음성으로부터 감성인식 요소 분석

Analyzing the element of emotion recognition from speech

박창현 · 심재윤 · 이동욱 · 심귀보
중앙대학교 전자전기공학부

Chang-Hyun Park, Jae-Yoon Sim, Dong-Wook Lee, and Kwee-Bo Sim
School of Electrical and Electronic Engineering, Chung-Ang University
E-mail : kbsim@cau.ac.kr

ABSTRACT

일반적으로 음성신호로부터 사람의 감정을 인식할 수 있는 요소는 (1)대화의 내용에 사용한 단어, (2)톤(Tone), (3)음성신호의 피치(Pitch), (4)포만트 주파수(Formant Frequency), 그리고 (5)말의 빠르기(Speech Speed) (6)음질(Voice Quality) 등이다. 사람의 경우는 주파수 같은 분석요소 보다는 톤과 단어, 빠르기, 음질로 감정을 받아들이게 되는 것이 자연스러운 방법이므로 당연히 후자의 요소들이 감정을 분류하는데 중요한 인자로 쓰일 수 있다. 그리고, 종래는 주로 후자의 요소들을 이용하였는데, 기계로써 구현하기 위해서는 조금 더 공학적인 포만트 주파수를 사용할 수 있게 되는 것이 도움이 된다. 그러므로, 본 연구는 음성 신호로부터 피치와 포만트, 그리고 말의 빠르기 등을 이용하여 감성 인식시스템을 구현하는 것을 목표로 연구를 진행하고 있는데, 그 1단계 연구로서 본 논문에서는 화가 나서 내뱉는 말과 기쁠 때 간단하게 사용하는 말들을 기반으로 하여 극단적인 두 가지 감정의 독특한 특성을 찾아낸다.

Keywords : 톤(Tone), 피치(Pitch), 포만트 주파수(Formant Frequency), 음질, 인두강

Ⅰ. 서 론

인간은 여러 가지 방법으로 상호간에 의견을 교환한다. 시각적, 청각적, 촉각적 방법 등의 방법이 일반적이다. 감정의 전달 또한 같은 방식으로 전달된다. 단순히, 말과 문자만 가지고도 상호간에 의사소통은 가능하다. 하지만 간단한 예로써 화난 상태에서 "가라" 고 하는 것과 그냥 아무런 감정상태 없이 "가라" 고 하는 것에는 청자의 행동방향에 차이가 생길 수밖에 없다. 이렇듯 대인관계에서 감정을 인식하느냐 못하느냐로 상대의 진정한 의도파악 여부가 결정된다. 그러므로, 감정의 인식이 말, 문자와 같이 중요한 의사소통 수단이다. 그런 의미에서 감성인식의 연구 또한 필연적이다.

Chan[1][3]에 의해 수행된 감성인식에 대한 연구를 살펴보면, 6가지 기본적인 감정인 행복, 슬픔, 분노, 증오, 놀람과 두려움을 음성모델과 시각 모델로 분류하여 놓고 음성모델만으로 알아본 인식률은 75%, 시각모델만으로 수행된 인식률은 70%라는 각각의 결과를 얻었다. 그리고 음성과 시각모델을 함께 표현하여 얻은 인식률은 97%에 이르렀다. 위의 연구를 통해서 음성을 통한 인식이 시각적 인식보다 조금 더 효과적이라는 것을 알 수 있고, 좀 더 높은 인식률을 위해서는 시각과 청각이 함께할 때라는 것을 알 수 있다.

본 논문에서는 감성 인식기를 만들기 이전에 음성 신호로부터 감성 특징을 나타내는 요소를 찾아내는 것을 목표로 하고 있다. 음성을 기반으로 감정을 인식하는 방법으로는 (1)대화의 내용에 사용한 단어, (2)톤(Tone), (3)음성신호의 피치(Pitch), (4)포만트 주파수(Formant Frequency), 그리고 (5)말의 빠르기(Speech Speed) (6)음질(Voice Quality) 등이 있지만 그 중에서 현재 가장 많이 접근하고 있는 방법은 피치(Pitch)에 의한 방법이다.

