모의 지능로봇에서 음성신호에 의한 감정인식

장광동, 권오욱
충북대학교 제어계측공학과

Speech Emotion Recognition by Speech Signals on a Simulated Intelligent Robot

Kwang-Dong Jang, Oh-Wook Kwon
Department of Control and Instrumentation Engineering, Chungbuk National University
{kdjang,owkwon}@chungbuk.ac.kr

Abstract

We propose a speech emotion recognition method for natural human-robot interface. In the proposed method, emotion is classified into 6 classes: Angry, bored, happy, neutral, sad and surprised. Features for an input utterance are extracted from statistics of phonetic and prosodic information. Phonetic information includes log energy, shimmer, formant frequencies, and Teager energy; prosodic information includes pitch, jitter, duration, and rate of speech. Finally a pattern classifier based on Gaussian support vector machines decides the emotion class of the utterance. We record speech commands and dialogs uttered at 2m away from microphones in 5 different directions. Experimental results show that the proposed method yields 59% classification accuracy while human classifiers give about 50% accuracy, which confirms that the proposed method achieves performance comparable to a human.

Keywords: emotion recognition, support vector machine, speech interface

Ⅰ. 서론

음성은 사람들 사이에 의사소통을 하는데 있어 단어만을 통해서 의미뿐만 아니라, 감정도 전달한다. 음성에 내포된 감정은 단어를 강조하거나 화자의 심리상태를 나타내어 의사소통을 더 자연스럽게 한다. 정서적 휴먼컴퓨터 인터페이스(affective human computer interface) 는 최근 들어 휴머노이드형 로봇의 관심에 힘입어 많은 관심의 대상이 되고 있다. 사람의 감정을 인식하는데 있어 영상을 이용한 얼굴의 감정표현 인식과 음성을 이용한 감정인식 이용한 연구가 많이 되고 있다. 영상을 이용한 경우는 사람의 얼굴 표정에서 주요 특징인 입술, 눈, 코의 위치를 찾고 모양과 감정간의 기하학적인 관계를 파악하여 감정을 인식하는 방법이 시도되었다.

음성에 있어서 사전적인 범주와 운율적인 범주로 나누어 볼 수 있다. 사전적인 범주는 사람들이 서로 이해할 수 있는 단어들이며, 각 언어권에서 사용하는 각각의 발화들로 구성되어 있고, 운율적인 범주는 음성에 내포되어 있는 운율, 즉 음악적인 요소들로 구성되어 있는 것을 말한다. 운율적인 요소들은 음성에서 청자는 화자의 감정을 예측할 수 있는 요소이다. 그러나 감정을 표현하는데 있어 일반적으로 감탄사를 말하는 경우가 있고, 일상적인 생활에서 사용되는 단어들에 감정이 표현된 경우가 있다. 이에 단어의 의미로부터 감정을 인식하는 방법과 단어의 의미와 상관없이 운율적인 정보만을 이용하는 방법, 두 가지를 모두 사용하는 방법 등에 대한 많은 연구가 있었다[3].

사람의 감정을 분류하는데 있어 여러 가지가 있지만 대개 화남, 기쁨, 감정이 없는 상태, 슬픔, 놀람, 지루함, 혐오등과 같이 분류하였다. 그러나 감정이 하나의 감정으로만 표현되는 경우만 있는 것이 아니라, 복합적으로 한 개 이상의 감정이 동시에 나타나기도 한다. 가령, 기쁨과 놀람이 같이 나타날 수 있는 경우가 있다[1]. 감정을 인식하는데 있어 감정이 없는 상태와 감정이 있는 상태로 분류하여 감정 인식 접근방법도 있다.

감정을 인식하는데 있어 사용하는 특징들은 에너지, 포만트, 템포, 지속시간, 주파수변이(jitter), 진폭변이(shimmer), mel frequency cepstral coefficient (MFCC), linear predictive coding (LPC)계수, Teager 에너지 등이 있다. 여러 특징들 중에 감정을 인식하는데 가장 큰 기여하는 특징은 피치와 에너지이다[2][4]. 추출된 특징들을 가지고 hidden Markov model (HMM)[5], support vector machine (SVM), neural network 등을 사용하여 감정을 분류한다.

