감성과학, Vol.16, No.1, pp.117-124, March 2013

자연스러운 정서 반응의 범주 및 차원 분류에 적합한 음성 파라미터*

Acoustic parameters for induced emotion categorizing and dimensional approach

박지은** · 박정식*** · 손진훈**†

Ji-Eun Park** · Jeong-Sik Park*** · Jin-Hun Sohn**†

**충남대학교 심리학과/뇌과학연구소

**Department of Psychology, Brain Research Institute, Chungnam National University

***목원대학교 공과대학 지능로봇공학과

***Department of Intelligent Robot Engineering, Mokwon University

Abstract

This study examined that how precisely MFCC, LPC, energy, and pitch related parameters of the speech data, which have been used mainly for voice recognition system could predict the vocal emotion categories as well as dimensions of vocal emotion. 110 college students participated in this experiment. For more realistic emotional response, we used well defined emotion-inducing stimuli. This study analyzed the relationship between the parameters of MFCC, LPC, energy, and pitch of the speech data and four emotional dimensions (valence, arousal, intensity, and potency). Because dimensional approach is more useful for realistic emotion classification. It results in the best vocal cue parameters for predicting each of dimensions by stepwise multiple regression analysis. Emotion categorizing accuracy analyzed by LDA is 62.7%, and four dimension regression models are statistically significant, $p<.001$. Consequently, this result showed the possibility that the parameters could also be applied to spontaneous vocal emotion recognition.

Key words : vocal emotion recognition, emotion recognition, dimensional approach

요 약

본 연구는 음성 인식기에서 일반적으로 사용되는 음향적 특징인 MFCC, LPC, 에너지, 피치 관련 파라미터들을 이용하여 자연스러운 음성의 정서를 범주 및 차원으로 얼마나 잘 인식할 수 있는지 살펴보았다. 자연스러운 정서 반응 데이터를 얻기 위해 선행 연구에서 이미 타당도와 효과성이 밝혀진 정서 유발 자극을 사용하였고, 110명의 대학생들에게 7가지 정서 유발 자극을 제시한 후 유발된 음성 반응을 녹음하여 분석에 사용하였다. 각 음성 데이터에서 추출한 파라미터들을 독립변인으로 하여 선형 판별 분석(LDA)으로 7가지 정서 범주를 분류하였고, 범주 분류의 한계를 극복하기 위해 단계별 다중회귀(stepwise multiple regression) 모형을 도출하여 4가지 정서 차원(valence, arousal, intensity, potency)을 가장 잘 예측하는 음성 특징 파라미터를 산출하였다. 7가지 정서 범주 판

* 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 휴먼인지환경사업본부-신기술융합형 성장동력사업의 지원을 받아 수행된 연구임(No. 2012K001339)

† 교신저자 : 손진훈 (충남대학교 사회과학대학 심리학과 및 뇌과학연구소)

E-mail : jhsohn@cnu.ac.kr

TEL : 042-821-6369

FAX : 042-823-9448

별율은 평균 62.7%이었고, 4 차원 예측 회귀모형들도 p<.001수준에서 통계적으로 유의하였다. 결론적으로, 본 연구 결과는 자연스러운 감정의 음성 반응을 분류하는데 유용한 파라미터들을 선정하여 정서의 범주와 차원적 접근으로 정서 분류 가능성을 보였으며 논의에 본 연구의 개선방향에 대해 기술하였다.
주제어: 음성 정서 인식, 정서 인식, 정서 차원 분류

1. 서론

HCI를 위해 인간의 소통 방법을 기계에 구현하려는 연구들이 활발히 진행되고 있다. 인간과 컴퓨터의 대화가 잘 진행되기 위해서는 컴퓨터가 상대의 말의 내용을 파악해야 할 뿐 아니라 그 속에 담겨있는 정서를 잘 인식해야 한다. 음성의 특성을 이용해서 대화의 내용을 파악하는 음성 인식 기술은 다양한 각도로 연구되어 스마트폰이나 콜센터 자동안내 시스템으로 상용화 되는 단계에 이르렀으나 정서의 인식은 아직까지 연구가 미미한 단계이다. 음성인식과 동시에 정서 인식이 가능하다면 콜센터나(Petrushin, 1999; Lee and Narayanan, 2005) 자동 예약 시스템에서(Ang et al., 2002; Schiel et al., 2002) 고객의 정서 상태를 고려하여 응대 방식을 유동적으로 할 수 있으며, 심리치료 장면에서도 상대방의 정서 상태를 인식한다면 치료의 효과를 높일 수 있다(France et al., 2000).

