웨이블렛 변환을 이용한 음성에서의 감정인식

고현주*, 이대종**, 전명근*
*충북대학교 제어계측공학과
**충북대학교 전기공학과
e-mail : mgchun@cbucc.chungbuk.ac.kr

Emotion Recognition of Speech Using the Wavelet Transform

Hyoun-Joo Go*, Dae-Jong Lee**, Myung-Geun Chun*
Chungbuk National University
School of Electrical and Electronic Engineering

요약

인간과 기계와의 인터페이스에 있어서 궁극적 목표는, 인간과 기계가 마치 사람과 사람이 대화하듯 자연스런 인터페이스가 이루어지도록 하는데 있다. 이에 본 논문에서는 사람의 음성속에 깃든 6개의 기본 감정을 인식하는 알고리듬을 제안하고자 한다. 이를 위하여 뛰어난 주파수 분해능력을 갖고 있는 웨이블렛 필터뱅크를 이용하여 음성을 여러개의 서브밴드로 나누고 각 밴드에서 특징점을 추출하여 감정을 인식하고, 이를 최종적으로 융합, 단일의 인식값을 내는 다중의사 결정 구조를 갖는 알고리듬을 제안하였다. 이를 적용하여 실제 음성 데이타에 적용한 결과 기존의 방법보다 높은 90%이상의 인식률을 얻을 수 있었다.

1. 서론

우리가 일상 생활에서 사용하는 언어, 음성, 제스처 그리고 시각, 청각을 이용한 인간과 컴퓨터간의 인터페이스는 정보통신의 한 분야로서 활발히 연구가 진행되고 있다. 이와 같은 사람과 기계와의 인터페이스의 궁극적 목적은 사람과 사람이 대화하듯이 휴먼인터페이스가 이루어지는 것이다. 이러한 연구에 있어서 인간의 감정을 어떻게 측정 및 인식하느냐가 가장 어려운 문제로 부각되고 있다.

이에 음성은 청각에 기반을 둔 가장 효율적이고 자연스러운 휴먼 컴퓨터 인터페이스로 기대되고 있는 분야로, 심리학자인 Ekman과 Friesen에 따르면 사람의 여섯 가지 감정인 행복, 슬픔, 화남, 놀람, 혐오, 공포는 각 문화에 영향을 받지 않고 공통으로 인식되는 기본감정이라 하였다[1]. 이러한 인간의 기본적인 감정을 인식하기 위한 컴퓨터 인터페이스에 관한 연구는 최근 들어 큰 관심의 대상이 되었다.

감사의 글 : 본 연구는 정보통신부 대학기초연구지원사업에 의해 일부 지원 받았습니다.

사람의 목소리를 이용한 휴먼인터페이스를 위해 음성 속에 내포된 감정을 추출하려는 연구가 최근들어 활발히 행해지고 있는 요즈음, Fukuda는 음성신호의 템포와 에너지를 가지고 여섯 개의 기본감정에 대한 분류를 시도하였는데, 녹음실과 같은 외부잡음이 전혀 없는 환경하에서 일본어와 이탈리어에 대한 음성신호를 녹음한 후 감정을 추출하였다[2]. Moriyama는 음성신호의 피치(pitch)와 전력의 포락선 검출을 통하여 20개의 일본어 샘플에 대하여 실험하였고, 실험 결과는 '화남' '슬픔' '놀람'에 대하여 비교적 양호한 인식률 획득을 보고하고 있다[3]. 또한, Silva는 음성신호의 피치와 HMM(Hidden Markov Model)을 이용하여 영어와 스페인어에 대하여 감정추출을 실험하였다[4].

한편, 국내에서도 음성을 이용한 감정인식 연구가 활발히 진행되고 있는 요즈음, 우리나라 국악의 창에서 인간의 희로애락을 표현하는 음의 고저와 장단을 기본으로 하여 분석하는 연구가 행하여졌다[5].

앞에서 살펴본 바와 같이 외국의 경우, 음성을 이

용한 감정인식은 보통 대화의 내용에 사용한 단어나 음성신호의 피치(Pitch), 음성의 톤(Tone), 포만트 주파수(Formant Frequency), 말의 빠르기(Speech Frequency), 음질(Voice quality)등의 방법을 사용하여 연구한다. 그러나, 모국어인 한국어의 경우 지역간의 억양차이와 개개인의 특성에 따라 피치나 말의 빠르기가 다르기 때문에 앞에서와 같은 방법을 사용하여 감정인식을 하기에는 매우 어려운 상황이다.

