화자인식 기술 및 국내외시장 동향

유하진
서울시립대학교 컴퓨터과학부

An Overview and Market Review of Speaker Recognition Technology

Ha-Jin Yu
School of Computer Science, University of Seoul

hjyu@venus.uos.ac.kr

Abstract

We provide a brief overview of the area of speaker recognition, describing underlying techniques and current market review. We describe the techniques mainly based on GMM(gaussian mixture model) that is the most prevalent and effective approach. Following the technical overview, we will outline the market review of the area inside and outside of the country.

I. 서론

화자인식은 음성을 입력으로 받아들여 이를 발성한 사람 즉, 화자가 누구인지 판별하는 과정이다. 최근 사회의 거의 전 분야에 정보화가 진행되면서 정보 보안에 대한 관심이 높아지고 있고, 사용자의 인증이 정보처리의 주요 과정이 되었다. 전통적으로 사용자 인증에 사용되는 방법은 비밀번호 인데, 이것은 추측 및 도용의 위험이 크고 또한 잊어버리기 쉬운 단점이 있다. JP Morgan H&Q 와 the Gartner Group 의 추정에 의하면 IT help desk에 전화하는 내용중 30%는 비밀번호를 잊어버리는 경우라고 한다.

이에 대한 대처방안으로 생체인식이 많이 고려되고 있다. 생체인식은 음성, 지문, 얼굴, 홍채 등 사람의 각 신체부위를 사용한다 사람의 신체부위는 분실 및 도용이 불가능하므로 안전하고, 사람이 사람을 식별하는 방식과 유사하므로 가장 자연스러운 방법이 될 수 있다. 이 중에서 음성은 기타 다른 방법에 비하여 비교적 저가의 하드웨어 장비로 구현할 수 있는 장점이 있다. 즉, 다른 부위는 고가의 장비인 카메라가 필요한데 반하여 음성은 비교적 저가인 마이크만 있으면 사용할 수 있으며, 흔히 보유하고 있는 전화를 사용하면 별도의 장비 추가 없이도 사용할 수 있다.

본 논문에서는 화자인식의 최근 연구 방향과 국내외 산업체 동향을 간략하게 정리하였다.

II. 화자인식 기술

1. 개요

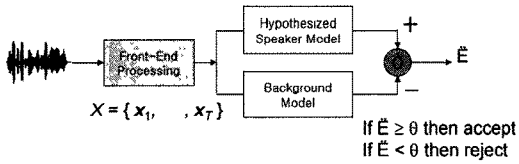
화자인식은 음성인식에서 단어와 화자가 서로 뒤바뀐 형태이므로 쌍대적(雙對的)으로 생각할 수 있다. 즉, 음성인식에서 단어와 화자의 역할을 서로 바꾸면 화자인식의 기능을 하게 되는 것이다. 음성인식에서는 하나의 음소(또는 단어)에 해당하는 여러 화자의 음성을 수집하여 음소(또는 단어) 모델을 작성하며, 화자인식에서는 이와 반대로 특정 화자가 발성한 다양한 음소를 수집하여 화자 모델을 만든다.

화자인식은 크게 화자식별과 화자인증으로 나눌 수

있다. 화자식별(Speaker Identification)은 입력된 음성이 여러 명이 포함된 화자그룹 내에서 누가 발성한 것인지를 판별하는 방법이다. 이것은 유한개수의 단어를 포함하는 집합 내에서 어느 단어를 발성한 것인지를 판별하는 단어음성인식과 유사하다고 볼 수 있다. 따라서 화자식별은 음성인식과 거의 동일한 알고리즘을 사용하여 구현 할 수도 있다. 화자인증 또는 화자확인 (speaker verification)은 입력된 음성이 특정화자의 음성인지 아닌지를 확인하는 과정이다. 화자인증의 결과는 "예"와 "아니오" 두 가지 이지만, 처리 과정은 화자식별보다 복잡한 문제를 가지고 있다. 화자식별과정에서는 입력 음성과 화자모델간의 우도(likelihood)를 구하여 우도가 가장 큰 모델을 선택하여 결과를 얻을 수 있지만, 화자인증 과정에서는 입력 음성과 특정 화자모델과의 우도를 구한 후 우도가 어느 정도 일 때 승인해야 할지를 결정해야 하는 어려운 문제가 있다. 임계값(threshold)을 정하여 우도가 임계값 이상이면 승인을 하도록 할 수 있지만, 이 임계값이 너무 높아지면 승인해야 할 화자에 대해서도 거부하는 오류율(False rejection rate)이 높아지고, 임계값이 너무 작아지면 거부해야 할 화자에 대해서도 승인하게 되는 오류율(false acceptance rate)이 높아지게 된다.

