한국음향학회지 제25권 제6호 pp. 269~276 (2006)

이동통신 환경에서 강인한 음성 감성특징 추출에 대한 연구

A Study on Robust Speech Emotion Feature Extraction Under the Mobile Communication Environment

조 윤 호*, 박 규 식*
(Youn-Ho Cho*, Kyu-Sik Park*)

*단국대학교 정보 · 컴퓨터과학과

(접수일자: 2006년 6월 7일; 수정일자: 2006년 7월 25일; 채택일자: 2006년 8월 2일)

본 논문은 이동전화 (Cellular phone)를 통해 실시간으로 습득된 음성으로부터 사람의 감성 상태를 평상 혹은 화남으로 인식할 수 있는 음성 감성인식 시스템을 제안하였다. 일반적으로 이동전화를 통해 수신된 음성은 화자의 환경 잡음과 네트워크 잡음을 포함하고 있어 음성 신호의 감성특징을 왜곡하게 되고 이로 인해 인식 시스템에 심각한 성능저하를 초래하게 된다. 본 논문에서는 이러한 잡음 영향을 최소화하기 위해 비교적 단순한 구조와 적은 연산량을 가진 MA (Moving Average) 필터를 감성 특징벡터에 적용해서 잡음에 의한 시스템 성능저하를 최소화하였다. 또한 특징벡터를 최적화할 수 있는 SFS (Sequential Forward Selection) 기법을 사용해서 제안 감성인식 시스템의 성능을 한층 더 안정화시켰으며 감성 패턴 분류기로는 k-NN과 SVM을 비교하였다. 실험 결과 제안 시스템은 이동통신 잡음 환경에서 약 86.5%의 높은 인식률을 달성할 수 있어 향후 고객 센터 (Call-center) 등에 유용하게 사용될 수 있을 것으로 기대된다.

핵심용어: 이동통신, 음성 감성인식, MA 필터링, SFS, 콜 센터

투고분야: 음성신호처리분야 (2.5)

In this paper, we propose an emotion recognition system that can discriminate human emotional state into neutral or anger from the speech captured by a cellular-phone in real time. In general, the speech through the mobile network contains environment noise and network noise, thus it can causes serious system performance degradation due to the distortion in emotional features of the query speech. In order to minimize the effect of these noise and so improve the system performance, we adopt a simple MA (Moving Average) filter which has relatively simple structure and low computational complexity, to alleviate the distortion in the emotional feature vector. Then a SFS (Sequential Forward Selection) feature optimization method is implemented to further improve and stabilize the system performance. Two pattern recognition method such as k-NN and SVM is compared for emotional state classification. The experimental results indicate that the proposed method provides very stable and successful emotional classification performance such as 86.5%, so that it will be very useful in application areas such as customer call-center.

Key words: Mobile communication, Speech emotion recognition, MA filtering, SFS, Call center

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

감성지능 (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. 인간의 감성 정보를 얻는 방법은 얼굴표정, 음성, 몸동작, 심장 박동 수, 체온, 혈압, 뇌파 등 다양한 수단을 통하여 얻을 수 있으나 이중에서도 음성을 이용한 감성 인식 시스템은 마이크로폰을 통한 음성 신호의 입력, 처리 등이 타 매체보다 상대적으로 편리하다

책임저자: 박 규 식 (kspark@dankook.ac.kr)
140-714 서울시 용산구 한남로 147번지 단국대학교
정보 · 컴퓨터학부 컴퓨터과학 전공
(전화: 02-709-2728; 팩스: 02-799-1343)

는 장점으로 최근에 활발한 연구가 이루어지고 있다. 이러한 음성 감성인식 시스템은 향후 유비쿼터스 (Ubiquitous) 환경에서 음성 정보를 이용하여 상대방의 감성 상태를 알고자하는 고객 센터 (Call Center), 결혼정보 회사, 모바일 콘텐츠 산업 등에서 다양한 형태로 서비스될 수 있을 것이다.