정서를 인식하기 위해서는 정서를 어떻게 기술할 것인지가 우선되어야 하고 이에 대한 설명 중 가장 전통적인 방식은 정서별 범주 분류이다(Ekman, 1971; Ekman and H. Oster, 1979). Ekman(1971, 1994)의 연구에 의하면 인간은 행복, 슬픔, 공포, 분노, 혐오, 놀람의 정서와 관련된 특정한 얼굴 표정을 가지는데, 이러한 얼굴 표정은 범문화적으로 표현되고 지각되므로 기본 정서라고 명명하였다. 이제까지 정서 인식에 관한 연구들은 주로 이러한 정서의 범주를 분류하는 것이고(Juslin & Laukka,2003), 분류하는 정서의 종류도 유발된 정서 데이터를 이용한 경우는 기쁨,분노,놀람,중성등의 3-4가지 정서나 긍정.부정 정서 분류에 국한되어 연구되었다(Zhihong Zeng, 2009). 본 연구에서는 기본정서 이외에 실 상황에서 유용한 정보를 처리하기 위해 통증을 느낄 때 표현하는 음성을 더 추가하여 7가지 정서 범주를 분류 하였다. 그러나 정서의 개별 범주는 자연스러운 의사소통 상황에서 발생하는 다양한 정서를 모두 포괄하지는 못한다. 이러한 범주적 설명 방식의 한계에 대한 대안 중 하나는 차원적 기술 방식이다(Greenwald, Cook & Lang, 1989 Russell & Mehrabian, 1977; Watson, Clark, & Tellegen, 1988). 이는 정서 상태를 정의된 차원 상에 위치시켜 각 차원의 특성으로 기술하는 방식이다. 정서의 차원을 몇 개로 할 것인지에 관해서는 다소 논란이 있지만(Russell & Barrett,1999) 본 연구에서는 현재 많은 연구들에서 다루고 있는 valence, arousal, potency, intensity 차원들을 사용하였다(Laukka, 2005). valence는 쾌한 정도를 나타내고, arousal는 각성 혹은 활성화 정도를 나타내며 potency는 대처할 힘 혹은 압도하는 정도, intensity는 정서의 강도를 나타내는 차원이다.

음성의 특성을 이용하여 기본정서 범주 중 한가지로 분류하는 방식의 한계는 정서가 고조된 상태에 이르러야 분류 성공률이 높아진다는 사실이다(Cowie & Cornelius,2003). 그러나 일상 대화 속에서 정서가 고조된 상태까지 올라가는 경우는 그리 많지 않다. 또 기본 정서 중 어느 한 가지 범주에 속하지 않는 정서 상태가 있거나 혹은 한 가지 범주에 속한다 하더라도 좀 더 세밀한 분류가 필요하다. 예를 들어 분노 정서의 경우 정서의 강도에 따라 매우 격한 분노에서 약한 분노로 나뉠 수 있고, 강한 분노라도 흥분해서 소리를 높이는 상태이거나 매우 침착하게 냉소적으로 낮은 소리를 내는 경우도 있다. 이러한 다양한 상태의 정서들을 범주 분류 방식으로 접근했을 때는 구별되지 못하고 모두 분노 범주로 분류될 뿐이다. 본 연구에서 도입한 4가지 차원으로는 앞서 설명한 분노의 쾌.불쾌감 정도, 각성의 정도, 강도 , 대처능력 정도의 세부적인 상태를 기술할 수 있다. 따라서 정서 상태를 세밀하게 파악하기 위해서는 정서의 범주적 분류와 더불어 차원적으로 기술하는 것이 더 효과적이다.

본 연구의 첫 번째 목적은 몇 가지 음향 특징 파라미터, 즉 멜-주파수 켑스트럼 계수(Mel-Frequency Cepstral Coefficient(MFCC)), 선형 예측 부호화(Linear Predictive Coding(LPC)) 계수, 에너지, 피치를 이용하여 정서의 범주 및 차원 분류가 가능한지 여부를 검토하고 그 중에 어떤 파라미터가 정서의 범주 뿐 아니라 차원을 예측하는데 더 유용한지를 알아보는 것이다.

음성의 음향적 특성을 이용한 정서 인식 연구에 주

로 사용되는 파라미터는 피치, 에너지, 말의 속도, 포만트 등이다(P.-Y. Oudeyer,2003). 그러나 본 연구에서는 정서인식이 아닌 음성인식에서 주로 사용되는 대표적인 특징 파라미터(J. W. Picone, 1993, O'Shaughnessy, 1999)인 MFCC 및 LPC 관련 파라미터를 사용하였다. 동일한 파라미터 추출을 통해 음성인식과 정서인식을 동시에 하는 것이 가능하다면 더 효과적일 것이기 때문이다.

