음성을 이용한 화자 및 문장독립 감정인식

강 면 구, 김 원 구
군산대학교 전자정보공학부

Speaker and Context Independent Emotion Recognition using Speech Signal

Myeong-Gu Kang, Weon-Goo Kim
School of Electronic and information Engineering, Kunsan National University
E-mail : {kkm9, wgkim}@kunsan..ac.kr

Abstract

In this paper, speaker and context independent emotion recognition using speech signal is studied. For this purpose, a corpus of emotional speech data recorded and classified according to the emotion using the subjective evaluation were used to make statical feature vectors such as average, standard deviation and maximum value of pitch and energy and to evaluate the performance of the conventional pattern matching algorithms. The vector quantization based emotion recognition system is proposed for speaker and context independent emotion recognition.

Experimental results showed that vector quantization based emotion recognizer using MFCC parameters showed better performance than that using the pitch and energy parameters.

I. 서론

인간의 감정을 인지하고, 그에 정서적인 반응을 하는 시스템의 개발은 보다 고차원적인 휴먼-컴퓨터 인터페이스 제품을 가능하게 한다. 인간의 감정 정보는 얼굴표정, 음성, 몸 동작, 심장 박동수, 체온, 혈압 등 다양한 방법으로 얻을 수 있고, 어플리케이션에 따라 감정 정보 취득 방법이 달라질 것은 자명하다. 음성을 이용한 시스템의 경우 센서가 신체부위에 직접 닫지 않거나, 전화와 같이 반드시 음성을 이용하여야 하는 시스템에 응용할 때 유리하다.

감성 인식 기술의 이미 일부 분야에서 사용화가 이루어지고 있다. 이러한 것으로 일본 Sony사가 개발하여 시판한 애완로봇 AIBO는 6가지 감정을 포함하는 감정 모델을 적용하여 주인과의 관계에 의해서 감정상태가 변화하고 반응하도록 만들어졌다. 또 IBM에서는 차세대 감정 인식 제품을 개발하고 있으며, 표정을 이용한 감정 인식시스템, 생체 신호를 이용한 감정 인식 마우스 등을 상용화하는 단계에 이르렀다[1].

현재까지 화자의 감성을 반영하는 요소로서 발음 속도, 피치 평균, 피치 변화 범위, 발음 세기, 음질, 피치의 변화, 발음법 등의 파라메터가 감성 인식 및 합성에 주로 사용되어오고 있다. 또한 이러한 파라메터를 바탕으로 감정 인식을 수행하기 위한 패턴 인식 방법으로는 MLB(Maximum Likelihood Bayes), KR(Kernel Regression), KNN(K-Nearest Neighbor) 분류기 등 기본적인 패턴 인식 기법이 사용되었고 음성 인식 방법과 결합된 감정 인식 방법도 사용되고 있다. 또한 음성과 표정 두 가지를 이용한 감성인식에 관한 연구도 수행되어 오고 있다[2-5]. 그러나 지금까지의 대부분의 연구는 화자 종속적이거나 문장 종속적인 환경을 대상으로 것이다. 또한 기존에 많이 사용하던 피치나 에너지 파라메터를 사용한 감정 인식 시스템은 화자 독립이나 문장 독립적인 음성을 대상으로 한 경우 인식 성능이 크게 저하되는 문제점이 있다.

본 연구에서는 감정 상태에 따라 분류된 한국어 음성 데이터 베이스를 이용하여 화자 및 문장 독립적인 감정 인식 시스템에 관하여 연구하였다. 이를 위하여 여러 가지 감정 상태에 따라 분류된 한국어 음성 데이터 베이스를 구축하고, 구축한 데이터 베이스를 이용하여 특징을 추출한 후, 피치와 에너지의 평균과 표준편차, 최대 값 등 통계적인 정보를 감성 인식의 특징으로 이용하는 기본적인 패턴 인식 시스템을 구축하고 음소의 특성을 표현하는 MFCC(Mel-Frequency

Cepstral Coefficient) 파라메터와 벡터 양자화를 이용한 문장 및 화자독립 감정 인식 시스템을 제안하여 인식 실험을 수행하였다.

