한국지능시스템학회 논문지 2008, Vol. 18, No. 5, pp. 586-591

강인한 음성 인식 시스템을 사용한 감정 인식

Emotion Recognition using Robust Speech Recognition System

김원구
Weon-Goo Kim

군산대학교 전자정보공학부

요 약

본 논문은 음성을 사용한 인간의 감정 인식 시스템의 성능을 향상시키기 위하여 감정 변화에 강인한 음성 인식 시스템과 결합된 감정 인식 시스템에 관하여 연구하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 인식 시스템의 성능에 미치는 영향에 관한 연구와 감정 변화의 영향을 적게 받는 음성 인식 시스템을 구현하였다. 감정 인식은 음성 인식의 결과에 따라 입력 문장에 대한 각각의 감정 모델을 비교하여 입력 음성에 대한 최종 감정 인식을 수행한다.

실험 결과에서 강인한 음성 인식 시스템은 음성 파라메터로 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 신호편의 제거 방법으로 CMS를 사용한 HMM 기반의 화자독립 단어 인식기를 사용하였다. 이러한 음성 인식기와 결합된 감정 인식을 수행한 결과 감정 인식기만을 사용한 경우보다 좋은 성능을 나타내었다.

키워드 : 음성 신호, 감정 인식, 음성 인식, 감정 변화, HMM, MFCC

Abstract

This paper studied the emotion recognition system combined with robust speech recognition system in order to improve the performance of emotion recognition system. For this purpose, the effect of emotional variation on the speech recognition system and robust feature parameters of speech recognition system were studied using speech database containing various emotions. Final emotion recognition is processed using the input utterance and its emotional model according to the result of speech recognition.

In the experiment, robust speech recognition system is HMM based speaker independent word recognizer using RASTA mel-cepstral coefficient and its derivatives and cepstral mean subtraction(CMS) as a signal bias removal. Experimental results showed that emotion recognizer combined with speech recognition system showed better performance than emotion recognizer alone.

Key words : speech signal, emotion recognition, speech recognition, emotional variation, HMM, MFCC

1. 서 론

컴퓨터가 인간의 생활에 미치는 영향이 커지면서 인간과 컴퓨터 간의 인터페이스에 대한 비중 또한 높아지고 있다. 특히, 컴퓨터가 인간의 감정을 인지하고, 그에 따른 정서적인 반응을 하는 시스템 개발은 보다 고차원적인 휴먼-컴퓨터 인터페이스 제품을 가능하게 한다. 인간의 감정이 나타나는 신호로는 얼굴, 음성, 제스쳐, 생체 신호 등 다양하다. 특히, 센서가 신체부위에 직접 닫지 않거나, 전화와 같이 음성신호에 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 감정이외의 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[1-14]. 이 연구결과를 바탕으로 다양한 어플리케이션의 개발이 시도되고 있으며, 특히 음향이나 시각 정보를 처리 및 저장하는 기술과 녹음하는 기술의 진보, 착용하는 컴퓨터의 개발, 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[3][8][9].

감정인식은 지금까지 많이 연구되어 온 음성 인식에서 그 시발점을 찾을 수도 있으나, 특징 파라메터 추출 및 패턴 인식 알고리듬 선택에 있어서 차이가 있다. 특징 선택에 있어서 음성 인식의 경우 음소를 모델링하는 요소를 주로 이용하는 반면, 감정 인식에 있어서는 피치, 에너지, 발음속도 등과 같은 운율적 요소를 활용하여 모델링한다. 패턴 매칭 알고리듬에서도 음성 인식에서는 HMM(Hidden Markov Model) 기법이 가장 우수한 방법으로 알려져 있는데 반해, 감정 인식을 위한 패턴 매칭 기법은 다양한 방법

접수일자 : 2008년 2월 20일
완료일자 : 2008년 3월 10일
이 논문은 정부(교육인적자원부)의 재원으로 한국학술진 재단의 지원을 받아 수행된 연구임(R05-2004-000-12491-0)

