감정 인식을 위한 음성신호 비교 분석

조동욱*, 김봉현**, 이세환**
*충북과학대학 정보통신과학과
**한밭대학교 컴퓨터공학과
e-mail:ducho@ctech.ac.kr

Comparison and Analysis of Speech Signals for Emotion Recognition

Dong-Uk Cho*, Bong-Hyun Kim**, Se-Hwan Lee**
*Dept of Information & Communication Science, Chungbuk Provincial University of Science & Technology
**Dept of Computer Engineering, Hanbat National University

요 약

본 논문에서는 음성 신호로부터 감정의 특징을 나타내는 요소를 찾아내는 것을 목표로 하고 있다. 일반적으로 감정을 인식할 수 있는 요소는 단어, 톤, 음성신호의 피치, 포만트, 그리고 발음 속도 및 음질 등이 있다. 음성을 기반으로 감정을 익히는 방법 중에서 현재 가장 많이 접근하고 있는 방법은 피치에 의한 방법이 있다. 사람의 경우는 주파수 같은 분석 요소보다는 톤과 단어, 빠르기, 음질로 감정을 받아들이게 되는 것이 자연스러운 방법이므로 이러한 요소들이 감정을 분류하는데 중요한 요소로 쓰일 수 있다. 따라서, 본 논문에서는 감정에 따른 음성의 특징을 추출하기 위해 사람의 감정 중에서 비교적 자주 쓰이는 평상, 기쁨, 화남, 슬픔에 관련된 4가지 감정을 비교 분석하였으며, 인간의 감정에 대한 음성의 특성을 분석한 결과, 강도와 스펙트럼에서 각각의 일관된 결과를 추출할 수 있었고, 이러한 결과에 대한 실험 과정과 최종 결과 및 근거를 제시하였다. 끝으로 실험에 의해 제안한 방법의 유용성을 입증하고자 한다.

1. 서론

정보화 사회의 급격한 발달로 고기능의 개인용 컴퓨터들이 각 가정으로 확산되어 보급됨이 따라 인간과 컴퓨터의 상호작용은 능동적인 양방향성 인터페이스로 변화 되어가면서 좀 더 자연스러운 사용성에 대해 진보되고 있으며, 쉬운 형태로 발전하고 있다[1]. 인간은 일반적으로 시각, 청각, 촉각 등을 다양한 방법을 통하여 상호간에 정보를 교환한다. 감정/감정의 전달 또한 이와 같은 방식으로 전달된다고 생각하는데[2], 이러한 휴먼 인터페이스 기술은 사용자의 감정 상태를 추출, 인식하는 것을 목적으로 하고 있다. 사용자의 감정 상태에 대한 인식을 설계하기 위한 도구로는 언어, 음성, 제스처, 시각, 청각 등을 이용하고 있다[1].

감정 인식에 대한 Chan etal 의 연구 결과에 의하면, 감정의 6가지 기본 요소인 행복, 슬픔, 분노, 증오, 놀람, 두려움을 음성모델과 시각 모델로 분류하여 놓고 음성 모델만으로 알아본 인식률은 75%, 시각 모델만으로 수행된 인식률은 70%라는 결과를 각각 얻었다. 그리고 음성과 시각 모델을 함께 표현하여 얻은 인식률은 97%에 이르렀다고 한다.

<표 1> Chan etal 의 감정인식 연구 결과

	음성	시각	음성+시각
인식률	75%	70%	97%

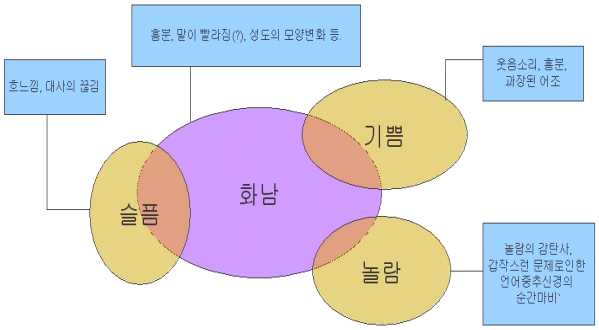
본 논문에서는 6 가지 기본 감정을 바탕으로 평상시, 기쁠 때, 화날 때, 슬플 때의 음성을 추출해 감정에 따른 음성의 변화를 비교·분석하고자 한다. 논문의 구성은 2장에서 음성을 기반으로 감정을 인식하는 방법에 대한 설명을 하였고, 3장에서 연구 환경 및 직접 추출한 음성 데이터를 비교 분석하여 실험 및 고찰을 설명하였다. 마지막으로 4장에서는

본 실험의 결과를 토대로 기존과의 차이점을 논하였고, 차후에 연구되어질 방향을 제안하고자 한다.