II. 감정 인식 알고리즘

2.1 감정 인식을 위한 특징 파라메터

감정 인식에 사용되는 음성의 특징 파라메터로는 운율적 특징으로 피치와 에너지에 관한 파라메터가 주로 사용된다. 음성 특징 파라메터는 음성신호의 단구간에 대해 구한 피치와 에너지 값으로부터 피치 평균 (pitch mean), 피치 표준편차 (pitch standard deviation), 피치 최대 값 (pitch maximum), 에너지 평균 (energy mean), 에너지 표준편차 (energy standard deviation) 등의 통계적 정보를 감정 인식을 위한 특징으로 사용하였다[6].

MFCC(Mel Frequency Cepstral Coefficient) 파라메터는 음소의 특성을 나타내는 특징으로 음성 인식에 널리 사용되고 있다. 이러한 파라메터는 같은 음소라도 포함된 감정에 따라 음소의 형태가 다르다는 점에서 감정인식에도 사용될 수 있다.

2.2 패턴 인식 알고리즘

2.2.1 KNN (K-Nearest Neighbor Classifier)

KNN 분류기는 기준 패턴의 분포 함수를 사용하는 대신에 미리 구하여 놓은 각각의 기준 패턴과의 거리를 계산하여 가장 가까운 기준패턴의 클래스를 입력 패턴의 클래스로 결정하는 방법이다[7]. 여기서 입력 패턴과 기준 패턴간의 거리는 특정한 거리 측정 방법을 사용하여 구하며 최소 거리는 계산된 거리 측정의 결과가 가장 작은 것을 의미한다. 기준 패턴 생성 방법은 적은 수의 패턴으로 클래스를 잘 표현할 수 있어야 한다. 일반적으로 기준 패턴 생성 방법으로는 k-means 알고리즘과 LBG 알고리즘이 많이 사용된다. 거리 측정 방법은 가장 기본적인 유클리디안 거리측정 (euclidean distance) 이외에도 음성 인식에서 사용되고 있는 많은 방법들이 사용될 수 있다[7].

사전에 클래스마다 기준이 되는 기준 패턴을 생성한 후 KNN 분류기는 전체 기준 패턴 중에서 미지의 입력 패턴 x 로 부터 가장 가까운 거리에 있는 K 개의 패턴을 x 의 K-NN이라 하며, K-NN규칙은 패턴 x 의 K-NN의 각 요소가 어느 클래스에 가장 많이 속하는가를 조사하여, 그 클래스를 x 의 클래스로 결정한다.

2.2.2 벡터 양자화를 이용한 인식기

벡터 양자화(VQ : Vector Quantization)를 이용한 인식 방법은 인식 대상마다 집단화(clustering)을 통하여 코드북을 만든 후 인식 시에 양자화 오차를 계산하여 가장 적은 오차를 갖는 코드북을 입력 대상으로 인식하는 방법으로 주로 음성인식 초기단계에 사용되었고 문장독립 화자 인식에도 사용되어 왔다.

벡터 양자화를 이용한 인식 시스템의 블럭도는 그림 1과 같다. 학습 과정에서는 각 감정마다 학습 데이터를 집단화하여 코드북을 만들고 인식 단계에서는 입력 음성을 각각의 코드북으로 양자화 한 후 양자화 오차를 계산하여 그 오차가 가장 적은 코드북의 감정을 입력 음성의 감정으로 결정한다. 양자화 이러한 방법은 입력 문장의 시간적인 변화에는 상관없이 동작하므로 이러한 특징을 이용하여 문장독립 감정 인식 시스템에 응용할 수 있다. 즉 감정의 구분된 학습데이터를 사용하여 감정별 코드북을 만들어 인식에 사용하는 것이다.

