퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 2, pp. 150-155

웨이블렛 변환을 이용한 음성에서의 감정 추출 및 인식 기법

Emotion Recognition Method from Speech Signal Using the Wavelet Transform

고현주* · 이대종* · 박장환** · 전명근*

Hyoun-Joo Go*, Dae-Jong Lee*, Jang-Hwan Park**, Myung-Geun Chun*

*충북대학교 전기전자컴퓨터공학부 컴퓨터 정보통신 연구소

**충주대학교 정보 제어 공학과

*Chungbuk National University School of Electrical and Computer Engineering
Research Institute for Computer and Information Communication

**Chungju National University School of Information & Control Engineering

요 약

본 논문에서는 사람의 음성속에 내포된 6가지 기본 감정(기쁨, 슬픔, 화남, 놀람, 공포, 혐오)의 특징을 추출하고 인식하고자 한다. 제안한 감정인식 알고리즘은 웨이블렛 필터뱅크를 이용하여 각각의 감정별 코드북을 만들고, 인식단계에서 필터뱅크별 감정을 확인한 후 최종적으로 다중의사결정기법에 의해 감정을 인식하는 구조로 이루어져 있다. 이와 같은 웨이블렛 필터뱅크와 다중의사 결정기법에 기반을 둔 알고리즘의 유용성을 보이기 위해 실험에 사용된 음성은 20명의 화자로부터 6가지의 감정을 대상으로 각각 3번씩 발음한 감정음성을 녹음하여 총 360개의 데이터베이스로 구성하고 실험하였다. 이와 같이 제안한 알고리즘은 기존의 연구에 비해 5% 이상 향상된 인식률을 보였다.

Abstract

In this paper, an emotion recognition method using speech signal is presented. Six basic human emotions including happiness, sadness, anger, surprise, fear and dislike are investigated. The proposed recognizer have each codebook constructed by using the wavelet transform for the emotional state. Here, we first verify the emotional state at each filterbank and then the final recognition is obtained from a multi-decision method scheme. The database consists of 360 emotional utterances from twenty person who talk a sentence three times for six emotional states. The proposed method showed more 5% improvement of the recognition rate than previous works.

Key words : 감정인식(Emotion Recognition), 웨이블렛 변환(Wavelet Transform), 음성인식(Speech Recognition), 휴먼인터페이스(Human Interface)

1. 서 론

디지털시대에 자주 접하는 모바일, 아바타, 인터랙티브 미디어, 휴먼 인터페이스 등은 사용자의 개성을 표현 및 인지하고, 상호간에 대화하며 반응할 수 있는 기술로 인간중심 장치를 구성하기 위해 사용되고 있다. 이렇게 고도로 발달된 정보화 시대는 인간을 모델링하는 것을 목표로 기술을 발전시키고 있으며, 인간을 모델링하기 위한 도구로는 언어, 음성, 제스처, 시각, 청각 등이 이용될 수 있다. 그리고 이와 관련된 연구가 최근 들어 활발히 진행되고 있다.

이에, 휴먼인터페이스 방법 중 하나로 얼굴표정을 이용한 감정인식을 들 수 있는데, 이는 얼굴의 주요 특징인 눈, 코, 입의 위치를 찾는 것으로 얼굴의 모양과 기하학적 관계를 파악하여 인식기를 구성하는 방법과 광 플로우(optical flow)나 포텐셜필드(potential field)등을 이용하여 얼굴감정을 인식하는 방법이 시도되었다[1][2][3][4]. 최근에는, 얼굴인식에 많이 사용되는 PCA(Principal Component Analysis), LDA(Linear Discriminant analysis)를 이용하여 웃는 얼굴이 무표정한 얼굴에 비해 더 좋은 인식률을 보일 수 있음을 제안하였다[5].

또한, 일본의 Fumio 교수팀은 역전파 학습알고리즘에 의한 인공신경망을 이용하여 여섯 가지 얼굴 표정을 인식할 수 있는 얼굴 로봇을 구현하였으며[6], Matsuno등은 에지(edge) 영상에서의 포텐셜필드(potential field) 개념과 Karluen-Loeve 변환을 사용하여 분노, 행복, 슬픔, 놀람의 4가지 얼굴표정에 대한 감정인식기법을 연구하였다[7].

음성은 청각에 기반을 둔 가장 효율적이고 자연스러운 휴먼 컴퓨터 인터페이스로 기대되고 있는 분야로, 심리학자인 Ekman과 Friesen에 따르면 사람의 여섯 가지 감정인 기쁨, 슬픔, 화남, 놀람, 공포, 혐오는 각 문화에 영향을 받지 않고 공통으로 인식하는 기본감정으로 분류하고 있다[8]. 이러한

접수일자 : 2003년 2월 14일
완료일자 : 2003년 3월 31일
감사의 글 : 본 연구는 한국과학재단 목적기초연구(R01-2002-000-00315-0) 지원으로 수행되었음.

