情 報 保 護 學 會 誌
第 12 券 第 2 號, 2002. 4

환경 변이에 강인한 화자 인식 기술

김 유 진*, 정 재 호*

요 약

음성 인식 기술과 뿌리를 공유하는 화자 인식 기술은 지난 수십 년간의 연구결과로 괄목할 만한 진보가 이루어졌으며 최근에는 일반화될 수 있으리라는 기대를 가지도록 하기에 충분했다. 하지만 이러한 기술이 실제 환경에 적용되었을 때, 발성 환경을 제어할 수 없으며 그 결과 훈련 환경과는 다른 환경에서 발성된 음성을 인식해야하는 이른바 '불일치 조건(mismatch condition)' 현상이 발생하게된다. 초기에는 이 현상을 극복하기 위해 잡음 자체를 모델링하고 제거함으로써 훈련과 인식 환경의 차이를 일정하게 정규화(normalization)해주는 연구가 진행되었다. 하지만 최근에는 잡음에 의한 왜곡의 모델이 복잡하고 실제 인식 성능에 직접적으로 나타나지 않는 문제점을 추가로 극복하기 위해, 훈련과 인식 환경의 차이를 보상해주는(compensation) 연구가 활발히 진행되고 있다. 본 논문에서는 기본적인 화자인식기술과 함께 성능저하를 일으키는 불일치 요인들 및 그것들을 극복하기 위한 기술들을 소개하고자 한다.

Ⅰ. 서 론

개인의 정보 욕구가 증가하고 유무선 통신망, 인터넷과 같은 통신 인프라가 급속도로 보급되면서 다양한 망을 통한 음성 인터페이스 기술의 실용화 요구가 증가하고 있다. 이러한 음성 인터페이스 기술은 인간의 가장 기초적인 의사 전달 수단인 음성을 이용함으로써 휴대성, 편리성 그리고 효율성 등을 높일 수 있다. 이미 증권정보, 114 전화번호 안내 등을 통해 음성 인식, 음성 합성 기술이 실용화된 예를 쉽게 경험할 수 있다.

한편 음성에는 언어 정보 외에 화자(speaker) 신원(identity)에 대한 정보도 포함되어 있음을 쉽게 깨달을 수 있다. 우리는 부지불식간에 음성에 포함된 발성 습관 또는 화자의 성도(vocal tract)의 구조적인 차이로 인한 특성을 통해 화자를 구분하는 것이다. 디지털 음성 신호 처리의 한 분야인 화자 인식 기술은 이러한 음성 내에 포함된 화자의 신원 정보를 알아내고 이용하는 기술이다.[1,2,3,4,5]

화자 인식은 사람마다 구분되는 특징, 특질을 이용하여 그 사람을 인식하는 생체 측정학(biometrics)의 한 분야로도 정의할 수 있다. 음성은 사람마다 고유한 성도에 의해서 생성되므로 지문, 얼굴, 홍채 등과 같이 생체 특징으로 사용될 수 있기 때문이다. 앞으로 화자 인식은 음성 인식 기술과 더불어 다양한 망을 통한 데이터와 시스템 접근을 보다 쉽게 하기 위한 기술로서 또는 다른 생체 특징과의 결합을 통해 보다 높은 신뢰도와 편리함을 제공하기 위한 기술로서 관심을 모을 것으로 예상된다.

본 논문에서는 기본적인 화자인식 기술 분류와 시스템 그리고 실제 환경에서 적용되기 위한 여러 가지 기술들을 살펴보고자 한다. 특히 대부분의 생체 측정학 기술들이 공통적으로 안고있는 환경 변이의 문제를 해결하기 위한 기술들을 중점적으로 소개한다.

Ⅱ. 화자 인식 기술

1. 화자 인식 기술의 분류[1-6]

화자 인식은 크게 어플리케이션의 성격에 따라 화자 식별(speaker identification)과 화자 확인(speaker verification)분야로 나눌 수 있다.

화자 식별 분야는 '말하는 사람이 누구인지'에 대한 답을 위한 분야라고 할 수 있다. 일반적으로 신원 확인을 청구한 화자가 등록된 화자들 중에 누구인지 또는 등록된 화자가 아닌지를 결정하게 된다.

* 인하대학교 전자전기공학부 DSP연구실(egkim@ieee.org, jhchung@inha.ac.kr, HTTP://DSP.inha.ac.kr)

한편, 화자 확인 분야는 '청구된 화자가 본인인가'라는 질문에 대한 답을 위한 분야이다. 따라서 본인임을 주장하는 사칭자와 실제 화자를 구분하는 대부분의 어플리케이션에 적용될 수 있다. 즉 화자 확인은 1명의 화자를 대상으로 하는 작업인 반면 화자 식별은 일반적으로 여러 명의 화자를 대상으로 이루어진다.

