논문 2009-46SP-4-10

최소 분류 오차 기법과 멀티 모달 시스템을 이용한 감정 인식 알고리즘

(Emotion Recognition Algorithm Based on Minimum Classification Error incorporating Multi-modal System)

이 계 환*, 장 준 혁**

(Kye-Hwan Lee and Joon-Hyuk Chang)

요 약

본 논문에서는 최소 분류 오차 기법 (Minimum Classification Error, MCE)에 기반한 감정 인식을 위한 알고리즘 멀티 모달 (Multi-modal) 시스템을 기반으로 제안한다. 사람의 음성 신호로부터 추출한 특징벡터와 장착한 바디센서로부터 구한 피부의 전기반응도 (Galvanic Skin Response, GSR)를 기반으로 특징벡터를 구성하여 이를 Gaussian Mixture Model (GMM)으로 구성하고 이를 기반으로 구해지는 로그 기반의 우도 (Likelihood)를 사용한다. 특히, 변별적 가중치 학습을 사용하여 최적화된 가중치를 특징벡터에 인가하여 주요 감정을 식별하는 데 이용하여 성능향상을 도모한다. 실험결과 제안된 감정 인식이 기존의 방법보다 우수한 성능을 보인 것을 알 수 있었다.

Abstract

We propose an effective emotion recognition algorithm based on the minimum classification error (MCE) incorporating multi-modal system The emotion recognition is performed based on a Gaussian mixture model (GMM) based on MCE method employing on log-likelihood. In particular, the proposed technique is based on the fusion of feature vectors based on voice signal and galvanic skin response (GSR) from the body senser. The experimental results indicate that performance of the proposal approach based on MCE incorporating the multi-modal system outperforms the conventional approach.

Keywords: 감정 인식, Minimum Classification Error (MCE), Gaussian Mixture Model (GMM)

I. 서 론

IT 기술은 기술 및 시설 인프라 구축 중심에서 인간 중심으로 진보되어 왔다. 인간 중심의 발전 방향은 계속적으로 지속될 것이고, 이에 따른 서비스의 중요성도 더욱 부각될 것이다. 이러한 서비스의 핵심기술 중 하나인 감정 인식은 현재 많은 연구가 진행되고 있으며 차세대 기술로 주목받고 있는 실정이다. 최근에는 휴대용 기기와 로봇 등의 분야에서 감정 인터페이스에 대한 관심이 고조되었으며, 국내에서는 물론 해외에서도 중요한 연구 주제로 부각되는 실정이다[1~3]. 이러한 감정 인터페이스에 관한 연구는 단순한 외적 요소의 감정 상태와 이를 넘어 선호 경향까지 파악할 수 있는 기술이 요구된다.

최근 감정 인식에 대한 연구들은 화남, 슬픔, 즐거움, 중립 감정 등의 기존 감정을 기반으로 그 범위를 확대

* 학생회원, ** 정회원, 인하대학교 전자공학부
(Department of Electronics Engineering, Inha University)

※ 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2008-C1090-0804-0007) 그리고 또한 본 연구는 지식경제부 출연금으로 ETRI, SoC산업진흥센터에서 수행한 IT SoC 핵심설계인력양성사업의 연구결과입니다.
접수일자: 2008년11월14일, 수정완료일: 2009년6월9일

해 나가고 있다. 특히, 음성신호에는 화자에 대한 고유한 정보는 물론 감정과 피로도 등 다양한 정보가 포함되어 있기 때문에 감정 인식 연구 분야에서 많이 사용되어 지고 있다. 음성 신호를 이용한 감정 인식에서는 화자의 감정 상태를 반영하는 효과적인 특징들을 추출하는 것이 성능을 결정하는 데 가장 중요한 요소이다. 많은 경우에서 음성에서 추출해낸 피치와 에너지 정보가 감정 인식에 매우 효과적인 특징임은 밝혀졌지만[4~5], 그 자체로 완벽한 분류가 어렵고 다양한 감정에 대한 분류에서 한계가 있기 때문에 보다 감정 상태를 잘 반영할 수 있는 특징에 대한 연구가 필요한 실정이다.

