음성인식을 위한 웨이블릿 필터 평가

김 기 대, 이 철 희
연세대학교 전기컴퓨터공학과
전화 : (02) 361-2779 / 팩스 : (02) 312-4584

Wavelet Filter Evaluation for Speech Recognition System

Kidae Kim, Chulhee Lee
Dept. Electrical and Computer Engineering, Yonsei University,
134 Shinchon-Dong, Seodaemoon-Gu, Seoul 120-749, KOREA,
E-mail: chulhee@yonsei.ac.kr

Abstract

In this paper, we explore the possibility to use wavelet decomposition based on modified octave structured 5-level filter banks as a set of features for speech recognition. The HMM (Hidden Markov Model) is used as a recognizer [1]. We compared the performance of the wavelet decomposition with the mel-cepstrum and LPC cepstrum. Experimental results show favorable results.

I. 서론

웨이블릿 변환은 다중해상도(multiresolution) 표현 및 이산신호의 부대역 분해방법 등에 대한 단일화된 이론을 제공하고 있으며 최근 신호처리 전반에 널리 사용되는 변환방법이다. 특히 음성과 영상 압축 분야에서 성공적으로 사용되고 있는데, 그 중 영상압축 분야에서 부대역 부호화(subband coding)방법은 높은 성능을 보여주고 있다 [2]. 최근에는 웨이블릿 분해(decomposition) 방법이 음성인식 시스템에 적용되고 있다 [3][4].

웨이블릿 변환은 음성인식에 사용되는 특징파라미터를 구할 때 필요한 국부적 정보를 제공한다. 그러나 웨이블릿 변환에 사용되는 필터의 성질과 필터뱅크의 구조에 따라 특징파라미터의 성능이 달라진다.

본 논문에서는 음성인식 시스템에 적합한 웨이블릿 필터들을 평가하기 위해, 먼저 대칭형 옥타브밴드 구조의 필터뱅크를 설계하여 저차원 음성특징파라미터를 구하고, 여러 가지 Daubechies 기저(basis)들을 적용하여 기존의 LPC, 멜켑스트럼의 성능과 비교하여 분석하였다.

II. 웨이블릿 분해(decomposition)

웨이블릿 분해(decomposition)는 그림 1.a와 같이 2채널 이산 웨이블릿 변환을 기본으로 한다. 신호 f(n)은 분석(analysis) low-pass 필터(scaling 함수) $h_0(n)$와 분석 high-pass 필터(wavelet 함수) $h_1(n)$에 입력된다. 그리고 출력을 다운샘플링(downsampling)하여 1레벨 분해(decomposition)의 기준(reference)신호 $r_1(n)$과 상세(detail)신호 $d_1(n)$을 구한다. 만일 scaling 함수와 wavelet 함수를 각각 $\Phi(t)$, $\Psi(t)$로 정의하면 $r_1(n)$과 $d_1(n)$은 다음 식 (1)과 같이 나타낼 수 있다.

$$r_1(n) = (\downarrow 2)\langle f(t), \Phi(t-n)\rangle$$
$$d_1(n) = (\downarrow 2)\langle f(t), \Psi(t-n)\rangle \quad (1)$$

여기서 $(\downarrow 2)$는 다운샘플링 연산자이고 $\langle\ \rangle$ 는 내적 연산자이다.

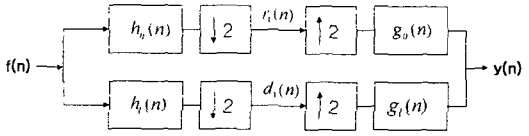
다단계 분해(multilevel decomposition)를 하기 위해서는, 기준(reference)신호를 다음 분석단계(analysis stage)에 반복적으로 입력한다. 다해상도(multiresolution) 분석은 식 (2)와 같이 구한다.

$$\Phi_{j,k}(t) = \sum_n h[n-2k]\Phi_{j-1,n}(t)$$
$$\Psi_{j,k}(t) = \sum_n g[n-2k]\Phi_{j-1,n}(t) \quad (2)$$

여기서 $\Phi_{j,k}(t) = 2^{-j/2}\Phi(2^{-j}t-k)$,
$\Psi_{j,k}(t) = 2^{-j/2}\Psi(2^{-j}t-k)$ 이다.

일반적으로 멀티레벨 트리구조는 2채널 웨이블릿 변환을 기본으로 하여 구현할 수 있으며 그 구성은 그림 1.b와 같다.