인간의 기본적인 감정을 인식하기 위한 컴퓨터 인터페이스와 관련한 연구 및 응용제품은 최근 들어 큰 관심의 대상이 되고 있으며, 이러한 기술발전의 변화는 당연한 것이라 할 수 있다.

이와 관련한 연구로 음성 속에 내포된 여러 가지 감정을 추출하려는 연구가 최근 들어 활발히 행해지고 있으며, Fukuda는 음성신호의 템포와 에너지를 가지고 여섯 개의 기본감정에 대한 분류를 시도하였는데, 녹음실과 같은 외부 잡음이 전혀 없는 환경 하에서 일본어와 이탈리아어에 대한 음성신호를 녹음한 후 감정 추출에 대한 연구를 하였다[9]. Moriyama는 음성신호의 피치(pitch)와 전력의 포락선 검출을 통하여 20개의 일본어 샘플에 대하여 실험하였으며, 실험결과 '화남' '슬픔' '놀람' 감정이 다른 감정보다 인식률이 비교적 높은 것으로 나타났다[10]. 또한, Silva는 음성신호의 피치와 HMM(Hidden Markov Model)을 이용하여 영어와 스페인어에 대하여 감정인식을 연구하였다[11].

한편, 국내에서도 음성을 이용한 감정인식 연구가 활발히 진행되고 있는 요즈음, 우리나라 국악의 창에서 인간의 희로애락을 표현하는 음의 고저와 장단을 기본으로 하여 분석하는 연구가 행하여졌으며[12], 화난 감정의 독특한 특성을 찾아내기 위해, 대화의 내용에 사용한 단어나, 톤(Tone), 음성신호의 피치(Pitch), 포만트 주파수(Formant Frequency), 말의 빠르기(Speech Speed), 음질(Voice Quality)등을 이용하는 연구가 진행 중에 있다[13]. 그러나, 모국어인 한국어의 경우 지역 간의 억양 차이나 방언, 사투리, 또는 개개인의 특성에 따라 피치나 말의 빠르기, 음성의 톤이 다르기 때문에 일반적인 방법을 사용하여 감정인식을 하기에는 매우 어려운 상황이다.

따라서, 본 논문에서는 널리 알려진 방법과는 달리 사람의 음성신호에 대하여 웨이블렛 서브밴드 필터뱅크를 적용하고 각 주파수 대역별로 음성신호를 분리하여 각각의 대역에 캡스트럼을 이용한 특징벡터를 산출한다. 이때, 감정이 섞인 음성에 대해 6개의 감정 코드북(Codebook)을 생성하여 저장하는 것을 등록과정으로 하였다. 인식하는 과정에서는 감정이 섞인 입력음성이 받아들여지면 등록과정에서 특징벡터를 산출했던 것과 같이 특징벡터를 산출하여 유사도를 측정하고 인식 대상 감정을 선정한다. 이때, 본 연구에서 제안하는 의사결정 방법을 적용하여 다중밴드로부터 최종 감정을 판별해 낼 수 있도록 구현하였다. 본 실험에 사용된 데이터로는 남성화자 10명과 여성화자 10명을 대상으로 심리학자 Ekman과 Friesen에 의해 분류된 기본감정 기쁨, 슬픔, 화남, 놀람, 공포, 혐오 6개의 감정에 대해서 분류하는 알고리즘을 제안하고 실험하였다. 음성파일의 형식은 샘플링 주파수 11kHz, 16bit, Mono를 사용하였으며, 웨이브 파일을 분석하기 위한 도구로는 상용중인 Cool Edit 2000을 사용하였다.

2. 감정인식을 위한 웨이블렛 신호해석

일반적으로 실생활에서 접하게 되는 대부분의 신호는 신호의 흐름을 나타내는 시간 축과 신호의 크기를 나타내는 진폭 축으로 표현된다. 이러한 신호를 시간영역에서만 분석하는 경우 신호가 포함하고 있는 정보를 충분히 해석하기 어렵기 때문에 신호분석은 시간영역의 신호를 주파수영역으로 변환하는 기법을 사용한다. 널리 사용되는 방법으로 퓨리에 변환은 오늘날 신호처리, 영상 압축 등 다양한 분야에서 원신호 데이터가 지니고 있는 특성을 추출해 내기 위해 기저함수의 가중 합으로 원 신호를 표현하는 기법으로 이용되고 있다. 그러나 퓨리에 변환은 정현파 함수가 무한한 범위를 갖는 신호이기 때문에 시공간 영역에서 발생하는 불규칙 신호의 발생 시점을 정확히 찾아낼 수 없으며 신호의 크기도 파악하기 어려운 단점을 가지고 있다. 이런 단점을 해결하기 위해 공간과 주파수 두 영역 모두에서 신호의 변화에 대한 정보를 표현할 수 있는 STFT가 제안되었으며, 식 (1)과 같이 정의할 수 있다. 식 (1)에서 w는 윈도우 함수이며, STFT는 신호의 모든 구간에서 동일한 윈도우가 적용되므로 시간과 주파수 영역에서 해상도가 같아지게 되어 구간마다 해상도가 변화하는 신호를 해석하기에는 어려운 한계를 가지고 있다.