일반적으로 음성 감성인식 시스템은 감성 특징벡터 추출과 감성 패턴 인식 2가지 단계로 구성된다. 감성 특징벡터 추출은 단구간 (short time) 음성 신호로부터 음성신호의 감성 상태를 대표할 수 있는 피치 (pitch), 포먼트 (formant), 에너지 (energy) 그리고 MFCC (Mel Frequency Cepstral Coefficient), LPC (Linear Predictive Coefficient) 같은 스펙트럼 정보 등을 구하는 과정이다. 한편 음성의 감성 상태를 분류하는 패턴인식 알고리즘으로는 k-NN (Nearest Neighbor), HMM (Hidden Markov Model), SVM (Support Vector Machine), NN (Neural Network)등 다양한 방법 [1]이 사용되고 있으나 일반적으로 음성 감성인식 시스템의 전체적인 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), 공포 (fear) 등의 4가지로 분류하였으며 k-NN 패턴 분류기를 사용해서 약 79.5%의 인식률을 달성하였다 [2]. Moriyama [3]는 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. 한편 A. Nogueiras는 논문 [4]에서 HMM을 이용한 화자종속 방식의 감성인식 시스템을 제안하였고 놀람, 기쁨, 화남, 공포, 혐오, 슬픔, 평상 (neutral) 등 6개 음성 감성상태를 분류하는데 평균 80%의 인식률을 보이고 있다. 또한 N. Amir [5]는 4개 감성상태를 분류하기 위해 피치와 음성신호의 신호 구간 개수 (number of voiced period)등의 특징벡터를 이용하여 k-NN과 NN 패턴 분류기의 성능을 비교하였다. C. Lee [6]는 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 이용해서 감성 분류 성능을 남자 음성의 경우 약 40%, 여성 음성의 경우 약 36% 향상시킬 수 있음을 밝히고 있다. 반면 G. Zhou는 논문 [7]에서 TEO (Teager Energy Operator)라는 새로운 특징벡터를 제안하여 평상 음성과 스트레스 음성을 구분하는 흥미로운 연구 결과를 보고하고 있다. 이외에도 미국의 Microsoft, HP, 일본의 SONY등의 산업계에서 음성 감성인식 기술을 HCI용 SW나 로봇 등의 응용분야에 적용하기 위한 활발한 연구를 진행하고 있다.

이상에서 살펴본 바와 같이 대부분의 기존 음성 감성인식 연구는 주로 PC 기반의 잡음이 없는 깨끗한 환경에서의 연구로서 이를 산업화할 때 필수적으로 고려해야 하는 환경 및 유무선 잡음을 고려한 연구는 거의 전무한 실정이다. 일반적으로 음성 감성인식 시스템은 시스템을 훈련시킬 감성음성 DB와 질의 (Query)로 입력되어질 음성 데이터간의 녹음환경 차이 (잡음 환경 등)로 인해 시스템의 감성인식 성능이 저하된다.

본 연구에서는 이동전화 (Cellular phone)를 통해 습득된 음성으로부터 음성의 감성 상태를 평상 혹은 화남 2가지로 구별할 수 있는 실시간 음성 감성인식 시스템을 제안하였다.

제안 시스템은 향후 고객 센터 (Call Center), 결혼정보 회사 등에서 상담원이 고객의 감성상태에 따라 적절한 서비스를 제공할 수 있게 하기 때문에 직접적인 서비스 품질 개선으로 연결될 수 있으며 마케팅 효과 향상에도 유용하게 사용될 것으로 기대된다. 본 논문의 핵심은 다음과 같이 요약될 수 있다.

첫 째, Intel Dialogic D/4PCI 음성 보드를 이용하여 이동전화로부터 음성을 실시간으로 습득하고 음성 감성상태를 평상 혹은 화남 2가지로 구별할 수 있는 감성인식 시스템을 제안한다.

둘 째, 이동통신 네트워크를 통해 수신된 음성은 화자의 환경 잡음과 네트워크 잡음을 포함하고 있어 심각한 시스템 성능저하를 초래함으로, 이러한 잡음 영향을 최소화하여 강인한 감성 특징벡터를 구축할 수 있는 방안을 제안한다. 본 논문에서는 특징벡터 영역에서 비교적 연산량이 간단한 MA (Moving Average) 필터를 적용하여 잡음에 의한 시스템 성능 저하를 최소화시켰다.

셋 째, 특징벡터 중 감성인식률에 기여가 높은 특징계수들만을 선별해서 시스템의 인식 성능을 향상시키고 또한 연산 복잡도를 낮출 수 있는 SFS (Sequential Forward Selection) [8] 기법을 적용하였으며 음성 감성분류를 위한 패턴 인식 기법으로는 k-NN과 SVM 2가지 방법을 비교하였다. 본 논문에서 사용된 SVM 분류기는

가우시안 커널 (Gaussian kernel) SVM이며 페널티 (penalty) 변수인 C값은 200 그리고 커널 민감도를 나타내는 감마 (gamma)값은 215이다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 제안된 음성 감성인식 시스템의 전반적인 구성에 대하여 설명하였다. 3장에서는 이동통신 잡음환경 하에서 음성 전처리, 강인한 감성 특징벡터를 추출하는 기법 그리고 특징벡터 최적화 기법을 제안하였고 4장에서는 다양한 실험을 통해 제안 시스템의 성능을 비교 분석하였다. 마지막으로 5장에서는 결론으로 끝을 맺는다.

II. 제안된 음성 감성인식 시스템

그림 1은 본 논문에서 제안한 음성 감성인식 시스템을 나타낸다. 제안 시스템은 크게 1단계 : SFS를 이용한 최적 감성 특징벡터 추출 단계, 2단계 : 감성 특징벡터 DB 구축 단계 그리고 3단계 : 입력된 질의 음성으로부터 감성 상태를 분류하는 단계 등 3가지 단계로 구성된다.

