한국음향학회지 제25권 제8호 pp. 375-382 (2006)

확률출력 SVM을 이용한 감정식별 및 감정검출

Identification and Detection of Emotion Using Probabilistic Output SVM

조 훈 영*, 정 규 준**
(Hoon-Young Cho*, Gue Jun Jung**)

*한국전자통신연구원 디지털콘텐츠연구단 HD게임연구팀, **한국과학기술원 전자전산학과

(접수일자: 2006년 9월 25일; 수정일자: 2006년 11월 3일; 채택일자: 2006년 11월 15일)

본 논문에서는 음성신호에 포함된 감정정보를 자동으로 식별하는 방법과 특정 감정을 검출하는 방법에 대해 다룬다. 자동 감정식별 및 검출을 위해 장구간 (long-term) 음향 특징을 사용하였고, F-score 기반의 특징선택 기법을 적용하여 최적의 특징 파라미터들을 선정하였다. 기존의 일반적인 SVM을 확률출력 SVM으로 변환하여 감정식별 및 감정검출 시스템을 구축하였으며, 가설검정에 기반한 감정검출을 위해 세 가지의 대수 우도비 (log-likelihood) 근사법을 제안하여 그 성능을 비교하였다. SUSAS 데이터베이스를 사용한 실험 결과, F-score를 이용한 특징선택 기법에 의해 감정식별 성능이 향상되었으며, 확률출력 SVM의 유효성을 검증할 수 있었다. 감정검출의 경우, 제안한 방법에 의해 91.3%의 정확도로 화난 감정을 검출할 수 있었다.

핵심용어: 감정식별, 감정검출, 확률출력 SVM, 특징선택

투고분야: 음성처리 분야 (2.5)

This paper is about how to identify emotional information and how to detect a specific emotion from speech signals. For emotion identification and detection task, we use long-term acoustic feature parameters and select the optimal parameters using the feature selection technique based on F-score. We transform the conventional SVM into probabilistic output SVM for our emotion identification and detection system. In this paper we propose three approximation methods for log-likelihoods in a hypothesis test and compare the performance of those three methods. Experimental results using the SUSAS database showed the effectiveness of both feature selection and probabilistic output SVM in the emotion identification task. The proposed methods could detect anger emotion with 91.3% correctness.

Key words: Emotion identification, Emotion detection, Probabilistic output SVM

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

음성신호는 언어정보, 화자의 개인성정보, 성별, 국적, 출신지역, 발성 당시 주변환경의 음향정보, 화자의 감정상태, 피로도 등 다양한 정보를 포함하고 있으며, 최근에 인간과 로봇 간의 감성 인터페이스에 대한 관심이 고조되면서 화자의 감정상태를 자동식별하는 방법에 관한 연구가 활발히 이루어지고 있다. 자동 감정식별 분야의 최근 연구들은 대개 네 종류에서 일곱 종류의 감정을 대상으로 하고 있으며, 화남, 슬픔, 지루함, 즐거움, 중립감정 등을 예로 들 수 있다. 이 분야의 주된 연구주제로 감정 분류를 위한 효과적인 특징추출, 감정식별기 설계, 감정식별을 위한 데이터베이스 구축 등을 들 수 있다 [1].

감정식별을 위한 특징 값으로는 피치, 에너지궤적, 지속길이정보 등의 운율정보 및 스펙트럼정보 등이 이용되고 있다. 이 중 다수의 연구결과에서 피치 및 에너지가 감정식별에 매우 효과적임이 밝혀졌으나, 그 이외의 특징표현에 대한 연구가 지속적으로 이루어지고 있다.

인식단계에서는 HMM (hidden Markov model), GMM (Gaussian mixture model), SVM (support vector machine), 신경회로망, 퍼지추론, k-nearest neighbor, LDA (linear discriminant analysis) 등 다

책임저자: 조 훈 영 (hycho@etri.re.kr)
305-700 대전시 유성구 가정동 161번지 한국전자통신연구원
디지털콘텐츠연구단 HD게임연구팀
(전화: 042-860-6591; 팩스: 042-860-1058)

양한 방식이 시도되었으며 [2][3][4][10], 각 분류방법에 따라 수십 ms구간에서 특징을 추출하는 단구간 특징 (short-term feature) 또는 수 초 이상의 음성 구간에서 특징을 추출하는 장구간 특징 (long-term feature)을 사용한다. 예를 들어, HMM의 경우 단구간 특징을 사용하며, SVM의 경우 장구간 특징벡터를 입력 패턴의 분류에 사용한다. 현재, 감정식별의 세계적인 수준은 네 종류의 감정에 대해 대략 60% 가량의 정확도를 보이지만, 데이터베이스의 종류 혹은 발성자의 국적 등에 따라 상당한 변이를 나타낸다.

