1993년도 한국음향학회 학술논문발표회 논문집(제 12권 1(s)호)

변형된 Wavelet 변환을 이용한 한국어 숫자음 인식에 관한 연구

지 상문*, 오 영환
한국과학기술원 전산학과

Isolated Korean Digits Recognition Using Modified Wavelet Transform

Sang-Mun Chi*, Yung-Hwan Oh
Dept. of Computer Science, KAIST

요 약

본 논문에서는 변형된 wavelet 변환을 통해 추출한 특징벡터를 이용하여 한국어 숫자음을 대상으로한 음성인식기를 구현하였다. wavelet 변환은 시간 및 주파수 영역에 대해 다중해상도(multiresolution)를 가지는 신호분석법이다. 본 연구에서는 계산량의 감소와 넓은 주파수 대역을 분석하기 위해, mother wavelet의 형태를 분석 주파수 대역에 따라 변화시키는 방법을 제안하였다. 기존의 wavelet 변환으로 실험한 결과 86.5%의 인식율을 얻었고, 변형된 wavelet 변환의 경우 96%의 인식율을 얻었으며 계산량이 감소하였다. 이와 함께 음성인식에서 널리 사용되는 특징파라미터인 멜켑스트럼과 FFT 멜스케일 필터 대역(mel scale filter bank)과 비교 실험한 결과 인식율의 향상을 보였다. 이는 제안한 방법이 고주파 대역의 세밀한 시간 해상도와 저주파 대역의 세밀한 주파수 해상도를 지니는데 기인하는 것으로 판단된다.

I. 서론

과학기술의 발전과 정보화 사회의 도래에 따라 인간과 기계간의 자유로운 의사소통에 대한 필요성이 증대되고 있다. 음성을 사용한 의사소통은 자연스러움, 신뢰성, 속도면에서 다른 방법들에 비해 우수한 면을 지니고 있다. 음성인식은 음성 신호로부터 언어적인 정보를 추출하여 인간이 이해 가능한 표현 방법으로 변환하는 과정으로, 음성합성과 함께 인간과 기계간의 정보교환을 가능하게 한다.

음성인식은 음성신호로 부터 특징을 추출하는 특징추출부와 추출된 특징벡터열을 이미 저장하고 있는 모델과 정합하여 인식하는 인식기로 나눌 수 있다. 특징추출을 위한 기존의 음성분석 방법은 분석구간내의 음성신호가 정상적(stationary)인 특징을 갖는다는 가정하에 음성을 모델링하여 특징을 추출한다. 음성인식의 특징추출을 위해 널리 사용되는 방법으로 LPC(linear predictive coding)분석과 퓨리에 변환이 있다. LPC분석과 퓨리에 변환은 정상적인 특징을 갖는 음성구간에서는 정확한 특징추출이 가능하지만, 음성의 시작부분, 또는 파열음이나 마찰음이 나타나는 음성구간에서는 분석의 전제에 어긋남으로 인해 신호에 대한 유효한 특징을 추출하기 어렵다. 퓨리에 변환은 모든 시간축 및 주파수 대역에 대해 같은 시간 및 주파수 해상도를 지니므로, 저주파의 긴 신호와 갑자기 변하는 짧은 고주파 신호로 구성된 신호의 분석에서 두신호에 대한 정확한 분석이 불가능하다. 이러한 해상도의 한계를 극복하고자 시간-주파수 평면에서 시간 해상도와 주파수 해상도를 변화시키는 다중 해상도 분석에 대한 연구가 진행되고 있다[4].

wavelet 변환은 시간 축 및 주파수 영역에 대해 다중해상도(multiresolution)를 가지는 신호분석법으로, 비정상적(nonstationary)인 신호의 분석에 유용하다. wavelet 변환은 고주파 대역에서는 세밀한 시간해상도와 거친(coarse) 주파수 해상도를 가지는 반면에, 저주파 대역에서는 세밀한 주파수 해상도와 거친 시간 해상도를 갖는 분석으로 사람의 귀가 가지는 주파수 응답(frequency response)과 유사한 특성을 가진 분석으로 알려져 있다[4].