화자인증은 음성인식에서의 미등록어 거부기능에 해당한다고 볼 수 있다. 즉, 단어 인식에서 사용자가 등록되지 않은 단어 또는 간투사를 발성하였을 때 등록된 단어 중의 하나로 인식하게 되는 것을 막는 기능이 있어야 하는 것과 같다. 이 문제는 현재 음성인식 시스템의 실용화를 어렵게 만드는 큰 문제 중의 하나로, 같은 어려움이 화자인증에서도 동일하게 존재하게 된다. 음성인식에서의 거부기능에 필러(filler)모델을 사용하는 것과 같이 화자인증에서는 배경모델(background model)을 사용하여, 입력된 음성에 대하여 배경모델과 특정 화자모델과의 우도비가 임계값보다 크면 거부하게 된다.

화자인식은 또 사용하는 문장에 따라 문장종속형과 문장독립형으로 구분할 수 있다. 문장종속(text-dependent)형은 학습과정과 인식과정에서 동일한 문장을 사용하는 것이고, 문장독립(text-independent)형은 인식과정에서 사용할 문장이 미리 정해지지 않는 것이다. 문장종속형을 사용하면 기존의 음성인식 방법을 그대로 사용할 수 있고, 비밀번호의 기능까지 포함할 수 있지만, 고성능 녹음기를 사용할 경우에 보안성을 보증할 수 없게 된다 문장독립형은 녹음기의 문제를 해결하는 외에, 사용자와 다른 목적의 대화를 하는 동안에 입력한 음성으로 화자인식을 할 수 있는 장점이 있지만 학습에 많은 데이터가 필요한 단점이 있다 이 두 가지 방법의 절충안으로 문장제시(text-prompted)



$$\Lambda(X) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{\overline{hyp}})$$

그림 1 화자 인식 과정

형이 있다. 이것은 인식하는 과정에서 발성해야 할 문장을 제시하여 녹음기의 문제를 해결하고 인식성능을 높일 수 있다.

앞서 기술한 바와 마찬가지로, 화자인식에서도 음성인식에서 사용하는 과정이 대부분 유사하게 사용된다. 특히, 문장종속형 화자인식에서는 음성인식과 동일한 알고리즘을 사용할 수 있으므로 현재는 주로 문장독립형 화자인식 방법이 연구되고 있다. 화자인식에서 사용되는 대표적인 음성인식 방법으로는 DTW(dynamic time warping), VQ(vector quantization), HMM(hidden Markov model)등이 있다[1]. 최근 가장 많이 사용되고 있는 방법은 GMM(Gaussinan mixture model)으로, 이것은 하나의 상태(state)를 가지는 연속밀도함수 HMM 또는 모든 상태와 상태가 모두 동일한 천이확률로 구성된 HMM이라고도 생각할 수 있다.

