훈련데이터 기반의 Temporal Filter를 적용한 4연숫자 전화음성 인식

정성윤(경북대), 배건성(경북대)

<차 례>

<Abstract>

Recognition of Korean Connected Digit Telephone Speech Using the Training Data Based Temporal Filter

Sungyun Jung, Keunsung Bae

The performance of a speech recognition system is generally degraded in telephone environment because of distortions caused by background noise and various channel characteristics. In this paper, data-driven temporal filters are investigated to improve the performance of a specific recognition task such as telephone speech. Three different temporal filtering methods are presented with recognition results for Korean connected-digit telephone speech. Filter coefficients are derived from the cepstral domain feature vectors using the principal component analysis. According to experimental results, the proposed temporal filtering method has shown slightly better performance than the previous ones.

* Keywords: Temporal filter, Principal component analysis, Telephone speech recognition

1. 서 론

유/무선 전화망 환경에서의 음성인식은 호가 형성될 때마다 변화하는 채널에 의한 왜곡과 통화 시 발생하는 불특정한 주변 잡음에 의해 인식성능이 크게 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 특징파라미터를 추출하는 연구가 이어져왔다[1,2,3]. 이러한 기법들은 채널왜곡이나 잡음을 제거하기 위해 음성의 특징파라미터 시계열에 HPF(High Pass Filter) 나 BPF(Band Pass Filter) 등의 필터링을 수행하여, 음성특징의 시계열 상에서 음성에 비하여 매우 느리게 변화하는 채널왜곡이나 빠르게 변화하는 잡음성분들을 감소시켜 전체적인 인식률을 증가시키고자 하는 기법이다. 또한 이들은 인식태스크에 독립적으로 수행되기 때문에 강인한 음성 특징파라미터 추출을 위해 일반적으로 사용되는 기법들이다.

만약 특정 인식태스크의 시계열 특성을 고려해 줄 수 있는 필터를 설계할 수 있다면, 채널왜곡 및 주변잡음에 대해 좀 더 효율적인 보상이 가능하게 된다. 적용할 인식태스크에 최적인 필터계수를 구하기 위해, PCA(Principal Component Analysis), LDA(Linear Discriminant Analysis) 그리고 MCE(Minimum Classification Error)등과 같은 데이터 기반의 접근방법들이 연구되고 있다[4,5,6].

본 논문에서는 한국어 4연숫자 전화음성의 인식성능 개선을 위해 훈련데이터에 기반한 temporal filter 적용기법을 제안하고 SITEC(Speech Information Technology & Industry Promotion Center)의 4연숫자 전화음성 DB를 사용하여 그 성능을 평가하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 훈련데이터 기반의 temporal filter들에 대해 기술한다. 그리고, 3장에서 기존의 특징파라미터 및 훈련데이터 기반의 temporal filter에 의한 특징파라미터에 대한 실험결과를 검토한 후, 4장에서 결론을 맺는다.

2. PCA를 적용한 temporal filter

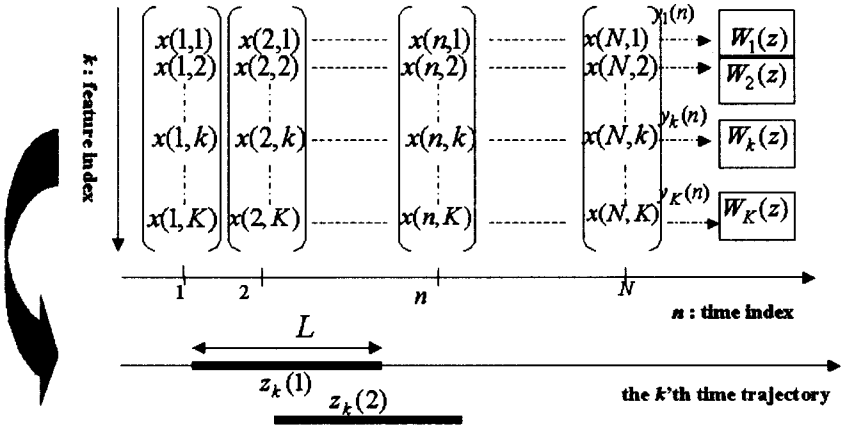
K 차원의 특징벡터 $x(n)$이 <그림 1>과 같이 프레임별로 시간축에 따라 순차적으로 나열되어 있다면, $x(n)$은 식 (1)과 같이 표현될 수 있고, $x(n)$의 k번째 시계열은 $[x(1,k), x(2,k), \ldots\ldots\ldots\ldots, x(N,k)]$로 나타낼 수 있으며, 이 때 $y_k(n) = x(n,k)$로 정의한다.

