Proceedings of KFIS Spring Conference 2004 Volume 14, Number 1

감정 인식을 위한 음성의 특징 파라메터 비교

The Comparison of Speech Feature Parameters for Emotion Recognition

김 원 구
군산대학교 전자정보공학부

Weon-Goo Kim
School of Electronic and Information Eng., Kunsan National University
E-mail : wgkim@kunsan.ac.kr

요 약

In this paper, the comparison of speech feature parameters for emotion recognition is studied for emotion recognition using speech signal. 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. MFCC parameters and their derivatives with or without cepstral mean subtraction are also used to evaluate the performance of the conventional pattern matching algorithms. Pitch and energy parameters were used as a prosodic information and MFCC parameters were used as phonetic information. In this paper,

In the Experiments, the vector quantization based emotion recognition system is used 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. The vector quantization based emotion recognizer achieved recognition rates of 73.3% for the speaker and context independent classification.

1. 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 박동 수, 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.

많은 심리학자, 음향학자들에 의해서 화자의 감성과 음성과의 관계가 연구되었는데[1,3,4], 1993년까지의 주요한 연구결과들은 감성음성 합성을 연구하는 Lain R. Murray가 발표한 논문[9]을 통하여 집대성되었다. 이 논문에서 화자의 감성을 반영하는 요소로서 발음 속도, 피치 평균, 피치 변화 범위, 발음 세기, 음질, 피치의 변화, 발음법 등으로 요약하고, 기쁨, 화남, 슬픔, 두려움, 혐오감 등의 주요 감성들을 표현할 때 이들 요소들의 차이점들을 정리하였다.

이렇게 밝혀진 감성 인식의 특징들을 바탕으로 최근에 몇 가지 감성 인식 및 합성 관련 실험 논문이 발표되었다. MIT 미디어랩의 Deb Roy[5] 및 Carnegie Mellon University의 Frank Dellaert[2]는 MLB(Maximum-Likelihood Bayes), KR(Kernel Regression), KNN(K-Nearest Neighbor)분류기 등의 패턴 인식 기법을 사용하여 실험하였다.

본 연구에서는 음성 신호 분석을 통해 화자의

감성 인식을 위한 연구를 수행하며, 이에 필수적인 감정 상태에 따라 분류된 한국어 음성 데이터 베이스를 구축 구축하고 인간의 감정을 잘 표현하는 특징 파라메터를 찾아 패턴 비교 알고리즘을 사용하여 그 성능을 평가하는 것에 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 정보와 MFCC(Mel-Frequency Cepstral Coefficient)와 그 미분을 CMS(Cepstral Mean Subtraction) 방법[7]과 합께 선택적으로 사용하였다.

2. 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, 피치 표준편차, 피치 최대 값, 에너지 평균, 에너지 표준편차 등의 통계적 정보를 감정 인식을 위한 특징으로 사용하였다[7].

MFCC 파라메터와 그 미분값들은 음소의 특성을 나타내는 특징으로 음성 인식에 널리 사용되고 있다. 이러한 파라메터는 같은 음소라도 포함된 감정에 따라 음소의 형태가 다르다는 점에서 감정인식에도 사용될 수 있다.

또한 분 연구에서는 음성 인식에서 채널의 차이에 따른 인식 성능 저하를 감소시키는 이유로 널리 사용되고 있는 CMS(Cepstral Mean Subtraction) 방법을 사용하여 이것이 감정 인식에 미치는 영향에 대하여 분석하였다.

2.2 패턴 인식 알고리즘

2.2.1 KNN(K-Nearest Neighbor) 분류기

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[11]. 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[8].

사전에 클래스마다 기준이 되는 기준 패턴을 생성한 후 KNN 분류기는 전체 기준 패턴 중에서 미지의 입력 패턴 x 에 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. K 가 2 이상인 경우에는 K-NN 규칙은 K=1인 규칙보다 많은 정보를 참조하기 때문에 인식결과가 보다 양호하다고 알려져 있지만, 패턴의 분포에 따라서 그 반대의 경우도 있다.

2.2.2 벡터 양자화를 이용한 인식기

벡터 양자화(VQ : Vector Quantization)를 이용한 인식 방법은 인식 대상마다 집단화를 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 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.

3. 실험 및 결과 고찰

3.1 감성 인식 시스템

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.

3.2 특징 추출

본 연구에서는 화자독립-문장독립형 인식 시스템을 구현하기 위해서 한국어 감정 음성 DB를 직접 구축하여 실험하였다. 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.

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 및 MFCC, 델타 MFCC를 파라메터로 사용하여 VQ를 사용한 감성 인식 실험도 수행하였다. 또한 CMS를 적용한 것과 적용하지 않은 MFCC 파라메터에 대해서도 비교 실험을 수행하였다.

