한국음향학회지 제26권 제2호 pp. 75~79 (2007)

음성신호 기반의 성별인식을 위한 Support Vector Machines의 적용

Voice-Based Gender Identification Employing Support Vector Machines

이 계 환*, 강 상 익*, 김 덕 환*, 장 준 혁*
(Kye-Hwan Lee*, Sang-Ick Kang*, Deok-Hwan Kim*, Joon-Hyuk Chang*)
*인하대학교 전자전기공학부

(접수일자: 2006년 11월 22일, 수정일자: 2007년 1월 11일, 채택일자: 2007년 2월 5일)

본 논문은 SVM (Support Vector Machines)을 이용한 음성신호 기반의 효과적인 성별인식 시스템을 제안한다. 분별적 이진 (binary) 패턴 분류기인 SVM은 특징 공간에서 비선형 경계를 찾아 분류하는 방법으로 우수한 성능을 보인다고 알려져 있다. 연구에서는 기존의 성별인식에서 널리 쓰이고 있는 MFCC (Mel Frequency Cepstral Coefficients)를 사용하여 SVM과 기존의 GMM (Gaussian Mixture Model) 알고리즘의 성별인식 성능을 비교하였고, 특히, 보다 향상된 SVM의 성별인식을 위해 MFCC와 Pitch를 이용한 결합 특징 벡터를 적용하였다. 실험결과 MFCC 파라미터를 사용했을 때 제안된 SVM이 GMM보다 우수한 성별인식 성능을 보였고, 제안된 결합 특징 벡터를 사용 했을 때 우수한 성능을 보였다.

핵심용어: SUM, GMM, 피치, MFCC, 음성기반, 성별인식

투고분야: 음성처리분야 (2.4)

We propose an effective voice-based gender identification method using a support vector machine (SVM). The SVM is a binary classification algorithm that classifies two groups by finding the voluntary nonlinear boundary in a feature space and is known to yield high classification performance. In the present work, we compare the identification performance of the SVM with that of a Gaussian mixture model (GMM) using the mel frequency cepstral coefficients (MFCC). A novel means of incorporating a features fusion scheme based on a combination of the MFCC and pitch is proposed with the aim of improving the performance of gender identification using the SVM. Experiment results indicate that the gender identification performance using the SVM is significantly better than that of the GMM. Moreover, the performance is substantially improved when the proposed features fusion technique is applied.

Key words: SUM, GMM, Pitch, MFCC, Voice based, Gender identification

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

음성신호를 기반으로 한 성별인식은 자동음성인식, 멀티미디어 및 인간과 컴퓨터와의 상호작용 (HCI, Human Computer Interaction) 등의 성능을 좌우하는 중요한 문제로 다루어져 왔다 [1, 2]. HCI를 위해서는 컴퓨터가 사람의 행동, 특히 음성신호를 제대로 인식하고 반응 하는 것이 필수적이다. 현재까지 효과적인 HCI를 위한 성별 인식에 관한 연구가 활발히 진행되어져 왔으며, 많은 부분의 연구가 필요하다. 일반적으로 성별 인식에 관한 연구는 HMM (Hidden Markov Model) 이나 GMM (Gaussian Mixture Model)과 같은 경험적 위험을 최소화하는 방법에 기초하고 있는 것들이 대부분이다 [3, 4].

본 논문에서는 고차원 공간으로의 확장을 통한 선형 패턴 분류에 있어서 좋은 성능을 보인 SVM (Support Vector Machines)을 성별인식에서 우수한 성능을 보인 GMM 알고리즘과 비교함으로써 음성신호 기반의 효과적

책임저자: 장 준 혁 (changjh@inha.ac.kr)
402-751 인천시 남구 용현동 253 인하대학교 전자전기공학부
(전화: 032-860-7423; 팩스: 032-868-3654)

인 성별인식 시스템을 제안 한다 [5]. 또한 정확한 성별 인식을 위해 음성신호의 효과적인 특징 추출 (Feature Extraction)은 인식 성능을 좌우하는 중요한 문제이다. 관련하여 본 논문에서는 효과적인 성별인식 을 위해 MFCC (Mel Frequency Cepstral Coefficients)를 특징으로 선택하여 실험하였다. 그리고 SVM의 성별 인식 성능의 향상을 위해 Pitch를 특징 벡터로 추출하고 기존의 MFCC와 결합하여 성별 인식 실험을 하였다 [6].

