SVM을 이용한 화자인증 시스템

최 우 용, 이 경 희, 정 용 화
한국전자통신연구원 정보보호연구본부
전화 : 042-860-1680 / 핸드폰 : 018-563-7242

Speaker Verification System Using Support Vector Machine

Woo-Yong Choi, Kyunghee Lee, Yongwha Chung
Information Security Research Division,
Electronics and Telecommunications Research Institute
E-mail : wychoi4@etri.re.kr

Abstract

There is a growing interest in speaker verification, which verifies someone by his/her voices. This paper explains the traditional text-dependent speaker verification algorithms, DTW and HMM. This paper also introduces SVM and how this can be applied to speaker verification system. Experiments were conducted with Korean database using these algorithms. The results of experiments indicated SVM is superior to other algorithms. The EER of SVM is only 0.5% while that of HMM is 5.4%.

I. 서론

최근 정보통신 기술이 급속도로 발전하고 인터넷의 이용이 확산됨에 따라 사용자 인증에 대한 관심이 높아지고 있다. 90년대까지 사용자 인증 수단으로 많이 사용되던 패스워드나 PIN (Personal Identification Number) 등은 타인에게 노출되거나 잊어버리는 등의 문제점을 가지고 있어 이를 대체하거나 보완하기 위한 방법으로 개인의 고유한 생체정보를 이용한 사용자 인증 방법에 관한 연구가 진행되고 있다. 이러한 생체인식 방법 중에서 사람의 음성을 이용하는 화자인식은 입력장치로 비교적 값이 싸고 손쉽게 구할 수 있는 마이크를 사용하며, 다른 생체인증방법에 비해서 사용자의 거부감이 적다는 장점이 있다. 전통적으로 많이 사용되어온 화자인증 방법으로는 DTW(Dynamic Time Warping)[1], HMM(Hidden Markov Model)[2], VQ(Vector Quantization)[3], GMM(Gaussian Mixture Model)[4] 등이 있다. DTW와 HMM은 주로 문맥종속 시스템에 많이 쓰이고, VQ와 GMM은 문맥독립 시스템에 많이 쓰이는 방법이다.

본 논문에서는 binary classifier로 최근 각광 받고 있는 SVM(Support Vector Machine)[5]을 화자인증에 적용하였다. 인증 실험을 위해서 다양한 종류의 문자열을 가진 데이터베이스를 이용하여 성능을 평가한 결과, DTW가 6.7%, HMM이 5.4%의 EER(Equal Error Rate)을 보인데 반해, SVM은 0.5%의 EER을 나타내었다.

본 논문의 구성은 2장에서 전통적인 문맥종속 화자인증 방법인 DTW와 HMM에 대해서 간단하게 설명하였고, 3장에서 SVM을 이용한 화자인증 방법을 설명하였다. 실험에 사용된 데이터베이스 및 실험 결과를 4장에서 기술하였으며, 마지막으로 5장에서 결론 및 향후과제를 제시하였다.

II. 화자인증 알고리즘

화자인증 시스템에는 인증하고자 하는 문자열의 고정여부에 따라 문맥종속과 문맥독립 시스템으로 나눌 수 있다. DTW와 HMM은 문맥종속 시스템에, VQ와 GMM은 문맥독립 시스템에 주로 사용되는 알고리즘이다. 본 장에서는 대표적인 문맥종속 화자인증 알고리즘인 DTW와 HMM에 대해서 알아본다.

2.1 DTW (Dynamic Time Warping)

문맥종속 화자인증 방법 중 전통적으로 가장 많이 사용된 방법은 template matching 방법이며[6], 이 중 DTW는 사람의 발성 속도의 차이를 보상해 주기 위해 가장 많이 사용되는 방법이다. Template은 특징벡터 $(\bar{\mathbf{x}}_1,\Lambda,\bar{\mathbf{x}}_N)$ 으로 이루어져 있는데, 이 template과 입력음성의 특징벡터 $(\mathbf{x}_1,\Lambda,\mathbf{x}_M)$ 과의 거리에 따라서 사용자의 인증 또는 거부가 결정된다. 일반적으로 M과 N이 같지 않기 때문에 이를 고려한 score를 식(1)과 같이 계산한다.

$$z=\sum_{i=1}^{M} d(\mathbf{x}_i,\bar{\mathbf{x}}_{j(i)}) \tag{1}$$

여기서 $j(i)$는 template의 index로 DTW 알고리즘에 의해서 계산되며, 거리 d는 식(2)에 의해서 계산된다.

$$d(\mathbf{x}_i,\bar{\mathbf{x}})=(\mathbf{x}_i,\bar{\mathbf{x}})^T\mathbf{W}(\mathbf{x}_i,\bar{\mathbf{x}}) \tag{2}$$

만약 $\mathbf{W}$가 항등행렬이면 두 벡터 사이의 거리는 Euclidean distance가 되고, $\mathbf{W}$가 공분산행렬의 역행렬이면 Mahalanobis distance가 된다.