기존의 wavelet 변환에서는 mother wavelet(a prototype wavelet)의 형태를 고정시키고 확대/축소에 의해서만 분석을 한다. 따라서, K옥타브 주파수 대역을 분석하려면 초기 wavelet의 2^K배 길이의 wavelet이 필요하게 되어 계산량이 많아지고, 저주파 대역의 시간 해상도가 거칠어지는 단점이 있다. 이러한 문제를 해결하고자 본 연구에서는 mother wavelet의 형태를 분석하고자 하는 주파수 대역에 따라 변화시키는 방법을 제안한다.

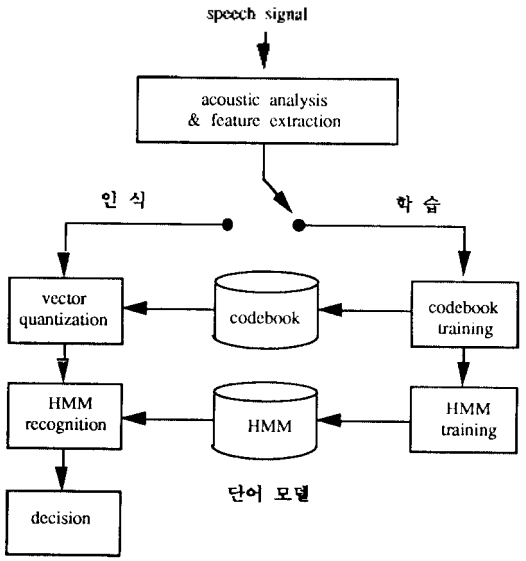
논문의 구성은 2절과 3절에서 음성인식 시스템의 구조와 특징추출방법 및 인식모델에 관해서 기술하고 4절에서는 실험에 사용한 자료를 설명하고 제안된 특징추출 방법과 기존의 특징추출 방법간의 성능비교 실험 및 결과를 분석한다. 마지막 5절에서는 결론및 추후연구 방향을 정리한다.

II. 인식 시스템의 개요 및 구성

음성인식을 위해서는 여러단계의 처리과정이 필요하다. 본 연구에서의 음성인식의 단계는 음성을 입력 받아 인식에 필요한 특징 파라미터로 변환하는 음향 분석 및 특징 추출 단계, 벡터양자화 를 통한 코드북의 작성 및 단어 모델을 학습하는 학습

단계, 저장된 모델로 부터 단어를 인식해 내는 인식 단계로 나눌 수 있다[그림 1].

첫번째, 전처리 및 음향분석단계에서는 음성신호를 16 kHz, 16 bit로 샘플링하여 고주파 영역을 강조하기 위한 preemphasis를 거친후 제안한 방법으로 wavelet 계수를 구하여 특징 파라미터로 사용한다. 두번째, 학습단계에서는 첫째 단계에서 구한 특징 파라미터를 LBG 알고리즘으로 초기화 한후 K-means 알고리듬을 사용하여 128개의 대표 패턴으로 양자화하였다. 단어 모델의 학습은 Baum-Welch가 제안한 reestimation 방법을 사용하여 단어 모델로 사용한 이산 HMM을 학습시켰다. 세번째, 인식 단계에서는 각모델에 의한 관측열의 생성 확률을 forward-backward 알고리즘을 사용하여 추정하였고, 최종 결정 단계에서는 최대 확률을 갖는 모델에 해당되는 단어를 인식 결과로써 출력하게 된다.



[그림 1] 시스템의 구성

III. 변형된 Wavelet변환을 이용한 특징추출

3.1 퓨리에 변환과 Wavelet 변환

특징추출을 위해 다양한 음성분석이 사용되고 있다. 변환을 통한 음성분석의 목적은 시간영역의 음성신호를 다른 영역으로 변환하여 음성신호에 관계된 유효한 특징을 추출하려는 것이다. 음성신호를 적당한 기저(basis)함수를 사용하여 이들 기저함수의 합으로 나타내는 방법 중 널리 사용되는 방법으로 퓨리에 변환이 있다.