다시 화자 식별 분야는 유한 집단(closed-set)과 무한 집단(open-set)에 대한 식별 어플리케이션으로 나눌 수 있다. 유한 집단에 대한 화자 식별 분야는 등록된 또는 알려진 N명의 그룹 내에서 누구인지를 식별한다.(다시 말해 반드시 N명에 속한 화자가 발성한다는 가정을 전제한다.) 따라서 N의 크기가 커지게 되면 매우 어려운 작업이 된다. 반면 무한 집단에 대한 화자 식별 분야는 N명의 그룹 내에 화자가 속하는지의 여부를 판단한다. 따라서 그룹 내에 속하지 않는 허가 받지 않은 사람을 구분하는 작업으로 볼 수 있다. 따라서 화자 확인 분야는 무한 집단 화자 식별의 특별한 경우로 볼 수도 있다.

화자 인식 분야는 다시 발성된 어구의 문맥 정보에 따라 크게 문장 독립(text-independent)과 문장 종속(text-dependent)으로 나눌 수 있다. 문장 독립 화자 인식 분야는 우리가 상대방의 대화를 듣고 인식하는 방식과 동일한 방식으로서 등록 과정과 - 학습과정으로 생각할 수 있다. - 확인 과정에서 임의의 문장을 처리한다. 따라서 화자가 발성한 문장의 의미와는 무관한 음성의 특징을 추출하여 화자의 특징을 훈련, 학습한다. 이러한 특징으로는 피치(pitch), 톤(tone) 그리고 음색 등이 있다. 반면 문장 종속 화자 인식 분야는 일반적인 패스워드 방식과 비슷한 방식으로 항상 동일한 내용의 문장을 발성할 것을 가정한다. 따라서 화자의 성문(voiceprint) 모델을 생성하는 훈련과정에서 발성한 패스워드, 카드 번호 또는 PIN(Personnel Identity Number) 번호 등을 확인 과정에서 기억하고 사용해야하는 제약점이 있다. 하지만 화자가 자유롭게 선택한 패스워드의 의미와 화자의 특징을 동시에 사용하므로 일반적으로 문장 독립 화자 인식 시스템에 비해 성능이 높은 것으로 알려져 있다.

일반적으로 화자 식별 어플리케이션에서는 문장 독립 방식을 사용하는데 훈련 과정에서 수초~수십초에 이르는 대화, 어구로부터 의미와는 관계없는 특징을 추출하여 사용한다. 화자 확인 어플리케이션에서는 문장 종속 방식이 선호되고 있다. 하지만 텔레뱅킹과 같은 강력한 보안이 필요한 분야에 화자 확인 어플리케이션이 적용될 경우 녹음에 의한 사칭자의 접근을 막을 수 없는 단점을 지닌다. 따라서 숫자음과 같은 제한된 단어들을 조합하여 매번 다른 발성을(예, 26-37-84, 42-63-91) 요구하는 변형된 문장 종속 방식인 문장 지시(text-prompt)방식이 사용된다. 특히 이러한 문장 지시 화자 확인 어플리케이션에서는 확인 때마다 달라진 발성이 녹음에 의한 발성이 아닌지를 먼저 확인하기 위해 음성 인식 기술이 사용된다. 문장 지시 화자 확인 분야와 같이 음성 인식 기술이 접목되는 점 때문에 화자 확인 분야는 화자 식별과는 다른 분야로 구분된다.

결론적으로 문장 독립 시스템은 보안보다는 화자 분류/검출(Speaker Classification, Detection) 등의 좀 더 완화된 보안 형태의 어플리케이션에 적용 가능하며 문장 지시 시스템은 강력한 보안이 필요한 신용 거래, 홈뱅킹 등의 분야에 적당하다. 반면 문장 종속 화자 확인 시스템은 비교적 간단한 등록, 확인 과정과 함께 음성을 통해 기존 암호나 물리적인 보안 수단을 대체할 수 있으므로 다양한 응용 분야에 적용될 수 있다.[6]

2. 화자 인식 기술의 구성

화자 인식 기술 또는 시스템은 일반적인 패턴 인식 시스템과 마찬가지로 크게 나누어 특징 추출기(Feature extractor)와 분류기(Classifier)로 구성되는데 좀 더 구체적인 과정을 덧붙인 화자 인식 시스템은 그림 1과 같이 구성될 수 있다.

전처리 알고리즘은 음성 특징 추출의 전 단계로서 변별력이 높은 음성 특징을 추출하기 위한 일련의 과정이다. 이 과정은 음성 신호 영역에서 처리되며 효과적인 음성 분석을 위한 음성 신호 강화(enhancement)와 변별력 높은 화자 특성을 추출하기 위한 음성 신호 분절(segmentation) 등으로 크게 나눌 수 있다.

