논문 2008-45SP-4-13

성별 구분을 통한 음성 감성인식 성능 향상에 대한 연구

(A Study on The Improvement of Emotion Recognition by Gender Discrimination)

조 윤 호*, 박 규 식*

(Youn-ho Cho and Kyu-sik Park)

요 약

본 논문은 남/여 성별에 기반해 음성을 평상, 기쁨, 슬픔, 화남의 4가지 감성 상태로 분류하는 감성인식 시스템을 구축하였다. 제안된 시스템은 입력 음성으로부터 1차적으로 남/여 성별을 분류하고, 분류된 성별을 기반으로 남/여 각기 최적의 특징벡터 열을 적용하여 감성인식을 수행함으로써 감성인식 성공률을 향상시켰다. 또한 음성인식에서 주로 사용되는 ZCPA(Zero Crossings with Peak Amplitudes)를 감성인식용 특징벡터로 사용하여 성능을 향상시켰으며, 남/여 각각의 특징 벡터 열을 최적화하기 위해 SFS(Sequential Forward Selection) 기법을 사용하였다. 감성 패턴 분류기로는 k-NN과 SVM을 비교하여 실험하였다. 실험결과 제안 시스템은 4가지 감성상태에 대해 약 85.3%의 높은 감성 인식 성공률을 달성할 수 있어 향후 감성을 인식하는 콜센터, 휴머노이드형 로봇이나 유비쿼터스(Ubiquitous) 환경 등 다양한 분야에서 감성인식 정보를 유용하게 사용될 수 있을 것으로 기대된다.

Abstract

In this paper, we constructed a speech emotion recognition system that classifies four emotions - neutral, happy, sad, and anger from speech based on male/female gender discrimination. At first, the proposed system distinguish between male and female from a queried speech, then the system performance can be improved by using separate optimized feature vectors for each gender for the emotion classification. As a emotion feature vector, this paper adopts ZCPA(Zero Crossings with Peak Amplitudes) which is well known for its noise-robustic characteristic from the speech recognition area and the features are optimized using SFS method. For a pattern classification of emotion, k-NN and SVM classifiers are compared experimentally. From the computer simulation results, the proposed system was proven to be highly efficient for speech emotion classification about 85.3% regarding four emotion states. This might promise the use the proposed system in various applications such as call-center, humanoid robots, ubiquitous, and etc.

Keywords: 음성 감성인식, 성별 구분, ZCPA, SVM, SFS

I. 서 론

감성 지능형(Emotional Intelligence) 컴퓨팅은 컴퓨터가 학습과 적응을 통하여 인간의 감성을 처리할 수 있는 감성인지 능력을 갖는 것으로 보다 효율적인 인간-컴퓨터 상호작용 (HCI: Human Computer Interaction)을 가능하게 한다. 음성은 사람들 사이의 의사소통에 있어 의미뿐만 아니라 감성도 전달한다. 음성에 내포되어 있는 감성은 특정 단어를 강조하거나 화자의 심리상태를 나타내어 의사소통을 더욱 자연스럽게 한다.

최근에는 휴머노이드형 로봇의 관심과 더불어 로봇의 감성인식에 대한 관심이 높아지고 있다. 인간의 감성을 인식하는 방법으로는 영상정보를 이용한 방법과 음성 정보를 이용한 방법이 주로 연구되고 있다. 이외에 심장 박동 수, 체온, 혈압, 뇌파 등 다양한 수단을 통하여 감성 정보를 얻을 수 있으나 이중에서도 음성을 이용한 감성 인식 시스템은 마이크로폰을 통한 음성 신호의 입력, 처리 등이 타 매체보다 상대적으로 편리하다는 장점으로 최근에 활발한 연구가 이루어지고 있다.

* 평생회원, 단국대학교 공과대학 컴퓨터학부
(Dept. of Computer Science & Engineering, Dankook University)

※ 본 연구비는 2006학년도 단국대학교 대학연구비의 지원으로 연구되었음.
접수일자: 2008년1월9일, 수정완료일: 2008년4월29일

이러한 음성 감성인식 시스템은 향후 유비쿼터스(Ubiquitous) 환경에서 음성 정보를 이용하여 상대방의 감성 상태를 알고자하는 고객 센터(Call Center), 결혼 정보 회사, 유비쿼터스(Ubiquitous) 환경 등에서 다양한 형태로 서비스될 수 있을 것이다.

