ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 64, No. 9, pp. 1363~1368, 2015
http://dx.doi.org/10.5370/KIEE.2015.64.9.1363

음성신호기반의 감정분석을 위한 특징벡터 선택

Discriminative Feature Vector Selection for Emotion Classification Based on Speech

최 하 나* · 변 성 우** · 이 석 필†
(Ha-Na Choi · Sung-Woo Byun · Seok-Pil Lee)

Abstract - Recently, computer form were smaller than before because of computing technique's development and many wearable device are formed. So, computer's cognition of human emotion has importantly considered , thus researches on analyzing the state of emotion are increasing. Human voice includes many information of human emotion. This paper proposes a discriminative feature vector selection for emotion classification based on speech. For this, we extract some feature vectors like Pitch, MFCC, LPC, LPCC from voice signals are divided into four emotion parts on happy, normal, sad, angry and compare a separability of the extracted feature vectors using Bhattacharyya distance. So more effective feature vectors are recommended for emotion classification.

Key Words :Bhattacharrya distance, Pitch, MFCC, LPC, LPCC, Emotion classification

1. 서 론

최근 컴퓨팅 기술의 발전으로 컴퓨터의 형태는 점점 소형화됐고, 항상 지니고 다닐 수 있는 각종 Wearable Device들이 생겨났다. 컴퓨터의 형태가 변함에 따라서 필요한 휴먼 인터랙션 작용의 종류도 다양해 졌고, 다양한 지능형 서비스가 요구되고 있다. 이에 따라 지능형 서비스를 위한 인공지능에 관한 연구가 활발하게 진행되면서 사람의 감정정보를 기기가 인식하여 사람과 적절한 인터랙션 작용을 하는 것 또한 중요해 지고 있다. 인간은 상대방에게 자신의 감정을 얼굴표정, 음성, 몸짓 등을 통한 다양한 방법으로 표현하는 이유로 영상, 음성, 생체신호 등의 매체를 통해 인간의 감정정보를 인식, 판별하기 위한 여러 분야에서의 연구가 활발히 진행되고 있다[1-5].

영상분야에서는 인간의 시각체계를 모사하여 인간의 감정을 인식하기 위해 인간의 얼굴 표정에서 여러 가지 특징(눈썹, 눈, 코, 입)의 움직임을 이용하여 감정을 인식 한다[1]. 생체신호 기반의 감정인식 분야에서 가장 많이 사용되는 생체 신호인 EEG 신호가 대뇌의 감정조절 영역에 관련이 있다는 것이 밝혀지면서 EEG 신호를 이용한 감정인식 연구가 진행 되고 있다[2, 3]. 뿐만 아니라 더 정확한 감정인식을 위해서 EEG신호와 몸짓데이터를 통합해서 연구가 진행되기도 했다[4]. 영상, 생체신호와 더불어 음성 신호 또한 인간의 감정 정보를 많이 가지고 있다. 감정별로 다른 음성의 주파수대역, 감정 별로 다른 음성의 크기 등으로 인간은 상대방의 감정을 인식한다. 이러한 이유로 음성신호기반의 감정인식 연구도 활발히 진행되고 있다[5].

이러한 연구에서 감정인식 정확도를 높이기 위해서는 정확한 분류엔진과 적절한 특징벡터를 선택하는 것이 중요하다[6]. 따라서 본 논문에서는 음성 신호기반의 감정을 분석하기 위한 가장 적절한 특징벡터를 선택하고자 한다. 이를 위하여 사람의 감정을 normal, happy, sad, angry 4가지로 분류한 뒤 방송매체를 통하여 각각의 감정에 대한 음성을 녹음하여 DB를 구성한다. 수집한 감정데이터들은 Pitch, MFCC, LPC, LPCC 4가지 특징벡터를 사용하여 분석한다. 이러한 특징벡터들의 감정분류에의 적합도를 측정하기 위해 분리도를 측정한다. 분리도는 Bhattacharyya 거리 측정[7, 8]을 이용하고 이를 통해 가장 적합한 특징벡터를 제시한다.

2. 실험 데이터

본 연구를 위한 실험 데이터는 사람의 감정에 대한 음성데이터이다. 실험을 위해 기준이 되는 감정들은 서로 명확하게 분리가 되는 것으로 normal, sad, happy, angry 4가지 감정으로 분류였다. 각각의 감정에 대한 데이터는 총 25명, 감정당 3~5개씩 수집하여 총 75~125개의 데이터를 취득을 한 뒤 음성구간을 추출하는 전 처리 과정을 거쳤다. 데이터 취득을 표 1에 상세 표기하였다.