단시간 퓨리에 변환 (short time fourier transform)은 기저함수로 시간창을 씌운 복소 sinusoid함수를 사용하여 식 3.1.1로 정의된다.

$$STFT(\omega, \tau) = \int_{-\infty}^{\infty} e^{-j\omega t} w(t-\tau)\, x(t)\, dt$$

$$= \langle e^{j\omega t} w(t-\tau),\ x(t) \rangle \qquad (3.1.1)$$

단, w(t)는 시간창이고, < >는 두 함수간의 내적이다.

퓨리에 변환은 고정된 길이의 기저함수를 사용하므로 모든 시간축 및 주파수 대역에 대해 같은 시간 및 주파수 해상도를 지닌다. 반면에 wavelet 변환은 고주파 대역의 분석을 위한 기저함수는 mother wavelet의 축소된 짧은 길이의 고주파 함수이고 저주파 대역의 분석을 위한 기저함수는 mother wavelet의 확대된 긴 길이의 저주파 함수이다. 이러한 기저함수를 사용하여 wavelet 변환은 고주파 대역의 세밀한 시간 해상도와 저주파 대역의 세밀한 주파수 해상도를 갖는다. 연속 wavelet 변환은 다음과 같이 정의된다.

$$CWT(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} h^*\left(\frac{t-b}{a}\right) x(t)\, dt$$

$$= \left\langle \frac{1}{\sqrt{a}} h\left(\frac{t-b}{a}\right),\ x(t) \right\rangle \qquad (3.1.2)$$

단, h(t)는 mother wavelet이며 매개변수 a, b가 실수($a, b \in R, a \neq 0$)이다. a값에 따라 mother wavelet이 확대/축소된 기저함수를 얻는다.

본연구에서 식 3.1.3의 modulated gaussian wavelet을 moher wavelet으로 사용하여 식 3.1.4의 연속 wavelet 변환을 얻었고, 실험시 이를 사용하였다.

$$h(t) = \exp(jct)\exp(-\tfrac{1}{2}t^2),\ c \geq 5 \qquad (3.1.3)$$

$$CWT(b, a) = \left\langle \frac{1}{\sqrt{a}} \exp\left(jc\tfrac{t-b}{a}\right)\exp\left(-\tfrac{1}{2}\left(\tfrac{t-b}{a}\right)^2\right),\ x(t) \right\rangle \qquad (3.1.4)$$

3.2 변형된 Wavelet 변환

단시간 퓨리에 변환과 연속 wavelet 변환에서의 기저함수를 $f(t) = K(t)\,L(t)$라 하고 표1과 같이 나 누어 보면 다음과 같은 유사성을 찾을 수 있다. 즉, K(t)는 기저함수의 형태를 결정하고 L(t)는 기저함수의 길이를 결정함을 알 수 있다.

[표 1] 기저함수의 비교

변환 \ 함수	k(t)	L(t)
fourier	$\exp(j\omega t)$	$w(t-\tau)$
wavelet	$\exp(jc\frac{t-b}{a})$	$\exp(-\frac{1}{2}(\frac{t-b}{a})^2)$

단시간 퓨리에 변환에서는 함수 L(t)의 길이(시간창 w(t-τ)의 길이)가 모든 분석에서 일정하고, 함수 K(t)의 형태(ω값)를 변화시킴으로써 분석을 수행한다. 연속 wavelet변환에서는 함수 L(t)의 길이(scale a)를 변화시키고 K(t)의 형태(C값)를 고정시킴으로써 분석을 수행한다. 다시 말하면, 단시간 퓨리에 변환은 기저함수의 형태만을 변화시키고 연속 wavelet변환은 기저함수의 길이만을 변화시킴으로써 분석함을 알 수 있다. 본연구에서는

wavelet 변환에서 기저함수의 길이와 형태를 변화시킴으로써 분석을 수행하는 방법을 제안한다.

제안된 방법을 구체화 하기위해 mother wavelet의 scale a와 C에서의 주파수 응답(frequency response)을 구하면 다음과 같다. (mother wavelet으로 modulated gaussian wavelet 을 사용했다.)

