화 자 인 식

▌이기용 / 숭실대학교 교수

화자인식이란 ?

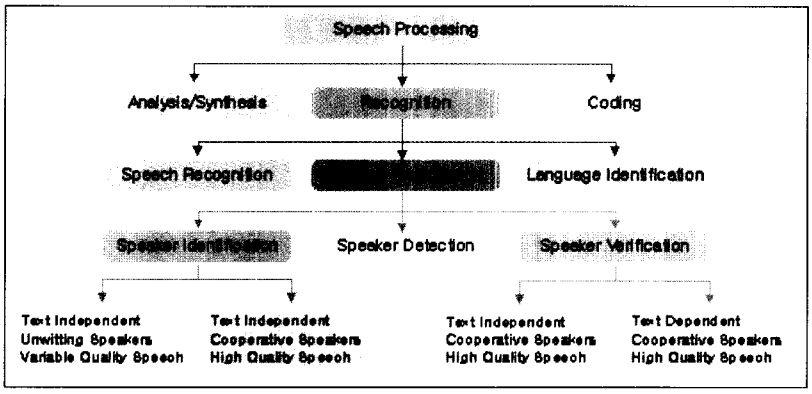

그림 1 음성 처리 분야 -[참고문헌 1]

음성신호에는 음운정보 뿐만 아니라, 각 사람 개개인을 구별할 수 있는 독특한 성문정보를 가지고 있다. 이러한 이유로 음성신호를 이용하여 그림 1과 같이 많은 음성처리를 이용하여 응용분야에 적용할 수 있다.

사람은 누군가의 음성을 듣고, 그 사람이 누구인지를 구별하는 능력을 가지고 있으며, 또한 그 음성의 주인공이 '갑' 이라는 사람인지 아닌 지를 확인하는 능력도 가지고 있다.

다시 말하면 사람에게는 각 개인의 음성에 포함된 특징, 즉 개인 정보를 추출하고 이를 인식하는 시스템을 가지고 있다는 것을 의미한다. 디지털화된 음성 신호를 이용하여 컴퓨터로 하여금 사람이 인식하는 것과 같이 화자의 신원을 식별(identification)하거나 확인(verification) 할 수 있도록 하는 연구를 화자인식(speaker recognition)이라고 하며, 이러한 화자인식을 수행하는 시스템을 화자인식 시스템이라고 한다.

사람은 귀를 통해 이 음성 신호를 받아들이고, 뇌로 전달시켜 누가 어떤 말을 했는지를 인식할 수 있다. 이러한 과정을 마이크를 이용해 음성 신호를 받아들이고, 컴퓨터 알고리즘을 동작시킴으로써 사람이 인식하는 것과 유사한 형태로 컴퓨터에서 수행할 수 있다. 즉, 컴퓨터를 통한 화자인식 과정은 사람이 뇌에서 하는 역할을 흉내 내는 과정이라고 할 수 있다. 이러한 화자인식을 다음과 같이 분류한다.

화자 식별은 발성된 음성 신호가 등록된 화자들 중에서 어떤 화자인지 골라내는 것이고, 화자 확인은 발성된 음성신호가 등록된 화자의 음성과 일치하는지를 판정하는 것으로, 발성한 화자와 등록된 화자와의 확인과정을 통하여 임계값보다 유사도가 큰 경우 수락(accept)하고, 임계값보다 유사도가 작은 경우 거절(reject)한다.

화자 인식 방법을 발성방법에 따라 분류하면 문장 종속(text-dependent)형과 문장 독립(text-independent)형이 있다. 문장 종속형 화자 인식은

학습과정과 테스트를 위하여 미리 정해놓은 단어나 문장을 사용하고, 문장 독립형 화자 인식은 학습과정과 테스트과정에서 발성하는 음성신호에 제한을 두지 않는다. 문장 종속 및 독립형은 녹음기 등을 이용하여 등록된 화자의 음성을 통해 등록을 시도할 때 등록된 화자로 인식하는 문제점이 발생하므로, 단어나 숫자의 나열이 아닌 임의의 문장을 제시하여 검증을 하는 문장 제시형(text-prompted)방식도 등장하였다.

