한국음향학회지 제23권 제8호 pp. 590~596 (2004)

한국어 연결 숫자음 인식을 위한 최대 사후 Eigenvoice에 근거한 자기적응 기법

Self-Adaptation Algorithm Based on Maximum A Posteriori Eigenvoice for Korean Connected Digit Recognition

전 형 배*, 김 동 국**
(Hyung Bae Jeon*, Dong Kook Kim**)

*한국전자통신연구원, **전남대학교 전자컴퓨터정보통신공학부
(접수일자: 2004년 8월 18일; 채택일자: 2004년 10월 28일)

본 논문에서는 한국어 연결 숫자음 인식을 위한 최대 사후 eigenvoice을 사용한 자기적응 기법을 제안한다. 제안된 최대 사후 eigenvoice 기법은 eigenvoice 계수에 대한 확률 밀도 함수를 가정함으로 구성된다. 제안된 알고리즘은 기존 eigenvoice 추정 과정에 선 분포 모델을 포함하는 일반적인 해를 제공하는 구조를 갖는다. 인식할 한 문장만을 사용하는 자기 적응 시스템을 위해 매우 강인한 특성을 갖는 최대 사후 eigenvoice 적응 기법을 사용하였다. 한국어 연결 숫자음에 대한 일련의 자기 적응 실험결과 제안된 알고리즘의 성능은 매우 적은 량의 적응 데이터에 대해 기존 eigenvoice 알고리즘에 비해 우수한 성능을 나타냈었다.

핵심용어: 최대 사후, 화자적응, 자기적응, 음성인식
투고분야: 음성처리 분야 (2.5)

This paper presents a new self-adaptation algorithm based on maximum a posteriori (MAP) eigenvoice for Korean connected digit recognition. The proposed MAP eigenvoice is developed by introducing a probability density model for the eigenvoice coefficients. The proposed approach provides a unified framework that incorporates the prior model into the conventional eigenvoice estimation. In self-adaptation system we use only one adaptation utterance that will be recognized, we use MAP eigenvoice that is most robust adaptation. In series of self-adaptation experiments on the Korean connected digit recognition task, we demonstrate that the performance of the proposed approach is better than that of the conventional eigenvoice algorithm for a small amount of adaptation data.

Keywords: Maximum a posteriori, Speaker Adaptation, Self-adaptation, Speech Recognition
ASK subject classification: Speech signal processing (2.5)

I. 서 론

현재 다양한 화자 적응 기술들은 음성인식의 학습과 인식 환경 사이의 불일치 (mismatch) 문제를 풀기 위해 많은 연구가 진행되어 왔다[1]. 실제 음성인식 시스템을 사용하기 위해서는 적은 량의 적응 데이터를 가지고 시스템을 적응하는 빠른 적응 기법이 필요하다[2]. 많은 모델기반 적응 기법들은 새로운 화자의 음향학적 특성들을 더 잘 일치시키기 위해 연속 밀도 은닉 마코프 모델 (CDHMM: continuous-density hidden Markov model)을 적응한다[1]. 일반적으로 모델기반 화자 적응 기법은 3가지 기법으로 분류된다[1]. 최대 사후 (Maximum a posteriori : MAP) 적응 기법[3], maximum likelihood linear regression (MLLR)을 포함한 변환기반 적응 기법[4], 그리고 eigenvoice[5]와 같은 화자 공간 (speaker space) 기법으로 분류된다. Kuhm[2] 등은 이러한 분류에 따라 빠른 적응 기법들에 대해 고찰하였고 화자 공간 기법들에 대해 자세히 기술하였다.

책임저자: 전 형 배 (hbjeong@etri.re.kr)
305-350 대전광역시 유성구 가정동 161
한국전자통신연구원
(전화: 042-860-5788; 팩스: 042-860-4889)

MLLR는 새로운 화자나 환경에 CDHMM 파라미터를 적응하기 위한 효율적인 빠른 적응 기법 중에 하나이다. MLLR는 maximum likelihood (ML) 기준에 의해 추정된 선형 regression 함수의 집합에 의해 CDHMM 파라미터가 적응되는 변환 기반 적응 기법이다[4,6]. 그러나 가용한 적응 데이터의 량이 매우 제한되어 있는 경우 많은 수의 파라미터를 강인하게 추정하는 것이 매우 어렵다. 이러한 문제를 극복하고 MLLR 적응 기법을 향상하기 위해 변환 파라미터에 대한 선 분포를 가정함으로 MAP 기준에 의한 maximum a posteriori linear regression (MAPLR)[7] 기법이 제안되었다.