정확한 데이터를 위해서 방송매체를 통해서 , 잡음이 섞이지 않도록 녹음하여 데이터를 수집하였다. 데이터 취득 대상은 여자/남자로 구분된다. 데이터를 취득할 때 데이터들은 음성만으로 감정구분이 명확한 데이터, 배경음이 없는 데이터들로 선정하였다. 각 음성 데이터들은 표본화를 16kHz로 하고 window size를 500으로 하여 각 윈도우당 32ms가 할당된다.

† Corresponding Author : Dept. of Media Software, Sangmyung University, Korea
E-mail : esprit@smu.ac.kr
* Dept. of Media Software, Sangmyung University, Korea
** Dept, of Computer Science, Sangmyung University, Korea
Received : June 8, 2015; Accepted : August 24, 2015

표 1 취득된 데이터
Table 1 Acquired data

	normal	sad	happy	angry
남자1	3	3	4	5
남자2	3	3	3	3
남자3	3	3	3	3
남자4	3	3	3	3
남자5	3	4	3	5
남자6	4	3	3	4
남자7	3	3	3	4
남자8	3	3	3	5
남자9	3	3	3	3
남자10	5	5	3	5
남자11	3	4	3	5
남자12	3	3	4	5
여자1	3	4	3	4
여자2	3	4	3	3
여자3	3	4	3	4
여자4	3	3	3	5
여자5	3	3	3	3
여자6	3	3	3	3
여자7	3	3	3	3
여자8	3	3	3	3
여자9	3	4	3	3
여자10	3	3	3	4
여자11	3	3	3	3
여자12	4	4	4	4
여자13	5	4	3	5
총 25명	81	85	78	97

녹음된 데이터들은 음성구간을 찾는 전 처리 과정을 거치게 된다. 음성구간을 찾는 것은 불필요한 정보가 될 수 있는 비 음성 구간을 제거하여 정확한 특징벡터를 추출하는 이유로 음성신호 기반의 감정인식에서 중요한 부분이다.

음성신호 구간은 비음성신호 구간에 비해 신호의 에너지 값이 크기 때문에 에너지 크기의 값을 반영하는 절대 적분치 (Integral Absolute Value) 특징벡터를 사용하였으며 식은 다음과 같다.

$$\bar{x} = \sum_{i=1}^{N} |X(i\triangle t)| \tag{1}$$

여기에서,
X : 측정된 신호,
△t : 샘플링 시간 간격,
N : 샘플의 수,
i : 샘플의 순서

IAV 임계값을 선택하는 과정은 신호에서의 IAV특징 벡터를 추출 한 후 최대값 최소값을 구한 후, 최대값 최소값 차의 10% 만큼 최소 값의 위로 잡는다. 만약 최소값이 최대값의 70%보다 크면 임계값은 최대값의 20% 아래로 잡는다. 신호의 크기 임계값은 IAV 임계값에서 프레임 크기로 나눠주어서 구하게 된다. IAV 특징벡터가 프레임내의 모든 신호 값의 절대치를 더한 값이기 때문에 프레임 크기로 나눠주게 되면 프레임의 신호 평균값이 나오게 된다. 따라서 이 값은 IAV 임계값을 신호의 크기 임계값으로 바꾼 값이 된다.

음성 구간을 추출하는 과정은 프레임 단위로 IAV 임계 값 보다 큰 구간이 나오면 해당 프레임 내에서 신호 에너지 임계 값 보다 커지는 지점을 시작 인덱스로 선정하고 시작 인덱스부터 IAV 임계치가 작아지는 구간이 나오면 그 지점을 끝 인덱스로 선정하게 된다. 위 방법을 사용하게 되면 정확하게 음성 구간을 추출할 수 있다.

