論文96-33B-4-14

청각 모델에 기초한 음성 특징 추출에 관한 연구

(A Study on the Speech Feature Extraction Based on the Hearing Model)

金 바 울 *, 尹 晳 鉉 *, 洪 光 錫 *, 朴 炳 哲 *

(Paul Kim, Seok Hyun Yoon, Kwang Seok Hong, and Byung Chul Park)

요 약

본 논문에서는 청각 모델의 기능을 신호처리를 통해서 음성의 특징을 추출하는 방법을 제안하였다. 제안한 방법에서는 음성을 소구간으로 분할하여 극대값으로 정규화 하는 과정과 이산 Wavelet 변환과 역 이산 Wavelet 변환을 이용하여 신호를 다해상도로 분해, 재합성하는 과정, 분해 후와 합성 후에 각각 신호를 미분하는 과정, 그리고 전파 정류하고 적분하는 과정을 포함하고 있다. 제안한 방법에 의해 추출된 음성 특징의 타당성을 조사하기 위해서 DTW와 VQ-HMM 알고리즘을 이용하여 고립 숫자음 인식을 각각 수행하였다. 인식 결과, DTW의 경우, 화자 종속은 99.79% , 화자 독립은 90.33%의 평균 인식률을 나타내었고, VQ-HMM의 경우, 화자 종속은 96.5%, 화자 독립은 81.5%을 나타내어 구현이 간단한 알고리즘과 적은 차수의 특징 파라미터로도 효과적인 인식 성능을 나타내었다.

Abstract

In this paper, we propose the method that extracts the speech feature using the hearing model through signal processing techniques. The proposed method includes following procedure ; normalization of the short-time speech block by its maximum value, multi-resolution analysis using the discrete wavelet transformation and re-synthesize using the discrete inverse wavelet transformation, differentiation after analysis and synthesis, full wave rectification and integration. In order to verify the performance of the proposed speech feature in the speech recognition task, korean digit recognition experiments were carried out using both the DTW and the VQ-HMM. The results showed that, in case of using DTW, the recognition rates were 99.79% and 90.33% for speaker-dependent and speaker-independent task respectively and, in case of using VQ-HMM, the rate were 96.5% and 81.5% respectively. And it indicates that the proposed speech feature has the potentials to use as a simple and efficient feature for recognition task.

I. 서 론

음성 신호의 특징분석에 있어서 인간의 발성 모델에 기초하여 진보되어 온 방법이 LPC 계열의 분석법이라면, 필터 뱅크 분석법은 청각 모델에 기초한 연구를 통해 발달한 분석법에 해당한다. 이 두 가지 분석 방법은 그 특성에 따라 서로 다양한 적용 가능성을 가지고 있으며 서로의 장단점을 가지고 있다. 필터 뱅크 분석을 이용하여 음성의 특징을 추출하는 경우 LPC계열의 분석법에 비해 계산량이 많아 구현이 어려웠으나 최근 DSP 기술의 발달로 구현상의 난점은 해결되었다. LPC 계열의 분석 방식의 경우 계산은 간단하나 기본적으로 AR모델을 사용하기때문에 성도에서의 되먹임이 있는 유음의 경우 정보의 손실을 피할 수 없으며 pitch 주기가 작은 여성의 경우 pitch 성분과 성도 성분의 deconvolution에 문제가 발생한다. 이러한 점을

* 正會員, 成均館大學校 電子工學科

(Dept. of Electronic Eng., Sung Kyun Kwan Univ.)

接受日字:1996年1月5日, 수정완료일:1996年2月16日

고려 할때 보다 robust한 음성 특징의 추출을 위해서는 발성 모델보다는 청각 모델에 기초한 음성 특징의 추출이 요구된다. 그동안 음성이 귀를 거쳐 전기적인 신호로 변환되고 특징 정보들이 추출되어 뇌로 전달, 인지되는 과정을 모델링하고 이를 적용한 연구가 다수 발표되었으나[1][2][3][4], 아직 음성인식에 뛰어난 성능을 발휘하지 못하고 있는 실정이다.

본 논문에서는 기존에 제안된 초기 청각 모델[3][4]의 부분적인 기능들을 이산 Wavelet 변환을 이용한 특징추출 과정에 적합하도록 변형시켜 적용하고, 이를 실제 음성 인식에 적용하는 문제에 대해 연구하여, 청각 모델에 기초한 신호처리를 통해 음성 인식을 목적으로 하는 새로운 특징 추출 방법을 제안하였다. 제안한 방법은 차분 방정식, 이산 Wavelet 변환, 그리고 정류 등의 신호처리로 음성 특징을 추출할 수 있는 방법이다.

제안한 방법에 의해 추출된 음성 특징의 타당성을 조사하기 위해서 가장 일반적으로 사용되고 있는 DTW와 VQ-HMM 알고리즘을 이용하여 한국어 숫자음 인식을 수행하였다.

Ⅱ. 청각 모델에 의한 음성 특징 추출

이 장에서는 인간의 청각 모델에 대한 설명과 청각에 관여하는 개개의 기능들을 실제적으로 모델링하고 구현하는 방법에 대하여 설명한다.