빠른 적응을 위한 다른 기법으로는 eigenvoice와 같은 화자 공간 기법이 있다. Eigenvoice 기법은 학습과정 중에 많은 화자종속 (SD:speaker-dependent) 모델로부터 서로 직교하는 기본 벡터들을 구해서 화자사이에 존재하는 변이의 가장 중요한 성분들을 표현한다[5]. Eigenvoice는 적응될 모델이 적은 수의 기본 벡터들의 선형 결합에 의해 표현되므로 적응 데이터로부터 추정될 파라미터 수를 상당히 줄인다.

최근에 eigenvoice 기법을 Bayesian 적응 구조로 확장하기 위해 은닉 공간 모델 (latent variable models)에 근거하여 화자 공간 모델 (speaker space model)이라 불리는 빠른 화자 적응 기법이 소개되었다[8,9]. 또한 eigenspace-based MLLR과 MAPLR 기법들은 기존 principal component analysis (PCA)[10]와 probabilistic principal component analysis (PPCA)을 사용하여 학습 화자와 관련된 변환 매트릭스을 분석함으로 제안되었다[8,11].

여러 가지 실험을 통해 eigenvoice 기법이 매우 적은 량의 적응 데이터가 주어지는 빠른 적응에 대해 매우 뛰어난 성능을 나타냈었다. 특히 1 문장만을 적응 데이터로 사용한 고속 화자적응 4연 숫자 문장인식률에서 1.6%의 성능향상을 보이며 다른 고속 화자 적응 방법보다 좋은 성능을 보였다[12]. 그러므로 매우 적은 량의 적응 데이터가 주어지는 실제 환경에서는 eigenvoice와 같은 빠른 적응 기법이 요구된다. 특히 유선 전화나 무선 전화 환경 하에서 한국어 연결 숫자 음을 인식하는 시스템이 많이 개발되어 있다. 이와 같은 실제 환경에서는 사용자의 불편함 때문에 많은 적응 데이터를 요구할 수 없으므로 빠른 자기-적응 (self-adaptation) 기법이 요구된다. 자기-적응 시스템은 인식되어질 발음 문장을 가지고서 적응 데이터로 사용하여 음향 모델을 적응한다. 일반적으로 자기-적응의 성능은 교사학습 화자적응에 비해 화자 독립 인식기의 성능에 따라 감소한다. 그러므로 신뢰 (confidence) 척도에 의해 비교사 (unsupervised) 적응에서 사용되는 데이터를 적당히 선택하여 사용해야 한다.

이 논문에서는 MAPLR 기법과 같이 eigenvoice 계수에 대해 선 확률 분포를 가정함으로 eigenvoice의 성능을 향상시키는 MAP eigenvoice 기법을 제안한다. Eigenvoice 계수 값들의 분포를 모델링 하는 선 확률 밀도를 갖게 되면 매우 적은 량의 적응 데이터가 주어지는 경우에 Eigenvoice 계수에 대한 강인한 추정이 가능하다. Nguyen[13]은 MLLR의 변환 행들과 eigenspace 계수에 대한 가우시안 특성을 언급하였다. 이를 기반으로 eigenvoice의 계수를 random 변수라 취급하고 임의의 확률 밀도 함수을 갖는다고 가정한다. 이러한 경우 eigenvoice 계수에 대한 추정은 MAP 기준에 의해 기존 eigenvoice 추정 과정을 포함한 일반적인 해를 갖게 된다. 이 기법의 성능은 적절한 선 분포의 모델링과 선 분포의 선택에 밀접하게 관련된다. 이 분포들은 전체 학습 데이터로부터 적절히 선택되어 훈련되어 진다.

본 논문 구성은 다음과 같다. 먼저 2장에서는 eigenvoice 알고리즘에 대해 서술한다. 그리고 3장에서는 제안된 MAP eigenvoice 알고리즘과 선 확률 분포 선택에 대해 제시한다. 4장에서는 자기-적응 구조와 신뢰 척도에 대해 설명한다. 5장에서는 실험 환경에 결과에 대해 기술하고 마지막 6장에서 결과를 요약한다.