$$STFT(\tau<f)=\int x(t)w(t-\tau)e^{-j2\pi ft}dt \qquad (1)$$

이에 비하여, 웨이블렛 변환(Wavelet Transform)은 비주기적인 신호분리가 가능한 기저함수를 사용하여 신호를 해석하는 것으로, 웨이블렛은 1980년대 초에 이론으로 정립되기 시작한 이후 순수·응용과학 및 공학 등에서 급속히 발달하였고, 현재 여러 분야에 스며들어 그 응용성이 확대되고 있는 분야이다[14]. 이러한, 웨이블렛 변환은 "시간-주파수" 공간에 걸쳐 크기가 변환하는 함수 즉, 웨이블렛을 적용하여 신호의 부분적인 스케일 성분을 추출한다. 웨이블렛 변환의 기저함수로는 Daubechies, Coiflet, Haar, Symmlet등과 같은 웨이블렛 계열의 기저함수를 사용하고 있으며, 자료를 해석하는 해상도가 시간 축과 진폭 축에 따라 다양한 형태의 윈도우를 이용하여 분석하기 때문에 원 신호로부터 다양한 주기와 진폭을 갖는 패턴을 동시에 해석할 수 있는 장점을 갖고 있다. 또한, 직교변환으로 식 (2)와 같이 정의 할 수 있으며, 신호 $x(t)$ 에 대하여 다중 윈도우(multi window) 기능을 제공함으로써 다중분해능 해석을 가능하게 한다[15].

$$CWT_x(\tau,a)=\frac{1}{\sqrt{a}}\int x(t)h^*(\frac{t-\tau}{a})dt \qquad (2)$$

$$x(t)=c\int_{a>0}\int CWT_x(\tau,a)\,h_{a,\tau}(t)\frac{dad\tau}{a^2} \qquad (3)$$

식 (2)의 웨이블렛 변환은 식 (3)과 같은 역변환 식으로 나타낼 수 있다. 식 (2)와 (3)에서 a는 웨이블렛의 크기에 영향을 미치는 압축계수이고 τ는 시간상으로의 이동을 나타내는 전이계수이다. 웨이블렛 변환은 신호 $x(t)$에 대하여 기저 웨이블렛 $h(t)$을 크기변환 하거나 이동시킨 함수 $h((t-\tau)/a)$에 대하여 내적 한 것과 같은 기능을 가지고 있다.

이산 웨이블렛 변환은 고역 통과 부분을 한 단계의 필터뱅크로 구성하고, 저역통과 부분을 계속적인 필터 뱅크로 확장하는 옥타브 밴드(octave-band)구조와 고역 통과 부분도 필터뱅크로 확장하는 구조를 가지는 웨이블렛 패킷(wavelet packet)구조로 구현될 수 있으며, 그림 1과 그림 2는 옥타브 밴드 구조와 웨이블렛 패킷 구조를 보이고 있다[16][17]. 여기서 $g[n]$은 저역 통과 필터를 $h[n]$은 고역통과필터를 각각 나타내고 마더 웨이블렛으로 부터 구성됨을 알 수 있으며, $\downarrow 2$는 샘플의 개수를 1/2로 줄이는 데시메이션(decimation)을 나타낸다. 그림 2에서 웨이블렛 패킷 필터뱅크를 통해 나오는 출력신호 A4는 고주파의 고주파신호며, A3은 고주파의 저주파신호, A2는 저주파의 고주파신호, A1은 저주파의 저주파 신호 성분을 가지고 있으며, 각 대역에서 얻어진 신호에 대해 특징이 되는 신호를 선택적으로 융합하여 사용할 수

도 있다. 본 논문에서는 다양한 음성을 이용하여 감정인식 실험을 한 결과 우수한 성질을 보인 웨이블렛 패킷 구조를 적용하고자 한다. 또한, 웨이블렛 패킷의 출력인 A4, A3, A2, A1 대역을 이용하고자 웨이블렛 서브밴드 필터뱅크를 구현하였으며, 각 주파수 대역별로 분리된 음성 신호로부터 멜켑스트럼 계수를 특징벡터로 추출하였다.