MFCC는 사람의 청각 기관이 지니는 비선형적인 주파수 특성을 표현하는 특징 파라미터이다. LPC는 사람의 발성 모델에 근거한 음성 부호화 방식으로, 음성 생성 모델에서 비롯된 값이다(J. D. Markel, A. H. Gray,1976). LPC는 음성의 발성 기관을 하나의 필터로 가정하며, 따라서 그 필터의 계수(LPC 계수)를 음성 특성을 나타내는 벡터로 활용한다. 음성 신호의 에너지 또한 음성 특성을 나타내는 중요한 정보이다. 본 연구에서는 두 종류의 에너지, 즉 필터 에너지와 로그 에너지를 특징 파라미터로 사용한다. 피치란 음의 높낮이를 나타내는 값으로 음성 성분 중 성대 진동으로 발생하는 유성음 구간에 대해서만 측정이 가능한 값이며, 음의 기본 주파수와 밀접하게 연관되어 있다 (D. Gerhard,2003).

본 연구와 유사한 파라미터를 사용하여 정서를 분류한 선행 연구들이 있으나 매우 제한적인 연구가 진행되었기에 본 연구에서는 최대한 제한점들을 완화시켜 선행 연구의 제약을 넘어설 수 있는 가능성을 보여주었다. 예를 들면, Yixiong Pan 등(2012)의 연구는 동일한 파라미터를 사용하여 90%이상의 분류율을 보고하였으나 정서의 종류가 기쁨, 슬픔, 중립으로 양극단에 속하는 두 가지의 정서 상태를 분류하였고, Francesco Beritelli(2006) 연구는 7가지 정서를 MFCC관련 파라미터로 분류하여 64%의 평균 분류율을 보고하였으나 직업 연기자의 강한 감정 데이터를 사용한 것이므로 본 연구와는 차별성이 있다.

본 연구의 중요한 목적 중의 하나는 실제 정서를 느끼는 상황과 가장 유사한 음성 자료를 이용하여 정서인식 가능성을 탐색하는 것이다. 음성 감정 인식 연구에 이용되는 대부분의 음성 자료는 배우들의 연기를 녹음해서 사용한다(Thurid Vogt, Elisabeth Andre, 2005). 이는 실제 정서를 느끼고 표현하는 것보다 과장되어 고조된 정서를 표현하므로 정서인식이 훨씬 용이하지만(Thurid Vogt, Elisabeth Andre, 2005) 실제 상황을 반영하지 못하는 치명적인 단점이 있다. 이러한 종류의 데이터를 사용하여 만든 정서 인식기는 강한 정서만을 인식하고 일반적인 대화 상황에서 나타나는 정서들은 인식할 수 없게 된다. 따라서 가장 자연스러운 정서표현 데이터를 얻기 위해서는 실제 상황에서 녹음하는 것이 최상이지만 잡음 등과 같은 현실적인 문제들로 인해 그것이 쉽지 않아 본 연구에서는 동영상이나 게임 등의 상황을 이용하여 해당 정서를 유발시키고(Mi-Sook Park et al., 2011a, Mi-Sook Park et al., 2011b) 그 반응을 녹음하여 최대한 자연스런 반응에 가까운 자료를 녹음하였다. 이러한 자연 감정 데이터를 잘 분류할 수 있는 파라미터를 찾아내어 분류 가능성을 제시하는 것이 음성 정서 인식 연구에 큰 기여를 할 수 있으리라 생각한다.

본 연구의 특징을 정리해보면 첫째, 자연스러운 정서 표현 음성을 사용하였다. 둘째, 7가지 정서를 다루었다. 셋째, 정서 범주 분류의 제한을 보완하기 위해 4가지 차원으로의 정서 분류 가능성을 탐색하였다. 넷째, 기존 정서 분류에서는 사용하지 않던 파라미터를 사용하였다.

이를 위하여 1) 정서 유발 자극을 사용하여 피험자의 음성을 녹음하고 2) 음향 특성 파라미터를 추출한 후 3) LDA로 정서 범주 분류를 하고 4) stepwise multiple regression으로 정서의 차원을 예측하였다.

2. 방법

2.1. 실험참가자

충남대학교에 재학중인 대학생 110(남:56, 여:54)명이 실험에 참여하였고 평균연령은 23.8세(19~29)였다. 실험 참가자들은 뇌손상 병력이나 시력, 청력에 문제가 없는 성인이었다.