본 논문의 구성으로, Ⅱ장에서는 음성의 특징 추출과 인식 알고리즘에 대해 기술하고, Ⅲ장에서는 실험 결과 비교 및 분석에 대해 기술하였으며, Ⅳ장에서는 관련하여 결론을 맺는다.

Ⅱ. 음성의 특징 추출과 인식 알고리즘

2.1. GMM (Gaussian Mixture Model)의 이해

GMM은 EM (Expectation Maximization) 알고리즘에 기반을 둔 패턴 분류기이다 [7]. 상태 열 N개의 특징 벡터를 $X=\{x_1,...,x_N\}$, $x_t \subseteq R^d$라 하면, 우도 (likelihood)는 다음과 같이 주어진다.

$$p(x_t|\lambda)=\sum_{i=1}^{M} p_i b_i(x_t) \tag{1}$$

$$b_i(x_t)=\frac{1}{(2\pi)^{\frac{D}{2}}\left|\sum_i\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x_t-\mu_i)^T(\sum_i)^{-1}(x_t-\mu_i)\right\}. \tag{2}$$

여기서 GMM 모델을 위한 파라미터는 가우시안 혼합 성분 밀도의 가중치 (mixture weight : p_i), 평균 벡터 (mean vector : $\vec{\mu_i}$), 공분산 행렬 (covariance matrix : $\sum i$)로 아래와 같이 구성된다.

$$\lambda=\{p_i, \vec{\mu_i}, \sum i\} \quad i=1,...,M \tag{3}$$

본 논문에서는 λ 추정을 위해, $P(X|\lambda') \geq P(X|\lambda)$가 되는 새로운 모델 λ'을 정해진 문턱값 (threshold)에 도달할 때까지 EM 알고리즘을 사용하여 i번째 사후확률을 구한다 [7]. 그리고, 구한 사후확률 중 가장 큰 확률을 가진 성별 인식 모델은 아래와 같이 주어진다.

$$\hat{G}=\arg\max_{1\leq g\leq S}\sum_{t=1}^{T}\log p(\vec{x_t}|\lambda_g). \tag{4}$$
$$S=2,\ (1:\ male,\ 2:\ female\ )$$

2.2. SVM (Support Vector Machines)의 이해

SVM은 SRM (Structural Risk Minimization) 이론으로부터 발전한 이진 패턴 분류기이다 [8]. 선형 SVM에 있어서 두 개의 클래스를 구분할 수 있는 초평면 (Hyperplane)은 무수히 많으나 그림 1에서 두 클래스 간 가장 가까운 점들의 거리 ρ (margin)를 최대화하도록 하면 유일한 해로 초평면을 구할 수 있다.

일반적으로 ρ를 최대화하는 초평면의 방정식은 최적의 가중벡터 w_*와 바이어스 b_*로 아래와 같이 표현된다.

$$w_*^T \cdot x_i + b_* = 0 \tag{5}$$

$$\rho = 2/\|w\|. \tag{6}$$

이때 거리 ρ를 최대화하기 위해서는 아래의 식(7)을 최소화하면서 식 (8)를 만족해야한다.

$$\Phi(w)=1/2\, w^T w \tag{7}$$

$$d_i(w^T x_i + b) \geq 1 \quad \text{for } i=1,2...,N. \tag{8}$$

식 (7)을 최소화하기 위해 Lagrangian의 안장점을 찾는 문제와 KKT (Karush-Kuhn-Tucker) 조건을 이용하여 Lagrange Multiplier를 찾는 Wolfe dual problem으로 바꿀 수 있다. 바뀐 식을 최대화하는 값을 가지고 최적 가중벡터 w_*와 바이어스 b_*를 구할 수 있다 [8]. 구해진 최적가중벡터와 바이어스에 따라 임의의 입력패턴 x는 아래와 같이 분류 된다 [9].

$$f(x)=\text{sign}(w_*^T x + b_*). \tag{9}$$

보통의 입력 패턴의 경우 아래의 그림 2와 같이 명확하게 선형분리가 되지 않는 경우가 대부분이다.