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

여기에서 N은 음성신호의 전체 프레임수이고, K는 특징벡터의 차수이다.

데이터기반의 temporal filter는 k번째 시계열 $y_k(n)$을 필터링하는 L샘플 FIR(Finite Impulse Response) 필터 $W_k(z)$이다. 이를 구하기 위해 먼저 k번째 특징파라미터의 시계열에 대해, L개의 특징파라미터를 취하여 식 (2)와 같이 L차원의 벡터 $z_k(n)$을 구한다. $z_k(n)$에 고유치 문제를 적용하여 고유벡터를 구한 후 이를 temporal filter로 사용한다.

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



<그림 1> 특징파라미터의 시계열 표현

인식실험에 사용된 데이터 기반의 temproral filter는 PCA에 의한 고유벡터의 적용방법에 따라 세 가지로 구분한다. 기존의 훈련데이터 기반의 temporal filter인 SETF(Single-eigenvector temporal filters)[4]와 METF(Multi-eigenvector temporal filters)[6], 그리고 본 논문에서 제안한 SVTF(Selective temporal filter)이다.

2.1 SETF(Single-eigenvector temporal filter)

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

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

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

식 (4)의 공분산행렬에 식 (5)와 같이 고유치 문제를 적용하여 가장 큰 고유치에 해당하는 고유벡터를 SETF의 필터계수로 선택한다.

$$\Sigma_{z_k}\Phi_k = \Lambda_k\Phi_k \tag{5}$$

여기에서 Λ_k와 Φ_k는 Σ_{z_k}의 고유치와 고유벡터들이다.

2.2 METF(Multi-eigenvector temporal filter)

PCA에 따라, 공분산행렬 Σ_{z_k}의 고유치 크기 순으로 L개, $\lambda_{i,k}$, $i=1,2,\ldots,L$ 에 해당하는 L개의 고유벡터들을 $\Phi_{i,k}$, $j=1,2,\ldots,L$ 라 정의할 수 있다. SETF는 가장 큰 고유치에 해당하는 고유벡터 하나만을 필터계수로 사용하였는데, 이것은 z_k의 가장 중요한 1차원의 표현으로 간주할 수 있다. 그러나 여전히 다른 고유벡터들도 음성인식의 성능향상에 도움을 줄 수 있는 정보를 가지고 있다고 볼 수 있다. 따라서 이러한 관점에서 METF를 식 (6)과 같이 고유치에 의한 가중치를 포함하여 정의한다.

$$w_k = \frac{\overline{w_k}}{|\overline{w_k}|} = \frac{1}{\sqrt{\sum_{i=1}^{M}\lambda_{i,k}^2}}\overline{w_k} \tag{6}$$

$$\overline{w_k} = \sum_{i=1}^{M}\lambda_{i,k}\Phi_{i,k}$$

여기에서, w_k는 차수 k의 특징파라미터 시계열에 대한 새로운 L차의 필터계수이고, 합은 크기 순으로 M개 $(1<M\leq L)$의 고유치에 해당하는 고유벡터들에 대

해 수행된다.

