최근 화자인식 기술 동향

서 영 주
한국과학기술원

김 회 린
한국과학기술원

Ⅰ. 서 론

자동 화자인식은 발성된 음성으로부터 그 화자에 대한 정보를 추출하는 기술로서 일반적으로 화자식별 (speaker identification)과 화자검증 (speaker verification)으로 나누어진다. 이 중에서 화자식별 기술은 〈그림 1〉과 같이 임의의 화자로부터 입력된 음성을 사용하여 등록된 화자들 중에서 발성화자를 찾아내는 기술이다[1]. 즉 이 기술은 여러 사람들 중에서 특정한 목소리의 사람을 찾아주는 기능을 한다. 이 기술은 입력된 음성을 등록된 화자들의 목소리와 비교하여 그 중에서 가장 일치하는 한 화자를 선택하기 때문에 등록되지 않은 임의의 화자가 음성을 입력하여도 등록된 화자들 중의 가장 유사한 화자로 인식되는 단점을 가지고 있다. 반면에 화자검증 기술은 〈그림 2〉와 같이 사전에 특정한 화자가 제시(claim) 되었을 경우에 발성된 음성이 그 제시화자(claimed speaker)의 목소리인지를 판단하여 발성화자가 제시화자 인지의 여부를 검증하는 기술이다[2]. 따라서 화자검증 기술은 화자식별 기술이 가지고 있는 미등록화자 오식별 문제를 해결할 수 있어서 화자식별 결과의 검증을 위한 후처리 기술로도 사용된다.

이러한 화자식별과 화자검증은 입력되는 음성의 내용에 대한 제한에 따라서 또 다시 문장독립 (text-independent) 방식과 문장종속 (text-dependent) 방식으로 나누어진다[1]. 문장독립 화자인식 방식에서는 화자식별이나 화자검증을 위하여 발성하는 음성의 문장 형식이나 종류에 제한이 없는 방식이다. 즉 임의의 형식의 문장으로 구성된 음성을 입력 대상으로 하여 화자인식을 수행한다. 따라서 이 방식에서는 화자인식기가 발성된 문장을 사전에 모르는 상태에서 인식을 수행할 수 있으므로 사용자가 임의로 선정한 어구나 대화음성

(conversational speech)을 입력 대상으로 한다. 이를 통해 사용자는 보다 편리하고 융통성 있게 화자인식기를 사용할 수 있다. 반면에 문장종속 화자인식 방식에서는 사전에 정해진 문장만으로 발성된 음성을 대상으로 화자인식을 수행한다. 즉 고정 어구나 제시 어구와 같이 화자인식기가 발성 가능한 문장 종류에 제한을 가하여 문장 내용을 사전에 알고 있다. 이 방식은 사용자가 입력 가능한 문장 내용을 사전에 숙지하고 그 내용에 맞게 발성해야 하기 때문에 사용자 편의성이 떨어지는 단점을 가지고 있다. 그렇지만 이 방식은 이러한 발성문장에 대한 사전지식을 토대로 화자의 음성에 대한 통계모델을 보다 신뢰성있게 구축할 수 있기 때문에 보다 높은 화자인식 성능을 제공할 수 있다. 또한 입력 가능한 문장을 구성하는 음성 정보에 대해서만 통계모델을 생성하기 때문에 화자등록을 위한 음성입력의 분량이 더 적어지는 장점을 가지고 있다.

이러한 두 가지 기술들로 구성되는 화자인식 기술은 접근제어, 디지털 포렌식, 음성 데이터 관리, 개인화, 지능형 로봇 제어, 보안인증 등의 광범위한 응용 분야를 가지고 있다. 또한 이 기술은 상대적으로 저용량 신호처리 특성을 가지고 있으면서 일상적으로 쉽게 발성할 수 있는 음성을 사용하기 때문에 다른 생체인식 방식에 비해 저렴한 비용의 시스템 구성, 높은 사용자 편의성을 가지고 있어서 최근 IT 기술의 발달에 따라 많은 주목을 받고 있는 생체인식 기술의 하나이다.

화자인식 기술은 접근제어, 디지털 포렌식, 음성 데이터 관리, 개인화, 지능형 로봇 제어, 보안인증 등의 광범위한 응용 분야를 가지고 있다.

이에 본 논문에서는 화자인식 기술의 주된 접근방법과 더불어 본 연구실에서 수행한 화자인식 연구의 결과, 그리고 최근의 화자인식 연구방향에 대하여 화자식별과 화자검증 두 분야에 걸쳐서 간략하게 다루어보기로 한다.

〈그림 1〉 화자식별 개념도

〈그림 2〉 화자검증 개념도

Ⅱ. 화자식별

1. 화자식별의 기술적 접근법

화자식별은 서론에서 간략하게 소개한 바와 같이 사용자로 하여금 일정시간 분량의 음성발성을 통해 입력된 음성신호로부터 사용자의 음성 특징에 해당하는 정보를 추출한 후, 이를 훈련 데이터로 사용하여 화자모델을 생성하고 해당 화자를 등록함으로써 화자식별기를 구성하고, 이를 토대로 사용자로부터 발성된 음성을 이용하여 등록 화자들 중에서 특정 화자를 식별하는 기술이다. 이러한 화자식별에 대한 기술적인 접근방법으로는 Gaussian mixture model (GMM)과 support vector machine (SVM) 기법을 대표적으로 들 수 있다.

