한국음향학회지 제21권 제5호 pp. 455~461 (2002)

음성 다이얼링을 위한 화자적응

Speaker Adaptation for Voice Dialing

김 원 구*, Chin-Hui Lee**
(Weon-Goo Kim*)

*군산대학교 전자정보공학부, **Bell Labs, Lucent Technologies
(접수일자: 2001년 7월 23일; 수정일자: 2002년 5월 11일; 채택일자: 2002년 7월 4일)

본 논문에서는 화자독립 음소 모델을 사용하는 개인용 음성 다이얼링 시스템의 성능 개선 방법을 제안하였다. 화자독립 음소모델을 사용한 음성 다이얼링 방법은 각 화자가 발성한 단어와 연관된 음소 열만을 저장하므로 저장 공간은 크게 줄일 수 있으나 화자독립 모델을 음소 인식에 사용할 때 발생하는 오차로 인하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점이 있다. 본 논문에서는 이러한 문제를 해결하기 위하여 학습과정에서 학습 데이터의 음소 열과 화자 적응을 위한 변환 벡터를 동시에 추정한 후 음소 열과 함께 저장하고, 인식 시에 화자독립 음소 모델을 각 화자의 변환벡터를 사용하여 변환한 후 인식을 수행하는 방법을 제안하였다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭 (stochastic matching)을 위한 최고 유사도 (maximum likelihood) 방법을 이용하여 구하였으며 음소 열과 함께 반복적으로 추정되었다. 인식 실험에서 제안된 방법은 음소 열만을 사용하는 기존 인식 시스템보다 우수한 성능을 나타내었다.
핵심용어: 음성인식, 음성 다이얼링, 화자적응, 확률적 매칭, 음소 HMM
투고분야: 음성처리 분야 (2.5)

This paper presents a method that improves the performance of the personal voice dialling system in which speaker independent phoneme HMM's are used. Since the speaker independent phoneme HMM based voice dialing system uses only the phone transcription of the input sentence, the storage space could be reduced greatly. However, the performance of the system is worse than that of the system which uses the speaker dependent models due to the phone recognition errors generated when the speaker independent models are used. In order to solve this problem, a new method that jointly estimates transformation vectors for the speaker adaptation and transcriptions from training utterances is presented. The biases and transcriptions are estimated iteratively from the training data of each user with maximum likelihood approach to the stochastic matching using speaker-independent phone models. Experimental result shows that the proposed method is superior to the conventional method which used transcriptions only.
Keywords: *Speech recognition, Voice dialling, Speaker adaptation, Stochastic matching, Phone HMM*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성 다이얼링은 임의의 단어나 문장을 사용하여 전화를 거는 방법으로 전화나 휴대폰에 성공적으로 상용화되어 사용되고 있다. 일반적으로 음성 다이얼일 시스템은 화자 종속형의 시스템을 사용하여 각 화자가 자동적으로 전화를 걸 때 사용할 명령이나 키워드를 포함하는 개인적인 목록을 사용한다. HMM (Hidden Markov Model)을 이용한 음성 다이얼링 시스템의 구조는 그림 1과 같다.

책임저자: 김원구 (wgkim@kunsan.ac.kr)
573-701 전북 군산시 미룡동 산 68
군산대학교 전자정보공학부
(전화: 063-469-4745; 팩스: 063-469-4699)