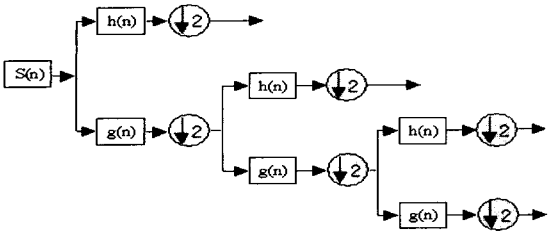

그림 1. 웨이블렛 옥타브 밴드의 구조
Fig 1. Structure of wavelet octave band

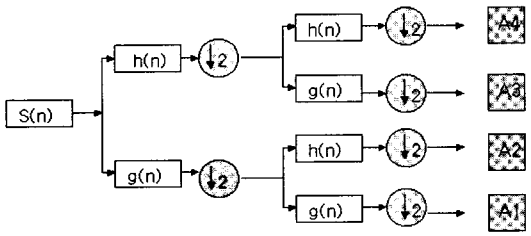

그림 2. 웨이블렛 패킷의 구조
Fig 2. Structure of wavelet packet

3. 웨이블렛 필터뱅크 기반 감정인식

본 연구에서는 주파수 대역을 균등하게 분할하는 방식인 웨이블렛 패킷구조방식과 기저함수 중 가장 널리 사용되고 있는 Daubechies를 사용하여 웨이블렛 필터뱅크 기반 감정인식 시스템을 구성하였다. 그림 3은 본 논문에서 제안한 감정인식 시스템 전체 구성도를 나타낸 것으로, 그림에서와 같이 시스템에 등록하는 과정은 감정이 섞인 음성을 입력으로 받아 음성 검출부를 통해 음성만을 검출하게 된다. 고립단어의 음성인 경우 한 단어의 음성이 묵음을 앞뒤로 하고 그 사이에 존재한다는 것을 전제로 하기 때문에 인식을 하기 위해서는 묵음으로부터 음성부분을 구별해 낼 필요가 있다. 이는 음성부분의 시작과 끝점을 찾아내는 과정으로 패턴인식의 문제로 생각할 수 있는데, 일반적인 방법으로 예측계수나 자기상관계수와 같은 음성특징 계수를 사용하는 방법이 있을 수 있다. 본 연구에서는 음성의 양 끝점을 검출하기 Rabiner와 Samburr에 의해 제안된 단시간 평균에너지(Short-time average energy)와 단시간 영교차율(Short-time zero crossing rate)을 사용한다.

음성 분석부에서는 음성 검출부를 통해 음성으로 분류된 부분을 웨이블렛 패킷구조방식을 이용하여 출력을 얻게 된다. 필터뱅크의 출력 수는 그림 2의 A4, A3, A2, A1와 같이 4개의 대역으로 나뉘어진 신호를 얻을 수 있으며, 이로부터 FFT기반 13차의 멜켑스트럼 계수를 특징벡터로 구할 수 있다. 그리고 음성 훈련부에서는 K-means 알고리듬을 이용하여 감정별 독립적인 코드북을 생성하는데 향상된 인식률을 얻고자 피치 분석 결과에 의한 남성화자와 여성화자를 구분하고 코드북을 각각 만든다. 이때, 여러 번의 실험에서 매우 낮은 인식률을 보인 고주파대역 즉, A4 고주파의 신호를 남자의 코드북에 대해서는 제외하였다.

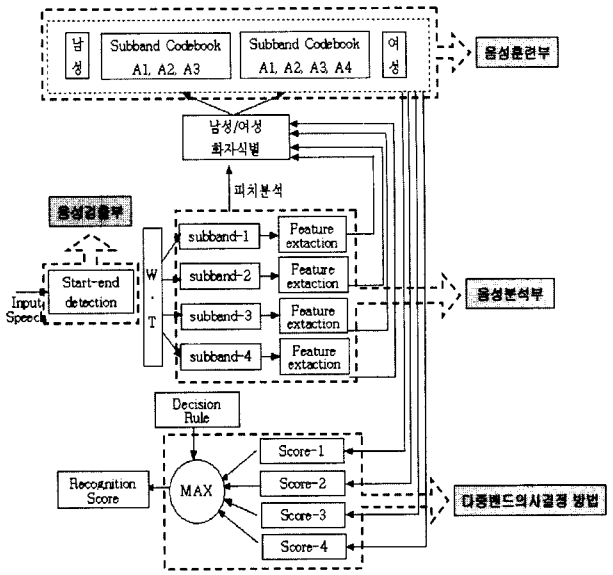

그림 3. 웨이블렛 필터뱅크를 이용한 감정인식기
Fig 3. Emotion recognition system using the wavelet filter banks

