감성적 인간 로봇 상호작용을 위한 음성감정 인식

장광동, 권오욱
충북대학교 제어계측공학과
{kdjang, owkwon}@chungbuk.ac.kr

Speech emotion recognition for affective human robot interaction

Kwang-Dong Jang and Oh-Wook Kwon
Department of Control Instrumentation Engineering, Chungbuk National University

요 약

감정을 포함하고 있는 음성은 청자로 하여금 화자의 심리상태를 파악할 수 있게 하는 요소 중에 하나이다. 음성신호에 포함되어 있는 감정을 인식하여 사람과 로봇과의 원활한 감성적 상호작용을 위하여 특징을 추출하고 감정을 분류한 방법을 제시한다. 음성신호로부터 음향정보 및 운율정보인 기본 특징들을 추출하고 이로부터 계산된 통계치를 갖는 특징벡터를 입력으로 support vector machine (SVM) 기반의 패턴분류기를 사용하여 6가지의 감정- 화남(angry), 지루함(bored), 기쁨(happy), 중립(neutral), 슬픔(sad) 그리고 놀람(surprised)으로 분류한다. SVM에 의한 인식실험을 한 경우 51.4%의 인식률을 보였고 사람의 판단에 의한 경우는 60.4%의 인식률을 보였다. 또한 화자가 판단한 감정 데이터베이스의 감정들을 다수의 청자가 판단한 감정 상태로 변경한 입력을 SVM에 의해서 감정을 분류한 결과가 51.2% 정확도로 감정인식하기 위해 사용한 기본 특징들이 유효함을 알 수 있다.

Keyword : 감정인식, SVM, 지능로봇, 음성인터페이스, 감성인터페이스

1. 서 론

음성은 사람들 사이에 의사소통을 하는데 있어 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.

운율요소들은 청자로 하여금 화자의 감정을 예측할 수 있도록 하는 요소이다. 그러나 감정을 표현하는데 있어 일반적으로 감탄사를 말하는 경우가 있고, 일상적인 생활에서 사용되는 단어들에 감정이 표현된 경우가 있다. 이에 단어의 의미로부터 감정을 인식하는 방법, 단어의 의미와 상관없이 운율적인 정보만을 이용하는 방법, 그리고 두 가지를 모두 사용하는 방법 등에 대한 많은 연구가 있었다[3].

사람의 감정을 분류하는데 있어 여러 가지 분류기준이 있지만 대개 화남, 기쁨, 중립(감정이 없는 상태), 슬픔, 놀람, 지루함, 혐오등과 같이 분류하고 있다. 그러나, 감정은 한순간에 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 감정인식 접근방법도 있다.

감정인식에 사용되는 특징들은 에너지, 포만트, 템포, 지속시간, 주파수변이(jitter), 진폭변이(shimmer), mel frequency cepstral coefficient (MFCC), linear predictive coding (LPC)계수, Teager에너지 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 한국인이 발화한 한국어 음성에 내포되어 있는 음향정보와 운율 정보들로부터 유효한 특징들을 추출하여 감정을 분류함에 있다. 또한, SVM을 사용한 감정인식 결과와 사람에 의해 판단된 결과를 비교 분석하였다.

2. 감정인식 방법

감정인식기는 입력 음성을 화남, 지루함, 기쁨, 중립, 슬픔 그리고 놀람의 6가지 감정상태를 분류한다. 감정을 인식하기 위해서는 기본 특징을 음성으로부터 추출하여 기본 특징들의 계산된 통계값을 입력으로 하는 SVM을 사용하여 음성신호에 표현되어 있는 감정을 인식한다.