일반적으로 음성 감성인식 시스템은 감성 특징벡터 추출과 감성 패턴 인식 2가지 단계로 구성된다. 감성 특징벡터 추출은 음성 신호로부터 감성 상태를 대표할 수 있는 피치(Pitch), 포먼트(Formant), 에너지(Energy), MFCC(Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coefficient)와 같은 특징 정보를 구하는 과정이다. 한편 음성의 감성 상태를 분류하는 패턴인식 알고리즘으로는 k-NN(Nearest Neighbor), HMM(Hidden Markov Model), SVM(Support Vector Machine), NN(Neural Network)[1]등 다양한 방법이 사용되고 있으나 일반적으로 음성 감성인식 시스템의 전반적인 인식 성능은 패턴인식 알고리즘보다는 음성 특징벡터에 더 의존하는 경향이 있다.

Dallaert는 피치 윤곽(pitch contour) 변화를 추출하여 음성 감성상태를 기쁨(happy), 슬픔(sad), 화남(anger), 두려움(fear) 등의 4가지로 분류하였으며 k-NN 패턴 분류기를 사용해서 약 79.5%의 인식률을 달성하였다[2]. Moriyama는 음성 신호의 피치 윤곽과 파워 포락선(power envelop)을 특징벡터로 사용하여 놀람(surprise), 화남, 기쁨, 공포, 슬픔 등의 5가지 음성 감성상태를 분류하였으며 이중에서 놀람, 화남, 슬픔 등의 3가지 감성에서 비교적 높은 인식률을 달성할 수 있음을 보였다[3]. C. Lee는 음성 신호의 음향학적 특징에 의미론적인 언어 특징 정보를 더하여 콜센터 같은 응용 시스템에서 부정적인 음성(negative)과 비-부정적인(non-negative) 음성을 분류할 수 있는 알고리즘을 제안하였다. 실험 결과 언어 특징 정보를 이용해 남성 음성의 경우 약 40.7%, 여성 음성의 경우 약 36.4% 향상시킬 수 있음을 밝히고 있다[4]. Rong은 Random Forest와 함께 최적화된 29차 특징벡터 열을 사용하여 부정적인 음성, 긍정적인 음성 그리고 중립적인 음성의 세 가지 상태에 대해 57.3%, 27.3%, 45.9%의 인식 성공률을 얻었다[5]. 한편, Vogt는 음성을 남성과 여성으로 먼저 분류한 후 인식 성능을 평가한 결과 성별 독립 시스템이 성별 종속 시스템에 비해 약 2~4%의 성능 향상이 있음을 밝히고 있다[6]. Lugger는 음성의 운율적 특징 정보를 이용해 음성을 High(화남, 기쁨, 근심), Low(평상, 지루함, 슬픔) 두 가지로 분류하는 감성인식 실험에서 66.7~74.5%의 인식 성공률을 보이고 있다[7].

이외에도 미국의 Microsoft, HP, 일본의 SONY등의 산업계에서 음성 감성인식 기술을 HCI용 SW나 로봇 등의 응용분야에 적용하기 위한 활발한 연구를 진행하고 있다.

본 연구에서는 음성의 감성 상태를 평상, 기쁨, 슬픔, 화남의 4가지로 구별할 수 있는 감성인식 시스템을 제안하였다. 본 논문의 주요 내용은 다음과 같다.

첫째, 음성을 남/여 성별로 1차적으로 구분한 뒤 남/여 성별에 따라 각기 다른 최적의 특징벡터 열을 사용함으로서 감성인식 성능을 향상 시켰다.

둘째, 음성의 운율적 특징을 설명하는 피치, 에너지, MFCC 외에 음성인식 분야에서 잡음에 강인한 특성을 가진 것으로 알려져 있는 ZCPA(Zero-Crossing with Peak Amplitudes) 특징[8]을 추가로 적용하여, 감성 인식 성능에 있어 ZCPA 특징의 유효성을 검증하였다.

셋째, 감성 특징으로 추출된 특징벡터 중 감성인식률 향상에 기여가 높은 특징계수들만을 선별해서 시스템의 인식 성능을 향상시키고 연산 복잡도를 낮출 수 있는 SFS(Sequential Forward Selection)[9] 기법을 남/여 별도로 적용하였으며, 음성 감성 분류를 위한 패턴 인식 기법으로는 k-NN, SVM 2가지 방법을 비교하였다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 제안된 시스템의 구조에 대해 알아보고, Ⅲ장에서는 감성 특징 추출 및 최적화 방법에 대해 설명하였다. Ⅳ장에서는 컴퓨터 모의실험 환경을 통해 제안된 시스템의 성능을 확인하고, 마지막 결론으로 끝을 맺는다.

Ⅱ. 제안된 음성 감성인식 시스템

그림 1은 본 논문에서 제안한 음성 감성인식 시스템을 나타낸다. 제안 시스템은 크게 1단계 - SFS를 이용하여 남/여 성별로 최적의 감성 특징벡터 추출하는 단계, 2단계 - 입력된 질의 음성으로부터 남/여 성별을 분류하고 분류된 성별을 바탕으로 1단계에서 구축된 성별 최적의 특징벡터 열을 적용하여 감성 상태를 분류하는 2가지 단계로 구성된다.