본 논문에서는 감정 인식을 위한 알고리즘을 멀티 모달 (Muti-modal) 시스템에 기반하여 최소 분류 오차 기법 (Minimum Classification Error, MCE)을 도입하여 향상된 기법을 제안한다. 사람의 음성 신호와 이를 보완하기 위해 바디센서로부터 얻어지는 피부 전기반응도 (Galvanic Skin Response, GSR)를 특징벡터로 구성한 뒤 Gaussian mixture model (GMM)를 구성한다. 특히, MCE기법을 통한 변별적 가중치 학습을 사용하여 최적화된 GMM기반의 감정 인식을 수행한다. 수행 결과 제안된 방법이 기존의 방법보다 우수한 성능을 보임을 알 수 있었다.

본 논문의 구성으로는, Ⅱ장에서는 실험에서 사용되어진 특징 벡터에 대해서 기술하고, Ⅲ장에서는 제안된 감정 인식 알고리즘에 대해 기술한다. Ⅳ장에서는 실험결과 비교 및 분석에 대해 기술하였으며, 마지막으로 Ⅴ장에서는 결론을 맺는다.

Ⅱ. 특징벡터의 소개

본 논문에서는 기본적인 감정 인식을 위해 기존의 연구에서 많이 사용되어진 특징벡터를 추출하였으며, 8 kHz로 샘플링 된 입력신호로 부터 추출 특징벡터는 다음과 같다.

- Mel Frequency Cepstral Coefficients (MFCC) 13차
- △MFCC 13차
- 피치 (Pitch) 1차
- 음악 연속성 계수 1차

여기서 음악 연속성 계수 (Music Continuity Counter, MCC)는 에너지의 이동 평균 (Running Mean Energy), 스펙트럼차이 (Spectrum Difference) 그리고 피치 상관도 (Pitch Correlation)와 설정된 문턱 값을 통해서 얻어지는 값이다[6].

특히, 이렇게 구성된 28차의 특징벡터에 기존의 음성 중심의 감정 인식을 보완하기 위해서 피부 전기반응도를 추가로 특징벡터로 선택하였다. 피부 전기반응도는 BodyMedia사의 SensorWear ISSPro를 사용하여 추출하였으며[7], 초당 4회의 데이터가 생성된다.

- 피부 전기반응도 (Galvanic Skin Response) 1차

따라서 최종적으로 감정인식에 사용되어지는 특징벡터는 마이크와 SensorWear ISSPro로 구성된 멀티 모달 시스템에서 총 29차의 특징벡터를 10 ms마다 추출되게 된다.

Ⅲ. 감정 인식을 위해 제안된 알고리즘

본 논문에서 제안된 감정 인식은 화남 (Angry), 즐거움 (Joy) 그리고 중립 (Neutral) 등의 총 3가지 감정에 대해서 인식을 한다. 이는 기존의 감정 인식 연구되어지던 5가지 또는 그 이상의 감정 분류를 3가지로 통합한 경우로, 기존의 감정 인식 분류가 억양이나 단순한 느낌에 의존한데 비해 실제 사람이 나타낼 수 있는 대표적인 감정을 중심으로 인식을 수행한 것으로 실제 로봇 및 휴대폰등의 상용시스템의 경우 3가지만의 감정에 특화되어 성능을 높이는 것은 의미있는 작업이라 고려될 수 있다.

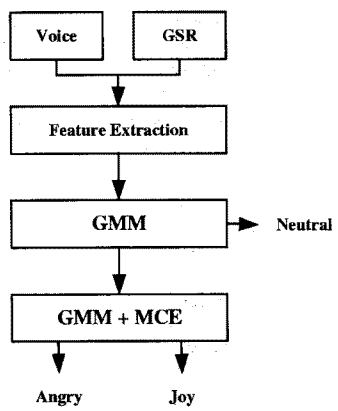

그림 1. 제안된 감정 인식의 전체 블록도
Fig. 1. Diagram of the proposed method.