본 연구의 목적은 음성에 내포되어 있는 음향정보와 운율 정보들로부터 유효한 특징들을 추출하여 지능 로

봇이 이러한 정보를 이용하여 상황에 맞는 음성 정보와 화자의 상태를 인식함에 있다.

Ⅱ. 감정인식

음성을 이용한 감정인식은 감정을 6가지 - 기쁨(happy), 슬픔(sad), 놀람(surprised), 지루함(bored), 화남(angry), 감정이 없는 상태(neutral)로 분류하여 음성으로 부터 피치(pitch), 에너지, 주파수변이(jitter), 진폭변이(shimmer), Teager에너지, rate of speech (ROS) 그리고 포만트(formant)는 평균(mean), 표준편차(standard deviation), 최대(maximum), 최소(minimum), 퍼센타일(percentile), range(MAX-MIN), linear regression coefficient, maximum gradient 특징을 추출하여 SVM 패턴분류기를 사용하여 인식한다.

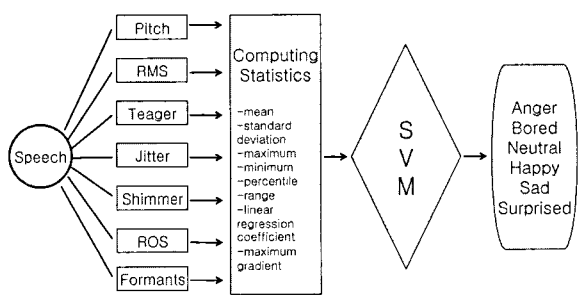

그림 1 음성 감정인식기

1. 특징 추출

음성신호에서 노이즈를 제거하기 위해서 Wiener 필터링한 후 끝점을 검출하여 음성부분만 추출하였다. 일반적으로 MFCC를 추출하여 음성인식을 하는 경우에는 해밍윈도우(hamming window)를 사용하지만 이 논문에서는 추출된 음성을 해닝윈도우(hanning window)를 사용하여 10ms단위로 오버랩하면서 매 프레임당 960 샘플로 하였다.

피치를 추출하기 음성을 저대역 필터링하여 10ms단위로 구간 이동하면서 average magnitude difference (AMDF)을 사용하여 피치 후보들을 구한 후, 피치 후보들에서 최소값을 피치로 결정한 후 smoothing하였다. AMDF를 사용하여 피치를 추출하는 것은 피치를 정확히 계산하여 주는 방법은 아니지만 노이즈에 강한 특성을 보이고 계산이 빠르고 간단하다[8].

$$AMDF_n(j)=\frac{1}{N}\sum_{i=1}^{N}|x_n(i)-x_n(i+j)|,\ 1\le j\le MAXLAG \quad (1)$$

여기에서 N은 음성 샘플 갯수, x_n은 음성 샘플 신호, MAXLAG은 AMDF의 최대값, 즉 피치 주기의 최대값이다

에너지는 일반적으로 많이 사용하는 에너지와 Teager에너지를 추출하여 사용하였다. Teager에너지는 기존의 추출 알고리듬[7]을 사용하여 추출하였다. Teager에너지는 음성신호가 복합 정현신호로 구성되어 있으므로 각 주파수 대역으로 분류한 후 계산한다. 단일대역 주파수대로 분류하기 위해서 필터뱅크를 사용하여 주파수 대역 별로 Teager에너지를 구하였다.

$$E_{Teager}(n,i)=x_n^2(i)-x_{n+1}(i)x_{n-1}(i), i=1...FB \quad (2)$$

여기에서 $x_n(i)$ 프레임 n번째 필터뱅크 계수가 i인 신호, FB는 필터뱅크의 최대 계수이다.

주파수변이(주파수변동률)와 진폭변이(진폭변동률)는 음성의 음질을 분석할 때 사용하는 특징으로서 주파수변이는 피치 주파수의 변동되는 정도를 진폭변이는 피치와 피치사이의 진폭이 변동되는 것을 나타낸다[9].

$$Jitt=\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_{0_i}-T_{0_{i+1}}|}{\frac{1}{N}\sum_{i=1}^{N}|T_{0_i}|} \quad (3)$$

$$Shimm=\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i-A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}|A_i|} \quad (4)$$

여기서 N은 AMDF를 사용하여 추출된 피치의 갯수, T_{0_i}는 i번째의 피치 주기이고 A_i는 i번째 프레임의 진폭이다.