3. 특징벡터

본 논문에서 쓰일 특징벡터 중 Pitch는 주기신호의 기본주파수를 의미한다. Pitch를 검출하기 전에 전 처리 과정이 있다. 음성신호는 사람의 발음에서 나오는 파열음, 파찰음, 마찰음, 경음 등 피치와 관련 없는 고주파 성분인 무성음 구간이 존재하게 된다. 이는 피치 검출에서 반드시 제거해 주어야 정확도를 높일 수 있으며 이 부분은 전체 시스템 정확도를 떨어뜨릴 수 있는 부분이다. 따라서 무성음은 자기상관 값을 정규화 한 값이 임계값보다 작으면 주기성이 약한 신호이기 때문에 무성음이라 할 수 있다. 본 연구에서는 경험적으로 임계값을 0.55로 정하였다. Normalized Autocorrelation는 식 (2)와 같이 나타낸다.

$$\text{Normalized Autocorrelation} = \frac{R_{s1s2}}{\sqrt{E1 * E2}} \tag{2}$$

여기에서
R : 자기상관 함수
S : 시간영역 신호
E : 신호의 에너지 값

전처리 과정을 거친 후 Pitch를 검출하게 되는데 일반적으로 사람의 Pitch는 80Hz~ 500Hz에 존재하게 된다. 따라서 신호는 80Hz~500Hz까지 주기를 늘리면서 자기상관 값이 가장 큰 주기를 찾게 된다. 상관도가 가장 높은 주기는 그 신호의 기본 주파수가 된다.

위 과정을 거쳐 Pitch를 검출하게 되면 유성음 검출 과정과 같은 여러 가지 이유로 정확하지 않은 Pitch 값이 존재하게 된다. 이러한 값들은 분류 시스템 전체의 성능을 낮출 수 있기 때문에 노이즈를 제거하는 Smoothing 작업이 필요하다. 본 연구에서는 일반적으로 많이 사용되는 평활화 기술인 median filter를 사용하였다.

다음으로 MFCC(Mel-Frequenct Cepstrum Coefficient)란 frame내의 음성 신호에 대하여 계산한 스펙트럼을 청각기의 주

파수 반응도를 모사한 Mel-scale 주파수 도메인에서 discrete cosine transform (DCT)를 취한 값이다. 일반적으로 이를 추출하기 위한 자세한 과정은 아래 순서도와 같다.

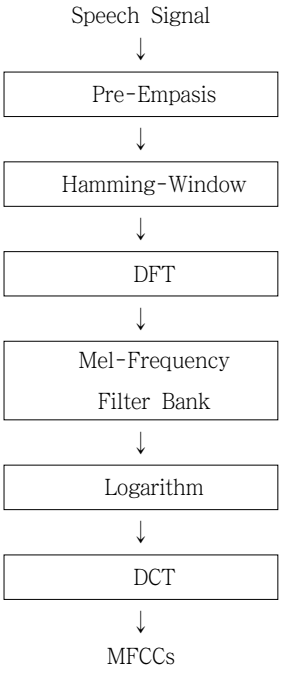

그림 1 MFCC 추출 순서도
Fig. 1 Flow chart of MFCC extraction

LPC는 선형 결합에 의해 과거의 신호에서 현재의 신호[n]을 예측하는 방법으로 전극(All-pole) 모델을 사용하여 식 (3)과 같이 차분 방정식의 형태로 나타낼 수 있다. 여기서 S_n은 입력신호, $\widetilde{S_n}$은 예측신호, a_i는 선형예측계수이며, p는 예측계수의 차수이다. 현재신호와 예측된 신호의 예측오차는 식 (4)와 같다.

$$\tilde{S}=-(a_1 S_{n-1}+a_2 S_{n-2}+\ldots+a_p S_{n-p}) \tag{3}$$

$$e_n = S_n - \widetilde{S_n} \tag{4}$$

식 (5)는 예측오차 에 대한 mean square error(MSE) J이며 J를 최소로 하는 선형예측계수를 찾기 위하여 식 (5)를 a_i에 대하여 편미분하여 0이 되는 p개의 선형 연립방정식 (6)을 얻을 수 있다[9]. 식 (6)은 식 (7)과 같이 나타낼 수 있고, 선형예측계수는 자기상관 행렬의 역행렬을 이용하여 구할 수 있다.

$$J=E[e^2(n)]=E[S_n-\widetilde{S_n^2}] \tag{5}$$

$$\sum_{j=1}^{p} a_j E_s(n-i)s(n-j)=E_s(n-i)s(n)$$
$$\text{for} i=1\ldots p \tag{6}$$

$$\begin{bmatrix} R_0 & \cdots & R_{p-1} \\ \vdots & \ddots & \vdots \\ R_{p-1} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix} \tag{7}$$