따라서, 본 논문에서는 음성신호를 웨이블렛 서브밴드 필터뱅크를 이용하여 각 주파수 대역별로 음성신호를 분리한 후 각각의 대역에서 인식 알고리즘을 수행하였고, 다중밴드에서의 의사결정 방법을 이용하여 최종 감정 추출법을 구현하였다. 본 실험에 사용된 데이터로는 남성화자 6명과 여성화자 4명을 대상으로 앞서 말한 기본감정인 화남, 혐오, 행복, 슬픔, 공포, 놀람 등 6개의 감정에 대해서 실험하였다.

2. 웨이블렛 필터뱅크를 이용한 음성 감정인식

2.1 웨이블렛 신호해석

일반적으로 실생활에서 접하게 되는 대부분의 신호는 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.

웨이블렛 변환은 비주기적인 신호분리가 가능한 Daubechies, Coiflet, Haar, Symmlet 등과 같은 웨이블렛 계열의 기저함수를 사용하여 신호를 해석한다. 또, 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. 웨이블렛 변환은 직교변환의 일종으로서 식 (1)과 같이 정의 할 수 있으며, 시평면 신호 $x(t)$ 에 대하여 다중 윈도우(multi window) 기능을 제공함으로써 다중분해능 해석을 가능하게 한다.

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

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

식 (1)의 웨이블렛 변환은 식 (2)와 같은 역변환 식으로 나타낼 수 있으며, 웨이블렛 변환된 신호는 마더 웨이블렛을 크기변환 하거나 이동시킨 함수 $h((t-\tau)/a)$에 대하여 신호 $x(t)$를 내적한 것과 같

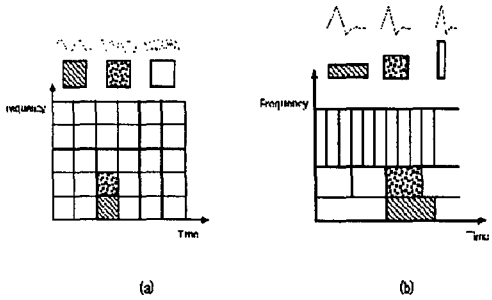

그림 1. 시간-주파수 평면에서 윈도우 형태
(a) 퓨리에 변환, (b) 웨이블렛 변환(Daubechies)

은 기능을 갖는다. 또한 웨이블렛 변환은 STFT(Short Time Fourier Transform)가 갖는 단점을 해소하기 위하여 신호의 특성에 따라 사용하는 위도우의 크기를 유동적으로 변화시킬 수 있는 기저함수를 사용한다. 즉, 그림 1과 같이 신호의 고주파 성분을 고립시켜 해석할 경우에는 주기가 짧은 기저함수를, 저주파 성분을 세밀하게 해석하고자 할 경우에는 주기가 긴 기저함수를 사용한다[6].

이산 웨이블렛 변환은 고역 통과 부분을 한 단계의 필터 뱅크로 구성하고, 저역통과 부분을 계속적인 필터 뱅크로 확장하는 옥타브 밴드(octave-vand) 구조와 고역 통과 부분도 필터뱅크로 확장하는 구조를 가지는 웨이블렛 패킷(wavelet packet)구조로 구현될 수 있다[7]. 그림 2에서는 옥타브 밴드 구조와 웨이블렛 패킷구조를 보이고 있는데, 여기서 g[n]은 저역 통과 필터를 h[n]은 고역통과필터를 각각 나타내며, 마더 웨이블렛으로 부터 구성됨을 알 수 있다. 또한 ↓2는 샘플의 개수를 1/2로 줄이는 데시메이션(decimation)을 나타낸다.

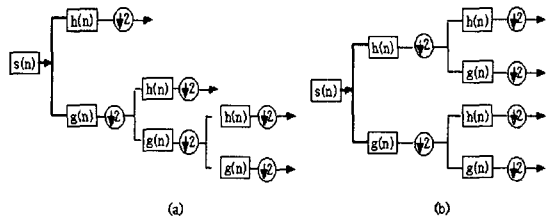

그림 2. 웨이블렛의 구조
(a) 옥타브 밴드 (b) 웨이블렛 패킷

2.2 웨이블렛 필터뱅크 기반 감정인식

본 연구에서는 주파수 대역을 균등하게 분할하는 방식인 웨이블렛 패킷구조방식을 사용하였으며, 필

터뱅크의 출력 수는 그림 2-(b)와 같은 4개의 필터뱅크로 구성되었고, 이 중 다양한 실험에서 매우 낮은 인식률을 보인 1개의 고주파대역을 제외하고자 한다. 그리고 가장 널리 사용되고 있는 Daubechies 기저함수를 이용하여 신호를 해석하였다.

