ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 65, No. 1, pp. 116~121, 2016
http://dx.doi.org/10.5370/KIEE.2016.65.1.116

IoT를 위한 음성신호 기반의 톤, 템포 특징벡터를 이용한 감정인식

Emotion Recognition Using Tone and Tempo Based on Voice for IoT

변 성 우* · 이 석 필†
(Sung-Woo Byun · Seok-Pil Lee)

Abstract - In Internet of things (IoT) area, researches on recognizing human emotion are increasing recently. Generally, multi-modal features like facial images, bio-signals and voice signals are used for the emotion recognition. Among the multi-modal features, voice signals are the most convenient for acquisition. This paper proposes an emotion recognition method using tone and tempo based on voice. For this, we make voice databases from broadcasting media contents. Emotion recognition tests are carried out by extracted tone and tempo features from the voice databases. The result shows noticeable improvement of accuracy in comparison to conventional methods using only pitch.

Key Words : Emotion recognition, Speech, Tone, Tempo

1. 서 론

컴퓨팅 기술이 발전하고 컴퓨터 사용이 대중화 되면서 컴퓨터와 사람간의 상호작용에 대한 관심이 증가됐다. 사람과 컴퓨터 사이에 주고, 받는 정보가 중요해짐에 따라 Internet of Things (IoT)와 Human Computer Interaction (HCI)에 대한 연구가 활발하게 진행되어 왔다. 이러한 연구 분야에서 사람의 감정을 컴퓨터가 인식하고 그에 따른 상호작용을 하기위한 생체신호 기반의 감정인식 기술들이 많이 연구되어 왔는데, 주로 얼굴인식 기반의 방법[1], EEG, PPG와 같은 생체 신호 기반의 방법[2, 3], 음성신호기반의 방법[4]들이 사용되었다. Carlos Busso외 8인은 normal, sad, happy, angry 4가지 감정을 얼굴과 음성을 이용하여 인식하는 멀티모달 방법[5]을 제안하였고, Daniel Neiberg외 2인은 normal, positive, negative의 3가지 감정의 음성신호에서 MFCC(Mel-frequency Cepstral Coefficients), 음성의 주파수 대역 20Hz~ 300Hz에서 계산되는 MFCC-low, Pitch 특징벡터를 추출하고, GMM(Gaussian Mixture Model) 분류기를 사용하여 분류하였다[6]. Xin Xu외 6인은 normal, sad, happy의 감정에 대한 음성신호에서 Duration, Energy, F0, Formant, HNR 특징벡터를 추출하여 분리도가 높은 특징벡터를 제안하였다[7]. Vijayan외 2인은 EEG 신호에서 Shannon Entropy 와 auto-regressive model 특징을 추출하여 happy, sad, fear 등의 감정을 분류하였다[8]. 얼굴인식 기반의 방법은 정확도가 비교적 높은 반면 실시간으로 데이터를 수집하기 어려운 단점이 존재하고, EEG, PPG와 같은 생체 신호 기반의 방법 또한 데이터를 수집하기 어렵고, 피 실험자에게 거부감이 들 수 있는 단점이 있다. 한편, 음성신호기반의 감정인식 기술의 경우 다른 생체신호를 이용한 방법보다 데이터 수집이 간편한 장점이 있어 IoT를 지원하는 착용형 기기에는 아주 적합한 기술이다.

본 논문에서는 음성신호에서 톤, 템포 정보를 이용하여 감정을 인식하였다. 이를 위하여 감정을 'normal', 'sad', 'happy', 'angry' 4가지로 분류하고, 방송매체를 통하여 각각의 감정에 대한 음성을 녹음하여 DB를 구성하였다. 수집한 음성신호에서 'normal', 'sad,' 'happy', 'angry' 감정에 대한 톤, 템포 정보를 추출하고 신경회로망을 이용하여 감정을 분류하였다.

본 논문의 구성은 다음과 같다. 2장에서는 실험에 사용된 실험 데이터에 대한 설명, 3장에서는 특징벡터, 4장에서는 실험 및 실험결과를 보여주고, 마지막으로 5장에서는 결과를 통해 결론을 맺도록 한다.