2. 음성에서의 감정 인식

2.1 음성의 특징

일반적으로 음성신호로부터 사람의 감정을 인식할 수 있는 특징 요소에는 대화의 내용에 사용한 단어, 톤(Tone), 음성신호의 피치(Pitch), 포만트 주파수(Formant Frequency), 그리고 발음 속도(Speech Speed), 음질(Voice Quality) 등이 있다[2]. 사람의 경우는 주파수 같은 분석요소 보다는 톤과 단어, 빠르기, 음질로 감정을 받아들이게 되는 것이 자연스러운 방법이므로 이러한 요소들이 감정을 분류하는데 중요한 인자로 쓰일 수 있다.



(그림 1) 인간의 감정 인식 요소

음성의 개인정보는 음성의 질, 높이, 강도, 속도, 템포, 억양, 악센트, 어휘의 사용 등에 따라 다르게 나타난다. 이들 특성들은 각종 물리적 특징들이 복잡한 상호작용을 거쳐 나타나는데, 성대의 길이, 성대 특성 등과 같이 선천적으로 타고나는 조음 기관의 개인적인 차이로부터 나타나며 말하는 습성 등으로부터도 나타난다. 또한, 각 개인의 가장 중요한 청각 정보인 음성의 질과 높이는 스펙트럼 포락선과 기본 주파수(피치)의 정적, 순시적 특성에 의존한다. 현재까지 화자의 감정을 반영하는 요소로서 발음 속도, 피치 평균, 피치 변화 범위, 발음 세기, 음질, 피치의 변화, 발음법 등의 파라미터가 감정 인식 및 합성에 주로 사용되어 오고 있다[3][4][5].

음성 파라미터는 음성신호의 단 구간에 대해 구한 피치와 에너지 값으로부터 피치 평균, 피치 표준편차, 피치 최댓값, 에너지 평균, 에너지 표준편차 등의 통계적 정보를 산출하여 감성 인식을 위해 사용한다. MFC(Mel Frequency Cepstrum Coefficient) 파라미터는 음소의 특성을 나타낸다. 같은 음소라도 포함된 감정에 따라 음소의 형태가 다르다는 점에서 감정 인식에 사용될 수 있다[6].

2.2 음성 분석학적 요소

본 논문에서는 음성 신호로부터 감정 특징을 나타내는 요소를 찾아내는 것을 목표로 하고 있다. 음성을 기반으로 특징을 추출하는 방법에는 여러 가지가 있다. 대표적인 요소들을 살펴보면 첫째로, 피치(Pitch)에 의한 방법이다. 피치는 사람이 귀로 들을 때의 음의 높낮이를 말하거나 준 주기적인 파형을 나타내는 유성음의 1 주기를 뜻한다. 보통, 인간의 언어에 사회성이 있듯이, 그 언어에 실리는 감정 또한 사회성을 갖고 있기 때문에 자신의 감정을 타인에게 정확히 알리기 위해선 보편적으로 인정되는 음의 높낮이를 보여야 한다. 이러한 관점에서 보면 거꾸로 음의 높낮이로부터 감정을 알아낼 수 있다는 의미를 내포하고 있다.

또한, 포만트 주파수(Formant Frequency)란 부분음 중에서 어느 특정 배음들이 강화되는 위치의 주파수를 말하고, 그 부근의 부분까지 포함해서 포만트(Formant)라고 한다. 그러므로, 주파수의 변화가 에너지 분포의 변화로 연결되고 또한 감정의 변화가 생기면 신체적/생리적인 변화가 발생하고 포만트의 변화가 연쇄적으로 일어나기 때문에 이들의 분석도 필수적이다[2][7].