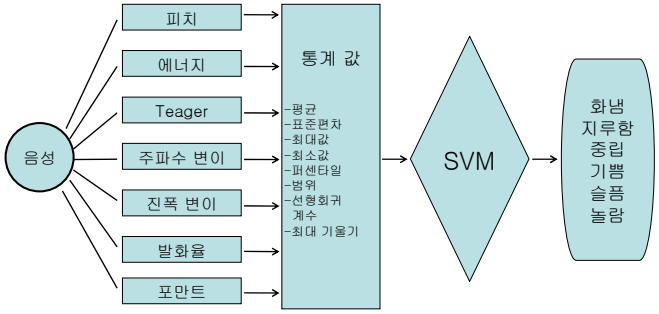

그림 1. 감정인식기

2.1 특징 추출

입력된 음성신호는 16kHz로 샘플링 되어 16비트 PCM(pulse code modulation) 방식으로 특징을 추출하기 위해서 사용된다. 샘플링 된 음성신호에 포함되어 있는 노이즈를 제거하기 위해서 위너 필터링(Wiener filtering)한 후, 음성신호에서 음성의 시작점과 끝점을 추출한다. 피치를 추출하는 경우에는 해닝 윈도우(Hanning window)를 사용하고, 이외에는 윈도우 크기가 25ms인 해밍 윈도우(Hamming window)를 10ms단위로 구간 이동하면서 기본 특징인 에너지, 주파수변이, 진폭변이와 발화율 등을 추출한다.

2.1.1 피치 추출

피치를 추출하는 방법은 크기가 60ms인 해닝 윈도우를 사용하여 한 프레임당 2~3개의 피치를 포함하게 하여 차단 주파수가 800Hz인 저대역 필터(low pass filter)를 통과한 후 AMDF(average magnitude difference function)을 사용하여 추출한 피치후보들 중에서 최소로 하는 값을 피치로 결정하는 방법을 사용하였다.

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

여기에서 N은 샘플의 개수, $x_n(i)$는 n번째 프레임의 i번째 샘플값을 나타내고, $MAXLAG$는 추출가능한 피치 주기의 최대값이다.

추출된 피치 후보는 프레임 간에 피치가 급격히 변하는 것을 방지하기 위해 스무딩(smoothing)한다. 그리고 짧은 구간의 무성음 프레임 구간(1~2프레임)이 유성음사이에 위치하여 있으면 이전과 이후 프레임의 평균 피치값을 갖는 유성음으로 처리한다.

2.1.2 에너지

에너지는 일반적으로 많이 사용하는 로그에너지와 Teager에너지를 사용한다. 로그에너지는 프레임에서 샘플 신호의 절대값들을 합하여 로그를 취해서 구하고, Teager에너지는 Kaiser[6]가 제안한 방법으로서 복합 정현신호에 필터뱅크를 적용하여 단일 주파수로 나눈 후 (2)에서와 같은 방법으로 구한다.

$$TE_n(i)=f_n^2(i)-f_{n+1}(i)f_{n-1}(i), i=1...FB \quad (2)$$

여기에서 $f_n(i)$는 n번째 프레임의 i번째 필터뱅크 계수이고 FB는 주파수 대역의 개수이다. Teager에너지는 신호가 잡음에 강한 특징을 보이고 음성신호를 동적으로 향상시킨다.

2.1.3 주파수 변이와 진폭 변이

주파수 변이(주파수 변동률)와 진폭변이(진폭 변동률)는 음질을 분석할 때 사용하는 특징[11]으로서 주파수 변이는 피치 주파수의 변동되는 정도를, 진폭 변이는 피치간에 진폭이 변동되는 정도를 나타낸다[8].

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

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

여기에서 N은 발화된 프레임의 개수, T_{0_i}는 i번째 프레임의 피치이고 A_i는 i번째 프레임의 진폭이다.

2.1.4 발화율

유성음/무성음/무음(voice/unvoice/silence)구간으로 음성을 나누어 단위 구간에서의 유성음 비율이 발화율이다. 발화율은 고정형과 가변형이 있는데 여기에서는 고정 발화율를 사용하였다.

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

여기에서 N은 유성음 구간의 개수이고 d_i는 i번째 유성음 구간의 지속시간이다.