1단계 SFS를 이용한 남/여 성별 최적 감성 특징벡터 추출에서는 평상, 기쁨, 슬픔, 화남 각 감성별로 남/여 구분된 훈련용 음성 DB의 음성 신호를 32ms 프레임 단위로 분할해서 해밍(Hamming) 윈도우를 적용한 후 끝점 검출(end-point detection) 등의 음성 전처리 과정을 거쳐 다음 3.2 절에 정의된 총 136차의 특징벡터를

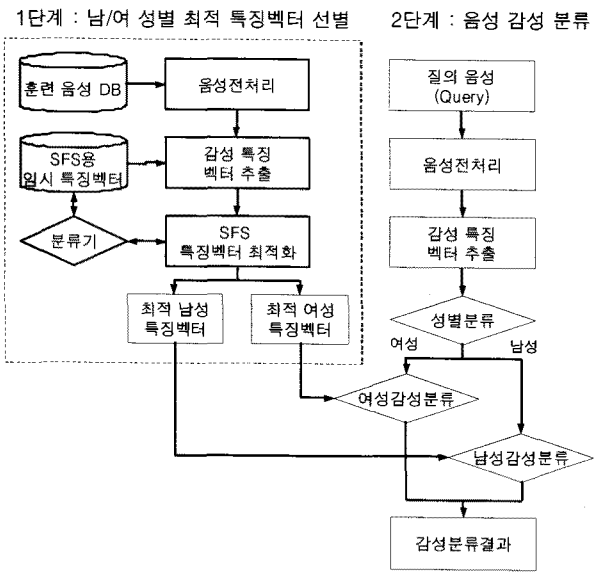

그림 1. 제안된 감성인식 시스템
Fig. 1. Proposed speech emotion recognition system.

추출한다. 추출된 특징벡터는 성별로 분류하여 SFS 특징벡터 최적화 과정(다음 3.3절 참조)을 거쳐 SVM 분류기로 감성인식 성공률에 가장 큰 기여를 하고 있는 특징계수들만을 선별해서 남/여 성별로 최적 감성 특징벡터를 구성하게 된다.

2단계 질의 음성에 대한 감성 분류는 1단계에서와 같은 음성 전처리 과정을 거쳐 특징벡터를 추출한다. 추출된 특징벡터 중에서 남/여 성별을 구분하는데 탁월한 성능을 나타내는 피치와 MFCC를 이용하여 남/여 성별을 구분한다. 구분된 성별에 따라 1단계에서 구축된 최적의 남/여 성별 특징벡터 열을 남/여에 각각 적용하여 k-NN, SVM 패턴 분류기를 이용하여 질의 음성의 감성 상태를 평상, 기쁨, 슬픔 또는 화남의 4가지 감성 상태로 최종 분류한다.

III. 감성 음성 특징벡터 추출 및 최적화

3.1. 음성 전처리

음성 전처리과정은 그림 2와 같이 신뢰성 있는 감성 특징벡터를 추출하기 위한 프레임 단위의 음성 신호 분할, 해밍 윈도우(Hamming window) 그리고 끝점 검출로 구성되어 있다.

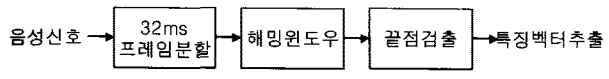

그림 2. 음성 전처리 과정
Fig. 2. Speech pre-processing.

입력되는 음성 신호는 먼저 32ms 프레임 단위로 분할되고 이웃한 프레임과 50% 중복되는 해밍 윈도우를 적용받게 된다. 이후 끝점 검출은 음성 신호에서 음성 구간(voiced period)과 비-음성 구간을 구별하여 음성 구간에서만 특징벡터를 추출하기 위한 것으로 이는 비-음성 구간에서 잘못된 음성 분석과 특징벡터 추출로 인해 야기되는 시스템 성능저하를 예방하기 위한 것이다. 본 연구에서 사용된 끝점 검출기는 TEO와 에너지 엔트로피(entropy)를 이용해 낮은 SNR 환경에서도 비교적 높은 성능을 낼 수 있는 L. Gu의 알고리즘[10]을 이용하였다. TEO는 기존의 에너지 값이 신호의 세기만을 나타내던 것에 비하여 신호의 세기와 주파수적 특성을 동시에 나타낼 수 있는 장점이 있다. TEO는 다음 수식 (1)과 같이 표현할 수 있다.

$$\begin{aligned}\psi[x(n)] &= x^2(n) - x(n+1)(n-1) \\ &= A^2\sin^2(\Omega) \cong A^2\Omega^2\end{aligned} \tag{1}$$

3.2. 감성 특징벡터 추출

감성 특징벡터는 매 프레임 단위로 피치, 에너지, MFCC와 ZCPA 특징을 추출하여 각 특징계수들의 평균(mean)과 표준 편차(standard deviation)를 구하고 피치, 에너지, MFCC에 대해서는 델타(Delta) 값을 구하여 최종적으로 총 136차의 특징벡터를 구성하였다.