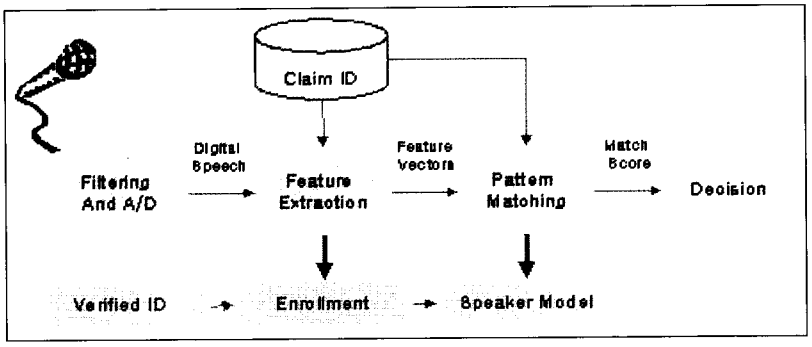

그림 2 화자인식 단계 – [참고문헌 1]

현재 화자 인식을 위한 방법으로 DTW, HMM, 그리고 GMM 방법 등이 많이 사용되고 있다. 각 방법을 표 1에 나타내었다.

이러한 화자 인식을 위하여 사람이 어떤 사람이 말을 했는지를 알아내는 과정을 간단하게 설명하면 다음과 같다. 어떤 사람과 대화를 나누면 사람은 그 화자의 특징을 분석하여 이를 뇌 속에 저장한다(훈련과정). 다시 그 화자와 대화를 나누게 되었을 경우에는 과거에 저장된 자료와의 매칭을 통하여 어떤 화자인 지 아니면 해당화자인지 아닌지를 찾아내는 것이다(인식과정). 컴퓨터를 통한 인식 과정도 이와 유사하게 진행된다. 그림 2는 컴퓨터를 통한 화자 인식 방법을 두 가지 과정으로 나누어 조금 더 자세히 본 것이다.

표 1 화자인식 방법들의 비교

	DTW(Dynamic Time Warping)	HMM(Hidden Markov Model)	GMM(Gaussian Mixture Model)
목 적	음성의 지속 시간의 차이로 인한 영향을 제거하기 위해 단순히 입력음성과 기준 음성의 양끝을 서로 맞추고 선형적으로 늘리거나 줄여 패턴을 비교하는 선형 시간 정렬 방식의 문제점을 해결하기 위해 사용	관측이 불가능한 프로세스를, 관측이 가능한 심벌로 발생시키는 프로세스를 통해 추정하는 이중 확률 프로세스이기 때문에 음성과 같이 다변성이 많고 발생 과정을 알 수 없는 프로세스를 표현하는데 적절한 모델링하기 위해 사용	화자의 목소리에 대응되는 음향 공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적 클래스는 화자를 구별하는데 이용되는 화자의 성도에 대한 정보를 이용하여 모델링하기 위해 사용
방 법	음성신호 내부의 각 부분도를 늘이거나 줄여가며 비교해 전역 유사도를 측정할 수 있는 알고리즘	음성 패턴의 각 특징을 상태 천이에 의해 서로 연결된 상태의 모임으로서 상태 천이 확률과 출력 확률로 표현 1) 현재의 상태는 바로 그 이전의 상태에만 의존 2) 각각의 출력은 서로 독립적.	출력 밀도 함수가 한 개의 상태로만 이루어진 CHMM (continuous HMM)의 한 형태로, 여러 mixture들의 Gaussian확률 분포의 최대 유사도를 측정하여 화자인식에 사용
응 용	소용량 어휘의 고립 단어 인식에 이용 인식 시간이 많이 소요된다는 단점이 있지만 인식률이 높아 집적회로 칩으로 구현돼 상용화	대용량 음성인식 시스템에 주로 이용 DTW 방식에서의 템플릿 구축보다 훨씬 크지만 인식 과정에서의 계산량은 훨씬 적고, 시간과 스펙트럼 양쪽에서 확률 통계적인 방법을 사용함으로써 클러스터링 효과를 얻을 수 있다	문장 독립형 사용 문장 제시형 사용 소용량 어휘에서도 사용이 가능하여 화자 인식에 주로 사용됨.
문제점	warping 함수를 찾는 과정에서는 많은 계산이 필요하므로 기준 음성의 개수가 많아지면 실시간 처리가 어려워짐.	파라메터수가 많이 필요하여 짧은 문장에 사용이 불가능 모든 확률이 현재 상태에만 의존한다고 처리하므로 조음효과를 충분히 모델링 하기에 부족	문장 독립형이라 다른 문장을 발성하여도 인증가능 EM알고리즘의 문제점 - 초기값의 영향을 많이 받음 - 잡음에 민감