2.3 SVTF(Selective temporal filter)

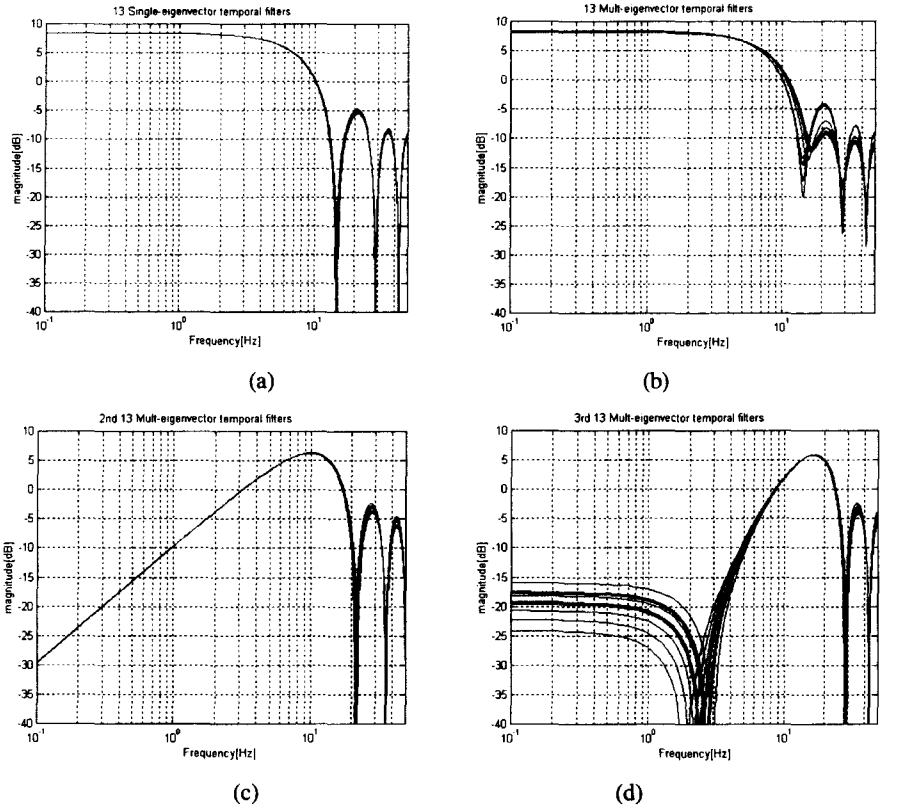
Kanedera[7]에 따르면, 깨끗한 환경에서는 대부분의 유용한 음성정보가 변조주파수 1 Hz에서 16 Hz 사이의 대역에 포함되어 있다. 그리고 잡음환경에서는 변조주파수 2 Hz 이하 그리고 20 Hz 이상의 성분들이 음성인지도에서 덜 중요하다도 알려져 있다. 일반적으로 중요한 변조주파수 영역은 특정 태스크 환경에 따라 결정되기 때문에 서로 다른 고유벡터들의 주파수 특성을 조사할 필요가 있다. 따라서 본 논문에서는 한국어 연속숫자음 전화음성 인식이라는 특정 태스크에서 여러 고유벡터들의 주파수 특성 및 인식성능을 분석한다.

기존의 SVTF나 METF는 하나의 고유벡터를 사용한 것인데 반해, 본 논문에서는 M개의 고유치에 해당하는 고유벡터들을 filter로 사용한다. 이러한 temporal filter를 SVTF라 명명한다. SVTF는 선택적인 temporal filter의 적용에 따라 SVTF01과 SVTF02로 구분한다. SVT01은 M개의 고유벡터를 사용하고 SVTF02는 $M-1$개의 고유벡터를 사용한다. 본 논문에서는 M을 3으로 설정하였기 때문에 SVTF01의 경우 가장 고유치가 큰 고유벡터로부터 순서대로 3개를 선택하였다.

2.4 Temporal filter의 주파수특성

본 논문에서는 데이터 기반의 temporal filter를 구하기 위해, SITEC의 4연숫자 전화음성 DB중 58,292개의 훈련 DB에 PCA를 적용하였다. 훈련 DB의 특징파라미터는 DWFBA(Direct Weighted Filter Bank Analysis)기반의 MFCC 13차를 사용하였고[8], temporal filter의 길이 L을 7로, 고유벡터의 갯수 M을 3으로 설정하였다. <그림 2>는 특징파라미터 13차에 대해 훈련 DB에서 구한 각 temporal filter의 주파수 응답을 13차의 모든 차수에 대해 누적하여 나타낸 것이다.

<그림 2>에서 가로축은 변조주파수로써, 100Hz 프레임율에 대해 50Hz 까지의 성분을 로그스케일로 나타낸 것이고, 세로축은 크기를 dB로 표시한 것이다. METF와 SETF는 차단주파수가 7Hz 부근인 LPF(Low Pass Filter)의 특성을 나타내고, SVTF의 첫 번째 filter는 SETF와 동일하고, 두 번째와 세번째 filter는 각각 4~17Hz, 10~24Hz의 통과대역을 갖는 BPF의 특성을 나타낸다.



