음성인식을 위한 화자적응 기술 동향

김 동 국 (전남대학교)

차 례

1. 서론

음성인식(speech recognition) 기술이란 사람이 말하는 음성을 기계나 컴퓨터가 이를 분석하고, 인식하여 단어나 문장형태로 변환하여 기계와 인간이 상호작용을 할 수 있도록 관련 알고리즘을 개발 및 구현하는 기술이다. 최근 음성인식 기술이 대두되는 가장 큰 이유는 인간과 기계간의 통신을 원활하게 하는 편리한 휴먼인터페이스 기능이라 할 수 있다. 즉 기존에는 복잡한 키 입력을 통해 정보를 얻거나 자동차 운전 중에 전화걸기와 같은 손과 눈을 사용하여 조작하는 경우 사고 위험성이 발생하는 단점을 가진다. 이에 반해 음성인식 기술을 사용하면 한번에 명령어를 말하므로 원하는 정보를 얻거나, 손과 눈을 사용하지 않고 음성을 통해 자연스럽게 전화 걸기 등을 할 수 있는 장점 등이 있다. 그러므로 음성인식 기술은 주식조회, 자동교환서비스, 보이스포탈 그리고 텔레매틱스 등의 다양한 응용 분야에서 사용되고 있다.

음성인식 성능은 기술적인 발달에 따라 계속 향상되고 있지만 기본적으로 인식 성능을 저하시키는 몇 가지 요소들이 존재한다. 이러한 요인으로는 화자간/화자내 발음 변이성, 마이크나 채널 특성, 주변 잡음 그리고 다른 엑센트/억양과 다른 발음 스타일 등이 있다[12][28][36]. 이는 학습과 인식 환경 사이에 존재하는 차이 또는 불일치(mismatch)에 기인하는 것으로 학습과 인식 과정 중에 언제나 존재하여 성능 저하에 큰 영향을 미친다. 특히 학습 데이터에 포함되는 않는 화자는 인식기의 성능을 크게 떨어뜨린다. 현재 사용되고 있는 화자독립 시스템은 매우 높은 인식 성능을 나타내지만 화자종속 시스템은 독립 시스템에 비해 2-3배 정도의 낮은 평균 인식 에러율을 나타내고 있다 [12][36].

현재 대부분의 인식 시스템은 음성신호의 통계적 특성을 나타내기 위해 hidden Markov

model(HMM)을 사용한다 [31]. HMM은 통계적 신호 모델에 근거하여 음성 신호를 표현하는 모델이다. 음성 인식기의 구조는 크게 특징 추출과 decoder로 나눌 수 있고, 음향모델, 발음사전, 언어모델은 decoding을 위해 필요한 정보를 제공하여 준다 [31]. 특징추출은 음성인식에 적당한 형태로 입력음성을 표현하는 과정이며, 오늘날mel-frequency cepstral coefficient(MFCC)가 많이 사용되고 있다 [8]. Decoder는 입력음성에 대해 최적의 인식결과를 찾는 과정으로 Viterbi 알고리즘이 사용된다 [35]. 음향모델은 학습 데이터로부터 학습된 HMM 파라메터로 정보를 저장하고 입력음성에 대해 각 HMM에서 확률 값을 계산할 때 이용한다. 발음사전은 인식하고자 할 단어들을 기본적인 subword 단위로 표현하는 것이다. 언어모델은 decoding 할 때 가능한 단어 열에 대한 제약을 주기 위한 확률적 모델이다.

2. 화자적응

적응(adaptation)은 주어진 적응 데이터를 사용하여 학습과 인식 환경 사이에 존재하는 차이를 줄여 인식 성능을 향상시키는 과정이다. 특히 화자적응(speaker adaptation)은 가능한 특정 화자로부터 적은 양의 데이터를 가지고 향상된 화자종속 인식 성능에 접근하도록 하는 것을 목표로 한다. 이러한 화자 적응 기술은 사용하는 방식에 따라 다음과 같은 3가지 모드로 구분된다 [12][36].

- 교사(supervised) 적응과 비교사(unsupervised) 적응
- 일괄(batch) 적응과 순차(online) 적응
- 자기적응(self-adaptation)

교사적응은 적응 데이터에 대한 전사(transcription)가 주어지는 경우이며, 주어지지 않은 경우를 비교사 적응이라 한다. 일괄적응은 모든 적응 데이터가 적응 전에 일괄적으로 주어진 경우이며, 순차적응은 시간에 따라 순차적으로 적응 데이터가 주어지는 경우이다. 자기 적응은 현재 인식하고자 하는 데이터를 이용하여 먼저 적응하고 적응된 모델을 이용해 다시 인식하는 경우를 말한다. 비교사 순차 적응은 부정확한 전사와 적은 시간의 음성 데이터가 주어지므로 교사 일괄 적응보다 인식 성능 향상이 보다 어렵다.