제안된 감정인식 알고리즘은 멀티 모달 시스템으로부터 추출된 총 29차의 특징 벡터와 GMM 그리고 MCE를 통한 변별적 가중치를 기반으로 구성되어진다. 일차적으로 뚜렷한 인식 성능을 보이는 중립 감정을 구별한 뒤 이차적으로 MCE를 이용하여 구해진 변별적 가중치를 적용하여 구별이 힘든 화난 감정과 즐거움 감정 사이의 인식을 수행 하였다. 관련하여 그림 1은 제안된 감정인식의 전체적인 블록도를 보여주며, 세부적인 사항은 다음과 같다.

1. Gaussian Mixture Model (GMM)

제안된 감정 인식에 사용되어지는 GMM은 알고리즘은 주어진 데이터에 대한 분포밀도를 복수개의 가우시안 확률밀도함수로 모델링하는 방법중하나이다. 인식에 사용되어지는 특징벡터가 N개의 D차원 특징벡터 $X=\{x_1,x_2,...,x_D\}$라고 하면, M개의 혼합성분(Mixture Component)으로 구성되는 감정 인식 모델의 우도 (Likelihood)는 다음과 같이 계산되어진다[8~9].

$$p(\vec{x}|\lambda)=\sum_{i=1}^{M}p_i b_i(\vec{x}),$$
$$b_i(\vec{x})=\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\vec{x}-\mu_i)^T(\Sigma_i)^{-1}(\vec{x}-\mu_i)\right\} \quad (1)$$
$$\sum_{i=1}^{M}p_i=1,\quad 0\le p_i\le 1,$$

여기서 감정 인식 모델 λ는 혼합 성분 밀도의 가중치 (Mixture Weight : p_i), 평균 벡터 (Mean Vector : μ_i) 그리고 공분산 행렬 (Covariance Matrix : Σ_i)로 구성된다. 구성된 감정 인식 모델은 Expectation Maximization (EM) 알고리즘을 사용하여 $p(x|\lambda')\ge p(x|\lambda)$가 되는 새로운 모델 λ'가 정해진 문턱값에 도달할 때까지 반복한다. 이때 구해진 사후확률 중 가장 큰 우도 값을 가지는 모델을 구한 뒤, 다음과 같이 입력신호에 대한 감정 모델별 우도 값을 비교하여 감정 인식을 하게 된다.

$$\hat{E}=\underset{1\le V\le S}{\operatorname{argmax}}\sum_{t=1}^{N}\log p(\vec{x}|\lambda_V) \quad (2)$$
$S=3$, (1: Angry, 2: Joy, 3: $\neq$ utral)

2. MCE를 통한 최적화된 GMM의 로그우도

제안된 방법은 일반적으로 감정 인식에서 많이 사용되어지는 GMM을 이용하여 3가지 감정 중 하나인 중립 감정을 분류해 내며 나머지 2가지 감정은 MCE를 사용한 변별적 가중치를 적용하여 보다 효과적인 인식을 수행한다[10]. 이는 감정인식에서 중립을 제외한 두 감정을 구별하는 데 있어 성능향상이 관건인 점을 고려하면 의미 있는 시도라고 고려될 수 있다. 식(2)와 같이 감정인식을 위해 사용되어진 GMM을 이용한 결정식 중, 화남과 즐거움만을 구분해 내는 결정식은 다음과 같이 나타낼 수 있다.

$$\Lambda=\log\frac{p(\vec{x}|\lambda_A)}{p(\vec{x}|\lambda_J)}\ \begin{matrix}\text{Angry}\\ >\\ <\\ \text{Joy}\end{matrix}\ \eta \quad (3)$$

여기서 η는 화남과 즐거움을 나누는 문턱값이며, λ_A는 화남 모델 그리고 λ_J는 즐거움 모델을 나타낸다.