피치는 일반적으로 사용되는 SHR(Subharmonic to harmonic ratio) 알고리즘[11]을 사용하였다. SHR은 음성 신호에 FFT를 취하여 2개의 피크 값 (f_1, f_2)을 피치 후보로 선정하고 아래의 수식 (3)에서 SHR 값을 특정 한계 값과 비교하여 SHR이 한계 값보다 작으면 f_2를 최종 피치로 선정하고, 아니면 f_1을 피치로 선정한다. 여기서 DA(·)는 논문 [11]에서 주어진 차분 함수이다.

$$\begin{aligned}SUMA(\log f)_{odd} &= \sum_{n=1}^{N} LOGA(\log f + \log(2n-1)) \\ DA(\log f) &= SUMa(\log f)_{even} - SUMA(\log f)_{odd} \\ SHR &= 0.5\frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)}\end{aligned} \tag{2}$$

음성신호에 있어서 데이터의 개수가 N인 프레임에서 단구간 에너지를 평가하기 위한 식은 다음 수식(3)과 같다.

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \tag{3}$$

여기서, $x(m)$는 m번째 음성데이터이고, n은 전체 음성데이터 수를 나타낸다. 일반적으로 수식(3)은 음성 신호의 크기에 너무 민감하다. 이와 같은 문제점을 해결하기 위해 아래의 수식(4)와 같이 단-구간 음성 에너지는 평균크기함수(average magnitude function)을 적용하여 에너지를 계산한다.

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m) \tag{4}$$

음성 인식 분야 등에서 널리 사용되는 MFCC(Mel Frequency Cepstral Coefficient)는 인간의 청각 시스템 특성과 유사한 멜-주파수(Mel-Frequency) 상에서의 음성 특성을 잘 표현할 수 있으며 총 12차를 추출하였다.

ZCPA 모델은 인간의 청각 특성을 모델링한 EIH(Ensemble Interval Histogram)에 기반한 것으로 그림 3은 ZCPA 모델 블록도를 나타낸다.

ZCPA 모델은 그림 3의 각 대역 통과 필터를 거친 신호의 상향 영교차점(upward zero-crossing point)을 파악하여 인접한 두 교차점간의 시간 간격을 측정하고, 그 역수를 주파수 히스토그램에 누적시킨다. 그리고 인접한 영교차점간 신호의 피크 값을 검출하여 비선형 함수를 통과한 값을 주파수 히스토그램에 가중치로 사용한다. 최종 ZCPA의 출력은 모든 채널의 히스토그램을 더함으로써 얻어지며, n번째 프레임의 ZCPA의 출력은 다음 수식 (5)과 같이 표현할 수 있다[8].

$$ZCPA(m,i) = \sum_{k=1}^{N_{ch}} \sum_{l=1}^{Z_{k-1}} \delta_{ijl} g(P_{kl}), \quad 1 \le i \le N \tag{5}$$

각 채널의 중심 주파수를 Fc라고 할 때, 시간 t에서의 ZCPA 출력을 얻기 위해서는 구간 $[\frac{t-10}{Fc}, \frac{t}{Fc}]$ 만큼

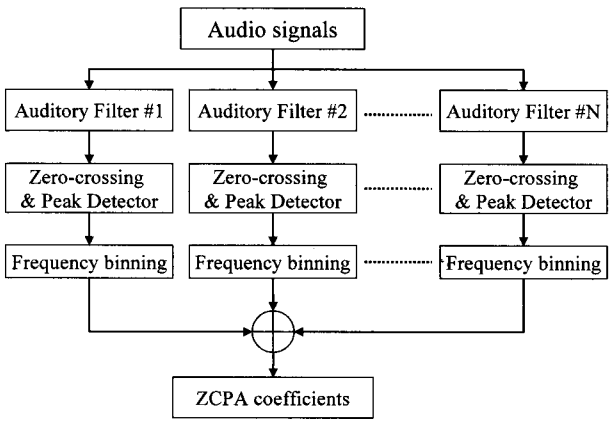

그림 3. ZCPA 모델의 블록도[8]
Fig. 3. Block diagram of ZCPA model[8].