2.2. 실험장치 및 절차

피험자는 방음 처리된 실험실 의자에 앉아 모니터를 통해 주어지는 자극을 보면서 내용에 몰입하고 느껴지는 정서를 표현하도록 지시 받았다. 자극은 기쁨, 슬픔,분노,혐오,공포,놀람,통증 유발 자극(Mi-Sook Park et al., 2011a, Mi-Sook Park et al., 2011b)을 이용하였다. 박 미숙 등(2011a,2011b)에 의하면 정서 유발 자극은 평균 83% 정도의 효과성을 보고하였고, 해당 정서를 평균 80% 이상으로 타당하게 유발한다고 보고하였다. 정서 유발 방법은 주로 2분 정도의 동영상이나 게임으로

이루어져 있어 영상을 보거나 게임에 참여하면서 느껴지는 감정을 표현하는 절차이다. 먼저 프리세션에서 연습자극을 통해 감정을 표현하는 상황에 익숙해지도록 한 후에, 본 세션이 시작되면 동영상 자극인 경우는 본인이 출연하고 있다고 생각하면서 상황에 몰입하도록 지시하였다. 기쁨 유발자극의 경우는 직접 룰렛 판을 돌려 게임에 참여하고, 통증 유발 자극은 직접 팔을 압박하여 통증을 유발시키도록 되어있다.

음성 녹음은 (주)HCI-LAB에서 개발된 Octopus Board 와 마이크를 사용하였고 16 kHz , 16bit 샘플링, PCM 형식으로 저장하였다. 마이크는 피험자의 오른쪽 어깨 옷깃에 고정하였다. 한 가지 자극이 끝날 때마다 피험자는 자신이 느낀 정서가 무엇이었는지 보고하였고(범주 평가) valence, arousal, intensity, potency 차원에 대해 자신이 느낀 정서가 어디에 해당되는지 각각의 차원에 대해 표시(차원 평가)하도록 하였다.

2.3. 음성 특징 추출

1) MFCC

각 Mel-scale의 주파수 대역에 해당하는 신호의 크기로부터 <그림 1>과 같은 과정을 통해 12차의 MFCC를 추출하였다(Picone, 1993, Rabiner and Juang, 1993).

2) LPC 계수

LPC 계수는 LPC Durbin 방법(J. D. Markel, A. H. Gray,1976)을 사용하였다. 일차적으로 각 프레임에 포함된 음성 샘플들의 자기상관(autocorrelation) 값을 계산하고 LPC Durbin 회귀(recursion) 알고리즘을 사용하여 LPC 계수를 얻는다.

3) 에너지

본 연구에서는 필터 에너지와 로그 에너지를 특징 파라미터로 사용하였다. 필터 에너지는 MFCC 계산 과정에서 측정한 각 필터 뱅크의 에너지를 모두 합산하였고, 로그 에너지는 프레임에 포함된 모든 음성 신호의 크기의 합을 계산한 후 로그를 취하여 얻어냈다.

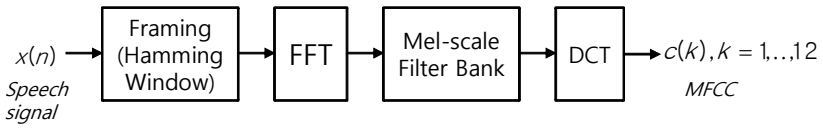
4) 피치(Pitch)

본 연구에서는 자기상관 방법을 통해 피치 정보를 추출하였다. 음성 샘플들의 자기상관(autocorrelation) 값으로부터 피치를 측정하는 이 방법은 샘플로부터 직접 측정 가능하며, 음성에 대해서도 비교적 정확한 값을 나타낸다. 자기상관 기반의 피치 측정 방법은 식(1)을 통해 구해진다. 즉, 한 프레임에 포함된 음성 샘플의 수가 N일 때, 각 i에 대해 i번째 샘플의 크기와 (i+m)번째 샘플의 크기의 곱을 계산한 후 얻어진 모든 값들의 합이 자기상관 값이며 이 때, 가장 큰 자기상관 값을 나타내는 m을 찾아 그것을 해당 프레임의 피치로 정한다(L. Rabiner, 1977).

$$r(m) = \sum_{i=0}^{N-1} x(i)x(i+m) \quad (1)$$

2.4. 분석

녹음한 음성자료의 프레임별로 추출한 12차 MFCC, LPC계수, 필터에너지, 로그에너지, 피치값으로부터 각 차수(*i*)별 MFCC의 평균(M*i*_mean), 분산 (M*i*_std), 최대(M*i*_max), 최소(M*i*_min), 범위(M*i*_range), 각 차수별 LPC계수의 평균(L*i*_mean), 분산(L*i*_std), 최대(L*i*_max), 최소(L*i*_min), 범위(L*i*_range), 필터에너지의 평균(FE_mean), 분산(FE_std), 최대(FE_max), 최소(FE_min), 범위(FE_range), 로그에너지의 평균(LE_mean), 분산(LE_std), 최대(LE_max), 최소(LE_min), 범위(LE_range), 피치의 평균(PH_mean), 분산(PH_std), 최대(PH_max), 최소값(PH_min), 범위(PH_range)를 계산하고 그 값을 이용하여 단계별 선형판별분석과 다중회귀 분석을 실시하여 정서의 7가지 범주를 구별하는 판별함수와 각 4가지 차원을 예측하는 회귀함수를 도출하였다.



