2001년도 한국음향학회 학술발표대회 논문집 제20권 제2(s)호

웨이블렛 변환을 이용한 음성특징 추출에 관한 연구

정의준*, 장성욱*, 양성일*, 권영헌**
*한양대학교 전자전기제어계측공학부 , **한양대학교 물리학과

A Study on Feature Extraction using Wavelet Transform for Speech Recognition

Eui-jun Joung*, Sung-wook Chang*, Sung-il Yang*, Y. Kwon**
*School of Electrical and Computer Engineering,
**Department of Physics Hanyang University
E-mail: bluerose76@ihanyang.ac.kr

요약

본 논문에서는 기존의 음성인식에서 사용하는 특징벡터인 MFCC(Mel-Frequency Cepstral Cefficients)를 대신하여 웨이블렛 변환을 이용한 새로운 특징벡터를 추출하는 방법을 제안한다. 새 특징벡터로는 MRA (Multi-Resolution Analysis)를 이용하여 구성하였다. 웨이블렛 변환을 이용한 새로운 특징벡터의 추출 목적은 시간축과 주파수축에서의 더 좋은 해상도를 가지는 성질을 이용하는 것이다. 실험결과에서 웨이블렛 변환을 이용한 새로운 특징벡터를 이용한 인식이 기존의 방식보다 더 좋은 인식률을 보이고 있음을 확인하였다.

1. 서론

MFCC(Mel-Frequency Cepstral Coefficients)를 특징벡터로서 이용한 방식은 음성인식에서 가장 보편적으로 사용되고 있는 방법이다. 이는 기존의 특징벡터인 LPC(Linear Prediction Coefficient)나 LFCC(Linear Filter Cepstral Coefficient)와 같은 것보다 더 좋은 인식능력을 갖추고 있다고 알려져 있다 [1].

웨이블렛 변환을 FT(Fourier Transform)를 이용하여 특징벡터를 계산하는 MFCC와 비교해 보면 다음과 같은 이점을 얻을 수 있다. 첫째, 웨이블렛 변환은 신호의 저주파 부분에서 좋은 주파수 분해능을 얻을 수 있다. 둘째, 웨이블렛 변환은 고주파 부분에서 좋은 시간 분해능을 얻을 수 있다 [2]. 이러한 장점으로 인하여 웨이블렛 변환을 이용한 특징벡터를 사용하게 되면 파열음이나 파찰음에서와 같이 시간/주파수 상에서 갑자기 튀는 국부적 특성을 잘 반영 할 수 있기 때문에 좀 더 나은 인식능력을 보일 수 있다.

본 논문의 음소모델은 한국어 음소로서 초성, 중성, 종성, 묵음, 잡음 등으로 지정한 50개의 음소모델을 가지고 인식을 하였다. 특징벡터로는 MRA (Multi-Resolution Analysis)를 이용하여 필터 에너지와 음성 프레임의 에너지를 구하여 9차의 특징벡터를 추출하고 이에 대한 1차, 2차 도함수를 계산하여 총 27차의 벡터를 이용하였다.

본 논문에선 MFCC 특징벡터를 이용한 인식 실험과 웨이블렛 변환을 이용한 특징벡터를 가지고 인식 실험

을 한 것을 비교해 보았다.

본 논문에선 웨이블렛 변환에 대해 서술하고, MFCC 특징벡터와 웨이블렛 변환을 이용한 특징벡터 추출에 대해 다루며, 특징벡터 추출에 대한 실험과정에 대해 기술한다. 마지막으로 실험결과 및 결론을 언급한다.

2. 웨이블렛 변환

웨이블렛 분석은 신호들을 분석하는데 있어 매우 능률적이다. 왜냐하면 웨이블렛은 퓨리에 변환에 비해 신호특성들의 부분적 해석과 분리를 더 정확히 할 수 있기 때문이다.

웨이블렛 변환은 퓨리에 변환과 같이 기저함수들의 집합에 의한 신호의 분해로서 이해할 수 있다. 이때 웨이블렛 변환에서 하나의 기저함수를 웨이블렛이라 부르며 웨이블렛은 하나의 대역통과 필터(band-pass filter)라고 할 수 있다. 웨이블렛 변환은 모 웨이블렛(mother wavelet)이라고 하는 원형 웨이블렛(prototype wavelet)을 정의한 후, 이를 시간축으로 이동(translation)시키고, 스케일링(scaling)하여 다양한 웨이블렛 기저들을 생성하여 신호를 분석한다.