2. 실험 데이터

본 연구를 위한 실험 데이터는 사람의 'normal', 'sad', 'happy', 'angry' 4가지 감정에 대한 음성데이터로 분류하고, 정확한 데이터를 위해 방송매체를 통해서 인물의 표정과 앞뒤 상황 등을 고려하여 판단하여 취득하였다. 이때, 배경음악이나 환경잡음은 데이터를 분석하는데 방해요소가 되므로 최대한 인물의 목소리만 녹음하도록 하였고, 컴퓨터 내부의 소리만 녹음되도록 설정하였다. 데이터는 총 200개로 남성, 여성 각 100개씩 취득하였으며,

† Corresponding Author : Dept. of Media Software, Sangmyung Univerity, Korea
E-mail : esprit@smu.ac.kr
* Dept. of Computer Science, Sangmyung University, Korea
Received : October 21, 2015; Accepted : December 24, 2015

감정 당 25개씩 취득하였다. 각 음성 데이터들은 3초 ~ 5초간 녹음하였으며, 샘플링 율은 16000Hz로 설정하였고, 프레임 크기는 500으로 32ms 단위로 분석하였다.

취득한 데이터들은 음성신호에서 음성구간만을 추출하는 전처리 과정을 거치게 된다. 이는 감정인식 시스템에서 불필요한 정보가 될 수 있는 비 음성 구간을 제거하는 이유로 음성신호 기반 감정인식에서 중요한 부분이다. 음성구간을 추출하는 flow chart는 다음 그림과 같다.

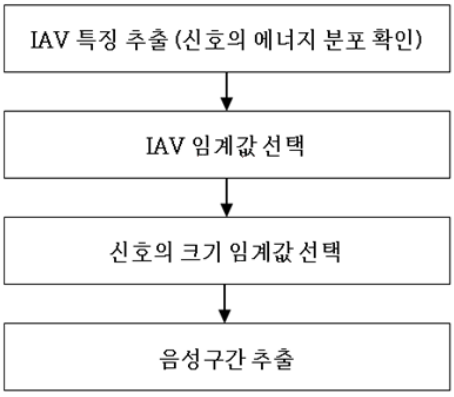

그림 1 음성구간 추출 순서도
Fig. 1 Flow chart of voice extraction

음성신호 구간은 비 음성신호 구간에 비해 신호의 에너지 값이 크기 때문에 에너지 크기의 값을 반영하는 절대 적분치 IAV (Integral Absolute Value) 특징 벡터를 사용 하였으며 식은 다음과 같다.

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

여기에서,

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

IAV 임계값을 선택하는 과정은 신호에서의 IAV특징 벡터를 추출 한 후 최댓값 최솟값을 구하고, 최댓값 최솟값 차의 10% 만큼 최솟값의 위로 잡는다. 만약 최솟값이 최댓값의 70%보다 크면 임계값은 최댓값의 20% 아래로 잡는다. 임계 값 선택하는 과정의 예시는 다음 그림 2와 같다.

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 임계값 보다

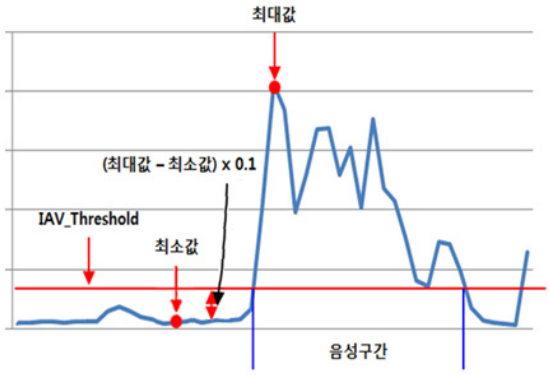

그림 2 IAV 임계 값 선택 예시
Fig. 2 Example of the IAV threshold selection

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.