첫 번째 과정은 많은 음성 데이터를 이용하여 훈련(training)을 시킨 후, 해당하는 화자에 적합한 모델(speaker model)의 대표 값을 구하는 부분이고(훈련과정), 두 번째 과정은 실제 음성 신호가 입력되었을 경우 이를 인식하는 단계이다(인식과정). 화자인식 단계는 입력 처리부, 특징 추출 (feature extraction), 패턴 매칭 (pattern matching), 그리고 판단 (decision) 단계로 나눌 수 있다. 물리적으로 음성 신호는 아날로그 형태를 취하기 때문에 입력 처리부에서는 A/D 변환기를 이용해서 디지털 신호로 변환한다. 특징 추출 단계에서는 각 개인의 음성의 특징을 잘 표현할 수 있는 변수를 구하여, 훈련을 통해 미리 얻어놓은 대표 값과 비교한다. 최종 판단 단계에서는 훈련을 통해 얻은 값과 새로 테스트하는 값과의 차이를 구한 후 이것을 임계값과 비교해서 최종 결론, 즉 식별인 경우에는 어떠한 화자인지를 구분하기 위한 ID를 결정하고 인증인 경우에는 그 사람인지 아닌 지를 판단한다.

화자 인식을 적용하는 방법은 매우 다양하지만, 가장 쉽게 생각할 수 있는 시나리오는 그림 3에 표시해 놓은 것과 같이 스마트카드 (smart card)를 이용하는 것이다. 훈련을 통해 화자들이 가지고 있는 특징 벡터들의 대표 값들을 구하고, 이를 스마트카드에 저장한다. 실제 인증 시에는 마이크로부터 얻어진 신호를 이용하여 특징 벡터를 얻고 이를 스마트카드에 이미 저장되어 있는 대표 값들과 비교하는 단계를 거치는 것이다.

화자인식 시스템의 성능은 훈련을 통해 얻은 특징 벡터들의 값과 실제 테스트가 수행되는 과정에서 얻은 특징 벡터들의 값이 일치할 경우에 최상의 결과를 얻을 수 있다. 그러나, 실제 시스템에서는 이러한 환경이 일치되기를 기대하기 어렵다. 예를 들어, 반향 (reflection)이나 주변 잡음이 존재하는 환경, 그리고 입력 마이크의 특성이 다를 경우에는 성능이 저하될 수밖에 없을 것이며, 화자의 노화나 감기 등 육체적 상태가 훈련 시에 비해 다르다면 역시 좋은 성능을 기대하기는 어려울 것이다. 이를 해결하기 위해 잡음을 전처리 단에서 미리 제거하는 방법, 그리고 채널 영향에 의해 발생하는 왜곡 등을 보상하기 위한 방법, 그리고 육체적 상황 변화에 강인한 음성 특징 벡터를 얻기 위한 연구가 활발히 진행 중이다.

응용 분야

본고에서는 현재 화자인식이 사용되어질 수 있는 응용 시스템들에 대해 살펴보려고 한다.