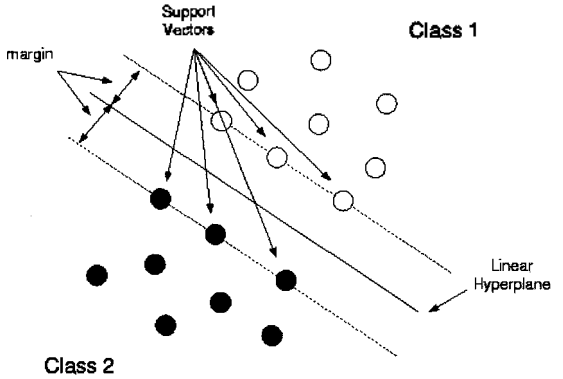


그림 1. 선형분리를 이용한 SVM
Fig. 1. linear separation using SVM.

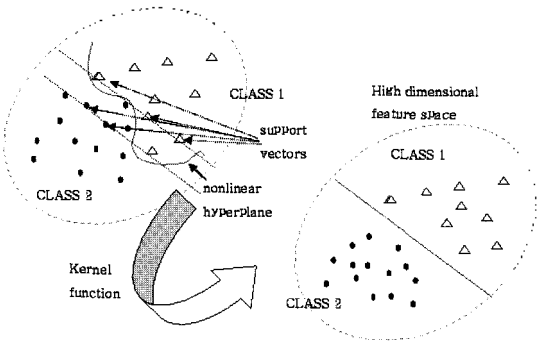


그림 2. 커널 함수에 기반 한 고차원 공간으로의 확장
Fig. 2. Extension of high dimensional feature space based on Kernel function.

그림 2처럼 입력 패턴의 선형 분리가 불가능한 경우 비선형 특성을 가진 SVM을 사용한다. 비선형 SVM은 커널 (kernel) 함수를 사용하여 그림 2와 같이 선형 분류가 가능한 고차원 공간으로 확장된 특징 공간을 가지고 ρ을 최대화 하는 값을 찾는다 [8, 9].

고차원의 공간으로 확장시킬 경우 어느 정도는 원 공간에서의 거리 관계를 보존 시킬 필요가 있기 때문에 커널 함수는 고차원 공간으로의 사상 함수 $\phi(x)$를 사용해 아래와 같이 정의 한다.

$$K(x,x') = \phi(x)^T \phi(x'). \tag{10}$$

여기서 중요한 점은 커널 트릭 (kernel trick)을 사용하여 사상 함수에 대한 구체적인 설정 없이도 분류함수를 구현 할 수 있다는 것이다. 본 논문에서는 다음과 같은 RBF (Radial-Basis Function) 커널 함수를 사용하였다 [10, 11].

$$K(x,x_i) = \exp(-\frac{1}{2\sigma^2} \| x - x_i \|^2). \tag{11}$$

그리고 커널 함수를 사용해서 선형 SVM과 마찬가지의 방법으로 w_*, b_*를 구할 수 있으며, 결론적으로 비선형 SVM은 다음과 같이 분류 된다.

$$f(x) = \text{sign}(w_*^T K(x_i,x) + b_*). \tag{12}$$

2.3. 성별 인식을 위한 특징 벡터 추출

효과적인 음성신호의 정보를 얻기 위해서는 효율적인 특징 벡터를 추출 하는 것이 중요하다. 특히, 성별 인식 성능 향상을 위해서는 효과적인 특징 벡터를 제시된 SVM과 기존의 GMM에 이용하는 것이 바람직하다. 본 논문에서는 전체적으로 음성파일에 8 kHz의 샘플링 주파수를 적용하였고 음성신호의 정보를 얻기 위해서 AURORA2를 이용하여 MFCC 계수 13개와 Δ MFCC 계수 13개를 추출 한 후 음성검출기 (VAD, Voice Activity Detection)를 사용해 음성구간의 정보를 추출하였다 [12]. 추출한 MFCC 계수는 25 ms의 Hamming 윈도우를 15 ms씩 이동하면서 계수를 추출하였다.