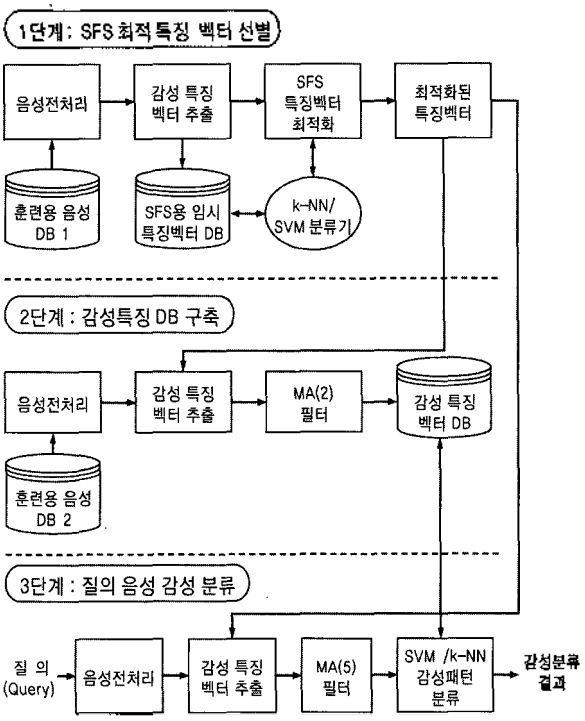


그림 1. 제안된 음성 감성인식 시스템
Fig. 1. Proposed speech emotion recognition system.

1단계 SFS를 이용한 최적 감성 특징벡터 추출에서는 평상, 화남 각 감성별로 구성된 훈련용 음성 DB 1의 음성 신호를 20ms 프레임 단위로 분할해서 해밍 (Hamming) 윈도우를 적용한 후 끝점 검출 (end-point detection) 등의 음성 전처리 과정을 거쳐 다음 3.2 절에 정의된 총 32차의 감성 특징벡터를 추출한다. 추출된 특징벡터는 SFS 특징벡터 최적화 과정 (다음 3.3절 참조)을 거쳐 각 k-NN/SVM 분류기별로 감성인식 성공률에 가장 큰 기여를 하고 있는 특징계수들만을 선별해서 최종 감성 특징벡터를 구성하게 된다. 본 논문에서 최적 특징벡터를 추출하기 위한 훈련용 음성 DB 1이 2단계의 감성 특징벡터 DB 구축에 사용된 훈련용 음성 DB 2와 다른 이유는 가능한 한 특정 DB에 의존하지 않는 범용성 있는 특징벡터를 추출하기 위해서다.

2단계 감성 특징벡터 DB 구축에서는 별도로 준비된 훈련용 음성 DB 2의 음성 신호를 음성 전처리 과정을 거쳐 1단계에서 각 k-NN/SVM 분류기별로 선별된 특징벡터만을 추출한 후 2차 MA 필터를 적용하여 감성 특징벡터 DB를 구축하게 된다.

3단계 질의 음성에 대한 감성 분류는 음성 전처리 과정을 거쳐 1단계에서 각 k-NN/SVM 분류기별로 선별된 특징벡터만을 추출한 후 특징벡터 영역에서 5차 MA 필터를 적용해 환경 및 네트워크 잡음에 의한 영향을 최소화시킨다. 최종적으로는 k-NN과 SVM 패턴 분류기를 이용하여 질의 음성의 감성 상태를 평상 혹은 화남으로 분류한다.

III. 강인한 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.

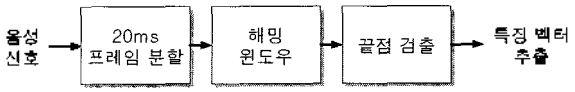


그림 2. 음성 전처리 과정
Fig 2. Speech pre-processing.

입력되는 음성 신호는 먼저 20ms 프레임 단위로 분할되고 이웃한 프레임과 50% 중복되는 Hamming 윈도우를 적용받게 된다. 이후 끝점 검출은 음성 신호에서 음성 구간 (voiced period)과 비-음성 구간을 구별하여 음

성 구간에서만 특징벡터를 추출하기 위한 것으로 이는 비-음성 구간에서의 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 [9]을 이용하였다.

3.2. 감성 특징벡터 추출

음성 특징벡터는 매 프레임 단위로 운율적 특징을 갖는 피치, 에너지와 음소 특징을 갖는 MFCC 그리고 각 특징계수의 델타 (Delta) 값을 추출하였고 최종적으로는 각 특징계수들의 평균 (mean)과 표준 편차 (standard deviation)를 구하여 총 32차의 특징벡터를 구성하였다.