음성 포탈

자료에 의하면 2003년과 2005년 사이에서 세계적으로 전개되는 휴대 전화의 수는 10억개 이상이 될 것이며, 특히, 2004년까지는 새로운 무선 전화기 70% 와 새로운 PDA 80 % 가 인터넷을 접속 가능하도록 하는 어떤 형태로 특징화 될 것이라고 한다. 이때 음성 신호처리 기술(음성 인식, 음성합성과 화자

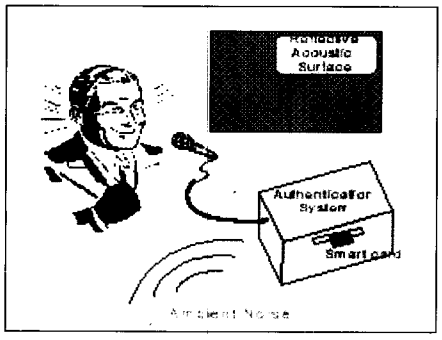

그림 3 화자 인식 시나리오 – [참고문헌 1]

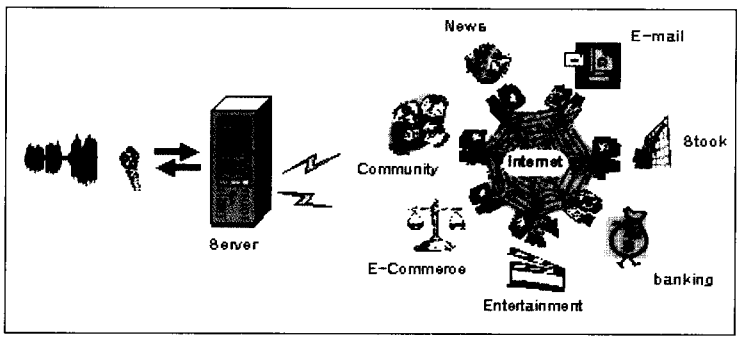

그림 4 음성 포탈 시스템 서비스 예

인증)은 가장 중요한 역할을 담당할 것으로 예상된다. 흔히 음성 포탈(voice portal)이라고 불리는 시스템은 음성을 이용하여 기존의 웹에서 제공하는 서비스를 대체하거나 전화망을 이용해 필요한 정보를 얻고자 하는 연구 분야이다.

지금까지 음성 포탈에 응용되는 대표적 연구 분야는 음성인식이었다. 즉, 음성 포탈 시스템은 어떤 사람이 말했는지에 관계없이 그 사람이 어떤 말을 했는지를 인식하고 이를 바탕으로 사용자가 원하는 결과를 대답해주는 방법이 주목적이었다. 그러나, e-commerce와 같이 신뢰성을 필요로 하는 경우에는 기존의 음성인식 시스템과 아울러 화자인증 시스템이 함께 결합된 형태를 취하는 시스템으로의 확장이 불가피한 형편이다. 이 경우에는 웹의 대중적인 부분과 개인적인 부분을 구분하고 사용자 각각에 따라 차별적으로 접근 가능하도록 함으로써 다양한 형태의 서비스 형태를 제공할 수 있다는 장점도 있다.

새로운 미개척분야(New Frontiers)

그 동안 상용화된 화자인식 시스템은 보안이나 모니터링 시 문장 종속형이나 문장 제시형 방법으로 화자를 인증하였다. 문장 종속형이나 문장 제시형 방법에서는 인증시 시스템에서 요구하는 문장을 말해야만 원하는 결과를 얻을 수 있었다. 그러나, 보안 정도를 강화하기 위해서는 문장에 독립적인 방법, 즉 정해진 문장이 아니라 사용자 편의에 따라 문장을 마음대로 변경하는 경우에도 동작할 수 있는 시스템이 필요하다.

음성 처리 기술과 음성에 기초를 둔 생체학(biometrics)에 관심이 많은 미국, 유럽 등에서는 이를 이용하여 여러 기술 프로젝트들을 지원하고 있다. EC에서 지원하는 Picaso 프로젝트, 미국의 NIST (National Institute of Standards in Technology)에서 지원하는 화자 인식 경쟁 연구는 그 대표적인 프로젝트라고 할 수 있다. 뿐만 아니라, 이스라엘과 다른 여러 나라의 정부에서 지원하고 있는 프로젝트들도 문장 독립 화자 확인의 상용화를 목표로 활발한 연구가 진행되고 있다.