다수의 감정을 식별하는 대신에 하나의 감정에 집중하여 이 감정의 표현여부를 검출하는 방법에 대한 연구도 수행된 바 있다. 평상시의 중립적 감정에서 화난 음성을 검출하는 태스크에서는 약 90%에 달하는 인식률을 얻을 수 있다 [2][3][4]. 이 경우, 자동 응답 시스템 (IVR; interactive voice response)에서 화가 난 고객의 전화를 자동으로 실제 상담원에게로 전환한다든지 [3], 또는 대용량의 방송 오디오 데이터베이스에서 특정 감정에 해당하는 부분만을 검색하는 데에 응용될 수 있다.

본 연구에서는 확률출력 SVM을 이용하여 감정식별 시스템을 설계하고 특징선택 기법을 적용하여 이 시스템을 최적화하며, 이 시스템을 기반으로 감정검출 (emotion detection) 방법을 제안한다. 기존의 SVM이 입력에 대해 클래스 ID만을 출력으로 내는 것에 비해 확률출력 SVM은 각 클래스 ID의 사후확률값 (a posteriori probability)을 알 수 있으므로, 이를 이용한 다양한 응용이 가능하다. 제안한 감정검출 방법은 입력 음성과 확인하고자 하는 감정에 대해 신뢰도를 측정하여 신뢰도가 높은 경우에만 수락 (accept)한다 [7]. 이를 위해 관심대상 감정 e 와 이를 제외한 모든 다른 감정 $\bar{e}$ 의 대수 우도비(log likelihood ratio)를 계산하고, 이 값을 정해진 임계치와 비교하여 수락 및 거절 여부를 결정한다. 관심대상 감정 e 의 여집합 (complimentary set) $\bar{e}$ 을 모델링하기 위한 모든 데이터를 획득하기란 불가능하므로 본 연구에서는 이를 근사하기 위한 세 가지 방법을 제시한다.

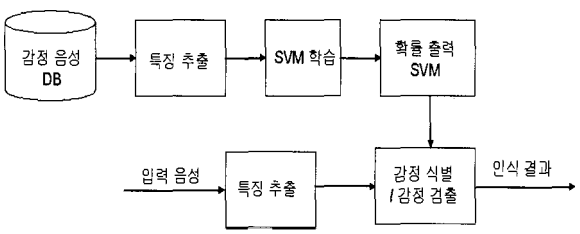


그림 1. 확률출력 SVM을 이용한 감정식별 및 감정검출 시스템
Figure 1. Emotion Identification and Detection system based on probabilistic output SVM.

본 논문의 제 2장에서는 사용된 특징추출 및 특징선택 방식을 기술하며, 제 3장에서는 분류기로 사용한 확률출력 SVM을 소개한다. 제 4장에서 제안한 감정검출 방식을 설명하고, 제 5장에서 실험 결과를 기술한 뒤, 제 6장에서 결론을 맺기로 한다.

II. 감정식별을 위한 특징추출

음성인식분야의 오랜 연구결과 멜 켑스트럼 (Mel Frequency Cepstral Coefficient; MFCC)과 같이 스펙트럼 정보에 기반한 특징 파라미터들이 음성인식에 효과적인 특징 파라미터로 알려졌다. 그러나, 비교적 최근에 활발히 연구되고 있는 감정식별 분야에서는 피치, 에너지, 지속길이 등의 운율정보에 기반한 특징 파라미터들이 널리 사용되고 있으며, 시스템의 성능을 향상시킬 새로운 특징들을 계속하여 연구하고 있다. 본 연구에서는 음성 신호의 에너지, 피치, 지속길이 및 스펙트럼 정보로부터 42 종류의 특징 파라미터를 추출하였다. 본 장에서는 이들 각각에 대해 간략히 기술한다.

1. 프레임 에너지 기반의 특징 파라미터

에너지 파라미터를 추출하기 위해 매 10ms마다 20 ms의 음성구간에서 로그 에너지를 계산하여 에너지 궤적 (energy trajectory)을 얻고, 0에서 1사이 값으로 정규화한다. 정규화한 에너지 궤적에서 임계치를 이용하여 음성 신호의 전후에서 무음 구간을 제거한다. 이렇게 얻은 음성 구간의 에너지 궤적으로부터 차분 (delta) 및 차차분 (delta-delta) 연산을 통해 에너지 속도 궤적 및 에너지 가속도 궤적을 획득한다. 에너지 궤적을 $E(t)$ 로 표현할 때, 본 연구에서 차분 연산은 $\Delta(t) = E(t+1) - E(t-1)$ 로 정의하였고, 차차분 연산은 $\Delta\Delta(t) = \Delta(t+1) - \Delta(t-1)$ 로 정의하였다. 각 궤적들에서 최소 및 최대값, 평균값, 표준편차 및 범위 (range)를 구하여 특징으로 사용한다.

2. 피치 및 지속길이 기반의 특징 파라미터

피치와 관련된 특징 파라미터를 추출하기 위해 본 연구에서는 바이터비 (Viterbi) 피치 추적 알고리즘을 이용하여 피치 궤적을 계산한다 [11]. 자기상관 계수 (autocorrelation coefficient)의 Lag값 τ 는 HMM의

상태번호로 사용되고, t 번째 프레임에서 자기상관 계수를 $ac_t[\tau]$라고 할 때, 각 상태에서의 출력확률은 $1-ac_t[\tau]/ac_t[0]$로 정한다. 피치 궤적으로부터 피치의 속도 궤적 및 가속도 궤적을 앞 절의 에너지와 동일한 방식으로 계산하고, 각각의 피치 궤적들로부터 최소값, 최대값, 평균, 표준편차 및 범위를 구하여 피치기반 특징으로 사용한다.