(1) GMM 기반 화자식별

GMM 기반 화자식별 기술은 화자를 모델링하기 위하여 GMM이라는 통계모델을 사용하며 화자식별을 위한 분류(classification)에서는 최대우도 (ML: maximum likelihood) 기법을 사용하는데 그 방식은 다음과 같다.

먼저 화자식별기에 등록된 화자들로 구성된 화자 그룹 $S = \{1, 2, 3, \cdots, S\}$가 이 화자들에 해당하는 통계음

향모델들의 집합 $\lambda_1, \lambda_2, \cdots, \lambda_S$로 구성될 경우에, 화자식별 과정은 입력된 음성발화로부터 특징추출 과정에서 추출한 특징벡터열 X에 대하여 최대사후확률(maximum posterior probability)을 나타내는 화자를 찾는 식 (1)의 베이즈 결정 규칙(Bayes' decision rule)으로 이루어진다[1].

$$\hat{s} = \arg\max_{1 \le s \le S} P(\lambda_s \mid X) \qquad (1)$$

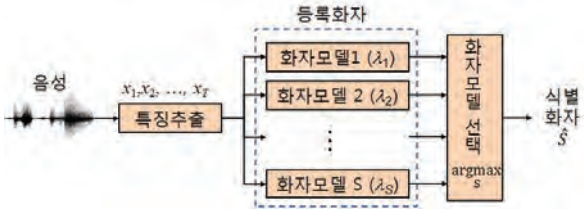
그러나 실제 화자식별에서는 이 사후확률 기법을 사용하기 보다는 몇 차례 수학적 전개를 통해 얻을 수 있는 최대우도 방식을 사용한다. 즉, 화자들의 사전확률(prior probability)이 균등분포를 이룬다고 가정하고 베이즈 정리를 적용한 다음, 화자들에 대한 특징열의 통계적 독립성을 이용하여 최대 우도 분류식으로 정리하고 이에 대해 추가적으로 특징열을 구성하는 특징벡터들 간에 통계적 독립성을 가정하고 로그를 적용한 식 (2)를 일반적으로 화자식별에 사용한다[1].

$$\hat{s} = \arg\max_{1 \le s \le S} \sum_{t=1}^{T} \log P(x_t \mid \lambda_s), \qquad (2)$$

여기서 T는 입력된 음성의 특징열을 구성하는 음성 특징벡터들의 수를 나타내고 x_t는 시간 t에서의 음성특징벡터이다. 이러한 화자식별 과정은 〈그림 3〉과 같이 나타내어질 수 있다.

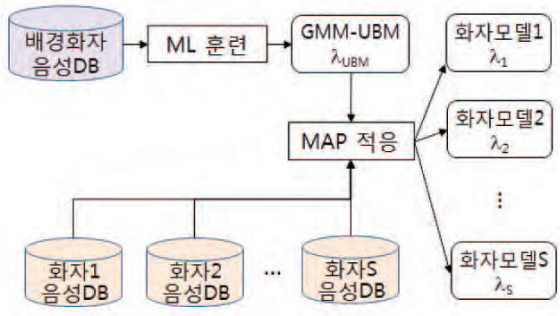
위와 같은 과정으로 화자식별을 수행하기 위해서는 개별 등록화자에 대한 통계음향모델을 생성해야 하는데 이를 위해 먼저 통계모델을 정의해야 한다. 이러한 통계모델로 화자인식에서는 다음과 같은 GMM을 가장 널리 사용한다.

$$p(x|\lambda) = \sum_{m=1}^{M} c_m p_m(x) \qquad (3)$$

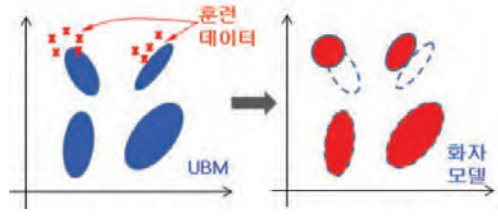


〈그림 3〉 화자식별 기술 구조도

여기서 M은 GMM의 혼합계수의 차원을 나타내고 c_m은 m번째 혼합계수를 나타내며, $p_m(x)$는 GMM의 개별 가우스 밀도함수를 나타낸다. 이들로부터 GMM의 모델 파라메터 $\lambda=\{c_m, \mu_m, \Sigma_m\}$으로 구성되는데 여기서 μ_m과 Σ_m는 m번째 혼합계수의 평균벡터와 공분산행렬을 의미한다. 이 화자모델은 훈련과정에서 expectation-maximization (E-M) 알고리즘을 이용하는 최대우도 추정 기법을 적용하여 훈련 데이타로부터 추정되는 것이 이론적으로 가장 타당하다. 그러나 정석적인 최대우도 기법으로 화자의 통계모델을 신뢰성 있게 추정하기 위해서는 개별 화자마다 다량의 훈련 데이터가 필요하다. 따라서 최대우도 기법이 가지는 이러한 단점 때문에 보다 표준화된 화자식별 훈련 방법으로 universal background model-maximum *a posteriori* (UBM-MAP) 적응 알고리즘을 이용하는 adapted GMM 기법을 사용한다[2]. 이 기법에서는 〈그림 4〉와 같이 다수 사람들(배경화자)의 보편적인 음성에 대한 통계모델을 구하기 위하여 불특정 다수 화자들로부터 수집된 배경화자 음성 데이터로부터 GMM 음향모델을 ML 방식으로 추정하여 GMM-UBM을 구성하고, 이 GMM-UBM에 대하여 특정 등록화자로부터 발성된 음성 데이터를



〈그림 4〉 Adapted GMM 기법에 의한 화자모델 생성 과정



〈그림 5〉 Adapted GMM에 의한 GMM 화자모델 적응

사용하여 〈그림 5〉와 같이 MAP 방식에 의해 그 화자의 GMM으로 적응시켜 개별화자모델을 구함으로써 화자식별기를 구축한다.