적응을 위해서는 가능한 많은 정보를 효율적으로 이용하여 성능을 높이는 것이 필요하다. 특히 빠른 순차 적응을 위해서는 적응 데이터뿐 아니라 학습에 사용된 화자로부터 다양한 지식을 추출하여 사용하는 것이 필요하다. 일반적으로 실용적인 적응을 위해 적응데이터, 모델 사이의 상관관계 정보, 파라메터에 대한 선분포 그리고 선지식 정보등과 같은 중요한 정보원을 사용한다 [18]. 그리고 좋은 화자적응 기술이 되기 위해서는 적은 양의 데이터로 순차적으로 큰 성능 향상이 이루어져야 하며, 많은 양의 데이터가 주어지는 경우는 화자종속 성능으로 수렴하는 것이 필요하다. 그러므로 실용적인 화자적응 기술이 가져야 하는 바람직한 기준은 빠른 적응성, 순차 적응성 그리고 수렴 적응성을 갖는다 [18].

최근의 적응기법은 HMM 파라메터를 포함하는 음향모델을 적응 데이터에 따라 적응된 음향모델을 얻기 위한 과정을 수행한다. 전사가 주어진 교사학습에 경우 주어진 전사정보를 이용하며, 비교사 학습은 decoder을 통해 얻어진 인식

결과를 전사정보로 이용한다. 적응을 위해 갱신되는 음향모델의 HMM는 각각의 상태(state), 상태 사이의 천이(transition), 그리고 각 상태에서 출력 값을 나타내는 출력분포(output probability)로 구성되어 있다 [31]. 인식과 적응 성능에 가장 큰 영향을 미치는 요소는 가우시안(Gaussian) 밀도 함수에 의해 표현되는 출력분포 함수이다. 일반적으로 출력분포 함수는 가우시안 mixture 연속밀도 함수가 사용되는데, 이를 continuous density HMM(CDHMM)이라 한다 [17][31]. 길이 T 개의 동등하고 독립적으로 분포된 관측 벡터 열 $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_T)$ 가 주어진 경우 CDHMM의 관측 확률분포는 다음과 같다.

$$f_j(\mathbf{x}\mid\lambda)=\sum_{k=1}^{K}\omega_{jk}N(\mathbf{x}\mid\mu_{jk},\Sigma_{jk}) \quad (1)$$

여기서 ω_{jk} 는 상태 j, 혼합성분 k 에 대한 가중치이며 $\sum_{k=1}^{K}\omega_{jk}=1$ 을 만족하고, $N(\mathbf{x}\mid\mu_{jk},\Sigma_{jk})$ 는 d 차원의 평균벡터 μ_{jk} 와 $d\times d$ 차원의 공 분산행렬 Σ_{jk} 을 갖는 다 변수 가우시안 분포이다. 일반적으로 위와 같이 HMM 파라메터를 추정하기 위해 maximum likelihood (ML) [17] 기준 하에서 반복적으로 파라메터를 추정하여 갱신하는 Expectation Maximization (EM) 알고리즘을 사용하여 파라메터를 추정한다[2][9].

최근 HMM의 파라메터를 적응 데이터에 따라 갱신하는 모델기반 적응 알고리즘이 많이 제안되었다. 모델기반 화자적응은 CDHMM의 관측 확률 분포의 파라메터 λ 을 적응 데이터에 따라 변화시켜 적응하는 것이다. 일반적으로 모델기반 적응기법은 다음과 같이 4가지로 형태로 분류된다 [27][36].

1. Maximum a posteriori (MAP) 적응기법 [11]
2. 변환기반 적응기법 [30]
3. 화자공간(speaker space) 적응기법 [10][18][26]
4. 순차적응 기법 [13]-[15],[21]-[22],[24]-[25]

MAP, 변환기반, 화자공간 기법들은 주로 일괄 적응에서 사용되며, 순차적응 기법은 순차모델에 근거하여 HMM 파라메터 λ 나 변환 파라메터를 순차적으로 적응을 수행한다.CDHMM 파라메터 벡터 가운데 일반적으로 데이터가 주어진 경우에 평균 벡터만이 적응된다.

3. MAP 적응기법

MAP 적응기법은 학습 데이터에 포함되어 있는 선 지식 정보를 선 밀도 함수에 포함시켜 이를 적응 데이터와 MAP 추정기법으로 결합하여 적응하는 방식이다 [11]. MAP에서는 CDHMM 파라메터 λ 가 hyperparameter ϕ 을 갖는 선 확률 밀도함수 $g(\lambda\mid\phi)$ 분포를 갖는 랜덤(random) 변수라 가정한다. MAP 기법은 유사도 $f(\mathbf{X}\mid\lambda)$ 를 갖는 관측 열로부터 추정된다면, 다음과 같이 λ 의 posterior mode로 정의 된다 [11].

$$\hat{\lambda}_{MAP}=\arg\max_{\lambda} f(\mathbf{X}\mid\lambda)g(\lambda\mid\phi) \quad (2)$$

학습화자의 데이터로부터 추정된 선 밀도 함수 $g(\lambda\mid\phi)$ 는 관심 있는 파라메터에 대한 통계적 특성을 포함하여 관측 열이 주어지기 전에 파라

메터가 어떤 제약된 값을 갖도록 한다. 일반적으로 HMM과 같이 상태와 혼합 성분이 내재된 은닉 과정을 포함하는 경우에 MAP 추정은 매우 어렵다. 그러나 HMM 파라메터의 선 밀도 함수가 완전데이터(complete-data) 밀도의 conjugate family에 속한다면 EM 알고리즘에 의해 MAP 추정을 쉽게 할 수 있다 [11]. 식 (1)에 있는 CDHMM의 평균과 precision 파라메터에 대한 선 밀도 함수가 normal-Wishart 밀도 형태로 주어지는 경우, 평균 벡터 파라메터에 대한 MAP 추정은 선 평균과 ML 추정된 값의 보간된 형태로 주어진다 [11].