다음으로 피치 정보로부터 주기성 (fundamentalness)을 계산하여 전체 음성 구간을 다시 유성음과 무성음 구간으로 구분한다. t 번째 프레임의 피치 주기를 $p(t)$라고 하고, 동일 프레임의 자기상관 계수를 $ac_t[\tau]$라고 할 때 주기성 정도 $fn(t)$는 다음과 같이 계산된다.

$$fn(t) = ac_t(p(t))/ac_t(0) \quad (1)$$

$fn(t)$의 값은 주기성이 강할수록 큰 값을 나타낸다. $fn(t)$의 값을 미리 정해놓은 임계치와 비교하여 각 프레임들의 유·무성음 여부를 표현하고, 지속길이 (duration)에 관한 특징을 계산할 수 있다. 본 연구에서는 전체 음성구간 중에서 유성음 및 무성음이 차지하는 구간의 비율을 0과 1사이의 값으로 표현하여 지속길이 특징으로 사용한다.

3. 스펙트럼 관련 특징 파라미터

스펙트럼 기울기 (spectral tilt) 특징을 계산하기 위해서 각 프레임의 2000~4000 Hz에 해당하는 음성대역의 에너지값을 0~1000 Hz 대역의 에너지값으로 나눈 후, 전체 프레임에 대한 평균값을 취한다. 또한, 13차의 멜 켑스트럼 계수 (Mel frequency cepstral coefficients; MFCC)를 유성음 구간에서 추출하여 각 차수의 평균값을 성도 (vocal tract) 정보를 나타내는 특징 파라미터로 사용한다.

이상에서 기술한 특징 파라미터들을 이용하여 본 연구에서 구성한 특징벡터는 표 1과 같다.

III. 특징 선택 기법

감정식별을 위한 효과적인 특징집합은 주어진 응용 시스템에 포함될 감정의 종류에 따라 달라질 수 있다. 예를 들어, 화난 감정과 중립 감정을 분류하는 시스템에서는 피치와 에너지 정보가 효과적일 가능성이 높고, 지루함 (boredom)과 중립 감정을 분류하는 경우에는 지속길이 정보가 효과적일 수 있다. 이는 감정검출의 경우도 마찬가지이며, 따라서 다수의 특징집합에서 주어진 태스크에 효과적인 특징을 선별하는 특징선택 (feature selection)이 중요하다 [4][5][6]. 본 연구에서는 F-score를 이용한 특징선택 기법을 적용하여 특징 벡터의 차수를 줄임과 동시에 감정식별의 정확도를 높이고자 한다. F-score는 벡터공간 상에서 두 개의 클래스에 대한 식별력을 측정하는 척도로서 두 클래스 A, B에 대한 학습벡터 $\mathbf{x}_k \in R^n, k = 1,\cdots,m$가 주어지고, 각각이 클래스 A 혹은 B에 속하며, 클래스 A 또는 B에 해당하는 학습벡터

표 1. 감정 인식 실험에 사용된 특징 파라미터의 차수 및 종류. Avg., Min., Std.는 각각 평균, 최소값, 표준편차를 의미하며, F0는 피치주기, Eng는 에너지, V 및 UV는 유성음 및 무성음을 의미한다. Δ 및 ΔΔ는 차분 (delta) 및 차차분 (delta delta) 특징 파라미터이다.

Table 1. Feature parameters used for emotion recognition task. Avg., Min., and Std. stand for average, minimum and standard deviation, respectively. F0 is pitch period and Eng. stands for energy. V and UV stands for voiced and unvoiced segments, respectively.

차수	특징종류	차수	특징종류	차수	특징 종류
(1)	Avg. F0	(15)	Min. ΔEng	(29)	Std. of Eng
(2)	Avg. Eng	(16)	Max. ΔF0	(30)	Avg. MFCC (0)
(3)	Min. F0	(17)	Max. ΔEng	(31)	Avg. MFCC (1)
(4)	Min. Eng	(18)	Range of ΔF0	(32)	Avg. MFCC (2)
(5)	Max. F0	(19)	Range of ΔEng	(33)	Avg. MFCC (3)
(6)	Max. Eng	(20)	Avg. ΔΔF0	(34)	Avg. MFCC (4)
(7)	Range of F0	(21)	Avg. ΔΔEng	(35)	Avg. MFCC (5)
(8)	Range of Eng	(22)	Min. ΔΔF0	(36)	Avg. MFCC (6)
(9)	Avg. Spectral Tilt	(23)	Min. ΔΔEng	(37)	Avg. MFCC (7)
(10)	Duration of V	(24)	Max. ΔΔF0	(38)	Avg. MFCC (8)
(11)	Duration of UV	(25)	Max. ΔΔEng	(39)	Avg. MFCC (9)
(12)	Avg. ΔF0	(26)	Range of ΔΔF0	(40)	Avg. MFCC (10)
(13)	Avg. ΔEng	(27)	Range of ΔΔEng	(41)	Avg. MFCC (11)
(14)	Min. ΔF0	(28)	Std. of F0	(42)	Avg. MFCC (12)