피치 (Pitch)는 일반적으로 많이 사용되는 HPS [10], AMDF [11], SHR [12] 방법을 비교 실험하여 잡음 환경에서 가장 높은 감성인식률을 나타낸 SHR (Subharmonic to harmonic ratio) 알고리즘을 사용하였다. SHR은 음성 신호에 FFT를 취하여 2개의 피크 값 ((f_1, f_2)을 피치 후보로 선정하고 아래의 수식 (1)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 [12]에서 주어진 차분 함수이다.

$$SHR = 0.5\frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)} \quad (1)$$

단-구간 음성 에너지는 수식 (2)를 이용하여 각 프레임에서의 에너지를 계산하였다.

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

MFCC (Mel Frequency Cepstral Coefficient)는 음성 인식 분야 등에서 널리 사용되는 특징으로 사람의 청각 특성과 유사한 멜-주파수 (Mel-Frequency)상에서의 음성 특성을 잘 표현할 수 있다. 본 논문에서는 MFCC 4, 6, 8, 12차에 대한 인식률 비교 실험을 통하여 연산량과 인식률에서 가장 적합한 MFCC 6차를 사용하였다.

3.3. SFS 특징벡터 최적화

SFS 기법 [8]은 총 32차 특징벡터 간에 중복된 상관성 (Correlation)을 제거하고 동시에 시스템의 감성 인식률에 가장 기여를 많이 하고 있는 최적의 특징계수들만을 선정하는 방법으로 시스템의 성능을 높여줌과 동시에 연산 복잡도를 낮출 수 있는 장점이 있다. SFS는 먼저 각 특징계수들을 개별적으로 사용하여 감성 분류를 한 후, 가장 좋은 감성 인식률을 나타내는 특징계수부터 순차적으로 하나씩 특징계수를 추가해 나가면서 감성 인식 정확도를 계산하게 된다. 이 과정을 거쳐 최적의 감성 인식률을 얻을 수 있는 처음 몇 개의 특징계수만을 선정해서 특징벡터 열을 새롭게 구성하게 된다.

3.4. 이동통신 환경에서 잡음 영향 최소화를 위한 MA 필터링

이동전화 (Cellular Phone)를 통해 수신된 음성은 화자의 환경 잡음과 네트워크 잡음을 포함하고 있어 추출되는 감성 특징벡터를 왜곡하게 되고 따라서 심각한 시스템 성능저하를 초래하게 된다. 일반적으로 이러한 복합 잡음들은 예측하기 어려운 패턴을 가지고 있기 때문에 잡음의 통계적 특성을 이용한 기존의 필터링 기법들을 사용하기에는 어려운 점이 있다. 따라서 본 논문에서는 그림 1에서와 같이 비교적 연산량이 적은 MA (Moving Average) 필터를 특징벡터에 적용하여 잡음에 의한 특징벡터 왜곡을 완화시켜 감성인식 시스템 성능저하를 예방하였다.

MA 필터는 기본적으로 저역 통과 필터 (Low Pass Filter)이므로 잡음에 의해 급격한 변화를 보이는 특징벡터 열 부분을 부드럽게 하는 역할을 한다. 물론 MA 필터 적용으로 인하여 원 음성 신호에 포함되어 있는 일부 음성 특징도 왜곡될 수 있으나, 일반적으로 잡음이 섞인 음성신호의 특징벡터 열에서 급격히 변하는 부분은 잡음에 의한 경우가 대부분이기 때문에 MA 필터를 사용하는 것이 타당하다 할 수 있다.

MA 필터를 특징계수에 적용하는 방법은 다음과 같다. 먼저 음성신호의 분석시간 동안 (본 논문에서는 2초) 매 20ms 단위로 추출한 후 SFS 특징벡터 최적화 과정으로 선정된 특징계수를 다음과 같이 $T \times D$ 행렬로 표현한다.

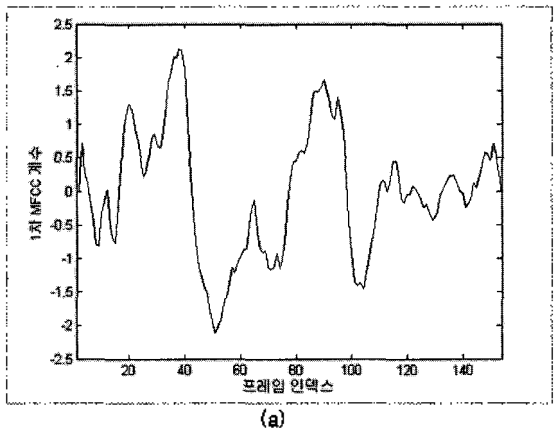
$$FM_{id} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \cdots & \cdots & \cdots & \cdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,D} \end{bmatrix} \quad (3)$$

여기서 t는 신호 분석시간 동안의 시간 순서별 프레임 번호 $t = 1, 2, \ldots, T$ 이고 D는 특징벡터의 차수이다. 따라서 수식 (3)의 행은 시간 순서별 각 T개 프레임에서