3. 실험 및 고찰

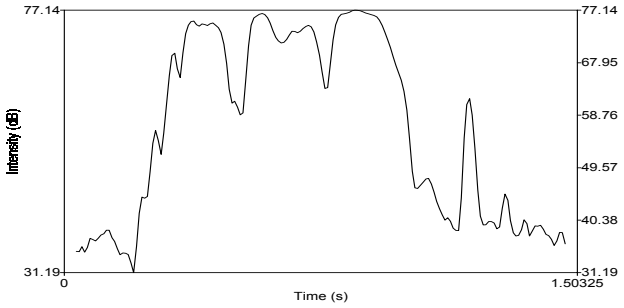
본 논문의 실험을 위하여 운영체제는 Microsoft Windows XP Professional 상에서 수행하였다. 음성 데이터 수집 장치로는 삼성 Voice Yepp을 사용하였고, 음성 비교 분석을 위해 사용한 도구(tools)는 Praat 4.2.07[8][9]를 사용하였다.

본 논문에서는 음성신호의 피치(Pitch), 포만트 주파수(Formant Frequency), 스펙트럼(Spectrum), 강도(Intensity)의 모든 부분에서의 각 하위범주의 특징을 분석하기 위하여 실험을 진행하였다. 감정에 따른 음성의 특징을 분석하기 위해 9명의 실험자를 대상으로 평상시, 기쁠 때, 화날 때, 슬플 때의 데이터를 추출하는 작업을 하였다. 그 중 2명의 여성 실험 데이터(실험자 H, I)에서 나머지 음성 데이터와는 달리 약간의 차이를 보였는데, 이는 2명의 여성 데이터가 사상체질의 진단 및 분류에서 소양인이라는 특성을 나타내었다. 이를 계기로 추후에 음성 분석을 통한 감정 인식에서 체질에 따른 변화를 연구할 계획이다.

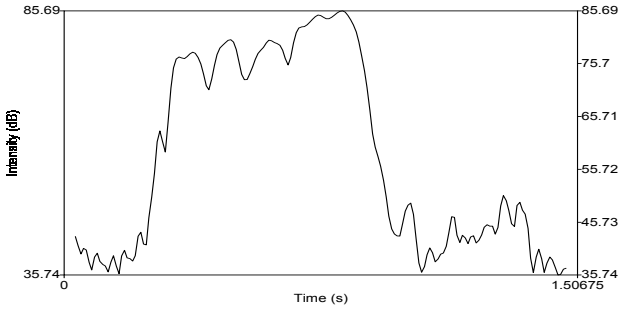
<표 2> 피실험자 정보

피실험자	성별	나이	직업
A	남	61	직장인
B	여	25	직장인
C	남	25	학생
D	남	26	학생
E	여	20	학생
F	남	26	학생
G	남	27	학생
H	여	23	학생
I	여	22	학생

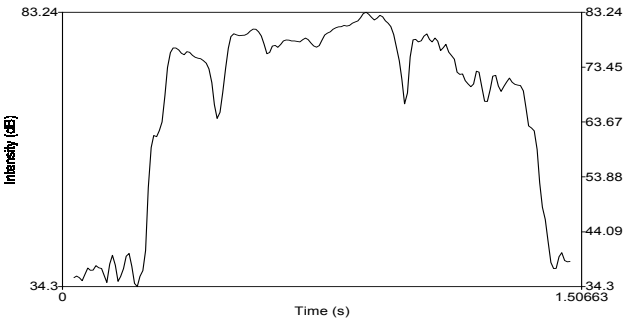
실험의 정확성을 위해 음성 데이터를 수집하는 과정에서 보다 실제와 근접한 감정의 음성 데이터를 추출해내기 위하여 각각 감정과 관련된 이야기책을 준비하였고, 실험 대상자들에게 이야기책을 읽게 한 후 감정이 어느 수준까지 올라와 있을 것이라고 판단될 즈음에 그 감정에 해당하는 음성을 녹음하는 방법으로 데이터를 추출 하였다. 음성을 녹음하는 과정에서는 화자의 입과 음성 저장 장치와 거리를 약 10cm의 간격을 유지하여 녹음을 하였으며, 일관성 있는 데이터의 분석을 위하여 실험자 모두 “감사합니다.” 라는 단어를 2초 이내에 말하게 하여 음성을 추출하였다. 강도(Intensity)의 측정 결과는 그림과 같으며, 인간의 감정이 화났을 때 보다 오히려 기쁠 때의 강도 파형 측정값이 가장 높이 나타났다.