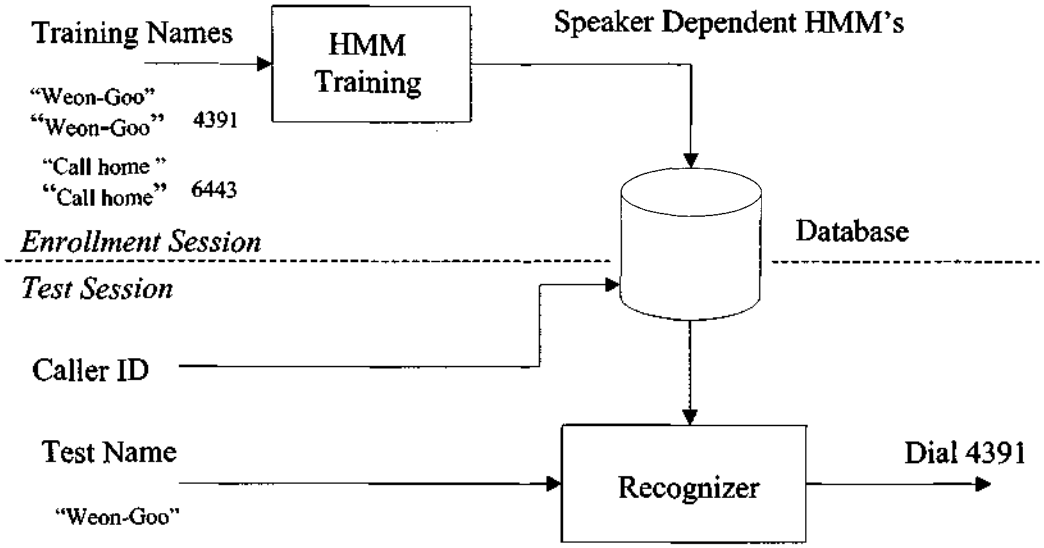

그림 1. HMM을 이용한 개인용 음성 다이얼링 시스템의 개념도
Fig. 1. The concept of personal voice dialing system using HMM.

이러한 시스템의 구성은 보통 다음과 같은 두 가지 과정을 거친다.

(1) 등록 단계: 사용자는 각각의 단어나 문장을 수 차례 발성하고 그에 해당되는 전화번호를 제공한다. 생성된 모델은 인식을 위하여 저장된다.

(2) 인식 단계: 사용자의 신원이 확인된 후 사용자는 단어 또는 문장을 발음하고 시스템은 인식을 수행하여 자동으로 해당된 전화번호로 전화를 건다.

이러한 형태의 시스템은 그 구조가 간단하고 화자종속의 형태를 갖기 때문에 인식 성능이 비교적 우수하지만 단어나 문장 단위로 모델을 저장해야 하기 때문에 저장공간이 많이 필요하고 인식 대상 단어수의 증가에 비례하여 필요한 저장도 증가하게 된다. 이러한 문제점은 핸드폰에 사용되는 음성 다이얼링 시스템과 같이 한 명의 사용자가 수십 단어 정도를 사용하는 경우에는 큰 문제가 되지 않지만 전화망이나 네트워크를 사용한 음성 다이얼링인 경우와 같이 수십 또는 수백만 명의 데이터를 서비스 사업자의 서버에 저장해야 하는 경우에는 음성인식을 수행하기 위한 데이터 저장공간의 크기가 매우 커지기 때문에 중요한 문제가 된다.

이러한 문제를 해결하기 위한 방법 중의 하나로 화자독립 음소모델을 이용한 방법들이 제안되었다[1-4]. 이러한 방법들은 화자독립 음소모델을 사용하여 학습 데이터의 음소 열을 구한 후 음소 열을 저장하고, 입력 음성을 인식할 때 저장된 음소 열과 화자독립 모델을 사용하는 것이다. 이러한 방법들의 장점은 각 화자마다 저장해야 할 정보가 각 화자가 발성한 단어와 연관된 음소 열이기 때문에 저장해야 할 데이터 양이 매우 적어진다는 점이다. 이러한 방법은 저장공간은 크게 줄일 수 있으나 다음과 같은 두 가지 문제점을 가지고 있다. 첫 번째는 화자독립 음소 HMM을 사용한 음소 열 추정 결과에 많은 오차가 발생하는 것이다. 음소 열 추정에는 문법적인 정보를 전혀 사용하지 않기 때문에 특히 음성 구간의 전,후 및 잡음 구간에서 무성음 오차가 많이 발생한다. 두 번째는 화자독립 모델을 음소 인식에 사용할 때 발생하는 오차로 인하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점이 있다.