Adapted GMM에 의한 화자모델 생성 과정을 세부적으로 설명하면 먼저 UBM에서 주어진 특징벡터 x_t에 대한 개별 혼합모델의 사후확률을 정의한 다음 이로부터 혼합계수, 평균벡터, 공분산행렬을 E-M 알고리즘의 expectation 과정에서 개별 혼합밀도의 길이의 기대치, 특징벡터의 1차와 2차 기대치를 충분통계량(sufficient statistics)으로 구한다. 이로부터 UBM 모델에서 개별화자 음향모델 추정치 $\hat{\lambda}=\{\hat{c}_m,\hat{\mu}_m,\hat{\Sigma}_m\}$를 혼합밀도 길이의 기대치의 함수로 평활화하여 (smoothing) 구한다[2].

GMM 기반 화자인식 방법은 개별 화자 자체에 대한 최적의 음향모델을 추정하게 하지만 화자들 간의 차이를 고려하지 않는 생성모델 (generative model) 접근법이다. 이를 보완하여 화자들 간의 차이를 고려하는 변별모델 (discriminative model) 접근법을 사용하면 보다 화자인식 성능을 높일 수 있다.

(2) 최소분류오류 스코어 가중 기반 화자식별

이러한 GMM 기반 화자인식 방법은 개별 화자 자체에 대한 최적의 음향모델을 추정하게 하지만 화자들 간의 차이를 고려하지 않는 생성모델 (generative model) 접근법이다. 이를 보완하여 화자들 간의 차이를 고려하는 변별모델 (discriminative model) 접근법을 사용하면 화자인식성능을 보다 높일 수 있다. 이에 따라 대표적인 변별모델 접근법인 최소분류오류 (MCE: minimum classification error) 기법을 GMM 방식에 적용하는 방법을 고려할 수 있다. 이러한 변별기법으로 식 (2)에서와 같이 GMM 방식으로 개별 프레임마다 화자식별 스코어를 구하고, 이 개별 프레임 스코어에 변별적인 가중치를 적용한 후 전체 프레임들에 대한 누적 스코어의 합으로써 오류를 최소화하는 식 (4)와 같은 화자식별 방식을 생각할 수 있다.

$$\hat{S}=\arg\max_{1\le s\le S}\sum_{t=1}^{T}w_{st}\log P(x_t\mid\lambda_s) \qquad (4)$$

여기서 w_{st}는 s번째 화자의 t 번째 음성프레임에서의 변별 가중치를 나타낸다.

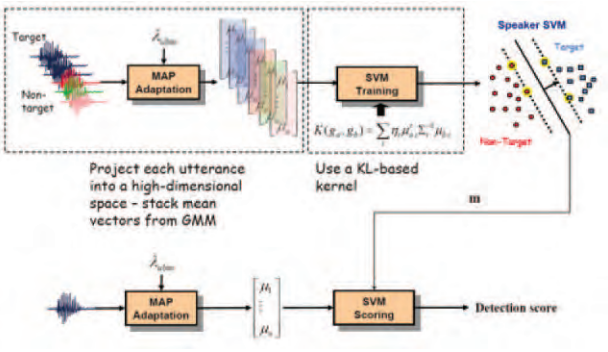
이 변별 방식에서 w_{st}를 직접 모델링하기보다는 음성클래스 별로 세분화하면 접근하기가 수월해진다. 이런 동기에서 먼저 음성 프레임들이 음향음성학적 (acoustic-phonetic) 특성과 더불어 화자 특성에 따라 각기 다른 양의 화자정보를 포함하고 있다고 가정할 수 있다. 이 가정하에 음성 프레임들을 음향음성학적 단위로 분류하면 식 (4)의 가중치를 변별적으로 구하기 위해 필요한 최소분류오류 기법에서의 변별함수(discriminant function)를 음향음성학적 클래스 기반의 함수로 정의하는 것이 가능해진다. 이로부터 가중치 파라메터는 오분류 척도 (misclassification measure) 함수로부터 유도된 sigmoid 계열 손실함수 (loss function) l_x에 generalized probabilistic descent (GPD) 알고리즘을 적용함으로써 구해질 수 있다. 이 과정에서 w_{st}는 가중치로서 $\sum_t w_{st}=1$, $w_{st}\ge 0$의 두 가지 조건을 만족해야 하기에 GPD에 직접 적용되지 않고 다음 식과 같이 변환된 형태로 구해진다.

$$\widetilde{w}_{st}^{(n+1)}=\widetilde{w}_{st}^{(n)}-\epsilon\nabla l_x\left(\widetilde{w}_{st}\right) \qquad (5)$$

여기서 $\hat{w}_{st}=\log w_{st}$로 주어진다. 식 (4)를 살펴보면 변별 가중치 w_{st}가 화자와 음성프레임의 종속변수이므로 GMM으로부터 구한 프레임별 로그 우도가 화자와 음성프레임에 따라 변별적으로 가중됨으로써 화자식별 성능을 최대화시킬 수 있음을 알 수 있다[3].