ROS는 음성을 voice/unvoice/silence (V/U/S)구간을 나누어 단위 구간의 모음비율을 나타낸 것이다. ROS는 fixed와 variable이 있는데 음성 감정 인식기에서는 fixed ROS를 사용하였다.

$$ROS=\frac{N}{\sum d_i} \quad (5)$$

여기에서 N은 유성음 구간 개수이고, d_i는 i번째 유성음 구간의 지속시간이다.

2. 감정 분류

위에서 추출된 특징들로부터 평균(mean), 퍼센타일(percentile), 표준편차(standard deviation), 최대(MAX), 최소(MIN), range(MAX-MIN), linear regression coefficient, maximum gradient등의 통계적인 값을 계산하여 인식 훈련과 테스트에 사용하였다.

SVM은 입력이 다차원인 경우 사용하여 최적의 분류를 할 수 있는 방법으로 Gaussian kernel SVM 을 사용하여 감정을 인식하였다. Mixture는 10으로 사용하였으며 감정을 훈련 및 테스트하는데 있어 6개의 감정 중 하나 이상의 감정이 동시에 인식되는 것은 배제하고 하나의 문장 또는 단어입력이 되면 하나의 감정만이 있는 것으로 가정하여 인식하였다.

Ⅲ. 실험 결과

1. 음성 데이타베이스

감정 음성 데이타베이스는 지능형 로봇의 감정인식 인터페이스 개발을 위해 성우가 아닌 일반인으로 20~30대 30명(남녀 각 15명)의 화자로부터 6가지의 감정-기쁨, 슬픔, 놀람, 지루함, 화남, 감정이 없는 상태를 녹음하였다.

마이크와 화자사이의 거리를 2m, 화자와 마이크 사이의 각도를 전방향 좌우 각도로 하여 스테레오로 녹음하였다. 무음구간은 300ms로 사용자 등록, 인사, 생활정보, 명령, 감정 등을 나타낼 수 있는 5개의 항목으로 구성되었으며, '사용자 등록'을 제외한 4개의 항목(50단어 및 문장)으로 6가지 감정을 발화하였다. 한 화자 당 발화량은 302개이며 총 9060개의 발화로 구성되어 있다.

표 1 감정 음성 데이타베이스 사용 단어

항목	단어 및 문장
사용자등록	사용자등록, 내 이름은 ㅇㅇㅇ입니다.
인사	안녕, 잘 지내어?, 보고 싶었어, 오랜만이야! 뭐하고 놀았어? 잘 있어, 갔다 올게, 나중에 봐 빨리 와, 어디 가?
명령	정지, 서, 위로 가, 아래로 가, 뒤로 가 왼쪽으로 가, 오른쪽으로 가, 앞으로 가, 돌아 하지마, 그만해, 안돼, 일어서, 앉아, 맘대로 해 일어서, 앉아, 맘대로 해, 가지고 와 가지고 가 이리 와, 저쪽으로 가, 이쪽으로 가
감정	이쁜 짓 해봐, 윙크해 봐, 춤춰 봐, 착하지, 괜찮니?, 좋아, 잘 했어, 못 했어, 사랑해, 예쁘다 혼날래? 어디 아프니?, 한번 더, 조용히 해
생활정보	오늘 날씨를 알려줘, 비오니? 시원하니?, 온도가 몇 도야?
날짜/시간	오늘은 몇월 몇일이야? 지금 시간이 몇시지?

2. 감정인식 실험 결과

이 논문은 지능로봇과 인간과의 음성인터페이스하는데 있어 언어의 정보만을 전달하는 것이 아니라 음성에 포함되어 있는 감정을 인식하여 현재의 발화한 사람의 상태에 따라 대응하는 것으로 전제하고 있다.

"사용자등록"이라는 감정이 없는 상태로 사용자를 등록한 후, 사용하는 것을 시나리오로 한다. 사람마다 발화시 운율 요소들이 변화하므로 기준이 될 수 있는 요소-감정이 없는 상태의 단어를 입력한다.