(a)



(b)

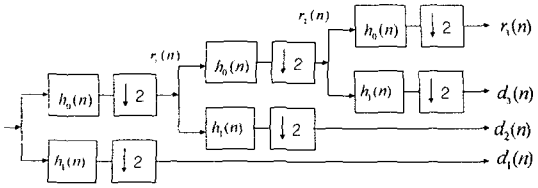


그림 1. (a) 웨이블릿 변환의 기본 필터뱅크.
(b) 3 레벨 웨이블릿 변환을 위한 트리구조.

III. 대칭형 옥타브 필터뱅크

필터뱅크의 구조는 시간과 주파수의 해상도를 결정한다. 전체 트리구조(tree-structured) 필터뱅크는 short-time 푸리에 변환과 비슷하게 스펙트럼을 선형적으로 분할하여 균등한 해상도를 제공한다. 그리고 옥타브밴드(octave-band tree structured) 필터뱅크는 앞서 설명한 바와 같이 2 채널 필터뱅크를 연결하여 구현할 수 있다. 이는 고주파수 영역에서 정밀한 시간 해상도를 제공하고 저주파수 영역에서 정밀한 주파수 해상도를 제공한다(그림 2.a). 그러나 옥타브밴드 트리구조는 고주파수 영역의 중요한 정보를 잃기 때문에 음성인식 시스템에 적합하지 않다. 이를 극복하기 위해 high-pass 필터(wavelet function)부분도 분해하며 수행하게 구성하면, 옥타브밴드 트리구조 보다 성능이 좋은 음성 특징파라미터를 제공하면서 전체 트리구조 보다 적은 데이터를 가지는 대칭형 옥타브 밴드 필터뱅크(그림 2.b)를 구할 수 있다.

IV. 특징 추출

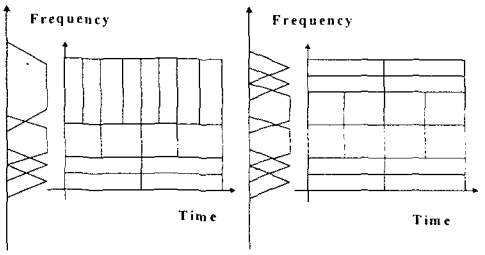


그림 2. 시간-주파수의 해상도. (a) 옥타브 밴드
(b) 대칭형 옥타브 밴드.

HMM(Hidden Markov Model)의 입력으로 사용할 특징 파라미터를 구하는 과정은 다음과 같다. 구간 [0,T]에서 샘플링된 음성신호 $S(t)$를 프레임 간격으로 분할하면 다수의 단위신호들을 구할 수 있다.

$$S(t) \rightarrow [S_0(t), S_1(t), \ldots, S_{N-1}(t)]. \quad (3)$$

이때 각 신호 $S_i(t)$는 앞서 설명한 대칭형 옥타브밴드(symmetric octave-band) 구조의 필터뱅크에 의해 웨이블릿 영역의 새로운 신호들로 분해된다. 이 신호들의 스펙트럼정보를 계수화하기 위해 각 서브밴드(subband) 에너지를 다음과 같이 구한다.

$$e_j = \sum_t |S_{i,j}(t)|^2 \quad (4)$$

이때 e_j 는 j번째 서브밴드 에너지를 나타낸다.

각 에너지는 입력신호의 에너지로 정규화 하였다. 위 과정은 그림 3과 같다.

V. 웨이블릿 필터 평가

음성인식에 적합한 특징파라미터를 구하기 위하여 웨이블릿 필터선택은 매우 중요하다. 본 논문에서는 음성인식 시스템에 사용될 수 있는 웨이블릿 필터를 조사하기 위해, 여러 가지 Daubechies필터들을 사용하여 한국어 숫자음 인식실험을 하였다. 11.025kHz로 샘플링한 총 2000개(남자 10명, 여자 10명이 0에서 9까지 각 10번씩 발음)의 데이터중 7명(남자 4명, 여자 3명)의 데이터는 HMM 학습에 사용하였고, 나머지 13명

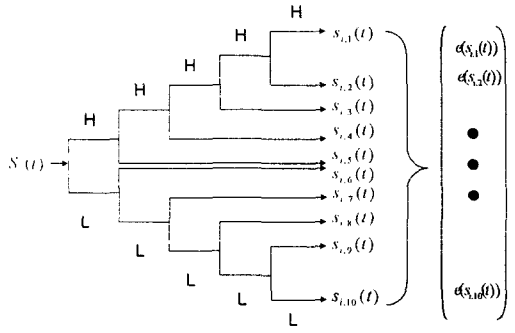


그림 3. 5레벨 대칭형 옥타브밴드 필터뱅크로부터 10차 특징벡터를 구하는 과정.