1. 귀의 구조 및 작용

그림 1은 인간의 귀의 구조를 보여 주며 그림 2는 이를 간략화 하여 음의 전달 과정에 참여하는 기관들만을 나타내고 있다. 귀는 크게 외이(Outer Ear)와 중이(Middle Ear), 그리고 내이(Inner Ear)로 나누어지며, 외이는 소리를 모아주는 귓바퀴(Pina)와 음파가 전달되는 외이도(Outer Ear Canal) 그리고 음파의 진동을 감지하는 고막(Ear Drum)으로 구성된다. 중이는 외이의 고막과 내이의 난원창(Oval Window)사이의 빈 공간으로 고막의 탄력을 유지시키기 위해 유스타키오관(Eustachio Tube)에 의해 일정한 압력을 유지하고 있다. 그리고 고막의 진동을 난원창에 전달해 주는 세 개의 뼈로 구성된 청소골(Ossicles)이 있다. 내이는 매우 복잡한 형태를 가지고 있으며 실제적인 음파의 분석과 특징 추출 부분에 해당한다. 달팽이처럼 둥글게 말린 형태의 와우각(Cochlear)은 음의 인지 과정에 있어 대단히 중요한 역할을 한다. 그 내부는 Reissner막과 기저막(Basilar Membrane)의 2개의 막과 이로 인해 형성된 Scala Vestibuli, Scala Media, Scala Tympani의 3개의 관(Tube)들이 있으며 각각의 관은 이온 용액(림프액)으로 가득차있다. 3개의 관 중에 가운데 부분의 Scala Media가 Corti의 기관(The Organ of Corti)이라고 불리우는 청음 기관을 가지고 있다. Corti의 기관은 부위에 따라 진동의 형태로 전달된 음의 특정한 주파수에 가장 높은 위치 변화를 보이는 기저막과 기계적인 위치 변화를 감지해 전기적인 신호로 변환하는 감각모세포(Hair Cell)들, 그리고 위치 변화의 정도를 청세포에 전달하는 섬모(Cilia)들로 구성되어 있다.[5][6] 이러한 조직들에 의한 음의 인지 과정을 간략하게 살펴보면 다음과 같다.

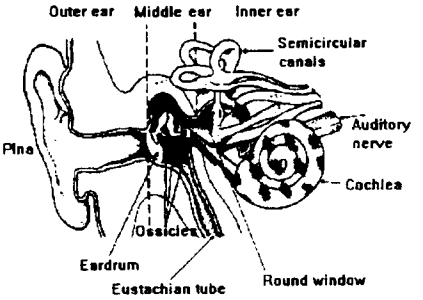


그림 1. 귀의 조직
Fig. 1. Ear structure.

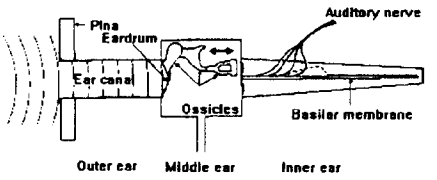


그림 2. 귀 조직의 간략화
Fig. 2. Simplified model of ear structure.

공기의 진동을 통해 전달된 음성신호는 외이도를 지나 귀의 고막을 진동 시키고 이 진동이 청소골의 증폭 작용을 통해 와우각에 전달된다. 와우각 내부에 있고위치에 따라 고유한 주파수에 반응하는 일종의 대역 통과 필터의 역할을 하는 기저막의 위치 변화는 섬모를 구부러뜨리고 감각모세포로의 이온의 유입을 비선형적으로 제어하여 신경세포인 감각모세포에 전기적인 신호를 유도한다. 여기서 기계적인 위치 변화가 전기적

신호로 바뀌게 되어 청각신경(Auditory Nerve)을 거치는 동안 다양한 정보들이 추출되어 뇌에 전달된다.[3][4]

2. 청각골(Ossicles)

청각골은 중이에 존재하며, 추골(Malleus), 침골(Incus), 등골(Stapes)이 차례로 연결되어 이루어진 일종의 기계적 변환 장치이다. 고막의 미세한 진동은 이 청각골에 의해 약 30배정도로 증폭되어 내이에 있는 와우각에 전달된다. 이러한 기능 외에 청각골은 매우 큰 음이나 갑작스러운 압력의 변화로부터 내이를 보호하고 압력을 일정하게 전달하는 기능을 수행한다.[5] 이는 입력된 음성신호를 일정한 레벨로 유지시켜 안정된 분석 출력을 내도록 음을 증폭하고 일정한 레벨로 유지시켜 주는 전처리 과정에 해당한다. 따라서 적정한 방법으로 높은 레벨의 음은 줄여 주고 낮은 레벨의 음은 높여 줌으로써 이 기능을 구현할 수 있다. 본 논문에서는 그림 3과 같이 입력된 음성신호의 소구간 극대값을 찾아 그 값을 기준으로 하여 정규화 하였다. 그 결과 그림 4와 같이 과대한 입력 값과 과소한 입력 값에 대해서 일정한 입력을 갖도록 하여 이산 Wavelet 변환을 이용한 특징 추출시 안정된 출력을 보장해 주게 된다.

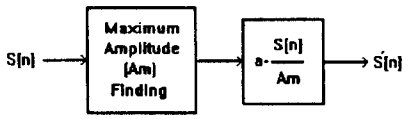


그림 3. 정규화 블럭도
Fig. 3. Block diagram of normalization process.

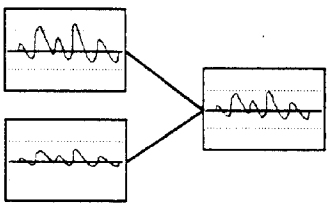


그림 4. 정규화 개념도
Fig. 4. Functional diagram of normalization process.