3개의 저주파 필터에서 출력되는 음성신호는 음성분석부에서 특징벡터로 FFT기반 멜켑스트럼 계수를 구한 후 K-means 알고리듬을 이용하여 독립적인 코드북을 미리 만들어 놓았다. 이 때, 감성인식률 향상을 위하여 남성화자와 여성화자용 코드북(Codebook)을 각각 만들었다.

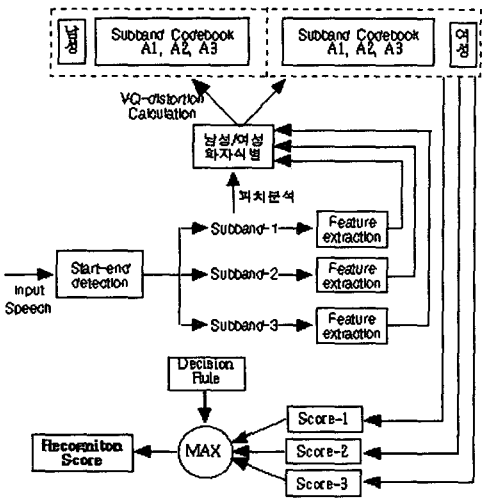

그림 4 웨이블렛 필터뱅크를 이용한 감정인식기

인식과정에서는 그림 3에서 보는 바와 같이 인식하고자 하는 음성신호가 입력되면 웨이블렛 변환하여 주파수별로 음성신호를 분할한다. 그리고 성별을 구분하여 만들어 놓은 코드북과의 비교하기 위해 저주파대역에서 피치를 이용한 성식별을 분석한 후 음성 분석부에서 각각의 주파수 대역에 대한 특징벡터를 계산한다. 이와 같이 음성 분석부에서 계산된 특징벡터는 미리 뱅크별로 만들어 놓은 코드북과의 거리를 계산한 후 독립적인 인식률을 산출한다. 여기서, 각 대역별에서 산출된 인식률은 음성신호를 프레임으로 나누고 각각의 프레임에서 얻어진 특징벡터와 코드북과의 거리계산에 의하여 산출되기 때문에 어느 특정 감정에 대한 정보만을 가진 것이 아니라 인식하고자 하는 각각의 감정들에 대한 소속정도를 모두 가지고 있다. 이러한 각 감정에 대한 소속도를 정규화 하기 위하여 각 감정의 선택된 프레임 수를 총 프레임으로 나누었다. 마지막으로 최종 감정 인식단계에서는 각각의 소속도를 계산한 후, 그 중에서 소속도가 가장 높은 감정을 인식하게 된다.

3. 실험 및 검증

3.1 실험환경 및 음성 데이터 구성

본 논문에서 제안한 알고리즘의 유용성을 평가하기 위하여 잡음이 억제된 상황에서 녹음된 우리말 "아! 그렇습니까 ?"를 대상으로 실험하였다. 이는 [5]에서의 실험 용어와 같은 것으로 남성화자 6명과 여성화자 4명이 각각 3회씩 발음하였으며, 음성 신호중 2개는 학습, 또는 기준패턴을 만들기 위해서 사용하였고, 나머지 1개는 인식실험을 위해서 사용하였다.

녹음된 음성데이터의 샘플링 주파수는 11.025kHz이며, 기준패턴인 코드북의 사이즈는 32로 하였다. 음성신호의 특징파라미터는 약 20ms 구간에서 음성신호가 정상(stationary)이라는 가정아래 20ms의 프레임 단위로 구하게 된다. 그러나 본 논문에서는 10ms의 Hamming window를 사용하고, 프레밍 양끝단의 신호정보를 보상하기 위하여 5ms씩 중첩을 시켜서 윈도우를 이동시켰다. 이렇게 Hamming window를 사용하여 원 신호를 프레임 단위로 분할한 후 각각의 프레임에 포함된 데이터에서 14차의 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(pre-emphasis) 과정을 거치는데, 웨이블렛 기법을 이용하는 경우 이와 같은 전처리과정을 하면 원 신호가 가지고 있던 각각의 대역별 신호가 유실되기 때문에 사용하지 않았다.

3.2 실험결과

표 1에서는 본 논문에서 제안한 웨이블렛 패킷구조의 웨이블렛 기법을 적용하여 남성화자 6명과 여성화자 4명의 음성신호를 대상으로 실험한 결과를 각 주파수 대역별로 구분하여 나타냈다. 표 1에서 알 수 있는 바와 같이 감정인식인 경우 고주파대역인 A4에서는 매우 낮은 인식률을 보인 반면에 저주파대

역인 A1, A2, A3에서는 상대적으로 높은 인식률을 보이고 있다. 성별로 코드북을 구분하여 실험한 경우 구분하지 않고 실험한 경우보다 15% 정도 인식률이 향상된 것으로 나타났다. 이와 같은 이유는 남성과 여성 음성의 주파수 대역폭이 어느 정도 차이가 발생하기 때문에 성별로 구분하지 않고 작성할 경우 데이터의 분포도 범위가 크기 때문에 최적의 코드북을 형성할 수 없기 때문이라 할 수 있다.