몇몇의 정부가 지원하는 또 다른 프로젝트로는 audio mining이 있다. Audio mining이란 영화, TV, 라디오 방송들과 같은 audio-visual 소스들을 찾은 후 이를 분류하는 작업을 말한다. 문서와는 다르게, 오디오 소스들을 모으는 일은 컨텐츠와 참여자들(화자들)에 따라 분류되어야만 한다. 결국, audio mining 시스템은 음성 인식을 기본으로 하며, 문장 독립 화자 식별 시스템과 정보 검색 시스템이 결합된 형태라고 할 수 있다. 음성인식은 소스들 가운데에 정보를 음성으로 표기하는데 사용되어지고, 화자식별은 화자들을 식별하는 데, 그리고 정보검색은 그것을 찾을 수 있는 정보를 분류하고 인덱싱하는데 사용된다. 미국의 DARPA(Defense Advanced Research Projects Agency)와 Esprit DiVan이 그 대표적인 프로젝트라고 할 수 있다.

기타 응용 분야로는 범죄 예방 (criminal justice), 고용자 확인 (authenticating employee)등을 들 수 있다. 고용자 확인 분야에서는 근무태도 관리를 위

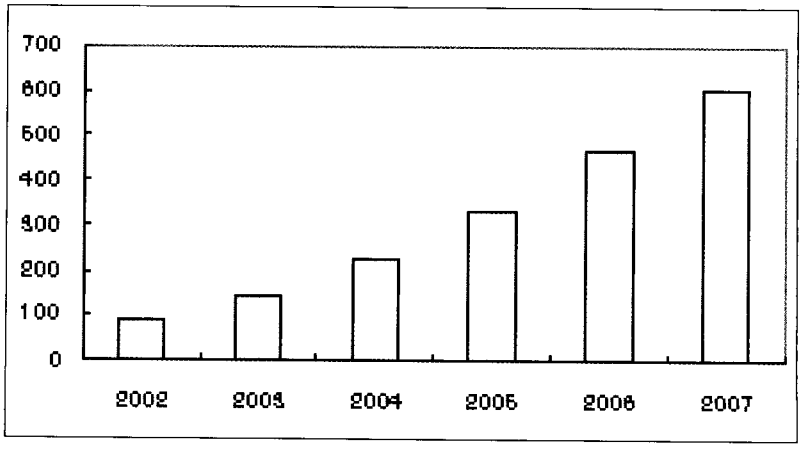

그림 5 년도별 음성 관련 시장 규모 현황 및 예측(단위:백만달러)
- copyright 2002 international biometric group -

한 표준 전화망과 내부 데이터망으로 사용하여 종업원들을 확인하는 것이다.

향후 전망

Biometrics산업의 전망을 예견하는 사람들은 음성을 이용한 시스템이 biometrics 시장Consultants는 170억에서 200억 달러의 규모로 내다보고 있다. Elsevier Advanced Technologies의 최근 biometrics 시장보고서에서는 2003년까지 75%의 급속한 성장률을 기반으로 9100만 달러의 수입을 내다보고 있다.

그림 5는 년도별 음성 관련 시장 규모 현황과 예측을 나타낸 것이다(international biometric group에서 발췌). 2002년부터 음성 산업은 매년 성장하여, 2007년에는 시장규모가 600만 달러에 달할 것으로 전망되고 있다. 또한 국내에서는 전화, 휴대전화, 신형 컴퓨터 등은 마이크로폰을 통해 상대방을 인식할 수 있는 소프트웨어를 갖추고 있는 음성인증 시장이 휴대전화 부문에 대한 인증 기술의 적용으로 향후 2년 내 꾸준히 성장할 것으로 기대하고 있다. 특히, 주식, 뱅킹 등에 화자인식 기술이 적용이 확산되면 점유율과 시장규모는 급격히 증가할 것이다. 음성인증은 주로 물리적 접근 제어 응용에 사용되어 왔으나, 콜센터와 같은 기능은 음성인증 시장의 강력한 유인책이 될 것이다. 표 2는 국?내외 음성 기술을 보유한 회사들이다.