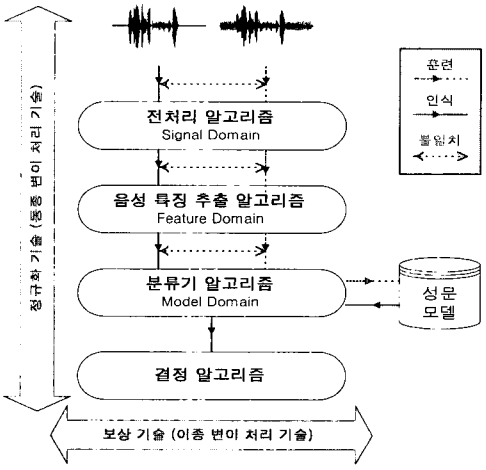
음성 신호 강화 처리는 주파수 특성의 균등화를 위한 고주파 대역 강조의 기능을 갖는 간단한 pre-emphasis와 잡음에 의해 왜곡된 신호로부터 잡음을 제거함으로써 음질을 향상시키는 알고리즘 등을 예로 들 수 있다. 음성 신호 분절은 화자 특성을 잘 포함한 유성음, 비음 등의 음향학적인 구간을 찾아내는 알고리즘들이다.

음성 특징 추출 알고리즘은 전처리 알고리즘에 의해 처리된 음성 신호 구간에 대해서 음성 특징을 추

출하는 과정이다. 가장 선호되는 음성 특징 분석 방법인 LPC(Linear Predictive Coding)는 과거 샘플들의 선형적인 조합으로 현재의 샘플을 예측하는 과정으로서 단구간에 대해 시불변 시스템으로 가정할 수 있는 음성 발성 시스템인 성도(vocal tract)를 효과적으로 모델링 할 수 있음이 증명되었다. 이 분석 결과는 AR(Auto-Regressive)형태의 간단한 all-pole 모델로서 다양한 음성 신호 처리 분야에서 응용되고 있다. 음성 인식 및 화자 인식에서는 단구간의 스펙트랄 포락선을 주요한 음성 특징으로 사용하며 *LPCC(LP Cepstrum Coefficient)*는 LP분석 방법에 의해 간단하게 추출 될 수 있다. 한편 FFT와 같은 주파수 분석법을 기반으로 잡음에 강인한 심리음향 이론을 첨가하여 스펙트랄 포락선을 추출하는 방법도 널리 사용되고 있으며 *MFCC (Mel-Frequency Cepstrum Coefficient)*가 대표적인 음성 특징이다.[7-9]

패턴 인식의 관점에서 분류기에 해당하는 과정으로서 화자 인식에서는 *DTW(Dynamic Time Warping), VQ(Vector Quantization), HMM (Hidden Markov Model),* ANN(Artificial Neural Network) 등이 사용되며, 음성 특징들의 시간적·공간적 분포를 모델링하고 기존의 모델들과 입력된 음성을 비교하여 유사도를 얻는다.[8,9] 최근에는 발성 환경, 시간, 감정 상태에 따른 변화를 정규화하고 적응하기 위해 전통적인 DTW, VQ 등의 알고리즘보다는 통계적이고 확률적인 알고리즘 들이 주로 사용된다. 따라서 문장 종속 알고리즘에서도 HMM 등의 통계적인 알고리즘이 효과적으로 사용되고 있으며 문장 독립 알고리즘의 경우 *GMM (Gaussian Mixture Modeling)*이 좋은 성능을 나타내는 것으로 나타났다.[10] 문장 지시 알고리즘은 하위 단어 단위(sub-word unit)의 화자 모델이 필수적이므로 HMM을 이용한 알고리즘이 선호되고 있다.[11,12]

일반적으로 HMM과 같은 통계적인 모델링 방법의 분류기는 결과가 화자 모델에 대한 입력 음성의 우도(likelihood)로 나타나며 그 분포는 화자의 발성 환경, 시간 경과에 따른 변화뿐만 아니라 발성된 어구에 따라서도 크게 달라진다. 따라서 화자와 사칭자를 구분하는 최적의 문턱값을 결정하기 위해 변이가 적은 우도 측정 방법을 정의하고 적용하는 결정 알고리즘을 적용한다. 일반적으로 이러한 화자간의 우도의 변이는 *pseudo-imposter*모델을 이용한 정규화 방법으로 해결하며, pseudo-imposter 모델은 cohort set을 이용하는 competition based 방법과 world model을 이용하는 qualifier based 방법으로 근사 될 수 있다. 또한 이러한 측정 방법을 절충한 형태도 연구되었다.[13,14,15]