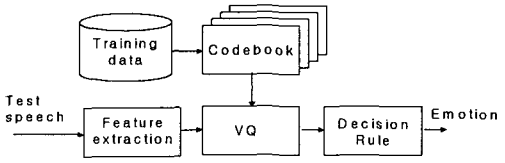

그림 1. 벡터 양자화를 이용한 감정인식 시스템 블록도

III. 실험 및 결과

3.1 감정 인식 시스템 구성

감정 인식 시스템 구현하기 위해서는 DB 구축 과정, 특징 추출 과정, 학습 및 인식 과정으로 구성된다. 특징 추출 과정에서 음성으로부터 감정 인식을 위하여 필요한 정보를 얻어내고, 이러한 정보를 이용하여 학습 과정에서 기준패턴을 생성하고, 인식 과정에서 결정법칙을 이용하여 최소 거리나 최대 확률을 갖는 기준패턴으로 인식을 한다. 기본적인 인식 시스템은 그림 2와 같다.

3.2 특징 추출

구축한 DB의 데이터를 이용한 특징 추출 과정은 다음과 같다. 전처리를 통하여 16KHz로 샘플링하고, 고주파 성분을 보강한다. 이렇게 샘플링된 신호를 20 msec씩 프레임별로 나누어 분석하여 특징벡터를 구한

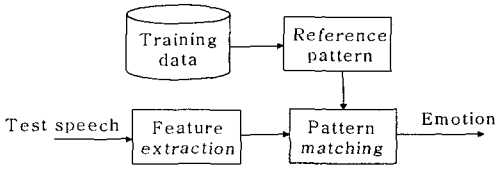

그림 2. 감정 인식 시스템 블록도

다. 본 연구에서는 음성의 특징벡터를 음소군 특징벡터와 감정 특징벡터로 구분하였는데, 음소군 특징벡터는 발성기관의 해부학적인 차이나 발성기관의 조음 방법 차이에서 나타나는 음소특징을 추출한 MFCC, 델타 MFCC와 같은 특징벡터이고, 감정 특징벡터는 감정의 표현에 기여하는 피치, 델타 피치, 델타 델타 피치, 에너지, 델타 에너지, 델타 델타 에너지 등으로 구성된 특징벡터이다. 지금까지 감정과 음성과의 상관관계에 대한 연구에 따르면, 운율적 요소 즉 감정 특징벡터가 감정을 표현하는데 많은 영향을 끼친다고 알려져 있다.

학습 및 인식 과정은 기준 패턴이나 기준 확률모델을 구하기 위해 사용되는 패턴 인식 기법에 따라 분류기가 달라진다. KNN 분류기는 제안된 방법과 비교하기 위하여 구성하였고, 음성의 감정 특징만을 이용하였다. 또한 피치, 델타 피치, 델타 델타 피치, 에너지, 델타 에너지, 델타 델타 에너지 및 MFCC, 델타 MFCC와 델타 델타 MFCC를 파라메터로 사용하여 벡터 양자화를 사용한 감정 인식 실험도 수행하였다.

3.3 데이터 베이스

데이터 베이스를 구성하기 위해서는 사용 용도를 고려한 감정 선정, 문장 선정, 녹음 대상 선정, 녹음 환경, DB 규모 등의 결정 작업이 필요하다. 본 연구에서는 인간의 주요 감정인 기쁨, 슬픔, 화남의 3가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4가지 감정을 인식 대상 감정으로 결정하였다. 음성의 녹음은 평소 감정 표현을 훈련하는 아마추어 연극단원 남/녀 각 15명을 대상으로 하였고, 모든 참여자에 대해서 표준어 사용여부 및 감정 표현능력을 심사하여 선별하였다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다. 각 화자는 45개의 문장을 4가지 감정으로 녹음하였고 녹음동안에 감정 표현이 미흡하다고 판단된 경우에는 다시 녹음을 하였다.