3.3 데이터 베이스

대상 감정은 인간의 주요 감정인 기쁨, 슬픔, 화남의 3가지 감정과 이들의 기준이 되는 평상 감정을 포함한 4가지 감정으로 결정하였다. 녹음은 평소 음성을 통한 감정 표현을 훈련하는 아마추어 연극단원 남/녀 각 15명을 대상으로 하였고, 모든 참여자에 대해서 표준어 사용여부 및 감정 표현능력을 심사하여 자격자를 선별하였다. 녹음 시에 감정별로 전체 문장세트를 4회씩 발음하여 그 중 우수한 3회분을 선택하였다. 제시된 문장이외에 감정에 수반될 수 있는 보조적인 효과음은 배제하였다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다.

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.

3.4 실험 결과

주관적 평가에서 100%의 정답률을 가진 데이터만을 선별(전체 5400문장의 38.2%)하여 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 감성인식 실험을 수행하였다.

3.4.1. KNN 분류기를 사용한 성능평가

KNN의 경우 특징 파라메터로 피치 평균, 피치 표준편차, 피치 최대값, 에너지 평균, 에너지 표준편차를 사용하였다. KNN을 이용한 실험에서 LBG 군집화 알고리듬을 사용하여 감정별로 기준 패턴을 생성하고 기준 패턴과의 거리측정을 위해 유클리디안 거리를 사용하였다. 코드북의 크기를 8, 16, 32, 64로 바꾸어 실험한 결과 인식률은 약 37.6~46.4%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.

Table 2. Recognition rates using KNN(%)

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. VQ를 사용한 인식기의 성능평가

피치, 델타 피치, 델타 델타 피치, 에너지, 델타 에너지, 델타 델타 및 MFCC, 델타 MFCC를 파라메터로 하여 각 감정별로 집단화을 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 그 때 사용된 코드북의 크기를 나타낸다.

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 있

Table 3. Recognition rates using VQ(%) (P : pitch, DP : delta pitch, DDP : delta delta pitch, E : energy, DE : delta energy, DDE : delta delta energy, M : mel-cepstrum, DM : delta mel-cepstrum, DDM : delta delta mel-cepstrum), CMS(Cepstral Mean Subtraction)

파라메터	코드북 크기	인식율(%)
P	32	42.2
P+DP	256	42.2
P+DP+DDP	64	45.7
E	256	41.4
E+DE	128	46.6
E+DE+DDE	128	51.7
M	64	67.2
M+DM	256	73.3
M+DM+DDM	127	71.6
M(CMS)	512	60.5
M+DM(CMS)	256	61.1
M+DM+DDM(CMS)	512	62.3

다. 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. 또한 CMS를 사용한 경우에는 인식 시스템의 성능이 저하되는 것을 알수 있다. 이러한 것은 CMS 과정이 채널의 차이를 보정해주는 것뿐만 아니라 각 감정의 차이점을 모호하게 하는 것을 알 수 있다.

4. 결 론

본 연구에서는 다양한 입력문장에 담긴 화자의 감정을 인식할 수 있는 문장독립 및 화자 독립 감정 인식 알고리듬으로 VQ를 이용한 방법을 사용하였다. 감정인식 시스템은 MFCC와 델타 MFCC를 파라메터로 하여 코드북을 만들고 양자화 오차를 사용하여 인식을 수행하였다. 피치 및 에너지 파라메터는 문장 종속이나 화자 종속인 경우에는 우수한 성능을 나타내지만 문장 독립 및 화자 독립인 경우에는 성능이 많이 저하되는 것을 알 수 있다. MFCC외 델타 MFCC를 결합한 파라메터로 하여 크기가 256개인 코드북을 사용한 경우 약 73.3%의 인식 성능을 나타내었다. 또한 MFCC 파라메터에 CMS를 사용한 경우에는 감정 인식 시스템의 성능이 저하되었다.

감사의 글

본 연구는 한국과학재단 목적기초연구 (R05-2003-000-12043-0) 지원으로 수행되었음

5. 참고문헌

[1] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates", Journal Acoustical Society of America, Vol. 52, No. 4, pp. 1238-1250, 1972.

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

[3] K. R. Scherer, D. R. Ladd, and K. E. A. Silverman, "Vocal cues to speaker affect: Testing two models", *Journal Acoustical Society of America*, Vol. 76, No. 5, pp. 1346-1355, Nov. 1984.

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

[5] D. Roy and A. Pentland, "Automatic spoken affect analysis and classification", in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 363-367, Killington, VT, Oct. 1996.

[6] 강봉석, *음성 신호를 이용한 감정 인식*, 석사학위논문, 연세대학교, 1999년 12월

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

[8] R.O. Duda, and P.E. Hart, *Pattern classification and scene anlaysis*, John Wiley & Sons Inc., 1973.

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