제안된 방법은 화남과 즐거움을 분류하는 결정식에 MCE 기법을 적용하여 각각의 모델별 혼합성분에 분별적 가중치를 적용한 최적의 감정 별 모델을 만드는 것이며, 제안된 최종 결정식은 다음과 같이 나타낼 수 있다.

$$\Lambda^{\omega}=\log\omega_i\frac{p_i^A b_i^A(\vec{x})}{p_i^J b_i^J(\vec{x})}\ \begin{matrix}\text{Angry}\\ >\\ <\\ \text{Joy}\end{matrix}\ \eta \quad (4)$$

여기서 Λ^{ω}는 제안된 MCE 기법을 통해 구해진 혼합성분별 가중치가 적용된 최종 결정식을 나타낸다. 최종 결정식을 위한 최적을 가중치 ω_i를 구하기 위해 Generalized Probilistic Descent (GPD) 기법을 사용하게 된다. 이러한 기법을 기반으로 다음과 최종 결정식에 대한 분류 오류 $D(t)$를 정의할 수 있다.

$$D(\Lambda^{\omega}(t))=\begin{cases}-g_A(\Lambda^{\omega}(t))+g_J(\Lambda^{\omega}(t)), & \text{if current frame is Angry frame}\\ -g_J(\Lambda^{\omega}(t))+g_A(\Lambda^{\omega}(t)), & \text{if current frame is Joy frame}\end{cases} \quad (5)$$

여기서 t는 프레임 인덱스이며, g_A와 g_J은 다음과 같이 화남과 즐거움 프레임을 분류하기 위한 함수이다.

$$g_A(\Lambda^{\omega}(t))=\Lambda^{\omega}(t)-\theta \quad (6)$$
$$g_J(\Lambda^{\omega}(t))=\theta-\Lambda^{\omega}(t)$$

오류 분류 함수는 음수 값을 가질 경우 올바른 분류로 판별하며 이를 기반으로 손실함수 L을 다음과 같이 Sigmoid 형태의 함수로 정의할 수 있다.

$$L=\frac{1}{1+\exp(-\beta D(\Lambda^{\omega}(t)))} \quad (7)$$

여기서 β는 sigmoid 함수의 기울기를 나타낸다. 구하고자 하는 최적 가중치 ω_i는 Generalized Probabilistic Descent (GPD) 알고리즘에 기반하여 손실함수 L의 값이 최소가 될 때 구해지게 된다.

Ⅳ. 실험결과 분석 및 비교

본 논문에서 제안한 감정 인식의 성능 평가를 위해서 실제 감정 분류에 따른 데이터를 수집하였다. 화남(Angry), 즐거움 (Joy), 중립 (Neutral)에 관련된 음성과 GSR 데이터를 남자 8명, 여자 4명에 대해 20분씩 정보를 모았으며, 표 1과 같이 모델 구성을 위한 Training과 테스트를 위해 사용되어 졌다.

첫 번째로 수집된 데이터를 사용하여 기존의 음성만을 이용한 특징벡터 (28차)를 사용할 경우 인식성능을 알아보기 위해 GMM을 구성하여 실험을 하였으며, 모든 실험에 사용된 GMM은 16개의 혼합성분을 사용하였다. 실험 결과 평균 79.62 %의 인식 성능을 볼 수 있었으며, 감정 별 인식 성능은 표 2에 나타내었다.

다음으로 음성만을 이용한 특징벡터를 보안하기 위해 바디센서로부터 얻어진 GSR을 추가하여 특징벡터 (29차)를 구성하였다. 실험 결과 87.73 %의 인식 성능을 볼 수 있었으며, 이는 음성만을 이용할 경우 보다 약 8 % 향상된 인식 성능을 보였다. 특히 중립 감정의 경우 엄청난 상승을 보였는데 이는 GSR이 화난 감정과 즐거운 감정을 구분하는데 그리 큰 영향을 주지는 않지만 중립 감정의 경우 뚜렷한 분포차를 나타내기 때문이다. 관련하여 표 3은 GSR이 추가된 인식 실험 결과를 보여주며, 그림 3은 3가지 감정에 대한 GSR의 히스토그램을 나타낸다.