다해상도 신호 해석에서 주어진 함수는 다른 해상도를 가진 연속적인 추정치들의 한계로써 표현된다. 이러한 smoothing은 스케일링 함수 $\phi(t)$라고 불리는 저주파 통과 커널과 convolution을 함으로써 이뤄진다.

다해상도 해석은 다음의 성질을 가지는 $L^2(R)$ 공간의 닫힌 부공간 $\{ V_m | m \in Z \}$의 열로서 구성된다 [3].

1 .부공간의 구성

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots \quad (1)$$

2. Compeleteness

$$\bigcap_{m \in Z} V_m = \{0\} \qquad \bigcup_{m \in Z} = L^2(R) \quad (2)$$

3. 스케일링 성질

어떤 함수 $f \in L^2(R)$에 대해

$$f(x) \in V_m \Leftrightarrow f(2x) \in V_{m-1} \quad (3)$$

4. 기저에 관한 성질

$\phi(t) \in V_0$인 스케일링 함수가 존재한다.

$$[\ \phi_{mn}(t) = 2^{-m/2} \phi(2^{-m} t - n)] \quad (4)$$

여기서 V_{m-1}에 속하는 V_m의 orthogonal complement인 W_m을 정의하며, 가능한 무한 공간인 W_m의 합은 $L^2(R)$에 스팬(span)한다.

$$V_{m-1} = V_m \oplus W_m \quad (5)$$

$$V_m \perp W_m \quad (6)$$

$\phi(t)$는 $\phi(2t)$의 이동시킨 것들의 선형 조합으로 나타낼 수 있다.

$$\phi(t) = 2 \sum_n h_0 \phi(2t - n) \quad (7)$$

그리고 대역통과 웨이블렛 함수도 아래와 같이 표현될 수 있다.

$$\phi(t) = 2 \sum_n h_1(n) \phi(2t - n) \quad (8)$$

실제적으로 웨이블렛 변환은 연속 웨이블렛 변환(Continuous Wavelet Transform)을 사용하지 않고 이산 웨이블렛 변환(Discrete Wavelet Transform)을 사용한다.

음성신호를 파라미터화 하기 위해서 Mallet Algorithm을 이용하여 신호를 분석해 나간다 [4]. 원래의 샘플 시퀀스를 x(k), 각 필터 뱅크의 임펄스 응답을 각각 $h_0(k)$ (low-pass filter), $h_1(k)$ (high-pass filter) 라고 하면 다음과 같은 방식에 의해 신호를 분석할 수 있다.

Level 0: $x_k^{(0)} = x(k)$

Level 1: $x_k^{(1)} = \sum_n h_{0(n-2k)}\, x_n^{(0)}(t)$

$$d_k^{(1)} = \sum_n h_{1(n-2k)}\, x_n^{(0)}(t)$$

.

.

.

Level L: $x_k^{(L)} = \sum_n h_{0(n-2k)}\, x_n^{(L-1)}(t)$

$$d_k^{(L)} = \sum_n h_{1(n-2k)}\, x_n^{(L-1)}(t) \quad (9)$$

와 같이 임펄스 응답과 신호의 convolution과

decimation의 과정을 거치면서 신호를 분석해 나간다.

Mallet Algorithm에 의해 전개된 다해상도 웨이블렛 변환 구조는 dyadic 계층구조로 이루어져있다. <그림 1>은 이러한 구조를 보여주는 블록 다이어그램이다.

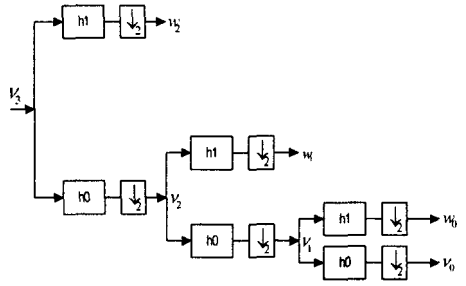


그림 1. Three-Stage Two-Band Analysis Tree

3. MFCC 특징벡터 및 웨이블렛 변환을 이용한 특징벡터

3.1 MFCC 특징벡터 추출

MFCC 특징벡터 추출은 <그림 2>에서와 같은 방식으로 이루어진다. 입력으로 얻어진 음성 데이터로부터 전처리로 Preemphasis와 Windowing을 한 후에 DFT (Discrete Fourier Transform)을 취한 후 Mel filter banks filtering을 한 값들을 에너지의 로그값을 취한 후 다시 IDFT를 취해서 얻는다.