(3) SVM 기반 화자식별

한편 최근에는 생성모델 방식인 GMM 기법보다 분류정확도 면에서 보다 유리한 변별모델 방식의 SVM 기반 화자인식 기술이 소개되어 광범위하게 연구되고 있다[4-5]. SVM 기법은 〈그림 6〉과 같이 화자로부터 발성된 음성발화 데이터로부터 추출된 특징 파라메터열 세트에 대하여 adapted GMM 기법으로 GMM 모델들을



〈그림 6〉 SVM 방식의 화자식별 훈련 및 학습 과정

추정한 후, 추정된 GMM의 평균벡터들을 모두 연결(concatenation)한 GMM 수퍼벡터(supervector)를 구하고, 이 수퍼벡터들로부터 최대마진 초평면(maximum margin hyperpalne)을 구성하는 support vectors를 SVM의 학습 알고리즘을 통해 구하여 화자모델들을 생성하는 훈련과정과, 구해진 support vectors를 이용하여 입력된 발화로부터 위에서 기술한 과정으로 얻어진 수퍼벡터를 SVM의 커널 (kernel) 척도에 입력하여 화자검증을 수행하는 시험 과정으로 구성된다.

이러한 SVM 기반 화자식별 기술에 사용되는 SVM의 원리는 다음과 같다.

$$f(X_{in}) = \sum_{i=1}^{l} a_i y_i K(X_{in}, X_i) + d \underset{H_{nontg}}{\overset{H_{tg}}{\gtrless}} \eta \qquad (6)$$

여기서 X_{in} 은 입력 음성발화를 나타내고 X_i는 i번째 support vector를 의미하며, y_i는 이상적인 출력값, 즉 레이블 정보를 의미하는데 X_i가 목표화자일 (target speaker) 경우에는 1을, 목표화자가 아닐 (nontarget speaker) 경우에는 −1의 값을 갖는다. SVM 파라메터 a_i와 d는 X_i와 함께 조건 $\sum_t a_i y_i = 0$과 $a_i > 0$에 따라 훈련 단계에서 구해진다. 식 (6)에서와 같이 기본적인 SVM 분류기는 이진 분류 (binary classification) 기능을 수행한다. 이 SVM이 화자식별과 같은 복수클래스 (multi-class) 분류 기능을 수행할 경우에는 기본적인 SVM을 변경해서 사용해야 한다. 이를 위해 여러 가지 방법들이 제안되었는데 그 중에서 대표적인 방식으로, 등록화자 사이에 조합 가능한 모든 이진 분류 경우를 SVM으로 구현하고 이들 SVM의 스코어가 해당 문턱치(threshold)보다 큰 경우가 최다인 화자를 식별화자로 선정하는 max-wins 방식의 one-versus-one 알고리즘이나 등록화자수만큼의 one-versus-remaining 개념의 이진분류를 SVM으로 구현하고 이들의 스코어로부터 winner-takes-all 방식에 의해 식별화자를 선정하는 one-versus-all 알고리즘을 들 수 있다[5].

2. 화자식별 실험

(1) 실험환경

GMM과 MCE 변별 스코어 가중 기반 GMM (DSW-GMM), 및 SVM 기법을 이용한 화자식별 기술의 성능평가를 위하여 200명의 화자들이 각각 10번씩 발성한 2,000 발화들로 구성된 TIMIT[6]음성 데이터베이스를 사용하였다. 이 TIMIT를 사용한 실험에서는 잡음으로부터 차단된 깨끗한 TIMIT 음성 데이터에 car, restaurant, subway, street의 4가지 Aurora[7]잡음들을 신호대잡음비 (SNR) 20dB, 10dB, 0dB가 되도록 부가하여 최종적으로 여러 가지 음향환경을 반영하는 26,000개의 문장 음성발화 데이터를 생성하였다. 이들 데이터 중에서 절반을 훈련에 사용하였고 나머지 절반을 평가에 사용하였다.

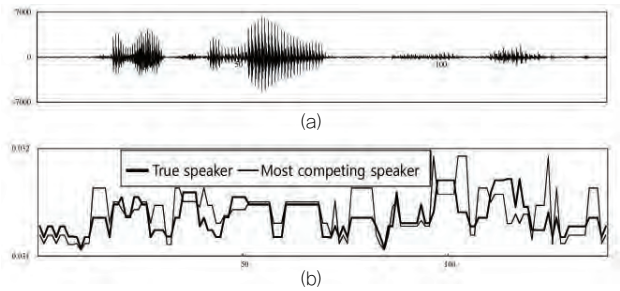
GMM 기반 화자식별 시스템은 GMM-UBM으로부터 adapted GMM 기법을 적용하여 구축되었다. 훈련에서는 5개의 서로 다른 TIMIT 음성발화들에 대하여 깨끗한 음향 조건과 12가지의 잡음에 오염된 음향 환경들을 반영하는 총 65개의 음성발화를 개별 화자모델의 훈련에 사용하는 다중조건 훈련 (multi-condition training) 기법을 적용하였다. 특징으로는 10ms의 간격마다 20ms 길이의 프레임으로부터 구한 12차 MFCC (mel frequency cepstral coefficient) 계수와 프레임 에너지를 사용하였다. GMM에서의 혼합밀도의 수는 512로 정하였고 SVM에서의 수퍼벡터의 차원은 13×512=6556으로 정하였고 선형 커널을 사용하였다.

