음성 신호를 사용한 GMM기반의 감정 인식

GMM-based Emotion Recognition Using Speech Signal

한국음향학회지 제23권 제3호 pp. 235~241 (2004)

음성 신호를 사용한 GMM 기반의 감정 인식

GMM-based Emotion Recognition Using Speech Signal

강 면 구*, 서 정 태**, 김 원 구*
(Myoun-Goo Kang*, Jeong-Tae Seo**, Weon-Goo Kim*)

*군산대학교 전자정보공학부, ** 충주대학교 정보제어공학과
(접수일자: 2003년 7월 24일; 채택일자: 2004년 4월 1일)

본 논문은 화자 및 문장 독립적 감정 인식을 위한 특징 파라메터와 패턴인식 알고리즘에 관하여 연구하였다. 본 논문에서는 기존 감정 인식 방법과의 비교를 위하여 KNN을 이용한 알고리즘을 사용하였고, 화자 및 문장 독립적 감정 인식을 위하여 VQ와 GMM을 이용한 알고리즘을 사용하였다. 그리고 특징으로 사용한 음성 파라메터로 피치, 에너지, MFCC, 그리고 그것들의 1, 2차 미분을 사용하였다. 실험을 통해 피치와 에너지 파라메터를 사용하였을 때보다 MFCC와 그 미분들을 특징 파라메터로 사용하였을 때 더 좋은 감정 인식 성능을 보였으며, KNN과 VQ보다 GMM을 기반으로 한 인식 알고리즘이 화자 및 문장 독립적 감정 인식 시스템에서 보다 적합하였다.

핵심용어: 음성 신호, 감정 인식, 화자 및 문장 독립, GMM, MFCC,
투고분야: 음성처리 분야 (2.5)

This paper studied the pattern recognition algorithm and feature parameters for speaker and context independent 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 its derivatives 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.

Keywords: speech signal, emotion recognition, speaker and context independent, GMM, MFCC
ASK subject classification: Speech signal processing (2.5)

I. 서 론

컴퓨터가 인간의 삶에 미치는 영향이 커지면서 휴먼-컴퓨터 인터페이스 시스템 (human computer interface system)에 대한 비중 또한 높아지고 있으며 인간의 감정을 인지하고, 그에 정서적인 반응을 하는 시스템의 개발은 보다 고차원적인 휴먼-컴퓨터 인터페이스 제품을 가능하게 한다. 인간의 감정에 대한 정보는 얼굴표정, 음성, 몸 동작, 심장 박동 수, 체온, 혈압 등의 다양한 방법으로 얻을 수 있고, 어플리케이션에 따라 감정 정보 취득 방법 또한 달라진다. 특히, 센서가 신체부위에 직접 닫지 않거나, 전화와 같이 음성 신호에 의존하여야 하는 어플리케이션의 경우, 음성을 이용한 시스템의 응용은 많은 이점을 가지고 있다.

음성에는 화자의 감정뿐만 아니라 전달하고자 하는 내용의 단어나 문법에서의 강세 부분, 지역적인 특성이 가미된 억양 등 정서이외의 것들이 많이 담겨져 있기 때문에, 음성에서 감정만을 따로 떼어서 분석하는데 어려움이 있다. 음성을 통한 감정 인식을 위해서는 각각의 감정이 음성에 어떠한 변화를 만들어내는가를 정확히 규명하여야 하는데, 이러한 음성과 감정과의 상관관계에 대한 연구는 서구의 음향학자들과 심리학자들에 의해 먼저 이루어졌다[2-5]. 이 연구결과를 바탕으로 다양한 어플리케이션의 개발이 시도되고 있으며, 특히 음향이나 시각 정보를 처리 및 저장하는 기술과 녹음하는 기술의 진보, 착용하는 컴퓨터의 개

책임저자: 강 면 구 (kkm9@kunsan.ac.kr)
573-701 전북 군산시 미룡동 산 68
군산대학교 전자정보공학부
(전화: 063-469-4701; 팩스: 063-469-4699)

발. point-and-click으로부터 sense-and-feel로의 휴먼-컴퓨터 인터페이스의 진행, Sony사의 Aibo와 Tiger Electronics의 Furby와 같은 여러 감정을 이해하고 표현하는 로봇의 제품화 등은 음성을 이용한 감정 인식과 감정 합성에 관련된 연구에 관심을 높이고 있다[1][6][7].