감정을 인식하는 실험은 사람의 판단에 의한 것과 SVM을 이용하여 감정을 분류하는 방법을 비교하였다. 실험 방법은 사람의 경우 6가지의 감정이 포함되어 단어 또는 문장을 발화한 단어를 들려주고 판단하였다.

표2와 3에서 데이터베이스 중에서 무작위로 추출하여 사람에 의해서 판단한 결과는 약 50%의 정확도를 보인 반면, SVM 감정 인식결과는 교차 검증으로 약 59%의 결과를 보여 감정을 분류하는데 있어 효과적임을 알 수 있었다.

사람의 감정분류 정확도보다 SVM에 의한 감정 정확보가 높게 나온 이유는 SVM의 경우 사전에 학습데이터를 가지고 있다. 감정을 분류하는 기준에서 SVM에 의한 감정인식기는 학습데이터를 가지고 일정한 기준을 가지고 감정을 분류하는 반면, 사람의 경우는 현재 몸 상태 또는 주변 상황에 따라 같은 감정 상태의 단어 또는 문장을 듣더라도 항상 같은 감정으로 분류할 수 없다. 그러나 사람의 의한 감정 분류는 화자가 음성에 표출한 다양한 감정을 복합적으로 한 개 이상의 감정이 있더라도 한 개 이상의 감정을 인식할 수 있다.

표 2 감정인식 confusion matrix
ang(angry), bor(bored), hap(happy), neu(neutral), sad(sad), sur(surprised), (average accuracy: 59%)

	ang	bor	hap	neu	sad	sur
ang	0.58	0.00	0.10	0.12	0.01	0.17
bor	0.00	0.65	0.02	0.05	0.28	0.00
hap	0.10	0.02	0.55	0.17	0.06	0.11
neu	0.09	0.03	0.13	0.63	0.09	0.03
sad	0.00	0.34	0.06	0.11	0.48	0.01
sur	0.21	0.00	0.10	0.03	0.00	0.66

표 3 사람 판단에 의한 감정 분류의 confusion matrix
(average accuracy: 50%)

	ang	bor	hap	neu	sad	sur
ang	0.31	0.03	0.07	0.42	0.05	0.08
bor	0.01	0.62	0.02	0.09	0.32	0.00
hap	0.16	0.03	0.59	0.09	0.01	0.14
neu	0.41	0.02	0.10	0.27	0.05	0.08
sad	0.03	0.29	0.04	0.09	0.53	0.00
sur	0.09	0.01	0.19	0.05	0.03	0.70

표2에서 사람의 의한 감정 분류는 화남과 감정이 없는 상태를 분류하는데 있어 화가 난 상태의 단어를 감정이 없는 상태로 또는 반대로 많이 발생하였다. 이러한 이유는 화남과 감정이 없는 상태의 발화의 경우, 에너지와 피치가 높게 나타는 경향을 보이기 때문이다. 그리고 또한, 지루함과 슬픔의 경우 분류하는데 있어 오판이 생기는데 이 경우 피치와 에너지가 다른 감정보다 피치와 에너지가 낮게 나타나는 경향이 있다.

표4~7은 단일 특징들 피치, 에너지, Teager에너지, 포만트를 사용하여 감정을 인식한 결과이며 인식률이 가장 높은 것은 피치였다. 단일 특징을 사용하여 분류한 결과, 피치는 화남, 지루함과 기쁨, 에너지의 경우는 화남과 지루함, Teager에너지의 경우는 지루함과 놀람을 그리고 포만트를 사용한 경우는 놀람을 분류하는데 있어 유효한 특징들이었다.