II. Eigenvoice

Eigenvoice 적응 기법은 새로운 화자 모델을 기존 화자 공간의 선형 결합에 의해 표현함으로서 화자 적응을 수행하는 효과적인 적응 기법이다[5]. Eigenvoice 기법을 사용하기 위해 두 가지 과정을 거친다. 첫 번째 단계에서는 eigenvoice를 학습하고, 두 번째 단계에서는 새로운 화자에 대해 적응된 새로운 모델을 구한다. 학습 단계에서는 먼저 학습 화자들에 대해서 화자 종속 모델을 학습한다. 그리고 화자 종속 모델의 평균 파라미터를 연결함으로 supervector을 만든다. 이러한 supervector들을 가지고서 PCA를 사용하여 기본 벡터들을 구한다. 이렇게 만들어진 PCA 공간을 화자 공간 (Speaker-space) 또는 eigenspace라 하며, 각각의 기본 벡터를 eigenvoice라 한다. 두 번째 적응 과정에서는 적응 데이터가 주어

지고, maximum likelihood eigen-decomposition (MLED)[5] 기법을 사용하여 eigenspace 내에서 새로운 화자의 위치를 구하여 적응하게 된다.

새로운 화자가 시스템에 주어지는 경우, 화자 적응 시스템은 적응 데이터 $o_t, t=1, \cdots, T$ 로부터 화자 적응된 화자 종속 모델을 구한다. 이때 m 번째 Gaussian mixture 평균 성분에 대한 갱신된 평균 벡터는 다음과 같다.

$$\mu_m = \sum_e w_e \quad u^m{}_e \tag{1}$$

여기서 m은 Gaussian mixture index, e는 eigenvoice index 그리고 w_e는 e 번째 eigenvoice $u^m{}_e$의 계수이다. 계수 w_e을 추정하기 위해서는 expectation maximization (EM) 알고리즘을 사용한다. Eigenvoice에 대한 보조 함수 Q는 다음과 같이 정의된다[7].

$$Q = -\frac{1}{2}\sum_{m,t}\gamma_m(t)\Big(o_t - \sum_e w_e \quad u^m{}_e\Big)^T \cdot R_m\Big(o_t - \sum_e w_e \quad u^m{}_e\Big) \tag{2}$$

여기서 $\gamma_m(t)$는 시간 t에서 Gaussian mixture 성분 m에 대한 occupation 확률이고, R_m는 Gaussian mixture 성분 m에 대한 precision 행렬이다. Gaussian mixture 성분 m에 해당하는 eigenspace의 한 부분에 해당하는 $D\times E$ eigenvoice 행렬을 다음과 같이 정의하자.

$$U_m = [\quad u_1{}^{(m)}, \cdots, \quad u_E{}^{(m)}] \tag{3}$$

그러면 Q-함수는 다음과 같이 다시 쓰여 진다.

$$Q = -\frac{1}{2}\sum_{m,t}\gamma_m(t)\big(o_t - U_m w\big)^T R_m\big(o_t - U_m w\big) \tag{4}$$

여기서 $w=[w_1, \cdots, w_E]^T$는 eigenvoice의 계수 벡터이다. 위 식을 기반으로 하여 계수 w 를 구하기 위해 w에 대해 미분 값을 취하여 0으로 놓으면 다음과 같다.

$$\frac{\partial Q}{\partial w} = -\sum_{m,t}\gamma_m(t) \quad U_m{}^T R_m\big(o_t - U_m w\big) = 0. \tag{5}$$

그러면 최종적으로 eigenvoice에 대한 계수 벡터 w에 대한 해는 다음과 같이 주어진다.

$$w = \Big(\sum_{m,t}\gamma_m(t) \quad U_m{}^T R_m \; U_m\Big)^{-1} \cdot \Big(\sum_{m,t}\gamma_m(t) \quad U_m{}^T R_m \; o_t\Big) \tag{6}$$

Eigenvoice 계수를 구한 후 이 계수를 이용하여 새로운 화자에 대한 적응된 평균 벡터는 식 (1)을 이용하여 구하게 된다.