감정 인식은 지금까지 많이 연구되어 온 음성 인식에서 그 시발점을 찾을 수도 있으나, 특징 추출 및 패턴 인식 알고리즘 선택에 있어서 차이가 있다. 특징벡터 선택에 있어서 음성 인식의 경우 음소를 모델링하는 요소를 주로 이용하는 반면, 감정 인식에 있어서는 운율적 요소를 활용하여야 한다. 특징 선택과 함께 패턴 매칭 알고리듬의 선택 또한 중요한 요소인데, 특징을 확률적으로 모델링하고 추출한 특징을 이용하여 감정을 모델링하는 방식에 따라 패턴 매칭 알고리듬이 달리 선택될 수 있다. MIT대학의 Deb Roy 및 Carnegie Mellon 대학의 Frank Dellaert는 MLB (Maximum-Likelihood Bayes), KR (Kernel Regression), KNN (k-Nearest Neighbor)분류기 등 기본적인 패턴 인식 기법을 이용하였고[8][9], 일본 Seikei University의 Jun Sato는 감정 합성에 있어서 Neural Network을 이용한 Emotion Space 개념을 선보였다[10]. 또한 Microsoft Research China의 Feng Yu는 SVM (Support Vector Machine) 분류기를 이용한 감정 인식을 선보였다[11].

감정 인식 시스템은 학습데이터와 시험데이터의 구성에 따라, 화자종속-문장종속 (speaker and context dependent), 화자독립-문장종속 (speaker independent, context dependent), 화자종속-문장독립(speaker dependent, context independent), 화자독립-문장독립 (speaker and context independent) 시스템 등으로 나눌 수 있다. 어떤 시스템을 선택하느냐에 따라서 패턴 인식 알고리듬의 선택과 적용 방법이 달라진다. 화자종속의 경우는 추출한 특징의 절대값이 주요 특징이 될 수 있으나, 화자 독립의 경우는 화자간의 차이를 보상하기 위하여 특징의 시간적인 미분값, 이중 미분값이 더 중요한 특징이 된다[12].

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 적합한 특징 파라메터를 구하였고, KNN 분류기 (K-Nearest Neighbor Classifier)와 VQ (Vector Quantization)를 이용한 기존 감정 인식 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 대표적인 특징이고, 운율적 요소로는 피치, 에너지, 발음속도 등이 있는데, 감정은 주로 이러한 운율 요소에 의해서 표현된다. 감정 인식을 위해서는 음성에서 이러한 운율 정보를 잘 반영하는 특징을 찾아내어 모델링을 해야 된다. 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 에너지 값으로부터 피치 평균 (pitch mean), 피치 표준편차 (pitch standard deviation), 피치 최대값 (pitch maximum), 에너지 평균 (energy mean), 에너지 표준편차 (energy standard deviation) 등의 통계적 정보가 감정 인식을 위한 특징 파라메터로 사용되어진다[12].

2.2 패턴 인식 알고리즘

2.2.1 KNN 분류기를 이용한 인식기

KNN 분류기는 기준 패턴의 분포 함수를 사용하는 대신에 미리 구하여 놓은 각각의 기준 패턴과의 거리를 계산하여 가장 가까운 기준 패턴의 클래스를 입력 패턴의 클래스로 결정하는 방법이다. 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 거리 측정 (euclidean distance) 이외에도 음성 인식에서 사용되고 있는 많은 방법들이 사용될 수 있다[13].

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 KNN이라 하며, KNN 규칙은 패턴 x 의 KNN의 각 요소가 어느 클래스에 가장 많이 속하는가를 조사하여, 그 클래스를 x 의 클래스로 결정한다.

