한 특징파라미터가 한국어 4연숫자 전화음성 인식 성능의 향상에 크게 기여하지 못함을 나타낸다.

표 1. 특징파라미터에 따른 인식결과
(4연속숫자열 인식률/개별숫자 인식률)

특징파라미터	인식률(%)
SETF	88.51 / 96.49
METF	88.37 / 96.44
PETF	89.02 / 96.72
MFCC+MRTCN	90.07 / 97.16

Ⅳ. 결론

본 논문에서는, 한국어 4연숫자 전화음성의 인식성능 향상을 위한 특징파라미터 연구를 위해, PCA를 적용하여 훈련 DB의 특징파라미터들로부터 데이터 기반의 temporal filter를 구한 후, 이를 사용하여 인식성능이 향상된 특징파라미터를 구하는 [1]의 논문을 검토하였다. 이를 위해 SITEC의 4연숫자 전화음성 DB에 [1]에서 제안한 방법을 적용한 인식실험을 수행하여, 훈련데이터 기반의 temporal filter에 의한 한국어 4연숫자 전화음성의 인식성능을 확인해 보았다.

인식실험결과, 3가지 temporal filter를 적용한 특징파라미터들이 기존의 특징파라미터보다 1%에서 1.7%의 인식율 감소를 나타내었으며, 따라서, [1]에서 제안한 방법이 한국어 4연숫자 전화음성 인식 성능의 향상에 별로 효과적이지 못함을 확인하였다.

참고문헌

[1] N.W.Wang,J.W.Hung,"Data-driven temporal filters based on multi-eigenvectors for robust features in speech recognition," *ICASSP,* 2003

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

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

[4] H. Hermansky, N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing 2,* 1994

[5] J.W.Hung, "Comparative analysis for data-driven temporal filters obtained via principal component analysis and linear discriminant analysis in speech recognition ," *Eurospeech,* 2001

[6] J.W.Hung, L.S.Lee,"Data-driven temporal filters obtained via different optimization criteria evaluated on AURORA2 database," *ICSLP,* 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