이 시도되고 있다. 이러한 시도로 기존적인 패턴 인식 기법을 이외에 Oh-Wook Kwon 등은 SVM(Support Vector Machine), LDA(Linear Discriminant Analysis)와 QDA(Quadratic Discriminant Analysis)를 사용하였다[1]. 또한 Thomas S. Huang은 음성과 표정 두 가지를 이용한 감정 인식을 실험하고, 음성과 표정 중 한 가지만을 이용하여 인식하였을 때보다 더 좋은 인식 성능을 나타내었다[14]. Carnegie Mellon 대학의 Thomas Polzin은 기존 논문들이 운율적 정보만을 이용하여 감정 인식 시스템을 구성하였던 것과는 달리, HMM 구조를 변형하여 운율적 정보와 음향학적 정보를 결합시키는 모델을 연구하였다[12].

본 연구에서는 음성을 사용한 인간의 감정 인식 시스템의 성능을 향상시키기 위하여 감정 변화에 강인한 음성 인식 시스템과 결합된 감정 인식 시스템에 관하여 연구하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 인식 시스템의 성능에 미치는 영향에 관한 연구와 감정 변화의 영향을 적게 받는 음성 인식 시스템을 구현하였다. 감정 인식은 음성 인식의 결과에 따라 입력 문장에 대한 각각의 감정 모델을 비교하여 입력 음성에 대한 최종 감정 인식을 수행하였다.

2. 감정 인식 시스템

본 연구에서는 기본 감정 인식 시스템의 성능 향상을 위하여 감정 변화에 강인한 음성 인식 시스템과 결합된 감정 인식 시스템 구조에 관하여 연구하였다. 이러한 감정 인식 시스템의 구조는 그림 1과 같다.

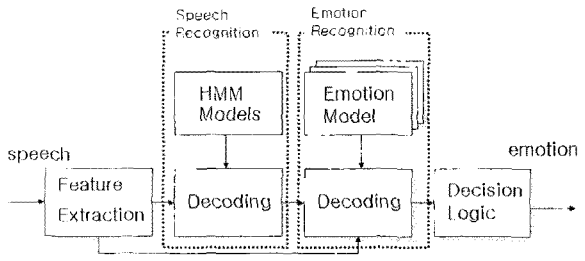

그림 1. 감정 인식 시스템
Fig. 1. emotion recognition system

그림 1에서 감정 인식 시스템은 학습과 인식의 두 단계로 나뉜다. 학습 과정에서에서 음성 인식 시스템은 입력 음성을 인식하기 위한 HMM 기반의 모델을 학습시키고 감정 인식 시스템은 문장 종속의 감정 모델을 학습한다. 인식 단계에서 음성 인식 시스템은 특징 파라메터로 변환되어진 입력 음성을 HMM 기반의 음성 모델을 사용하여 인식한다. 이렇게 음성 인식된 결과는 감정 인식 단계에 전달되어 감정 인식에 사용될 감정 모델을 선택하고 문장 종속 감정 인식을 수행하게 된다.

이러한 구조의 감정 인식 시스템이 우수한 성능을 얻기 위해서는 우선 음성 인식 시스템이 감정 변화에 강인해야 한다. 입력 음성은 감정이 포함되어 있기 때문에 기존 음성 인식 시스템의 인식 성능은 상당히 저하될 수밖에 없다. 이러한 상황에서는 다음 단계의 감정 인식 시스템의 성능이 향상될 수 없는 것이다. 감정 변화에 강인한 음성 인식 시스템을 구현하는 방법은 감정 변화에 강인한 음성 파라메터를 사용하거나 감정 변화를 포함하는 음성 모델을 사용하여 성능을 향상 수 있다.

음성 인식에 널리 사용되고 있는 특징 벡터로는 오래 전부터 사용되어온 LPC 켑스트럼 계수와 멜(mel) 켑스트럼 계수가 주로 사용되고 있으며 잡음에 강인한 특징 벡터로 루트(root) 켑스트럼 계수, PLP(Perceptually Linear Prediction) 계수와 RASTA (RelAtive SpecTrAl) 처리를 한 특징 파라메터 특징 벡터들이 있다[12-18]. 잡음에 강인한 거리 측정 방법으로는 가중 켑스트랄 거리 측정 방법(weighted cepstral distance measure) 방법이 주로 사용되고 있다. 또한 음성에 포함된 편의(bias)를 제거하는 방법으로 켑스트럼 평균 차감법(CMS : Cepstral Mean Subtraction)와 SBR (Signal Bias Removal) 방법을 등이 사용되고 있다.