LPCC는 C(z)의 inverse z-transform으로 정의되고 식은 다음과 같다.

$$C(z)=\sum_n c(n) z^{-n} \tag{8}$$

전극(All-pole) z=z_i가 unit cycle 안에 있고, gain값을 1로 주면 LPCC (c_{ip}(n))는 다음과 같이 정의 된다.

$$n>0\; C_{lp}(n)\begin{cases} \frac{1}{n}\sum_{i=1}^{p} z_i^n \\ 0 \end{cases} \qquad n<0 \tag{9}$$

LPCC는 recursive에 의해 예측계수 값으로부터 구한다.

$$c_1 = a_1 \tag{10}$$

$$c_n\begin{cases} \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{n-k}+a_n & 1<n<p \\ \sum_{k=1}^{n-1}\frac{k}{n}c_k a_{n-k} & n>p \end{cases}$$

4. 실 험

본 논문에서 쓰이는 Pitch, MFCC, LPC, LPCC 4가지의 특징벡터들의 분리도를 비교하기 위하여 Bhattacharyya 거리측정을 이용하였다.

Bhattacharyya거리 측정은 오류 율을 측정하여 거리를 계산하는 방법으로 각 클래스의 분포가 가우시안 형태를 가질 때 가장 좋은 평가기준이 된다. Bhattacharyya거리의 식은 다음과 같다.

식에서 M_1, M_2 : 클래스 1, 2의 평균, $\sum_1$,$\sum$: 클래스 1, 2의 공분산이다.

$$\mu(1/2)=\frac{1}{8}(M_2-M_1)^T\left\{\frac{\sum_1+\sum_2}{2}\right\}^{-1}(M_1-M_2)+\frac{1}{2}ln\frac{\left|\frac{\sum_1+\sum_2}{2}\right|}{\sqrt{|\sum_1||\sum_2|}} \tag{11}$$

Bhattacharyya 거리가 가장 큰 값이 나온 특징벡터가 클래스 간의 거리가 가장 멀리 떨어져있다는 의미로, 감정인식에 가장 적합하다고 할 수 있다.

실험데이터들의 음성구간을 구하는 전 처리 과정을 거친 후 실험데이터들의 음성구간으로부터 특징벡터 Pitch, MFCC, LPC, LPCC를 추출하였다. MFCC, LPC와 LPCC의 차수는 기존 연구들과 같이 10차로 정하였다[10]. 이렇게 추출된 특징벡터들의 분리도를 Bhattacharyya 거리 측정을 통하여 비교하였다.

그림 2는 남자의 음성신호에서 Pitch특징벡터를 ,그림 3은 MFCC특징벡터를, 그림 4는 LPC를 그림 5는 LPCC를 추출한 후 Bhattacharyya 거리를 구하여 분리도를 비교한 결과이다.

남성음성의 경우 sad와 normal, sad와 angry 사이에 분리도가 Pitch 특징벡터를 이용했을 때 약 25로 다른 특징벡터를 이용했을 때 보다 크게 나왔다. normal과 angry 사이에 분리도는 Pitch를 제외한 3가지 특징벡터 모두 0.5이하로 나타났지만, Pitch에서는 약 20으로 분리도가 높게 나타났다. happy 는 normal, sad와의 분리도가 Pitch를 이용했을 때 약 25로 다른 특징벡터를 사용했을 때 보다 나타났다. normal과 sad, sad와 angry, happy와 angry 이 3가지 경우는 4가지 특징벡터 모두