수가 각각 N_A, N_B 라면 i 번째 특징 파라미터의 F-score는 다음과 같이 정의된다.

$$F(i) = \frac{(\bar{x}_i^A - \bar{x}_i)^2 + (\bar{x}_i^B - \bar{x}_i)^2}{\frac{1}{N_A - 1}\sum_{k=1}^{N_A}(x_{k,i}^A - \bar{x}_i^A)^2 + \frac{1}{N_B - 1}\sum_{k=1}^{N_B}(x_{k,i}^B - \bar{x}_i^B)^2} \quad (2)$$

여기서 $\bar{x}_i, \bar{x}_i^A, \bar{x}_i^B$ 는 각각 전체 학습벡터, 클래스 A, 클래스 B에 대한 i 번째 특징 파라미터의 평균이며, $x_{k,i}^A$ 는 클래스 A의 k 번째 학습 벡터에서 i 번째 특징 파라미터를 의미한다. 분자항은 두 클래스 간의 변별력을 나타내며, 분모항은 각각의 클래스 내에서의 변별력을 의미한다. 본 연구에서 사용한 F-score기반의 특징 선택 기법은 먼저 전체 특징 파라미터를 F-score가 높은 특징 파라미터부터 순차적으로 정렬하고, 변별력이 높은 특징 파라미터로부터 순서대로 $(n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \cdots, 1)$개의 특징 파라미터 집합을 선택한다. 다음으로 학습자료를 무작위로 재분할하여 5종류의 학습 자료와 검증(validation)자료를 얻은 후, 각각의 특징 파라미터 집합에 대해 SVM을 학습하고 검증자료에 대해 감정식별을 수행하여 성능의 평균을 구한다. 평균 식별성능이 최고인 특징 파라미터 집합을 최종적으로 선택한다.

IV. 확률 출력 SVM

본 논문에서는 감정식별을 위해 SVM을 사용하였다. SVM은 학습벡터들을 고차원 공간으로 사상시킨 후, 최대의 마진 (margin)을 갖는 분리평면 (hyperplane)을 구하는 것이다. HMM (hidden Markov model)은 수십 밀리 초의 단구간 프레임에서 추출한 단구간 특징 (short term feature)을 사용하므로 평균 피치 등의 장구간 특징 (long term feature)을 적용하기 어려운 반면, SVM은 전체 음성구간의 평균 피치, 에너지 표준편차와 같이 감정식별에서 빈번히 사용되는 장구간 특징을 다룰 수 있다.

두 개의 클래스에 대한 학습벡터 $\mathbf{x}_k \in R^n, k = 1, \cdots, m$ 와 각 벡터의 클래스를 나타내는 벡터 $\mathbf{y} \in R^m, y_k \in \{1, -1\}$ 가 주어졌을 때, SVM은 다음과 같은 최적화 문제를 푼다 [8].

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{k=1}^{m}\xi_k, \\ &s.t. \quad y_k(\mathbf{w}^T(\phi(\mathbf{x}_k) + b) \geq 1 - \xi_k, \\ &\qquad \xi_k \geq 0, k = 1, \cdots, m \end{aligned} \quad (3)$$

학습벡터는 ϕ 에 의해 고차원 공간으로 사상되며, C 는 학습오류에 대한 손실 (penalty) 파라미터이다. 계산된 지지벡터 (support vector)를 이용한 SVM의 출력은 다음 식과 같이 나타낼 수 있다.

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (4)$$

위 식에서 $\mathbf{x}$ 는 입력 특징, l 은 지지벡터의 수를 나타내며, K 는 커널함수 (kernel function)로서 $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}')$ 이다. 일반적인 SVM의 출력은 식 (4)의 결과에 따른 -1 또는 1로서, 두 클래스 중 하나로 결정된다. 세 개 이상의 클래스에 대해 SVM을 이용하여 분류기를 설계하는 경우에는 모든 클래스의 쌍 (pair)에 대해 SVM 학습 및 인식을 수행한 후, 다수결 (majority vote)등에 의해 입력에 해당하는 클래스 ID (identifier)를 출력으로 얻는다. 지금까지 기술한 일반적인 SVM분류방식은 입력에 대해 단지 클래스 ID를 얻는 방식이지만 결과로 클래스 ID 뿐만 아니라 출력확률을 얻을 수 있다면 패턴분류 결과에 대한 신뢰도 분석 등 다양한 응용이 가능하다.

출력결과에 대한 확률정보를 얻기 위해서 시그모이드 함수 (sigmoid function)를 이용하여 식 (4)의 출력값을 확률값으로 변환하는 방법이 제안되었다 [8][9]. 식 (4)에서 i 번째 특징에 대한 출력치 $f(\mathbf{x}_i)$를 f_i로 표기하고, $t_i = (1 + f_i)/2$ 라고 할 때, 다음 식에 의해 확률값으로 변환한다.

$$g_i = \frac{1}{1 + \exp(A \cdot t_i + B)} \quad (5)$$

이 식에서 A 와 B 의 값은 학습자료에 대한 최대우도 추정 (maximum likelihood estimation)에 의해 구해진다.