<그림 1> MFCC 계산 과정

3. 결과

3.1. 정서의 범주 분류

12차 MFCC, LPC계수, 필터에너지, 로그에너지, 피치 관련 파라미터들을 독립 변인으로 하여 단계 선택적 판별분석을 한 결과, 선택된 파라미터는 M1_mean, M0_std, FE_max, L8_range, M0_mean, LE_max, M4_mean, LE_range, M9_std, M2_min, M5_min, L6_mean, M10_range, M9_min이고 각 정서의 분류 성공률은 분노 80.5%, 혐오 45.8%, 놀람 80.0%, 공포 51.2%, 기쁨 46.0%, 통증 56.8%, 슬픔 72.5% 로 7가지 정서에 대해 62.7%의 평균 판별률을 보였다. 표 1에 7가지 각 정서 범주의 판별 결과를 제시 하였다.

Table1. Accuracies of emotional category discrimination

	정서	예측 소속집단							전체
		분노	혐오	놀람	공포	기쁨	통증	슬픔	
빈도	분노	62	5	0	0	5	0	5	77
	혐오	7	22	3	2	2	5	7	48
	놀람	0	1	16	1	2	0	0	20
	공포	3	1	4	22	4	7	2	43
	기쁨	4	2	9	6	23	4	2	50
	통증	0	2	2	7	2	25	6	44
	슬픔	6	12	0	0	1	0	50	69
%	분노	80.5	6.5	.0	.0	6.5	.0	6.5	100.0
	혐오	14.6	45.8	6.3	4.2	4.2	10.4	14.6	100.0
	놀람	.0	5.0	80.0	5.0	10.0	.0	.0	100.0
	공포	7.0	2.3	9.3	51.2	9.3	16.3	4.7	100.0
	기쁨	8.0	4.0	18.0	12.0	46.0	8.0	4.0	100.0
	통증	.0	4.5	4.5	15.9	4.5	56.8	13.6	100.0
	슬픔	8.7	17.4	.0	.0	1.4	.0	72.5	100.0

3.2. 정서의 차원 예측

정서 범주 변인을 통제한 후에 12차 MFCC, LPC계수, 필터에너지, 로그에너지, 피치 관련 파라미터들을 독립 변인으로 하여 정서 차원 예측 회귀함수를 도출한 결과, valence 차원을 가장 잘 예측하는 파라미터는 L10_max, M10_range, L7_max, L8_std, M11_max이 선택되었고 이들이 음성 자료의 쾌.불쾌 정도를 83.1% 설명할 수 있었다. arousal 차원은 L7_min, L9_max, L3_std, L10_max이 각성 수준의 높고 낮음을 27.0% 설명하고, potency 차원은 L7_min, M6_min, L0_max, L3_std, M6_std이 통제능력을 18.6%, intensity 차원은 M1_min, L2_min, M0_max, L3_min, M5_min, L7_mean, M7_std, L0_std, M0_min이 최적의 예측 파라미터로 선택되어 정서의 강도를 22.5%를 설명한다. 각 모형의 설명량(R^2)과 선택된 파라미터들의 회귀계수(B) 및 유의도는 표2 에 상세히 표시하였다.

3.3. 정서범주와 정서차원간의 관계

각 차원에 미치는 범주 변인의 영향을 통제하기 위해 단계적 회귀분석에서 정서 범주 더미 변인을 사용하였다. 7개의 정서 범주이므로 6개의 더미 변수를 추가하였고 D1~D6는 각 각 분노, 혐오, 놀람, 공포, 기쁨, 통증의 범주 더미 변인을 의미하며 코딩 값은 표2의 각주1)에 제시하였다. 각 차원별로 범주만의 설명량에 비해 음성 파라미터들을 추가함으로 인해 증가된 설명량(R^2증가량)을 표2에 제시하였다. 정서의 범주에 차원적 기술을 추가함으로써 각 모형은 1.8%, 4.8%, 7.1%, 13.2% 설명량이 증가하는 결과를 확인할 수 있었다. 이 결과로 각 차원마다 정서 범주가 미치는 영향이 다름을 알 수 있다. 표3에 제시한 차원을 평가한 평균값을 보면 valence차원의 경우 긍정적 정서와 부정적 정서의 평균값이 양극단에 위치하여 두 부류의 정서 범주가 valence 차원에 영향을 많이 미칠 것을 예측할 수 있고 그 설명량도 범주 변인이 80%이상 설명하며 1.8%정도만큼만 개인의 쾌.불쾌 정도가 다름을 알 수 있다.