화자인증 시스템의 개략적인 구성은 그림 1과 같다. 입력된 음성신호는 잡음신호 구간을 제거하는 끝점 검출 과정을 거쳐 특징추출 단계에서 특징 벡터로 변환된다. 등록단계에서는 사용자가 발성한 음성에서 추출된 특징벡터열을 이용한 학습과정에 의하여 화자모델이 만들어진다 사용단계에서는 사용자가 사용자 번호를 제시하고 음성을 발성하면, 이 음성에서 추출된 특징 벡터열과 화자모델을 비교하여 유사도를 측정한다. 이 유사도와 배경화자와의 유사도와의 비율 등을 이용하여 사용자의 음성이 제시한 사용자 번호와 일치하는지를 결정한다 이때, 유사도가 미리 정해놓은 임계값보다 높으면 승인하고, 그렇지 않으면 거부하게 된다.

2. 특징추출

화자인증에서는 음성인식과 반대로 화자내 변이보다 화자간 변이가 큰 특징을 사용해야 한다. 화자인증을 위한 특징에 관하여 많은 연구가 진행되어 왔으나, 대부분의 경우에는 음성인식에서 좋은 효과를 내는 특징이 화자인식에서도 높은 성능을 보이고 있다. 주로 사용되는 특징으로는 MFCC(Mel-frequency cepstral

coefficients)와 이의 1차 및 2차 미분이 사용되고, 잡음을 감소시키기 위하여 CMS (Cepstral mean subtraction) 방법을 사용한다.

3. GMM(Gaussina Mixture Model) [2]

통계적인 화자 모델에서는 화자를 특징벡터를 출력하는 랜덤 소스로 가정한다. 이 랜덤 화자 모델 안에는 성도의 특성을 나타내는 숨겨진 상태(state)들이 있다. 이 랜덤 소스가 특정한 상태에 있을 때 특정한 성도 특성에 해당하는 특징 벡터를 출력하게 된다. 각각의 상태는 각각 평균과 공분산을 가지고 다차원 가우시안 확률 분포 함수에 의하여 특징 벡터를 출력한다 입력 **x**에 대하여 상태 i의 D차원 확률분포함수 pdf 는 다음과 같이 계산된다

$$g_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'(\Sigma_i)^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)'\right\}$$

여기서 μ_i 는 상태의 평균 벡터이고, Σ_i는 상태의 공분산행렬이다. 공분산행렬은 상태에서의 특징벡터의 상관관계와 변이성을 표현한다. 또한, 다음식과 같이 M개의 상태에 연관되어 있는 이산 확률분포함수w_i는 각각의 상태에 있게 될 확률이다.

$$\sum_{i=1}^{M} w_i = 1$$

상태 i에서 상태 j로 천이하는 확률을 표현하는 상태 천이 이산 확률분포함수는 다음과 같이 정의된다.

$$a_{ij} = P(i \rightarrow j),\ \text{for } i,j = 1, \ldots, M$$

이와 같은 확률 모델의 정의는 HMM(Hidden Markov Model)이라고 부른다. 문장 독립 화자 모델에서는 모든 상태 천이 확률이 같도록 한다. 즉

$$a_{ij} = 1/M$$

으로 한다. 상태 천이 확률은 화자에 종속적인 특징을 표현할 수 있지만, 주로 언어적인 정보가 많고, 실제로 문장 독립 화자인식에서는 불필요한 것으로 알려져 있다.

M개의 상태를 가진 화자 모델에서 가우시안 혼합 모델(Gaussian Mixture Model)은 다음과 같이 표현된다.

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g_i(\mathbf{x})$$

여기서 λ는 화자 모델의 파라미터를 나타낸다

$$\lambda = (w_i, \mu_i, \Sigma_i),\ \text{for } i=1,..,M$$

따라서, 파라미터λ의 화자모델에서 특징벡터 **x**를 관측할 확률은 **x**가 각각의 상태에서 출력될 확률을 그 상태에 있을 확률로 가중하여 합한 것이다. 모델의 학습에는 EM(Expectation-Maximization) 알고리즘을 이용한다. 등록화자로부터 발성된 음성에서 추출된 특징벡터가 주어지면 EM알고리즘은 반복적으로 모델 파라미터를 다듬어서 학습데이터와 모델파라미터가 잘 정합되도록 한다.