화자인증 시장 형성에 있어서 원동력은 음성인식 기술의 발달에 기인한다. 예를 들어 음성인식 시스템을 성공적으로 사용하고 있는 몇몇 미국의 투자 전문 회사들에서는 이미 그 시스템에 보안 기능을 확장 적용하려고 노력하고 있다. 이러한 응용 시스템에서는 음성인식 시스템의 첫 번째 단계에서 발성된 입력 (예를 들면 사용자 번호)을 이용하여 음성인식과 화자인증을 동시에 수행한다. 결론적으로 두 기술을 함께 사용하는 전반적인 음성 솔루션으로 단 시간에 더 안전하게 인증을 할 수 있도록 도와주는 방향으로 응용 분야가 확장되고 있는 것이다. 홈쇼핑 네트워크에서 자동화된 주문 시스템이 그러한 시스템의 또 다른 실례이다. 전화-중심 큐(queue)에서는 집에서 기다리고 있는 고객들을 위해 전화상에서 자동적으로 제품을 주문하도록 디자인 되어 있으며, 음성 인식과 화자인증이 결합된 시스템을 이용해 자동적으로 보완 인증(backup verification)을 할 수 있도록 되어 있다. 이 시스템에서는 기본적인 질문 내용에 대한 해답을 비교해서, 만일 최초의 인증 결과가 맞지 않으면 시스템은 기재되었던 항목에 링

표 2. 국·내외 음성 기술 보유 회사

	회사명	분 야	홈페이지 주소
국 외	Nuance	TTS(text to speech), Speech Recognition	http://www.nuance.com
	Scansoft	TTS(text to speech), Speech Recognition	http://www.lhsl.com
	SpeechWorks	TTS(text to speech), Speech Recognition, Speaker Verification	http://www.speechworks.com
	Sensory	STT(speech to text), TTS(text to speech), Speech Recognition	http://www.sensoryinc.com
국 내	voiceware	TTS(text to speech), Speech Recognition, Speech Synthesis, Speaker Recognition	http://www.voiceware.co.kr
	corevoice	Speech Recognition, Speech Synthesis	http://www.corevoice.co.kr
	voicepia	Speech Recognition, Speech Synthesis, Speaker Recognition, Speech Coding	http://www.voicepia.co.kr
	webprotek	Speaker Detection, Speaker Recognition	http://www.webprotek.com

크되어 있는 질문을 닥치는 대로 선택하는 질문-응답 세션을 불러낸다. 호출 루틴이 믿을 수 있는 것임이 증명되면, 보완이 요구되는 고객 관계 관리 시스템을 불러내어 해당 요구 사항을 처리해 줄 수 있다. 즉, 음성 인식과 화자 인식을 결합한 시스템은 그 응용 분야가 다양해질 수 있으며, 보완이 필요한 경우에도 효과적으로 대체할 수 있는 방법이라고 볼 수 있다.

[참고문헌]

[1] J.P. Campbell, "Speaker Recognition : A Tutorial", IEEE, Proceeding, Vol. 85, No. 9, pp. 1437~1462, 1997.

[2] S. Furui, "Recent advances in speaker recognition", Pattern Recognition Letters, 18, pp. 859~872, 1997.

[3] R.J. Mammone, X.Zhang and R.P. Ramachandran," Robust Speaker Recognition-a feature-based approach," IEEE Signal Processing Magazine, Vol. 13, No.5, Sep. 1996.

[4] H. Bourlard, N. Morgan," Speaker Verification a Quick Overview," IDIAP, pr 98-12, 1998.

[5] D.A. Reynolds, and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. SAP, vol. 3, No. 1, pp. 72-83, 1995.

[6] D.A. Reynolds," An overview of Automatic Speaker Recognition Technology", ICASSP 2002, vol 4, pp. 4072-4075, 2002.

[7] 한국정보보호진흥원 : http://www.kisa.or.kr/K_trend/KisaNews/200112/special_report_04.html

[8] International biometric Group http://www.biometricgroup.com/reports/public/market_report.html