한편, 음성신호를 이용하여 성별을 결정지을 수 있는 일반적으로 알려진 가장 우수한 특징은 Pitch이다 [1]. 실제로 Pitch는 남자 50 kHz에서 250 kHz, 여자는 120 kHz에서 500 kHz의 분포를 보인다 [13]. 그림 3은 실험에서 사용한 남, 녀 파일에서 추출한 Pitch를 대비하여 보여준다. 따라서 기존의 음성 인식 시스템에서 가장 널리 고려되는 MFCC 외에 성별 인식 성능 향상을 위해 Pitch를 추출하여 기존의 MFCC와 결합하여 결합 특징

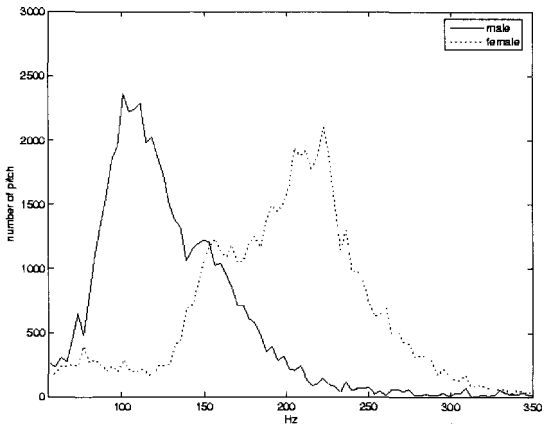


그림 3. 성별에 따른 Pitch의 분포
Fig. 3. Pitch distribution according to gender.

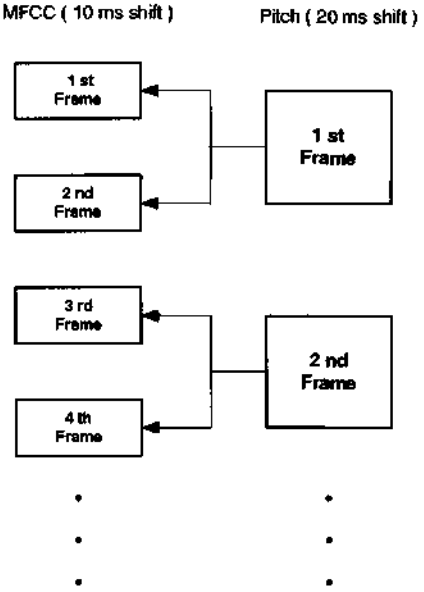


그림 4. MFCC와 Pitch의 결합 방법
Fig. 4. Fusion method for combine MFCC with Pitch.

벡터를 구성하였다.

실제로 결합 특징 벡터를 구성할 때, 3GPP2 SMV (Selectable Mode Vocoder)를 이용해 추출한 Pitch의 프레임 길이는 20 ms이고, MFCC의 프레임의 길이는 10 ms이기 때문에 결합 특징 벡터를 구성하기 위해 그림 4와 같은 방법을 사용하였다.

III. 실험 결과 비교 및 분석

실험에 쓰인 남, 녀 음성 파일은 OGI database를 사용 하였다 [14]. 각각의 파일은 약 5 sec 정도의 전화 음성 신호이며, 한사람이 여러 가지 문장과 단어를 영어로 읽는 정보를 담고 있다. Training은 남, 녀 한명 당 10개의 파일을 선택해 각각 10명씩 구성했으며, Test는 남, 녀 각각 1000개의 파일을 사용 하였다.