이러한 파라메터와 잡음 제거 방법들은 음성에 포함된 잡음이나 채널 왜곡 등의 제거하여 음성 인식 시스템의 성능을 향상시킨다. 따라서 이러한 파라메터는 감정 변화에 따라 발생된 음성 변화에도 강인하다. 따라서 이러한 파라메터와 평의 제거 방법을 결합한 음성 인식 시스템을 사용하여 감정 변화에 강인한 음성 인식 시스템을 구현한다.

감정 인식 시스템은 음성 인식 결과를 이용하기 위하여 문장 종속 형태를 사용하였다. 즉 다양한 감정이 포함된 학습 데이터를 사용하여 각 문장별 감정 모델을 구현하였다. 이러한 것은 문장 종속 형태의 감정 인식 시스템이 문장 독립 형태의 것보다 우수한 성능을 나타내기 때문이다.

2.1 음성 파라메터

멜(mel)을 기반으로 한 켑스트럼은 DFT 또는 FFT 크기를 멜과 주파수 사이의 대응 관계에 따라 주파수 축에서 와핑(warping)하여 이의 대수 값을 역 DCT하여 8에서 14차 정도의 계수를 구한다. 예를 들어, 로그 에너지 출력을 X_k라 하면 M개의 멜 켑스트럼 계수는 다음과 같이 나타내어진다.

$$c_n = \frac{1}{20}\sum_{k=1}^{20} X_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{20}\right], n=1,\cdots,M. \quad (1)$$

델타 켑스트럼 계수 $d_k(t)$는 t 번째 구간의 k번째 켑스트럼 계수를 $c_k(t)$라 할 때, 다음과 같이 나타낼 수 있다. 여기서, δ는 시간 간격을 나타낸다.

$$d_k(t) = c_k(t+\delta) - c_k(t-\delta) \quad (2)$$

Lockwood 등은 멜 기반의 루트 켑스트럼 계수가 잡음에 의한 변형에 강인한 것을 관찰하였고 루트 함수로 일반적인 로가리듬 역컨볼루션을 근사화하였다[17].

PLP 분석 방법은 Hermansky에 의해 제안되었으며, 음성 신호의 파워 스펙트럼을 변화시켜 청각 특성이 고려된 스펙트럼을 이용한다. 이러한 단계를 거쳐 얻어지는 저차의 스펙트럼은 인간이 실제 감지하는 소리와 유사한 특성을 갖게 되며, 음성인식에 적용되어 좋은 성능을 보여주었다[16].

RASTA 분석 방법에서는 단구간 스펙트럼을 사용하는 대신 스펙트럼 성분 중 시간에 따라 천천히 변화하는 성분을 배제하는 대역 통과 스펙트럼(band-pass filtered spectrum)을 사용한다[15,16]. 이러한 대력 통과 필터는 각 주파수 대역을 IIR 필터를 사용하여 대역 통과 필터링하는 것과 같다. 이 대역 통과 필터의 전달 함수는 다음과 같다.

$$H(z) = 0.1 \times \frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}(1-0.98z^{-1})} \quad (3)$$

2.2 신호 편의 제거

음성에 포함된 채널 왜곡 특성이 음성 신호의 관찰구간에 대해서 일정하고 그 구간이 충분히 길다면, 왜곡 켑스트럼의 추정치는 관찰된 신호의 켑스트럼의 평균으로 구해질 수 있다. 이와 같이 긴 시간 동안의 켑스트럼의 평균을 빼줌으로써 채널왜곡의 영향을 제거하는 방식을 CMS 이라고 부르며, 다음 수식으로 표현될 수 있다. 여기서 m_y 는 음성의 모든 프레임에서 켑스트럼의 평균이고, $N(s)$는 입력음성의 전체 프레임 수이며, C^t_{comp}는 t 번째 프레임에서 CMS를 통해 보상된 켑스트럼을 의미한다.