3. 기저막(Basilar Membrane)

내이의 와우각 내에 존재하는 기저막은 난원창쪽에 가까운 기저부위는 좁고 딱딱하며, 꼭지쪽로 갈수록 느슨한 형태를 가지고 있다.[5] 이러한 모양으로 인해 청각골을 통해 전달된 음파의 영향은 꼭지 근처에서는 100Hz 내외의 저주파에서 공진이 일어나 최대 위치 변화로 나타나며 기저쪽에서는 10,000Hz 이상의 고주파에서 그리고 꼭지와 기저의 중간지 점에서는 2,000Hz의 주파수에 기저막의 최대 위치 변화가 일어난다.[6]

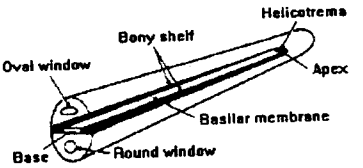
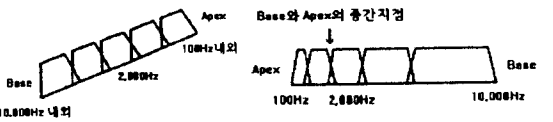


그림 5. 기저막
Fig. 5. Basilar membrane.



(a) (b)

그림 6. 기저막의 주파수 반응
Fig. 6. Frequency response of basilar membrane.

이처럼, 기저막의 기능은 마치 그림 5와 같이 음성신호의 각 주파수대를 다해상도로 분해하는 필터 뱅크로 해석할 수 있다.[3] 그림 6(a)는 그림 5의 기저막을 균일한 간격의 길이로 잘라 구성한 필터 뱅크로 가정한 것이고, 이는 주파수축을 Log Scale로 표현한 것으로 볼 수 있다. 그림 6(b)는 그림 6(a)의 주파수 축을 다시 선형적으로 나타낸 것이다. 이러한 기저막의 특성을 통해 인간의 청각 체계가 저주파에 민감하고 고주파에는 덜 민감하게 작용함을 알 수 있다. 또한 실제로 음성의 파형을 분석해 보면, 고주파 부분 보다는 상대적으로 저주파 부분에 정보가 집중되어 있는 것을 볼 수 있다. 따라서 저주파 부분은 주파수 해상도가 높은 필터를 사용하고 고주파 쪽에는 주파수 해상도가 낮은 필터를 사용하면 효과적으로 음성을 분석해 낼 수 있다.

본 논문에서는 이와 같은 기저막의 기능을 신호를 주파수 다해상도 분해하는 이산 Wavelet 변환과 역 이산 Wavelet 변환을 이용하여 구현하였다. 이산 Wavelet 변환 알고리즘은 연속적으로 Decimation을 수행하여 저주파쪽으로 갈 수록 데이타가 줄어드는 특

성으로 인해 제한된 연산으로 수행 속도가 빠르고 효율적인 Octave 대역 필터링을 수행할 수 있다.

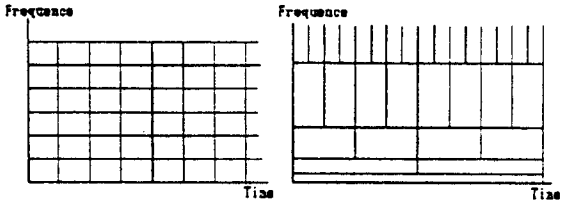
4. 이산 Wavelet 변환

Wavelet 변환은 원형(prototype) Wavelet을 정의하고, 이 원형 Wavelet의 Time Scaling과 Translation(Time Shift)을 통해 다양한 Wavelet을 구성하여 이러한 Wavelet들로 신호를 분해한다.[7]

$$\text{분해(WT)}: \quad W_\psi f(a,b) = |a|^{-\frac{1}{2}} \int f(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{1}$$

$$\text{합성(IWT)}: \quad f(t) = \frac{1}{c_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_\psi f(a,b)\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)\frac{dbda}{a^2} \tag{2}$$

기존에 음성의 분석에 사용되어 왔던 Short Time Fourier Transform(STFT)이 주파수(시간)에 대한 해상도가 일정한 반면, Wavelet 변환은 높은 주파수 대역은 주파수 해상도가 낮고(시간 해상도는 높다) 낮은 주파수 대역은 주파수 해상도가 높은(시간 해상도는 낮다) 다해상도 분해 특성을 나타낸다(그림 7).[9] 따라서 Wavelet 변환의 다해상도 분해 특성을 이용하면 그림 7에서 보인 기저막의 다해상도 필터 뱅크들을 구현할 수 있다.



(a) STFT (b) Wavelet 변환

그림 7. STFT 와 Wavelet 변환의 시간-주파수 해상도 비교

Fig. 7. Comparison of time frequency resolution of STFT and wavelet transformation.