MAP 적응된 평균값은 관측된 파라메터에 대해서만 적응되며, 관측되지않은 파라메터는 추정된 평균은 선 평균값이 된다. MAP 추정은 ML 추정에 비해 적은 적응 데이터에 대해 더 강인하게 파라메터를 추정하며, 적응 데이터의 양이 증가함에 따라 MAP은 ML 추정치로 수렴하는 장점을 갖는다. 그러나 관측된 파라메터에 대해서만 적응되기 때문에 많은 수의 파라메터를 갖는 대용량 인식기의 경우 적응 속도가 매우 느리다는 단점을 갖는다.

Extended MAP(EMAP)는 파라메터 사이의 상관관계를 이용하여 MAP 적응 방식의 단점을 보안하는 적응 기법이다 [37]. EMAP는 모든 가우시안 평균값이 상관관계를 갖는다고 가정하고, 관측되지않은 파라메터를 적응하기 위해 이 상관관계 정보를 이용한다. Ahadi와 Woodland는 EMAP과 비슷한 regression-based model prediction (RMP) 기법을 제안했다 [1]. 이 방법은 파라메터 사이의 상관관계를 추정하기 위해 linear regression을 사용하고, 잘 적응된 파라메터에 근거하여 관측되지 않은 파라메터를 적응한다. Shinoda와 Lee는 structured MAP (SMAP) 기법을 제안하였는데, 이는 가우시안들이 계층적 tree 구조로 조직화되어 MAP 기법이 각 계층에서 파라메터를 적응하기 위해 사용된다 [33].

4. 변환기반 적응기법

변환기반 적응기법의 대표적인 방식은 Maximum Likelihood Linear Regression (MLLR) 방식이 있다 [30]. MLLR은 CDHMM의 평균 벡터를 적응하기 위해 선형 변환 행렬(linear transformation matrix)을 사용하는 적응 기법이다. MLLR에서 평균벡터의 적응은 다음과 같은 선형 변환에 의해 이루어진다.

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b} = \mathbf{W}\xi \tag{3}$$

여기서 $\mathbf{A}$는 $d \times d$ 행렬이고 $\mathbf{b}$는 d 차원 벡터이다. 또한 $\mathbf{W} = [\mathbf{A}, b]$는 $d \times (d+1)$ regression 행렬이고 ξ는 확장된 평균벡터 $\xi = [1 \ \mu_1 \cdots \mu_d]$이다. MLLR에서는 regression 행렬 $\mathbf{W}$을 추정하기 위해 적응 데이터의 유사도가 최대가 되도록 다음과 같이 추정된다 [30].

$$\hat{\mathbf{W}}_{MLLR} = \arg\max_{\mathbf{w}} f(\mathbf{X} | \mathbf{W}, \lambda) \tag{4}$$

$\mathbf{W}$의 대한 해를 구하기 위해EM 알고리즘을 사용하여 반복적으로 갱신하여 구한다. 데이터 양의 따라 변환 행렬의 수를 조절하기 위해 regression class tree 기법이 사용된다 [27]. 변환 행렬은 각각의 가우시안에 대해 구하거나 여러 가우시안을 묶여서 구할 수 있다. 모든 가우시

안에 대한 전체 변환 행렬을 사용하면 적은 양의 적응 데이터에 대해 적응을 수행할 수 있다. 반면에 각각의 가우시안에 대한 변환 행렬을 사용하는 경우는 많은 양의 적응 데이터에 대해 강인한 파라메터를 추정할 수 있다. MLLR 적응은 MAP 적응기법에 비해 관측되지 않은 모델 파라메터도 변환 행렬에 의해 모두 적응할 수 있는 장점이 있다. 그러나 데이터의 양의 증가하여도 화자종속 모델로 수렴하지 않는 단점을 갖고 있다.

MLLR 파라메터를 강인하게 추정하기 위해서는 많은 양의 적응 데이터가 필요하다. Maximum a posteriori linear regression (MAPLR) 기법은 변환 행렬에 대한 선 확률분포 $p(\mathbf{W})$을 포함하여 MAP 기준에 의해 더 강인하게 추정하므로써 MLLR 성능을 향상하기 위해 제안되었다 [7].