아울러 500 화자 규모의 사무실 환경 한국어 음성을

대상으로 하는 화자식별 실험도 수행하였는데 이 실험에서는 GMM 및 SVM 기반의 문장종속과 문장독립 방식의 화자식별 평가를 위하여 5,000 발화를 사용하였다. 문장종속에서는 개별 화자의 등록에 18초의 음성 분량을 사용하였고 평가에서는 6초를 사용하였다. 문장독립에서는 등록에 120초, 평가에 10초 길이의 음성을 사용하였다.

(2) 성능평가 및 검토

〈그림 7〉은 식 (4)에서 정의된 최소분류오류 기반 변별적 스코어 가중 기법에 적용되는 변별 가중치를 식 (5)의 GPD 알고리즘으로 구한 후 20 dB 신호대잡음비 조건으로 Aurora 잡음에 오염된 음성신호에 적용한 변별 스코어 가중치의 시계열 윤곽선이다. 〈그림 7(b)〉에서 보는 바와 같이 제시화자인 정화자(true speaker)와 가장 경쟁화자(most competing speaker)의 가중치가 음성 부분에 따라 서로 뚜렷하게 다름을 알 수 있다. 이는 음성프레임이 속한 음소나 음운 등 음향음성학적 클래스에 따라 화자별로 해당 GMM 로그 우도 스코어를 변별적으로 가중하는 방식이 최종적인 화자식별 정확도를 높일 수 있음을 의미한다.



〈그림 7〉 20 dB 신호대잡음비로 Aurora 잡음에 오염된 음성 (a)에 대한 변별 스코어 가중치의 시계열 윤곽선 (b)

〈표 1〉 Aurora 잡음 환경에서 DSW 기법에 의한 화자식별 오류율 (%)

SNR	GMM	DSW-GMM	SVM
Clean	10.90	8.30	9.0
20 dB	17.70	15.80	25.43
10 dB	30.02	26.45	42.45
0 dB	72.13	59.22	69.53

〈표 1〉은 다양한 Aurora 잡음환경에서의 화자식별 성능평가 결과이다. 시험 데이터 세트들은 깨끗한 음성 데이터와 세 가지 신호대잡음비 조건에서 네 가지 Aurora 잡음(car, restaurant, subway, street)을 부가하여 생성된 잡음에 오염된 음성 데이터들로 구성되었다. 이 표를 보면 깨끗한 음성 환경에서 SVM 방식이 GMM 방식에 비해 더 우수한 화자식별 성능을 나타냄을 알 수 있다. 반면에 GMM 방식은 잡음환경에서 보다 더 우수하다. 이는 변별학습 방식인 SVM 기법이 훈련환경과 시험환경에서 음향 불일치가 일어날 경우 성능저하가 더 클 수 있음을 암시한다. 전체 신호대잡음비 조건에서 제안된 DSW-GMM 기법이 기존의 GMM 방식에 비해 10% 이상의 식별오류 개선을 보이며 가장 큰 오류 개선은 깨끗한 음성에서 이루어짐을 알 수 있다. 깨끗한 음성에서 가장 우수한 개선을 얻는 이유로는 이 조건에서 음향음성학적 분류가 보다 정확하게 이루어지기 때문일 것으로 추측된다. 〈표 1〉의 화자식별 실험결과는 변별적 스코어 가중 기법이 다양한 잡음환경에서 화자식별 성능을 개선시키는데 효과가 있음을 확인시켜준다.

〈표 2〉는 500 화자 규모의 사무실 환경 한국어 음성 데이터에 대한 화자식별 실험결과이다. 문장종속 및 문장독립 방식 실험에서 화자 수가 비교적 큼에도 불구하고 GMM과 SVM 모두 사무실 환경의 음성발화에 대해 매우 우수한 성능을 나타냄을 확인할 수 있다. 오류율 상으로는 비슷하지만 두 방식에서 등록과 평가에 사용된 음성발화의 길이가 다르게 주어졌음을 주목할 필요가 있다. 즉 문장종속의 경우 제한된 문장을 사용하기 때문에 등록과 평가에 사용되는 발화의 길이가 훨씬 작아질 수 있다. GMM과 SVM의 성능은 거의 비슷하지만 문장종속의 경우 등록에서 과도하게 제한된 음성발화의 길이로 인하여 support vectors가 충분히 만들

〈표 2〉 한국어 음성에 대한 화자식별 오류율 (%)

화자식별 방식	GMM	SVM
문장종속	0.14	0.26
문장독립	0.14	0.16

어질 수 없었으며 이로 인하여 예상과 달리 SVM 방식의 오류율이 더 높게 나타났다.