입력음성과 그에 해당하는 template이 정해지면, DTW는 시간축에 대해서 piece-wise linear mapping을 수행함으로써 z가 최소가 되도록 두 신호를 정렬한 후, z 값이 임계치보다 작으면 사용자를 허용하고 그렇지 않으면 거부한다. 또한 warping path에 제약조건을 두어서 발음속도의 차이의 최대 허용치를 지정할 수도 있다.

2.2 HMM (Hidden Markov Model)

HMM은 대표적인 통계적 모델로 특징벡터가 확률분포를 가지는 random 벡터라고 가정한다. HMM은 통계적 방법을 통해서 특징벡터의 변화를 효과적으로 모델링하기 때문에, 일반적으로 DTW보다 높은 성능을 나타내는 반면에 계산량이 많다는 단점이 있다[7]. 본 논문에서는 N개의 상태를 가지며 skip path가 없는 left-to-right HMM을 사용하였다.

HMM은 상태천이확률 $\mathbf{A}$와 관측치의 확률밀도함수 $\mathbf{B}$, 그리고 초기상태확률 π로 구성되며, 식(3)과 같이 나타낼 수 있다.

$$\lambda=\{\mathbf{A},\mathbf{B},\pi\}=\{a_{i,j},b_i,\pi_i;\ i,j=1,\Lambda,N\} \tag{3}$$

관측치를 $\mathbf{O}=(\mathbf{o}_1,\Lambda,\mathbf{o}_T)$ 라고 두면, 시간 t에서의 j 번째 상태의 확률밀도는 다음과 같다.

$$b_j(t)=P(\mathbf{o}_t\mid q_t=j)=\sum_{m=1}^{M} c_{jm}N(\mathbf{o}_t;\mu_{jm},\mathbf{R}_{jm}) \tag{4}$$

여기서

$$N(\mathbf{o}_t;\mu_{jm},\mathbf{R}_{jm})=(2\pi)^{-d/2}\mid\mathbf{R}_{jm}\mid^{-1/2} \times\exp\left\{\frac{1}{2}(\mathbf{o}_t-\mu_{jm})^T\mathbf{R}_{jm}^{-1}(\mathbf{o}_t-\mu_{jm})\right\} \tag{5}$$

이고, M은 Gaussian mixture의 개수, μ_{jm}와 $\mathbf{R}_{jm}$은 각각 상태 j의 m번째 Gaussian의 평균과 공분산 행렬이며, d는 특징벡터의 차원이다.

이때, 모델 파라메타 $\{\mathbf{A},\mathbf{B},\pi\}$ 는 훈련데이터의 likelihood, $P(\mathbf{O}\mid\lambda)$를 최대로 하는 추정치로서, Baum-Welch 반복추정 알고리즘을 이용하여 구할 수 있으며, 그 반복 추정식은 식(6)-(8)과 같다[8].

$$\bar{\pi}_j=\gamma_1(i) \tag{6}$$

$$\bar{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{7}$$

$$\bar{b}_j(k)=\frac{\sum_{t=1\ s.t.\ \mathbf{o}_t=\mathbf{v}_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{8}$$

여기서

$$\gamma_t(i)=P(q_t=i\mid\mathbf{O},\lambda) \tag{9}$$

$$\xi_t(i,j)=P(q_t=i,q_{t+1}=j\mid\mathbf{O},\lambda) \tag{10}$$

이다.

III. SVM을 이용한 화자인증 시스템

SVM의 목적은 두 class를 분류하는 hyperplane을 설계하는 것으로, structural risk minimization 기법에 그 기초를 두고 있다[5]. 훈련데이터가 다음과 같이 주어졌다고 가정하자.

$$(\mathbf{x}_1, y_1), \Lambda, (\mathbf{x}_N, y_N) \in \Re^d \times \{\pm 1\} \tag{11}$$

여기서 $\mathbf{x}_i$ 는 input pattern이고, y_i 는 target output이다. 만약 두 class가 linearly separable하다면 식 (12)의 hyperplane에 의해서 두 class를 구분할 수 있다.

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{12}$$

이러한 hyperplane 중에서 hyperplane과 가장 가까운 데이터 포인트와의 거리를 최대로 하는 hyperplane을 optimal hyperplane이라고 하고, optimal hyperplane과 거리가 가장 가까운 데이터를 support vector라고 한다. SVM은 이러한 support vector를 이용하여 optimal hyperplane을 나타내는 방법으로 식(13)과 같은 constrained optimization 문제를 풀면 식(14)와 같은 해를 구할 수 있다[9].

$$\begin{aligned} &\min \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, \Lambda, N \end{aligned} \tag{13}$$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \tag{14}$$

여기서 α_i 는 Lagrange multiplier이다.