III. 최대 사후 Eigenvoice

3.1. 최대 사후 Eigenvoice 유도

Eigenvoice 알고리즘은 ML 기준을 사용하여 화자 공간 내에 위치하는 해를 구한다. 그러나 ML에 의한 추정은 파라미터 w에 대한 어떤 제약 조건도 주어지지 않으며, 단지 화자 공간과 적응 데이터 o_t 만을 이용하여 적응시키는 알고리즘이다. 그러므로 eigenvoice의 성능을 향상시키기 위해 계수 w의 가능한 값에 대한 제약 조건을 주는 것이 바람직하다. 이를 위해 계수 w가 random 변수라고 가정하고 어떤 선 확률 밀도 함수, $P(w)$에 의해 기술된다고 가정한다. 이러한 제약 조건이 주어진 경우 eigenvoice 계수에 대한 추정은 MAP 기준, $\hat{w} = \arg\max_w P(O|w)P(w)$을 사용함으로서 할 수 있다. 본 논문에서는 선 분포 $P(w)$가 다음과 같은 Gaussian 분포를 갖는다고 가정한다[1].

$$P(w) \propto \Sigma^{-1}\exp\Big(-\frac{1}{2}(w-\overline{w})^T\Sigma^{-1}(w-\overline{w})\Big) \tag{7}$$

여기서 $\overline{w}$, Σ는 random 변수 w에 대한 평균 벡터와 공분산 행렬이며, 공분산 행렬 Σ은 대각 행렬이라고 가정한다. 그러면 MAP 추정을 위한 EM 알고리즘의 보조 Q-함수는 다음과 같이 정의된다.

$$Q = -\frac{1}{2}\sum_{m,t}\gamma_m(t)\Big(o_t - \sum_e w_e \quad u^m{}_e\Big)^T R_m \cdot \Big(o_t - \sum_e w_e \quad u^m{}_e\Big) + \log P(w) \tag{8}$$

위 식에 식 (7)에 정의된 w에 대한 선 분포를 대입하

여 전개한 후, 계수 w에 대해 Q-함수를 최대화하기 위해 $\partial Q/\partial w = 0$라 놓고 해를 구하면 다음과 같다.

$$w_{MAP} = \left(\sum_{m,t}\gamma_m(t)\ \ U_m{}^T R_m\ U_m + \Sigma^{-1}\right)^{-1} \cdot \left(\sum_{m,t}\gamma_m(t)\ \ U_m{}^T R_m\ o_t + \Sigma^{-1}\overline{w}\right) \quad (9)$$

기존 eigenvoice 해 식 (6)과 비교하면, MAP eigenvoice는 계수에 대한 선 확률 분포 모델을 기존 eigenvoice 추정 과정에 포함하는 통합된 형태의 구조를 보임을 위식을 통해 알 수 있다. 만약 w가 고정되고 알려지지 않은 상수라 하면, w에 대한 어떤 정보도 주어지지 않게 되며 이는 noninformative prioir 또는 improper prior, 즉 $P(w)$=상수라고 가정하는 것과 같다. 이러한 가정하에 MAP추정은 ML 추정과 같게 되며, 결국 MAP eigenvoice 해 (9)는 ML에 의한 해 식 (6)을 구하는 것과 같게 된다. 그러므로 MAP eigenvoice은 기존 eigenvoice을 포함하는 일반적인 적응 기법임을 알 수 있다.

3.2. 선 확률 밀도 선택

MAP 기준을 따라 파라미터를 추정하는 경우 알고리즘의 성능은 선 확률 밀도 분포의 적절한 선택과 매우 밀접한 관계가 있다. 일반적으로 MLLR의 변환 행렬에 대한 선 확률 분포 $P(W)$은 matrix variate normal 선 분포를 사용한다[7]. 그리고 eigenvoice 계수 w는 추정 과정 중에 가우시안 특성을 갖으며, mutigaussian 선 분포 사용을 제안하였다[13].

본 논문에서는 계수 w에 대한 선 확률 밀도 함수는 계산과 모델링을 간단히 하기 위해 single Gaussian 함수라 가정한다. 선 밀도 함수를 모델링하기 위해 eigenvocie 학습에 사용된 학습 화자의 데이터로부터 Gaussian 분포의 파라미터를 추정한다. 모든 학습 화자의 여러 학습 문장으로부터 단지 한 문장씩만을 사용한 교사 eigenvoice 적응 기법을 사용하여 계수 w에 대한 값들을 구한다. 이렇게 구한 w에 대한 값들을 가지고 각각의 학습 화자에 대한 single Gaussian에 대한 파라미터 값을 추정한다.

비교사 적응 과정 중에는 적응 데이터를 이용하여 eigenvoice 계수를 구하고 이 계수와 가장 가까운 Gaussian 모델을 현재 화자에 대한 선 밀도 함수로 선택한다. 가장 가까운 Gaussian을 선택하기 위한 거리 척도로서 추정된 계수와 Gaussian 분포사이의 Mahalanobis 거리 척도를 사용하였다.