그림 5는 SVM과 GMM에 대한 성별 검출확률 (P_d)을 이용한 ROC (Receiver Operating Characteristic) 곡선을 이용하여 보여주고 있다. SVM의 경우 식 (12)의 바이어스값 b_*을 변화시키면서 인식 성능을 비교 했으며 [15], GMM의 경우 식 (4)의 구해진 사후확률과 비교되는 문턱 값을 변화함으로써 인식 성능을 비교하였다. 사용된 GMM은 16개의 Mixture를 사용하였으며, SVM에 사용된 특징 벡터들은 모두 mean과 variance로 정규화 하였다.

실험결과 MFCC를 특성을 사용한 경우 대부분의 구간에서 GMM보다 제안된 SVM이 우수한 성별 인식 성능을 보였다. 제안된 SVM을 이용하여 MFCC와 Pitch의 인식 성능을 비교하였을 때, MFCC를 사용하면 남성 인식 성능이 우수하고 Pitch를 사용하면 여성 인식 성능이 우수함을 알 수 있었다. 그리고 SVM에 제안된 결합 특징 벡터를 사용한 경우 MFCC와 Pitch를 단독으로 사용할 때보다 우수한 성능 인식을 보였다. 특히 제시된 결합 특징 벡터를 이용하는 경우 기존의 GMM과 SVM의 성능이 매우 우수하면서도 비슷하게 나타났는데, 이것은 Pitch가 단독으로 사용될 때는 상대적으로 성능이 떨어진 점을 고려하면 Pitch 특징 벡터가 효과적으로 성능 보완 역할을 한다고 예측된다.

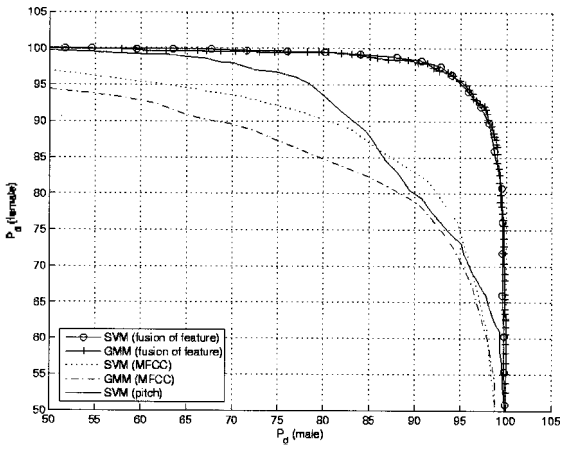


그림 5. ROC에 기반한 GMM과 SVM의 성별인식 성능 비교
Fig. 5. Gender identification performance of GMM and SVM based on ROC.

IV. 결 론

본 논문에서는 SVM을 이용한 성별 인식 기법을 제안하였고, 성별 인식 성능의 향상을 위해 MFCC와 Pitch를 이용한 결합 특징 벡터를 제시하였다. 기존의 성별 인식에서 널리 쓰이고 있는 MFCC 파라미터를 가지고 SVM과 기존의 GMM의 성별 인식 성능을 비교한 결과 대부분의 구간에서 새로이 제안된 SVM이 GMM보다 좋은 성능을 보였다. 특히, 제시된 결합 특징 벡터를 사용했을 경우, SVM과 GMM의 성별 인식 성능 향상을 보임으로써 Pitch를 이용한 결합 특징 벡터가 SVM과 GMM등의 패턴인식기를 이용하여 성별을 구분 지을 수 있는 우수한 특징임을 보였다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신 선도기반기술개발사업의 연구결과로 수행되었습니다.

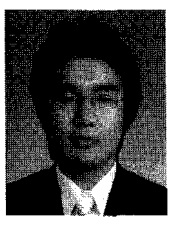
참 고 문 헌

1. H. Harb, L. Chen, "Voice-based gender identifivation in multimedia applications," Journal of Intelligent Information System, 24 179-198, May 2005.
2. M. Wald, "Using automatic speech recognition to enhance education for all student : Turning a vision into reality," 34th ASEE/IEEE "Frontiers in Education" Conference S3G, 22-25, Oct. 2004.
3. E. S. Parris and M. J Carey, "Language independent gender identification," 1996 International Conference on Acoustics, Speech and Signal Processing, 2 685-688, May 1996.
4. H. Harb and L. Chen, "Gender identification using a general