5. 화자공간 적응기법

화자공간 기법은 학습 화자에 대한 선 지식 (a priori knowledge) 정보를 활용하는 방법이다. 학습화자에 대한 선 지식은 화자종속 모델이나 화자종속 변환 파라메터로부터 추출된 대표적인 화자 벡터에 의해 정의된다. 새로운 화자에 대한 적응 데이터가 주어지는 경우 적응될 모델은 화자 공간내에서 화자벡터의 선형적인 결합에 의해 표현되어진다. 그러므로 추정되어야 할 파라메터 수가 매우 적으므로 빠른 화자적응에 매우 적합하다. 화자공간 적응기법으로 eigenvoice [26], 화자공간모델 [18]-[20],[23] 그리고 변환공간모델 [22]이 있다.

5.1 Eigenvoice

Eigenvoice은 화자종속 모델로부터 principal component analysis(PCA)[16]을 통해 화자공간의 화자벡터를 구하고, ML 기준을 통해 화자벡터의 선형 결합 계수를 추정하여 적응하는 기법이다.

$\{\boldsymbol{\mu}_1,\cdots\boldsymbol{\mu}_R\}$을 학습 데이터로부터 R개의 잘 학습된 화자종속 모델이라고 하자. 여기서 $\boldsymbol{\mu}=[\mu_{00}^T,\cdots\mu_{NK}^T]$는 특정 화자종속 모델로부터 모든 가우시안 평균 벡터를 모아서 만든 D차원의 supervector이다. Eigenvoice는 모든 학습 화자 모델로부터 P차원의 선형공간 (또는 eigenspace) $\mathbf{U}_P$ 을 PCA을 통해 구하게 된다. 그러면 새로 적응된 모델 $\hat{\boldsymbol{\mu}}$은 다음과 P개의 주요 화자 벡터의 선형 결합에 의해 구하여 진다.

$$\hat{\boldsymbol{\mu}} = \mathbf{U}_P\mathbf{y} + \overline{\boldsymbol{\mu}} \tag{5}$$

여기서 $\mathbf{y}$는 추정될 가중치 벡터로 ML eigen-decomposition (MLED) 기법에 의해 구해진다 [26].

Eigenvoice 적응기법은 적은 양의 적응 데이터가 주어지는 경우에도 추정해야 할 파라메터 수가 적어 강인한 파라메터 추정이 가능하여 매우 높은 성능의 빠른 적응을 수행하는 장점을 갖는다. 반면에 적응 데이터가 많아져도 화자공간 내에서만 적응 모델이 추정되므로 화자종속 모델로 수렴하지 못하며, 용량의 경우 저장해야 할 화자공간 파라메터의 수가 많다는 단점을 갖는다.

Gales는 eigenvoice와 비슷한 cluster adaptive training(CAT) 기법을 제안하였는데, eigenvoice와 주요 차이점은 CAT 기법은 화자공간의 화자벡터를 학습화자 모델의 clustering

함으로 구한다는 것이다 [10].

5.2 화자공간모델

화자공간모델 (speaker space model : SSM) [18]-[20],[23] 기법은 eigenvoice에 단점을 극복하기 위해 제안된 방법이다. 화자공간을구할 때 eigenvoice와 다르게 factor analysis (FA)[32]나 probabilistic principal component analysis(PPCA)[34]와 같은 은닉변수모델 (latent variable model)을 사용한다. SSM 기법에서는 각 화자종속 모델 $\boldsymbol{\mu}$가 다음과 같은 선형 모델에 의해 발생한다고 가정한다 [34].

$$\boldsymbol{\mu} = \mathbf{U}\mathbf{y} + \bar{\boldsymbol{\mu}} + \varepsilon \tag{6}$$

여기서 $\bar{\boldsymbol{\mu}}$는 supervector에 대한 평균이고, $\mathbf{y}$는 P 차원의 은닉변수, $\mathbf{U}$는 $D \times P$ 차원의 화자공간을 나타내는 행렬이며, ε는 $\mathbf{y}$와 독립적인 가우시안 랜덤 잡음이다. FA 모델에서는 잡음은 대각 공분산 행렬 ?를 가진 가우시안, $p(\varepsilon) \sim N(0, ?)$으로 정의 된다. PPCA는 잡음 공분산 행렬이 isotropic, 즉 $? = \sigma^2\mathbf{I}$로 정의 된다.

위 식 (6)는 CDHMM 평균 벡터와 관련된 학습 화자에 대한 선 지식을 표시하는 SSM이다. 이 SSM는 학습화자와 관련된 평균 supervector을 분석함으로써 화자 변이에 대한 중요한 정보를 추출한다. [18]의 연구 결과로부터 SSM는 화자적응을 위해 사용될 수 있는 다른 음성 단위사이에 존재하는 상관관계 정보, 선 확률분포 그리고 화자공간과 같은 선 지식을 동시에 포함하고 있음을 알 수 있다.