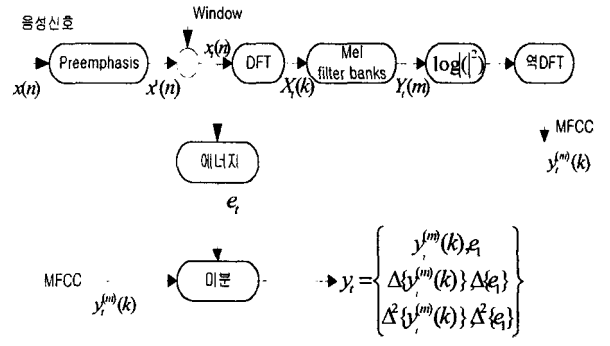


그림 2. MFCC 연산의 개략도

실험에서 사용한 특징 파라미터는 MFCC(12차)+음성 프레임 에너지(1차), 1차, 2차의 도함수 값을 포함하여 모두 39(13*3)차를 사용하였다. MFCC의 도함수 값들은 음성의 시변 특성을 보다 잘 반영하기 위해 사용되는 값들이며 그것들은 다음과 같은 식으로 나타낼 수 있다.

$$\Delta^i\{u_t\} = \Delta^{i-1}\{u_{t+1}\} - \Delta^{i-1}\{u_{t-1}\}$$
$$\Delta^0\{u_t\} = u_t \qquad (10)$$

3.2 웨이블렛 변환을 이용한 특징벡터 추출

위에서 언급하였듯이 웨이블렛 변환을 이용한 특징벡터는 <그림 2>에서와 같은 tree 구조로 신호를 low pass fiter와 high pass filter를 이용하여 신호를 분석하고 decimation을 취한 후 다시 low pass filtering한 값을 high pass filter와 low pass filter를 이용하여 신호를 분석해 나간다. 이러한 Mallet tree를 이용하여 신호의 다해상도 해석을 통한 분석을 한 후 그 계수 벡터의 에너지값의 로그값을 구한다. 다음으로 DCT를 취해서 특징벡터로 구성을 한다. 또한 음성 프레임의 전체 에너지 값을 구한 후 로그값을 취한 값을 특징벡터에 포함시킨다.

Mallet tree는 six-stage two-band로 구성하였으며 음성프레임의 에너지 값을 포함해서 총 9차의 벡터를 구성하였다. 또한 이들 값의 1차, 2차 도함수를 구하여 전체 특징벡터의 차수는 27차로 구성하여 인식에 사용하였다.

4. HMM 훈련

HMM에서의 훈련은 이전의 음성 모델 Θ'이 있을 때 주어진 관측 벡터 y와 해당 상태 s의 더욱 최적화된 likelihood 값 $P(y|s,\Theta)$을 찾는 과정이다. 확률 추정은 전향 절차, 후향 절차에 의해서 이루어지고, 관측열이 j 상태에 있을 때 관측열 y를 발견할 확률 $b_j(y)$는 다음과 같이 구한다 [5].

$$b_j(y) = \frac{1}{\sqrt{(2\pi)^D \det U_j}} \exp\left\{-\frac{1}{2}(y-\mu_j)^T U_j (y-\mu_j)\right\} \quad j = 1,\ldots,N \qquad (11)$$

여기서 μ_j는 mean 벡터이고 U_j는 covariance 행렬이다. 식(11)의 경우는 다중화자 시스템에는 부적합하므로 다음 식(12)과 같이 M mixture의 확률밀도 함수를 사용한다.

$$b_j(y) = \sum_{n=1}^{M} c_j^n b_j^n(y) \quad \sum_{n=1}^{M} c_j^n = 1 \quad c_j^n \le 1 \qquad (12)$$

5. 실험 및 결과

실험에 사용한 음성데이터는 남성 화자가 신문을 낭독한 것으로 실험했다.

음성 녹음은 16kHz sampling rate, 16bit resolution으로 실험실 환경에서 녹음했으며, 윈도우 함수는 Hamming Window를 사용하였다. 앞에서 언급한 50음소의 분류는 아래와 같다.