2.1.5 포만트

포만트는 음성을 켑스트럼(cepstrum)으로 변환하여 피크를 구해서 추출하는 방법과 LPC(linear predictive coding)으로부터 구하는 방법 등이 있는데 여기에서는 LPC를 이용하여 구한다. 성도가 선형 시스템이라고 전제하며 계산된 LPC계수를 All-pole다항식의 분모로 이용하여 인수분해하면, pole에 해당하는 포만트를 알 수 있다[12]. 여기서는 14차 선형 예측을 이용하여 포만트(F1,F2,F3)를 계산한다.

2.2 SVM을 이용한 감정분류

추출된 기본 특징들에서 통계 값인 평균, 퍼센타일, 표준편차, 최대, 최소, 범위, 선형회귀계수, 최대 기울기등로 이루어진 73차의 벡터를 입력으로 하는 SVM감정분류기를 사용하였다. SVM은 두 클래스 사이의 결정오류(decision error)의 개수를 최소화하게 하는 hyperplane을 찾을 뿐만 아니라 신경망과 비교하여 매우 간단한 구조이고 일반화의 장점을 가지고 있기 때문에 많은 응용분야에 사용된다[9][10]. 커널 함수(kernel function)로는 가우시안(Gaussian), 다항식(polynomial) 등이 있는데, 이 논문에서 사용한 가우시안 커널 함수은 이전 실험에 의하면 다른 커널 함수 중에서 가장 높은 인식률을 보였다[2].

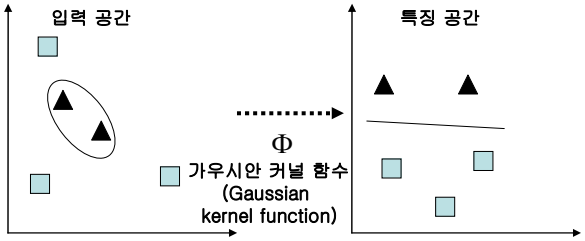

그림 2. SVM의 개념

3. 실험 결과

3.1 음성 데이터베이스

음성 데이터베이스는 정서적 인터페이스를 위해서 감정이 있는 음성으로 구성하였다. 데이터베이스의 화자는 나이는 20~30대이며 남 15명과 여 15명으로 총 30명의 성인으로 구성되어 있고 화남, 지루함, 기쁨, 중립, 슬픔 그리고 놀람 6가지의 감정상태의 명령어 또는 대화로 구성되어 있다. 음성 데이터는 6개의 항목-사용자 등록, 인사, 명령, 감정, 생활 정보, 날짜/시간으로 구성하였으며, '사용자 등록'을 제외한 5개의 항목(50 단어 및 문장)으로 6가지 감정을 발화하였다. 한 화자당 발화수는 302개이며 총 9,060개의 발화로 구성되어 있다.

3.2 화자가 정의한 감정을 SVM에 의한 감정분류

음성 데이터베이스를 구성하는 감정음성들은 제시한 5개의 항목의 단어들을 화자가 정의한 감정으로 표현한 음성들로 이루어져 있다. 화자가 감정상태에 따라서 적절한 발화를 했는지 알아보기 위해서 30명의 화자 중 29명은 훈련, 1명은 테스트로 하는 교차검증(cross validation)을 적용하여 SVM을 사용하여 감정을 분류하였다.

표 1. SVM에 의한 혼돈행렬(confusion matrix) (%)

	angry	bored	happy	neutral	sad	surprised
angry	58.6	0.3	9.0	12.2	1.1	18.8
bored	0.1	64.1	2.4	5.5	27.9	0.1
happy	11.8	2.2	54.0	16.7	5.5	9.7
neutral	8.5	3.1	13.4	64.0	8.6	23.8
sad	0.3	35.6	5.8	12.5	45.2	0.7
surprised	19.5	0.0	11.5	2.6	0.5	66.0

표1은 화자가 발화한 음성을 SVM을 사용한 인식률 58.64%였다. 다른 관점에서 바라보면 화자들이 감정을 표현하는 58.64%의 정확도를 나타내는 것이기도 하다. SVM이 주로 오인하는 감정은 지루함과 슬픔이고 가장 인식이 잘되는 감정은 놀람이었다.