2.2.2 VQ를 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 방법

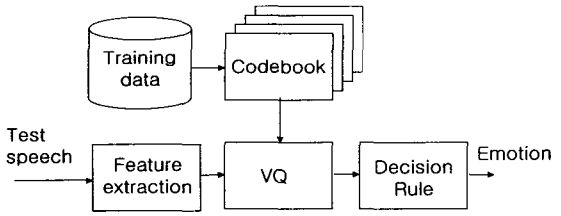

그림 1. VQ를 이용한 감정 인식 시스템 블록도
Fig. 1. Block diagram of emotion recognition system using VQ.

으로 주로 음성인식 초기단계에 사용되었고 문장독립 화자 인식에도 사용되어 왔다.

벡터 양자화를 이용한 인식 시스템의 블록도는 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. 이러한 방법은 입력 문장의 시간적인 변화에는 상관없이 동작하므로 이러한 특징을 이용하여 문장독립 감정 인식 시스템에 응용할 수 있다. 즉 감정의 구분된 학습 데이터를 사용하여 감정별 코드북을 만들어 인식에 사용하는 것이다[15].

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 (1)로 표현된다[14].

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

여기서 bi(·)는 가우시안 확률 분포이고 pi는 가중치 (mixture weight)이다. 가우시안 혼합 분포를 표현하기 위해서는 평균 벡터 (means vector) μ_i 들과 공분산 행렬 (covariance matrix) Σ_i 그리고 가중치 p_i 파라메터가 필요하다. 이들 세 가지 파라메터의 집합이 어떤 화자나 감정의 가우시안 혼합 분포를 표현할 수 있는 모델이 되며 이 집합을 GMM이라고 하고 식 (2)와 같이 표현된다.

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

GMM을 이용한 인식 시스템은 그림 2와 그림 3의 블록도와 같으며, 학습 과정에서 감정별 학습 데이터마다 ML (Maximum Likelihood) Estimation과 EM (Expectation Maximization) 알고리즘을 이용하여 최대 가우시안 혼합 분포값을 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 확률값을 가지는 GMM의 감정을 입력된 음성 데이터의 감정으로 선택하게 된다.

III. 실험 및 결과

3.1 특징 추출

구축한 데이터 베이스를 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 음소 특징을 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.

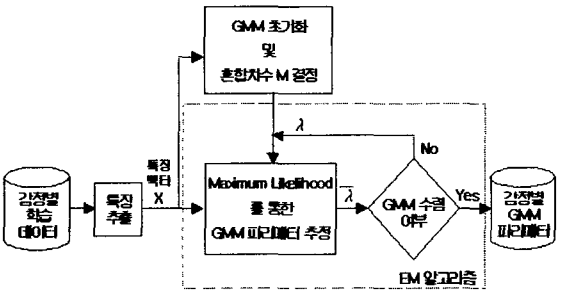

그림 2. GMM 파라메터의 학습과정 블록도
Fig. 2. Block diagram of GMM parameter training processing.

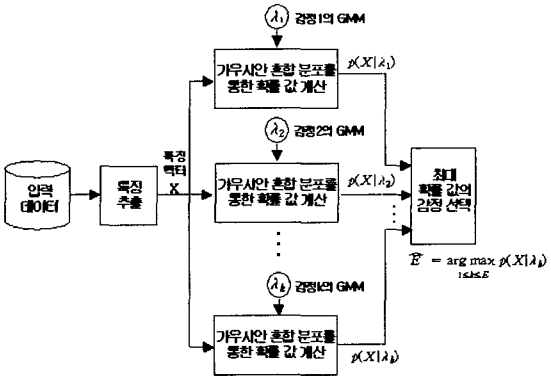

그림 3. GMM을 이용한 감정 인식 과정의 블록도
Fig. 3. Block diagram of emotion recognizing processing using GMM.