의 신호가 고려되는데 그 결과, 저주파 채널에서는 긴 신호가, 고주파 채널에서는 상대적으로 짧은 신호가 사용되므로 저주파 영역에서는 주파수 분해능(Frequency resolution)이 좋아지는 반면 시간 분해능이 나빠지게 되고, 고주파 영역에서는 이와 반대의 특성이 나타난다. 이는 실제 인간의 청각 특성과 일치하는 특성이다.

위의 여러 가지 특징벡터 추출 과정을 통해 전체 특징 벡터의 구성은 에너지(1), 피치(1), MFCC(12) 값에 대한 평균, 표준편차 그리고, 각각의 델타(Delta) 값으로 56차, ZCPA(40차)에 대한 평균과 표준편차로 80차를 구성하여 총 136차의 특징벡터를 추출한다.

3.3. SFS 특징벡터 최적화

SFS 기법은 총 136차 특징벡터 간에 중복된 상관성(Correlation)을 제거하고 동시에 시스템의 감성 인식률 향상에 기여도가 높은 최적의 특징계수들만을 선정하는 방법으로 시스템의 성능을 높여줌과 동시에 연산 복잡도를 낮출 수 있는 장점이 있다. SFS는 아래와 같은 순서에 의해 먼저 각 특징 계수들을 개별적으로 사용하여 감성 분석 한 후, 가장 좋은 감성 인식률을 나타내는 특징계수부터 순차적으로 하나씩 특징계수를 추가해 나가면서 감성인식 정확도를 계산한다.

1. Start with the empty set $Y_0 = \{\varnothing\}$
2. Select the next best feature $x^+ = \arg max[J(y_k + x)]$
3. Update $Y_{k+1} = Y_k + x^+; k = k+1$
4. Go to 2

본 논문에서는 남/여 성별 차이에 따라 서로 다른 최적의 특징벡터 열을 갖는다는 가정하에 남/여 성별로 가장 높은 감성 인식률을 나타내는 특징계수부터 순차적으로 하나씩 특징계수를 추가해 가면서 감성 인식 정확도를 계산하여 남/여 성별로 최적화된 특징벡터 열을 구성하게 된다.

Ⅳ. 실험 환경 및 결과

4.1. 실험 환경

본 연구에서는 논문 [12]에서 기 구축된 DB를 이용하여 평상, 기쁨, 슬픔, 화남 4가지 감성에 대한 감성 음성 DB를 구축하였다. 구축된 감성 음성 DB는 평소 감성 음성 발성을 훈련하는 아마추어 연극단원 남/여 각 15명이 45개 문장에 대하여 4가지 감성 상태별로 발성한 음성을 총 5,400개 16kHz, 16bit로 녹음한 것이다. 본

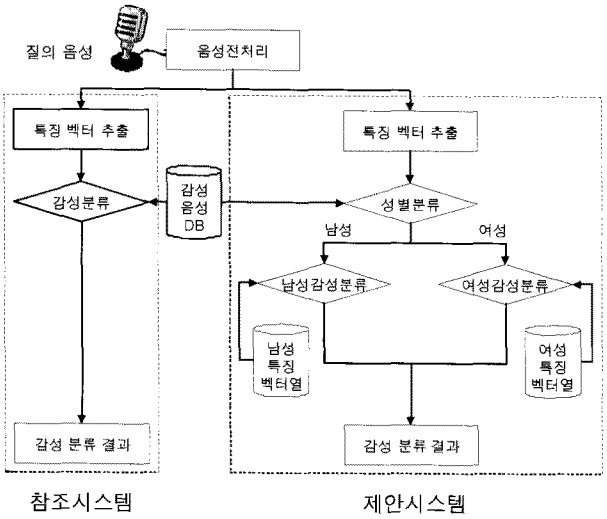

그림 4. 제안 시스템의 성능비교를 위한 실험환경
Fig. 4. Experimental setup for performance comparison of the proposed system.

연구에서는 8kHz, 16bit로 다운샘플링 한 뒤에 사용하였다.

5,400개 음성 중 무작위로 1,600개(남/여 각 800개, 감성별 200개)의 음성을 선택하여 SFS를 이용한 최적의 감성 특징벡터를 선별하는데 사용하였고, 이와 문장이 중복되지 않는 400개(남/여 각 200개, 각 감성별 50개)의 음성을 무작위로 선택하여 질의 데이터로 사용하였다.

그림 4는 본 논문에서 제안한 시스템의 성능비교를 위하여 구축된 실험환경을 보이고 있다.