표 4 피치 특징을 이용한 감정인식의 confusion matrix
(average accuracy: 45%)

	ang	bor	hap	neu	sad	sur
ang	0.54	0.00	0.18	0.12	0.00	0.16
bor	0.00	0.56	0.12	0.14	0.18	0.00
hap	0.02	0.02	0.52	0.20	0.08	0.16
neu	0.19	0.00	0.27	0.44	0.02	0.08
sad	0.00	0.76	0.00	0.00	0.24	0.00
sur	0.22	0.00	0.26	0.10	0.02	0.40

표 5 에너지 특징을 이용한 감정인식의 confusion matrix
(average accuracy: 40%)

	ang	bor	hap	neu	sad	sur
ang	0.51	0.01	0.10	0.15	0.02	0.20
bor	0.00	0.51	0.04	0.10	0.34	0.00
hap	0.13	0.05	0.26	0.34	0.10	0.12
neu	0.11	0.06	0.16	0.45	0.14	0.08
sad	0.00	0.39	0.06	0.16	0.38	0.02
sur	0.30	0.01	0.15	0.22	0.02	0.31

표 6 Teager에너지 특징을 이용한 감정인식의 confusion matrix
(average accuracy: 39%)

	ang	bor	hap	neu	sad	sur
ang	0.41	0.05	0.15	0.15	0.01	0.22
bor	0.01	0.70	0.04	0.11	0.14	0.01
hap	0.12	0.13	0.27	0.27	0.04	0.17
neu	0.10	0.17	0.16	0.38	0.08	0.11
sad	0.01	0.63	0.06	0.15	0.12	0.03
sur	0.15	0.04	0.16	0.13	0.01	0.51

표 7 F1,F2,F3 특징을 이용한 감정인식의 confusion matrix
(average accuracy: 29%)

	ang	bor	hap	neu	sad	sur
ang	0.22	0.13	0.15	0.21	0.06	0.23
bor	0.08	0.29	0.11	0.23	0.19	0.10
hap	0.16	0.15	0.18	0.25	0.09	0.18
neu	0.12	0.17	0.15	0.35	0.11	0.09
sad	0.08	0.27	0.09	0.23	0.20	0.12
sur	0.14	0.06	0.10	0.07	0.06	0.56

Ⅲ. 결론

이 논문은 지능로봇과 음성 인터페이스시 음성에 포함되어 있는 감정을 인식하는 시뮬레이션 결과를 기술하였다. 입력된 음성의 운율적인 정보만을 사용하여 감정을 분류하였다. SVM에 의한 인식결과는 사람의 판단에 의한 인식률보다 높게 나타므로 감정을 인식하는데 있어 효과적이었음을 알 수 있었다. 피치, 에너지, Teager에너지가 약 40%정도의 정확도로 포만트는 약 29%의 정확도를 보였으며 반면, 주파수변이, 진폭변이는 인식률 향상에 크게 기여도가 크지 않았다. 인식률을 향상하기 위해서 특징들을 조합하여 유효한 감정 인식 특징들을 추출하거나 피치를 추출하는데 있어 음성의 발화 방법 예측 또는 ROS에 따라 가변적인 피치 추출 알고리즘을 적용하여 검증하는 것이 추가적으로 필요하다.

감사의 글

"이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음."

참고문헌

[1] Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech," IEEE Int'l. Conference on Multimedia Computing and Systems, pp. 840-844, 1999.

[2] O.-W. Kwon, K. Chan, J. Hao, T.-W. Lee, "Emotion recognition by speech singnals," Proc. Eurospeech, Geneva, pp. 125-128, 2003.

[3] B. Schuller, G. Rigoll, M. Lang, "Speech emotion recognition combing acoustic features and linguistic information in hybrid support vector machine-belief network architecture," Proc. ICASSP, pp. 577-580, 2004.

[4] B. Schuller, G. Rigoll, M. Lang, "Hidden Markov Mdoel-based speech emotion recognition", Proc. ICASSP, pp. 401-404, 2003.

[6] T.-L. Pao, Y.-T. Chen, "Mandarin emotion recognition in speech," IEEE workshop on Automatic Speech Recognition and Understanding, pp. 227-230, 1999.

[7] J.F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," Acoustics, Speech, and Signal Processing, ICASSP-90, pp. 381-384, 1990.

[8] G.S. Ying, L.H. Jamieson, C.D. Michell, "A probabilistic approach to AMDF pitch detection," Spoken Language, Proc. ICSLP-96, pp. 1201-1204, 1996.

[9] R.E. Slyh, W.T. Nelson, E.G. Hansen, "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database," Acoustics, Speech, and Signal Processing, Proc. ICASSP-99, pp. 2091-2094, 1999.