EM알고리즘은 E단계와 M단계로 나누어 진다. E(expectation) 단계는 현재의 모델 파라미터와 관측데이터 (observations) 를 이용하여 숨겨진 구조 (hidden structure) 를 예측한다. GMM에서는 다음과 같이 표현할 수 있다.

$$p(i|\vec{x}_t, \lambda) = \frac{w_i g_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k g_k(\vec{x}_t)}$$

M(maximaization)단계에서는 예측된 숨겨진 구조를 이용하여 파라미터를 재추정한다 모델의 가중치, 평균, 분산은 각각 다음과 같이 재추정된다.

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \overline{\mu}_i^2$$

4. 인증기술

화자 식별 시스템에서 최종 결과의 결정은 비교적 간단해 진다. 고립단어 음성 인식 시스템에서와 같이 여러 후보 단어들 중에서 가장 유사도가 높은 단어를 선택하면 되기 때문이다. S명의 화자의 모델 λ_1, λ_2, ⋯, λ_s 이 있을때 화자식별은

$$p(\vec{x}_t|\lambda_k) = \sum_{t=1}^{T} w_i g_i(\vec{x}_t)$$

을 최대로 하는 k를 찾는 것이다.

입력 x가 k번째 화자의 음성인지 아닌지를 판별하는 화자인증에서는 배경화자 모델λ_B을 사용하여 $p(x|\lambda_s)$와 $p(x|\lambda_B)$를 비교하여 승인여부를 결정한다. 화자인증 시스템에서는 입력된 음성이 제시된 화자의 음성인지 아닌지만 판별하면 되므로 매우 단순해 보이지만 실제로는 화자 식별 시스템 보다 훨씬 복잡해진다. 그것은 제시된 화자에 해당하는 모델은 등록과정에서 잘 모델링 되지만 이와 반대되는 모델이 정

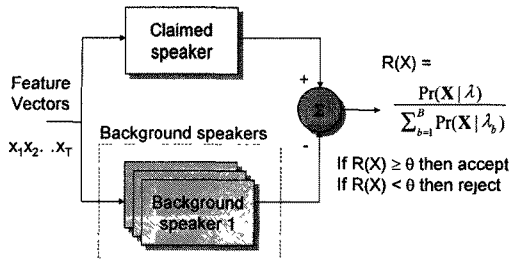


그림 2. 화자인증 시스템의 결정 구조

확히 정의되기 어렵기 때문이다. 화자인증 시스템은 이와 같이 잘 정의된 모델과 잘 정의되지 않은 모델간의 구별을 해야 한다. 화자인증 시스템의 결정 과정은 그림 2와 같다. 테스트 과정에서는 입력된 음성의 화자가 제시된 화자와 같을 때의 선택은 H0이 되고, 다를 때는 H1이 된다.

H0과 H1을 결정하는 유사도비(likelihood)를 결정하는 데는 제시된 화자이외의 가능한 모든 사람에 대한 모델이 있어야 한다. 입력된 음성 X에 대하여 제시된 화자의 모델이 λc이고, 제시된 화자가 아닌 모델이 λĉ이라면, 유사도비는 다음과 같이 계산된다.

$$\frac{Pr(X \text{ 가 제시된 화자의 음성})}{Pr(X \text{ 가 제시된 화자 이외의 음성})} = \frac{Pr(\lambda c|X)}{Pr(\lambda \hat{c}|X)}$$

여기에 베이스(Bayes')규칙을 적용하고 상수항인 사전확률을 제거하면 로그영역의 유사도비는 다음과 같이 된다.

$$R(X) = \log P(\lambda c|X) - \log P(\lambda \hat{c}|X)$$

$P(\lambda c|X)$는 입력된 음성이 제시된 화자에 속할 유사도이고, $P(\lambda \hat{c}|X)$는 음성이 제시된 화자에 속하지 않을 유사도가 된다. 이 유사도비는 임계값(θ)과 비교되어, $R(X) \geq \theta$ 이면 제시된 화자는 승인되고, $R(X) < \theta$ 이면 거부된다. 이 결정 임계값을 변경함으로써 올바른 화자를 거부하는 오류거부율(false rejection rate)과 사칭자를 승인하는 오류승인율(false acceptance rate)을 조절하게 된다.