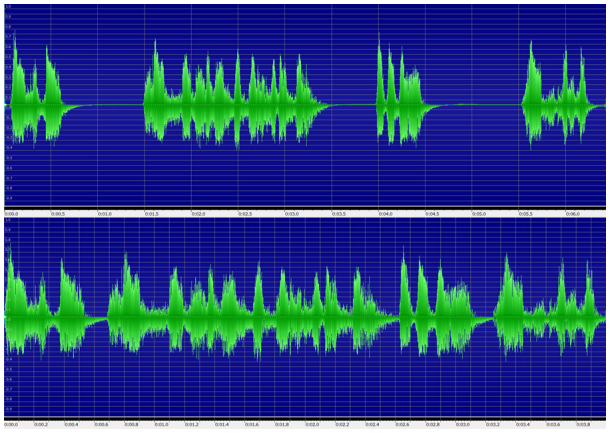

그림 3 음성 구간 추출 결과
Fig. 3 Result of voice extraction

3. 특징 벡터

3.1 톤 특징벡터

일반적으로 사람의 음성신호는 성대가 진동하여 발생하며, 준주기성 신호이다. 이러한 신호의 진동의 주기를 기본주파수(F0) 혹은 피치, 톤 이라고 한다. 음성신호의 톤은 여러 음성 신호처리 분야에서 이용되고 있는 중요한 특징으로 주로 자기상관함수 (Autocorrelation) 혹은 AMDF(Average Magnitude Difference Function) 방법을 사용하여 구한다. 본 연구에서는 음성신호에서의 톤을 구하기 위해 선행연구에서 연구된 방법을 사용하였다 [9].

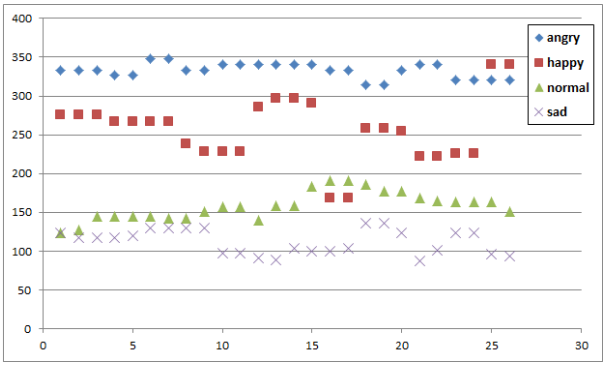

그림 4 남성 음성데이터의 톤 정보 추출 결과
Fig. 4 Result of tone extraction from male voice data

3.2 템포 특징벡터

템포는 BPM(beat per minute) 단위를 사용하며 일정한 간격으로 규칙을 띠고 반복되는 소리의 움직임의 단위인 비트(beat)가 1분 내에 들어있는 비트의 수를 의미 한다. 하지만 사람의 음성의 경우, 템포는 정량적인 리듬 혹은 사람의 말하는 빠르기를 의미한다. 따라서 음절은 사람의 말 빠르기를 인식하는 척도이며 각각의 음절을 이용하여 음성의 리듬 혹은 빠르기를 측정할 수 있다.

음절은 하나의 종합된 음의 느낌을 주는 말소리 단위이며, 하나의 자음과 모음 혹은 하나의 모음으로 이루어진 한 음의 단위를 의미한다. 본 연구에서는 음성신호에서의 모음과 자음을 추출하고 모음의 길이를 하나의 음절로 가정하였다. 음절 추출 알고리즘의 flow chart는 다음과 같다. 만일 256개의 의사 무작위 패턴이 가해졌을 때, 66개의 결정 패턴만을 가하면 고장 검출률 100%를 보장할 수 있다. 따라서 총 322개의 패턴으로 고장 검출률 100%를 얻을 수 있는 것이다. 그리고 의사 무작위 패턴을 더 많이 가하면 총 테스트 수는 길어지지만 Ld 값을 줄일 수 있다.

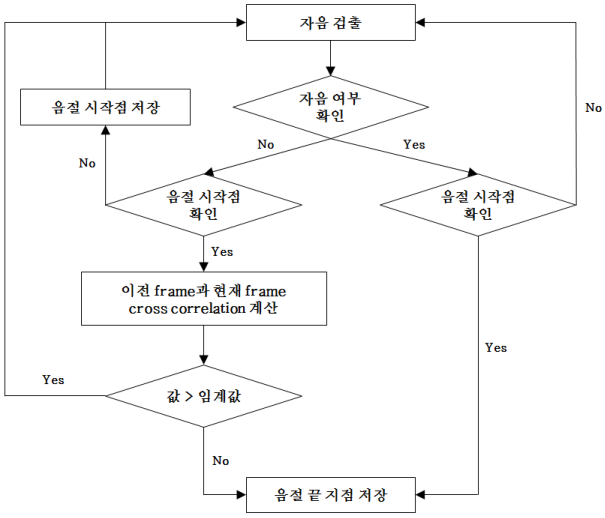

그림 5 템포 추출 순서도
Fig. 5 Flow chart of tempo extraction

일반적으로 모음은 준 주기성을 가지는 유성음 신호이고 자음은 주기성이 없는 무성음 신호이다. 따라서 원 신호와 그 신호를 지연시킨 지연신호의 에너지 값으로 두 신호의 자기상관함수 값을 정규화 했을 때, 그 값이 임계값보다 작으면 주기성이 약한 신호이므로 자음이라 할 수 있다. 임계값은 경험적으로 0.55로 정하였으며, 자기상관함수의 정규화 식은 다음과 같다.