[단위 : %]

성별 코드북 구분	유·무	Band A1	A2	A3	A4	Decision Rule
유	남성	94	75	86	61	92
	여성	79	83	92	71	88
	종합	87	79	89	66	90
무	종합	95	68	63	47	75

표 1. 대역별 감정인식률 비교

	행복	슬픔	화남	놀람	공포	혐오
행복	**5**	0	0	1	0	0
슬픔	0	**6**	0	0	0	0
화남	0	1	**5**	0	0	0
놀람	0	0	0	**6**	0	0
공포	0	0	0	0	**6**	0
혐오	0	0	0	0	1	**5**

표 2. 남성화자에 대한 인식결과

	행복	슬픔	화남	놀람	공포	혐오
행복	**3**	0	0	0	1	0
슬픔	0	**4**	0	0	0	0
화남	0	0	**3**	0	0	1
놀람	1	0	0	**3**	0	0
공포	0	0	0	0	**4**	0
혐오	0	0	0	0	0	**4**

표 3. 여성화자에 대한 인식결과

	행복	슬픔	화남	놀람	공포	혐오
행복	**8**	0	0	1	1	0
슬픔	0	**10**	0	0	0	0
화남	0	1	**8**	0	0	1
놀람	1	0	0	**9**	0	0
공포	0	0	0	0	**10**	0
혐오	0	0	0	0	1	**9**

표 4. 인식결과 종합

감정별 인식률을 알아보기 위하여 표 2와 3에서는 남성화자와 여성화자에 대한 각각의 인식결과를 나타냈었고, 표 4에서는 최종인식결과를 감정별로 구분하여 나타냈다. 표 2~4에서 알 수 있는 바와 같이 "슬픔"과 "공포"에 관련된 감정추출능력은 100%로서 매우 높은 인식률을 보인 반면 "행복"과 "화남"에 관련된 감정추출능력은 80%로서 상대적으로 다른 감정보다 인식률이 저조함을 알 수 있다. 그리고 사람의 음성을 이용한 감정인식방법에 대해 연구한 [5] 논문에서의 전체 인식율 89% 보다 우수한 성능을 보였을 뿐 만 아니라 간단한 알고리즘과 수행시간 또한 단축할 수 있었다.

4. 결론

본 논문에서는 음성을 이용한 감정인식을 향상시키기 위해 웨이블렛 패킷구조를 기반으로 하여 각각의 주파수 대역별로 분할한 후 벡터양자화 기법을 이용하여 개별적인 인식률을 조사한 후 최종적으로 감정인식을 결정하는 방법에 대해 연구하였다.

실험결과, 음성을 이용한 감정추출인 경우 고주파 대역보다 저주파대역에서 인식률이 높게 나타났으며, 성별로 코드북을 구분하여 실험한 경우 구분하지 않고 실험한 경우보다 15% 정도 인식률이 향상된 것으로 나타났다. 또한, 인간의 감성중 "슬픔"과 "공포"에 관련된 감정추출능력은 100%로서 매우 높은 인식률을 보인 반면 "행복"과 "화남"에 관련된 감정추출능력은 80%로서 상대적으로 다른 감정보다 인식률이 저조함을 알 수 있었다.

참고문헌

[1] P.Ekman and W.V. Friesen. *Emotion in the human face System.* Cambridge University Press, San Francisco, CA, second edition, 1982.

[2] V.Kostov and S.Fukuda, Emotion in User Interface, Voice Interaction System, IEEE In시 Conf. on Systems, Man, Cybernetics Representation, no. 2, pp. 798-803, 2000.

[3] T. Moriyama and S. Oazwa, Emotion Recognition and Synthisis System on Speech IEEE In시. Conference on Multimedea Computing and Systems, pages 840-844, 1999.

[4] 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.

[5] 김이곤, 배영철, "퍼지 로직을 이용한 감정인식 모델설계", 한국퍼지 및 지능시스템 춘계학술대회, 2000.

[6] 이대종, 곽근창, 유정웅, 전명근, "웨이블렛 필터뱅크를 이용한 자동차 소음에 강인한 고립단어 음성인식" 퍼지 및 지능시스템학회 논문지, 2002, 4월

[7] Stephane Mallat, A wavelet tour of signal processing, Academic press, 1999.