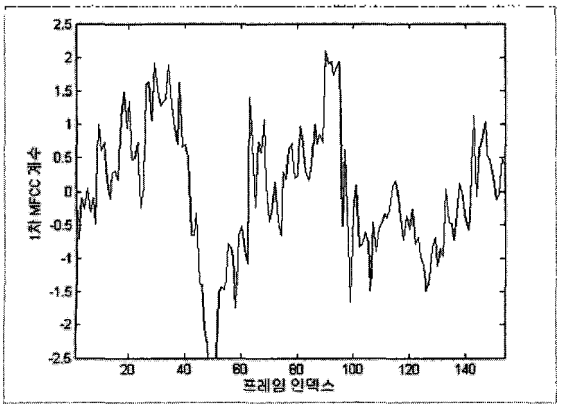
추출된 32개의 특징벡터를, 그리고 열은 각각의 특징계수들이 시간 순서별 프레임에 따른 변화를 나타낸다. 수식 (3)의 특징계수 행렬은 각 특징계수들 간의 편차로 인한 오 분류 동작을 방지하기 위해 평균 0, 표준편차 1이 되도록 각 열별로 정규화 하였다. 여기서 MA 필터는 정규화된 각 특징계수 열별 프레임 방향으로 적용되며 본 논문에서는 수식 (4)의 MA 필터에 대한 다양한 실험을 거쳐 필터 차수를 M=5로 하였다.

$$\widehat{x_{i,f}} = \frac{1}{M}\sum_{i=0}^{M} x_{(i-n),f} \qquad x_{i,f} = \text{특징계수} \qquad (4)$$

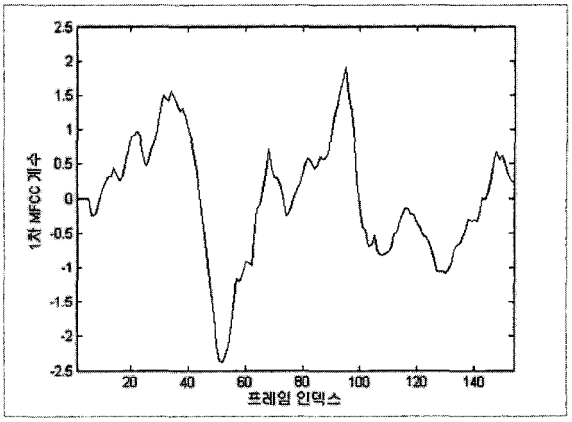
그림 3은 1차 MFCC 계수에 대해 잡음으로 인한 왜곡을 최소화하는 MA (5) 필터의 성능을 비교하고 있다. 그림 3에서 (a)는 깨끗한 음성 신호에서 추출한 1차 MFCC 계수의 프레임에 따른 변화, (b)는 이동전화를 통해 수신된 잡음이 섞인 음성 신호에서 추출한 1차 MFCC 특징 계수 열, (c)는 (b)의 특징계수 열에 MA (5) 필터를 적용한 결과이다. 그림에서 보듯이 MA (5) 필터가 그림 3 (b)의 특징벡터 열의 급격하게 변하는 부분을 부드럽게 해서 깨끗한 음성에 대한 특징벡터 열과 잡음 음성 신호에 대한 특징벡터 열 차이를 최소화하는 것을 볼 수 있다.



(a)



(b)



(c)

그림 3. MA 필터를 이용한 특징벡터 열 왜곡 완화 : (a) 깨끗한 음성 신호에서 추출한 1차 MFCC 계수의 프레임에 따른 변화, (b) 이동전화를 통해 수신된 잡음이 섞인 음성 신호에서 추출한 1차 MFCC 특징 계수 열, (c)는 (b)에 MA(5) 필터를 적용한 결과

Fig. 3. Alleviation of feature distortion using MA filter: (a) 1st MFCC coefficient over the frame index for clean speech, (b) feature sequences from the noisy speech captured by cellular phone, (c) resulting feature sequences by applying the MA(5) filter to (b).

한편 본 연구에서는 잘의 음성의 특징벡터 열에 MA(5) 필터를 적용할 뿐만이 아니라 감성 특징벡터 DB 구축 단계에서도 2차의 MA 필터를 적용하게 되는데 이는 질의 음성의 특징벡터 열과 감성 DB 특징벡터 열의 차이를 가능한 최소화하여 패턴 매칭에 의한 감성 분류 시 오동작을 완화하기 위한 것이다.