표 1. KNN 분류기를 이용한 인식기의 인식률(%)
Table 1. Recognition rate of recognizer using KNN classifier(%).

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			

3.2 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 "여보세요" 등의 한국어 45개의 문장을 4가지 감정으로 녹음하여 총 5400개 (30명×4감정×45문장×1회)의 문장을 구성하였다[15].

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 감정인식 실험의 데이터로 사용하였다. 또한 화자 및 문장 독립적 시스템을 구현하기 위하여 학습과정과 인식과정의 데이터를 화자와 문장을 각각 다르게 하여 학습에 2237문장과 인식에 274 문장을 각각 사용하였다.

3.3.1 KNN 분류기를 이용한 인식기의 성능 평가

KNN 분류기는 특징 파라메터로 피치 평균, 피치 표준편차, 피치 최대값, 에너지 평균, 에너지 표준편차가 감정 인식 시스템의 파라메터로 사용되는 경우에 감정 인식 성능을 평가하기 위하여 실험되었다. KNN 분류기는 감정의 운율적 정보를 잘 반영하는 특징인 피치와 에너지에 관한 파라메터들이 적합함을 보이는 기존의 대표적인 감정 인식 알고리듬이다.[3] 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 인식률은 표 1과 같다.

3.3.2. VQ를 이용한 인식기의 성능 평가

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 집단화를 통한 코드북을 만든 후 입력을 테스트 입력을 양자화하여 최소의 거리를 갖는 코드북을 입력의 감정으로 인식하는 인식 시스템을 구성하여 성능을 평가하였다. 표 2는 각종 파라메터에 따른 인식 성능과 그때 사용된 코드북의 크기를 나타낸다.

표 2. VQ를 이용한 인식기의 인식률(%)
('+'기호는 파라메터들의 결합을 의미)
Table 2. Recognition rate of recognizer using VQ(%)
(Symbol of '+' means to parameters combination.)

파라메터	코드북 크기	인식률(%)
P	16	36.40
P+DP	16	42.05
P+DP+DDP	128	44.59
E	16	31.94
E+DE	512	37.18
E+DE+DDE	256	42.95
M	32	66.77
M+DM	512	67.86
M+DM+DDM	512	68.12

표 2는 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 $\sigma^2_{\min}$의 결정은 인식의 성능에 영향을 미치기 때문에 성분 혼합 차수 M은 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 으로 하고 최소 분산 $\sigma^2_{\min}$ 은 0.002, 0.005, 0.01, 0.05, 0.1 의 값으로 선정하여 각각에 대해 인식 성능을 평가하였다.

그림 4는 $\sigma^2_{\min}$가 0.002일 때 각 혼합 차수 M에 대한 특징별 평균 인식 성능을 보여준다. 그림 4(a)는 피치와 관련된 파라메터에 대한 인식 성능으로 30~45%의 낮은 인식률을 보였으며 혼합 차수가 증가하여도 인식률에는 영향을 주지 못하였다. 그림 4(b)는 에너지와 관련된 파라메터들에 대한 인식 성능으로 그림 4(a)에서와 마찬가지로

30~45%의 낮은 인식률을 보였으며, 최소 분산 σ^2_{min}의 값을 변화시켜도 인식률에는 영향을 주지 못하였다. 그림 4(c)는 MFCC와 관련된 특징 파라메터들은 피치나 에너지에 비해 높은 인식률과 안정적인 인식 성능을 보였으며 혼합 차수가 증가할수록 인식률이 증가함을 볼 수 있다.