입력된 음성 X가 제시된 화자의 음성에 속할 유사도는 다음과 같다.

$$\log P(X|\lambda_C) = \frac{1}{T}\sum_{t=0}^{T} \log P(x_t|\lambda_C)$$

여기서 1/T 항은 발성 길이에 대하여 유사도를 정규화하기 위해 사용된다. 입력 음성이 제시된 화자의 음성이 아닐 유사도는 배경 화자 모델의 집합에 의하여 정해진다 배경화자 집합 B = {$\lambda_1, \lambda_2, \ldots, \lambda_B$} 에 대하여 배경화자의 로그 유사도는 다음과 같이 계산된다]

$$\log P(X|\lambda_{\bar{C}}) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} P(x_t|\lambda_b)\right\}$$

성능척도로는 동일오류율(Equal Error Rate)을 사용한다. 동일오류율은 오류승인률과 오류거부율이 같아지도록 임계값을 조절했을 때 두 오류율이다.

4. UBM(Universal Background Model) [4]

배경화자 모델을 생성하는 방법에는 두 가지 접근방법이 있다. 첫 번째는 여러 개의 대표적인 화자모델을 만드는 것이다. 이 방법은 Likelihood ratio sets, cohorts, background speakers 등 여러 가지로 불린다. 배경화면의 우도는 다음과 같이 계산된다.

$$P(X|\lambda_{hyp}) = F(p(X|\lambda_1), \ldots, p(X|\lambda_N))$$

여기서 F( )함수로는 평균, 최대값 등이 사용된다. 두 번째 방법은 음성을 모두 모아 하나의 모델을 만드는 것이다. 이것은 General model, world model, universal background model등으로 불리는데, 전자보다 높은 성능을 얻을 수 있고, 화자 적응에 이용할 수 있는 장점이 있다. 배경화자를 표현하기 위한 모델은 특정 화자에 상관없이 모든 화자의 공통적이고 일반적인 특징을 표현하도록 학습된다. 즉, 입력된 음성이 ID가 제시된 특정 화자의 음성인지 아닌지를 판별할 때, 입력 음성이 화자모델의 특성에 더 가까운지, 일반적인 화자의 모델인 UBM에 더 가까운지를 비교하여 일반 화자 모델에 더 가까우면 거절하게 되는 것이다. 배경화자 모델은 시스템이 사용되는 환경을 고려하여 만들 수도 있다. 주로 사용되는 환경이 전화선이라면 전화선을 통해 녹음된 음성자료를 이용하여 UBM을 만들게 되고, 남녀 구분이 가능한 환경이라면 남녀 각각의 모델을 만들면 더욱 정밀한 배경화자모델을 만들 수 있게 된다.

UBM을 생성하는 방법에는 두 가지 접근방법이 있다. 첫째는 가용한 모든 데이터 동시에 사용하여 모델을 학습하는 것이다. 이때 주의해야 할 것은 마이크, 성별, 전송선 등 여러 환경의 데이터를 균형 있게 사용해야 한다. 그렇지 않으면 배경모델이 특정한 환경에 치우치게 되어 올바른 비교를 할 수 없게 되기 때문이다. 예를 들어 학습데이터에서 남녀의 화자비율에 불균형이 있으면 배경모델이 어느 한 성별의 특성을 주로 갖게 된다. 두 번째는 각 종류별로 모델 생성한 다음 조합하는 것이다. 이것은 데이터 불균형을 해소하는 장점이 있다. 예를 들어 남녀 화자 배경 모델을 각각 따로 만들고 두 모델을 합하면 균등한 특성을 가지는 모델을 합성할 수 있다. 예를 들어, Gender-

independent UBM을 생성할 때 남녀 각각 1024 mixture를 가지는 GMM을 만들고 두 모델의 Gaussian들을 합하여 정규화하면 2048 mixture를 가지는 하나의 UBM을 생성할 수 있다.