피치[5]는 사람이 귀로들을 때의 음의 높낮이를 말하거나 준 주기적(Quasiperiodic)인 파형을 나타내는 유성음의 1주기를 뜻한다. 감정의 경우 그 심리적 변화가 생리적/신체적 변화로 이어져 음성 스펙트럼에 변화가 오므로 주파수의 변화, 피치의 변화로 귀결된다.

실제로, 사회화 과정에서 감정표현도 학습되어 감정의 표현이 인간의 원초적인 성질에 의존하지 않고 상호간에 인정되는 수준에서 유사한 패턴을 보이고 있고, 가장 분석이 쉬운 화난 감정의 경우 표현 방식의 분류와 그 범주 내에서 보이는 일반적인 성향을 보인다[2].

그리고 포만트 주파수란 부분 음 중에서 어느 특정 배 음들이 강화되는 위치의 주파수를

말하고, 그 부근의 부분까지 포함해서 포만트라고 한다. 그러므로 주파수의 변화가 에너지 분포의 변화로 연결되고 또한 감정의 변화에 따른 포만트의 변화가 일어나기 때문에 이들의 분석도 필수적이다.

II. 신체적 변화와 포만트의 관계

포만트는 주파수가 낮은 쪽에서부터 F1, F2, F3, …로 이름을 붙인다. 영어 모음의 예를 들어 포만트를 살펴보면 그림 1(A)의 모음(前舌 모음)은 F1과 F2의 차이가 크고, 그림 1(B)의 모음(後舌 모음)은 그 차이가 작다. 따라서 여기서 전설·후설 모음의 음향학적 차이를 알 수 있다. 신체적 변화와 연관지어 살펴보면, 전설모음의 경우 혀의 가장 높은 부분과 입 천정과의 간격이 벌어질수록 F1의 주파수는 높아지고 F2는 낮아진다. 후설모음은 인두의 간격이 가장 좁아진 부분이 성문으로부터 멀어질수록 F1의 주파수는 낮아지며, 입술이 점점 오무라질수록 F2의 진폭이 줄어드는 것을 볼 수 있다. F1은 입안의 뒤쪽 및 목구멍에서 나는 공명에 기인하는데, 이것은 인두강의 공간에 기인함을 의미한다. 이 인두강은 혀의 높이에 따라 달라지며, 혀의 높이가 높을수록 인두강은 넓어지고 F1이 낮아진다. 그리고 F1은 모음의 고·저 자질과 관계가 깊어서 F1이 낮을수록 고모음이다. 한편 F2는 혀의 가장 높은 부분을 기준으로 하여 입안의 앞쪽 부분의 공명에 기인하므로 공명실의 길이에 좌우됨을 알 수 있다. 즉, 입안의 앞쪽이 넓을수록 F2는 낮아진다. 그러므로 F2는 모음 전·후 자질과 관련이 있어서 F2가 낮을수록 후설 모음이라고 할 수 있다[4].

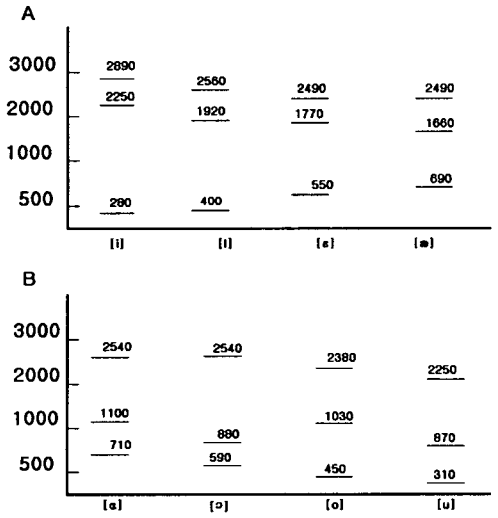


그림 1. 영어 모음의 제1, 제2, 제3 포만트 주파수

III. 화(Anger)의 표현분류

본 논문에서는 화나는 감정의 특징을 찾기 위한 방법으로 먼저 감정을 하위범주로 분류하고 범주의 최소단위로부터 감정의 특성을 찾아내었다.