3.3 화자가 정의한 감정을 사람에 의한 감정분류

화자가 표현한 감정이라고 해도 청자가 화자의 감정상태를 파악하기는 어렵다. 따라서 화자가 표현한 감정을 사람에 의해 판단되는 감정상태의 결과를 알아 보기 위해서 남녀 12명인 청자들을 구성하여 판단하였다. 청자들에 의한 감정을 판단하기 위한 기준으로는 감정이 중립인 문장 '사용자 등록'과 '내 이름은 OOO입니다'를 들을 수 있도록 하였다.

재구성한 음성 데이터베이스는 화자 30명을 가지고 실험한 결과에서 얻은 인식률을 기준으로 인식률이 가장 높은 남녀 화자 각 2명, 중간 정도의 남녀 화자 각 2명, 가장 낮은 남녀 각 1명을 선택하였고, 비교적 발화길이가 긴 음성 25개를 선택하여 구성하였다. 재구성한 음성 데이터베이스는 화자당 150개의 음성으로 이루어져 '사용자 등록' 항목의 두 개의 중립 음성을 포함하여 총 1,520개로 구성되어 있다. 이 음성 데이터 베이스를 가지고 사람이 판단한 결과 화자가 표현한 결과를 기준으로 정확도 60.4%였으며 30명의 화자를 SVM에 의한 판단 결과(58.6%)와 비교하여 볼 때 사람이 약 2%의 차이를 보였지만 신뢰구간(confidence interval)이 ±3.82로서 의미있는 차이는 아닌 것으로 판단된다.

표 2. 사람 판단에 의한 혼돈행렬 (%)

	angry	bored	happy	neutral	sad	surprised
angry	68.7	1.9	4.1	15.6	0.7	20.2
bored	1.2	56.9	1.9	5.3	42.2	0.4
happy	4.5	2.8	62.7	4.5	2.6	9.9
neutral	18.4	6.2	26.4	70.9	11.0	7.9
sad	0.4	32.0	2.0	3.0	42.4	0.4
surprised	6.8	0.2	2.9	0.8	0.1	61.2

사람 판단에 의한 결과에서 남자청자와 여자청자는 같은 감정을 듣더라도 판단 기준에 차이가 있다는 것을 알 수 있었다. 여자 청자의 경우는

63.4%의 인식률을 보이는 반면 남자의 경우는 58.3%의 인식률을 보였다.

표 3. 성별 화자에 따른 성별 청자 감정 인식률 비교 (%)

	남자 청자	여자 청자
남자 화자	60	67
여자 화자	57	60

남자와 여자의 화자인 경우 같은 성별의 화자가 발화한 감정을 판단하는 결과는 같았으며, 남자와 비교하여 여자는 감정이 표현된 음성에서 화자가 의도한 감정을 더 잘 분류함을 알 수 있었다.

3.4 청자의 판단으로 재구성한 감정을 SVM에 의한 감정분류

감정이라는 것은 화자와 청자에 따라 감정을 판단하는데 있어 화자가 정확한 감정을 했는지는 알 수가 없으므로 청자의 입장에서 볼 때, 감정이라는 것은 발화한 관점에 볼 수도 있지만 듣는 입장 즉, 청자가 관점에서 감정을 판단할 수 있다. 앞의 재구성한 음성 데이터베이스의 감정을 사람이 판단하는 결과들의 다수 결정에 따라 음성 데이터베이스를 구성하는 감정을 재정의하였다. 화자의 판단결과에 따라 재정의한 감정 인식률은 51.2%였다.