5. 화자 모델 적응 (Adaptation)을 이용한 화자모델 생성 [3]

일반적으로 화자모델을 생성할 때는 각각의 화자별로 ML (Maximum Likelihood) 학습 방법을 이용하여 각각의 모델을 생성한다. 그런데, 화자인증시스템의 경우 대부분 학습자료를 충분히 얻기 어려운 경우가 많다. 학습 자료를 많이 요구하면 사용자가 불편함을 느끼기 때문이다. 따라서, 적은 학습 자료를 이용하여 높은 인식성능을 얻는 방법이 중요하다.

적은 학습자료를 효과적으로 이용하는 방법에는 화자적응을 이용한 방법이 있다. 이것은 UBM으로부터 적응을 통하여 화자모델을 학습하여 각각의 화자모델을 생성하는 것이다. UBM은 화자 독립의 광범위한 영역을 모델링하게 되고, 화자적응된 모델은 화자 데이터에 의한 화자종속 튜닝이 된다. 기존의 방법을 사용할 때, 학습에 사용되지 않은 음운환경이 테스트에서 나타나면 이것은 본인의 음성이 아닌 것으로 간주되어 우도값이 매우 낮아지므로, 정상 화자에게 나쁜 영향을 미쳐 시스템의 성능에 치명적인 영향을 미치게 된다. 적응 방법을 사용하게 되면, 미지의 음성에 대해서는 일반적인 화자의 우도값이 되므로 급격한 성능저하를 막을 수 있다

UBM의 적응을 통한 화자모델의 생성은 성능향상 외에도 기억용량 요구량 감소와 속도개선의 효과가 있다. 적은 양의 학습 데이터만을 사용하기 때문에 화자적응은 화자모델의 일부만 적응이 되므로 각각의 화자모델에서 UBM과의 차이만을 저장하는 것이 가능하다. 적응 UBM에서 인식속도를 향상시키는 방법은 다음과 같다. 각각의 화자모델의 mixture는 UBM의 mixture에 화자적응을 적용하여 만들어지므로 mixture간에 일대일대응이 가능하다. 따라서, UBM의 mixture 와 적응된 GMM 의 mixture 와의 관계 이용하면 인식 속도를 향상시킬 수 있다 먼저 UBM에서의 우도 $\log p(X \mid \lambda_{ubm})$ 를 계산하여 최대 우도를 가지는 C개의 mixture를 결정한다. 그리고, 적응된 화자모델에서 관련된 C개만을 이용하여 우도 $\log p(X \mid \lambda_{hyp})$ 를 계산한다 이 방법을 사용하면 UBM의 mixture 수가 M 일때 M+C Gaussian 계산이 필요하지만, 비적응 시스템에서는 2M 번의 계산이 필요하게 된다.

UBM에서 화자 적응하는 과정은 다음과 같다. 먼저 다음과 같이 각 mixture 의 우도를 구한다.

$$\Pr(i \mid x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}$$

그리고, 각 mixture 의 가중치, 평균과 분산은 다음 식으로 갱신된다.

$$\hat{w}_i = [\alpha_i^w n_i / T + (1-\alpha_i^w) w_i]\gamma$$
$$\hat{\mu}_i = \alpha_i^m E_i(x) / T + (1-\alpha_i^m)\mu_i$$
$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1-\alpha_i^v)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2$$

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho}, \quad \rho \in \{w, m, v\}$$

여기서 $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$는 적응계수 (Adaptation coefficients)라고 하며, r^ρ = scale factor 는 mixture 의 합이 0이 되도록 한다.

6. 최근의 연구방향

최근의 연구방향은 주로 GMM을 기본으로 하고, 속도 개선이나 모델 구조변경 등에 의한 성능 개선에 주력하고 있다 속도 개선에는 GMM에서 혼합물을 적절히 선택하여 계산량을 줄이는 방법을 사용하며[5][6], 모델 구조 변경에는 트리형태의 모델 등이 있다[7]. 그밖에 support vector machine 이나 신경회로망과 GMM의 조합을 사용하는 다양한 방법 등이 연구되고 있다.