$$C^t_{comp} = c^t_y - m_y, \text{ where } m_y = \frac{1}{N(s)}\sum_{t=1}^{N(s)} c^t_y \quad (4)$$

음성에 포함된 편의를 제거하기 위한 또 다른 방법으로 SBR 방법이 있다. 이 방법은 ML(Maximum Likelihood) 추정에 의해 유사도를 최대화하는 방법을 이용한다[18]. 현재 추정된 바이어스를 b 라고 하면, b를 이용하여 보상된 신호 $\tilde{x}_t$ 는

$$\tilde{x}_t = y_t - b \quad (5)$$

이고 보상된 신호에 대한 가장 가까운 모델과 추정된 편의는 다음과 같다.

$$z_t = \mu_i = \arg\max_j p(y_t|b,\lambda_j) = \arg\max_j p(\tilde{x}_t|\lambda_j) \quad (6)$$

$$\bar{b} = \frac{1}{T}\sum_{t=1}^{T}(y_t - z_t) \quad (7)$$

위의 방법을 이용하여 반복적으로 편의를 구하면 편의는 어떠한 값에 수렴하게 된다.

2.3 GMM 기반 감정 인식 시스템

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 다음과 같이 표현된다[8].

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (8)$$

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 9와 같이 표현된다.

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M. \quad (9)$$

GMM을 이용한 인식 시스템은 학습 과정에서 감정별 학습 데이터마다 ML(Maximum Likelihood)추정과 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.

3. 실험 및 결과 고찰

3.1 데이터베이스

감정 인식 시스템의 성능을 평가하기 위해서는 다양한 감정이 포함된 음성 데이터베이스가 필요하다. 이러한 데이터베이스는 다음과 같은 과정으로 구성되었다.

데이터베이스를 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 선별되었다. 녹음작업은 조용한 사무실 환경에서 이루어졌고, DAT를 이용하여 녹음되었다. 각 화자는 45개의 문장을 4가지 감정으로 녹음하였고 녹음 동안에 감정 표현이 미흡하다고 판단된 경우에는 다시 녹음을 하였다. 본 연구를 위하여 사용된 데이터의 규모는 5400(30명×4감정×45문장×1회)문장이다.

3.2 특징 파라메터 추출

음성 신호의 특징 파라메터 추출 과정은 다음과 같다. 전처리를 통하여 16kHz, 16비트로 샘플링하고, 고주파 성분을 보강한다. 이렇게 샘플링된 신호는 음성 구간과 묵음 구간을 구별하기 위하여 음성 구간 검출을 수행하고 특징 벡터를 구한다. 검출된 음성 신호는 20ms(320샘플)의 길이를 갖는 해밍 창(Hamming window)을 사용하여 10ms씩 이동하면서 특징 파라메터를 구한다. 본 연구에서는 음성의 특징 파라메터로 LPC 켑스트럼 계수, 멜 켑스트럼 계수, 루트 켑스트럼 계수, PLP 계수와 RASTA 처리를 한 멜 켑스트럼 계수와 음성의 에너지를 사용하였다. 또한 특징 파라메터의 시간적인 변화에 대한 정보를 포함하는 델타 켑스트럼과 델타 에너지를 사용하였다. 실험에 사용된 켑스트럼 계수는 12차를 사용하였고 PLP 계수는 5차를 사용하였다. 또한 음성에 포함된 편의를 제거하는 방법으로 CMS와 SBR 방법을 사용하였다.

3.3 음성 인식 시스템의 구성

본 연구에서는 우선 감정 변화에 강인한 음성 인식 시스템 개발을 위하여 우선 반연속 HMM을 기본으로 하는 화자 독립 단독음 인식 시스템을 구현하였다. 음성 신호는 샘플링되어 고주파 성분이 보강된 후 음성구간 검출을 수행된다. 검출된 음성 신호를 사용하여 음성 파라메터를 구하고 음성에 포함된 편의를 제거하기 위한 편의 제거 방법을 사용하였다.