IV. 자기-적응 시스템

4.1. 자기-적응 구조

자기-적응 시스템은 인식되어질 발화를 사용하여 먼저 인식하고 인식된 결과와 인식 문장을 사용해 음향 모델을 적응하는 구조를 갖는다. 일반적으로 자기-적응 시스템은 2-pass 탐색 구조를 갖는다. 첫 번째 pass 탐색은 화자 독립 모델을 사용하여 탐색 과정을 수행하여 phone 정렬 정보와 적응 과정 동안에 필요한 통계 데이터 정보를 구한다. 첫 번째 pass 탐색에서 얻어진 정렬 정보와 통계 데이터 정보를 이용하여 원하는 적응 기법을 사용해 특정 화자 모델을 얻게 된다. 그리고 두 번째 pass 탐색에서는 적응 과정에서 얻어진 적응된 특정 화자의 음향 모델을 가지고 다시 탐색 과정을 수행하여 최종적인 인식 결과를 얻게 된다.

본 실험에서 사용되는 인식기는 한국어 4연 숫자음을 인식하는 시스템이다. 그러므로 인식과 적응 과정에서 사용 될 데이터는 4개의 연결 숫자음으로 구성된 단 한 발화이다. 인식 대상 발화를 사용하여 먼저 화자 독립 1-pass 탐색을 이용하여 4연 숫자음에 대한 인식 결과를 얻게 된다. 그러나 이 결과는 첫 번째 pass 탐색기의 성능에 따른 인식 오류를 갖게 된다. 오류를 포함한 정보를 이용하여 적응을 수행하게 되면 잘못 적응된 모델을 얻게 되므로 적응 과정 동안에 성능을 향상시키기 위해 신뢰 척도를 이용하여 오류가 없는 데이터를 선택해야 한다. 그러나 이런 경우 유용한 적응 데이터의 량은 점점 줄어들게 된다. 이와 같이 매우 적은 량의 적응 데이터가 주어지는 자기-적응 환경 하에서 시스템은 빠르고 강인하게 잘 동작하는 적응 알고리즘이 필요하게 된다. 그러므로 이 논문에서 제안된 MAP eigenvoice 기법은 이와 같은 자기-적응 환경에 매우 적합한 알고리즘이라 할 수 있다.

4.2. 신뢰 척도 (Confidence Measure)

자기-적응의 경우 비교사 적응을 수행하기 때문에 적응 성능은 화자 독립 인식기의 성능이 저하됨에 따라 감소하게 된다. 특히 인식 결과가 많은 오류를 포함하게

되면 적응 성능은 현저하게 떨어진다. 그러므로 적응 과정 동안에 사용될 오류가 없는 신뢰성있는 인식 데이터를 신뢰 척도에 의해 선택해야 한다. 이를 위해 신뢰 척도를 위한 시스템 구성이 필요하게 된다.

본 논문에서 신뢰 척도를 위한 구조는 다음과 같다. 먼저 tri-phone score와 anti-phone score사이의 log-likelihood ratio (LLR)을 신뢰 척도로 정의한다. Anti-phone을 정의하기 위해 phone 에 따라 cohort 집합을 정의한다. Cohort 집합에 따라 다양한 수의 mixture을 가지는 anti-phone을 만들었다. 신뢰 값을 얻기 위해 서로 다른 수의 mixture을 가지는 anti-phone과 비교하기 위해 anti-phone mixtur중에 12개의 활성화된 mixture를 사용하였다.

적응을 위해 임계치보다 더 큰 신뢰 값을 가진 데이터만을 적응 데이터로 선택했다. 임계치 값은 오류가 최소가 되도록 경험적으로 설정하였다. 최상의 임계치 값을 구하기 위해 학습 데이터로부터 LLR 분포를 살펴서 각각의 tri-phone에 대해 임계치를 결정하였다.

V. 실험 및 결과

5.1. 실험 환경 및 데이터 베이스

실험은 한국어 4연 숫자 음으로 구성된 두개의 다른 데이터베이스를 가지고 수행되었다. 실험 1에서는 동일한 유선 전화선 채널을 통해 294명의 화자로부터 얻어진 데이터베이스가 사용되었다. 각 화자 당 데이터는 200 문장 내외로 구성되었다. 245 화자의 약 50,000 문장이 학습을 위해 사용되었고, 나머지 49 화자의 약 10,000 문장을 가지고 테스트를 하였다.