〔그림 1〕 화자 인식 시스템

Ⅲ. 정규화(Normalization) 기술

1. 환경 불일치

훈련, 테스트 그리고 인식으로 이어지는 개발 과정을 거치는 대부분의 패턴 인식 시스템이 그렇듯이, 훈련과 테스트 데이터의 제한된 환경에서 발성된 음성만으로 훈련된 화자 인식 기술이 실제 환경에 적용된 경우, 심각한 성능저하를 초래하게 된다. 이는 모든 어플리케이션 환경을 미리 모델링하거나 제어할 수 없고 결과적으로 훈련 환경과는 다른 환경에서 발성된 음성을 인식해야하는 이른바 '환경 불일치(Environmental mismatch)' 현상이 발생하기 때문이다. 예를 들어 등록을 위한 훈련 과정에서는 조용한 사무실 환경에서 발성하지만 실제 사용은 자동차 소음, 주변 사람들의 대화, 음악 소리 등의 주변 잡음에 노출된 환경에서 이루어지는 것을 말한다. 한편 훈련과 인식 환경의 불일치를 유발하는 **근본적인 환경 변이** 뿐 아니라 인식 환경이 일정하지 않은 **실용적인 환경 변이**에 의해서도 불일치 현상이 발생하게 된다. 예를 들어 유무선 전화망을 통한 어플리케이션 경우, 전화선 환경에서의 훈련이 이루어지더라도 인식에서 접속될 때마다의 채

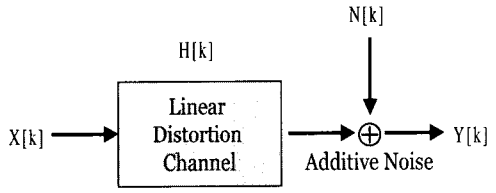
널의 특성이 일정하지 않기 때문에 성능이 일정하게 유지되지 않게 된다. 화자 인식에서의 이러한 불일치 현상은 다른 생체 인식 기술에 비해 매우 일반적이라는 점에 심각성이 있다. 또한 덜 심각하지만 시간, 환경 그리고 심리적인 상태에 따라 화자의 발성이 일정하지 않은 **화자 내(intra-speaker) 변이**로 인한 불일치도 발생한다.[16,17]

일반적으로 '불일치'를 유발하는 환경 변이의 원인은 크게 배경 잡음과 채널 잡음의 2가지로 나눌 수 있다.[18] 일반적으로 배경 잡음은 가산 잡음(additive noise)의 형태로 음성 신호를 왜곡시키게 되고 $y[k]$를 왜곡된 음성신호, $x[k]$를 왜곡되기 전의 음성신호 그리고 $n[k]$를 가산 잡음이라고 할 때, 다음과 같은 식으로 나타낼 수 있다.

$$y[k]=x[k]+n[k] \tag{1}$$

한편 채널에 의한 음성 신호의 왜곡은 시간축에서 채널 성분과 컨벌루션된 형태로 나타나며 다음과 같은 식으로 표현된다. 이는 $h(k)$를 채널 성분이라고 할 때, 음성 신호가 채널에 의해 필터링된 효과를 나타낸다.

$$y[k]=x[k]*h[k] \tag{2}$$



〔그림 2〕 환경 변이 원인의 모델링

표 1에서는 채널 잡음에 의한 인식률의 저하를 보여준다. 실험 결과는 TIMIT 데이터베이스의 DR1 테스트 영역의 38명의 화자에 대해서 8초 가량의 음성 데이터를 통해 화자 모델을 훈련하고 약 3초 가량의 음성 데이터로 문장 독립 화자 인식 실험을 거쳐 얻었다. 음성 특징은 12차 LP 켑스트럼, 분류기로서 46개의 정규분포를 포함한 GMM을 사용하였다.[19]

Clean은 모의 채널을 거치지 않은 경우의 인식률을 나타내며 CMV와 CPV는 각각 실제 전화선 채널을 모사한 필터를 거친 경우이다. 표에서 훈련과 인식에서 동일한 채널일 경우 인식률 저하는 미미하지만 서로 다른 채널일 때 매우 큰 인식률 저하를 보이는 것을 알 수 있다. 따라서 채널 잡음 자체보다는 환경 불일치에 따른 성능 저하가 심각함을 보여준다.

〔표 1〕 모의 채널 불일치에 따른 인식 실험

Clean	CMV CMV	CPV CPV	CMV CPV	CPV CMV
74.7%	74.2%	73.2%	17.4%	15.3%

그림 1에서 볼 수 있듯이 이러한 불일치 현상은 화자 인식 시스템의 전처리, 특징 추출 그리고 분류기 알고리즘에서 각각 나타날 수 있으며 이를 극복하기 위한 직접적인 방법은 잡음 자체를 제거함으로써 불일치를 유발하는 환경 변이를 정규화 하는 방법이다. 이러한 기술은 크게 나누어 특징 기반 접근과 모델 기반 접근 방법으로 분류할 수 있다.