그림 4에서 참조 시스템은 음성 전처리 과정을 거쳐 136차 감성 특징벡터 전부를 추출하여 감성 패턴 분류기로 음성 감성상태를 인식하는 시스템이다. 본 논문에서 제안한 시스템은 음성 전처리 과정을 거쳐 피치, 에너지, MFCC, ZCPA 등의 136차 특징벡터를 추출하고 1차적으로 피치, MFCC 특징벡터 열을 이용하여 성별을 분류한 다음, 2차적으로 훈련용 음성 DB에서 SFS 기법으로 남/여 성별에 따라 선정된 최적화된 특징벡터 열을 남/여 음성에 각각 적용하여 최종 감성 상태를 분류하여 감성인식 성공률을 향상시킨다.

4.2. 실험 결과

다음의 그림 5는 제안 시스템에서 사용될 남/여 성별로 최적화된 특징벡터 계수를 선정하기 위한 SFS 실험결과를 보이고 있다.

그림 5와 같이 k-NN과 SVM 분류기에서 SFS를 이용한 최적 특징 벡터 열을 추출할 때 SVM 분류기의

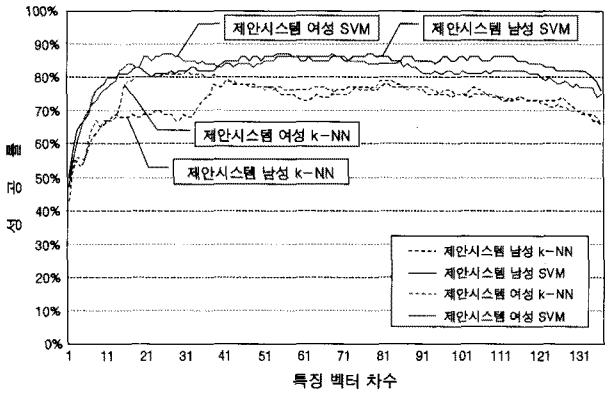

그림 5. SFS 특징벡터 최적화 과정
Fig. 5. SFS feature vector optimization process.

표 1. 제안시스템의 1차 남/여 성별 분류 성공률
Table 1. Male/female gender classification of the proposed system.

분류기 / 질의	k-NN		SVM	
	남	여	남	여
남	196	4	199	1
여	12	188	1	199
성공률	96.0%		99.5%	

성능이 우수하였으며, 여성이 남성에 비해 낮은 차수에서 높은 성공률을 나타냈다. 남/여 성별로 각각 수행된 SVM-SFS 결과 여성은 25차에서 85%의 최대 인식 성공률을 남성은 54차에서 85.5%의 최대 인식 성공률을 각각 나타냈다. 제안 시스템은 그림 5에서 추출된 최적 특징 벡터 열을 남/여 성별로 각각 구성하여 사용한다.

표 1은 제안 시스템에서 1차적으로 수행하게 되는 남/여 성별 분류에 대한 실험 결과이다.

표 1에서와 같이 피치와 MFCC로 구성된 특징 벡터를 이용한 SVM 분류기가 99.5%로 우수한 성별 분류 성능을 나타냄을 알 수 있다. 위 결과에 따라 본 논문의 제안 시스템에서 1차 성별 분류과정에서는 피치와

표 2. 제안 감성 인식 시스템의 성능 비교
Table 2. Performance comparison of proposed emotion recognition system.

	성별	k-NN	SVM	SFS 여부
참조 시스템	-	65.3% (136)	72.0% (136)	×
제안 시스템	남	65.5% (136)	76.5% (136)	×
		79.0% (43)	85.5% (54)	○
	여	65.0% (136)	73.0% (136)	×
		81.5% (31)	85.0% (25)	○

MFCC로 특징 벡터 열을 구성하고 분류기로 SVM을 사용하였다.

표 2는 위의 그림 5에서 최고의 감성 인식 성공률을 나타내는 최적 특징 차수와 해당 인식률 나타내고 있다. 표에서 괄호 안의 숫자는 사용된 특징 벡터의 차수를 말한다.

표 2에서 보듯이 SFS 기법을 사용하지 않고 136차 특징 벡터 전체를 사용한 참조 시스템에서 k-NN 분류기는 65.3%, SVM 분류기는 72%의 인식 성공률을 보이고 있다. 반면, 남/여 성별을 구분한 제안 시스템의 경우는 이 보다 약 1~4% 정도 향상된 결과를 보이고 있다.

한편, SFS기법을 사용하여 최적화된 특징 벡터를 사용한 제안시스템은 참조시스템에 비해 k-NN 분류기에서는 약 14~16%, SVM 분류기에서는 약 14~15% 정도의 성능 향상을 보이고 있다.

그림 6은 본 논문에서 사용한 ZCPA 특징벡터가 전체 시스템 성능향상에 미치는 영향을 비교하고 있다. 본 논문에서 사용된 ZCPA 특징은 80차로, ZCPA를 포함한 전체 136차 특징벡터와 ZPCA를 제외한 56차 특징벡터의 인식 결과를 비교하고 있다.