SSM에 의한 적응 기법은 선 분포를 갖고 있기 때문에 MAP 추정기법에 의해 다음과 같은 형태로 나타난다 [19][20][23].

$$\hat{\boldsymbol{\mu}}_{SSM} = \alpha\,\boldsymbol{\mu}_{ML} + (1-\alpha)\,\boldsymbol{\mu}_{SS} \tag{7}$$

SSM의 기법은 간단히 ML 추정된 평균값 $\boldsymbol{\mu}_{ML}$과 화자공간 안에서 추정된 선 평균 값 $\boldsymbol{\mu}_{SS}$과의 결합된 형태로 나타난다. 그러므로 eigenvoice와는 다르게 적응 데이터가 적은 경우에는 화자공간 내에서 추정된 선 평균에 의존하고 데이터가 많아짐에 따라 ML 평균값에 수렴한다. SSM 기법의 장점은 빠른 화자 적응 특성과 화자 종속 모델로 수렴하는 특성을 동시에 갖는다. 단점으로는 대용량 인식기와 같이 파라메터수가 매우 크고, 적응 데이터가 매우 작은 경우에 MAP 특성상 적응이 다소 느리다는 것이다.

Eigenvoice 기법은 PCA 방법을 사용하여 가장 큰 eigenvalue에 해당한eigenvector을 사용하여 화자공간을 만든다. 이런 공간은 서로 직교하며 학습 화자 사이의 가장 중요한 성분들을 추출한다. 반면 SSM은 화자공간을 EM 알고리즘을 통해 구한다. FA를 통해 구한 화자공간은 학습화자의 주요 준공간에 해당하지 않고 단지 span한다. 그러나 Tipping과 Bishop은 관측 데이터의 유사도가 최대가 되는 점에서 PPCA에 의해 구한 준공간이 주요 공간을 span함을 증명하였다 [34].

5.3 변환공간모델

변환공간모델 (transformation space model : TSM) [18][22][25]은 MLLR 적응기법에서 얻어진 regression 행렬 파라메터 $\mathbf{W}$을 SSM과 같이 은닉변수 모델에 근거하여 적응하는 기법이

다. $\{\mathbf{W}_r\}, \mathrm{r}=1,\cdots,\mathrm{R}$ 을 MLLR 적응 기법에 의해 잘 학습된 화자종속 regression 행렬이라 하자. $\mathbf{w}_r$ 는 regression 행렬 $\mathbf{W}_r$ 의 열 벡터들을 모아서 만든 $D(=d(d+1))$ 차원의 변환 supervector, $\mathbf{w}_r=[W_{r1},\cdots,W_{rd}]$ 라 하자. TSM은 SSM과 같이 변환supervector들이 FA나 PPCA와 같은 은닉변수 모델에 의해 다음과 같이 정의된다 [18][22][25].

$$\mathbf{w}=\mathbf{V}\mathbf{z}+\overline{\mathbf{w}}+\delta \tag{8}$$

여기서 $\overline{\mathbf{w}}$ 는 변환supervector에 대한 평균이고, $\mathbf{z}$ 는 P 차원의 은닉변수, $\mathbf{V}$는 $D\times P$ 차원의 변환 공간을 나타내는 행렬이며, δ 는 $\mathbf{z}$ 와 독립적인 가우시안 랜덤 잡음 $p(\delta)\sim N(0,?)$ 이다. 위 식 (8)은 regression 행렬 $\mathbf{W}$와 관련된 학습 화자의 선 지식을 특징짓는 TSM을 나타낸다. 이 TSM는 변환 파라메터에 대한 상관관계 정보, 선 분포에 대한 정보 그리고 변환공간에 대한 선 지식을 동시에 나타내고 있다. 기본적으로 선 확률분포를 갖기 때문에 MAP 추정을 통해 TSM에 대한 적응 과정을 수행할 수 있다. TSM 적응 기법은 MAPLR 적응기법과 같은 형태를 나타내며, TSM 적응된 변환 파라메터는 변환공간 내에서 추정된 선 변환 모델을 MAPLR 추정 과정에 포함함으로 얻을 수 있다 [18].

TSM의 장점은 대용량 인식 시스템에서도 변환 파라메터에 의해 빠른 화자 적응을 수행할 수 있다는 것과 eigenvoice 보다 변환공간 파라메터 수가 훨씬 작다. 그러나 많은 적응 데이터에 대한 성능 향상을 위해 각 regression class tree에 따른 변환공간 모델을 요구한다는 단점을 갖는다.

TSM과 비슷하게 eigenspace-based MLLR과 eigenspace-MAPLR기법은 PCA와 PPCA에 근거하여 MLLR regression 파라메터에 대해 학습화자의 선 지식이 포함되도록 제안되었다 [3][4].

6. 순차 적응기법

음성인식 시스템은 시간에 따라 다양하게 변화하는 화자나 환경으로부터 음성 데이터를 받아 인식함으로 이를 순차적으로 처리하는 것이 바람직하다. 화자 적응에서도 화자가 시간에 따라 다른 음성 데이터를 발음함으로 이를 순차적으로 적응에 사용하는 것이 요구된다. 이를 위해 순차적응 기법은 선 진화 (prior evolution)라는 불리는 개념에 기초한 Bayesian 추정기법 하에 발전되어 왔다 [13][29]. 이 기법은 음성인식 시스템이 모든 데이터를 모으지 않고 새로운 화자 또는 환경에 순차적으로 적응하도록 한다.