실험 2에서는 다양한 곳에서 발신한 유선/무선 전화와 이동 전화 환경에서 2,800명의 화자로부터 모아진 데이터베이스가 이용되었다. 이동 전화 환경에서 모아진 데이터베이스는 5가지 다른 이동 전화 채널을 통해 다양한 환경 하에서 수집되었다. 각 화자 당 30개 내외의 문장이 발성되었다. 학습은 2,600명의 화자가 발성한 약 70,000 문장으로 훈련하였고, 테스트는 200명의 화자가 발성한 약 4,600 문장을 사용하였다.

실험 1에서는 다양한 화자간의 불일치에 대한 적응 실험을 수행하고 실험 2에서는 화자뿐 아니라 환경에 대한 불일치에 대한 적응 실험을 수행하였다.

특징 벡터로는 일반적인 mel-frequency cepstral coefficient (MFCC)을 사용하였다. 각 특징 벡터는 C0을 포함한 13차 cepstral 계수, 1차 미분치 그리고 2차 미분치로 구성된 39차의 성분으로 구성되었다.

CDHMM 음향 모델은 한국어 숫자음에서 관찰된 342개의 tri-phone과 두 개의 묵음 모델을 가지고 학습하였다. 각 tri-phone CDHMM은 5개의 상태와 6개의 mixture 성분을 가지며, 묵음 CDHMM은 1개의 상태와 16개의 mixture 성분을 각각 갖는다. 그러므로 eigenvoice를 위해 각 supervector는 401,388개의 파라미터 수를 갖는다.

화자 종속 모델을 얻기 위해 전체 변환 MLLR과 MAP이 결합된 적응 기법을 사용하였고 supervector는 각각의 화자 종속 모델로부터 구성되었다. 각 supervector로부터 공분산 행렬을 만들고 이에 PCA을 적용하여 K개의 eigenvoice를 얻었다.

실험 1에서는 학습 화자 중 화자종속 모델을 훈련하기에 충분한 화자 244명으로부터 244개의 화자 종속 모델을 훈련하고, 훈련된 화자 종속 모델로부터 20개의 eigenvoice을 얻었다. 실험 2에서는 매우 많은 학습 화자와 다양한 채널과 배경 잡음이 존재한다. 그러므로 26,000명의 학습 화자로부터 최대한 많은 화자 종속 모델을 훈련하고, 이로부터 eigenvoice를 구하는 것이 다양한 주변 환경을 적응 하는데 유리하게 된다. 그러나 훈련에 사용된 화자 중에는 화자 종속 모델을 훈련하기에 충분한 데이터를 가지지 못하는 화자가 상당수 포함되어 있다. 그렇기 때문에 훈련 화자 중 800명의 화자를 선택하여 화자 종속 모델을 훈련하고, 훈련된 화자종속 모델로부터 20개의 eigenvoice을 구하였다.

실험 2에서는 추가로 multi-path 탐색에 대한 화자적응 실험을 수행하였다. 첫 번째 pass decoder에서는 남자와 여자의 각각 다른 음향 모델을 갖는 두 개의 decoder를 사용하였다. 이러한 multi-path decoder을 위해 남자와 여자의 학습 화자로부터 남, 녀 각각 다른 20개의 eigenvoice을 구하였다. 첫 번째 pass 탐색에서는 acoustic score에 의해 어떤 모델이 맞는 모델인지 결정하였고 이 결과를 바탕으로 해당된 모델을 eigenvoice로 학습하였다. 그리고 두 번째 pass 탐색에서는 적응된 모델을 가지고 최종 결과를 탐색하였다.

5.2. 실험 결과

실험 1에 대한 결과는 그림 1에 주어졌다. 그림은

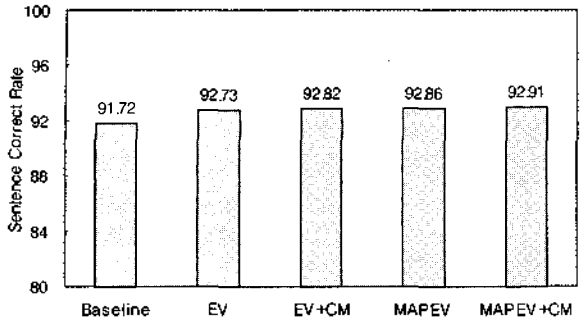

그림 1. 동일한 유선 전화선 환경에서의 eigenvoice와 MAP eigenvoice를 사용한 자기-적응 실험 결과

Figure 1. The experimental results of self-adaptation with Eigenvoice and MAP Eigenvoice in same telephone line environment.

eigenvoice와 제안된 MAP eigenvoice 알고리즘을 문장 인식률에 따라 비교하였다.