III. 국내외 시장동향

본 장에서는 주로 대중매체를 통해 알려진 국내외 업체들의 동향을 살펴본다

1. 국내시장동향

화자인식 기술은 기본적인 알고리즘과 특징 등이 음성인식 기술과 유사하므로 현재 음성인식 기술을 보유하고 있거나 개발 중인 대부분의 업체가 기본적으로 취급하고 있다. 대표적인 국내 기업으로는 보이스웨어 (VoiceCop™) , SL2, 디앤엠테크놀로지 (SV-100 SDK), MPC(Say②me-pass), 보이스텍 (Saygaurd), 웹프로텍 등 100개에 가까운 업체가 있다. 그렇지만 화자인증은 보안이라는 특성상 실패했을 때의 비용이 커질 수 있고, 아직 기술적으로도 사용자가 충분히 만족할 만한 수준에 이르지 못하고 있어 크게 활성화 되지는 못하고 있다. 최근 화자인식으로 가장 활발한 활동을 보이

는 대표적인 업체는 예스테크놀로지와 한국파워보이스가 있다.

예스테크놀로지는 화자인증 출입통제관리 시스템 'SSV 937'을 출시하고 관련 특허를 출원하였다. 예스테크놀로지가 개발한 시스템은 네트워크 기반으로 전화망과 IP망을 이용하며 기존 보안 업체가 설치한 출입 시스템에 사람 음성의 특징을 판단하는 모듈과 출입을 제어하는 출입통제 모듈만 있으면 돼 별도의 추가 장치가 필요없다.

한국파워보이스는 오픈세사미라고 하는 화자인식솔루션을 2002년 출시하였다. 이 제품은 '안녕하세요'와 같은 자연스러운 단어발성만으로 96%이상의 인식율과 0.08% 이하의 사칭율을 자랑하며 특히 녹음시와 다른 마이크를 사용해도 92%의 인식률이 가능하다고 회사 측은 설명했다. 또한 녹음한 목소리의 경우 고음과 저음에서 찌그러짐이 발생하고 주파수적인 특징이 달라지기 때문에 실제로 말하는 것과 구별해 낼 수 있다고 설명했다. 이 회사는 2003년 2월 광운대학교 종합정보서비스에 화자인증솔루션을 공급하였다. 이 솔루션은 학생·교수·교직원 등이 종합정보서비스를 이용하기 위해 로그인 할 때 본인의 신원을 자신의 음성으로 확인하게 된다. 종합정보서비스는 교수·학생·교직원 등 학교 구성원들에게 수강신청 및 조회, 성적처리 및 조회, 학사일정 확인 등의 서비스를 제공하는 것으로, 기존에는 ID와 패스워드를 직접 입력해 사용했다. 그러나 이 서비스는 음성인식 기술을 적용한 후 ID만 입력하 한 후 자신의 음성 패스워드를 PC나 노트북의 마이크를 통해 얘기하면 바로 본인 확인 과정을 거쳐 로그인 할 수 있도록 바뀌었다. 이 서비스를 이용하기 위해 사용자들은 최초 접속시 자신의 음성 패스워드를 3회에 걸쳐 녹음하게 된다. 이는 학교 음성 데이터베이스(DB)에 저정되며 로그인 시 사용자의 음성과 대조하는 데 사용된다.

한국파워보이스는 또한 음성인증 기술을 이용한 응용소프트웨어 솔루션을 일본 시스템케이코퍼레이션에 공급하기로 하고 50만 달러 규모의 공급계약을 체결하였다. 제품을 구매한 시스템케이코퍼레이션은 일본내에서 ASP(Application Service Provider) 사업, IDC(Internet Data Center) 운영과 소프트웨어 유통을 수행하는 중견기업으로 한국파워보이스의 음성인증솔루션을 자사의 서비스시스템에 적용하고, 일본 내 솔루션 공급과 유통과정을 통해 마케팅 채널로 활동하기로 하였다.