IV. 실험 환경 및 결과

4.1. 실험 환경

본 연구에서는 논문 [13]의 기 구축된 DB를 이용하여 평상, 화남 두 가지 감성에 대한 감성 음성 DB를 구축하였다. 논문 [13]의 감성 음성 DB는 평소 감성 음성 발성을 훈련하는 아마추어 연극단원 남, 여 각 15명이 45개 문장에 대하여 발성한 음성을 총 2700개 8kHz, 16bit로 녹음한 것이다. 본 연구에서는 2700개 문장 중 비교적 감성이 잘 표현되어 있고 문장 길이가 2초 이상이 되는 1200개 문장만을 선정해서 본 연구에 사용하였다. 1200개 문장 중 무작위로 200 (평상-100, 화남-100)개의 음성 (그림 1에서 훈련용 음성 DB 1)을 선택하여 SFS를 이용한 최적의 감성 특징벡터를 선별하는데 사용하였고, 이와는 별도로 중복되지 않는 800개 (평상-400개, 화남-400개)의 음성을 선택하여 감성 특징벡터 DB를 구축

하기 위한 훈련데이터 (그림 1에서 훈련용 음성 DB 2)로 사용하였다. 제안 시스템의 평가에 사용될 질의 데이터는 훈련용 음성 DB 1, DB 2와 중복되지 않는 별도의 200 (평상-100, 화남-100)개 음성을 무작위로 선택하여 사용하였다.

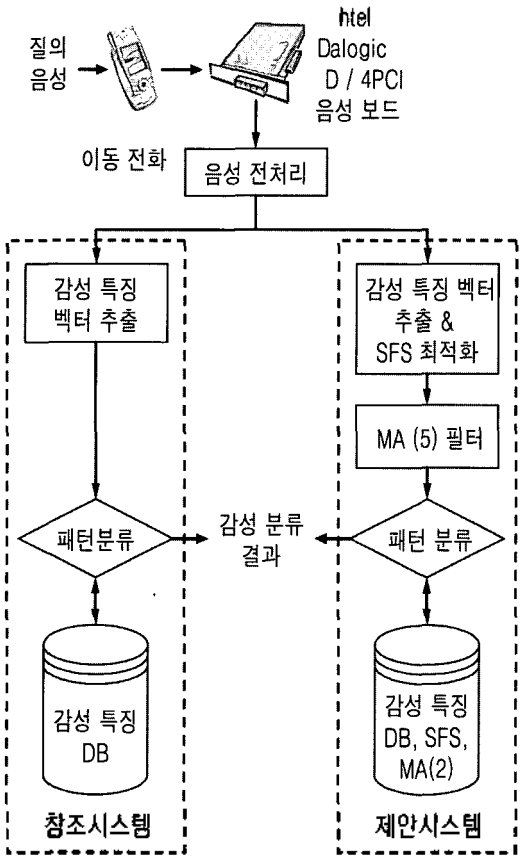


그림 4. 제안 시스템의 성능비교를 위한 실험환경
Fig. 4. Experimental setup for performance comparison of the proposed system.

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

그림 4에서 실제 질의 음성은 이동전화를 거쳐 실시간으로 습득되기 때문에 수신된 음성에는 잡음이 섞이게 된다. 그림에서 참조 시스템은 음성 전처리 과정을 거쳐 32차 감성 특징벡터 전부를 추출하여 k-NN과 SVM 패턴 분류기로 음성 감성상태를 인식하는 시스템이다. 반면에 제안 시스템은 본 논문에서 제안된 시스템을 나타낸 것으로 음성 전처리 과정을 거쳐 그림 1의 훈련용 음성 DB 1에서 SFS 기법으로 선정된 최적화된 특징벡터 만을 추출하여 MA (5) 필터를 적용한 후 k-NN과 SVM 패턴 분류기로 최종 음성 감성 상태를 분류하게 된다. 제안 시스템의 경우 감성 특징 DB의 특징벡터 또한 SFS로 최적화된 특징벡터 열에 MA (2) 필터를 적용한 것이다.

4.2. 실험 결과

다음의 그림 5는 제안 시스템에서 사용하게 될 최적화된 특징벡터 계수를 선정하기 위한 SFS 실험 결과를 보이고 있다. 그림 5의 SFS 특징 벡터 최적화 과정은 훈련용 DB 1 (깨끗한 음성)에 대해 3.2절에서 정의된 32차 특징벡터 중 최적의 특징벡터 열만을 선정하기 위한 것으로 그림에서 점선은 k-NN 분류기에 의한 특징벡터 최적화 과정을 그리고 실선 부분은 SVM 분류기에 의한 최적화 과정을 나타낸다.

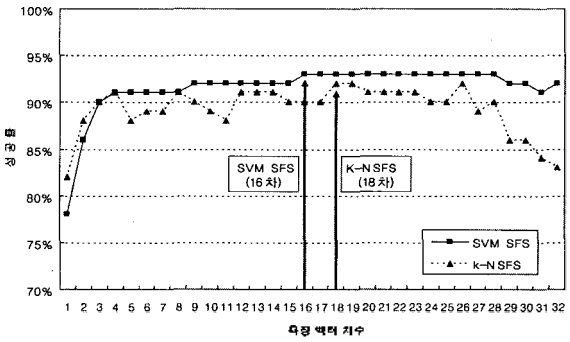


그림 5. SFS 특징벡터 최적화 과정
Fig. 5. SFS feature vector optimization process.