audio classifier," In Proceeding of IEEE 2003 International Conference, 2 733-736, July 2003.
5. S. Slomka, and S. Sridharan, "Automatic gender identification optimised for language independence," In Proceeding of IEEE TENCON - Speech and Image Technologies for Computing and Telecommunications, 1 145-148, Dec. 1997.
6. K. R. Farrell and R. J. Mammone, Data fusion techniques for speaker recognition, In R. V. Ramachandran and R. J. Mammone, editors, Modern Methods of Speech Processing, chapter 12 279-297, Kluwer Academic Publishers, Boston, Massachusetts, 1995.
7. G. Xuan, W. Zhang and P. Chai, "EM algorithms of gaussian mixture model and hidden markov model," In Proceeding of International Conference on Image Processing, 1 145-148, October 2001.
8. V. N Vapnic, "An overview of statistical learning theory," IEEE Transactions on Neural Networks, 10 (5) 988-999, Sept. 1999.
9. B. Boser, I. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," In Proceeding of 5th Annu. Wkshp. Comput. Learning Theory, Pittsburgh, PA : ACM, 144-152, 1992.
10. J. C. Palatt, *Advances in kernel methods - Support vector learning*, (MIT Press, February 1999)
11. J. Ramírez, P. Yélamos, J. M. Górriz, J. C Segura and L.García, "Speech/Non-speech discrimination combining advanced feature extraction and SVM learning," In Proceeding of the INTERSPEECH'2006 International Conference on Spoken Language Processing, 1662-1665, Pittsburgh, Sept. 2006.
12. N. S Kim and J. -H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Processing Letters, 7 (5) 108-110, May 2000.
13. W. C. Chu, *Speech coding algorithms : Foundation and evolution of standardized coders*, (John wiley & Sons, INC., May 2003) chapter 2, pp. 33-43
14. Y. K. Muthusamy, R. A. Cole and B. T. Oshika, "The OGI multi-language telephone speech corpus," In Proceeding of the 1992 International Conference on Spoken Language Processing, 2 895-898, October 1992.
15. J. Ramírez, P. Yélamos, J. M. Górriz and J. C Segura, "SVM-based speech endpoint detection using contextual speech features," IEE Electronics Letters, 42 (7) 426-428, Mar. 2006.

저자 약력

• **이 계 환 (Kye-Hwan Lee)**



2007년 2월: 인하대학교 전자공학과 학사
2007년 3월~현재: 인하대학교 전자공학과 석사과정

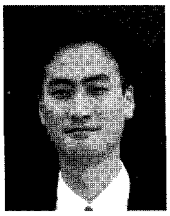
• **강 상 익 (Sang-Ick Kang)**

2007년 2월: 인하대학교 전자공학과 학사
2007년 3월~현재: 인하대학교 전자공학과 석사과정

• **김 덕 환 (Deok-Hwan Kim)**

1987년 2월: 서울대학교 계산통계학과 학사
1995년 8월: 한국과학기술원 컴퓨터공학과 석사
2003년 2월: 한국과학기술원 컴퓨터공학과 박사
1987년~1997년: LG전지 통신기기연구소 선임연구원
2006년 3월~현재: 인하대학교 전자전기공학부 부교수
※ 주관심분야: 임베디드시스템, 멀티미디어 정보검색, 데이터마이닝

• **장 준 혁 (Joon-Hyuk Chang)**



1998년 2월: 경북대학교 전자공학과 학사
2000년 2월: 서울대학교 전기공학부 석사
2004년 2월: 서울대학교 전기컴퓨터공학부 박사
2000년 3월~2005년 4월: (주)넷더스 연구소장
2004년 5월~2005년 4월: 캘리포니아 주립대학, 산타바바라 (UCSB) 박사후연구원
2005년 5월~2005년 8월: 한국과학기술연구원 (KIST) 연구원
2005년 9월~현재: 인하대학교 전자전기공학부 조교수
※ 주관심분야: 음성향상, 음성부호화, 음성인식, 적응신호처리, 오디오부호화, 멀티미디어시스템