다음은 시스템에 등록되어 있는 코드북을 이용한 인식과정으로 입력으로 인식하고자 하는 음성이 들어오면 등록과정과 같이 음성 검출부, 음성 분석부를 통해 특징벡터를 추출하고 성별을 분리하여 만들어 놓은 코드북과 거리를 계산한 후 독립적인 감정 인식률을 산출한다. 각 대역별에서 산출된 인식률은 음성신호를 프레임으로 나누고 각각의 프레임에서 얻어진 특징벡터와 감정별 코드북과의 거리계산에 의하여 산출하였으며, 각 감정에 대한 소속도를 정규화 하기 위하여 각 감정의 선택된 프레임 수를 전체로 나누었다. 이는 어느 특정 감정에 대한 정보만을 가진 것이 아니라 6가지의 모든 감정에 대한 소속정도를 가지고 있어 다른 정보를 이용하는 데에도 사용할 수 있다.

최종 감정 인식단계에서는 각각의 소속도를 계산한 후, 계산된 소속도가 가장 높은 감정을 인식하게 되는데, 이때 사용된 다중 밴드 의사 결정 방법을 그림 4와 같이 나타내었다 [16]. 하나의 음성에 대해 4개의 밴드로 분리된 특징벡터는 각각의 밴드별로 6개의 감정에 대한 소속정도를 μ_{ij}으로 표현할 수 있으며, μ_{ij}는 4개의 서브밴드 $i(i=1,2,3,4)$에서 6개의 감정 $j(j=1,2,\cdots,6)$로 선택할 소속도를 나타낸다. 최종인식 단계에서는 각각의 소속도를 합산하여 나타낸 U_j $(j=1,2,\cdots,6)$, 즉, j의 감정으로 인식할 소속도를 얻을 수 있으며, 소속도가 가장 높은 U_j를 인식 결과로 선정한다. 본 논문에서는 6개의 기본감정(기쁨, 슬픔, 화남, 놀람, 공포, 혐오)을 사용하였으므로, 6개의 감정에 대한 소속도중 가장 높은 값을 갖는 감정을 인식하고자 하는 감정으로 삼는다.

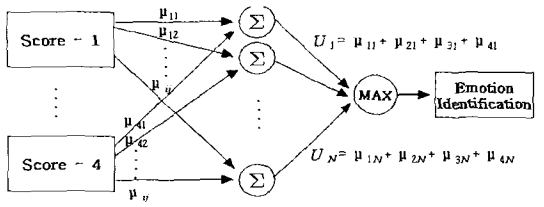

그림 4. 다중 밴드 의사결정 방법
Fig 4. Multiple band decision-making method

이와 같은 시스템은 일반적으로 인식대상의 특징벡터로 이루어지는 코드북 및 특징벡터의 종류에 따라 인식률에 큰 차이를 보인다. 일반적으로 코드북 사이즈가 클수록 인식률이 향상되지만, 인식속도의 저하 및 메모리상의 문제로 인하여 사전에 코드북의 적정 사이즈를 결정할 필요가 있다. 또한 인식대상을 잘 표현해 줄 수 있는 특징벡터의 선정도 중요한데, 본 연구에서는 일반적으로 사용되는 멜켑스트럼 계수 13차를 이용하여 특징벡터의 종류별 인식률을 조사하였다.

4. 실험 및 고찰

4.1 실험환경 및 음성 데이터 구성

본 논문에서 제안한 알고리즘의 유용성을 평가하기 위하여 인위적인 잡음을 발생하지 않는 환경(학교 연구실)에서 우리말 "아! 그렇습니까 ?"를 대상음성으로 녹음하여 실험하였다. 이를 위해 도구로는 MAGIC HEADSET SMH-M100을 사용하였으며, 음성파일의 형식은 샘플링 주파수 11kHz, 16bit, Mono를 사용하였으며, 웨이브 파일을 분석하기 위한 도구로는 상용 소프트웨어인 Cool Edit 2000을 사용하였다. 대상용어는 [12]에서의 실험 용어와 같은 것으로 남성화자 10명과 여성화자 10명이 각각 3회씩 발음하였으며, 음성 신호 중 2개는 학습, 또는 기준패턴을 만들기 위해서 사용하였고, 나머지 1개는 인식실험을 위해서 사용하였다.