2. 특징 기반 접근(Feature-based Approach)

특징 기반 접근 방법은 전처리 또는 특징 추출 알고리즘에서 환경 변이의 원인을 제거하는 방법이다. 배경 잡음의 경우, 주파수 분석 후 잡음의 파워 스펙트럼을 추정·제거하는 전처리 알고리즘으로서의 *주파수 차감법(Spectral Subtraction)*이 대표적이며 채널 잡음의 경우, 호모몰픽 분석에 의한 켑스트럼의 특성을 이용하여 특징 추출 후 장구간 평균을 차감하는 *켑스트랄 평균 차감법(Cpestral Mean Subtraction)*에 의해 효과적으로 채널 잡음을 최소화할 수 있다.[18-21]

각각의 방법은 선형적인 특성을 갖고 천천히 또는 거의 변하지 않는 안정적인 잡음을 효과적으로 제거할 수 있다. 특히 CMS는 전화선 채널을 거친 음성에 대해 음성 인식 및 화자 인식에서 매우 효과적으로 사용되고 있다. 표 2는 표 1의 실험과 동일한 조건에서 CMS를 적용하여 얻은 결과이다.

〔표 2〕 모의 채널 불일치에서 CMS에 의한 인식 실험

CMV CMV	CPV CPV	CMV CPV	CPV CMV
57.9%	54.7%	51.1%	47.4%

하지만 주파수 차감법은 특징 추출의 전처리 알고

리즘이므로 특징 추출 또는 분류기 알고리즘과 달리 인식 성능에 직접 영향을 주지 못하고 부가적인 연산량에 대한 부담을 가지며, 켑스트랄 평균 차감법은 장구간 평균을 얻을 수 없는 실시간 처리에서 구현되기 어렵다. 이러한 단점을 극복하기 위해 고주파 통과 필터링에 의해 저주파 잡음 성분을 걸러냄으로써 CMS의 효과를 얻을 수 있는 *RASTA (RelAtive SpecTrAl)* 방법이 제안되기도 했다.[18,21]

특징 기반 접근의 다른 방향으로 잡음의 영향에 강인한 음성 특징 추출 기법들이 제안되기도 했다. *ACWC(Adaptive Component Weighting Cepstrum)*는 상대적으로 채널 잡음에 강인한 LP 기반의 켑스트럼의 포먼트 영역을 강조하고 그외의 영역은 감쇄시킴으로써 전화선 환경의 문장 독립 화자 인식 실험에서 향상된 결과를 보여주었다.[22]

*PLP(Perceptually Linear Prediction)*는 스펙트럼 정보에 청각 특성을 고려한 분석을 통해 잡음 환경에서 강인한 특성을 가진 것으로 알려졌다. 그러나 PLP는 음성 인식을 위해 고안된 음성 특징이므로 상대적으로 화자 인식에는 유용한 정보들이 정규화되는 단점을 가지며 특히 전화선 채널 환경에서는 기존의 LPCC, MFCC 등에 비해 두드러진 성능향상을 보이지 못하는 것으로 나타났다.[23,24]

또한 널리 사용되는 켑스트럼의 경우 상대적으로 잡음 성분에 민감한 저주파 또는 고주파 대역을 감쇄시키는 효과를 가진 켑스트랄 가중 기법들이 제안되기도 했다. 특히 band-pass liftering은 잡음에 민감한 저차와 고차 영역을 상대적으로 감쇄시킴으로써 음성 인식에서 좋은 성능을 나타낸 것으로 알려졌으며 화자 인식에서도 일부 환경에서 성능 향상에 도움이 되는 것으로 나타났다.[25]

3. 모델 기반 접근 (Model-based Approach)

모델 기반 접근은 음성 신호 또는 음성 특징에 대한 처리가 아닌 분류기에서 화자의 성문 모델에 대한 처리 방법이다. 특징 기반 접근과는 반대로 왜곡된 음성 신호에서 잡음 영향을 제거하는 것이 아니라 기존의 훈련된 모델에 잡음 영향을 반영하여 불일치를 최소화하는 방법이라고 할 수 있다.

대표적인 방법으로서 *PMC(Parallel Model Combination)*는 훈련 과정에서 잡음 모델과 음성 모델을 분리하여 훈련시키고 인식 과정에서 다시 두 모델을 합성하여 인식하는 방법이다. 따라서 새로운 환경 변이의 잡음 모델을 갱신할 경우 불일치에 따른 잡음에 적응하는 효과를 가지게 된다.[26]

한편 Rose에 의해 제안된 방법은 화자 인식 어플리케이션을 위한 PMC 방법이라고 할 수 있다. 이 방법에서는 훈련과정에서 배경 잡음에 대한 사전 정보를 이용하여 깨끗한 환경에서의 화자 음성 모델과 잡음 모델을 융합시킨 방법을 제안하였다.[27]

Ⅳ. 보상(Compensation) 기술

지금까지의 살펴본 기술들은 신호 또는 특징 영역에서 잡음을 제거하거나 모델 영역에서 오히려 잡음 영향을 추가함으로써 불일치를 최소화하는 방법으로써 공통적으로 환경 변이에 의한 불일치보다는 그 원인인 잡음 자체에 대한 처리에 중점을 둔 것이라 할 수 있다. 또한 이러한 처리들은 동일한 신호, 특징 그리고 모델에 대한 처리로서 **동종(同種, homogeneous) 변이**에 대한 처리라고 할 수 있다.