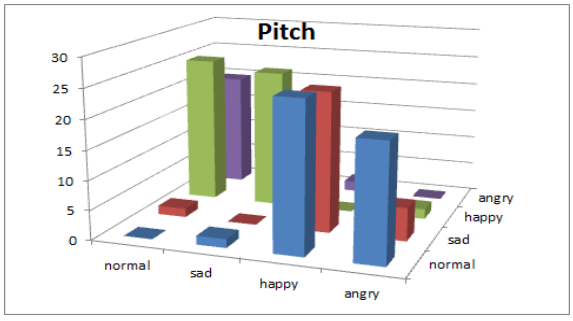

그림 2 남자음성 Pitch 분석도
Fig. 2 Pitch analysis of man's voice

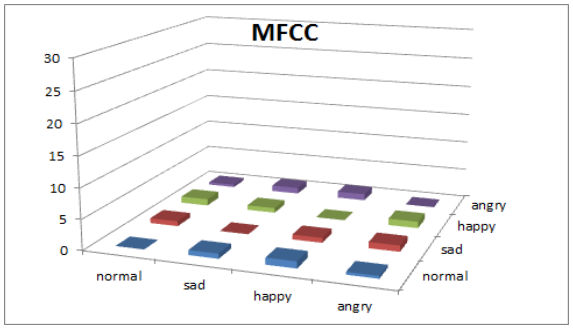

그림 3 남자음성 MFCC 분석도
Fig. 3 MFCC analysis of man's voice

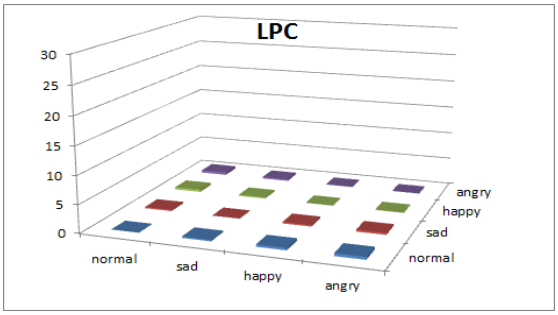

그림 4 남자음성 LPC 분석도
Fig. 4 LPC analysis of man's voice

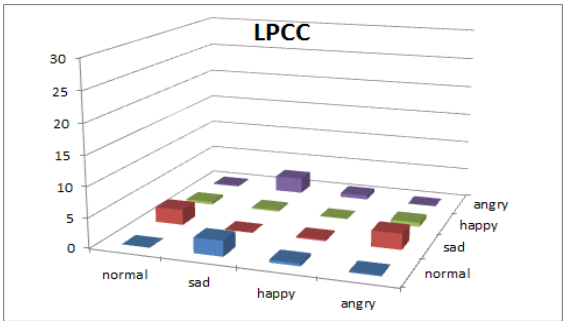

그림 5 남자음성 LPCC 분석도
Fig. 5 LPCC analysis of man's voice

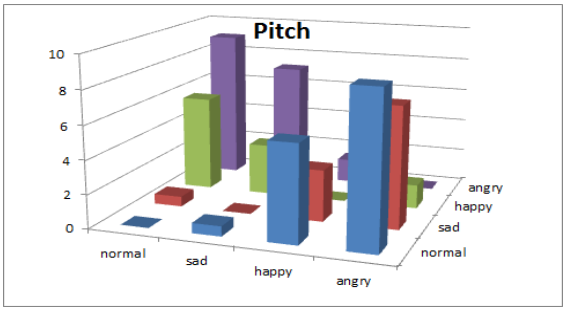

그림 6 여자음성 Pitch 분석도
Fig. 6 Pitch analysis of woman's voice

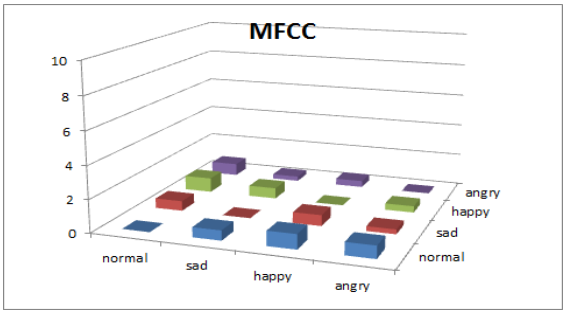

그림 7 여자음성 MFCC 분석도
Fig. 7 MFCC analysis of woman's voice

분리도가 6 이하로 낮게 나타났다. 전체적으로 Pitch 특징벡터를 이용했을 때 분리도가 큰 것으로 나타난다.

Pitch 다음으로는 LPCC 분리도가 가장 큰데, 남성의 normal과 sad의 경우는 Pitch를 이용한 Bhattacharyya 거리 측정값이 1.48738, LPCC를 이용한 Bhattacharyya 거리측정값이 2.577081으로 LPCC의 분리도가 더 크다.