4. 논의

본 연구의 목적은 MFCC, LPC계수, 에너지, 피치 관련 파라미터를 이용하여 자연스러운 음성의 정서를 범주 및 차원으로 분류하는 것이 가능한지 탐색하는 것이다. 단계 선택적 판별분석으로 범주 분류를 한 결과, 7가지 정서 평균 판별률이 62.7% 이었다. 이는 7가지 정서의 우연 판별률 14.4%보다 높을 뿐 아니라, 직접적인 비교는 어렵더라도 정서의 종류와 사용 파라미터가 가장 근접한 선행 연구들과 비교해 보면 J. Nicholson(2000) 연구는 중립을

Table 2. Regression model of Emotional Dimension using acoustic cue

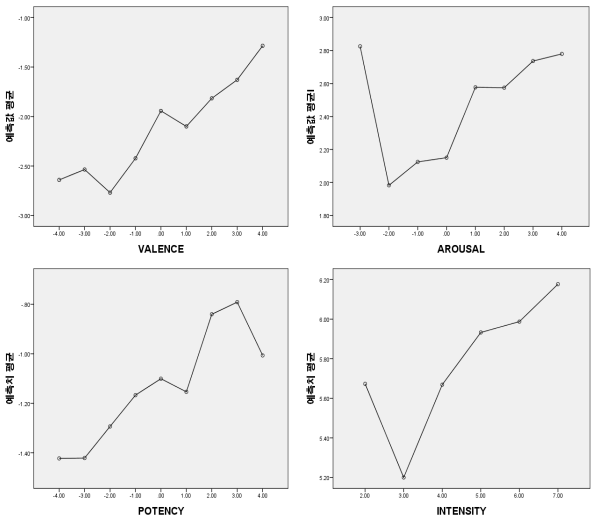
차원	Valence		Arousal		Potency		Intensity	
	(상수)	-2.847***	(상수)	2.265***	(상수)	-2.648***	(상수)	5.304***
	D1[1)]	-.861***	D1	1.837***	D1	.426	D1	.530*
	D2	-1.138***	D2	1.473***	D2	-.874*	D2	.353
	D3	.514*	D3	1.371***	D3	.169	D3	-.143
	D4	-.418*	D4	1.607***	D4	-.653	D4	-.089
	D5	6.075***	D5	.825***	D5	1.972***	D5	-.062
	D6	.204	D6	.779**	D6	.168	D6	-.585**
	L10_max	1.721*	L7_min	1.325*	L7_min	-2.910**	M1_min	-.048**
	M10_range	-.038**	L9_max	1.749**	M6_min	.120**	L2_min	.465*
	L7_max	-1.400***	L3_std	-1.673*	L0_max	-1.019**	M0_max	-.034
	L8_std	6.124**	L10_max	-1.550*	L3_std	3.610*	L3_min	-1.091**
	M11_max	.029*			M6_std	.291*	M5_min	-.030**
							L7_mean	1.345
							M7_std	-.090*
							L0_std	-1.337**
							M0_min	-.037*
모형	F(11,339) =151.857***		F(10,340) =12.561***		F(11,339) =7.023***		F(15,335) =6.476***	
R^2	.831		.270		.186		.225	
R^2증가량	.018		.048		.071		.132	

1) D1~D6: dummy variables D1:Anger=1, D2:Disgust=1, D3:Surprise=1, D4:Fear=1, D5:Jo=1, D6:Pain=1
*p<.05 **p<.01 ***p< .001

Table 3. Mean of dimension rating by emotional categories

	Valence	Arousal	Potency	Intensity
분노	-3.47	3.29	-.87	6.42
혐오	-3.67	3.13	-2.40	6.29
놀람	-2.20	3.15	-1.40	5.95
공포	-3.07	3.33	-2.05	6.07
기쁨	3.40	2.52	.32	5.98
통증	-2.27	2.43	-.98	5.41
슬픔	-2.64	1.58	-1.51	5.75

제외한 7가지 정서를 LPC등의 파라미터를 사용하여 평균 50%의 분류율을 보고했고, Francesco Beritelli(2006) 연구에서도 7가지 정서를 MFCC 파라미터로 64%의 평균 분류율을 보고한 것과 비교할 수 있다. 이들은 모두 연기자들의 데이터를 사용하였기 때문에 자연스런 음성 반응의 인식 결과라는 점을 감안해 보면 본 연구 분류 성공률이 낮지 않은 결과이다.