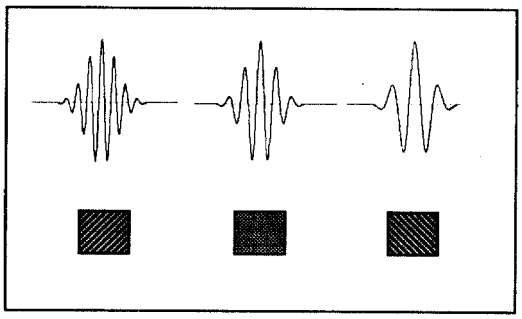
$$H_{a,c}(\omega) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} \exp(jc\frac{t}{a})\exp(-\frac{1}{2}(\frac{t}{a})^2)\exp(-j\omega t)dt$$

$$= \sqrt{2\pi a}\,\exp(-\frac{a^2}{2}(\omega-\frac{c}{a})^2) \qquad (3.2.1)$$

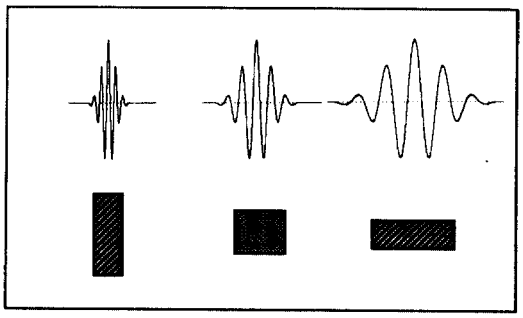
scale a와 C에서의 주파수 응답은 주파수 $\frac{c}{a}$를 중심으로 하고 주파수 해상도는 a와 비례한다. wavelet변환에서는 C는 고정시키고 scale a만을 조정하여 중심 주파수 $\frac{c}{a}$를 가지는 주파수 대역을 분석하므로, N옥타브의 주파수영역을 분석하기 위해서는 mother wavelet의 길이의 2^N배의 확대된 wavelet까지 필요하다. 긴 길이의 wavelet은 계산시간의 증가와 시간 해상도의 거칠음을 야기하므로 인식율의 저하를 가져온다.

본 연구에서는 이러한 문제를 해결하고자 중심 주파수$\frac{c}{a}$를 a만을 이용해서 변화시키지 않고, 고주파 대역에서는 C가 큰값을 갖고 저주파수 대역에서는 작은값을 갖게 하여 적은 계산량으로 넓은 주파수대역을 분석할 수 있다.

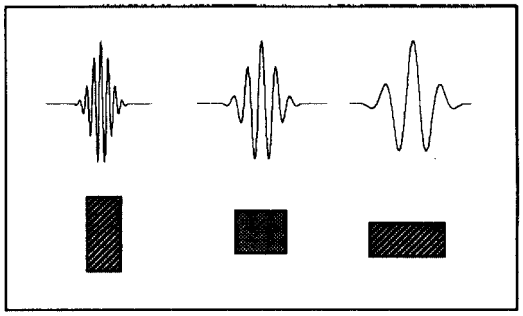
아래의 그림2, 3, 4는 단시간 퓨리에 변환, wavelet 변환과 제안한 방법의 기저함수와 그에 해당하는 시간 및 주파수 해상도를 나타낸다.(해상도를 나타내는 그림에서 가로축은 시간 해상도를, 세로축은 주파수 해상도를 나타낸다.)



[그림 2] 단시간 퓨리에 변환



[그림 3] wavelet 변환



[그림 4] 제안한 변환

샘플링된 이산적인 음성신호를 분석하기위해 연속 wavelet 변환에서 일정한 간격으로 표본추출된 wavelet 변환을 이용하여 특징벡터를 추출한다[1]. 표본추출된 wavelet 변환은 modulated gaussian wavelet을 mother wavelet으로 하여 다음과 같이 정의된다.

$$SCWT(iT_s, a) = T_s\frac{1}{\sqrt{a}}\sum_n h^*\left(\frac{(n-i)T_s}{a}\right)x(nT_s) \qquad (3.2.2)$$

단 T_s는 샘플링 주기이고, h(t)는 mother wavelet이다.