기준패턴인 코드북의 사이즈는 64로 하고, 음성신호의 특징파라미터는 약 20ms 구간에서 음성신호가 정상(stationary)이라는 가정아래 20ms의 프레임 단위로 구하게 되나, 본 논문에서는 10ms의 Hamming window를 사용하고, 프레밍 양 끝단의 신호정보를 보상하기 위하여 5ms씩 중첩을 시켜서 윈도우를 이동시켰다. 이렇게 Hamming window를 사용하여 원 신호를 프레임 단위로 분할한 후 각각의 프레임에 포함된 데이터에서 13차의 멜켑스트럼 계수를 구하였다. 벡터 양자화 과정에서 음성의 시작점과 끝점을 정확하게 검출하는 것은 매우 중요한데, 본 논문에서는 Raviner와 Sambur에 의해 제안된 단시간 평균에너지(Short-time average energy)와 단시간 영교차율(Short-time zero crossing rate)을 이용한 알고리즘을 사용하였다.

음성의 시작점과 끝점을 검출한 후 음성의 고주파성분을 나타내기 위하여 일반적으로 $H(z)=1-0.95\,z^{-1}$과 같은 고역통과 필터를 이용한 전처리(preprocessing) 과정을 거치는데, 웨이블렛 기법을 이용하는 경우 이와 같은 전처리과정을 하면 원 신호가 가지고 있던 각각의 대역별 신호가 유실되기 때문에 사용하지 않았다.

4.2 실험결과

본 논문에서 제안한 웨이블렛 패킷구조를 갖는 웨이블렛 서브밴드 필터뱅크를 적용하여 남성화자 10명과 여성화자 10명의 음성신호를 대상으로 실험한 결과를 각 주파수 대역별로 구분하여 나타내었다.

우선, 표 1은 그림 2의 A4 대역의 출력으로부터 얻어진 특징벡터를 이용한 기쁨감정 인식을 10명(a~j)에 대해 나타낸 것으로, 음성신호를 프레임으로 나누고 각각의 프레임에서 얻어진 특징벡터와 코드북과의 거리계산에 의하여 산출된 인식률이다. 이때, 동일한 문장일지라도 사람의 특성에 따라 음성신호의 길이가 달라 질 수 있으므로 프레임의 수는 다르게 표현되며, 표의 결과 값은 대상 감정 프레임이 전체 음성의 프레임에서 차지하는 개수의 %를 나타낸 것이다. 표 1의 결과에서 알 수 있듯이 전체 음성의 프레임을 볼 때, 특정 감정에 대한 정보만을 가지지 않고 대상감정 이외의 감정에 대한 정보도 가지고 있음을 알 수 있다. 그러나 가장 많은 %를 차지하고 있는 것을 인식 대상 감정으로 한다.

표 2 A1 대역에 대한 기쁨 감정인식
Table 1. Emotion recognition with A1 subband

[단위 : %]

감정 / 대상자	기쁨	슬픔	화남	놀람	공포	혐오
a	35.2	9.5	12.8	21.7	7.6	13.2
b	25.6	21.8	16.3	11.9	9.9	14.5
c	29.9	14.3	16.4	18.9	6.2	14.3
d	20.2	7.9	14.9	20.2	25.0	11.8
e	38.7	9.9	17.6	6.4	6.7	18.7
f	45.4	14.4	13.7	10.2	6.4	9.9
g	28.9	8.4	14.8	14.8	7.2	25.9
h	38.2	15.0	18.1	7.8	3.4	17.5
i	30.3	10.6	19.7	13.6	7.0	18.8
j	35.5	12.8	16.0	9.2	8.0	18.5

표 2는 대역별, 그리고 남자, 여자 코드북 구분 여부에 따른 인식 결과로, 대역별로 살펴본 감정인식인 경우 고주파대역인 A4에서는 매우 낮은 인식률을 보인 반면에 저주파대역인 A1, A2, A3에서는 상대적으로 높은 인식률을 보이고 있다. 그러나 여자의 경우 감정을 표현할 때 상대적으로 높은 주파수를 가지는 음성을 표현하므로, 다중 밴드 의사결정 방법에서 A4 밴드를 사용하였으며, 남자의 경우 A4 밴드가 여자의 경우와 달리, 인식률을 저하시키므로 A4 밴드를 사용하지 않았다. 또한, 성별로 코드북을 구분하여 실험한 경우 구분하지 않고 실험한 경우보다 14% 정도 인식률이 향상된 것으로 나타났다. 이와 같은 이유는 남성과 여성 음성의 주파수 대역폭이 어느 정도 차이가 발생하기 때문에 성별로 구분하지 않고 작성할 경우 데이터의 분포 범위가 크기 때문에

표 2. 대역별 감정인식률 비교
Table 2. Emotion recognition rate according to each subband

[단위 : %]

성별 코드북 구분 유·무		Band				밴드 3개	밴드 4개	최종 인식률
		A1	A2	A3	A4			
유	남성	88	78	73	55	90	87	90
	여성	82	73	82	67	88	95	95
	평균	85	76	78	61	89	91	92.5
무	평균	73	65	62	54	74	79	79

최적의 코드북을 형성하는데 문제가 있는 것으로 볼 수 있다.