그림 6에서 보듯이 ZCPA 특징이 전반적으로 감성 인식률을 약2~3% 향상시키는 것을 확인할 수 있다. 아래의 표 3에서와 같이 남/여 성별 최적의 특징벡터 열에 ZCPA 특징벡터가 다수 포함되어 있으며 특히 남성의 감성을 분류하는데 크게 기여하고 있음을 알 수 있다.

표 3은 그림 5에서 남/녀 성별 SFS 기법으로 선정된 최적화된 25차 특징벡터를 비교하고 있다(단, 남성의 경우 총 54차 중 처음 25차만을 나열하였음).

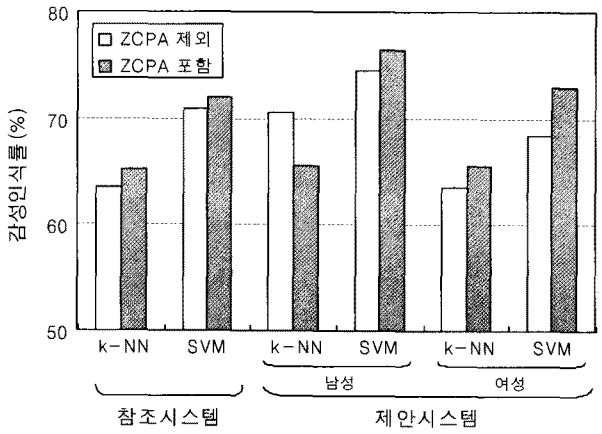

그림 6. ZCPA 성능비교
Fig. 6. Performance comparison of ZPCA feature.

표 3. 남/여 성별 SFS 최적 특징벡터 열
Table 3. SFS Optimized feature vector of male/female.

No.	남성 최적 특징벡터	여성 최적 특징벡터
1	std.ZCPA4	mean.ZCPA2
2	mean.Pitch	mean.MFCC9
3	std.MFCC12	mean.MFCC7
4	mean.MFCC9	mean.ZCPA24
5	std.ZCPA3	mean.MFCC1
6	mean.MFCC10	mean.MFCC2
7	std.Pitch	mean.DeltaMFCC6
8	mean.MFCC6	mean.Pitch
9	mean.MFCC2	mean.ZCPA30
10	std.ZCPA20	mean.ZCPA17
11	std.DeltaMFCC11	std.DeltaMFCC1
12	mean.DeltaEnergy	std.ZCPA22
13	mean.DeltaMFCC4	std.MFCC7
14	mean.DeltaPitch	mean.MFCC11
15	mean.ZCPA16	mean.MFCC4
16	mean.ZCPA18	std.DeltaMFCC5
17	mean.ZCPA33	mean.DeltaEnergy
18	mean.DeltaMFCC3	mean.DMFCC7
19	mean.ZCPA17	mean.DMFCC1
20	mean.ZCPA20	std.ZCPA3
21	mean.MFCC1	std.ZCPA15
22	std.DeltaMFCC5	std.MFCC9
23	mean.MFCC4	mean.Energy
24	mean.MFCC7	std.ZCPA35
25	std.DeltaMFCC6	std.DeltaEnergy

표에서 보듯이 남성의 경우에는 피치, 에너지 및 ZCPA 특징벡터가 중요한 역할을 하고 있다. 특히 ZCPA 특징은 저주파 영역에서 주파수 분해능(Frequency resolution)이 좋아지는 특성으로 인해 고주파 성분이 많은 여성에 비해 남성음성에서 중요한 역할을 하고 있으며, 여성의 경우는 MFCC가 가장 중요한 역할을 하고 있다.

본 논문의 제안시스템과 같이 1차적으로 남/여 성별을 분류한 뒤 각 성별로 최적화된 특징벡터 열을 적용하여 감성 분류를 수행하는 것이 성공률 향상에 크게 기여함을 알 수 있다.

표 4는 위 표 2의 각 감성 분류 결과를 구체적으로 나타낸 것으로 400(각 감성별 100개)개 질의 음성에 대한 결과를 나타낸다. 표에서 제안 시스템은 남(총 54차)/여(총 25차)별 최적화된 특징벡터 열과 SVM 분류기에 대한 평균 결과를 참조 시스템은 136차 특징벡터가 적용된 SVM 결과를 나타낸다.

표 4에서 보듯이 참조 시스템에서는 평상을 슬픔으로, 화남을 기쁨으로 오분류하는 경우가 많아 72.0%의 감성인식 성공률을 나타냈다. 하지만 남/여 성별을 1차로 분류한 뒤 남/여 각 성별에 따라 다른 최적의 특징

표 4. SVM을 이용한 감성분류 결과 비교
Table 4. Comparison of emotion classification using SVM.