그러나 이러한 기술들은 대부분 복잡하고 다양한 원인에 의해서 발생하는 잡음의 모델을 단순한 선형 시불변 모델로 가정함으로 처리에 한계가 있고 따라서 여전히 잡음에 의한 불일치 현상은 남게 된다. 결국 기존의 잡음 처리 기술과 함께 불일치 자체를 최소화하기 **이종(異種, heterogeneous) 변이**에 대한 처리인 보상 기술이 최근 활발히 연구되고 있다. 일반적으로 이러한 보상 기술은 확률적 정합(Stochastic Matching)기법으로 일반화될 수 있으며 확률적인 이론에 바탕을 두고 음성 인식을 위한 화자 적응 기법을 응용한 형태로 적용되는 추세이다.[17]

1. 특징 보상(Feature Compensation)

특징 보상 방법은 인식 환경에서의 특징을 보상함으로써 훈련 환경에서 훈련된 모델의 환경에 일치하도록 한다. 대표적인 특징 보상 기술은 Carnegie Mellon 대학에서 꾸준히 연구되었으며 미리 수집된 stereo 데이터베이스를 통해 훈련된 특징 벡터에 의해 훈련과 인식 환경의 불일치를 보상해주는 방법으로서 훈련 기반의 보상 방법이라고 할 수 있다.

왜곡된 음성 특징의 보상 벡터를 얻어내는 기준에 따라 SNR에 따라 보상 벡터를 얻어내는 *SDCN (SNR-Dependent Cepstral Normalization)*, 미리 훈련된 VQ table로부터 보상벡터를 얻어내는

CDCN(Codeword-Dependent) 그리고 인식된 음소열의 결과에 따라 보상벡터를 얻어내는 *PDCN (Phone Dependent)* 등의 방법들이 일련의 연구 결과로서 제안되었다.[18]

그러나 환경 불일치에 대한 stereo DB를 얻을 수 없거나 훈련 환경이 일정하지 않은 화자 인식의 경우 소량의 화자 음성만으로 훈련되어야 하므로 적용되기 힘든 단점을 가진다.

한편 왜곡되지 않은 음성으로 생성된 VQ를 이용하여 전화선 채널 잡음에 대한 음성 특징을 보상하는 방법인 SBR(Signal Bias Removal)이 제안되기도 하였다. 이 방법은 기존에 훈련된 환경과 인식 환경의 차이에 의한 특징 영역의 불일치와 인식 어휘에 차이에 의한 모델 영역의 불일치를 확률적인 방법으로 동시에 최소화할 수 있는 CMS의 EM (Expectation Maximization) 버전이라고 할 수 있다.[28]

2. 모델 보상(Model Compensation)

모델 보상은 이미 훈련된 모델을 일련의 변환 과정에 의해 인식 환경의 특징으로 훈련된 모델에 가깝도록 보상해주는 방법이다. 이 방법은 특징 보상 방법에서 환경 불일치를 모델링하기 위한 데이터 수집의 어려움과 보상된 특징의 분포가 훈련된 모델의 특징 분포와는 다르게 왜곡될 수 있다는 단점을 극복할 수 있다.

특히 평균과 분산에 의해 정의되는 정규 분포로 모델링되는 HMM과 GMM의 경우 특징 보상으로 얻어지는 효과를 직접 모델의 각 상태 또는 mixture의 정규 분포를 수정함으로써 더욱 효과적이고 직접적인 효과를 얻을 수 있다. 예를 들어 CMS, CDCN 그리고 SBR의 경우 모두 특징 보상을 위해 음성 특징과 동일한 차원의 편향 벡터 b_t를 사용하므로 왜곡된 음성 특징 x_t의 선형적인 특징만을 보상하여 음성 특징 y_t를 다음과 같은 식에 의해 얻을 수 있다.

$$y_t = x_t - b_t \tag{3}$$

반면 모델 보상 기법을 응용할 경우 평균(μ_x)과 분산(σ_x)을 수정하여 다음과 같은 affine transform 형태의 보상된 음성 특징을 얻는 것과 동일한 효과를 얻을 수 있다.

$$\mu_y = \mu_x + \mu_b \tag{4}$$

$$\sigma_y = \sigma_x + \sigma_b \text{ 또는 } \sigma_y = \alpha_b \cdot \sigma_x \tag{5}$$

$$y_t = A \cdot x_t + b_t \tag{6}$$

이러한 모델 보상 기법은 음성 인식의 화자 적응을 위해 널리 사용되는 *MAP(Maiximum A Posterior), MLLR(Maximum Likelihood Linear Regression)* 등의 알고리즘을 이용하여 구현될 수 있다.