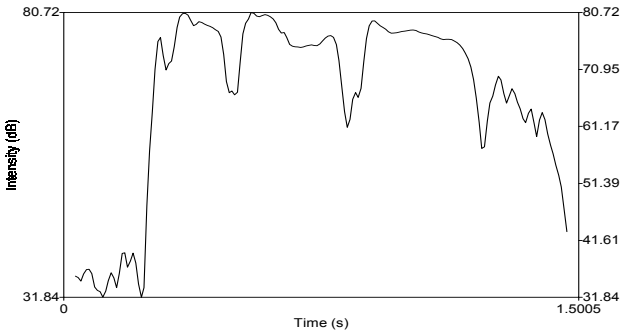
(그림 2) 피실험자 A 의 평상시 강도



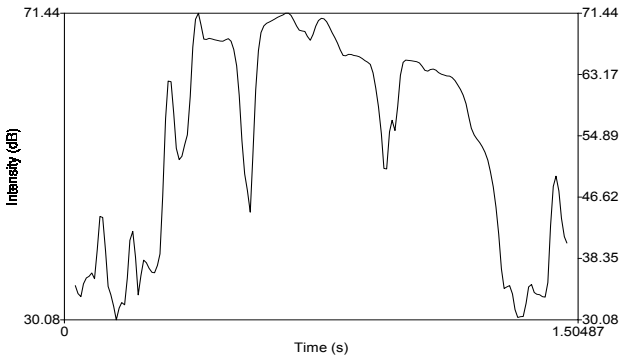
(그림 3) 피실험자 A 의 기쁠 때 강도



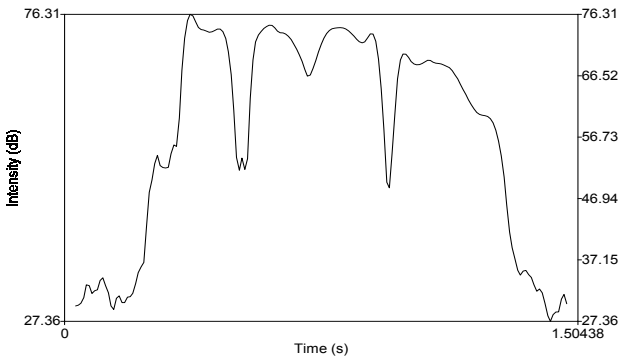
(그림 4) 피실험자 A 의 화날 때 강도



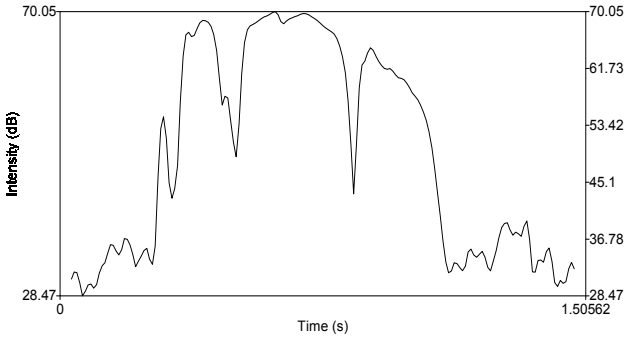
(그림 5) 피실험자 A 의 슬플 때 강도



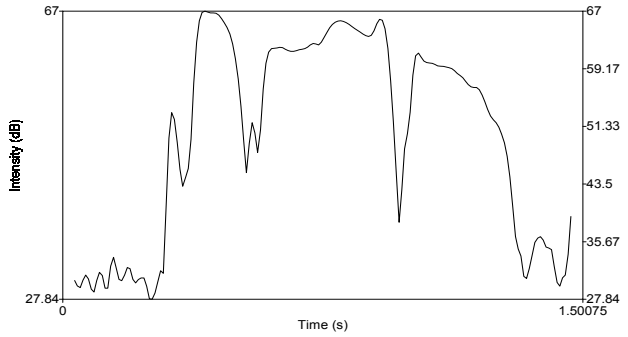
(그림 6) 피실험자 B 의 평상시 강도



(그림 7) 피실험자 B 의 기쁠 때 강도



(그림 8) 피실험자 B 의 화날 때 강도



(그림 9) 피실험자 B 의 슬플 때 강도

<표 3> 감정에 따른 강도 측정 값 비교

(단위 : Hz)