이산 Wavelet 변환은 식 3과 같은 Dilation Equation을 정의한 후 이것으로부터 Orthogonal Wavelet들을 만들어 낼 수 있다.

$$\phi(x) = \sum_n h_0(n)\phi(Mx-n) \tag{3}$$

윗 식에서 $h_0(n)$는 Wavelet Basis로서 분해와 합성에서 사용되는 Basis가 같을 경우인 Orthogonal Wavelet Basis와 분해와 합성에서 사용되는 Basis가 다른 경우인 Biorthogonal Wavelet Basis로 구분된다. 이를 구하는 방법과 과정들은 이미 다양한 논문들을 통해 제시되어 졌으며, Quadruture Mirror Filter(QMF)의 관점에서 설계하는 방법도 가능한데 그 결과는 기존의 방법과 일치한다.[11]

이와 같이 Dilation Equation이 구해지면, 다음과 같은 방법으로 Dilation Wavelet들을 정의한다.

$$\phi(x) = \sum_n h_1(n)\phi(Mx-n) \tag{4}$$

여기서, $h_0(n)$는 저역필터에 해당하며, 고역필터 역할을 하는 $h_0(n)$와는 앞서 언급한 바와 같이 서로 QMF 관계가 있다. 다음식 5에 의해 그 관계가 정의된다.

$$h_1(n) = (-1)^{n+1}h_0(L-1-n) \tag{5}$$

여기서, L은 Wavelet Basis의 길이

이와 같이 정의된 방법으로 이산 Wavelet 변환을 수행하기 위해서는 가장 기본적인 형태로부터 Recursive하게 파형이 변화가 없을 때까지 구해야 한다. 그러나 실제로는 다음에 제시하는 방법을 통해 이산 Wavelet 변환을 매우 효과적으로 구현하게 된다.

그림 8은 이산 Wavelet 변환 방법으로 Mallat의 Tree 알고리즘(Subband Coding Scheme)이다.[8][9][10]

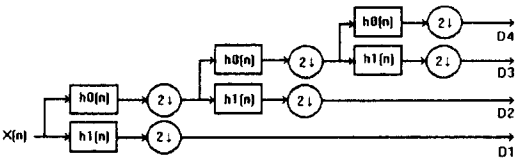


그림 8. 이산 wavelet 변환

Fig. 8. Discrete Wavelet Transform(DWT).

여기서 사용된 Wavelet Basis의 차수가 높을 수록 구현 속도는 떨어지지만 Regularity가 증가하여 보다 좋은 필터링 특성을 나타내게 된다.[9] 본 논문에서는 완전 재생 조건을 갖는 Daubechies 20차 Orthogonal Wavelet Basis를 사용하였다.

그림 9는 주파수 다해상도로 분해된 신호를 다시 합성하는 과정으로 역 이산 Wavelet 변환을 나타낸다. 여기서는 Decimation 대신 Interpolation을 수행한다.

DWT의 알고리즘의 특성상 연속적인 Decimation 과정을 거치게 되므로 소요되는 총 연산 횟수는 식 6과 같이 제한된다.[9]

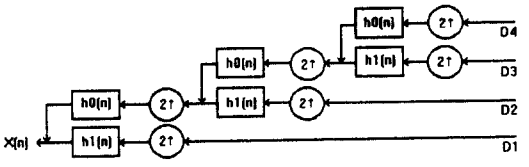


그림 9. 역 이산 wavelet 변환
Fig. 9. Inverse Discrete Wavelet Transform (IDWT).

따라서, 입력 샘플 수의 2배 이내의 연산만으로도 효과적으로 신호를 주파수 다해상도 분해할 수 있다.

$$C_{total} = C_0 + \frac{C_0}{2} + \frac{C_0}{4} + \ldots\ldots < 2C_0 \qquad (6)$$

여기서, C_0는 입력 신호의 샘플 수.

그림 10은 위에서 언급한 이산 Wavelet 변환과 역 이산 Wavelet 변환을 이용하여 신호를 분해, 합성하는 과정을 통해 음성신호를 주파수 다해상도로 분해하는 과정을 나타낸 블럭도 이다.

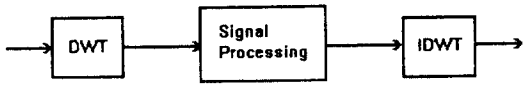


그림 10. Wavelet 필터 뱅크
Fig. 10. Wavelet Filter bank.

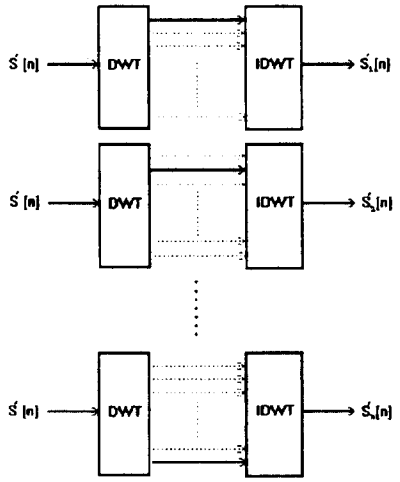


그림 11. DWT와 IDWT를 이용한 필터 뱅크
Fig. 11. Filter banks using DWT and IDWT.

그림 8의 방법만으로도 음성 신호의 주파수 다해상도 분해가 가능하나, Decimation 수행으로 인해 각 주파수대 출력 샘플 수가 다르고 청각 모델의 특성인 시변에서의 미분을 고려하기 힘들게 된다. 따라서, 본 논문에서는, 이산 Wavelet 변환에 의해 분해된 각각의 신호들을 역 이산 Wavelet 변환으로 재합성하는 방식으로 동일한 샘플 수를 갖는 필터 뱅크 수 만큼의 주파수 다해상도 분해된 시변 신호들을 얻는다. 그림 11은 그림 10의 구현 과정 중 중간의 신호처리 과정을 제외하고 나머지 부분을 보다 구체적으로 나타낸 그림이다.