모델 보상 기술이 화자 인식에 적용된 대표적인 연구는 Reynolds에 의해 제안되었다. 제안된 시스템은 2048개의 mixture로 구성된 GMM을 다양한 화자 음성 데이터에 의해 생성한 후 소량의 등록 화자의 음성을 적응시켜 화자 모델을 생성하였다. 적응 기법은 MAP을 사용하였으며 화자 확인 실험에서 화자종속 모델에 비해서 적응모델이 보다 높은 성능을 보임으로써 적응 기법이 화자 인식을 위해 효과적으로 적용될 수 있음을 보여주었으며, 한편 전화선 환경의 실험에서도 환경 변이에 대한 불일치도 적응 모델에 의해 최소화시킴으로써 성능을 향상시키는 결과를 제시하였다.[15] 또한 마이크로폰과 같은 채널 잡음의 비선형 왜곡으로 인한 화자 인식의 성능 저하를 해결하기 위한 모델 보상 기법에 대한 연구를 발표하기도 했다.[29]

지금까지 살펴본 환경 변이에 대한 정규화와 보상 기술은 융합되어 보다 효과적으로 불일치를 완화시킬 수 있으며, 특징 보상과 모델 보상 방법을 융합함으로도 좀 더 나은 성능을 얻을 수 있다. 적응 기법을 적용한 GMM방법과 SBR과 결합하여 얻은 문장 독립 화자 인식 결과를 표 3에 나타내었다.

〔표 3〕 보상 기법에 의한 인식 실험(모의 채널)

	NONE	MAP	MLLR	MLLR-MAP
Clean	81.7%	87.8%	94.8%	95.7%
None	18.4%	15.8%	15.8%	16.7%
CMS	53.5%	59.7%	53.5%	61.4%
SBR-GMM	-	62.3%	63.2%	65.8%

실험은 TIMIT 데이터베이스의 DR3 테스트 영역의 23명의 남자에 대해서 수행되었으며 훈련과 인식에서 각각 CMV와 CPV채널을 거쳤다. 또한 동일한 방법을 강인한 음성 특징인 ACWC와 융합하

여 실제 채널 환경인 NTIMIT 데이터 베이스에서 수행한 경우에도 표 4에서 보인 것과 같은 성능 향상을 나타내었다.[30]

〔표 4〕 ACWC와 보상 기법을 병행한 인식 실험 (실제 채널, NTIMIT)

	NONE	MAP	MLLR	MLLR-MAP
None	53.0%	36.5%	24.6%	38.3%
CMS	47.0%	45.2%	40.8%	47.8%
SBR-GMM	-	60.9%	55.7%	59.1%

V. 결 론

일반적인 화자 인식 시스템과 환경 변이에 강인한 화자 인식 시스템을 위한 기술들을 살펴보았다. 화자 인식 시스템의 성능을 저하시키는 주요한 원인은 잡음에 의한 신호, 특징 그리고 모델 영역의 왜곡을 일으키는 환경 변이이며, 동시에 이에 따른 훈련과 인식 환경의 불일치이다. 특히 화자 인식 기술의 경우 소량의 화자 음성 데이터만으로 화자의 음성 모델을 생성해야하므로 사소한 잡음에 의한 왜곡도 큰 훈련과 인식 환경의 불일치를 초래하게 된다.

최근의 연구는 환경 변이의 원인인 잡음 자체를 정규화 하는 연구에서 특징 및 모델 영역에서의 불일치를 최소화하기 위한 보상에 대한 연구로 진행되고 있다. 특히 확률적 정합 기법으로 일반화되는 보상 기술은 앞으로도 다양한 응용 분야에서 불일치에 대한 해결 방법으로 사용될 것으로 전망된다. 하지만 비선형적인 왜곡에 대한 정확한 모델링이 부족한 점과 보상 기법의 연산량에 대한 부담 등은 앞으로 연구되어야 할 과제로 남아있다고 사료된다.

참 고 문 헌

〔1〕 Bishnu S. Atal, "Automatic Recognition of Speakers from Their Voices," Proceedings of the IEEE, Vol.64, No.4, pp.460-474, April, 1976

〔2〕 D. O'Shaughnessy, "Speaker Recognition", IEEE ASSP Magazine, pp.4-17, October 1986

〔3〕 Herbert Gish and Michael Schmidt, "Text-Independent Speaker Identification", IEEE Signal Processing Magazine, pp.18-31, October 1994

〔4〕 AARON E. Rosenberg, "Automatic Speaker Verification: A Review," Proceedings of the IEEE, vol. 64, No. 4, pp. 475-487, April 1976

〔5〕 Jayant M. Naik, "Speaker Verification : A Tutorial," IEEE Communications Magazine, pp.42-48, Jan., 1990

〔6〕 K.-U. Mazel and H.-P. Frei, "CAVE - Speaker Verification in Banking and Telecommunications," Computer Science Research at Ubilab, pp.153-162, Nov. 1996

〔7〕 L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals, Prentice-Hall*, Englewood cliffs, N.J., 1978

〔8〕 L. R. Rabiner, B. H. Juang, *Fundamentals Of Speech Recognition*, Prentice-Hall, 1994