본 연구를 위하여 사용된 데이터의 규모는 5400(30명×4감정×45문장×1회)문장이다. 향후 실험에서는 제작된 DB 중 감정이 적절히 반영되었다고 판단되는 문장을 선별하는 주관적 평가를 거쳐 선택하였다. 구축된 DB가 화자의 감정을 어느 정도로 정확히 반영하는 지를 판단하기 위해서 평소 음성 신호 처리 실험에 숙련된 연구원들을 대상으로 주관적 평가를 실시하였다. 주관적 평가는 5400문장을 문장 당 10명이 청취한 후 평가를 하였다. 주관적 평가를 통한 데이터 베이스의 감정 평가는 다음 표와 같다. 표에서 알 수 있듯이 화자가 의도적으로 특정한 감정을 담아 발음한 음성이라 실제로 청취자들은 다르게 느낄 수 있는 것이다.

표 1. 주관적 평가 결과

정수	문장수	백분율(%)	누적백분율(%)
10	2063	38.20	38.20
9	1031	19.09	57.30
8	592	10.96	68.26
7	415	7.69	75.94
6	296	5.48	81.43

3.4 실험 결과

주관적 평가에서 100%의 정답률을 보인 데이터만을 선별하여 실험하였고, 20명의 화자(남성 10명, 여성 10명)는 학습 데이터, 10명의 화자(남성 5명, 여성 5명)를 인식 데이터로 사용하였다. 또한 총 45문장 중에 35문장은 학습에 나머지 10문장은 인식에 사용하여 화자 및 문장독립 감정인식 실험을 수행하였다.

3.4.1. KNN 분류기를 사용한 성능평가

KNN 분류기는 기존의 감정 인식 알고리듬으로 제안된 알고리듬과 비교하기 위하여 실험되었다. KNN의 경우 특징 파라메터로 피치 평균, 피치 표준편차, 피치 최대값, 에너지 평균, 에너지 표준편차를 사용하였다. KNN을 이용한 실험에서 LBG 군집화 알고리듬을 사용하여 감정별로 기준패턴을 생성하고 기준 패턴과의 거리측정을 위해 유클리디안 거리를 사용하였다. 코드북의 크기를 8, 16, 32, 64로 바꾸어 실험한 결과 인식률은 약 37.58 ~ 46.44%의 인식률을 보였으며 그 중 32일 때의 결과는 표 2와 같다. 클러스터 크기의 변화에 따른 인식률 편차는 인식률 대비5% 미만으로 크기를 최적화함에 따른 인식률 향상은 크게 기대되지 않았다.

표 2. KNN 분류기를 이용한 감성 인식 성능(%)

감정	평상	기쁨	슬픔	화남
평상	32.1	21.4	25.0	21.4
기쁨	18.2	72.7	9.1	0.0
슬픔	20.0	20.0	40.0	20.0
화남	18.2	36.4	4.5	40.9
평균	46.4			

3.4.2. 벡터 양자화를 이용한 인식기의 성능평가

피치, 델타 피치, 델타 델타 피치, 에너지, 델타 에너지, 델타 델타 및 MFCC, 델타 MFCC를 파라메터로 하여 각 감정별로 집단화(clustering)을 통한 코드북을 만든 후 입력을 테스트 입력을 양자화하여 최소의 거리를 갖는 코드북을 입력의 감정으로 인식하는 인식 시스템을 구성하여 성능을 평가하였다. 표 3은 각종 파라메터에 따른 인식 성능과 그때 사용된 코드북의 크기를 나타낸다. 여기서 사용된 파라메터의 기호는 다음과 같다.