피실험자	기쁠 때	평상시	화날 때	슬플 때
A	85.69	77.14	83.24	80.72
B	76.31	71.44	70.05	67.00
C	87.78	77.51	78.44	73.33
D	79.85	79.12	75.55	76.71
E	89.49	79.58	87.52	78.99
F	81.76	72.89	80.22	76.32
G	86.44	79.79	76.60	76.50
H	84.92	83.76	85.83	79.10
I	83.48	76.87	84.78	80.93

4. 결론

본 논문에서는 음성 신호로부터 감정 특징을 나타내는 요소를 찾아내는 것을 목표로 하였다. 기존의 이러한 연구들은 통상적으로 연기자에게 연기를 요구하는 방식이거나 TV 드라마를 통해 샘플을 얻는 방식으로 각각 상황에 맞는 다른 뜻의 단어들 즉, 일관성이 없는 단어들을 사용하여 연구를 하였는데, 본 논문에서는 위에서 언급한 것처럼 사람들에게 각 감정별 이야기가 담긴 이야기책을 읽게 하였고, 동일한 단어를 사용하게 하여 보다 정확한 감정의 비교 분석에 대한 통계를 낼 수 있도록 노력하였다. 이러한 연구를 위해서는 우선적으로 주위의 잡음을 최소화 시켜야 한다.

본 논문에서는 주위의 잡음을 최소화 시킬 수 있는 적절한 장소가 미흡했던 점이 많은 아쉬움으로 남았으며, 체질과 실제의 감정 상태를 표현할 수 있는 상황적인 설정의 부족함이 연구의 보완점으로 나타났다. 향후 이러한 사항들을 보완하여 다음 논문에서는 보다 더 정확한 분석 통계자료를 얻어야 할 것이다. 이러한 연구들이 현재 많은 곳에서 활발하게 진행되고 있다. 하지만, 아직까지도 인간에 비교할 만한 완전한 감정 지능형 컴퓨팅 기술은 개발되지 않고 있는 것으로 조사된다[3][4]. 이번 논문의 성과와 더불어 현재 활발히 진행되고 있는 다른 연구 결과를 토대로 더 나아가 인간의 감정에 지능적으로 대앙할 수 있는 최종의 목표인 지능형 의료 진단기기 시스템 개발[10][11]이 본 논문에서 지향해야 할 목표이다.

참고문헌

[1] 고현주, 이대종, 전명근, “얼굴표정과 음성을 이용한 감정인식”, 한국정보과학회 논문지, 31권, 6호, pp 799-807, 2004.

[2] 심귀보, “음성으로부터 감정인식 요소분석”, 퍼지 및 지능시스템학회 논문지, 11권, 6호, pp 510-515. 2001.

[3] 김정한, “음성에서의 감정인식”, 성균관대

[4] http://hci.skku.ac.kr/

[5] 김원구, “음성신호를 사용한 감정인식의 특징 파라메터 비교”, 대한전자공학회, 2005.

[6] “음향적 요소분석과 DRNN을 이용한 음성신호의 감정 인식”, 퍼지 및 지능시스템학회 논문지, 13권, 1호, pp. 45-50.

[7] 박경범, ‘선형예측분석법에 의한 음성의 압축과 재생’, 도서출판 하늘소, pp. 53-60, 1994.

[8] 양병곤, ‘프라트(praat)를 이용한 음성분석의 이론과 실제’, 만수출판사, 2003.

[9] http://www.fonetiks.info/

[10] 성하경, “HCI 및 감정로봇을 위한 오감 인식 기술 동향”, 전자공학회지, 28권, 12호, pp. 26-31, 2001.

[11] 권기상, 이필규 “표정에 강인한 가보 웨이블릿 기반 얼굴인식에 대한 연구”, 한국정보과학회 2004 추계학술대회, 31권, 2호, pp.724-726, 2004.