식3.2.1의 주파수 응답을 이용하여 TFQ Hz의 중심 주파수를 갖는 wavelet의 형태는 식 3.2.3로 구할 수 있다.

$$a = \frac{FLT_s}{2\sqrt{-2\log WS}},\quad c = 2\pi\,TFQ\,a \qquad (3.2.3)$$

여기서 FL은 필터의 길이, T_s는 샘플링 주파수, WS는 필터의 양끝단에서의 wavelet의 크기이다.

IV. 실험 및 결과

4.1 실험 환경

실험자료는 ETRI음성 데이터 베이스중에서 남성 화자 20인과 여성 화자 20인이 4회 발성한 10개의 숫자음(영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)을 대상으로 인식 실험을 수행하였다. 이중 남여 각20인이 2회 발성한 자료를 학습에 사용하였으며, 나머지 2회분은 인식 실험을 위해 사용하였다.

비교 실험에 사용한 특징 파라미터는 표2와 3과 같다.

표 2

	분석 방법	i 번째 필터길이	필터 대역
MWT	제안한 방법	$2^{i/10}$ x 64	22개
WT	wavelet 변환	$2^{i/5}$ x 64	22개

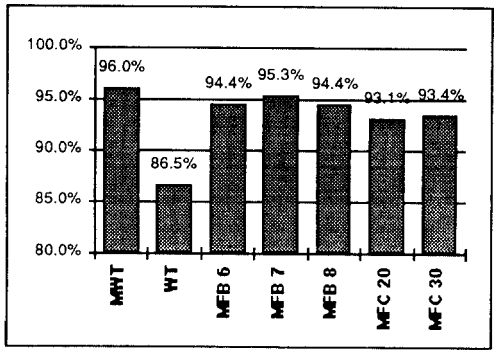
표 3

	분석 방법	분석 프레임 길이	특징벡터 차원
MFBk	FFT	2^k 샘플	22차
MFCCk	LPC 켑스트럼	k ms	12차

본 논문에서는 단어모델로 5 state를 가지는 left-to-right 이산형 HMM을 사용하여 패턴을 모델화 한다.

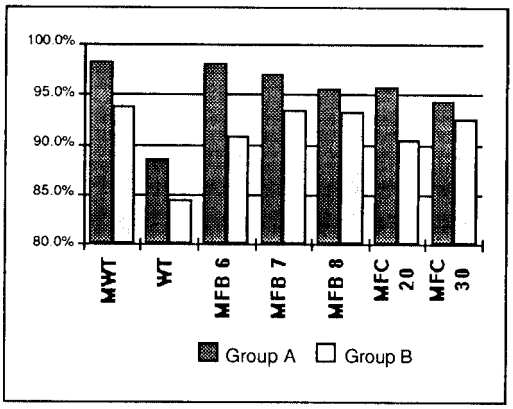
4.2 실험 및 결과

4.2절에서 설명한 여러 특징 파라미터 추출에 의한 비교 실험 결과를 그림 6에 나타내었다. 제안한 방법이 wavelet 변환에 대해 9.5%의 인식율의 향상을 보였다. 이는 wavelet 변환이 저주파 대역에서 가지는 거친 시간 해상도(가장 긴 필터의 길이가 74ms이다.)가 제안된 방법에서는 개선되었기 때문이다. 다른 특징 파라미터와의 비교 실험 결과 인식율이 가장 높은 MFB7에 대해 0.7%의 인식율의 향상을 보였다. 이는 제안한 방법이 고주파영역에서 세밀한 시간해상도를 가지고 저주파영역에서는 세밀한 주파수 해상도를 지니는데 기인하는 것으로 판단된다.



[그림 6] 전체 단어의 평균인식율

제안한 방법으로 추출된 특징 파라미터의 특성 고찰을 위해 음성신호의 음향학적 특징에 따라 대상 어휘를 비정상적인 특징을 가진 마찰음과 파열음으로 시작되는 숫자음(삼, 사, 칠, 팔, 구; A그룹)과 안정적인 모음으로 시작되는 숫자음 (영, 일, 이, 오, 육; B그룹)으로 나누어 결과를 고찰 하였다[그림 7].