그림에서 "Baseline"은 화자 독립 모델을 사용하는 경우의 성능이다. "EV"는 첫 번째 pass의 결과가 정확하다는 가정 하에 모든 데이터를 적응을 위해 사용한 eigenvoice에 근거한 비교사 적응 실험의 결과이다. "EV + CM"는 신뢰 척도에 의해 적응을 위한 데이터를 선택하는 eigenvoice 기반 비교사 적응 실험에 대한 결과이다. 일반적인 eigenvoice (EV) 화자적응 방법에 의한 성능은 기본 성능에 대해 12.2%의 error reducation rate (ERR) 성능 향상을 나타내었다. "MAPEV"와 "MAPEV+CM"은 제안된 MAP eigenvoice 기법과 신뢰 척도를 각각 사용한 실험에 대한 결과이다. 신뢰 척도를 사용함으로서 eigenvoice에 비해 1%의 ERR 성능개선을 나타내었다. MAP eigenvoice를 사용하는 경우에는 일반적인 eigenvoice에 비해 1.1%~1.5%의 향상된 ERR 성능향상을 나타내었다.

실험 2에서 데이터베이스는 사무실, 집, 공중전화, 거리, 지하철 등 다양한 환경 하에서 유무선 전화를 사용한 공중망과 5개의 이동 통신 전화망을 통해 수집되었다. 이러한 데이터베이스를 사용한 실험 결과는 그림 2에 주어졌다. 다양한 채널 변이에 의해 기본 시스템의 성능은 실험 1에 비해 저하되었다. 일반적인 eigenvoice (EV) 화자적응 방법에 의한 성능은 기본 성능에 대해 10.1%의 ERR 성능향상을 나타내었다. 제안한 MAP eigenvoice에 비해 1.1%-1.5%의 ERR의 향상을 보이면서 동일한 환경에서의 성능개선과 동일한 성능 개선을 나타내었다.

실험 3에서는 시스템의 성능 향상을 위해 좀 더 정교한 음향 모델을 사용하였다. 첫 번째 pass 탐색에서는 남자와 여자 모델을 각각 사용하였다. Multi-path 탐색을 통해 기본 성능으로 1.1%의 향상된 문장 인식률을 얻었다. 이와 같은 시스템을 기반으로 신뢰 척도를 갖는 eigenvoice을 사용하는 경우 4.37%의 ERR 향상을 얻을 수 있었다. 일반적으로 eigenvoice의 첫 번째 계수는 성별 특성을 나타내는 걸로 알려졌다. 그러므로 남녀 음향 모델을 사용하는 multi-path시스템에서의 eigenvoice 화자적응을 통한 ERR 향상 정도는 화자 독립 시스템에서의 eigenvoice 화자적응을 통한 ERR 향상 정도 보다 낮게 나타났다. 이 실험에서는 MAP eigenvoice가 기존 eigenvoice에 비해 2.4%의 향상된 ERR 성능 향상을 나타내었다.

VI. 결 론

본 논문에서는 고속 화자적응 방법 중 성능이 가장 우수 하였던 eigenvoice의 성능을 향상시키기 위해서 선

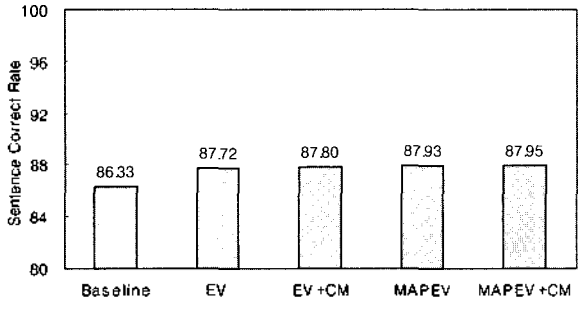

그림 2. 다양한 유무선 및 이동 통신 환경하에서 eigenvoice와 MAP eigenvoice을 사용한 자기-적응 실험 결과

Figure 2. The experimental results of self-adaptation with Eigenvoice and MAP Eigenvoice in various mobile and telephone line environment.

