GMM을 이용한 화자 및 문장 독립적 감정 인식 시스템 구현

강 면 구, 김 원 구
군산대학교 전자정보공학부

Speaker and Context Independent Emotion Recognition System using Gaussian Mixture Model

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

Abstract

This paper studied the pattern recognition algorithm and feature parameters for emotion recognition. In this paper, KNN algorithm was used as the pattern matching technique for comparison, and also VQ and GMM were used for speaker and context independent recognition. The speech parameters used as the feature are pitch, energy, MFCC and their first and second derivatives.

Experimental results showed that emotion recognizer using MFCC and their derivatives as a feature showed better performance than that using the pitch and energy parameters. For pattern recognition algorithm, GMM based emotion recognizer was superior to KNN and VQ based recognizer

I. 서론

컴퓨터가 인간의 삶에 미치는 영향이 커지면서 휴먼-컴퓨터 인터페이스 시스템 (Human-Computer Interface System)에 대한 비중 또한 높아지고 있다. 특히, 인간의 감정을 인지하고, 그에 정서적인 반응을 하는 시스템의 개발은 보다 고차원적인 휴먼-컴퓨터 인터페이스 제품을 가능하게 한다. 특히 센서가 신체 부위에 직접 닫지 않거나, 전화와 관련된 어플리케이션의 경우, 음성을 이용한 시스템의 응용은 더 많은 이점을 가지고 있다.

음성을 통한 감정 인식을 위해서는 각각의 감정이 음성에 어떠한 변화를 만들어내는가를 정확히 규명하여야 한다. 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 사용되었다[1-4].

본 연구에서는 GMM(Gaussian Mixture Model)을 이용한 화자 및 문장 독립적인 감정 인식 시스템을 제안하였다. 또한 화자 및 문장 독립적 감정 인식 시스템에 적합한 특징 파라메터를 찾기 위하여 인식 실험을 통하여 제안한 시스템에 최적의 특징 파라메터를 구하였고, KNN분류기(K-Nearest Neighbor Classifier)와 VQ(Vector Quantization)를 이용한 기존의 감정인식 시스템과 함께 인식 실험을 수행하여 제안된 시스템의 인식 성능을 평가하였다.

II. 감정 인식 시스템

2.1 음성의 특징 파라메터

음성의 음소를 나타낼 때 사용되는 파라메터로는 MFCC(Mel-Frequency Cepstral Coefficient)가 대표적

인 특징이고, 운율적 요소로는 피치, 에너지, 발음속도 등이 있는데, 감정은 주로 이러한 운율 요소에 의해서 표현된다.

MFCC 파라메터는 음소의 특성을 나타내는 특징으로 음성 인식에 널리 사용되고 있으며, 같은 음소라도 포함된 감정에 따라 음소의 형태가 다르다는 점에서 감정 인식에도 사용될 수 있다.

운율적 특징은 단구간에 대해 구한 피치와 에너지의 평균(mean), 표준편차(standard deviation), 최대 값(maximum) 등의 통계적 정보가 감정 인식을 위한 특징 파라메터로 사용되어진다[5].

2.2 패턴 인식 알고리즘

2.2.1 KNN 분류기를 이용한 인식기

KNN 분류기는 기준 패턴의 분포 함수를 사용하는 대신에 클래스마다 기준이 되는 기준 패턴을 생성한 후, 전체 기준 패턴 중에서 미지의 입력 패턴 x 로 부터 가장 가까운 거리에 있는 K 개의 패턴을 x 의 KNN으로 정하고 패턴 x 의 KNN의 각 요소가 어느 클래스에 가장 많이 속하는가를 조사하여 그 클래스를 x 의 클래스로 결정하는 방법이다. 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[6].

2.2.2 VQ를 이용한 인식기

VQ를 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[8].

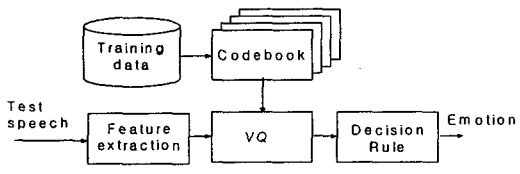

그림 1. VQ를 이용한 감정인식 시스템 블록도

2.2.3 GMM을 이용한 인식기

가우시안 혼합 분포(Gaussian mixture density)는 음성 신호를 M개의 각 성분 분포(component density)들의 선형 조합으로 근사화를 할 수 있으며 긴 구간의 신호에 대해서도 표현이 가능하다. 가우시안 혼합 분포는 식 2-1로 표현된다[7].