<그림 2> 차원 평가치와 예측치의 관계

표1에 제시된 분류 결과에 따르면 보면 분노, 놀람, 슬픔 정서는 그 성공률이 각각 80.5%, 80%, 72.5%로 상당히 높으나 혐오나 기쁨은 50% 미만으로 낮은 분류율을 보이고 있다. 잘못 분류된 경우들을 분석한 결과, 혐오 음성의 경우 분노나 슬픔으로 오분류 되는 경우가 가장 많았고, 기쁨 음성의 경우는 놀람과 공포

로 오분류 되는 경우가 많았으며, 통증 음성도 공포로 분류된 경우가 가장 많았다. 이는 분석에 사용된 음성 데이터가 인위적으로 만든 강한 반응이 아니고 자연스러운 반응에 가깝기 때문에 생긴 결과로 보인다. 일반적으로 기쁠 때 표현하는 소리는 인위적으로 기쁨을 표현하지 않는 한 놀란 소리처럼 들릴 수 있어, 이러한 오분류 결과는 인위적으로 강하게 만들어진 데이터가 아닌, 실제 정서반응 데이터를 범주로 분류하는 방식이 한계가 있음을 잘 드러내는 결과로 보인다. 따라서 이러한 단점을 극복하기 위한 대안이 범주 정보와 함께 차원적 정보를 함께 제공하는 것이다. 이를 위해 본 연구에서는 앞서 범주 분류에 사용되었던 동일한 파라미터들을 사용하여 차원의 인식이 가능한지 알아보았다. 동일한 파라미터를 독립 변인으로 하여 단계별 회귀분석 결과 valence, arousal, potency, intensity 4가지 차원에 관한 회귀모형은 모두 통계적으로 유의미하였다. 그림 2는 도출된 회귀식에 의한 차원 예측값 평균과 차원 측정값과의 관계를 그래프로 나타낸 것이다. 설명량이 높은 valence차원은 선형 회귀선의 형태를 보이나 그렇지 못한 나머지 차원들은 선형 회귀선에서 조금 벗어난 형태를 보인다. Arousal은 매우 각성이 낮은 상태의 값(-4점)을 잘 예측하지 못하고, intensity도 낮은 강도의 값(2점 이하)일 때, potency는 높은 통제력(4점)을 잘 예측하지 못하였다. 이는 차원을 보다 정확하게 예측할 수 있는 추가적인 파라미터의 도입이 필요함을 의미한다고 할 수 있다.

그러나 차원적 접근을 함으로써 자연스런 감정 처리를 할 때, 개인별 정서를 얼마나 더 세밀하게 설명할 수 있는지 결정계수의 증가량을 통해 알 수 있었고 상세 자료를 표2에 제시하였다. 범주 변인의 영향을 가장 많이 받은 차원은 앞서 언급했듯이 valence차원이므로 이 차원을 도입하는 것은 정서를 범주로 분류하는 방식에서 크게 더 나은(1.8%) 설명을 제시하지는 못하는 것으로 보이나 arousal, potency, intensity차원은 범주적 접근으로는 설명하지 못하는 개인이 느끼는 정서의 세밀한 차이를 더 풍부하게 설명해 줄 수 있음을 의미한다. 차원 예측력이 좀 더 우수한 파라미터의 개발을 통해 더 많은 설명력의 증가를 기대할 수 있으리라 생각된다.

본 연구의 결과를 기존 다른 파라미터를 사용한 연구(Laukka, 2005)와 비교해 볼 때, valence의 설명량은 83.1%로 Laukka의 연구 25%에 비해 월등히 높고, 나머지 3개의 차원은 Laukka의 연구에서 사용되었던 파라미터의 설명력이 더 높은 것으로 확인되었다. 배우들의 연기 데이터를 사용하는 경우는 자연스러운 감정 데이터에 비해 감정이 과장되어 표현되므로 분류결과가 더 높게 나올 수 있다. 이 점에서 본 연구의 차별 점을 들 수 있다. 따라서 본 연구에서 사용한 파라미터들과 기존 감정인식에 사용된 파라미터들을 적절히 조합하여 자연스런 음성의 감정 분류를 잘 할 수 있는 최적의 파라미터를 찾아내는 것이 추후 연구에서 수행되어야 할 것이다. 뿐만 아니라 감정의 범주도 좀 더 세분화 하고 이를 더 정확하게 인식을 할 수 있는 알고리즘을 개발해야 할 필요가 있다.

REFERENCES

C. M. Lee, S. S. Narayanan(2005). Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Process. 13 (2),* 293 – 303.

Cowie R.; Cornelius R.R. (2003). Describing the emotional states that are expressed in speech, Speech Communication, 40(1), 5-32.