그림 12는 위의 방법으로 구현한 Octave 대역 필터의 이상적인 결과를 나타낸다. 여기서, 구현 가능한 필터 뱅크 수는 Tree 알고리즘의 Decimation 특성에 의해 다음 식 7과 같이 제한된다.

$$\text{필터 뱅크 수} = \log_2(\text{분석 Frame 수}) + 1 \qquad (7)$$

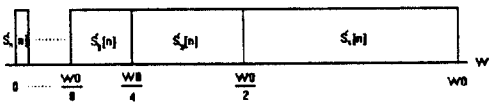


그림 12. 이상적인 Octave 대역 분할
Fig. 12. Ideal octave band division.

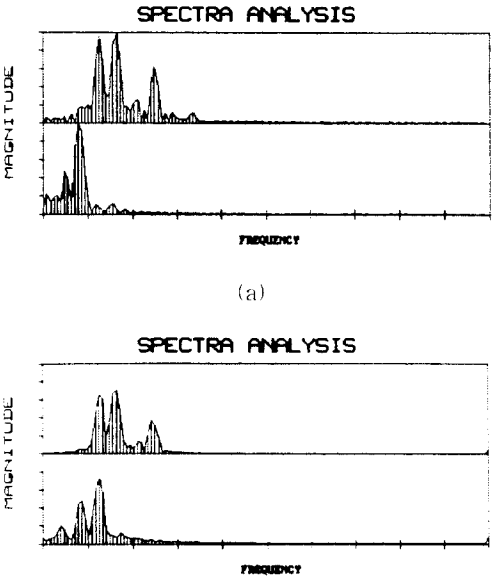


그림 13. DWT-IDWT 필터 와 FIR 필터와의 비교 (a) DWT-IDWT로 필터링한 발음 /아/의 2개 인접 주파수 대역 (b) 65차 Kaiser Window FIR 필터를 사용한 결과
Fig. 13. Comparison of DWT-IDWT filter and FIR filter.

그림 13은 DWT-IDWT를 이용하여 저주파 부분의 2개 인접 주파수대를 필터링한 결과를 Kaiser Window Method FIR 필터와 비교한 것이다. 이 그림은 본 논문에서 사용한 방법(그림 13(a))이 인접 주파수대와 겹치는 부분이 적고 매우 효과적으로 원 신호를 필터링 했음을 보여 준다.

5. Fluid-Cilia Coupling

기저막의 위치 변화에 의해 발생한 림프액의 흐름은 감각모세포와 연결된 섬모를 구부러 뜨린다. 림프액의 흐름이 섬모를 구부러 뜨리는 현상은 그림 14와 같이 시간에 대한 기저막 필터 출력들의 미분 형태로 모델링 한다.[13][14]

시변에서의 미분은 그림 15와 같이 주파수변에서의 고주파 강조 특성을 나타내며 이산 신호의 경우 미분은 다음과 같은 식 8의 차분 방정식으로 간단하게 구현할 수 있다.

$$\dot{S}_n[n] = S_n[n] - S_n[n-1] \quad (8)$$

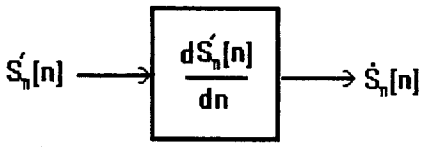


그림 14. Fluid Cilia 커플링 블럭도
Fig. 14. Block diagram of fluid-cilia coupling.

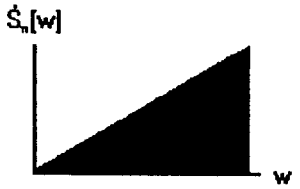
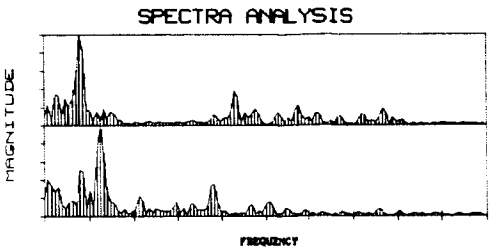
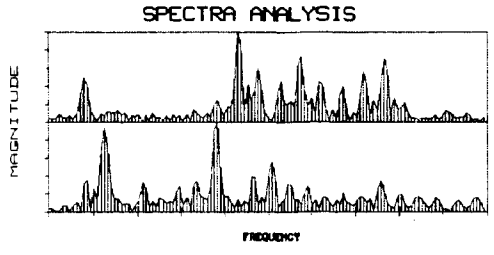


그림 15. Fluid-Cilia 커플링 개념도
Fig. 15. Functional diagram of fluid cilia coupling.

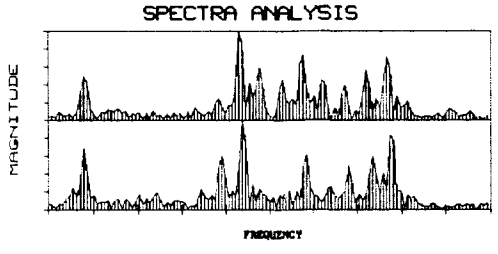
제안한 방법으로 필터 뱅크 출력을 얻은 후 차분 방정식을 적용하면, 그림 16(a)와 같이 유사한 주파수 특성을 나타내는 두 발음 /아/와 /에/는 그림 16(b)와 같이 더욱 명확해지고 다른 화자라 하더라도 같은 발음 /이/에 대해서는 그림 16(c)와 같이 유사하게 나타난다. 이 결과를 통해, 음성 인식 적용시 인식률 향상에 기여할 것으로 기대할 수 있다.