지금까지 기술한 내용에서는 두 개의 클래스를 분류하는 문제에 대한 확률출력 SVM을 다루었으나, 세 개 이상의 클래스에 대해 적용하기 위해 이를 확장할

필요가 있다 [9]. 한 쌍의 클래스들에 대한 확률값 $\mu_{ij} = p(y = i \mid y = i \text{ or } j, \mathbf{x})$ 의 추정치 γ_{ij} 가 주어졌고 γ_{ij} 는 이진 패턴분류기 (binary classifier)로부터 구했다고 할 때, 클래스 i 에 대한 확률값 $p_i = p(y = i \mid \mathbf{x}), i = 1, \cdots, k$ 는 다음의 두 식 (6)과 (7)에 의해 식 (8)과 같이 구할 수 있다.

$$\left(\sum_{j:j\neq i} p(y = i \text{ or } j \mid \mathbf{x})\right) - (k-2)p(y = i \mid \mathbf{x}) = \sum_{j=1}^{k} p(y = i \mid \mathbf{x}) = 1 \quad (6)$$

$$\gamma_{ij} \approx \mu_{ij} = \frac{p(y = i \mid \mathbf{x})}{p(y = i \text{ or } j \mid \mathbf{x})} \quad (7)$$

$$p_i \approx \frac{1}{\sum_{j:j\neq i} 1/\gamma_{ij} - (k-2)} \quad (8)$$

위 식 (8)에서 얻게 되는 확률값은 SVM의 클래스별 최종 출력확률이며, 본 연구에서는 이 확률정보를 다음 장에서 기술하는 바와 같이 감정검출에 응용하였다.

V. 감정검출 기법

감정검출은 사용자가 관심을 갖고 있는 감정 e 에 대하여 시스템에 입력된 음성신호가 e 를 포함하고 있는지의 여부를 결정하는 문제이다. 본 논문에서는 가설검정 (hypothesis test)에 기반한 몇 가지 검증방법을 적용해 본다[7]. 주어진 입력신호로부터 추출한 특징벡터를 $\mathbf{x}$, 검출하고자 하는 감정 클래스를 C_e, 클래스 C_e 의 여집합을 $C_{\bar{e}}$ 라고 할 때, 대수 우도비를 이용한 검증함수 (verification function) $V(\mathbf{x})$ 는 식 (9)와 같이 정의될 수 있다.

$$V(\mathbf{x}) = \log\left(\frac{\Pr(\mathbf{x} \mid C_e)}{\Pr(\mathbf{x} \mid C_{\bar{e}})}\right) \quad (9)$$

위 식에서 분모 및 분자항의 확률값은 입력 $\mathbf{x}$ 에 대한 HMM 출력확률 또는 SVM 출력확률 등에 의해 근사할 수 있다. 벡터 $\mathbf{x}$ 가 감정 클래스 C_e 에서 발생한 경우, 즉, 입력 신호가 검출하고자 하는 감정에 해당할 경우에는 분자항의 값이 커지므로 $V(\mathbf{x})$ 가 큰 값을 갖게 되며, 그 외의 경우는 $V(\mathbf{x})$ 가 작은 값을 갖는다.

제 6장의 그림 2에서 실선은 화난 감정에 대한 발성자료로부터 V 값들을 계산하여 확률밀도 함수로 나타낸 것이며, 점선은 그 이외의 발성자료로부터 구한 V 값들의 확률밀도 함수이다. 그림에서 보듯이 화난 감정에 대한 V 값은 그 외의 발성자료에 비해 더 높은 값을 가진다. 따라서, 임의의 입력음성에서 추출한 특징벡터 $\mathbf{x}$ 에 대해 $V(\mathbf{x})$ 값을 계산한 후, 이 값을 미리 정해둔 임계치 θ 와 비교함으로써 입력음성을 감정 e 로서 수락 또는 거절할 것인지의 여부를 결정할 수 있다. 임계치의 값이 높을수록 FA (false acceptance)가 줄어드는 반면, FR (false rejection)이 높아진다.