$$Nomalized\ Autocorrelation = \frac{R_{s1s2}}{\sqrt{E_1 \times E_2}}$$
$$E_1 = \sum S_1^2 \quad E_2 = \sum S_2^2 \tag{2}$$

여기에서,

R : 자기상관함수
S_1 : 원 신호
S_2 : 원 신호의 지연된 신호
E : 신호의 에너지 값

모음은 각 모음이 가지고 있는 신호의 envelope와 주파수 성분이 다르다. 따라서 음성신호에서 음성이 변하면 시간영역의 envelope 또한 변하게 된다. 따라서 현재의 프레임과 이전 프레임간의 교차상관함수 값을 구하여 두 신호의 에너지 값으로 정규화 한 값이 임계값보다 작으면 두 신호의 상관도가 떨어지기 때문에 envelope가 변했다고 할 수 있다. 임계값은 경험적으로 0.55로 정하였으며, 교차상관함수의 정규화 식은 다음과 같다.

$$Nomalized\ Crosscorrelation = \frac{R_{s1s2}}{\sqrt{E_1 \times E_2}}$$
$$E_1 = \sum S_1^2 \quad E_2 = \sum S_2^2 \tag{3}$$

여기에서,

R : 교차상관함수
S_1 : 이전 프레임의 신호
S_2 : 현재 프레임의 신호
E : 신호의 에너지 값

음절을 추출하게 되면 결과는 한 모음에 대한 프레임 개수가 구해진다. 한 프레임은 32ms이기 때문에 프레임단위를 ms 단위로 변환이 가능하고, 한 문장에서 추출된 음절의 길이의 평균값을 템포로 추출하였다. 템포는 신호의 envelope를 직접 반영하여 구하기 때문에 음성의 주변 환경에 민감하지만, 음성신호 기반의 감정인식에서 중요한 요소이다[10].

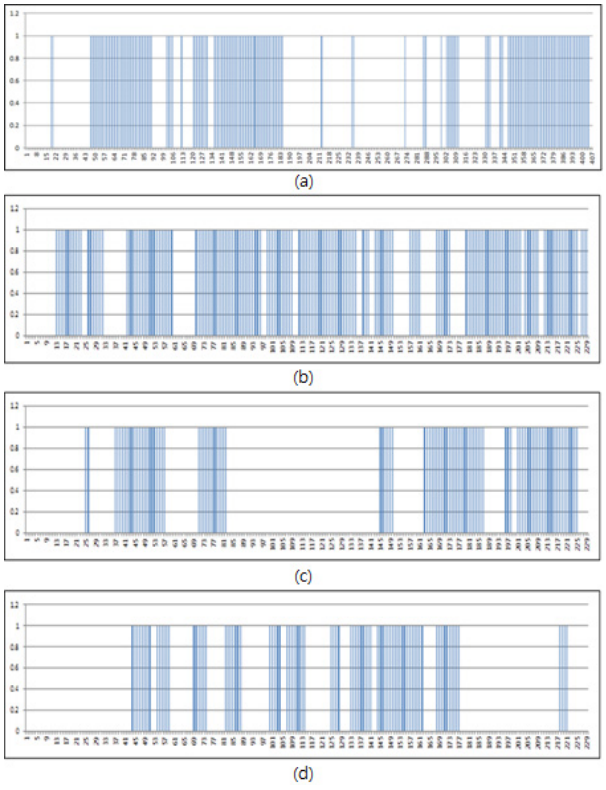

그림 6 남성 음성데이터의 음절 추출 결과 위에서부터 (a) 'normal' 감정에 대한 결과, (b) 'sad' 감정에 대한 결과, (c) 'happy' 감정에 대한 결과, (d) 'angry' 감정에 대한 결과

Fig. 6 Result of tempo extraction using male voice data. (a) nomal emotion, (b) sad emotion, (c) happy emotion, (d) angry emotion

4. 실험 및 실험결과

4.1 실험 구성

본 논문에서는 톤, 템포 정보를 이용한 감정인식 성능을 평가하기 위해 신경회로망 (Neural Network) 기반의 분류기를 사용하여 감정인식 시스템을 구현하였다. 시스템의 구성은 그림 7과 같다.