반연속 HMM 모델은 256개의 코드을 갖는 코드북을 사용하였고 반연속 HMM은 상태 당 4개의 가우시간 결합 분포를 사용하였다. 또한 각 모델의 상태 수는 학습에 사용된 문장의 평균길이에 비례하게 할당하였다. 모델의 학습에는 20명(남성 10명과 여성 10명)의 음성이 사용되었고 인식에는 학습에 참여하지 않은 10명(남성 5명과 여성5명)을 사용하였다.

입력 특징 파라메터는 다양한 거리 측정 방법과 반연속 HMM을 사용하여 기준 패턴과 유사도를 측정한다. 이때 기준 패턴은 각 문장마다 4가지 감정이 모두 포함된 하나의 HMM 모델을 사용하는 경우와 각 문장마다 각각의 감정으로 학습된 4개의 모델을 사용하는 경우로 구분하였다. 스펙트럼간의 비교 또는 매칭 방법으로 가중 켑스트럼에 의한 거리 측정 방법이 사용되었고 결정 법칙은 비교된 결과를 각 단어당 기준 패턴 수를 고려하여 최종 인식을 결정하는 단계로서 최대 확률을 갖는 기준 패턴을 입력 음성의 단어로 결정한다.

3.4 실험 결과

본 실험에서는 우선 감정이 포함되지 않은 음성으로 학습한 인식 시스템을 대상으로 테스트 음성에 4가지 감정이 포함된 음성을 사용하여 각각의 감정 변화에 따른 시스템의 성능 변화를 관찰하였다. 그림 2는 각 음성 파라메터와 감정별 인식 성능을 나타낸다. 여기서 음성 인식 시스템은 평상의 감정만 포함된 데이터로 학습되었기 때문에 인식 데이터가 평상인 경우에 가장 성능이 우수하고 감정이 포함되면 인식 성능이 급격히 저하된다. 그림은 4가지 감정에 대한 평균 인식률을 나타낸다. 실험에 사용된 4가지(멜 켑스트럼, 루트 켑스트럼, RASTA 멜 켑스트럼, PLP 계수)의 음성 파라메터 중에서는 RASTA 멜 켑스트럼이 89.6%로 가장 우수한 성능을 나타내었다. 이러한 것은 RASTA 처리 과정이 음성의 감정 변화에 따른 스펙트럼의 변화를 보상해 주는 효과가 있다고 볼 수 있다. 또한 음성 파라메터로 델타 켑스트럼을 추가하여 사용했을 때의 성능 평가 실험을 수행하였다. 여기서 멜 켑스트럼의 경우에는 델타 켑스트럼과 결합하여 사용한 경우에 평균 인식률이 1.5%정도 감소하였으나 RASTA 멜 켑스트럼과 멜 켑스트럼을 결합하여 사용한 경우에는 인식률이 91.4%로 약 1.8% 정도 성능이 향상되었다.

MEL : 멜 켑스트럼 계수,
ROOT_MEL : 루트 켑스트럼 계수
RASTA_MEL : RASTA 처리를 한 멜 켑스트럼 계수
PLP : PLP 계수
DMEL : 델타 켑스트럼 계수
SBR : 신호 편의 제거 방법(SBR)
CMS : 신호 편의 제거 방법(CMS)

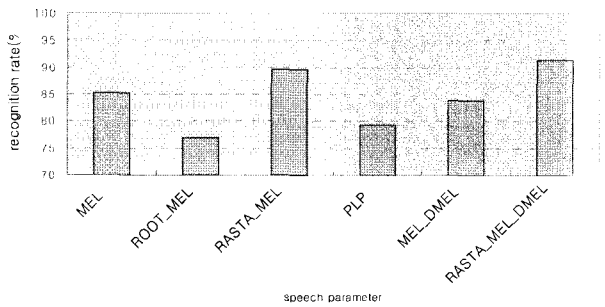

그림 2. 음성 파라메터에 따른 음성 인식 시스템 성능 평가
Fig. 2. Performance of speech recognition system according to the speech parameters