문장의 구성에도 의문형, 명령형, 평서형의 구분이 되어 있듯이 감정의 표현 또한 문장으로 구성되어 있기 때문에 화나는 감정 한가지를 표현하는 경우에도 표 1과 같이 3개의 범주로 구분할 수 있다. 그리고 각각의 형에 따라서 각기 다른 특징을 보인다. 본 논문에서 화나는 감정의 특징을 찾기 위해 10명의 연기자에게 몇 가지 대사를 연기하도록 하여 녹음을 하였고, 그 외에도 TV드라마에서 연기자들의 대화를 녹음하였다. 녹음된 파일의 형식은 16bit, mono, 22kHz이고, 이 웨이브 파일을 분석하기 위한 도구로는 Praat 3.9.20(made by Paul Boersma and David weenink)를 사용하였다. 이 도구들을 이용하여 피치와 포만트에 초점을 맞추어 각 하위범주의 특징을 찾아내었다.

표 1. 화(anger)표현 분류(TV드라마로부터 발췌)

Category	Examples
의문형	"이제 눈에 뵈는 게 없냐" "일루 못와" "내가 그렇게 ...한 사람인줄 알어" "니가 뭘 안다고 함부로 떠들어" "당신 어딜 그렇게 싸돌아 다니는 거야"
설명형	"형은 형 방식대로 살어 난 내 방식대로 살테니까."
외치는 (Shout)형	"야이 양아치야!" "자꾸 부르지마!" "빨리 해!" "너 이리 못와" "너나 잘해!"

3.1 의문형의 특징

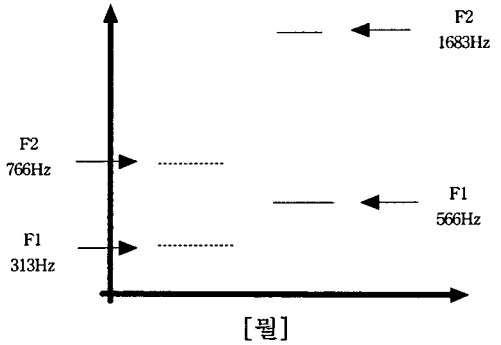
의문형의 특징은 감정 없는 대사에서와 마찬가지로 말끝을 올리는 것이다. 그러나 감정 없는 대사와의 차이는 말끝에서의 피치의 상승 편차이다. 그런데, 여기서 주의할 점은 대사의 내용이 의문형이라고 해서 화난 감정의 분류에서 반드시 의문형이 되는 것은 아니다. 그 이유는 내용상으로 의문형의 문장이더라도 외치는 형으로 표현될 수 있기 때문이다. 피험자로부터 얻은 웨이브 파일을 분석한 결과 문장의 마지

막 극소 피치와 끝 피치의 차를 살펴보면 다음의 표 2와 같다.

표 2. 화난 감정의 피치 특징

	대사	차
화난 감정	이제 눈에 뵈는게 없냐	135Hz
	내가 그렇게...한 사람인줄 알어	128Hz
	니가 뭘안다고 함부로 떠들어	116Hz
감정 없음	이거 얼마에요	51Hz
	이거 어때요	11Hz
	뭐라고 해야되지	46Hz
	그럼 어떻게 되는 거야	33Hz

위의 표에서 보는 바와 같이 화난 감정에서는 화자의 격앙된 느낌을 표현하고, 상대를 위협하기 위해서 문장 끝을 감정 없이 말할 때 보다 훨씬 과장되게 말하는 것을 피치의 차로 알 수 있다. 이것 한가지만으로 감정을 알 수 있는 것은 아니지만 여러 가지 요소 중의 한가지로써 중요한 역할을 한다. 그리고, 감정이 섞인 경우에는 무의식적으로 마디마다 강세를 섞는 경우가 많은데 그중 한 음절을 살펴보았을 때 그림 2와 같은 결과를 얻을 수 있었다. 앞에서 언급한 바와 같이 인두강의 공간 변화가 F1에 영향을 미치기 때문에 위 그림 2에서 사용된 '뭘'이란 음절에 감정을 넣게되면 인두강이 좁아지면서 F1이 높아지게 되는 것이다. 그리고, 입안의 앞쪽 공간에 따라서 F2가 영향을 받게되는데 이 경우는 입안의 앞쪽이 좁아지기 때문에 F2가 높아졌다.