입력된 음성신호에서 프레임단위로 톤, 템포 정보를 추출한다. 추출된 톤, 템포는 별도의 파라미터를 추출하게 되는데 본 연구에서는 한 문장에서의 톤의 평균과 분산, 템포의 평균과 분산을 사용하였다. 이렇게 구성된 톤의 특징벡터와 템포의 특징벡터는 각각의 신경회로망을 사용하여 결과를 출력하고 2개의 결과를 또 다른 신경회로망을 거쳐 최종 감정을 인식하게 된다. 신경회로망의 학습률 ρ는 0.6으로 설정하였으며 은닉 층은 1개, 은닉 층의 노드 p는 5개로 설정하였다.

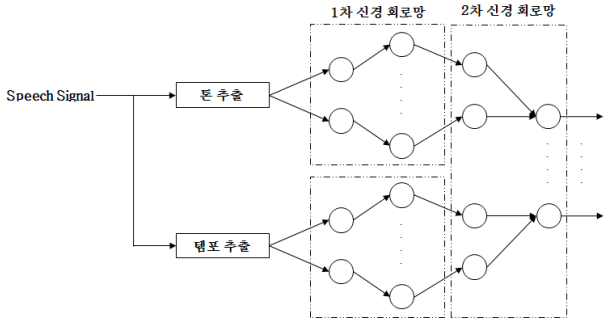

그림 7 음성신호 기반 감정인식 시스템 구성

Fig. 7 Emotion recognition system architecture based on speech.

4.2 실험 결과

실험은 남성, 여성에 대하여 2번 진행하였다. 'normal', 'sad', 'happy', 'angry' 4가지 감정에 대한 100개의 데이터에서 각 감정의 데이터 15개, 총 60개의 데이터로 학습을 진행하였고, 각 감정의 데이터 10개, 총 40개의 데이터로 인식을 진행하였다. 또한, 교차검증을 통해 총 5번 수행하였으며 300개의 데이터가 학습에 사용되었으며, 200번 인식과정을 진행하였다.

표 1 전체 데이터에 대한 인식률

Table 1 Recognition rate for total data

	normal	sad	happy	angry	accuracy
normal	68%	20%	11%	1%	68%
sad	12%	69%	19%	0%	69%
happy	8%	10%	70%	12%	70%
angry	2%	5%	18%	75%	75%
total	70.5%				

표 1은 남성 과 여성 실험 데이터에 대한 인식률을 나타내는 표이다. 'normal' 감정의 인식률은 68%, 'sad' 감정의 인식률은 69%, 'happy' 감정의 인식률은 70%, angry 감정의 인식률은 75%로 'normal', 'sad', 'happy', 'angry' 감정에 대한 최종 인식률은 70.5%로 나타났다. 결과에서 'normal','sad' 감정과 'angry' 감정 간의 분리는 명확하였으나 'normal'과 'sad', 'happy'와 'sad', 'happy'와 'angry'간의 분리에서는 약간의 오차가 존재하였다.

표 2는 본 논문에서 제안한 톤, 템포 특징벡터를 이용한 방법과 기존의 피치 특징벡터를 이용한 방법에 대한 감정 인식률을 나타낸다. 참고문헌[11]의 피치 파라미터를 사용한 감정인식 방법은 음성신호에서 기본주파수, 피치 성분을 추출하여 피치에서 56개의 파라미터를 추출하였고 그 중 15개의 파라미터 조합을 이용하여 감정은 인식하였다. 본 논문에서 제안한 방법이 'normal', 'happy', 'angry' 인식 성능이 개선된 것을 알 수 있다.

피치, 즉, 톤의 파라미터만 이용한 방법보다 템포 특징까지 사용하였을 때 감정 인식하는데 용이하다고 할 수 있다.

표 2 톤과 템포를 사용한 방법과 피치를 사용한 방법과의 비교
Table 2 Comparison of method using tone and tempo and method using pitch

	Tone + Tempo	Pitch[11]
normal	68%	65.8%
sad	69%	71.4%
happy	70%	52%
angry	75%	64.9%
total	70.5%	63.5%

5. 결 론

본 논문은 IoT를 위한 음성신호의 톤, 템포 특징기반의 감정 인식에 대한 연구이다. 이를 위하여 실험데이터를 사람의 normal, sad, happy, angry 4가지 감정에 대한 음성데이터로 분류하고, 정확한 데이터를 위해 방송매체를 통해서 인물의 표정과 앞뒤 상황 등을 고려하여 판단하여 취득하였고, 데이터는 총 200개로 남성, 여성 각 100개씩 취득하였으며, 감정 당 25개씩 취득하였다. 실험 데이터는 비 음성구간을 없애는 전처리 과정을 거치고 톤, 템포를 추출하였으며 신경회로망 분류기를 이용하여 인식 시스템을 구현하였다.