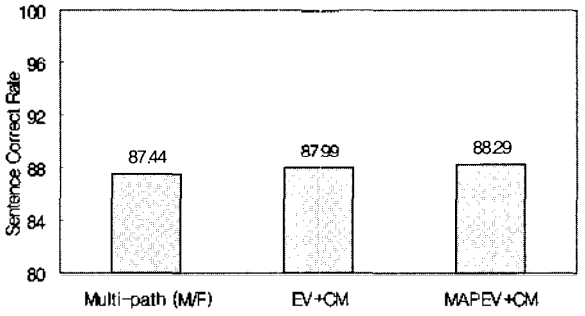

그림 3. Multi-path 탐색을 통한 다양한 유무선 및 이동 통신 환경에서 eigenvoice와 MAP eigenvoice의 자기-적응 실험 결과

Figure 3. The experimental results of self-adaptation with Eigenvoice and MAP Eigenvoice in various mobile and telephone line environment with multi-path decoder.

확률 밀도를 가지고 eigenvoice의 계수를 추정하는 MAP eigenvoice 기법을 제시하였다. MAP eigenvoice는 기존 eigenvoice 추정 기법에 선 모델 특성을 반영하도록 하는 일반적인 형태의 구조를 갖는다. 제안된 알고리즘은 한국어 4연 숫자 인식을 위한 자기-적응 시스템에 고속 화자 적응 기법으로서 적용되었다. 실험 결과 MAP eigenvoice 기법은 단 한 문장의 적응 데이터가 주어지는 자기-적응 환경에서 기존 eigenvoice 화자적응 방법보다 1.1%~1.5% 정도의 ERR이 향상 되었다. 또한 유선, 무선 및 이동 통신 환경에서도 기존 eigenvoice보다 좋은 성능을 나타내었다.

앞으로 본 논문에서 사용한 선 확률 밀도 함수보다 정교한 선 확률 밀도 함수에 대한 연구를 통하여 더욱 개선된 MAP eigenvoice 기법에 대한 연구를 수행할 예정이다.

참 고 문 헌

1. P. C. Woodland, "Speaker adaptation for continuous density HMMs: a review," in Proc. Adaptation Methods for Speech Recognition, ISCA ITR-Workshop, Sophia-Antipolis, France, pp. 11-19, 2001.
2. R. Kuhn, F. Perronnin and J. -C. Junqua, "Time is money: why very rapid adaptation matters," in Proc. Adaptation Methods for Speech Recognition, ISCA ITR-Workshop, Sophia-Antipolis, France, 33-36, 2001.
3. 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., 2, 291-298, 1994.
4. C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, 9, 171-185, 1995.
5. R. Kuhn, J. -C. Junqua, P. Nguyen and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice Space," IEEE Trans. Speech and Audio Proc., 8(6), 695-707, 2000.
6. Ho-Young Jung, Mansoo Park, Hoi-Rin Kim, and Minsoo Hahn, "Speaker Adaptation Using ICA-Based Feature Transfomation," ETRI J., 24(6), pp.469-472, Dec. 2002.
7. W. Chou, "Maximum a posteriori linear regression with elliptically symmetric matrix priors," in Proc. Euro. Conf. Speech Commun., Technology, 1, 1-4, 1999.
8. D. K. Kim and N. S. Kim, "Rapid speaker adaptation using probabilistic principal component analysis," IEEE Signal Processing Letters, 8(6), 180-183, June 2001.
9. 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.
10. I. T. Jolliffe, Principal Component Analysis. Springer-Verlag, 1986.
11. 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, 742-745, Oct. 2000.
12. 전형배, 김동국, "연결 숫자음 인식에서의 고속 화자 적응", 제 20회 음성통신 및 신호처리 학술대회 논문집, pp 441-444, 2003.
13. P. Nguyen, "Speaker adaptation: Modeling variabilities," Ph.D. thesis, 2002.

저자 약력

● 전 형 배 (Hyung Bae Jeon)

1976년 7월 1일생
1999년 2월 : 연세대학교 전자공학과 학사
2001년 2월 : 한국과학기술원 전기및전자공학과 석사
2001년 3월현재 : 한국전자통신연구원 연구원

● 김 동 국 (Dong Kook Kim)

1966년 8월 11일생
1989.2: 전남대학교 전자공학과 학사
1991.2: 포항공과대학 전자전기공학과 석사
2003.2: 서울대학교 전기컴퓨터공학부 박사
1991.2~1993.3: 삼성전자 정보통신 연구원
1993.3~1999.2: 삼성종합기술원 전문연구원
2000.2~2002.12: ㈜넷더스 기술이사
2003.4~2004.2: 한국전자통신연구원 선임연구원
2004.2~현재: 전남대학교 전자컴퓨터정보통신공학부 전임강사