(a) /이/와 /에/에 대해 DWT-IDWT를 수행한 결과



(b) 그림(a)의 결과를 미분한 경우



(c) 서로 다른 화자가 발음한 /이/를 미분한 경우

그림 16. Fluid-Cilia 커플링의 고주파 강조 특성
Fig. 16. High-frequency emphasis of fluid cilia coupling.

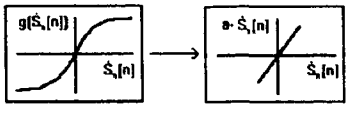


그림 17. Ionic Channel을 선형으로 가정할 때의 개념도
Fig. 17. Functional diagram of ionic channel assuming the liearity.

6. 이온 채널(Ionic Channel)

섬모의 구부러진 정도에 따라 감각모세포로 유입되는 이온의 양이 비선형적으로 조절된다. 이온의 유입은 감각모세포 내부와 외부와의 전위차를 유발시키고, 이렇게 형성된 전위차는 청각 신경 섬유(Auditory Nerve Fiber)를 통해 청각 중추로 전달된다.[13][14] 실제 이온 유입량은 그림 17과 같은 Sigmoid 함수의 비선

형 특성을 나타내지만, 본 논문에서는 포화상태와 포화 상태 직전을 제외한 나머지 부분을 선형으로 가정하고 모델링 한다.

이렇게 하면 그림 17에서 볼 수 있듯이 증폭인자 a 만이 이온 채널의 특성을 나타내는 요소가 된다. 증폭인자 a도 1(Lossless Transmitter)로 가정하여 이 특성 자체를 고려 대상에서 제외하면 이산 Wavelet 변환 적용시 유리하게 작용하게 된다.

7. Lateral Inhibitory Network(LIN)

청각 중추에서는 음색, Pitch, 그리고 시간상, 주파수상의 특징 등 다양한 정보들이 추출된다. 이러한 청각 체계를 모델링하기 위해 모든 감각 체계에서 발견되는 LIN의 간단한 기능들을 이용한다. LIN은 청신경으로부터의 출력을 받아 구조와 기능에 따른 특징을 나타내는 것으로 추측된다. 이러한 기능들은 LIN Neuron들 사이의 측면 상호작용(Lateral Interaction)을 모방한 주파수변에서의 미분과 LIN 뉴런들의 비선형성을 고려한 반파 또는 전파 정류 작용, 그리고 출력 값들의 시변 적분의 세단계 과정으로 모델링 할 수 있다.[13][14]

그림 18에 그 과정들을 나타내었다.

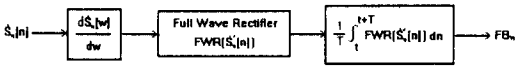


그림 18. LIN의 모델링
Fig. 18. Modelling of the LIN.

주파수변에서의 미분은 시변에서의 미분과 마찬가지로 차분 방정식으로 식9와 같이 구현한다.

$$S'_n[w] = S_n[w] - S_n[w-1] \quad (9)$$

특별히 이산 Wavelet 변환을 이용할 경우 주파수변에서의 미분은 매우 간단하고 효과적으로 적용할 수 있게 된다. 즉, 그림 11에서 제시한 알고리즘에서 DWT 를 수행하여 대역 분할한 결과들에 곧바로 차분 방정식을 적용하는 것이다.

앞 절에서 제시한 바와 같이 이온 채널을 선형으로 가정하고 증폭 인자를 1로 가정하여 무시하였으므로, 식 10에 의해 시변에서의 미분 과정과 주파수 변에서의 미분 과정의 순서를 바꿔 구현할 수 있다.

$$\frac{\partial}{\partial w}\left(\frac{\partial S'_n[n,w]}{\partial n}\right) = \frac{\partial}{\partial n}\left(\frac{\partial S'_n[n,w]}{\partial w}\right) \quad (10)$$

그림 18의 최종 출력은 전파 정류기 출력의 소구간 평균값으로서 음성 인식을 위한 특징으로 사용한다.

Ⅲ. 모의 실험 및 결과

본 논문에서 제안한 음성 인식을 위한 특징 추출 과정을 그림 19에 나타내었다.

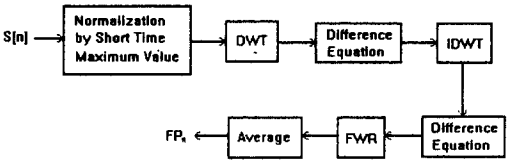


그림 19. 음성 특징 추출 과정 블럭도
Fig. 19. Block diagram of speech feature extraction process.

제안된 방법으로 얻을 수 있는 음성 특징의 갯수는 구현 가능한 필터 뱅크 수와 동일하며 앞서 언급한 식 7과 같이 제한된다. 본 논문에서는 1 Frame당 256 샘플을 사용하므로,

$$\log_2 256 + 1 = 9$$

즉, 총 9개의 필터 뱅크 구성이 가능하며, 이들 중 가장 저주파 부분의 출력은 왜곡을 고려해 제외하고 나머지 8개 필터 뱅크 출력을 이용한 특징만을 사용한다.