최종적으로 멀티 모달 시스템으로 부터 입력받은 신호로 부터 추출된 29차의 특징벡터를 사용한 감정인식에서 중립 감정을 분류한 뒤 성능 향상을 위해 제안된 MCE를 이용한 변별적 가중치가 적용된 GMM의 방법을 사용하여 화남과 즐거움을 구분한다. MCE를 사용하여 구해낸 GMM의 혼합성분 별 가중치는 그림 4와 같다. EM 알고리즘으로부터 구해진 GMM의 혼합성분에 대한 가중치는 1, 2, 9, 10, 11에서 혼합성분에서 큰 가중치를 나타냈으며 5, 6, 12에서 중간 정도의 가중치를 보였고 나머지의 경우 작은 가중치를 보였다. 구해진 가중치를 사용하여 구성된 GMM과 멀티 모달 시스템을 기반으로 추출된 29차의 특징벡터를 사용하여 그림 1과 같이 첫 번째 인식 결과가 중립의 감정이 아닐 경우 화남과 즐거움에 대한 감정 인식을 다시 실시하였다. 실험 결과 표 4와 같이 MCE를 사용하기 전보다 평

표 1. 수집된 데이터에 대한 세부 정보
Table 1. Information of collected data.

	Database (Voice & GSR)	
	Training	Test
인원	4 Male, 2 Female	4 Male, 2 Female
시간	Angry 20 min, Joy 20 min, Neutral 20 min	Angry 20 min, Joy 20 min, Neutral 20 min

표 2. 음성만을 이용한 감정 인식 결과 (%)
Table 2. Emotion recognition result base on voice.

Accuracy	Angry	Joy	Neutral
Angry	90.45	2.71	6.84
Joy	9.42	79.04	11.54
Neutral	19.96	10.66	69.38
Average accuracy : 79.62 %			

표 3. 음성과 GSR을 사용한 감정 인식 결과 (%)
Table 3. Emotion recognition result using fusion feature vector based on voice and GSR.

Accuracy	Angry	Joy	Neutral
Angry	83.71	13.22	2.07
Joy	15.06	83.36	1.58
Neutral	2.65	1.23	96.12
Average accuracy : 87.73 %			

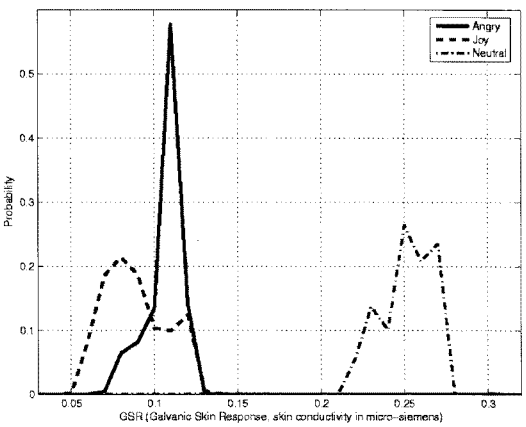

그림 3. 감정에 따른 GSR 분포
Fig. 3. Histogram of GSR according to emotion.

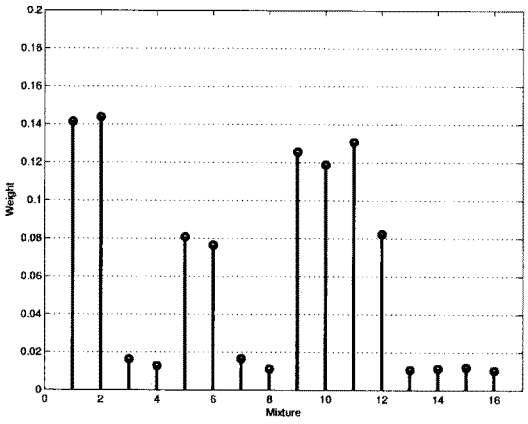

그림 4. GMM의 혼합성분에 따른 가중치 분포
Fig. 4. Weights distribution according to Gaussian mixtures.