감정별 인식률을 알아보기 위하여 표 3은 남성화자 10명에 대해 각 감정별 인식하는 사람 수를 나타낸 것이고, 표 4는 여성화자 10명에 대해 감정별 인식하는 사람 수를 나타내었다. 그리고 표 5는 표 3과 표 4를 더한 것으로 최종인식결과를 감정별로 구분하여 나타냈다. 표 3~5에서 알 수 있는 바와 같이 "슬픔"에 관한 감정추출능력은 100%로서 매우 높은 인식률을 보인 반면 "행복"과 "놀람"에 관련된 감정추출능력은 90% 이하로 상대적으로 다른 감정보다 인식률이 저조함을 알 수 있다.

표 3. 남성화자에 대한 인식결과
Table 3. Recognition rate for male voice

[단위 : 명]

	기쁨	슬픔	화남	놀람	공포	혐오
기쁨		0	0	1	1	0
슬픔	0		0	0	0	0
화남	1	0		0	0	0
놀람	1	0	0		0	0
공포	0	0	0	1		0
혐오	0	0	1	0	0	

표 4. 여성화자에 대한 인식결과
Table 4. Recognition rate for female voice

[단위 : 명]

	기쁨	슬픔	화남	놀람	공포	혐오
기쁨		0	0	0	1	0
슬픔	0		0	0	0	0
화남	0	0		0	0	0
놀람	0	0	0		2	0
공포	0	0	0	0		0
혐오	0	0	0	0	0	

표 5. 인식결과 종합
Table 5. Total emotion recognition rate

[단위 : 명]

	기쁨	슬픔	화남	놀람	공포	혐오
기쁨		0	0	1	2	0
슬픔	0		0	0	0	0
화남	1	0		0	0	0
놀람	1	0	0		2	0
공포	0	0	0	1		0
혐오	0	0	1	0	0	

그림 5는 표 5에 대한 최종인식결과를 차트로 표현한 것으로, 평균적으로 92.5%의 인식률을 얻음으로서 음의 고저와 장단을 고려한 감정인식 방법[10]에서 보였던 89% 보다 높은 인식률을 얻었다. 하지만, 음성을 이용한 감정인식을 연구하는 경우 공인된 표준 데이터가 없음으로 상대적으로 결과를 비교하기에는 어려움이 있다.

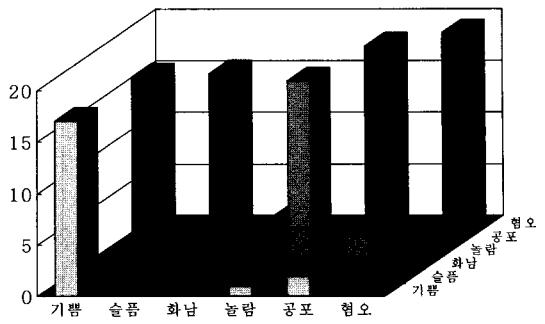

그림 5. 감정인식 결과 차트
Fig 5. Recognition rate for each emotion

5. 결 론

본 논문에서는 사람의 음성을 이용하여 감정을 인식할 수 있는 시스템을 구현하기 위해 웨이블렛 패킷구조를 기반으로 하는 웨이블렛 서브밴드 필터뱅크를 제안하고 구현하였다. 구현된 알고리즘은 음성신호를 서로 다른 대역으로 분리한 후 개별적인 대역별 인식 알고리즘을 수행하기 때문에 특정 웨이블렛 서브밴드에 노이즈 영향이 있더라도 다른 서브밴드에 영향을 미치지 않으므로 외부의 영향에 둔감한 좋은 성능을 보임을 알 수 있었다. 또한, 다중 밴드 의사 결정 방법을 이용하여 각각의 밴드별 6개의 감정에 대한 소속도 정보를 가짐으로서 보다 우수한 인식률을 얻을 수 있었다.

제안된 방법을 이용한 실험 결과를 보면, 고주파 대역보다 저주파 대역에서 높은 인식률을 얻을 수 있었으며, 성별을 구분하여 코드북을 작성한 경우 구분하지 않고 실험한 경우보다 14% 정도 인식률이 향상시킬 수 있었다. 또한, 분리된 성별에 대해 대역을 달리하여 코드북을 작성하였을 때에도 인식률을 향상시킬 수 있었다. 제안한 알고리듬을 사용하는 경우 인간의 감정 중 "슬픔"에 관한 감정추출능력은 100%로서 매우 높은 인식률을 보인 반면 "행복"과 "놀람"에 관련된 감정추출능력은 90% 이하로, 상대적으로 다른 감정보다 인식률이 저조함을 알 수 있었다. 향후에 여러 가지 언어적 표현과, 더 많은 음성 데이터를 확보하여 알고리즘의 일반성을 높이는 연구가 있어야 할 것으로 생각된다.