6.1 순차 적응

먼저 순차 적응을 위한 Bayesian 추정과 quasi-Bayes(QB) 학습 이론에 대해 알아본다. Bayesian 추정 이론에 있어서 HMM 파라메터 λ 의 불확실성을 나타내기 위해 이 값을 랜덤이라고 가정한다. λ에 대한 선 지식은 hyperparameter $\phi^{(0)}$ 을 갖고 알려진 결합 선분포(joint prior distribution) $g(\lambda|\phi^{(0)})$에 포함된다고 생각한다. $\mathbf{X}^n=\{\mathbf{X}_1,\cdots,\mathbf{X}_n\}$을 CDHMM 파라메터를 갱신하기 위해 순차적으로 모아진 독립적인 관찰 데이터 집합이라 하자. 그

러면 λ 의 posterior 확률밀도 함수에 대한 반복 추정 식은 다음과 같이 주어진다.

$$p(\lambda \mid \mathbf{X}^n) = \frac{p(\mathbf{X}_n \mid \lambda) p(\lambda \mid \mathbf{X}^{n-1})}{\int p(\mathbf{X}_n \mid \lambda) p(\lambda \mid \mathbf{X}^{n-1}) d\lambda} \quad (9)$$

위 식으로부터 파라메터 λ 에 대한 반복적인 Bayesian 추정이 가능하다. 그러나 CDHMM 파라메터에 대한 이러한 형태의 추정은 구현하기에 매우 어렵다. 이런 문제를 쉽게 해결하기 위해 QB 학습 알고리즘이 제안되었다 [13]. QB 학습 알고리즘은 각 시간에서 실제 posterior 밀도함수 $p(\lambda \mid \mathbf{X}^n)$ 을 가장 근사한 선 밀도함수 $g(\lambda \mid \phi^{(n)})$ 로 두 밀도함수의 최대치가 같도록 근사화 한다. 여기서 $\phi^{(n)}$ 는 $\mathbf{X}^n$ 가 관측된 이후에 갱신된 hyperparameter을 나타낸다. 초기값 $\phi^{(0)}$ 로부터 출발하여 적응 데이터 $\mathbf{X}_1, \cdots, \mathbf{X}_n$ 을 순차적으로 적용함으로 모델 파라메터 $\lambda^{(0)}, \cdots, \lambda^{(n)}$ 의 순차적 추정과 hyperparameter의 진화 $\phi^{(0)}, \cdots, \phi^{(n)}$ 를 수행한다. 이러한 방식으로 현재의 적응데이터 $\mathbf{X}_n$ 에 의해 이전 prior 밀도함수 $g(\lambda \mid \phi^{(n-1)})$ 가 새로운 밀도함수 $g(\lambda \mid \phi^{(n)})$ 로 진화하게 된다. 즉 현재 시간 $\mathbf{X}_n$ 에 의해 주어진 새로운 정보는 과거 정보와 결합되어 $g(\lambda \mid \phi^{(n)})$ 에 갱신되어 저장된다. 이렇게 진화하는 밀도함수를 사용하여 순차 적응을 수행하게 된다.

Huo와 Lee는 QB 학습 기법에 기반한 선 진화를 통해 CDHMM 파라메터와 그에 대한 hyperparameter을 동시에 순차적으로 갱신하였다 [13]. 또한 모든 CDHMM 평균 벡터가 결합 선 분포를 갖는 경우, 즉 상관관계를 갖는 CDHMM 파라메터에 대한 순차적응 기법으로 확장하였다 [14]. 최근에는 multi-stream 선 진화와 posterior pooling 라 불리는 적응기법을 다양한 양의 적응 데이터에 대해 잘 동작하도록 제안하였다[15]. Chien는 CDHMM 파라메터에 대한 간단한 변환을 취하고 변환 파라메터에 대한 선 밀도를 가정함으로 순차 변환 기반 QB 알고리즘을 제안하였다 [5]. 또한 CDHMM의 순차 선형 regression 적응을 위한 quasi-Bayes linear regression (QBLR) 기법을 제안하였다 [6]. 그리고 본 저자들은 SSM과 TSM에 근거한 순차적응 기법인 SSM evolution과 TSM evolution를 각각 제안하였다 [18][21]-[22], [24]-[25].

6.2 SSM evolution

SSM evolution 기법은 SSM에 근거하여 QB 추정 이론을 적용한 순차 적응 알고리즘이다 [18][21][24]. SSM에서는 파라메터간의 상관관계, 선 지식정보뿐 아니라 선 분포 정보를 포함하고 있으므로 Bayesian 추정 이론을 적용할 수 있다. 식 (6)로부터 $\mathbf{y}$ 가 주어진 경우 $\boldsymbol{\mu}$ 의 조건부 확률 $p(\boldsymbol{\mu} \mid \mathbf{y}) \sim N(\mathbf{U}\mathbf{y} + \bar{\boldsymbol{\mu}}, ?)$ 로 주어진다. 또한 $\boldsymbol{\mu}$ 에 대한 선 확률 분포는 hyperparameter $\phi = \{\bar{\boldsymbol{\mu}}, \mathbf{U}, ?\}$ 을 가지며 $p(\boldsymbol{\mu} \mid \phi) \sim N(\bar{\boldsymbol{\mu}}, ? + \mathbf{U}\mathbf{U}^T)$ 로 주어진다. 시간 n 에서 적응 데이터 $\mathbf{X}^n$ 가 주어짐에 따라 이전 선 확률 분포의 hyperparameter $\phi^{(n-1)}$ 가 진화하도록 함으로 각 시간에서 순차적응을 수행할 수 있다. QB 이론에 근거한 SSM의진화과정은