식 (9)에서 $C_{\bar{e}}$ 에 해당하는 모든 특징벡터들을 수집하기는 불가능하므로, 분모항 $\Pr(\mathbf{x} \mid C_{\bar{e}})$ 를 한정된 데이터를 사용하여 근사해야 한다. 본 논문에서는 세 가지 방법을 제안하여 비교해 본다. 첫 번째 방법은 식 (10)과 같이 여집합 $C_{\bar{e}}$ 이 중립감정 클래스 C_n 에 의해 근사 될 수 있다고 가정한다. 본 논문에서는 이 방법을 AppxNeut로 표기한다.

$$V(\mathbf{x}) = \log\left(\frac{\Pr(\mathbf{x} \mid C_e)}{\Pr(\mathbf{x} \mid C_{\bar{e}})}\right) \approx \log\left(\frac{\Pr(\mathbf{x} \mid C_e)}{\Pr(\mathbf{x} \mid C_n)}\right) \quad (10)$$

두 번째 방식은 클래스 $C_{\bar{e}}$ 에 해당하는 가능한 모든 발성자료를 수집한 뒤, 이를 이용하여 확률출력 SVM 모델 파라미터를 학습하는 것이다. 본 논문에서는 이 방법을 AppxCmpl로 기술하기로 한다. 세 번째 방식은 발성검증 (utterance verification)에서 제안된 방식으로 N 개의 감정들에 대해 발성자료를 수집하고, 각각에 대해 확률모델을 학습한 뒤 다음 식 (11)과 같이 $\Pr(\mathbf{x} \mid C_{\bar{e}})$ 를 근사하는 방법이다.

$$\begin{aligned} V(\mathbf{x}) &= \log \Pr(\mathbf{x} \mid C_e) - \log \Pr(\mathbf{x} \mid C_{\bar{e}}) \\ &\approx \log \Pr(\mathbf{x} \mid C_e) \\ &\quad - \log\left[\frac{1}{N}\sum_{i\neq e} \exp(\gamma \cdot \log \Pr(\mathbf{x} \mid C_i))\right]^{1/\gamma} \end{aligned} \quad (11)$$

식 (11)에서 $\gamma > 0$ 이며, 실험적으로 결정된다. 본 논문에서는 이 방법을 AppxVrfy로 기술하기로 한다.

VI. 실험 및 결과

감정식별 및 감정검출의 성능을 알아보기 위해 본 논문에서는 특징선택 실험, 기존 SVM과 확률출력 SVM의 성능 비교 실험, 그리고 감정검출에 대한 실험을 수행하였다. 본 장에서는 먼저 실험에 사용한 데이터베이스를 살펴보고, 수행한 실험 및 결과를 기술하기로 한다.

1. 데이터베이스

본 연구에서는 실험을 위해 SUSAS (speech under simulated and actual stress) 데이터베이스를 사용하였다. 이 데이터베이스는 스트레스 상황에서 음성인식의 성능개선 연구를 위해 구축되었으나, 화난 음성, 롬바드 (Lombard) 음성, 빠른 발성, 느린 발성, 중립음성 등 감정식별 태스크와 흡사한 발성 변이들을 포함하고 있어 감정식별 연구에도 활용될 수 있다 [4]. 8kHz로 샘플링된 이 데이터베이스는 9명의 화자가 11 종류의 다른 발성 스타일로 35 단어를 발성하였다. 본 연구에서는 SUSAS 데이터베이스 중에서 화남 (anger), 의문 (question), 빠름 (fast), 중립 (neutral) 및 느림 (slow) 등 다섯 종류의 발성 스타일에 대해 540개의 단어 발성음을 이용하여 감정 식별기를 학습하였다. 또한, 평가를 위해서는 동일한 종류의 발성 스타일에 대해 2430개의 음성 자료를 사용하였다.

2. 감정식별 및 특징선택 실험

첫 번째 실험에서는 2장에서 기술한 42차의 특징벡터를 이용하여 감정식별을 수행하였으며, 식별 정확도는 65.7%를 보였다. 표 2에서 이 결과를 혼동행렬 (confusion matrix)로 표시하였다. 실험에 사용한 평가자료에 대해 본 논문의 시스템은 의문 (Question)과 화남 (Angry)을 가장 잘 식별하였으며 중립 (neutral) 감정의 경우, 빠른 발성 (Fast)과 혼동하기 쉬웠다.

표 2. 42차의 특징벡터를 사용한 경우, 5종류의 감정에 대한 혼동행렬
Table 2. A confusion matrix for five-emotion identification using feature vectors of 42-dimension.

	Question	Fast	Angry	Neutral	Slow
Question	0.85	0.04	0.03	0.05	0.01
Fast	0.01	0.73	0.05	0.17	0.04
Angry	0.03	0.09	0.84	0.02	0.01
Neutral	0.07	0.38	0.05	0.36	0.13
Slow	0.03	0.18	0.04	0.24	0.50

두 번째 실험은 F-score에 의한 특징선택 기법을 적용하여 식별성능을 향상시킴과 동시에 특징차원을 줄이고자 하였다. 학습자료에 대해 F-score를 적용한 결과 21차가 최적으로 나타났으며, 이 때의 식별 정확도는 67.8%로서 첫 번째 실험에 비해 절반의 특징을 사용하면서도 성능이 향상되었다. 선택된 21차의 특징은 표 1에서 중요도가 가장 높은 특징으로부터 평균 차분 피치 (12), 피치의 표준편차 (28), 차분 피치의 범위 (13), 차분 피치의 최대값(16), 피치의 범위 (7), 차차분 피치의 범위 (26), 피치의 최대값 (5), 피치의 평균 (1), 차차분 피치의 최소 (22), MFCC 1차 계수 평균 (31), 에너지 최대값 (6), 차차분 피치의 최대값 (24), 차분 피치의 최소 (14), 차차분 에너지의 평균 (21), MFCC 8차 계수 평균 (38), MFCC 7차 계수 평균 (37), 차분 에너지의 범위 (19), 차분 에너지의 최소 (15), 차차분 피치의 평균 (20), 피치의 최소 (3), 차분 에너지의 평균 (13)이다. 이 실험 결과에서 피치 및 에너지 관련 정보가 대부분을 차지하여 감정 식별 태스크에서 이들 정보가 매우 효과적임을 알 수 있다.