다음은 신호 편의 제거 방법에 따른 인식 성능 평가를 수행하였다. 편의 제거 방법으로는 ML 방법을 사용한 SBR과 CMS 방법을 사용하였다. 여기에서도 음성 인식 시스템은 감정이 포함되지 않은 음성(평상)으로 학습되었다. 그림 3에서 알 수 있듯이 편의 제거를 수행하면 인식 성능이 향상되는 것을 알 수 있다. 특히 CMS가 SBR에 비하여 우수한 성능을 나타내어서 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 신호편의 제거 방법으로 CMS를 사용한 경우에 94.0%로 가장 우수한 성능을 나타내었다. 이러한 것은 감정의 변화에 따라 음성에 편의가 발생한다는 것을 의미하고 편의 제거 과정을 통하여 이러한 변화를 정도 보상해 주는 효과가 있다고 볼 수 있다. 이러한 것은 멜 켑스트럼을 사용한 경우의 인식 성능 85.3%를 기준 시스템으로 할 때 8.7%의 인식률 향상을 나타내고 오차의 감소율로 계산하면 약 59%정도 오차가 감소된다고 볼 수 있다.

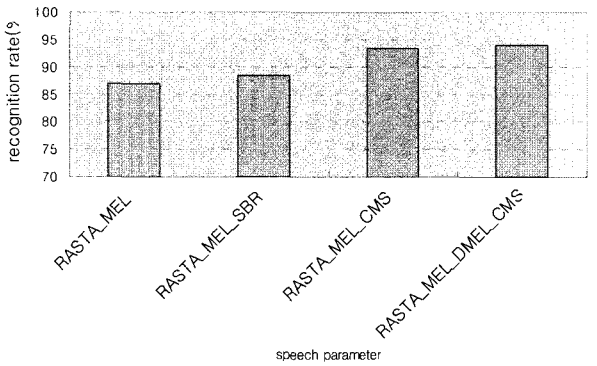

그림 3. 신호 편의 제거 방법에 따른 음성 인식 시스템 성능 평가
Fig. 3. Performance of speech recognition system according to the signal bias removal methods

마지막으로 감정 변화에 강인한 음성인식 시스템과 감정 인식 시스템을 결합한 감정 인식 실험을 수행하였다. 감정 인식은 음성 인식의 결과에 따라 입력 문장에 대한 각각의 감정 모델을 비교하여 입력 음성에 대한 최종 감정 인식을 수행하였다.. 따라서 감정 모델은 각 단어 또는 문장에 대하여 4가지 감정별로 학습이 되어있는 모델을 가지고 인식을 다시 수행하여 최고의 확률 값을 갖는 모델의 감정을 입력 음성의 감정으로 결정하였다. 그림 4는 실험 결과를 나타낸다. 이때 감정 인식만 사용한 시스템(A)은 음성 파라메터로 멜 켑스트럼, 델타 멜 켑스트럼, 델타 델타 멜 켑스트럼, 델타 에너지, 델타 델타 에너지를 사용하고 GMM 기반의 감정 모델을 사용했을 때 가장 우수한 성능으로 73.8%를 나타내었다. 한편 특징 파라메터로 RASTA 멜 켑스트럼, 델타 켑스트럼과 편의 제거 방법으로 CMS를 사용한 감정 변화에 강인한 시스템과 결합한 감정 인식 시스템의 경우에는 1.5% 정도의 감정 인식 성능 향상을 나타내었다.

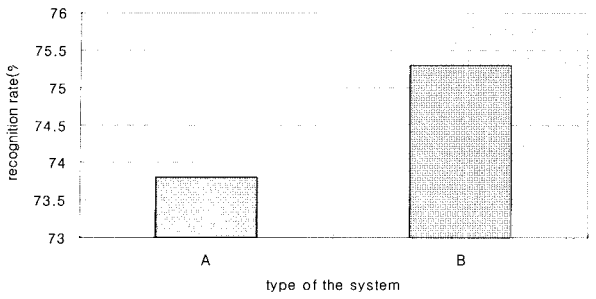

그림 4. 제안된 시스템의 성능평가 (A: 감정 인식만 사용, B: 음성인식과 결합한 감정 인식 사용)

Fig. 4. Performance of the proposed emotion recognition system(A: emotion recognition only, B: emotion recognition combined with speech recognition)