<그림 2> Temporal filter의 주파수 응답 : (a) MFTE
(b) SETF, SVTF(M=1번째), (c) SVTF(M=2번째), (d) SVTF(M=3번째)

3. 실험 및 결과

3.1 SITEC 4연숫자 전화음성 DB

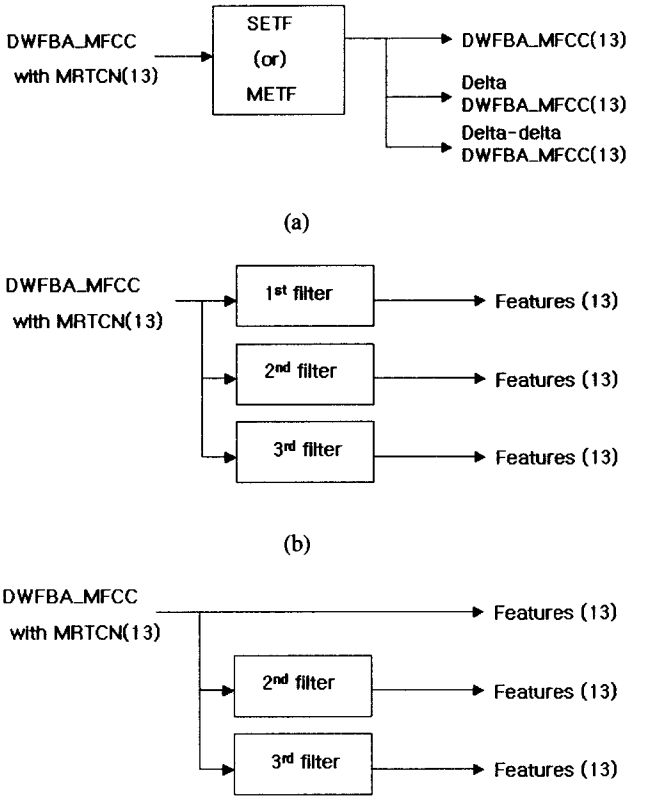
음성정보기술산업지원센터(SITEC)에서 제작된 한국어 4연숫자음 전화음성 DB는 2,000명 화자의 음성으로 이루어져 있으며, 유선전화, 무선전화, cellular, PCS 전화음성이 모두 포함되어 있다[9]. 녹음 환경은 연구실과 사무실, 가정집 환경으로 이루어져 있고, 모든 전화음성은 8kHz 샘플링에 16bits/sample linear PCM으로 저장되어 있다. SITEC 전화음성 DB는 훈련 데이터로 1,800명 화자의 58,292개, 테스트 데이터로 200명 화자의 6,468개의 4연숫자 전화음성으로 구성되어 있으며 모두 1,620 종류의 4연숫자음으로 이루어져 있다. 그리고, "륙"과 "육"은 서로 다른 단어로 구분되어 레이블링 되어있다.

3.2 인식실험 및 결과

4연숫자 전화음성 인식기는 HTK(Hidden Markov Tool Kit)를 사용하여 구현하였다[10]. 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 특징파라미터는 기본 특징파라미터로 DWFBA(Direct Weighted Filter Bank Analysis) 39차(DWFBA_MFCC), 그리고 DWFBA에 temporal filter를 적용한 특징파라미터이다. 모든 특징파라미터에는 기본 특징파라미터 인식실험에서 가장 인식률이 높았던 보상기법인 MRTCN(Modified Real Time Cepstral Normalization)이 적용되었다. 인식실험에 적용된 temporal filter의 종류는 SETF, METF, SVTF01, SVTF02의 네 가지이다. 각 temporal filter의 특징파라미터 구성은 <그림 3>과 같다.

<표 1>은 특징파라미터 종류에 따른 인식실험의 결과를 나타낸 것이다. Temporal filter를 사용했을 때의 실험결과와 비교하기 위해, DWFBA와 여기에 MRTCN를 적용한 인식결과를 함께 나타내었다. Temporal filter를 적용한 실험들 중에서는 METF보다 SETF가 0.14%의 4연숫자음 인식률 증가를 나타내었고, SETF보다 SVTF01이 0.34%의 인식율 증가를 나타내었다. 그리고, SVTF02는 SVTF01보다 0.52%의 인식률 증가를 나타내어 temporal filter들 중에서 가장 높은 인식성능을 보였다. 이것은 temporal filter를 적용하지 않은 DWFBA_MFCC+MRTCN 보다 0.03%의 근소한 4연숫자 인식률 증가치에 해당된다. 이러한 결과는 기존의 temporal filter인 SETF 나 METF를 적용하여 구한 특징파라미터가 한국어 4연숫자 전화음성의 인식성능 향상에 크게 기여하지 못함을 나타낸다. 그러나 본 논문에서 제안한 SVTF02는 기존에 SITEC DB에서 가장 좋은 인식성능을 나타낸 DWFBA_MFCC+MRTCN 기법[2]과 비슷하거나 성능이 약간 개선된 결과를 나타내었다. 따라서 4연숫자 전화음성인식 태스크에서는 temporal filter 선택에 따라 인식성능에 차이가 있을 뿐만 아니라, 인식 태스크에 적합한 temporal filter 조합이 중요함을 알 수 있다.