Ⅲ. 화자검증

1. 화자검증의 기술적 접근법

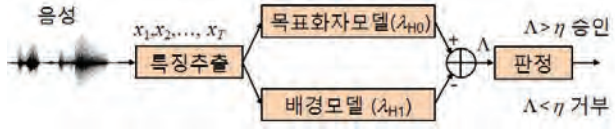
(1) GMM 기반 화자검증

화자검증 기술은 앞서 기술한 바와 같이 입력된 음성을 분석하여 발성화자가 제시화자 또는 목표화자 인지의 여부를 검증하는 기술이다. 현재 화자검증을 위한 표준적인 방법은 앞서 기술한 화자식별과 동일하게 훈련 음성 데이터를 사용하여 GMM-UBM으로부터 개별 화자의 음향모델을 구하는 베이즈 학습에 기반을 둔 adapted GMM 기술이다. Adapted GMM 방법에서 화자검증은 식 (7)과 같은 우도비 시험 (LRT: likelihood ratio test) 기법을 활용하는데, 이 방법은 〈그림 8〉과 같이 음성이 입력되면, 목표화자의 음향모델에 대한 입력음성의 우도와 GMM-UBM에 대한 입력음성의 우도 간의 비율을 사전에 정한 문턱치와 비교하는 방식이다.

GMM의 개념적인 측면에서 GMM 수퍼벡터 선형커널을 살펴보면 이 커널함수가 수퍼벡터에 대해 GMM의 혼합계수와 대각선 공분산행렬을 고려하지만 이 값들이 화자검증의 성능 측면에서 최적화되었다고는 볼 수 없다.

$$\Lambda = \frac{\text{Likelihood S came from target speaker model}}{\text{Likelihood S came from nontarget speaker model}} = \frac{p(X|\lambda_{H0})}{p(X|\lambda_{H1})} \underset{H_1}{\overset{H_0}{\gtrless}} \eta \quad (7)$$

여기서 H_0는 대상화자가 목표화자일 경우인영가설(null hypothesis)이고 H_1은 대상화자가 목표화자가 아닐 경우 또는 사칭자(imposter)일 경우인 대립가설(alternative hypothesis)을 나타낸다. η는 이 우도비 시험법에서의 문턱치를 나타낸다. 화자검증 서비스에서 문턱치 η의 선정은 개별 서비스의 특성에 따라 달라지는데 높은 보안성이 요구되는 분야에서는 오승인(false acceptance) 비율을 최소로 하기 위하여 이 값을 높게 잡는 경향이 있고 반대로 보안성이 덜 요구되면서 사칭자를 일부라도 자동적으로 걸러낼 때 얻게 되는 효과가 큰 분야에서는 오거부 (false rejection) 비율을 낮추기 위하여 이 값을 낮게 정한다.



〈그림 8〉 화자검증 기술의 구조도

(2) SVM 기반 화자검증

화자검증은 이진 분류의 문제이므로 SVM 기법에서는 SVM의 원리인 식 (6)을 그대로 적용할 수 있다. 즉, SVM 기반 분류는 입력 벡터와 support vectors 간의 커널 거리척도 (distance metric) 방식의 비교를 통해 이루어지며 커널은 입력공간(input space)을 고차원 특징공간(feature space)으로 사상시키는데 특정한 커널함수를 통해 입력공간이 선형비분리(linearly nonseparable) 형태의 경우에도 선형분리 형태의 특징공간으로 변환시킬 수 있다. 따라서 SVM 기법에 어떤 분류 문제가 주어졌을 때 적절한 커널함수의 선정이 그 분류의 정확도에 큰 영향을 미친다. 화자검증 문제에서는 초기에 GMM의 평균벡터 만으로 구성된 수퍼벡터의 내적으로 구성되는 선형커널이 주로 사용되다가 여기에 GMM 혼합계수와 대각선 공분산행렬을 통합하는 식 (8) 형태의 GMM 수퍼벡터 선형커널이 제안되었고 기존 선형커널에 비해 개선된 성능을 나타내는 것으로 보고되었다[4].

$$K(utt_a, utt_b) = \sum_m c_m (\mu_m^a)' \Sigma_m^{-1} \mu_m^b \quad (8)$$

여기서 utt_a와 utt_b는 각각 음성발화 a와 b를 나타낸다.

(3) 최소분류오류 가중커널 SVM 기반 화자검증

SVM 방식 화자검증에서는 커널함수가 성능에 직접적인 영향을 미치기 때문에 GMM에 기반을 둔 SVM의

경우 GMM의 특성을 고려하여 커널함수를 선정하면 보다 높은 성능을 기대할 수 있다. GMM의 개념적인 측면에서 GMM 수퍼벡터 선형커널을 살펴보면 이 커널함수가 수퍼벡터에 대해 GMM의 혼합계수와 대각선 공분산행렬을 고려하지만 이 값들이 화자검증의 성능 측면에서 최적화되었다고는 볼 수 없다. 따라서 수퍼벡터의 개별 계수에 대한 최적화를 추가로 진행한다면 화자검증 성능의 개선이 가능할 것으로 예상된다. 이런 이유에서 식 (9)와 같이 최소분류오류 최적화에 기반을 둔 GMM 혼합계수와 평균벡터 계수에 대한 가중커널을 고려할 수 있다[8].

$$K_\alpha(X_i,X_j)=\sum_m \alpha_m^{mc} c_m (\mu_m)^t A \Sigma_m^{-1} \mu_m^j \qquad (9)$$

여기서 α_m^{mc}는 혼합계수에 대한 가중치로서 조건 $\sum_m \alpha_m^{mc}=1$, $\alpha_m^{mc}\geq 0$를 만족해야 하고 대각선 정방행렬 A의 대각선 요소는 변별 가중치 $\{\alpha_d^{mv}, 0\leq d\leq D\}$로서 $\sum_d \alpha_d^{mv}=1$, $\alpha_d^{mv}\geq 0$의 조건을 만족해야 하며 여기서 D는 화자검증에 사용되는 특징벡터의 차원을 의미한다.