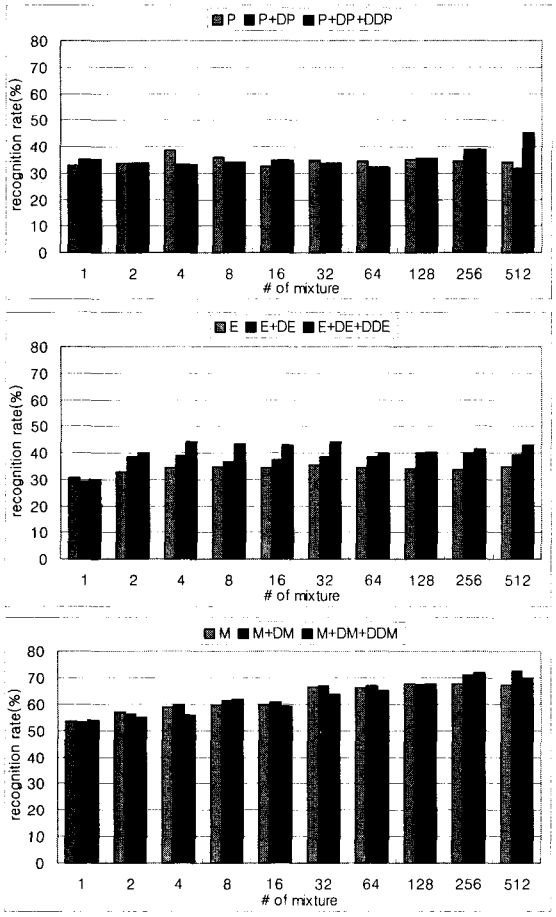

그림 4. 혼합 차수 M에 따른 파라메터들의 인식 성능 (σ^2_{min} =0.002)
(a) 피치 (b) 에너지 (c) MFCC

Fig. 4. Recognition performance of parameters according to mixture weights M (σ^2_{min} = 0.002).
(a) pitch (b) energy (c) MFCC

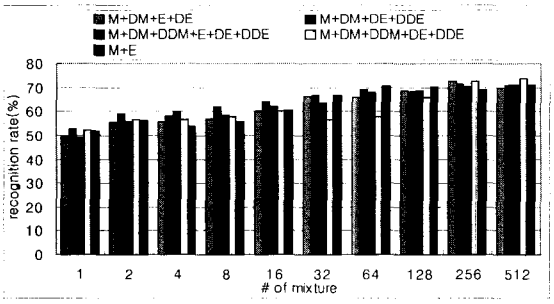

그림 5. 혼합 차수 M에 따른 MFCC와 에너지 파라메터들의 결합 형태별 인식 성능(σ^2_{min} =0.002)

Fig. 5. Recognition performance of MFCC and energy parameters combination according to mixture weights M (σ^2_{min} =0.002)

표 3. M+DM+DDM+DE+DDE의 감정별 인식률(%) (σ^2_{min} =0.002, M=512)

Table 3. Recognition rate of M+DM+DDM+DE+DDE about each emotion(%) (σ^2_{min} =0.002, M=512).

감정	평상	기쁨	슬픔	화남
평상	71.05	14.47	6.58	7.90
기쁨	6.00	64.00	12.00	18.00
슬픔	12.09	12.09	75.82	0.00
화남	3.51	12.28	0.00	84.21
평균	73.77			

표 4. GMM을 이용한 인식기의 인식률(%)

Table 4. Recognition rate of recognizer using GMM(%).

파라메터	# of M	σ^2_{min}	인식률(%)
P	4	0.002	38.43
P+DP	256	0.005	40.60
P+DP+DDP	512	0.01	46.69
E	32	0.002	35.41
E+DE	128	0.002	40.07
E+DE+DDE	32	0.002	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

인식률이 증가하여 혼합 차수가 64개 이상에서 65%이상의 인식률이 보였다. 또한 MFCC에 델타 파라메터들을 결합하여 최대 인식률이 70%이상으로 증가함을 보였다.

다음으로 MFCC 파라메터에 에너지 파라메터를 결합시켜 그에 대한 인식 성능을 평가하였으며 그 결과는 그림 5와 같다. 대부분의 파라메터가 혼합 차수가 증가할수록 인식률이 증가하였고 혼합 차수 64개 이상에서 65%이상의 인식률을 보였다. 여기서 최대 인식률은 모두 70%이상의 인식률을 보였고 MFCC만을 사용할 때보다 조금 더 우수한 결과를 보였다.