그 결과, 남성과 여성의 음성데이터 실험에서는 'normal' 감정의 인식률은 68%, 'sad' 감정의 인식률은 69%, 'happy' 감정의 인식률은 70%, angry 감정의 인식률은 75%, 평균 인식률은 70.5%가 나왔다. 톤과 템포정보를 사용한 방법이 톤 정보만 사용한 방법보다 평균 6.5%의 인식률이 개선되었다.

향후, 본 논문에서 제안한 톤, 템포 특징외의 적용할 수 있는 새로운 특징벡터를 찾고 인식성능을 개선하는 연구가 필요하다고 판단된다.

References

[1] Chung-Hsien Wu, Wen-Li Wei, Jen-Chun Lin, Wei-Yu Lee "Speaking Effect Removal on Emotion Recognition From Facial Expressions Based on Eigenface Conversion", Multimedia, IEEE Transactions on, Vol. 15, pp. 1732-1744, July 2013.

[2] Sung-Woo Byun, So-min Lee, Seok-Pil Lee, "A Selection of Optimal EEG Channel for Emotion Analysis According to Music Listening using Stochastic Variables", KIEE, Vol. 62, No. 11, pp. 1598-1603, November 2013.

[3] So-min Lee, Sung-Woo Byun, Seok-Pil Lee, "Comparison of EEG Feature Vector for Emotion Classification according to Music Listening", KIEE, Vol. 63, No. 5, pp 696 - 702, May 2014.

[4] Jung-In Lee, Hong-Goo Kang, "On the Importance of Tonal Features for Speech Emotion Recognition", JBE, Vol. 18, No. 5, pp. 713-721, September 2013.

[5] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information", ICMI '04 Proceedings of the 6th international conference on Multimodal interfaces, pp. 205-211, October 2004.

[6] Daniel Neiberg, Kjell Elenius, Kornel Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs", Proc. Int',l Conf. Spoken Language Processing (ICSLP ',06), pp. 809-812, 2006.

[7] Xin Xu, Ya Li, Xiaoying Xu, Zhengqi Wen, Hao Che, Shanfeng Liu, Jianhua Tao, "Survey on discriminative feature selection for speech emotion recognition", Chinese Spoken Language Processing (ISCSLP), 2014 9th IEEE International Symposium on, pp. 345-349, 2014.

[8] Vijayan, A.E, Sen, D, Sudheer, A.P, Hao Che, Shanfeng Liu, Jianhua Tao, "EEG-Based Emotion Recognition Using Statistical Measures and Auto-Regressive Modeling", Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on, pp. 587-591, 2015.

[9] Sung-Woo Byun, Seok-Pil Lee, "A study on pitch detection for RUI emotion classification based on voice", 2015 Conference on The Korean Society Of Broad Engineers, pp. 421-424, July 2015.

[10] Kostov, V, Fukuda, S, "Emotion in user interface, voice interaction system", Systems, Man, and Cybernetics, 2000 IEEE International Conference on, pp. 798-803, October 2000.

[11] Guehyun Lee, Weon-Goo Kim, "Emotion Recognition using Pitch Parameters of Speech", KIISS, Vol. 25, No. 3, pp. 272-278, June 2015.[1] A. Ghosh, S. Devadas, K. Keutzer and J. White, "Estimation of Average Switching Activity in Combinational and Sequential Circuits," ACM/IEE Design Automation Conf., pp. 253-25.

저 자 소 개

변 성 우 (Sung-Woo Byun)

2014년 상명대학교 디지털미디어학과 이학사. 2014년～현재 상명대학교 컴퓨터과학과 석・박사 통합과정. 〈주관심분야〉 멀티미디어처리, 인공지능, 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 디지털미디어연구센터 센터장. 2010년～2011년 미국 Georgia Tech. 방문연구원. 2012년～현재 상명대학교 미디어소프트웨어학과 교수
〈주관심분야〉 멀티미디어 검색, 방송통신시스템, 인공지능