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (2\text{-}1)$$

가우시안 혼합 분포를 표현하기 위해서는 평균 벡터(means vector)들과 공분산 행렬(covariance matrix), 그리고 가중(mixture weights) 이 세 가지의 파라메터가 필요하다. 이들 세 가지 파라메터의 집합이 어떤 화자나 감정의 가우시안 혼합 분포를 표현할 수 있는 모델이 되며 이 집합을 GMM이라고 하고 식 2-2와 같이 표현된다.

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M. \qquad (2\text{-}2)$$

GMM을 이용한 인식 시스템은 학습 과정에서 감정별 학습 데이터마다 ML(Maximum Likelihood) Estimation과 EM(Expectation Maximization) 알고리즘을 이용하여 최대 가우시안 혼합 분포 값을 갖는 GMM의 파라메터를 추정하고 인식 과정에서는 추정된 감정별 GMM 파라메터를 이용하여 입력된 음성 데이터의 특징 벡터에 대한 각각의 가우시안 혼합 분포를 구하여 그 중 가장 큰 확률 값을 가지는 GMM의 감정을 입력된 음성 데이터의 감정으로 선택하게 된다.

Ⅲ. 실험 및 결과

3.1 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 구한다. 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 구

성된 특징벡터이다.

3.2 데이터 베이스

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, 각 화자는 45개의 문장을 4가지 감정으로 녹음하여 총 5400개(30명×4감정×45문장×1회)의 문장을 구성하였다.

3.3 실험 결과

DB 중 주관적 평가를 수행하여 감정이 적절히 반영되었다고 판단되는 문장만을 선별하였다. 주관적 평가는 5400문장을 문장 당 10명이 청취한 후 감정 평가를 하여 70%이상의 정답률을 보인 데이터만을 선별하여 4100문장을 감정인식 실험의 데이터로 사용하였다.

3.3.1. KNN 분류기를 사용한 성능평가

KNN 분류기는 기존의 감정 인식 알고리듬으로 제안된 알고리듬과 비교하기 위하여 실험되었다. 특징 파라메터로 피치 평균, 피치 표준편차, 피치 최대값, 에너지 평균, 에너지 표준편차를 사용하였다. 기준패턴을 생성하기 위해 LBG 군집화 알고리듬을 사용하였고, 기준 패턴과의 거리측정은 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.

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			

3.3.2. VQ를 이용한 인식기의 성능평가

피치(P), 델타 피치(DP), 델타 델타 피치(DDP), 에너지(E), 델타 에너지(DE), 델타 델타 에너지(DDE) 및 MFCC(M), 델타 MFCC(DM), 델타 델타 MFCC(DDM)를 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.

표 3. 벡터 양자화를 이용한 감성 인식 성능(%)
('+'기호는 파라메터의 결합을 의미)

파라메터	코드북 크기	인식률(%)
P	16	36.40
P+DP	16	37.18
P+DP+DDP	128	42.95
E	16	42.05
E+DE	512	66.77
E+DE+DDE	256	45.64
M	32	31.94
M+DM	512	67.86
M+DM+DDM	512	68.12

표 3은 각 특징 파라메터마다의 VQ 최대 인식률을 나타낸 것으로 그 중 코드북이 512일 때 M+DM+DDM에서 68.12%로 가장 우수한 인식 성능을 보였고 피치와 에너지에서는 30~45%정도의 낮은 인식 성능을 나타내고 있다. 이러한 것은 시스템의 형태가 감정 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.

3.3.3 GMM을 이용한 인식기의 성능평가

혼합 차수 M의 개수와 분산의 최소값 σ^2_{min}의 결정은 인식의 성능에 영향을 미치기 때문에 다음과 같이 몇 가지 값을 선정하여 각각에 대해 인식 성능을 평가하였다.

성분 혼합 차수 M : 1, 2, 4, 8, 16, 32, 64, 128, 256, 512

최소 분산 σ^2_{min} : 0.002, 0.005, 0.01, 0.05, 0.1

최소 분산의 값은 0.002부터 실험에 적용하였다. 그것은 0.002보다 작은 0.001의 실험에서 분산이 너무 작아 가우시안 성분 분포가 너무 예리해져 모델에 대한 확률 값을 구하지 못하고 overflow가 발생했기 때문이다. 그리고 이러한 특이성(singularity)은 GMM을 이용한 화자인식에 관한 논문에서도 언급되었었다[7].