표 3은 가장 높은 인식률을 보인 MFCC와 델타 MFCC, 델타 델타 MFCC 그리고 에너지와 델타 에너지가 결합된 파라메터 (M+DM+DDM+DE+DDE)의 감정별 인식률을 나타낸 것이다. 여기서 화남이 가장 우수한 성능을 보였고 기쁨이 가장 낮은 성능을 나타내었다.

표 4는 GMM을 이용한 인식기에서 특징별 인식 성능 중 가장 좋은 인식률들만을 나타낸 것이다. 표에서 알 수 있듯이 피치와 에너지는 35.41~46.69% 정도의 낮은 인식률을 보였고, 혼합 차수 M의 선정에서는 256이나 512

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

표 2와 4를 비교한 경우, 같은 특징 파라메터를 사용하는 VQ를 이용한 인식기의 성능에 비해서 GMM을 이용한 인식기가 더 좋은 인식률을 보였으며, MFCC와 에너지를 결합한 파라메터는 모두 71%이상의 최대 인식률로 다른 특징 파라메터보다 좋은 성능을 보였다.

VI. 결 론

본 연구에서는 GMM을 이용한 감정 인식 시스템을 제안하였다. 감정 인식 시스템은 감정별로 GMM을 추정한 후, 입력된 음성 신호에 대한 GMM의 관찰 확률을 통해 입력된 음성 신호의 감정을 인식하였다. GMM은 가우시안 분포들을 성분으로 하여 각 성분을 표현하는 파라메터를 집합으로 음성 신호 전체를 모형화하기 때문에 긴 구간의 음성 신호의 표현이 가능하며 음성 신호의 시간적 변화와는 무관한 특성을 가지고 있어 문장 독립적 시스템에 적합하다고 판단하였다.

GMM을 이용한 제안된 시스템은 모델 파라메터의 학습을 위해 성분 혼합 차수 M과 최소 분산 σ^2_{min}의 값을 선정해야 했으며 실험을 통하여 M은 256과 512에서, σ^2_{min}은 0.002와 0.005에서 좋은 인식 성능을 보였고, 특징 파라메터별 성능 평가 실험에서는 MFCC와 델타 MFCC, 델타 델타 MFCC 그리고 델타 에너지와 델타 델타 에너지를 결합하였을 때 (M+DM+DDM+DE+DDE) 최대 73.77%의 인식률을 얻었다.

제안된 GMM 기반의 감정 인식 시스템과 기존 시스템을 비교하기 위하여 감정 인식 및 음성 신호 처리에 널리 사용되고 있는 음성 신호의 피치와 에너지의 평균, 표준편차와 최대 값 등의 통계적인 파라메터 사용하는 KNN 분류기를 이용한 시스템과 MFCC와 같은 음소적 특징 파라메터와 VQ를 사용하여 화자 및 문장 독립적 시스템을 사용하여 성능을 평가하였다. 그 결과 운율적 특징을 이용하는 KNN분류기의 성능이 가장 떨어졌으며, 화자 및 문장 독립적 시스템에 적합하지 않다고 판단되었다. VQ를 이용한 시스템의 경우, 화자 및 문장 독립적 시스템에 좋은 적응을 보였으며 최대 68.12%의 인식 성능을 보였다. GMM을 이용한 인식기는 최대 73.77%의 인식 성능을 보였고, 대부분의 MFCC에 관련한 특징 파라메터들에서 71%이상의 인식 성능을 보여 VQ를 이용한 인식기와 비교하여 볼 때 화자 및 문장 독립적 감정인식 시스템에 대해 보다 적합한 인식 시스템으로 적용할 수 있다고 판단되었다.

향후 HMM (Hidden Markov Model)과 같은 모델링 기법들과 GMM을 병행한 형태의 시스템을 감정 인식에 적용하는 연구도 좋은 결과를 얻을 수 있을 것으로 생각된다. 또한 보다 자연스럽고 사실적인 데이터를 취득하여 데이터 베이스를 구축하고 이를 통해 실제 환경에 적용할 수 있는 어플리케이션에 적합한 시스템이나 특징 파라메터에 관한 연구가 필요하다.