4. 결 론

본 연구에서는 음성을 사용한 인간의 감정 인식 시스템의 성능을 향상시키기 위하여 감정 변화에 강인한 음성 인식 시스템과 결합된 감정 인식 시스템에 관하여 연구하였다. 이를 위하여 우선 다양한 감정이 포함된 음성 데이터베이스를 사용하여 감정 변화가 음성 인식 시스템의 성능에 미치는 영향에 관한 연구와 감정 변화의 영향을 적게 받는 음성 인식 시스템을 구현하였다. 감정 인식은 음성 인식의 결과에 따라 입력 문장에 대한 각각의 감정 모델을 비교하여 입력 음성에 대한 최종 감정 인식을 수행하였다. 실험 결과에서 감정 변화에 강인한 음성 파라메터로는 RASTA 멜 켑스트럼과 델타 켑스트럼을 사용하고 편의 제거 방법으로 CMS를 사용한 경우에 가장 우수한 성능을 보였다. 또한 감정 변화에 강인한 음성인식 시스템과 감정 인식 시스템을 결합한 경우에 1.5%정도의 감정 인식 오차를 감소시킬 수 있었다.

참 고 문 헌

[1] Oh-Wook Kwon, etc "Emotion Recognition by Speech Signal", *Proceedings of Eurospeech '2003*, Vol. 1, pp. 125-128, Geneva, 2003

[2] K. R. Scherer, "Adding the Affective dimension : A New Look in Speech Analysis and Synthesis", *Proceedings of ICSLP*, 2002

[3] Noam Amir,"Classifying Emotions in Speech: a Comparison of Methods", *Proceedings of Eurospeech '2001*, Vol. 1, pp. 127-130, Aalborg, Denmark, 2001

[4] A. Nogueiras, etc,"Speech Emotion Recognition using Hidden Markov Models", *Proceedings of Eurospeech '2001*, Vol. 4, pp. 2679-2682, Aalborg, Denmark, 2001

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

[6] Janet E. Cahn, "The Generation of Affect in Synthesized Speech", *Journal of the American Voice I/O Society*, Vol. 8, pp. 1-19, July 1990.

[7] 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.

[8] Iain R. Murray and John L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A review of the literature on human vocal emotion", *Journal of Accoustal Society of America*, pp. 1097-1108, Feb. 1993.

[9] 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.

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

[11] Rainer Banse and Klaus R. Scherer, "Acoustic Profiles in Vocal Emotion Expression", *Journal of Personality and Social Psychology*, Vol. 70, No. 3, pp. 614-636, 1996.

[12] Frank Dellaert, Thomas Polzin, Alex Waibel, "Recognizing Emotion in Speech", *Proceedings of the ICSLP 96*, Philadelphia, USA, Oct. 1996

[13] Jun Sato, and Shigeo Morishima, "Emotion Modeling in Speech Production using Emotion Space", *Proceedings of the IEEE International Workshop 1996*, pp. 472-477, IEEE, Piscataway, NJ, USA., 1996.

[14] 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.

[15] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong, "Integrating RASTA-PLP into Speech Recognition", in *Proc. ICASSP*, pp. 421-424, 1994.

[16] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech(RASTA-PLP)", in *Proc. EUROSPEECH*, vol.3, pp. 1367-1370, Sep. 1991.

[17] P.Alexandre, ect. "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments", *Speech Communication*, vol. 12, no. 3, pp. 277-288, 1993.

[18] M. G. Rahim, B. H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", *IEEE Trans. Speech & Audio Processing*, vol. 4, No. 1, pp. 19-30, 1996.

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

저 자 소 개

김원구(Weon-Goo Kim)
1987년 2월 : 연세대 전자공학과 학사
1989년 8월 : 연세대 전자공학과 석사
1994년 2월 : 연세대 전자공학과 박사
1994년 9월 ~ 현재 : 군산대 전자정보공학부 교수
1998년 9월 ~ 1999년 9월 : Bell Lab, Lucent Technologies(USA) 객원연구원

관심분야 : 음성 신호처리, 음성 인식, 감성 인식, 음성 변환, 화자 인식
Phone : 063) 469-4745
Fax : 063) 469-4699
E-mail : wgkim@kunsan.ac.kr