(남자 6명, 여자 7명)의 데이터는 테스트에 사용하였다. 이때 모든 실험데이터는 23ms 길이의 Hamming Window를 86.9/sec로 이동하며 프레임단위로 분석하였다. Daubechies의 orthogonal basis와 biorthogonal basis로 5레벨 대칭형 옥타브밴드 필터뱅크를 구성하였고, 이로부터 구한 10차원 특징파라미터로 인식평가 실험을 하였다. 제안된 구조로부터 구한 특징파라미터를 기존의 LPC 켑스트럼, 멜켑스트럼과 비교하였다. 또한 제안된 대칭형 옥타브밴드 필터뱅크에 여러 가지 웨이블릿 basis(Daubechies's)를 사용한 결과, 일반적으로 orthogonal basis보다 홀수의 biorthogonal basis가 보다 우수함을 관찰하였다. 특히 영상압축분야에서 높은 압축률을 보이는 Daubechies's 9-7 tab [2]은 기존의 LPC 켑스트럼과 멜켑스트럼 [6]보다도 우수한 성능을 보여주고 있으며, 다른 웨이블릿 필터들 보다 뛰어난 성능을 보여주고 있다. 여러 가지 필터들의 인식결과는 표 1과 같다.

VI. 결론

본 논문에서는 음성인식시스템에 응용될 수 있는 새로운 구조의 필터뱅크를 제안하였다. 그리고 여러 가지 웨이블릿 필터들을 평가하였다.

향후 음성인식시스템에 보다 적합한 필터를 선택하기 위해서는 본 실험에 사용되지 않은 다른 웨이블릿 필터들에 대해서 광범위한 평가가 필요하다. 또한 웨이블릿 필터들로부터 구한 특징파라미터의 성능향상에 대해서도 더 연구가 필요하며 각 서브밴드의 에너지의 성능을 향상시키는 프로세싱을 추가하면 [4][5] 보다 향상된 특징파라미터를 구할 수 있을 것으로 예측된다.

참고문헌

[1] Rabiner L., "Tutorial on Hidden Markov Models." Proceedings of the IEEE Vol. 77 No. 2 pp.11-20, 1988.

[2] Villasenor J.D., Belzer B., Liao J. "Wavelet filter evaluation for image compression." IEEE Transaction on image processing Vol. 4 No. 8

표 1. 13명의 화자에 대한 인식률(%).

	LPC Cepstrum	Mel Cepstrum	D6	D8	D10	D12	D14	D97	D1311	D610
Speaker1	94	100	90	84	95	91	96	95	97	80
Speaker2	83	73	79	76	71	73	73	73	80	77
Speaker3	97	87	80	83	90	84	78	89	92	88
Speaker4	74	80	93	85	91	90	87	93	88	91
Speaker5	90	97	87	85	88	90	91	84	90	84
Speaker6	66	82	74	71	75	76	69	81	73	69
Speaker7	60	63	71	69	72	59	69	74	67	74
Speaker8	92	74	88	79	94	87	83	94	85	89
Speaker9	55	55	60	71	63	67	60	57	60	71
Speaker10	63	92	80	84	80	81	85	83	87	82
Speaker11	88	67	91	82	88	85	90	88	85	62
Speaker12	74	82	67	68	67	73	75	69	79	53
Speaker13	44	52	55	62	58	57	52	61	51	51
Mean	75.38	77.23	78.08	76.85	79.38	77.92	77.54	80.08	79.54	74.69
STD.	16.84	15.12	12.13	7.82	12.51	11.58	12.86	12.44	13.40	13.10

pp. 1053-1060 1995.

[3] Favero. R.F., King R.W. "Wavelet parameterization for speech recognition." ICSPAT Vol. 2 pp. 1444-1449 1993.

[4] Kaisheng Y., Zhigang C. "A wavelet filter optimization algorithm for speech recognition." ICCT Vol. 2 pp.5 1998.

[5] Wesfreid. E., Wickerhauser. M.V. "Adapted local trigonometric transform and speech processing" Technical Report, Washington University, St. Louis, 1992.

[6] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" IEEE Trans. Acoustic, Speech, Signal Processing, ASSP-28(4), pp.11-20 1998.