시스템 / 질의	참조시스템 (전체:136차)				제안시스템 (남:54 / 여:25차)			
	평상	기쁨	슬픔	화남	평상	기쁨	슬픔	화남
평상	66	7	21	6	87	5	6	2
기쁨	3	77	6	14	4	77	8	11
슬픔	12	6	82	0	1	5	92	2
화남	6	27	4	63	2	12	1	85
인식률	72%				85.3%			

벡터 열을 적용하여 2차 감성 분류를 하는 제안 시스템에서는 참조 시스템에서 보였던 평상을 슬픔으로 오분류하는 경우와 화남을 기쁨으로 오분류하는 경우가 많이 줄어들어 85.3%의 감성인식 성공률을 나타내어 참조 시스템에 비해 약 13% 이상 감성인식 성공률이 크게 향상 되었다.

V. 결 론

본 논문은 음성으로부터 평상, 기쁨, 슬픔, 화남의 4가지 감성 상태를 인식할 수 있는 음성 감성인식 시스템을 제안하였다. 제안 시스템은 1차적으로 남/여 성별을 구분한 뒤 각 성별에 따라 최적화된 특징벡터 열을 이용하여 2차적으로 음성 감성 인식을 수행함으로써 약 85.3%의 높은 인식률을 달성할 수 있었다. 향후 제안시스템은 고객센터(Call center), 유비쿼터스(Ubiquitous) 환경 등에서 감성 상태에 따라 다른 맞춤형 서비스를 제공하는데 유용하게 사용될 것으로 기대된다.

참 고 문 헌

[1] Duda R., Hart P. and Stork D., "Pattern Classification, Second Edition," John Wiley & Sons, 2000.

[2] Dellaert, F., Polzin, T., and Waibel, A., "Recognizing Emotion in Speech," *Proceedings of the International Conference on Spoken Language*, Vol. 3, pp. 1970-1973, Oct. 1996.

[3] Moriyama T. and Oazwa, "Emotion Recognition and Synthesis System on Speech," *IEEE International Conference on Multimedia Computing and Systems*, Vol. 1, pp.840-844, Florence, Italy, 1999.

[4] Lee C. M. and Narayanan S. S., "Towards Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 2, March 2005.

[5] Rong J., Chen Y., Chowdhury M. and Li G., "Acoustic Features Extraction for Emotion Recognition," *Computer and Information Science, ICIS 2007. 6th IEEE/ACIS International Conference*, pp.419-424, July 2007.

[6] Vogt T and Andre E., "Improving Automatic Emotion Recognition from Speech via Gender Differentiation," *Proceedings of Language Resources and Evaluation Conference '06*, Italy, May 2006.

[7] Lugger M. and Yang B., "The relevance of voice quality features in speaker independent emotion recognition," *Acoustics, Speech and Signal Processing, ICASSP 07.*, Vol. 4, pp.IV-17~IV-20, April 2007.

[8] Doh-Suk Kim, Soo-Young Lee and Rhee M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 1, pp.55-69, Jan. 1999.

[9] Jain A. and Zongker D., "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE transactions on pattern analysis and machine intelligence*, Vol. 19, No. 2, pp.153-158. 1997.

[10] Lingyun Gu and Stephen A. Zahorian, "A New Robust Algorithm for Isolated Word Endpoint Detection," *Acoustics, Speech and Signal Processing, ICASSP 02.*, Orlando, FL, May 2002.

[11] Xuejing Sun, "A Pitch Determination Algorithm Based On Subharmonic-to-Harmonic Ratio," *International Conference on Spoken Language Processing '2000*, pp.676-679, 2000.

[12] 강봉석, "음성 신호를 이용한 문장독립 감정 인식 시스템," 석사학위 논문, 연세대학교, 2001.

저 자 소 개

조 윤 호(평생회원)

1994년 단국대학교 농학과 학사졸업.

2004년 단국대학교 멀티미디어과 석사졸업.

2006년 8월 단국대학교 정보 컴퓨터과학과 박사수료

2007년~현재 단국대학교 공과대학 컴퓨터학부 강의전임강사

<주관심분야 : 음성 및 음향 신호처리, 감성인식, 멀티미디어 신호처리>

박 규 식(평생회원)

1986년 Polytechnic University 전자공학과 학사 졸업.

1988년 Polytechnic University 전자공학과 석사 졸업.

1993년 Polytechnic University 전자공학과 학사 졸업.

1994년~1996년 삼성전자 마이크로사업부, 선임연구원

1996년~2001년 상명대학교 컴퓨터·정보통신 공학부 조교수

2001년~현재 단국대학교 공과대학 컴퓨터학부 교수

<주관심분야 : 음성 및 음향 신호처리, 멀티미디어 신호처리, DSP 시스템 구현, 감성인식>