참고문헌

[1] C.L. Huang and Y.M. Huang, "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification", J. Visu Comm. And Image Representation, vol. 8, no. 3. 3, pp. 278-290, 1997

[2] M.J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of Singl Facial Images", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 12 pp. 1357-1362, 1999

[3] M. Pantic and L, Rothkrantz, "Automatic Analysis of Facial Expression : The sta of the Art", IEEE Trans. Pattern Analysis Machine Intelligence, Vol. 22, No 1424-1445, 2000

[4] Hyoun-Joo Go, Keun-Chang Kwak, Dae-Jong Lee, Myung-Geun Chun, "Emotion Recognition From the Facial Image and Speech Signal", SICE Annual Conference in Fukui, August 4-6, 2003
[5] Yaser Yacoob, Larry Davis, "Smiling Faces and Better for Face Recognition" Proceedings of the Fifth IEEE Intl Conf on Automatic Face and Gesture Recognition, 2002
[6] H. Kobayashi and F.Hara, "Facial Interaction between Animated 3D Face Robot and Human Beings", Proc. Intl Conf. Systems, Man, Cybernetics, pp.3732-3737, 1997
[7] Katsuhiro Matsuno and Saburo Tsuji, "Recogni zing human facial expressions in a potential field", In Proc. CVPR, pages 44-49, 1994
[8] P.Ekman and W.V. Friesen, "Emotion in the human face System", Cambridge University Press, San Francisco, CA, second edition, 1982.
[9] V.Kostov and S.Fukuda, "Emotion in User Interface", Voice Interaction System, IEEE Intl Conf. on Systems, Man, Cybernetics Representation, no. 2, pp. 798-803, 2000.
[10] T. Moriyama and S. Oazwa, "Emotion Recog nition and Synthisis System on Speech", IEEE Intl. Conference on Multimedea Computing and Systems, pages 840-844, 1999.
[11] L.C. Silva and P.C. Ng, "Bimodal Emotion Recognition", Proceeding of the 4th International Conference on Automatic Face and Gesture Recognition, pp. 332-335, 2000.
[12] 김이곤, 배영철, "퍼지 로직을 이용한 감정인식 모델설계", 한국퍼지 및 지능시스템 춘계학술대회, 2000.
[13] 심귀보, 박창현, "음성으로부터 감성인식 요소 분석" 퍼지 및 지능시스템학회 논문지, 2001.
[14] 강현배, 김대경, 서진근, "웨이블릿 이론과 응용", 대우학술총서, 2001
[15] 이승훈, 윤동한, "알기쉬운 웨이브렛 변환", 진한도서, 2002
[16] 이대종, 곽근창, 유정웅, 전명근, "웨이블렛 필터뱅크를 이용한 자동차 소음에 강인한 고립단어 음성인식" 퍼지 및 지능시스템학회 논문지, 2002.
[17] Stephane Mallat, "A wavelet tour of signal processing", Academic press, 1999.

저 자 소 개

고현주(Hyoun Joo Go)
1999년 : 한밭대학교제어계측공학과(학사)
2002년 : 충북대학교제어계측공학과 (공학석사)
2002년~현재 : 충북대학교 제어계측공학과 박사과정

관심분야 : Biometrics, Computer vision, 감정인식

이대종(Dae Jong Lee)
1995년 : 충북대학교 전기공학과(학사)
1997년 : 충북대학교 전기공학과(공학석사)
2002년 : 충북대학교 전기공학과(공학박사)
2003년~현재 : 충북대학교 컴퓨터정보통신연구소

관심분야 : 음성신호처리, 서명인식, 다중생체인식

박 장 환(Jang Hwan Park)
1991년 : 충북대 전기공학과 (학사)
1993년 : 충북대학교 전기공학과(공학석사)
1999년 : 충북대학교 전기공학과(공학박사)
현 재 : 충주대학교 정보제어공학과 계약교수(BK21)

관심분야 : 강인제어, 음성인식 및 확률계통 해석

전명근(Myung Geun Chun)
1987년 : 부산대학교 전자공학과(학사)
1989년 : 한국과학기술원 전기 및 전자공학과(공학석사)
1993년 : 한국과학기술원 전기 및 전자공학과(공학박사)
1993년~1996년 : 삼성전자 자동화연구소 선임연구원
2000년~2001년 : University of Alberta 방문교수
1996년~현재 : 충북대학교 전기전자 컴퓨터공학부 교수

관심분야 : Biometrics, 감정인식, 지능시스템