그림 5에서 보듯이 k-NN 분류기를 이용할 경우 약 18차 특징벡터 차수에서 최대 92%의 감성 인식률을 나타내고 있고 반면 SVM 분류기의 경우 약 16차 특징벡터 차수에서 최대 93%의 감성 인식률을 나타내고 있다. 따라서 k-NN 분류기 경우 처음 18차 특징 계수로 특징벡터 열을 구성하고 SVM 분류기의 경우 처음 16차 특징 계수로 특징벡터 열을 구성하는 것이 바람직한 것을 알 수 있다. 한편 감성 인식률에 가장 많은 영향을 미치는 특징계수는 k-NN, SVM 분류기 공통으로 MFCC, 피치, 에너지 순으로 나타났다.

그림 5의 실험에서 SFS 특징 벡터 최적화를 실제 이동전화로 수신된 잡음이 섞인 음성 신호가 아닌 깨끗한 음성의 훈련용 DB 1을 대상으로 적용한 주된 이유는 잡음 패턴의 불규칙성으로 인해 안정된 특징 벡터 열을 선정할 수 없기 때문이다. 또한 이 실험에서 유의해야할 사항은 그림 5의 실험 과정은 단지 각 분류기별 최적화된 특징벡터 열만을 선정하기 위한 실험으로 이 실험에서 구해진 감성 인식률은 깨끗한 음성에 대한 인식결과로 실제 이동통신 환경에서 습득된 질의 음성에 대한 인식률과는 아무런 관련이 없다는 점이다. 그림 5에서 SFS 기법으로 선정된 특징벡터는 그림 4와 같이 제안 시스템에서 이동전화로 수신된 음성 신호로부터 특징벡터를 추

출할 때 사용하게 된다.

다음의 표 1은 그림 4의 이동통신 실험 환경에서 참조 시스템의 감성 분류 결과와 제안 시스템의 감성 분류 결과를 비교한 것이다.

표 1. 제안된 감성 인식 시스템의 성능 비교
Table 1. Performance comparison of proposed emotion recognition system.

	참조 시스템		제안 시스템	
	k-NN (32차)	SVM (32차)	k-NN (18차)	SVM (16차)
평균 감성 인식률	67.5%	72%	73.5%	86.5%

표 1에서 알 수 있듯이 제안 시스템이 참조 시스템에 비해 평균 8% ~ 14.5%의 성능 향상을 가져옴을 볼 수 있으며 SVM 분류기는 k-NN 분류기에 비해 약 13% 정도의 높은 인식률을 제공함을 알 수 있다. 이미 잘 알려져 있는 바와 같이 이진 분류 SVM 알고리즘은 k-NN보다 빠른 연산속도를 가지고 있어 본 제안 시스템의 경우에는 감성 인식률이나 연산량 측면 모두에서 SVM 분류기를 사용하는 것이 바람직한 것을 알 수 있다.

표 2는 위 표 1의 각 감성 분류 결과를 구체적으로 비교한 것으로 200 (평상-100, 화남-100)개 질의 음성에 대한 결과를 나타낸다. 표 2의 (a) 와 (b)는 각각 k-NN과 SVM을 이용한 참조 시스템의 분류 결과표를 그리고 (c)와 (d)는 각각 k-NN과 SVM을 이용한 제안 시스템의 감성분류 결과표를 나타낸다.

표 2. 감성 분류 결과 비교: (a) k-NN을 이용한 참조 시스템, (b) SVM을 이용한 참조 시스템 (c) k-NN을 이용한 제안 시스템, (d) SVM을 이용한 제안 시스템
Table 2. Comparison of emotion classification result: (a) reference system with k-NN, (b) reference system with SVM, (c) proposed system with k-NN, (d) proposed system with SVM.

시스템 / 분류기 / 질의	참조 시스템				제안 시스템			
	(a) k-NN		(b) SVM		(c) k-NN		(d) SVM	
	평상	화남	평상	화남	평상	화남	평상	화남
평 상	73	27	46	54	68	32	80	20
화 남	38	62	2	98	21	79	7	93
평균 감성 인식률	67.5%		72%		73.5%		86.5%	

표 2에서 보듯이 참조 시스템에서 k-NN 분류기를 이용한 경우에는 평상, 화남 감성에 대한 오분류율이 각각 27%, 38%로서 유사한 정도를 보이지만 SVM 분류기에서는 대부분의 오분류가 평상 감성 (54%)에서 발생하고 화남 감성은 거의 완벽하게 분류해 내는 것을 알 수 있다. 반면 제안 시스템의 k-NN 분류기 경우 평상, 화남 감성에 대한 오분류율은 각각 32%, 21%로 화남 감성에 대한 부분적인 성능 개선을 볼 수 있으며 SVM 분류기의 경우 평상 감성에 대한 오분류율이 참조 시스템의 54%에서 20%로 줄어들어 획기적으로 개선된 것을 알 수 있다.