[그림 7] 두그룹의 인식율 비교

그림 7에서 보는 바와 같이 A그룹의 음성신호는 같은 종류의 특징벡터를 사용할 경우 분석 프레임의 길이가 짧을수록, 즉 시간 해상도가 세밀 할수록 인식율이 좋았다. 이 결과를 고찰해보면 고주파 대역에 에너지가 집중되어 있고 비정상적인 음성신호인 마찰음과 파열음으로 시작되는 음성신호에 대해서는 주파수 해상도보다 시간 해상도가 인식에 중요한 영향을 끼침을 알 수 있다. 제안한 방법은 고주파 대역에서의 시간 해상도가 세밀하기 때문에 시간 해상도가 가장 세밀한 MFB6보다 0.3%의 인식율 향상이 있었다.

B그룹의 음성신호는 같은 종류의 특징벡터를 사용할 경우 분석 프레임의 길이가 길수록 인식율이 좋았다. 즉, 정상적인 모음으로 시작되고 A그룹에 비해 상대적으로 저주파에 에너지가 분포하는 음성신호에 대해서는 시간 해상도보다 주파수 해상도가 인식에 중요한 영향을 끼침을 알 수 있다. 제안한 방법은 저주파 대역에서의 주파수 해상도가 세밀하기 때문에 주파수 해상도가 세밀한 MFB7이나 MFB8보다 각기 0.3%와 0.5%의 인식율 향상을 보였다.

기존의 wavelet변환과 제안된 방법의 계산량을 비교하기 위해 옥타브를 v개의 보이스로 나누어 n개의 필터 대역일때 분석 프레임당 계산되는 필터길이의 합은 아래와 같다.

$$\text{초기 필터의 길이} \times \sum_{k=0}^{n-1} 2^{k/v} \qquad (4.3.1)$$

본 연구에서의 wavelet변환은 v=5, n=22이므로 분석 프레임당 계산되는 필터길이의 합은 8650이다. 제안한 방법에서는 v=10, n=22이므로 3196이 되어 제안된 방법의 계산량이 기존의 wavelet변환의 1/2.7로 감소 하였다.

V. 결론 및 검토

본 논문에서는 변형된 wavelet 변환을 통해 추출한 특징벡터를 이용하여 한국어 숫자음을 대상으로한 음성인식기를 구현하였다. 제안한 방법의 특징은 다음과 같다. 첫째, 기존의 wavelet 변환 방법보다 계산량이 적고 저주파 대역에서의 시간 해상도의 향상으로 인식율의 향상을 가져왔다. 둘째, 음성인식에서 널리 사용되는 특징파라미터인 멜캡스트럼과 FFT 멜스케일 대역과 비교 실험결과, 제안된 방법이 고주파 영역의 세밀한 시간 해상도와 저주파 영역의 세밀한 주파수 해상도를 자님으로써 인식율의 향상을 얻었다.

앞으로의 연구 방향으로는 mother wavelet으로 modulated gaussian wavelet만을 사용했는데 다른 형태의 mother wavelet에 대한 연구와 변형된 wavelet 변환을 통해 얻은 다차의 계수를 이용하여 새로운 계수(예, wavelet cepstrum)를 유도하여 이를 이용한 실험등이 필요하다 판단된다.

참고 문헌

[1] R. F. Favero, R. W. King; "Wavelet Parameterization for Speech Recognition", ICSPAT 93, Vol. 2, pp. 1444-1449

[2] A. Grossman, R. Kronland-Martinet, J. Morlet; "Reading and Understanding Continuous Wavelet Transforms" , Wavelets, Springer-Verlag Berlin, 1990

[3] L. R. Rabiner, B. H. Juang; "An introduction to hidden Markov Models", IEEE ASSP Magazine,Vol. 3, No. 1, pp. 4 - 16, 1986

[4] O. Rioul, M. Vetterli; "Wavelets and Signal Processing", IEEE Signal Processing Magazine, October, 1991

[5] M. Vetterli, C. M Herley; "Wavelets and Filter Banks: Theory and Design", IEEE Trans. SP., Vol. 40, No 9, pp. 2207-2232, 1992