2. 화자검증 실험

(1) 실험환경

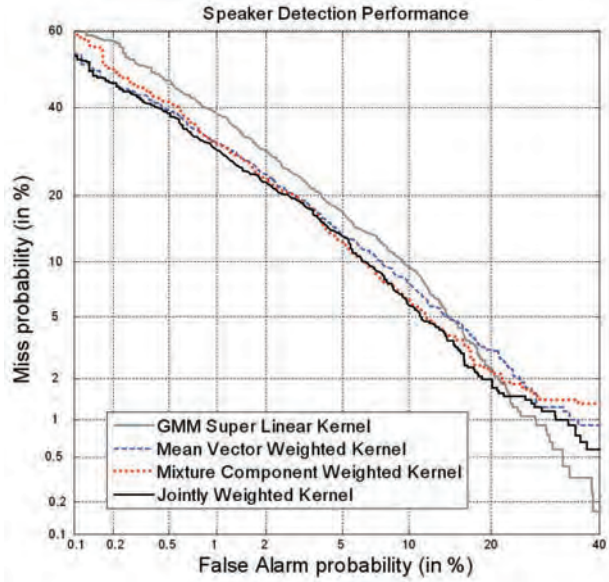
화자검증 기술의 성능평가를 위한 실험도 두 가지 음성 데이터에 대하여 수행되었다. 최소분류오류 가중커널을 사용하는 SVM 방식 화자검증 실험에서는 2008 NIST speaker recognition evaluation (SRE)[9] 데이터베이스를 사용하였는데 실험 조건은 8 대화 훈련-1 대화 시험 (8 conversation training - 1 conversation test) 규약을 적용하였다. 이 조건은 635 목표화자 모델에 대해 훈련에서 5080 개의 목표 발화와 시험에서 1870 개의 목표 발화와 14700 개의 비목표 발화로 구성된다. 특징벡터로는 19차 MFCCs와 1차 프레임 에너지와 이에 대한 델타 계수들로 구성된다.

이와 더불어 화자식별 실험에서와 같이 500 화자 규모의 사무실 환경 한국어 음성을 대상으로 화자검증 실험도 수행하였다. GMM 및 SVM 기반의 문장종속과 문장독립 방식의 화자검증 평가를 위하여 목표화자 대비목표화자의 비율이 1:3으로 조정된 20,000 발화로 구성된 데이터 셋을 사용하였다. 문장종속에서는 등록 음성의 분량을 18초로 정하였고 평가에서는 6초로 정하였다. 문장독립에서는 등록에서 120초로 정하였고 평가에서는 10초로 정하였다.

(2) 성능평가 및 검토

〈그림 9〉는 2008 NIST SRE 과제에서 최소분류오류 가중커널에 기반한 SVM 기법에 대한 화자검증 실험결과를 detection error tradeoff (DET) 곡선으로 그린 것이다. 기존의 GMM 수퍼벡터 선형커널 (GMM super linear kernel) 보다 최소분류오류에 기반을 둔 세 가지 가중 커널들, 즉 혼합계수 가중커널, 평균벡터 가중커널, 이 둘을 통합한 가중커널이 보다 우수한 성능을 나타냄을 알 수 있다.

〈표 3〉은 500 화자 규모의 사무실 환경 한국어 음



〈그림 9〉 2008 NIST SRE 과제에서 GMM 수퍼벡터 선형커널과 최소분류오류 가중커널을 사용한 SVM 기반 화자검증 기술의 DET 곡선

〈표 3〉 한국어 음성에 대한 화자검증 오류율 (%)

화자검증 방식	GMM	SVM
문장종속	0.20	0.28
문장독립	0.21	0.42

성 데이터에 대한 화자검증 실험결과이다. 성능은 miss probability 와 false alarm probability가 같은 값에서의 오류율을 나타내는 equal error rate로 나타내었다. GMM 방식이 SVM 방식에 비해 약간 우수하지만 두 방식 모두 매우 우수한 성능을 나타냄을 알 수 있다.

Ⅳ. 최근 화자인식 연구방향

현재까지 화자인식에 대한 전세계적 연구는 화자식별보다는 주로 화자검증 분야에서 활발히 진행되어 왔다. 이는 화자검증이 화자식별보다 실제 적용 분야에 있어서 우위를 점하고 있기 때문인 것으로 보고되고 있다. 그러나 화자검증 기술도 다중 클래스 분류의 개념을 도입하면 화자식별 분야에 바로 적용가능하기 때문에 화자검증에 대한 연구를 화자인식 분야 전체에 대한 연구로 이해해도 큰 무리가 없을 것이다. 앞에서 언급한 바와 같이 화자인식에서의 표준적인 방법으로 GMM 기법이 연구 초기 단계에서 집중적으로 연구되었고 그 후에 SVM 기법의 등장으로 GMM 수퍼벡터에 기반을 둔 SVM 방식의 화자검증이 현재까지 화자인식의 주된 기술적 접근방법이 되고 있다.