P : 피치
DP : 델타 피치
DDP : 델타 델타 피치
E : 에너지
DE : 델타 에너지
DDE : 델타 델타 에너지
M : 멜 켑스트럼
DM : 델타 멜 켑스트럼
DDM : 델타 델타 멜 켑스트럼

표 3. 벡터 양자화를 이용한 감성 인식 성능(%)

파라메터	코드북 크기	인식율(%)
P	32	42.24
DP	64	42.24
DDP	64	38.39
E	256	41.38
DE	128	33.62
DDE	256	23.28
M	64	67.24
DM	256	56.03
DDM	256	56.03
P+DP	256	42.24
P+DP+DDP	64	45.69
DP+DDP	64	40.52
E+DE	128	46.55
E+DE+DDE	128	51.72
DE+DDE	4	41.38
M+DM	256	73.28
M+DM+DDM	127	71.55

표에서 알 수 있듯이 가장 우수한 성능을 나타낸 것은 MFCC와 델타 MFCC를 결합한 MFCC+DMFCC로 73.28%의 인식 성능을 나타내었다. 피치와 에너지는 화자종속 또는 문장종속 형태의 시스템에서는 비교적 우수한 성능을 나타내지만 문장독립 및 화자독립 감정 인식 시스템에서는 40-50%정도의 낮은 인식 성능을 나타내고 있다. 이러한 것은 시스템의 형태가 문장독립 및 화자독립 감정 인식 시스템이기 때문으로 코드북에 다양한 화자와 다양한 문장이 포함되어 있기 때문이다. MFCC의 경우에는 오히려 피치나 에너지의 영향보다는 각 감정상태에서 발음한 음성의 스펙트럼 차이를 표현하기 때문에 인식 성능이 더 우수한 것으로 판단된다.

Ⅳ. 결론

본 논문에서는 여러 가지 감정 상태에 따라 분류된 한국어 음성 데이터 베이스를 이용하여 화자 및 문장 독립 감정 인식 시스템에 관하여 연구하였다. 본 연구에서는 MFCC 파라메터와 벡터 양자화 방법을 사용한 감정인식 시스템을 제안하였으며 이것을 감정 인식 및 음성 신호 처리에 널리 사용되고 있는 음성 신호의 피치와 에너지의 평균, 표준편차와 최대 값 등 통계적인 파라메터 사용하는 시스템과 성능을 비교 평가하였다.

성능 평가를 위한 실험에서는 운율적 특징으로 입력 신호에 대한 피치 평균, 피치 표준편차, 피치 최대 값, 에너지 평균, 에너지 표준 편차를 이용하여 KNN 방법은 46.44%의 인식률을 나타내었다. 인식 실험에서 문장 종속이나 화자 종속인 경우에는 우수한 성능을 나타내는 것으로 알려진 피치 및 에너지 파라메터는 문장 독립 및 화자 독립인 경우에는 성능이 많이 저하되는 것을 알 수 있다. MFCC와 델타 MFCC를 결합한 파라메터로 하여 크기가 256개인 코드북을 사용한 경우 약 73.3%의 인식 성능을 나타내었다. 이러한 것은 기존에 사용하던 KNN 방법보다 우수한 성능을 나타내고 구조도 간단한 장점이 있다.

참고문헌

[1] Rosalind W. Picard, *Affective Computing*, The MIT Press 1997.

[2] Lain R. Murray and John L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", in *J. Accoust. Soc. Am.*, pp. 1097-1108, Feb. 1993.

[3] Frank Dellaert, Thomas Polzin, Alex Waibel, "Recognizing emotion in speech", *Proceedings of the ICSLP 96*, Philadelphia, USA, Oct. 1996

[4] Michael Lewis and Jeannette M. Haviland, *Handbook of Emotions*, The Guilford Press, 1993

[5] Thomas S. Huang, Lawrence S. Chen and Hai Tao, Bimodal emotion recognition by man and machine, *ATR Workshop on Virtual Communication Environments - Bridges over Art/Kansei and VR Technologies*, Kyoto, Japan, April 1998.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice-Hall Inc., 1993.

[7] Earl Gose, Richard Johnsonbaugh, and Steve Jost, *Pattern Recognition and Image Analysis*, Prentice Hall Inc., 1996.