감사의 글

본 연구는 정보통신부 정보통신연구진흥원에서 지원하고 있는 정보통신기초연구지원사업 (과제번호: 2002-036-034-3)의 연구결과입니다.

참 고 문 헌

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., 1097-1108, Feb. 1993.
3. Frank Dellaert, Thomas Polzin, Alex Waibel, "Recognizing emotion in speech", in 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, in ATR Workshop on Virtual Communication Environments-Bridges over Art/Kansei and VR Technologies, Kyoto, Japan, April 1998.
6. V. A. Petrushin, "Emotion Recognition Agents in Real World", 2000 AAAI Fall Symposium on socially Intelligent Agents: Human in the Loop, 136-138, Nov. 2000.
7. V. A. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers", Artificial Neu. Net. In Engr (ANNIE '99), 7-10, Nov. 1999.
8. Frank Dellaert, Thomas Polzin, Alex Waibel, Recognizing emotion in speech, in Proceedings of the ICSLP **96**, Philadelphia, USA, Oct. 1996
9. D. Roy and A. Pentland, Automatic spoken affect analysis and classification, in Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, 363-367, Killington, VT, Oct. 1996.
10. Jun Sato, and Shigeo Morishima, Emotion Modeling in Speech Production using Emotion Space, in Proceedings of

the IEEE International Workshop 1996, 472-477, IEEE, Piscataway, NJ, USA., 1996.
11. Feng Yu, Eric Chang, Ying-Qing Xu, Heung-Yeung Shum, "Emotion Detection from Speech to Enrich Multimedia Content", IEEE Pacific-Rim Conference on Multimedia, 550-557 Beijing, Oct. 2001.
12. L. R. Rabiner and B. H. Juang, Fundamentals of speech recognition, Prentice-Hall Inc., 1993.
13. Earl Gose, Richard Johnsonbaugh, and Steve Jost, Pattern Recognition and Image Analysis, Prentice Hall Inc., 1996.
14. Douglas A. Renolds and Ricard C. Rose, "Robust Text-Independent Speaker Indentification Using Gaussian Mixture Speaker Models", IEEE Trans. on Speech and Audio Processing, **3** (1), 72-83, Jan. 1995.
15. 강면구, 김원구, "음성을 이용한 화자 및 문장 독립 감정 인식", 대한전자공학회 하계학술대회 **25** (1), 377-380, 2002년 6월.

저자 약력

• 강 면 구 (Myoun-Goo Kang)

1993년 3월~2000년 2월: 군산대학교 전기공학과 학사
2000년 3월~2003년 8월: 군산대학교 전기전자제어공학과 석사
※주관심분야: 음성 및 디지털 신호처리, 음성 인식, 감성 인식, 음성 통신 등임

• 서 정 태 (Seo, Jeong-Tae)

1985년: 연세대학교 전자공학과 (학사)
1987년: 연세대학교 본 대학원 전자공학과 (공학석사)
1988년~1990년: 삼성전자 정보통신 연구소 주임연구원
1990년~1995년: 연세대학교 본 대학원 전자공학과 (공학박사)
2003년: Virginia Polytechnic Institute and State University Visiting Scholar
1995년~현재: 충주대학교 정보제어공학과 부교수

• 김 원 구 (Weon-Goo Kim)

1983년 3월~1987년 2월: 연세대학교 전자공학과 학사
1987년 9월~1989년 8월: 연세대학교 전자공학과 석사
1989년 9월~1994년 2월: 연세대학교 전자공학과 박사
1994년 9월~현재: 군산대학교 전자정보공학부 부교수
1998년 9월~1999년 9월: Bell Lab, Lucent Technologies(USA) 객원연구원
※주관심분야: 음성 및 디지털 신호처리, 음성 인식, 감성 인식, 음성 통신 등임