표 1. 음소 기호의 분류

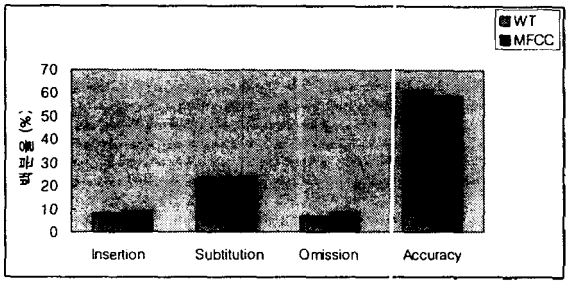
분류	기호	음소	분류	기호	음소
파열음	g	ㄱ 초성	유성자음	n	ㄴ
	gx	ㄱ 종성		m	ㅁ
	g+	ㄱ 유음사이		ng	ㅇ 종성
	G	ㄲ	모음	a	ㅏ 장음
	k	ㅋ		ax	ㅏ 단음
	d	ㄷ 초성		ya	ㅑ
	dx	ㄷ 종성		eo	ㅓ
	d+	ㄷ 유음사이		veo	ㅕ
	D	ㄸ		o	ㅗ 장음
	t	ㅌ		ox	ㅗ 단음
	b	ㅂ 초성		yo	ㅛ
	bx	ㅂ 종성		u	ㅜ
	b+	ㅂ 유음사이		yu	ㅠ
	B	ㅃ		eu	--
	p	ㅍ		i	ㅣ 장음
마찰음	s	ㅅ		ix	ㅣ 단음
	S	ㅆ		eui	ㅢ
	h	ㅎ 초성		wa	ㅘ
	h+	ㅎ 유음사이		wi	ㅟ
파찰음	z	ㅈ		weo	ㅝ
	Z	ㅉ		e	ㅐ,ㅔ
	c	ㅊ		ye	ㅒ,ㅖ
유성자음	r	ㄹ 초성		we	ㅚ,ㅙ,ㅞ
	rx	ㄹ 종성	묵음	sil	묵음
	r+	ㄹ 유음사이	잡음	#	잡음

웨이블렛 변환시 사용된 웨이블렛은 Coifman 24차 계수를 필터 계수로 사용하였다.

인식실험은 음소 인식을 했으며 인식에 사용된 문장은 총 87문장을 인식하였다.

다음 <표 2>에 실험에 대한 결과를 나타내었다.

표 2 인식결과 비교



6. 결론

위 결론에서 보듯이 전체적 인식률은 MFCC를 특징벡터로 했을 때보다 더 나은 인식률의 향상을 보여주고 있다. 이는 파열음에서 갑자기 튀는 부분에 대해서 웨이블렛 변환이 Fourier 변환보다 좀 더 잘 검출할 수 있기 때문인 것으로 보인다.

반면 유성자음에 대해서는 MFCC를 특징벡터로 한 인식의 경우가 웨이블렛 변환을 이용하여 인식한 경우보다 인식률이 약간 더 나은 것을 볼 수 있었다.

이와 같은 결과를 미루어 우리는 웨이블렛 변환의 장점과 MFCC의 장점을 살린 더 나은 특징벡터의 구성을 제시해 볼 수 있을 것으로 생각된다.

7. 참고 문헌

[1] Z. Tufekci, J.N. Gowdy, "Feature Extraction using Discrete Wavelet Transform for Speech Recognition," Proceedings of the IEEE, 2000.

[2] Yu Hao and Xiaoyan Zhu, "A New Feature in Speech Recognition Based on Wavelet Transform," Signal Processing Proceedings, 2000.

[3] C. Sidney Burrus and Ramesh A. Gopinath Haitao Guo, Introduction to Wavelets and Wavelet Transforms, (Prentice-Hall International, Inc, 1998), Chap.2, pp. 11-13.

[4] Mark J. Shensa "The Discrete Wavelet Transform: Wedding the À Trous and Mallet Algorithms," IEEE Transactions on Signal Processing, VOL. 40. NO. 10. October 1992.

[5] Claudio Becchetti and Lucio Prina Ricotti, Speech Recognition, (John Wiley & Sons, 1999), Chap.4, pp.179-182.