지금까지는 input space에서의 데이터들이 linearly separable한 경우에 대해서 살펴보았는데, input space에서의 데이터들은 linearly separable하지 않은 경우에는 input space에서는 문제를 풀 수가 없고, 더 높은 차원의 공간(feature space)으로 변환해서 문제를 해결해야 한다. 이때 사용하는 함수가 kernel 함수이다. Kernel 함수를 이용하면 식(15)와 같은 optimal hyperplane을 얻을 수 있다.

$$\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) = 0 \tag{15}$$

Kernel 함수에는 여러 가지가 있으나, SVM에서 많이 쓰이는 kernel 함수는 다음과 같다.

- Polynomial learning machine

$$k(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T\mathbf{x}_i + 1)^p \tag{16}$$

- RBF (Radial Basis Function) network

$$k(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right) \tag{17}$$

- Two-layer perceptron

$$k(\mathbf{x}, \mathbf{x}_i) = \tanh(\beta_0 \mathbf{x}^T\mathbf{x}_i + \beta_1) \tag{18}$$

IV. 데이터베이스 및 실험 결과

본 장에서는 DTW, HMM 및 SVM을 이용한 화자인증 시스템의 성능을 비교하였다. 실험에 사용된 데이터베이스는 사무실 환경에서 녹음된 음성으로 4연숫자, 고립단어 및 단문으로 구성되어 있다. 훈련용 데이터는 각 utterance를 27명(남성 17명, 여성 10명)이 각각 6회씩 발성한 음성데이터를 사용하였고, 인식용 데이터는 동일인이 각 utterance를 각각 6회씩 발성한 음성데이터를 사용하였다. 음성 특징 파라메타로는 12차 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하였다. 모든 음성데이터는 16kHz로 샘플링되었으며 16bit로 양자화하였다.

표 1에서 DTW, HMM 및 SVM의 화자인증 성능을 비교하였다. 그 결과 사용한 kernel에 관계없이 SVM이 가장 좋은 성능을 나타내었고, HMM, DTW 순이었다. 또한, kernel의 선택에 따라서도 인식성능에 영향을 미침을 알 수 있는데, RBF kernel을 사용하였을 때 0.5%로 가장 좋은 성능을 나타내었고, polynomial kernel을 사용하였을 때가 2.7%로 가장 성능이 떨어졌다. RBF kernel을 사용한 SVM의 EER은 0.5%로 HMM의 5.4%에 비해 91%의 에러율 감소를 나타내었다.

표 1. DTW, HMM 및 SVM의 성능 비교

알고리즘	EER (%)
DTW	6.7
HMM	5.4
SVM (RBF kernel)	0.5
SVM (perceptron kernel)	0.7
SVM (polynomial kernel)	2.7

V. 결론 및 향후과제

본 논문에서는 기존의 화자인증 방법인 DTW와 HMM에 대해서 기술하고, binary classifier인 SVM을 화자인증에 적용한 방법을 설명하였다. 또한 화자인증 실험을 통하여 SVM을 이용한 화자인증 방법이 기존의 방법보다 우수한 성능을 나타냄을 알 수 있었다. 특히 HMM이 5.4%의 EER을 나타낸 데 반해, RBF kernel을 사용하였을 경우 SVM의 EER은 0.5%로 91%의 에러율 감소를 나타내었다. 본 논문에서 사용한 데이터베이스는 잡음이 없는 사무실 환경에서 녹음한 음성을 사용하였는데, 잡음 환경에서의 인증실험을 통하여 SVM의 성능을 알아보고 그 성능을 높이는 연구가 진행되어야 하겠다.

참고문헌

[1] Joseph P. Campbell, "Speaker recognition: a tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sep. 1997.

[2] Qi Li, Biing-Hwang Juang, Chin-Hui Lee, Qiru Zhou, Frank K. Soong, "Recent advancements in automatic speaker authentication," *IEEE Robotics and Automation Magazine*, pp. 24-34, Mar. 1999.

[3] Jialong He, Li Liu, Gunther Palm, "A New Codebook Training Algorithm for VQ-based Speaker Recognition," *Proc. ICASSP*, vol. 2, pp. 1091-1094, 1997.

[4] C. Martin del Alamo, F. J. Caminero Gil, C. de la Torre Munilla, L. Hernandez Gomez, "Discriminative training of GMM for speaker identification," *Proc. ICASSP*, vol. 1, pp. 89-92, 1996.

[5] Simon Haykin, *Neural Networks*, Prentice Hall, 1999.

[6] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 29, no. 2, pp. 254-272, 1986.

[7] Jayant M. Naik, Lorin P. Netsch, George R. Doddington, "Speaker verification over long distance telephone lines," *Proc. ICASSP*, vol. 1, pp. 524-527, 1989.

[8] Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[9] Bernhard Scholkopf, Christopher J. C. Burges, Alexander J. Smola, *Advances in Kernel Methods*, The MIT Press, 1999.