남성과 마찬가지로 전체적으로 Pitch를 이용하여 구한 Bhattacharyya 거리가 다른 특징벡터들에 비해서 월등히 높은 것을 알 수 있다. Pitch 특징벡터로는 남성의 감정분리와 비슷하지만 남성음성은 sad와 happy가 sad와 angry보다 분리가 잘 되는 것과 반대로 여성음성은 sad와 happy보다 sad와 angry가 분

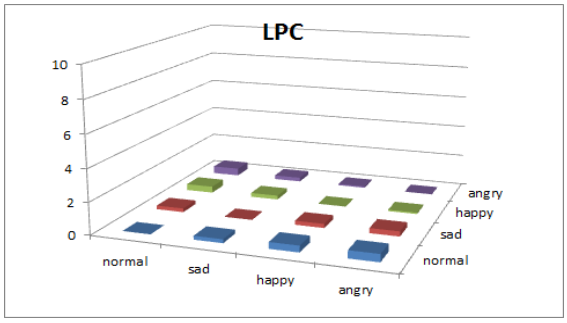

그림 8 여자음성 LPC 분석도
Fig. 8 LPC analysis of woman's voice

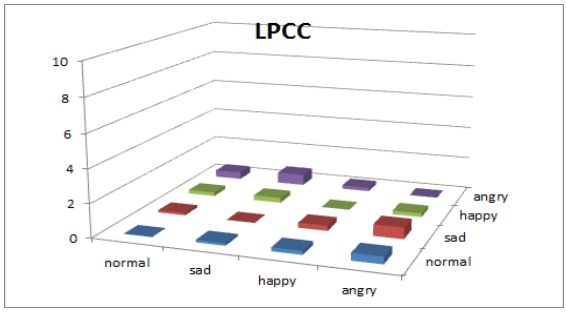

그림 9 여자음성 LPCC 분석도
Fig. 9 LPCC analysis of woman's voice

리가 더 잘된다.

Pitch 다음으로는 다 비슷하기 때문에 Bhattacharyya 거리 값의 평균으로 비교해 볼 때 MFCC 특징벡터에서 0.60618, LPC 특징벡터에서 0.286124, LPCC 특징벡터에서 0.337511으로 MFCC를 통한 여성음성 분리도가 2번째로 높다.

표 2 특징벡터에 따른 Bhattacharyya 거리 값의 평균
Table 2 Average value of Bhattacharyya distance by discriminative feature vectors

	normal	sad	happy	angry
남성	3	3	4	5
여성	3	3	3	3

표 2는 남성과 여성의 각각의 특징벡터에 따른 Bhattacharyya 거리 값의 평균이다. 남성과 여성 모두 Pitch 특징벡터를 이용했을 때 분리도가 가장 크게 나타났다. 따라서 남성과 여성 모두 Pitch 특징벡터가 감정분리를 위해 적합하다고 판단된다. 그리고 여성의 Pitch 평균 값과 남성의 Pitch 평균 값을 비교해 보았을 때 여성보다 남성의 음성 분리도가 크게 나왔다. 남성의 음성이 여성의 음성보다 감정 분리가 잘 되는 것으로 판단된다.

Pitch 다음으로는 다 비슷하기 때문에 Bhattacharyya 거리 값의 평균으로 비교해 볼 때 MFCC 특징벡터에서 0.60618, LPC 특징벡터에서 0.286124, LPCC 특징벡터에서 0.337511으로 MFCC를 통한 여성음성 분리도가 2번째로 높다. 따라서 Pitch 특징벡터가 음성신호기반의 감정분석을 위해 가장 적합한 것으로 판단된다.

5. 결 론

본 논문에서는 음성신호 기반의 감정인식을 위한 적합한 특징벡터를 선택하였다. 이를 위해 감정의 종류를 normal, sad, angry, happy 4가지로 분류하고 각각의 감정에 대한 음성데이터를 수집하였다. 데이터는 총 25명, 감정 당 3~5개씩 수집하여 총 75~125개의 데이터를 취득한 뒤 음성구간을 추출하는 전 처리 과정을 거쳤다. 취득한 데이터에서 각각 Pitch, MFCC, LPC, LPCC 특징벡터를 추출하여 각 특징벡터간의 분리도를 비교, 분석 하였다. 그 결과, 음성신호 기반의 감정인식을 위한 특징벡터로는 남성과 여성 모두 평균적으로 Pitch 특징벡터가 분리도가 크게 나와 음성신호기반의 감정분석을 위해 가장 적합한 것으로 나타났다. 하지만 감정에 따라서 Pitch를 사용하더라도 분리도가 낮은 감정들이 있었다. 남성의 경우는 normal과 sad, sad과 angry, happy과 angry의 분리도가 비교적으로 낮게 나타났고, 여성의 경우는 normal과 sad, sad과 happy, happy과 angry의 분리도가 비교적으로 낮게 나타났다. 그리고 전체적으로 여성보다 남성의 음성이 분리도가 크게 나타났다.