-(실선) : 화남, …(점선) : 무감정

그림 2. 음절 '뭘'에서의 감정유무에 따른 F1과 F2의 비교

그리고, 또 한가지 예로써, 그림 3을 보면 똑같이 '너'라는 음절에 대해서 별 감정 없이 조용히 말한 가장 좌측의 부분과 화내는 감정으로 크게 외친 가운데 부분을 비교해보면, 앞에서 말한 바와 같이 F1과 F2에서 변화를 보임을 알 수 있다(그림에서 F1은 점들 중 가장 낮은 주파수대역을 말하고, F2는 두 번째로 낮은 주파수대역을 말한다).

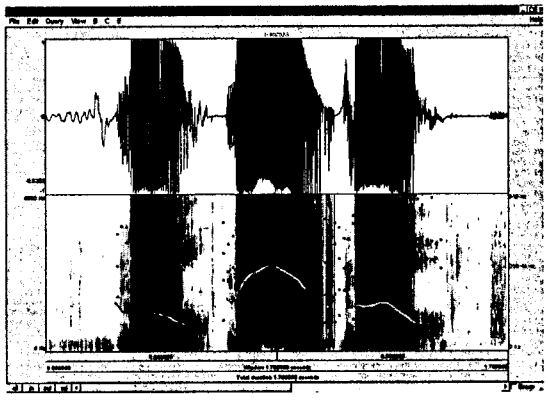


그림 3. '너'음절의 감정에 따른 비교

이러한 특징은 위의 한 음절에만 해당되는 것이 아니고 이외의 다른 음절에 대해서도 성립하는 것을 반복된 실험을 통해 알 수 있다.

3.2 설명형의 특징

설명형의 경우는 표 1의 간단한 예를 통해서 나왔듯이 훈계하는 형식으로 의문형이나 외치는 형과는 달리 긴 문장으로 이루어져 있다. 긴 문장으로 이루어져 있기 때문에 짧은 경우처럼 갑자기 큰 소리를 내는 경우는 드물다. 이런 타입의 경우는 완전히 주기적이진 않지만 준 주기적으로 피치의 변화가 주어진다. 그 이유는 화를 내면서 설명하는 경우 상대에게 명확하게 자신의 감정을 표시하기 위해서 한 마디마다 강조를 하기 때문이다. 즉, 예를 들면 드라마 상에서 나온 대화 중 『당신한텐 엄마가 아무 내색도 안한 모양인데 …』의 피치모양을 살펴보면 그림 4의 아래쪽 부분과 같다. 그림에서 원으로 표시된 부분이 위에서 설명한 마디마다의 강조된 부분으로써 강조를 위한 피치의 변화가 일정하다는 것을 알 수 있다.

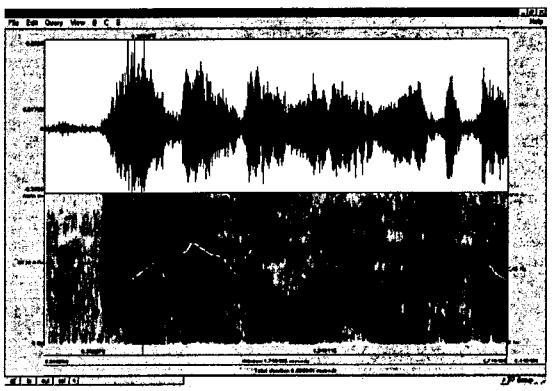


그림 4. 『당신한텐 엄마가 아무..』의 신호분석

그림 5도 동일한 설명을 보충해주는 그림으로써 다른 드라마로부터 녹취한 대사로써 강조되는 부분을 원으로 표시해 놓았다.