<그림 3> Temporal filter 사용에 따른 특징파라미터 추출
(a) SETF, MFTE (b) SVTF01 (c) SVTF02

<표 1> 특징파라미터 종류에 따른 인식결과

특징파라미터 종류	인식률(%)	
	4연숫자 인식률	개별숫자 인식률
DWFBA_MFCC	88.64	96.72
DWFBA_MFCC with MRTCN	91.48	97.55
SETF with MRTCN	90.65	97.27
METF with MRTCN	90.48	97.22
SVTF01 with MRTCN	90.99	97.43
SVTF02 with MRTCN	91.51	97.63

4. 결론

본 논문에서는 한국어 4연숫자 전화음성인식과 같은 특정 태스크에서의 성능 향상을 위한 특징파라미터 추출기법 연구를 위해, 훈련 DB의 특징파라미터의 시계열 집합에 PCA를 적용한 데이터 기반의 temporal filter 적용방법을 검토하고, 새로운 temporal filter 적용방법을 제안하였다. 비록, 기존의 temporal filtering 방법이 잡음환경에서 중요한 인식성능 향상을 가져왔다 하더라도 전화망 환경에서는 효과적인 성능향상을 볼 수 없었다. 그러나 실험결과, 본 논문에서 제안한 SVTF02가 기존의 훈련데이터 기반의 temporal filter인 SETF, METF 보다 약간의 인식성능 향상을 확인할 수 있었다.

참고문헌

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

[2] 최종연구보고서, 전화망 환경에서의 연속숫자음 신호왜곡 연구, 전자통신연구원, 2002.

[3] H. Hermansky, N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578-589, 1994.

[4] J. W. Hung, "Comparative analysis for data-driven temporal filters obtained via principal component analysis and linear discriminant analysis in speech recognition", *in Proc. of EUROSPEECH*, pp. 1959-1962, 2001.

[5] J. W. Hung, L. S. Lee, "Data-driven Temporal Filters Obtained via Different Optimization Criteria Evaluated on Aurora 2 Database", *in Proc. of ICSLP*, vol. 1, pp. 441-444, 2002.

[6] N. W. Wang, J. W. Hung, "Data-driven temporal filters based on multi-eigenvectors for robust features in speech recognition", *in Proc. of ICASSP*, vol. 1, pp. 400-403, 2003.

[7] N. Kanedera, H. Hermansky, T. Arai, "On properties of modulation spectrum for robust automatic speech recognition", *in Proc. of ICASSP*, vol. 2, pp.613-616, 1998.

[8] 정성윤, 김민성 외, "한국어 연속숫자음 전화음성의 인식성능 개선", *대한전자공학회 추계학술발표대회, 제 25권 제 2호*, pp. 582-585, 2002.

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

[10] S. Young, G. Evermann, D. Kershaw, *The HTK Book (HTK Version 3.1)*, Cambridge University Engineering Department, 2000.

접수일자 : 2005년 2월 10일
게재결정 : 2005년 3월 15일

▶ 정성윤(Sung-Yun Jung)
주소 : 대구광역시 북구 산격동 1370번지 경북대학교
소속 : 경북대학교 공과대학 전자공학과
전화 : 053) 940 - 8627
FAX : 053) 950 - 5505
E-mail : yunij@mir.knu.ac.kr

▶ 배건성(Keun-Sung Bae)
주소 : 대구광역시 북구 산격동 1370번지 경북대학교
소속 : 경북대학교 공과대학 전자공학과
전화 : 053) 950 - 5527
FAX : 053) 950 - 5505
E-mail : ksbae@ee.knu.ac.kr