표 4. 최종적으로 제안된 감정 인식 결과 (%)
Table 4. Emotion recognition result according to the proposed method.

Accuracy	Angry	Joy	Neutral
Angry	86.71	13.22	2.07
Joy	12.91	85.51	1.58
Neutral	2.65	1.23	96.12
Average accuracy : 89.45 %			

균 인식률이 향상된 것을 보였다. 또한 음성만의 특징 벡터에 MCE를 적용한 평균 인식 결과인 82.83 %와 비교하였을 때 감정인식에서 GSR이 훌륭한 특징벡터로 사용될 수 있음을 나타낸다.

V. 결 론

본 논문에서는 최소 분류 오차 기법과 멀티 모달 시스템을 기반으로 감정 인식 알고리즘을 제안하였다. 기존의 음성만을 이용한 감정 인식을 보안하기 위해 GSR을 이용한 멀티 모달 시스템을 적용하였으며, 구분 짓기 어려운 감정의 경우 MCE를 이용한 변별적 가중치를 적용하는 방법 등을 제시하였다. 실험 결과 제안한 방법이 기존의 방법들에 비해 우수한 성능을 보인 것을 알 수 있었다. 또한 감정 인식 전단부에 우수한 성능의 성별인식기를 추가한다면 보다 효과적인 감정 인식을 수행할 수 있을 것이라 생각되며, 보다 효과적인 특징 벡터와 인식 기법에 대한 다양한 연구와 시도가 진행되어야 할 것이라 생각 된다.

참 고 문 헌

[1] Q. Ji, P. Lan and C. Looney, "A Probabilistic Framework for Modeling and Real-Time Monitoring Human Fatigue," *IEEE Transaction on systems, man, and cybernetics Part A : Systems and humans*, vol. 36, no. 5, pp. 862-875, Sep. 2006.

[2] S. Casale, A. Russo and S. Serrano, "Multi-Style Classification of Speech Under Stress Using Feature Subset Selection Based on Genetic Algorithms," *Speech Communication*, vol. 49, no. 10-11, pp. 801-810, Oct. 2007.

[3] R. Faltlhauser, T. Pfau, G. Ruske, "On-line Speaking Rate Estimation Using Gaussian Mixture Models," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1355-1358, June 2000.

[4] O. Kwon, K. Chan, J. Hao and T. Lee, "Emotion Recognition by Speech Signals," *Eurospeech*, pp. 125-128, Sep. 2003.

[5] S. Ramamohan and S. Dandapat, "Sinusoidal Model-Based Analysis and Classification of Stressed Speech," IEEE Transactions on audio, speech, and language processing, vol. 14, no. 3, pp. 737-746 , May 2006.

[6] J. -H. Song, K. -H. Lee, J. -H. Chang, J. K. Kim and N. S. Kim, "Analysis and Improvement of speech/music classification for 3GPP2 SMV based on GMM," *IEEE Signal Processing Letters*, vol. 15, pp. 103-106, Jan. 2008.

[7] BodyMedia Armband http://www.bodymedia.com

[8] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, UK, 1995.

[9] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, John Wiley & Sons, INC., 2001.

[10] Kang S. -I., Jo Q. -H., Chang J. -H., "Discriminative Weight Training for A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, vol. 15, pp. 170-173, Feb. 2008.

저 자 소 개

이 계 환(정회원)
2007년 인하대학교 전자전기 공학부 학사.
2007년~현재 인하대학교 전자공학과 석사과정.
<주관심분야 : 디지털신호처리>

장 준 혁(정회원)
1998년 경북대하교 전자공학과 학사.
2000년 서울대학교 전기공학부 석사.
2004년 서울대학교 전기컴퓨터 공학부 박사.
2000년~2005년 (주)넷더스 연구소장
2004년~2005년 캘리포니아 주립대학, 산타바바라(UCSB) 박사후연구원
2005년 한국과학기술연구원(KIST) 연구원
2005년~현재 인하대학교 전자공학부 조교수
<주관심분야 : 음성 신호처리, 오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 인터페이스>