표 3은 이 경우의 혼동행렬을 나타낸다. 표 2와 비교하면 화남 (Angry)과 느린 발성 (Slow)에 대한 식별능력이 보다 향상되었음을 알 수 있다.

표 3. 선택된 21차의 특징벡터를 사용한 경우, 5종류의 감정에 대한 혼동행렬
Table 3. A confusion matrix for five-emotion identification using selected feature vectors of 21-dimension.

	Question	Fast	Angry	Neutral	Slow
Question	0.84	0.03	0.06	0.06	0.01
Fast	0.01	0.71	0.07	0.16	0.05
Angry	0.01	0.08	0.87	0.02	0.00
Neutral	0.06	0.33	0.08	0.35	0.17
Slow	0.03	0.16	0.06	0.15	0.60

마지막으로 두 번째 실험과 동일한 21차의 특징 파라미터를 사용하고 확률출력 SVM에 의해 인식한 결과는 67%였으며, 앞의 실험결과와 비교할 때 구현한 확률출력 SVM 모델이 기존 SVM을 잘 근사함을 알 수 있다.

3. 감정검출 실험

본 연구에서는 감정검출의 대상으로 화난 감정을 사용하였으며, 실험에는 앞 절에서 기술한 바와 동일한 학습자료를 사용하고, 앞 절의 평가자료 중 1890개를 감정검출의 평가자료로 사용하였다.

그림 2는 평가자료에 대하여 제 5장에서 기술한 AppxCmpl 근사방법에 의해 구한 화난 음성과 여집합 감정의 대수 우도비 (LLR) 값의 분포를 나타낸다. 그림에서 두 클래스에 대한 분포가 비교적 명확히 구분되며, 대수 우도비 값이 -2와 2 사이에서는 두 분포가 겹쳐있어 이 영역을 최소화할 필요가 있음을 알 수 있다.

그림 3은 X축 상에서 임계값을 변경해가면서 구한 제 5장의 세 가지 대수 우도비 근사법에 대한 ROC (receiver operating characteristic) 곡선이다. 그림에 AppxCmpl 근사방식이 EER (equal error rate)이 가장 낮아서 비교한 방법들 중에 제일 좋은 성능을 보임을 알 수 있다. AppxCmpl 근사방식은 EER에서 약 92%의 정확도 (correctness)를 보였고, AppxNeut 근사방식은 약 86%, AppxVrfy 근사방식은 약 88%의 정확도를 나타냈다. 여기에서 정확도란, 화난 감정으로 승인 (accept)된 발성들 중에 실제 화난 감정 발성의 비율을 뜻한다.

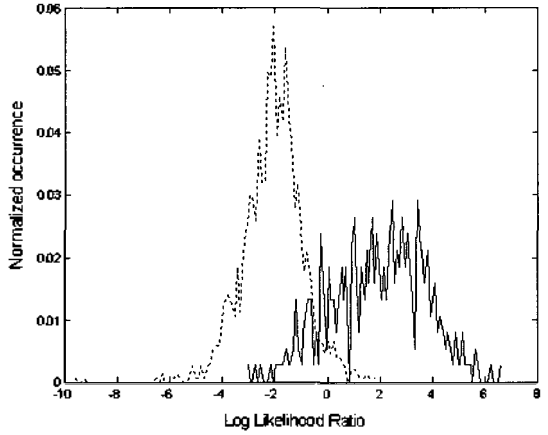


그림 2. 화난 감정의 여집합 클래스에 대한 화난 감정 클래스의 대수 우도비 (log likelihood ratio: LLR) 분포: 실선은 화난 음성 자료에 대한 LLR의 분포이며, 점선은 화난 감정 이외의 음성 자료에 대한 LLR의 분포를 나타낸다.

Figure 2. Distributions of the log-likelihood ratio (LLR) of angry emotion class to its complementary class: the solid line is LLR distribution for angry emotion data and the dotted line is for all other emotion data.

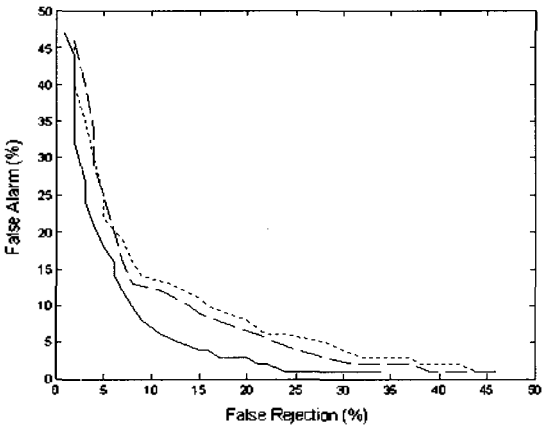


그림 3. 세 가지 종류의 LLR 근사방법에 대한 ROC 곡선: 실선은 AppxCmpl 근사법의 결과이고, 점선은 AppxNeut 근사법의 결과이며, 파선 (dashed line)은 AppxVrfy 근사법에 의한 결과를 나타낸다.