2. 해외 시장 동향

해외에서도 화자인식 기술의 실용화는 크게 활성화되어 있지 않으나, 콜센터에 적용한 사례들이 다수 보고되고 있다.

DST Retirement Solutions 은 TRAC™ 이라고 하는 recordkeeping platform 에 화자인증을 적용하였다.

Maxxar Corporation도 자사의 Natural Language Speech Recognition platform 에 화자인증 기능을 추가하였다.

2003년 3월, Persay LTD는 Persay FreeSpeech™ 이라고 하는 음성 인증 솔루션을 Leumi 은행에 IBM Global Services Israel 이 구축한 콜센터의 일부로 설치하기로 했다고 발표하였다. 이 시스템은 사용자의 트랜젝션이 진행되는 동안 수초동안에 사용자를 인증하여 PIN 코드나 비밀번호없이 사용할 수 있게 해준다.

2003년 4월, SpeechWorks International, Inc., 는 SpeechSecureTM speaker verification software 에 대하여 두개의 특허를 획득하였다. 한가지는 화자모델의 갱신에 관한 것이고, 또 한가지는 녹음된 음성을 탐지하는 기술이다.

2002 년 12월, 화자인증 업체인 Vocent 는 음성으로 비밀번호를 초기화 해주는 Voice Secure™-Password Reset 2.0 을 발표하였다.

2002년 11월, Scandinavian IT Group 과 Voice Provider 는 항공사인 SAS에 화자인증시스템을 설치하였다. 이 시스템은 스웨덴, 노르웨이, 덴마크의 SAS 직원들이 비밀번호를 잊어버렸을 경우 음성으로 인증하도록 하여 새로운 비밀번호를 주는 방법으로 처리시간을 절약할 수 있게 해준다.

2002년 10월, Convergys Corporation 은 SpeechWorks International Inc. 의 솔루션을 이용하여 통신회사들의 이용자가 회선 제공자를 바꿀 때 인증해주는 시스템을 개발하였다.

Aculab 은 SpeechTek 2002 에서 speaker verification and identification (SVI) software 를 발표하였다.

Ⅳ. 결론

본 논문에서는 화자인식의 기술을 대략적으로 기술하고 국내외 시장동향을 살펴보았다. 최근의 음성인식 방법과 마찬가지로 화자인식도 주로 통계적인 모델을 이용하고 있다. 음성인식과 같은 특징을 사용하고, 유사한 통계적인 모델을 사용하여 기존의 방법에 비하여 성능은 크게 향상시킬 수 있었지만, 아직도 어려운 환경에서 사용자가 안심하고 사용할 수 있을 만큼의 성능은 얻지 못하고 있다 화자인식의 실용화에 어려

음을 주는 요인도 음성인식과 마찬가지로 잡음과 전송 선로의 차이 등이지만, 그 정도는 더 심하다고 할 수 있다. 화자인식의 실용화를 위해서는 음성인식에서와는 달리 화자간 변이를 잘 표현할 수 있는 새로운 특징의 추출과 잡음 및 전송선로의 차이를 극복할 수 있는 모델링이 선행되어야 할 것이다.

참고문헌

[1] Joseph P Campbell, JR, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, Vol 85, No 9, September 1997, pp. 1437-1462

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

[3] D A Reynolds, T F Quatieri and R B Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing 10, 19-41, 2000

[4] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in Proc. Eurospeech, 1997.

[5] R. Auckenthaler and J. Mason, "Gaussian selection applied to text-independent speaker verification," in Proc. A Speaker Odyssey-Speaker Recognition Workshop, 2001

[6] B. L. Pellom and J. H. L. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," IEEE Signal Processing Lett., vol. 5, no. 11, pp. 281 - 284, 1998.

[7] Bing Xiang, "Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, September 2003 pp447-456