향후, 본 논문의 실험 결과를 바탕으로 음성신호 기반의 감정인식 시에 감정별로 분리도가 높은 적합한 특징벡터를 사용하여 더 효과적인 감정 인식 결과를 기대 할 수 있다.

감사의 글

본 연구는 2015년도 상명대학교 교내연구비를 지원받아 수행하였음

References

[1] Dong-Hoon Lee and Kwee-Bo Sim "Emotion Recognition and Expression System of Robot Based on 2D Facial Image" Journal of Control, Automation, and Systems Engineering Vol.13, No.4, April 2007.

[2] Petrantonakis, P.C, Hadjileontiadis, L.J, "Emotion Recognition From EEG Using Higher Order Crossings", IEEE Transactions on Information Technology in Biomedicine, Volume 14, Issue 2, Pages 186-197, March 2010.

[3] Kwang-Eun Ko, Hyun-Chang Yang, Kwee-Bo Sim, "Emotion recognition using EEG signals with relative power values and Bayesian network", International Journal of Control, Automation and Systems, Volume 7,

Issue 5, pp. 865-870, October 2009.
[4] Ho-Duck Kim, Tae-Min Jung, Hyun-Chang Yang, and Kwee-Bo Sim “Emotion Recognition Method using Gestures and EEG Signals” Journal of Institute of Control, Robotics and Systems Volume 13 Issue9, 2007. 9, 832-837.
[5] Jae Hun Bang, SungYoung Lee "Call Speech Emotion Recognition for Emotion based Services" Journal of KIISE: software and application Volume 41 Issue 3 (2014.3) pp. 208-213.
[6] Hyun Woo Kim, Sung Yong Lee “The Phoneme Kernel Technique based on Support Vector Machine for Emotion Classification of Mobile Texts“ Journal of KIISE: software and application Volume 40 Issue 6 (2013.6) 350-355.
[7] Sang-Yeob Oh, “Improving Phoneme Recognition based on Gaussian Model using Bhattacharyya Distance Measurement Method”, Journal of Korea Multimedia Society Vol 14. No.1. Jenuary 2011 pp. 85-93.
[8] Seok-Pil Lee, Sang-Hui Park, Jeong-Seop Kim, Ig-Jae Kim “EMG pattern recognition based on evidence accumulation for prosthesis control”, Proc Ann Intl Conf IEEE Eng Med Biol 4, pp. 1481-1483, 1996.
[9] Padmasai, Y외 3인, “Linear Prediction Modelling for the Analysis of the Epileptic EEG”, IEEE-Advances in Computer Engineering (ACE), 2010 International Conference on, pp. 6-9, June 2010.
[10] Kwang-Seung Heo, Chang-Hyun Park, Dong-Wook Lee, and Kwee-Bo Sim "speaker identification using incremental neural network and LPCC" Journal of The Korean Institute of Intelligent Systems Volume 12 Issue 2, 2002.12. pp. 341-344.

저 자 소 개

최 하 나 (Ha-Na Choi)

2013년 ~ 현재 상명대학교 미디어소프트 웨어학과 학부 과정
〈주관심분야〉 멀티미디어처리, 인공지능, 음성신호처리

변 성 우 (Sung-Woo Byun)

2014년 상명대학교 디지털미디어학과 이학사. 2014년 ~ 현재 상명대학교 컴퓨터과학과 석·박사 통합과정
〈주관심분야〉 멀티미디어처리, 인공지 능, 음성신호처리

이 석 필 (Seok-Pil Lee)

1990년 연세대학교 전기공학과 공학사. 1992년 연세대학교 전기공학과 공학석사. 1997년 연세대학교 전기공학과 공학박사. 1997년 ~ 2002년 대우전자 영상연구소 선 임연구원. 2002년 ~ 2012년 KETI 디지털미디어 연구센터 센터장. 2010년 ~ 2011년 미국 Georgia Tech. 방문연구원. 2012년 ~ 현재 상명대학교 디지털미디어학과 교수
〈주관심분야〉 멀티미디어 검색, 방송통신시스템, 인공지능