Figure 3. ROC curves for three LLR approximation methods: solid line is for the AppxCmpl method, dotted line for the AppxNeut, and dashed line for the AppxVrfy method.

마지막으로, 앞 실험의 1890개 평가자료에 SUSAS 데이터베이스의 SOFT와 LOMBARD 음성자료 756개를 추가한 총 2646개 평가자료에 대해 앞 실험에서 얻어진 세 방식의 EER에서의 임계값을 적용하여 감정 검출 실험을 수행하였다. 실험 결과 AppxCmpl의 경우, 91.3%의 정확도에 FA는 18.3%, FR은 8.7%였고, AppxNeut는 86.5%의 정확도에 FA는 15.5%, FR은 13.5%였다. AppxVrfy는 88.1%의 정확도에 FA는 7.3%, FR은 11.9%였다.

VII. 결 론

본 논문에서는 입력 음성에 포함된 감정을 식별하거나, 특정 감정을 검출하는 방법에 대하여 다루었다. 이를 위해 확률 출력 SVM을 도입하고, 대수 우도비에 기반한 세 가지 검증 함수의 성능을 비교하였다. SUSAS 데이터베이스에 대한 실험결과, 5가지 종류의 발성 스타일에 대해 65.7%의 성능을 보였으며, 특징 선택 기법을 적용하여 특징 차수를 줄임과 동시에 성능을 더 향상시킬 수 있었다. 또한, 화난 감정의 검출을 위해 가설 검정 기법을 응용하였으며, 여집합 클래스의 우도값 근사를 위한 세 가지 방법을 비교하였다. 화난 감정을 제외한 모든 발성 자료로 여집합 클래스의 SVM을 학습한 경우에 검출 능력이 최대였고, 91.3%의 검출 정확도를 얻을 수 있었다.

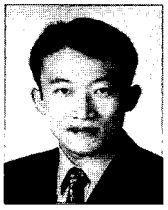
참 고 문 헌

1. K. Scherer, "Vocal communication of emotion: A review of research paradigms", Speech Communications, 40 227-256, 2003.
2. N. Amir, S. Ziv, and R. Cohen, "Characteristics of authentic anger in Hebrew speech", In Proc. Eurospeech, 713-716, 2003.
3. S. Yacoub, S. Skimske, and J. Burns, "Recognition of emotions in interactive voice response systems", In Proc. Eurospeech, 729-732, 2003.

4. O.-W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals", In Proc. Eurospeech, 125-128, 2003.
5. I. Guyon and A. Elisseef, "An introduction to variable and feature selection", Journal of Machine Learning Research, 3 1157-1182, 2003.
6. A. Rakotomamonjy, "Variable selection using SVM based criteria", Journal of Machine Learning Research, 3 1357-1370, 2003.
7. R. A. Sukkar and C. H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", IEEE Speech and Audio Processing, 4(6) 420-429, 1996.
8. J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods", Advances in Large Margin Classifiers, MIT Press, 2000.
9. D. Price, S. Knerr, L. Personnaz, and G. DreyFus, "Pairwise neural network classifiers with probabilistic outputs", Advances in Neural Information Processing Systems, 1109-1116, 1995.
10. R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition", In Proc. of International Conference on Spoken Language Processing, 2029-2032, 2002.
11. T. Robinson, "Cookbook: Speech processing functions in C", CD ROM: Prime Time Freeware for AI, Issue 1-1, 1994.

저자 약력

• 조 훈 영 (Hoon-Young Cho)



1991. 3 ~ 1995. 8 : KAIST 전자전산학과 학사
1996. 3 ~ 1998. 2 : KAIST 전자전산학과 석사
1998. 3 ~ 2003. 2 : KAIST 전자전산학과 박사
2003. 3 ~ 2003. 9 : KAIST 정보전자연구소, Post Doc.
2003. 10 ~ 2004. 9 : Univ. of California San Diego, 방문연구원
2004. 10 ~ 2006. 1 : LG 전자기술원, 모바일 멀티미디어 연구소 선임연구원
2006.2 ~ 현재 : 한국전자통신연구원 디지털콘텐츠연구단 선임연구원
※ 주관심분야 : 잡음에 강한 음성인식, 기계학습, 3차원 입체음향 신호처리, 3D 게임, 대용량 멀티미디어 검색

• 정 규 준 (Gue Jun Jung)

2000년 2월 : 경북대학교 컴퓨터공학과 (공학사)
2000년 3월 ~ 2002년 2월 : KAIST 전자전산학과 전산학전공 대학원 (공학석사)
2002년 3월 ~ 현재 : KAIST 전자전산학과 전산학전공 대학원 (박사과정)