D. Gerhard(2003). Pitch Extraction and Fundamental Frequency: History and Current Techniques, *Technical report,* Dept. of Computer Science, University of Regina.

D. O'Shaughnessy(1999). *Speech Communication: Human and Machine,* 2nd ed. Wiley-IEEE Press.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion, *Proceedings of the 1971 Nebraska Symposium on Motivation*, 207-283.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique, *Psychological Bulletin, 115(2),* 268-287.

Ekman, P. & Oster, H. (1979). Facial Expressions of Emotion, *Annual Review of Psychology, 30,* 527-554.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M.,Wilkes, M.(2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomedical Engineering 7,* 829 – 837.

F. Beritelli, S. Casale, A. Russo, S. Serrano, and Donato Ettorre(2006). Speech Emotion Recognition Using MFCCs Extracted from a Mobile Terminal based on ETSI Front End, 8th International Conference on Signal Processing(ICSP), Beijing.

Greenwald, M., Cook, E., & Lang, P. (1989). Affective

judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli, *Journal of Psychophysiology, 3(1)*, 51-64.

J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog, *Proceedings of International Conference Spoken Language Processing (ICSLP '02), 3,* 2037 – 2040.

J. D. Markel, A. H. Gray(1976). *Linear Prediction of Speech,* Springer-Verlag, New York.

J. Nicholson, K. Takahashi and R. Nakatsu (2000). Emotion Recognition in Speech using Neural Networks, *Neural Computing & Appliccation, 9,* 290-296.

J. W. Picone(1993). Signal modeling techniques in speech recognition, *Proc. IEEE, 8(9),* 1215-1247.

L. Rabiner, B.H. Juang(1933). *Fundamentals of Speech Recognition,* Prentice-Hall, Englewood Cliffs, NJ.

L. Rabiner(1977). On the Use of Autocorrelation Analysis for Pitch Detection, *IEEE Transactions on acoustics, speech and signal processing, 25(1),* 24-33.

Mi-Sook Park, Hyo-Eun Kim, Jin-Hun Sohn (2011a). Development of emotion-evoking stimuli to provoke spontaneous emotions. *Proceedings of the Korean Society for emotion & sensibility : Emotion & Smart Interface,* 42-43.

Mi-Sook Park, Ji-Eun Park,Jin-Sup Eom, Jin-Hun Sohn (2011b). Development of stimuli for spontaneous emotion-evoking II, *Proceedings of Korean Society for emotion & sensibility,* 22-23.

Patrik Juslin., Petri Laukka (2005). Communication of emotions in vocal expression and music performance: Different channels, same code?, *Psychological Bulletin, 129(5),* 770-814.

Petri Laukka, Patrik Juslin & Roberto Bresin (2005). A dimensional approach to vocal expression of emotion, *Cognition & Emotion, 19(5),* 633-653.

Petrushin, V. A.(1999). Emotion in speech recognition and application to call centers, *Proceedings of Artificial Neural Networks in Engineering (ANNIE 99), 1,* 7 – 10.

P.-Y. Oudeyer (2003). The production and recognition of emotions in speech: features and algorithms, Int. *Journal of Human-Computer Studies, 59(1 – 2),* 157 – 183.

Russell JA, Barrett LF(1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant, *Journal of Personality and Social Psychology, 76(5),* 805-19.

Russell, J., Mehrabian, A. (1977). Evidence for a three-factor theory of emotions, *Journal of Research in Personality, 11(3),* 273-294.

Schiel, F., Steininger, S., Turk, U.(2002). The Smartkom multimodal corpus at BAS, *Proc. Language Resources and Evaluation (LREC '02),* Las Palmas (Spain)

Smith CA, Ellsworth PC(1985). Journal of personality and social psychology, *Patterns of cognitive appraisal in emotion,48(4),* 813-38.

Thurid Vogt, Elisabeth Andre. (2005). Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition, *In Proceedings of IEEE International Conference on Multimedia & Expo(ICME 2005),* Amsterdam, The Netherlands

Todor Ganchev, Nikos Fakotakis, George Kokkinakis (2005). Comparative evaluation of various MFCC implementations on the speaker verification task, *10th International Conference on Speech and Computer,1,* 191 – 194.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales, *Journal of Personality and Social Psychology, 54(6)*, 1063-1070.

Yixiong Pan, Peipei Shen, and Liping Shen (2012). Speech Emotion Recognition Using Support Vector Machine, *International Journal of Smart Home. 6(2)*, 101-108.

Zhihong Zeng, Maja Pantic,Glenn I. Roisman, Thomas S. Huang, (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence , 31(1), 39-58.

원고접수: 2013.02.27
수정접수: 2013.03.29
게재확정: 2013.03.29