표 2와 같이 이동통신 잡음 환경에서 평상, 화남 감성 분류 시스템의 주된 오분류는 평상 감성에서 발생하고 있으며 화남 감성은 비교적 잡음에 강인한 특성을 가지고 있음을 알 수 있다. 이러한 현상은 논문 [3]에서 지적한 바와 같이 화남 감성의 피치, 에너지, MFCC 같은 특징 계수들이 비교적 큰 값을 가지고 있어 다른 감성과 구별이 잘되는 반면 평상 감성은 상대적으로 작은 값을 갖고 있어 잡음의 영향을 많이 받기 때문인 것으로 해석될 수 있다.

V. 결 론

본 논문은 이동전화 (Cellular phone)를 통해 실시간으로 습득된 음성으로부터 사람의 감성 상태를 평상 혹은 화남 2가지로 인식할 수 있는 음성 감성인식 시스템을 제안하였다. 일반적으로 이동전화를 통해 수신된 음성은 화자의 환경 잡음과 네트워크 잡음을 포함하고 있어 음성 신호의 감성특징을 왜곡하게 되고 이로 인해 인식 시스템에 심각한 성능저하를 초래하게 된다. 따라서 본 논문에서는 이러한 잡음 영향을 최소화하고 강인한 감성 특징벡터를 추출하기 위해 비교적 단순한 구조의 MA (Moving Average) 필터를 제안하였으며 SFS 특징벡터 최적화 기법을 적용하여 시스템 성능을 한층 더 안정화시켰다. 제안 시스템은 실제 이동통신 잡음 환경에서의 실험 결과 SVM 패턴 분류기를 이용해서 약 86.5%의 높은 인식률을 달성할 수 있어 향후 다양한 산업 분야에 적용할 수 있는 가능성을 보이고 있다. 제안 시스템은 특히 고객 센터 (Call center), 결혼정보회사 등에서 상담원이 고객의 감성상태에 따라 적절한 서비스를 제공할 수 있게 하기 때문에 직접적인 서비스 품질 개선으로 연결될 수 있으며 마케팅 효과 향상에도 유용하게 사용될 것으로 기대된다.

참 고 문 헌

1. M. Liu and C. Wan, "A Study on Content-based

Classification Retrieval of Audio Database," Proc. of the International Database Engineering & Applications Symposium, 339-345, 2001.
2. F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech", In Proc. International Conf. on Spoken Language Processing, 1970-1973, 1996.
3. T. Moriyama and Oazwa, "Emotion Recognition and Synthesis System on Speech", IEEE International Conference on Multimedia Computing and Systems, 1 840-844, Florence, Italy, 1999.
4. A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Marino, "Speech Emotion Recognition Using Hidden Markov Models," presented at Eurospeech 2001, Scandinavia, 2001.
5. Noam Amir, "Classifying Emotion in Speech : a Comparison of Methods", Proceedings of Euro Speech'2001, 1 127-130, Aalborg, Denmark, 2001.
6. C. M. Lee and S. S. Narayanan, "Towards Detecting Emotions in Spoken Dialogs," in IEEE Transactions on Speech and Audio Processing, 13 (2) 2005
7. Guojun Zhou, John H. L. Hansen, and James F. Kaiser, "Nonlinear Feature Based Classification of Speech Under Stress" IEEE Transactions on Speech and Audio Processing, 9 (3) 2001.
8. Anil Jain and Douglas Zongker, "Feature Selection : Evaluation, Application, and Small Sample Performance", IEEE Pattern Analysis and Machine Intelligence, 19 (2) 153-158, 1997.
9. Lingyun Gu and Stephen A. Zahorian, "A New Robust Algorithm for Isolated Word Endpoint Detection," IV-4161 International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 13-17, 2002.
10. P. de la Cuadra, A. Master and C. Sapp, "Efficient Pitch Detection Techniques for Interactive Music", International Computer Music Conference, 403-406, Havana, Cuba, September, 2001.
11. M.J. Ross, H.L. Shaer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor", Acoustics, Speech, and Signal Processing (see also IEEE Transactions on Signal Processing), IEEE Transactions on 22 (5) 353-362, Oct. 1974.
12. Xuejing Sun, "A Pitch Determination Algorithm Based On Subharmonic-to-Harmonic Ratio", International Conference on Spoken Language Processing '2000, 676-679, 2000.
13. 강봉석, "음성 신호를 이용한 문장독립 감정 인식 시스템", 석사학위 논문, 연세대학교, 2001.

저자 약력

• 조 윤 호 (Youn-Ho Cho)

1970년 4월 25일생
1994년 2월: 단국대학교 농학과 (농학사)
2000년 12월~현재: (주)리얼커뮤니케이션
2003년 9월~현재: 경복대학 겸임교수
2004년 2월: 단국대학교 멀티미디대학원 멀티미디어학과(공학석사)
2006년 8월: 단국대학교 정보 · 컴퓨터학과(박사수료)

• 박 규 식 (Kyu-Sik Park)
한국음향학회지 제23권 7호 참조