다음과 같이 주어진다 [18][21][24].

$$\begin{aligned}\overline{\boldsymbol{\mu}}^{(n)} &= \overline{\boldsymbol{\mu}}' \\ ?^{(n)} &= ?' \\ \mathbf{U}^{(n)} &= \mathbf{U}^{(n-1)}\end{aligned} \tag{10}$$

일반적으로 화자공간은 $\mathbf{U}$는 학습 화자에 대한 선 지식을 나타내므로 시간에 따라 변하지 않는다. 새롭게 진화하는 SSM의 평균값 $\overline{\boldsymbol{\mu}}'$과 $?'$는 SSM에 의한 평균 및 분산 추정 값이 의해 주어진다 [18][21][24].

SSM 적응기법은 주어진 초기 화자모델이 새로운 데이터가 주어짐에 따라 갱신된 평균 파라메터로 위치로 이동하여 화자 모델을 갱신한다. 데이터가 많아짐에 따라 화자모델은 점점 화자종속 모델로 순차적으로 수렴해 간다. 그러므로 SSM evolution 기법은 일괄적응 기법인 SSM을 포함한 일반적인 적응 구조를 갖고 있다. SSM evolution의 장점은 빠른 화자 적응 특성, 순차적응 특성 그리고 화자 종속 모델로 수렴하는 특성을 동시에 갖는다. 반면에 일괄SSM 적응기법과 같은 단점을 갖는다.

6.3 TSM evolution

TSM evolution 기법은 SSM evolution과 같이 TSM에 근거하여 QB 학습 알고리즘을 적용한 순차적응 알고리즘이다 [18][22][25]. 식 (8)로부터 $\mathbf{w}$에 대한 선 확률 분포는 hyperparameter $\theta = \{\overline{\mathbf{w}}, \mathbf{V}, ?\}$을 가지며 $p(\mathbf{w}|\theta) \sim N(\overline{\mathbf{w}}, ? + \mathbf{V}\mathbf{V}^T)$로 주어진다.

시간 n에서 적응 데이터 $\mathbf{X}^n$가 주어짐에 따라 이전 선 확률 분포의 hyperparameter $\theta^{(n-1)}$가 진화하도록 함으로 각 시간에서 순차적응을 수행할 수 있다. QB 이론에 근거한 TSM의진화 과정은 SSM 진화 과정과 같은 형태로 주어진다 [18][22][25].

$$\begin{aligned}\overline{\mathbf{w}}^{(n)} &= \mathbf{w}' \\ ?^{(n)} &= ?' \\ \mathbf{V}^{(n)} &= \mathbf{V}^{(n-1)}\end{aligned} \tag{11}$$

TSM evolution 기법은 일괄적응 기법인 TSM을 포함한 일반적인 적응 구조를 갖고 있다. TSM evolution의 장점은 대용량 인식 시스템에서도 변환 파라메터에 의한 빠른 화자 적응 및 순차적응을 동시에 수행할 수 있다.SSM과 마찬가지로 일괄 TSM 적응기법과 같은 단점을 갖는다.

7. 결론

본 논문에서 음성인식에서 성능향상을 위해 사용된 다양한 적응기법에 대한 기술동향을 살펴보았다. 화자적응에 대한 일반적인 개념에 대해 소개하고 음성인식과의 관계를 기술하였다. 적응기법은 MAP, 변환기반, 화자공간 그리고 순차적응 기법의 4가지로 분류하고 각각의 특징들에 대해 기술하였다. 여러 가지 적응기법 가운데에 SSM evolution과 TSM evolution 기법은 실용적인 화자적응의 기준에 만족하면서 좋은 적응 특성을 나타내었다. 현재 음성인식의 성능 향상을 위해 다양한 환경에서 화자적응 기술이 응용되고 있으며 음성인식 기술이 실용화할 수 있도록 여러 가지 기법들이 제안되고 사용되고 있다. 결론적으로 음성인식을 위한 화자 적응 기술은 앞으로 성능향상에 많은 기여를 할 것으로 예상된다.

참고문헌

[1] S. M. Ahadi and P. C. Woodland, "Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models," Computer. Speech and Lang., vol. 11, 1997, pp. 187-206.

[2] L. E. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov process," Inequalities., vol. 3, 1972, pp. 1-8.

[3] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in Proc. Int. Conf. Spoken Language Processing, Beijing, China, Oct. 2000,pp. 742-745.

[4] K.-T. Chen and H.-M. Wang, "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, USA, May 2001.

[5] J.-T. Chien, "On-line hierarchical transformation of hidden Markov models for speech recognition," IEEE Trans. Speech and Audio Proc., vol. 7, Nov. 1999, pp. 656-667.

[6] J.-T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models," IEEE Trans. Speech and Audio Proc., vol. 10, July 2002, pp. 268-278.