표 4는 GMM을 이용한 인식기에서 특징별 인식 성능 중 가장 좋은 인식률들만을 나타낸 것이다. 표에서 알 수 있듯이 피치와 에너지는 35.41~46.69% 정도

의 인식률을 보였으며, σ^2_{min}을 바꾸었을 때 성능의 차이는 없었다. 혼합 차수 M의 선정에서는 256이나 512에서 다른 차수들에 비해 대체적으로 인식률이 높았으며, σ^2_{min}의 선정에서는 0.002와 0.005에서 우수한 인식 성능을 보였다. M+DM+DDM+DE+DDE에서 혼합 차수가 512이고 σ^2_{min}가 0.002일 때 73.77%로 가장 좋은 인식률을 보였다.

표 4. GMM을 이용한 인식기의 성능 평가

파라메터	# of M	σ^2_{min}	인식률(%)
P	4	-	38.43
P+DP	256	0.005	40.60
P+DP+DDP	512	0.01	46.69
E	32	-	35.41
E+DE	128	-	40.07
E+DE+DDE	32	-	44.11
M	256	0.002	67.81
M+DM	256	0.005	73.37
M+DM+DDM	256	0.002	71.71
M+DM+DDM+E+DE+DDE	512	0.002	71.13
M+DM+DDM+DE+DDE	512	0.002	73.77
M+DM+DE+DDE	512	0.005	72.87
M+DM+E+DE	256	0.002	72.53
M+E	512	0.002	71.16

표 3과 4를 비교한 경우, 같은 특징 파라메터를 사용하는 VQ를 이용한 인식기의 성능에 비해서 GMM을 이용한 인식기가 더 높은 인식률을 보였으며, MFCC와 에너지를 결합한 파라메터는 모두 71%이상의 최대 인식률로 다른 특징 파라메터보다 좋은 성능을 보였다.

Ⅳ. 결론

본 연구에서는 GMM을 이용한 화자 및 문장 독립 감정 인식 시스템을 제안하였다. GMM은 가우시안 분포들을 성분으로 하여 각 성분을 표현하는 파라메터를 집합으로 하여 음성 신호 전체를 모형화하기 때문에 긴 구간의 음성 신호의 표현이 가능하며 음성 신호의 시간적 변화와는 무관한 특성을 가지고 있어 문장 독립적 시스템에 적합하다고 판단하였다.

감정 인식 및 음성 신호 처리에 널리 사용되고 있는 피치와 에너지의 통계적인 파라메터를 사용하는 KNN 분류기를 이용한 시스템과 MFCC와 같은 음소적 특징 파라메터를 사용하여 화자 및 문장 독립적 시스템을 구현한 VQ를 이용한 시스템으로 성능을 비교 평가하였다. 운율적 특징을 이용한 KNN분류기의 성능이 가장 떨어졌으며 화자 및 문장 독립적 시스템에 적합하지 않았다. VQ를 이용한 시스템의 경우 화자 및 문장 독립적 시스템에 좋은 적응을 보였으며 최대 68.12%의 인식 성능을 보였다. VQ와 비추어 볼 때, GMM을 이용한 인식기는 최대 73.77%의 인식 성능을 보여 화자 및 문장 독립적 감정인식 시스템에서 보다 적합한 시스템으로 적용할 수 있음을 보였다.

향후 HMM(Hidden Markov Model)과 같은 모델링 기법들과 GMM을 병행한 형태의 시스템을 감정 인식에 적용하는 연구도 좋은 결과를 얻을 수 있을 것이다.

감사의 글

본 연구보고서는 정보통신부 정보통신연구진흥원에서 지원하고 있는 정보통신기초연구지원사업의 연구결과입니다.

참고문헌

[1] 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.

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

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

[4] 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.

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

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

[7] Douglas A. Renolds and Ricard C. Rose, "Robust Text-Independent Speaker Indentification Using Gaussian Mixture Speaker Models", *IEEE Trans. on Speech and Audio Processing*, vol. 3, No. 1, pp.72-83, Jan. 1995.

[8] 강면구, 김원구, "음성을 이용한 화자 및 문장 독립 감정 인식", 대한전자공학회 하계학술대회 제25권 1호 pp. 377-380, 2002년 6월.