〔9〕 Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, 1996

〔10〕 Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. on Speech and Audio Processing, Vol.3, No.1, pp.72-83, Jan., 1995

〔11〕 S. Euler, R. Langlitz, J. Zinke, "Comparison of Whole Word and Subword Modeling Techniques for Speaker Verification With Limited Training Data," Proc. of ICASSP, pp.1079-1082, 1997

〔12〕 Tomoko Matsui, Sadaoki Furui, "Concatenated Phoneme Models for Text-variable Speaker Recognition," Proc. of ICASSP, pp. II-391-394, 1993

〔13〕 Yong Gu and Trevor Thomas, "A Hybrid Score Measurement For HMM-

Based Speaker Verification," Proc. of ICASSP, pp.317-320, 1999

[14] J. B. Pierrot, J. Lindberg, J. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, F. Bimbot, "A Comparison of A Priori Threshold Setting Procedures For Speaker Verification in the CAVE Project," Proc. of ICASSP, pp.11-14, 1998

[15] Douglas Reynolds, Thomas Quatieri, and Robert Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol. 10, pp.19-41, 2000

[16] R. J. Mammone, X. Zhang, R.P. Ramachandran, "Robust Speaker Recognition", IEEE Signal Processing Magazine, pp.58-71, Sep., 1996

[17] Chin-Hui Lee, "On Stochastic Feature and Model Compensation Approaches to Robust Speech Recognition", Speech Communication, Vol. 25, pp.29-47, 1998

[18] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Dordrecht

[19] Yu-Jin Kim, Hea-Kyoung Jung, Jae-Ho Chung, "Formant-Broadened CMS Using Peak-Picking in Log Spectrum", Proc. of Eurospeech-2001, Vol. 4, 2829-2832, 2001.6

[20] Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," IEEE Trans. on Acoustics, Speech, And Signal Processing, vol. 29, pp. 113-120, April 1979.

[21] Hynek Hermansky, Nelson Morgan, "RASTA Processing of Speech," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, Oct. 1994

[22] K.T. Assaleh, R.J. Mammone, "New LP-Derived Features for Speaker-Identification", IEEE Trans. on Speech and Audio Processing Vol. 2, No. 4, Oct. 1994

[23] Hynek Hermansky, "Perceptual linear predictive(PLP) analysis of speech," Journal of Acoustical Society of America, Vol. 87, No. 4, Apr. 1990

[24] 조태현, 김유진, 이재영, 정재호, "전화선 채널이 화자확인 시스템의 성능에 미치는 영향", 한국음향학회지, 제 18권 5호, pp. 12-20, 1999.7.

[25] Sadaoki Furi, "Cepstral Analysis Technique for Automatic Speaker Verification", Proc. of ICASSP, pp.254-271, 1981

[26] M.J.F. Gales, S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination", Computer Speech and Language 9, pp. 289-307. 1995

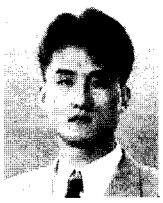
[27] R.C. Rose, E.M. Hofstetter, D.A. Reynolds, "Integrated models of speech and background with application to speaker identifiation in noise", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 245-257, 1994.

[28] Mazin G. Rahim, Biing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", IEEE. trans. on Speech and Audio Processing, Vol. 4, No. 1, Jan. 1996

[29] T. F. Quatieri, D. A. Reynolds, "Estimation of Handset Nonlinearity with Application to Speaker Recognition", IEEE Trans. on Speech and Audio Processing, Vol. 8, No. 5, pp. 567-584, 2000

[30] Yu-Jin Kim, Jae-Ho Chung, "Signal Bias Removal Based GMM for Robust Speaker Recognition", Proc. of ICASSP, Student-Forum, May. 2002

〈著 者 紹 介〉



김 유 진 (Yu-Jin Kim)
정회원

1995년 2월 : 인하대학교 전자공학과 졸업
1997년 2월 : 인하대학교 전자공학과 석사
1995년 6월~1996년 12월 : 한국전자통신연구소 음성언어처리연구실 위촉 연구원
1997년 2월~1998년 5월 : LG반도체 System Device 연구소 DSP 그룹 연구원
1998년 9월~현재 : 인하대학교 전자공학과 박사과정(수료)
관심분야 : 패턴인식, 음성인식, 화자인식

정 재 호 (Jae-Ho Chung)
정회원

1982년 : University of Maryland (BSEE)
1984년 : University of Maryland (MSEE)
1990년 : Georgia Institute of Technology (Ph. D.)
1984년~1985년 : 미묵 국방성 산하 해군 연구소, 신호처리실 연구원
1991년~1992년 : AT&T Bell Labs, 음성신호처리 연구실 연구원 (MTS)
1992년~현재 : 인하대학교 공과대학 전자공학과, (현)정교수
관심분야 : 음성코딩, 음성인식, 화자인식, 화자적응