[7] W. Chou, "Maximum a posteriori linear regression with elliptically symmetric matrix priors," in Proc. Euro. Conf. Speech Commun., Technology, 1999, pp. 1-4.

[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoustics, Speech, Signal Proc., vol. ASSP-28, no. 4, Aug. 1980, pp. 357-366.

[9]P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, 1977, pp. 1-38.

[10] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," IEEE Trans. Speech and Audio Proc., vol. 8, no. 4, 2000, pp. 417-428.

[11] J. L. Gauvain and C. -H.Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Proc., vol. 2, 1994, pp. 291-298.

[12] X. D. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," IEEE Trans. Speech and Audio Proc., vol. 1, April 1993, pp. 150-157.

[13] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," IEEE Trans.Speech and Audio Proc., vol. 5, Mar. 1997, pp. 161-172.

[14] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," IEEE Trans. Speech and Audio Proc., vol. 6, July 1998, pp. 386-397.

[15] Q. Huo and B. Ma, "Online adaptive learning of continuous-density hidden Markov models based on multiple-stream prior evolution and posterior pooling," IEEE Trans. Speech and Audio Proc., vol. 9, May 2001, pp. 388-398.

[16] I. T. Jolliffe, Principal component analysis, Springer-Verlag. 1986.

[17] B.-H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," AT & T Technical Journal, vol. 64, no. 6, July-August 1985, pp. 1235-1248.

[18] D. K. Kim, Rapid speakeradaptation using speaker and transformation space models, PhD thesis, Seoul National Univ. Feb. 2003.

[19] D. K. Kim and N. S. Kim, "Bayesian speaker adaptation based on probabilistic principal component analysis," in Proc. Int. Conf. Spoken Language Processing, 2000, pp. 734-737.

[20] D. K. Kim and N. S. Kim, "Maximum a posteriori adaptation of HMM parameters based on principal component analysis," in Adaptation Methods for Speech Recognition, ISCA ITR-Workshop, Sophia-Antipolis, France, 2001, pp. 25-28.

[21] D.K. Kim and N. S. Kim, "Online adaptation of continuous density hidden Markov models based on speaker space model evolution," in Proc. Int. Conf. Spoken Language Processing, Denver, USA, 2002.

[22] D. K. Kim, Y. J. Kim, W. H. Lim, and N. S. Kim, "Online adaptation using transformation space model evolution," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 2003.

[23] D. K. Kim and N. S. Kim, "Maximum a posteriori adaptation of HMM parameters based on speaker space projection," Speech Communication, vol. 42, 2004, pp. 59-73.

[24] D. K. Kim and N. S. Kim, "Rapid online adaptation using speaker space model evolution," Speech Communication, vol. 42, 2004, pp. 467-478.

[25] D. K. Kim and N. S. Kim, "Rapid online adaptation based transformation space model evolution, " to be published in IEEE Trans. Speech and Audio Proc.,2004.

[26] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech and Audio Proc., vol. 8, no. 6, 2000, pp. 695-707.

[27] R. Kuhn,F. Perronnin and J.-C. Junqua, "Time is money: why very rapid adaptation matters," in Adaptation Methods for Speech Recognition, ISCA ITR-Workshop, Sophia-Antipolis, France, 2001, pp. 33-36.

[28] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," Speech Comm., vol. 25, 1997, pp. 29-47.

[29] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," Proc. of IEEE, vol. 88, no. 8, 2000, pp. 1241-1269.

[30] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, 1995, pp. 171-185.

[31] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of IEEE, vol. 77, Feb. 1989, pp. 257-286.

[32] D. Rubin and D. Thayer, "EM algorithms for factor analysis," Psychometrika, vol. 47 1982, pp. 69-76.

[33] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech and Audio Proc., vol. 9, 2001, pp. 276-287.

[34] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analysers," Neural Computation, vol. 11, no. 2, 1999, pp. 443-482.

[35] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Trans. Inform. Theory, vol. 13, Apr. 1967, pp. 260-269.

[36] P. C. Woodland, "Speaker adaptation for continuous density HMMs; a review," in Adaptation Methods for Speech Recognition,

ISCA ITR-Workshop, Sophia-Antipolis, France, 2001, pp. 11-19.

[37] G. Zavaliagkos, Maximum a posteriori adaptation techniques for speech recognition, PhD thesis, Northeastern Univ., Boston, MA, 1995.

저자소개

● 김동국(Dong-kook Kim)

1989년 2월 : 전남대학교 전자공학과 학사

1991년 2월 : 포항공과대학 전자전기공학과 석사

2003년 2월 : 서울대학교 전기컴퓨터공학부 박사

1991년 2월~1993년 3월 : 삼성전자 정보통신 연구원

1993년 3월~1999년 2월 : 삼성종합기술원 전문연구원

2000년 2월~2002년 12월 : ㈜넷더스 기술이사

2003년 4월~2004년 2월 : 한국전자통신연구원 선임연구원

2004년 2월~현재 : 전남대학교 전자컴퓨터정보통신공학부 전임강사

<관심분야> : 음성인식, 음성신호처리, 패턴인식