표 4. 청자가 판단한 감정으로 재정의한 데이터베이스에서 SVM에 의한 혼돈행렬 (%)

	angry	bored	happy	neutral	sad	surprised
angry	57.5	2.8	6.8	19.2	0.0	13.8
bored	0.8	81.0	2.9	15.0	0.3	0.0
happy	13.7	7.0	38.3	33.5	0.4	7.2
neutral	9.6	18.7	12.3	54.4	0.0	5.1
sad	0.0	71.6	11.2	15.8	1.3	0.0
surprised	29.4	1.0	16.0	7.9	0.0	45.8

감정을 인식하는 구조에서 사람의 경우와 SVM에 의한 경우 차이점이 있는데 사람의 경우 이전에 들었던 음성에 포함되어 있는 감정관련 정보를 가지고 다음에 입력되는 감정이 판단할 수 있지만 SVM은 이전에 듣고 판단한 감정에 대한 정보를 가지고 있지 않기 때문에 오인하는 확률이 높아진다.

3.5 토의

기존 연구에서의 인식률(50~70%)[4][13][14]보다 낮은 결과이었으며, 기존의 연구의 결과와 본 논문의 결과를 비교해 본 결과 음성 데이터베이스를 만들 때 명령어와 인사 등과 같이 비교적 짧은 문장들을 사용한 것과 훈련 및 테스트하는데 있어 너무 적은 화자를 사용했다는 점이다. 이러한 점들을 보완 및 수정하면 인식률의 향상이 기대된다.

4. 결 론

이 논문은 지능로봇과 음성 인터페이스시 음성에 포함되어 있는 감정을 인식하는 시뮬레이션 결과를 기술하였다. 6가지(화남/지루함/기쁨/중립/슬픔/놀람)의 감정 상태를 표현하는 음성으로부터 음향 및 운율 정보를 추출하여 SVM을 사용하여 분류하였다. 화자가 정의한 감정 상태에 대하여 실험한 사용된 음성 데이터베이스에서 전체 화자의 감정 인식률과 비교해 보면 감정 표현이 비교적 쉬운 발화를 선별하여 사람이 테스트해본 결과 인식률이 60.4%였다. 청자가 판단한 감정들의 결과들을 다수결의 법칙에 따라 음성데이터베이스의 감정으로 다시 정한 후, SVM을 사용하여 테스트한 결과 51.2%였다. 기본 특징들과 SVM을 사용한 감정인식기는 효과적임을 알 수 있다.

감사의 글

“이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음.”

참고문헌

[1] T. Moriyama and S. Ozawa, “Emotion recognition and synthesis system on speech,” *IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, pp. 840-844, 1999.

[2] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech signals,” *Proc. Eurospeech*, Geneva, Switzerland, pp. 125-128, 2003.

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

[4] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” *Proc. ICASSP*, Hongkong, China, pp. 401-404, 2003.

[5] T.-L. Pao and Y.-T. Chen, “Mandarin emotion recognition in speech,” *IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO, pp. 227-230, 1999.

[6] J.F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” *Proc. ICASSP*, Albuquerque, NM, pp. 381-384, 1990.

[7] G.S. Ying, L.H. Jamieson, and C.D. Michell, “A probabilistic approach to AMDF pitch detection,” *Proc. ICSLP*, Philadelphia, PA, pp. 1201-1204, 1996.

[8] R.E. Slyh, W.T. Nelson, and E.G. Hansen, “Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database,” *Proc. ICASSP*, Phoenix, AZ, pp. 2091-2094, 1999.

[9] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machine*s, Cambridge University Press, 2000.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

[11] J. Clark and C. Yallop, *An Introduction to Phonetics and Phonology*, Blackwell, 1995.

[12] L.R. Rabiner and R.W. Shafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[13] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” *Proc. ICASSP*, Montreal, Canada, pp. 593-596, 2004.

[14] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ““Of all things the measure is man” Automatic classification of emotions and inter-labeler consistency,” *Proc. ICASSP*, Philadelphia, PA, pp. 317-320, 2005.