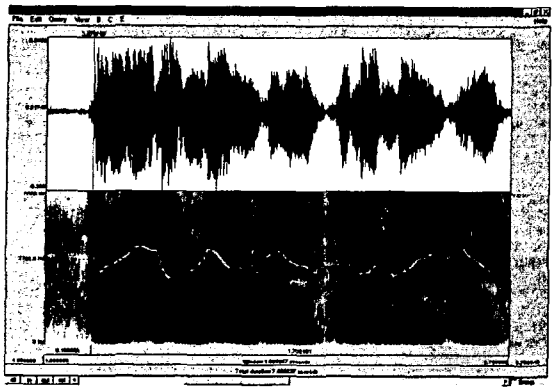


그림 5. 『엄만 엄마의 인생이..』 신호분석

3.3 외치는 형(Shout type)의 특징

외치는 형은 3.1절에서 설명한 의문형과 어미가 거의 유사하다. 즉, 의문형의 경우도 기본적으로 화내는 중의 격렬한 의사표현의 한가지 방법이기 때문에 그 신호의 특성은 외치는 형과 같다. 단지, 의문형의 고유특징인 어미에서의 피치의 변화가 다를 뿐이다. 포만트 주파수에서의 특성이 같다. 이 경우에도 피치궤적으로 봤을 때, 강조되는 부분(그림 4, 그림 5참조)에서는 신체적 변화가 평정이 유지되고 있을 때보다 크게 발생하기 때문에 그때의 포만트 변화가 두드러지게 나타난다. 이것에 대한 예는 그림 3에서 설명하였다. 즉, 본 논문에서 외치는 형과 의문형의 구분만은 어미에서의 피치 변화 성향이 의문형인 것과 그렇지 않은 것을 외치는 형으로 정했다.

Ⅳ. 결론 및 향후과제

본 논문에서는 '화(angry)'의 좀 더 용이한 분석을 위해서 3가지 타입으로 분류를 하였고, 감정의 변화에 따라 신체적으로 생기는 변화가 주파수에 영향을 주는 것을 드라마와 직접 연기하여 녹음한 대화내용을 분석하여 확인하였다. 그리고, 감정을 분석하는 요소로써 피치나 음질, 포만트 등 각각의 분석도구를 독립적으로 사용하지 않고 피치와 포만트를 연합하여 사용하였다. 본 논문에서는 'angry'의 경우의 특징을 주로 살펴보았는데, 강조 점에서의 포만트의 특징은 'angry'에서 뿐만이 아니라 기쁜 감정에서도 나타날 수 있다. 이런 중복되는 기본적인 특징들 때문에 다른 요소들과 함께 감정을 인식해야 할 것이다. 그리고, 차후에는 이들을 좀 더 체계화시켜서 감상인식 시스템을 구축할 수 있도록 할 것이다.

감사의 글 :

본 연구는 산업자원부의 2000년도 차세대신기술개발사업인 『수퍼지능칩 및 응용기술개발』 과제의 제5세부과제인 「Autonomous Family Machine(AFM) 요소기술개발(N09-A08-4301-05)」의 위탁연구로 이루어졌으며, 산업자원부의 연구비지원에 감사 드립니다.

Ⅴ. 참고문헌

[1] F. Dellaert, T. Polzin, and A. Waibel, *"Recognizing Emotion in Speech,"* Fourth International Conference on Spoken Language Processing, vol. 3, pp. 970-1973, 1996.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsi, S. Kollias, W. Fellenz, and J. G. Taylor ,"Emotion recognition in Human Computer Interaction" , *IEEE Signal Processing Magazine*, pp. 33-80, January, 2001

[3] T.L. Nwe, F.S. Wei, L.C.De Silvia,"Speech Based Emotion Classification" ,IEEE, 2001.

[4] 이규식, 석동일 "*청각학*", 대구대학교출판부, pp. 49-60, 1996.

[5] 박경범, "*선형예측분석법에 의한 음성의 압축과 재생*", 도서출판 하늘소, pp. 53-60, 1994