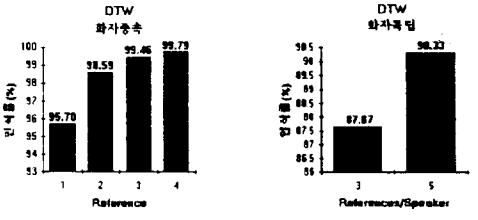
표 1. 실험 조건
Table 1. Conditions for experiments.

인식 대상 어휘	한국어 고립 숫자음 0-9
샘플링 주파수	16 KHz
양자화 레벨	16 bits
화자	20대 성인 남자 8명
어휘 당 발음 횟수	10회
총 데이타 수	800개(8*10*10)
인식기	DTW, VQ-HMM
VQ 방식	LBG 알고리즘

제안된 음성 특징의 성능 평가를 위해 본 논문에서는 인식기로서 많이 사용되고 있는 DTW와 VQ-HMM을 사용하여 인식 실험을 수행 하였다. VQ-HMM의 경우 LBG 알고리즘을 사용하여 우선 32레벨로 벡터 양자화하여 실험하였고, 후에 64 레벨 양자화

한 결과도 비교하여 표2에 나타내었다. HMM 모델은 기본적으로 6-State Left-to-Right 모델이며, 상태수를 가변하여 비교 실험한 결과를 표2와 표3에 나타내었다. 인식 대상 어휘는 고립 숫자음 0에서 9까지이며, 총 8명의 20대 성인 남자로부터 각 단어 당 10회씩 발음한 것을 일상적인 실험실 환경하에서 채취하였다. 실험 조건은 표1에 나타내었다.

DTW의 경우 Reference로 사용한 데이타는 인식에서 제외하였으며 Reference 수를 가변하여 인식 실험을 수행하였다. 그림 20(a)에 화자 종속 인식 결과를 나타내었다. 화자 독립의 경우에는 1번화자와 2번화자의 음성을 Reference로 사용하였고, 각 화자 당 사용 Reference 수를 달리하여 인식 실험을 수행하였다. 그림 20(b)에 인식 결과를 나타내었다. VQ-HMM으로 인식을 수행한 경우에도 마찬가지로 훈련에 사용한 데이타는 인식에서 제외하였다. 화자 종속의 경우 2가지의 실험을 수행하였는데, 첫번째는 각 화자별 인식 실험을 한 경우(방법I)로서 단어 당 5회씩 발음한 데이타를 훈련 데이타로 사용하고 나머지 5회 발음한 데이타로 인식을 수행하였다. 두번째는 모든 화자에 대해 학습을 수행한 경우로서 8명의 화자가 5회씩 발음한 데이타 총 400개로 학습하고 나머지 5회씩 발음한 총 400개의 데이타로 인식(방법II)을 수행하였다. 화자 독립의 경우 총 8명의 화자 중 4명의 화자 총 400개의 데이타로 학습하였고, 나머지 4명의 화자 총 400개의 데이타로 인식을 수행하였다. 각각의 인식 결과를 그림 21에 함께 나타내었다.



(a) (b)

그림 20. DTW를 이용한 화자 종속 및 화자 독립 인식 결과

Fig. 20. The results of recognition experiments using DP matching for speaker dependent and speaker-independent task.

그림 22는 청각 모델의 적용이 인식률에 미치는 영향을 비교하여 나타낸 그래프이다. Normalization process의 적용 시와 그렇지 않은 경우 평균 5.48%의 인식률 차이를 보였고, 주파수 변에서의 미분(LIN) 적용 시 평균 0.72%, 그리고 시변에서의 미분(Fluid-Cilia Coupling) 적용 시에는 평균 2.13% 정도의 인식률 증가를 나타내었다. 또한 Reference 수가 적을수록 청각 모델의 적용은 더 효과적으로 나타났다.

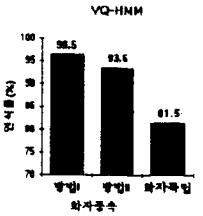


그림 21. VQ-HMM을 이용한 화자 종속 및 화자 독립 인식 결과

Fig. 21. The results of recognition experiments using VQ-HMM for speaker-dependent and speaker-independent task.

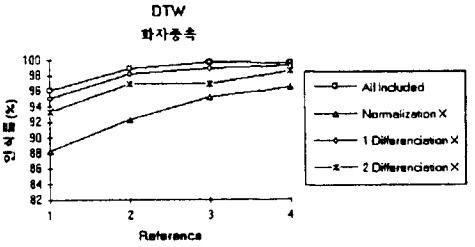


그림 22. 청각 모델 적용이 인식률에 미치는 영향 비교

Fig. 22. Effects of the hearing models on recognition rate.

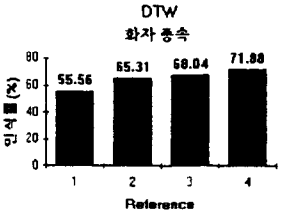


그림 23. IDWT를 사용하지 않은 경우의 인식률

Fig. 23. Recongnition rate for each reference in case of not using the IDWT.

그림 23은 역 이산 Wavelet 변환을 이용하여 재합성하는 방법으로 얻는 필터 뱅크 출력 방법의 효율성을 평가하기 위해 이산 Wavelet 변환만을 사용하여 음성 신호를 주파수 다해상도 분해한 후 LIN을 적용하

여 특징을 추출하여 DTW(화자 종속)에 적용한 결과이다. 결과적으로 이산 Wavelet 변환만을 사용할 경우에는 그다지 좋은 인식률을 얻을 수 없는 것을 알 수 있다.