최근에는 요인분석 (factor analysis) 기법의 일종인 joint factor analysis (JFA) 또는 i-vector 기법을 SVM과 결합시키는 방식이나 JFA나 i-vector를 cosine 커널을 통해 직접 거리 계산하는 방식 등이 제안되어 이를 통해 화자인식 성능을 보다 높이는 방향으로의 연구가 활발히 진행되고 있다[10]. 한편 지금까지 이러한 화자검증 연구는 주로 조용한 환경에서 발성된 음성에 대한 채널, 핸드셋, 화자 목소리의 시간적 경과에 따른 변화 등과 같은 음성녹음에서의 변이요인을 의미하는 session variability에 대처하는 연구가 주로 이루어져왔다. 그러나 화자검증 기술도 실제 환경에 도입되기 위해서는 주변 잡음의 영향으로부터의 강인성이 확보되어야 한다. 이러한 잡음에 강인한 화자검증 방법으로는 probabilistic linear discriminant analysis (PLDA)를 i-vector에 적용하는 기법이 제안되었으며 현재 이 분야에 대한 연구가 활발히 진행되고 있다.

최근에는 요인분석 (factor analysis) 기법의 일종인 joint factor analysis (JFA) 또는 i-vector 기법을 SVM과 결합시키는 방식이나 JFA나 i-vector를 cosine 커널을 통해 직접 거리 계산하는 방식 등이 제안되어 이를 통해 화자인식 성능을 보다 높이는 방향으로의 연구가 활발히 진행되고 있다.

Ⅴ. 결 론

지금까지 화자인식에서의 표준적인 기술적 접근방법인 GMM과 SVM 기법을 이용한 화자식별과 화자검증 기술에 대해 간단하게 살펴보았다. 또한 최소분류오류 방식에 기반을 둔 변별 가중 기법을 GMM의 프레임 스코어에 적용하는 화자식별 방법과 SVM의 커널함수에 적용하는 화자검증 방법을 소개하였다. 아울러 화자인식 분야의 성능평가에서 표준으로 사용되는 TIMIT 및 NIST SRE 데이터베이스와 한국어에 대한 성능평가를 위하여 자체적으로 구축한 500 화자 규모의 한국어 데이터베이스를 사용하여 소개된 개별 화자인식 기술들의 성능평가를 수행하였고 그 결과를 소개하였다.

본 논문에서 소개된 GMM 및 SVM 기반의 화자인식 기술들은 기술적 구현의 용이함으로 인하여 실제 서비스를 위한 기술개발 과정에서 폭넓게 채택되고 있다. 특히 GMM 방식은 SVM 방식에 비해 일반적으로 인식 성능이 떨어진다고 알려졌지만 프레임 동기 방식의 실시간 처리 구현 측면이나 잡음에 대한 강인성 측면에서 상대적인 우월성을 가지고 있다. 현재 음성인식 기술이 모바일 기기 환경에서 요구되는 서비스 수요와 기술 자체의 발전, 그리고 서버 컴퓨터의 계산능력 개선에 힘입어 실제 서비스에 폭넓게 적용되는 것과 같이 화자인식 기술도 향후 보안, 지능형 로봇 등의 주요 화자인식 서비스 분야에서 보다 광범위하게 이용될 것이며 이를 뒷받침하는 기술적 연구도 더욱 활발히 수행될 것으로 예상된다.

참 고 문 헌

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. Vol. 3, no. 1, pp. 72-83, 1995.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Process., Vol. 10, pp. 19-41, 2000.

[3] 서영주, 김회린, "음향음성학적 분류 기반 변별적 스코어 가중 기법을 이용한 화자식별," 제 25회 신호처리합동학술대회, pp. 24-26, 2012년 9월.

[4] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, Vol. 13, no. 5, pp. 308-311, 2006.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Trans. Intelligent Systems and Technology, 2:27:1-27:27, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[6] W. Fisher, M. G. R. Doddington, and K.M. Goudie-Marchall, "The DARPA speech recognition research database: Specifications and status," in Proc. DARPA Workshop on Speech Recognition, pp. 93-99, 1986.

[7] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. Int. Conf. Spoken Language Processing, pp. 16-20. Oct. 2000.

[8] Y. Suh and H. Kim, "Minimum classification error-based weighted support vector machine kernels for speaker verification," The Journal of The Acoustical Society of America, Vol. 133, no. 4, pp. EL307-EL313, April 2013.

[9] NIST Speaker Recognition Evaluation is available at http://www.itl.nist.gov/iad/mig/ tests/sre/2008/ (Last viewed 3/11/2013).

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech and Language Processing. Vol. 19, no. 4, 788-798, 2011.

서 영 주

1991년 2월 경북대학교 전자공학과 학사
1993년 2월 경북대학교 전자공학과 석사
2006년 8월 한국과학기술원 공학부 박사
1993년 2월~1998년 12월 한국전자통신연구원 연구원
2000년 3월~2002년 5월 ㈜코아보이스 선임연구원
2006년 9월~현재 한국과학기술원 전기및전자공학과 연구 부교수

〈관심분야〉
화자인식, 음성인식, 음성 및 음향 신호처리

김 회 린

1984년 2월 한양대학교 전자공학과 학사
1987년 2월 한국과학기술원 전기및전자공학과 석사
1992년 2월 한국과학기술원 전기및전자공학과 박사
1987년 11월~1999년 12월
한국전자통신연구원 선임연구원
1994년 6월~1995년 5월
일본 ATR-ITL 방문연구원
2006년 7월~2007년 7월
미국 UCSD INC 방문학자
2000년 1월~현재
한국과학기술원 전기및전자공학과 교수

〈관심분야〉
음성인식, 화자인식, 음성 및 음향 신호처리