표 2는 VQ-HMM을 이용한 화자 종속 음성 인식의 경우 VQ의 레벨과 HMM의 State 수를 가변하여 인식률의 변화를 나타낸 것이다. 결과적으로 한국어 숫자음 0에서 9까지의 경우 제안한 방법으로 인식을 수행하는데 있어 VQ 레벨은 32, HMM 상태수는 6인 경우 가장 높은 인식률을 나타내었다.

표 2. VQ 레벨과 HMM State 수에 따른 화자 종속 인식률

Table 2. Recognition rate for speaker-dependent experiment.

HMM state			3	4	5	6
VQ level	32	방 법 I	95.75 %	96 %	95.5 %	96.5 %
		방 법 II	90.5 %	91 %	92.5 %	93.5 %
	64	방 법 I	96 %	95.25 %	95 %	95.25 %
		방 법 II	92 %	92 %	92 %	92 %

표 3은 VQ-HMM을 이용한 숫자음 화자 독립 음성 인식의 경우, 앞에서와 마찬가지로 VQ의 레벨과 HMM의 State 수를 가변 하는데 따르는 인식률 변화를 나타낸 것이다. 그 결과 역시 VQ 레벨은 32, HMM 상태수는 6에서 가장 높은 인식률을 나타내었다.

표 3. VQ 레벨과 HMM State 수에 따른 화자 독립 인식률

Table 3. Recognition rate for speaker-independent experiment.

HMM state		3	4	5	6
VQ level	32	80.75 %	78 %	80.25 %	81.5 %
	64	79 %	79.25 %	78.75 %	78.75 %

Ⅳ. 결 론

본 논문에서는, 이산 Wavelet 변환과 청각 모델을 이용하여 음성 인식을 위한 음성 특징을 추출하는 알고리즘을 제안하였다. 제안된 방법으로 추출된 특징이 음성 인식에 적합한가를 확인하기 위해 DTW와 VQ-HMM을 이용하여 인식 실험을 수행하였다. 인식 결과 DTW의 경우, 화자 종속은 최고 99.79%, 화자 독립은 최고90.33%의 인식률을 나타내었다. 또한, VQ-HMM의 경우 화자 종속은 96.5%, 화자 독립은 81.5%을 나타내어 구현이 간단한 알고리즘과 적은 수의 인식 파라미터로도 효과적인 인식 성능을 나타내었다. 또한, 청각 모델의 적용이 인식률에 미치는 영향을 알아보기 위해 DTW를 이용하여 각 단계의 적용 시와 삭제 시의 인식률을 비교하였다. 그 결과 청각 모델의 모든 단계의 적용이 인식률 향상에 도움을 주는 것으로 나타났다.

참 고 문 헌

[1] Jont B. Allen, "Cochlear Modeling", IEEE ASSP Magazine, January 1985.

[2] Stephanie Seneff, *A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing*, Academic Press, 1988.

[3] Xiaowei Yang, Kuansan Wang, Shihab A. Shamma, "Auditory Representation of Acoustic Signals", IEEE Trans. Information theory, Vol 38, No. 2, 1992.

[4] Kuansan Wang, Shihab A. Shamma, "Self-Normalization and Noise-Robustness in Early Auditory Representations", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, July 1994.

[5] Thomas D. Rossing, *The Science of Sound*, Addison-wesley Publishing Company, 1990.

[6] David Ottoson, *Physiology of the Nervous System*, Macmillan Press, 1983.

[7] Randy K. Young, *Wavelet Theory and Its Application*, Kluwer academic publishers, 1993.

[8] D.E. Newland, *An Introduction to Random Vibration, Spectral & Wavelet Analysis*, Longman Scientific & Technical, 1993.

[9] Olivier Rioul, Martin Vetterli, "Wavelets and Signal Processing", IEEE SP magazine, pp 14-38, Oct., 1991.

[10] Mac A. Cody, "The Fast Wavelet Transform", Dr. Dobb's Journal, pp 16-24, April, 1992.

[11] 권상근, "QMF banks의 PR조건을 이용한 웨이브렛 기저의 설계 및 응용", Telecommunications review, Vol 3, pp 38-71, 1993.

저 자 소 개

金 바 울(正會員)

1994年 2月 성균관대학교 공학사. 1996年 2月 성균관대학교 대학원 공학석사. 1995年 12月 ~ 현재 (주)신도리코 기술연구소 연구원. 주관심 분야는 음성 및 신호처리

洪 光 錫(正會員)

1985年 2月 성균관대학교 공학사. 1988年 2月 성균관대학교 대학원 공학석사. 1992年 2月 성균관대학교 대학원 공학박사. 1990年 3月 ~ 1993年 2月 서울보건전문대학 전산정보처리학과 전임강사. 1993年 3月 ~ 1995年 2月 제주대학교 정보공학과 전임강사. 1995年 3月 ~ 현재 성균관대학교 전자공학과 조교수. 주관심 분야는 음성 및 신호처리, 패턴인식 등임

尹 晳 鉉(正會員)

1992年 2月 성균관대학교 공학사. 1996年 2月 성균관대학교 대학원 공학석사. 1996年 3月 ~ 현재 성균관대학교 대학원 박사과정. 주관심 분야는 음성 및 신호처리, 적응필터 등임

朴 炳 哲(正會員) 第 31卷 B編 第 1號 參照
성균관대학교 전자공학과 명예교수