본 논문에서는 화자독립 음소 모델을 사용한 음성 다이얼링 시스템의 성능을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환 함수를 동시에 추정하는 방법을 제안하였다. 제안된 방법은 학습과정에서 학습 데이터에 대하여 음소 열과 화자적응을 위한 변환 벡터를 동시에 추정한 후 변환 벡터와 음소 열을 함께 저장하고, 인식 과정에서는 화자독립 음소 모델을 각 화자의 변환벡터를 사용하여 변환한 후 입력 음성에 대한 인식을 수행한다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭 (stochastic matching) 방법을 위한 최고 유사도 (maximum likelihood) 방법[5,6]을 이용하였으며 음소 열과 함께 반복적으로 추정되었다. 이러한 변환 벡터는 크기가 작아서 적은 저장공간을 사용하면서도 인식 성능을 화자종속 시스템에 근사하도록 향상시킬 수 있었다.

Ⅱ. 화자독립 HMM을 이용한 음성 다이얼링 시스템의 화자적응

본 논문에서는 화자독립 음소모델을 사용한 음성 다이얼링 시스템의 성능을 향상시키기 위하여 화자적응 방법을 사용하여 성능을 향상시키는 방법을 제안하였다. 기존 화자독립 모델을 사용하는 음성 다이얼링 시스템은 학습단계에서 등록에 사용되는 음성으로부터 음소인식을 수행하여 음소 열을 구한 후 이 음소 열을 저장한다[1-4]. 인식 단계에서는 저장된 음소 열과 화자독립 음소 HMM을 연결한 모델을 만든 후 입력 음성에 대한 확률을 구한다. 이러한 방법은 저장해야 할 데이터가 음소 열이므로 필요한 저장공간이 매우 작아지는 장점이 있다. 그러나 이러한 방법은 저장공간은 크게 줄일 수 있으나 화자독립 모델을 음소 인식에 사용할 때 발생하는 오차로 인하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점이 있다.

본 논문에서는 화자독립 음소모델을 사용한 음성 다이얼링 시스템의 성능을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환함수를 동시에 추정하는 방법을 제안하였다. 제안된 시스템의 구조는 그림 2와 같다.

제안된 방법은 등록 단계 (enrollment session)인 학습 과정에서 학습 데이터와 화자독립 음소 HMM을 사용하여 학습 데이터의 음소 열과 화자적응을 위한 변환 벡터 (bias)를 동시에 추정한 후 음소 열과 함께 저장하고, 인식 단계 (test session)에서 화자독립 음소 HMM을 각 화자의 변환벡터를 사용하여 변환한 후 입력 음성에 대한 인식을 수행한다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭 (stochastic matching)을 위한 최고 유사도 (maximum likelihood) 방법[5,6]을 이용하였으며 음소 열과 함께 반복적으로 추정되었다.

확률적 매칭을 위한 최고 유사도 방법은 다음과 같이 적용될 수 있다. 우선 일련의 특징 벡터 $Y=\{y_1, y_2, \cdots, y_T\}$와 화자독립 음소 HMM의 집합을 Λ_X라고 할 때, 화자적응되어 변형된 모델 Λ_Y를 위한 모델 공간 변환은 다음과 같은 식으로 이루어진다.

$$\Lambda_Y = G_\eta(\Lambda_X) \tag{1}$$

여기서 $G_\eta(.)$는 모델 변환 함수이고 η는 변환 파라메터이다. 변환 파라메터 η와 음소 열 W의 동시 최대화는 다음과 같이 정의된다.

$$\begin{aligned}(\eta', W') &= \arg\max_{(\eta, W)} p(Y, W \mid \eta, \Lambda_X) \\ &= \arg\max_{(\eta, W)} p(Y \mid W, \eta, \Lambda_X) P(W)\end{aligned} \tag{2}$$

따라서 추정된 음소 열 W와 확률적 매칭 방법을 사용하여 구한 변환 벡터 η'는 다음과 같다.

$$\eta' = \arg\max_{\eta} p(Y \mid W, \eta, \Lambda_X) P(W)$$

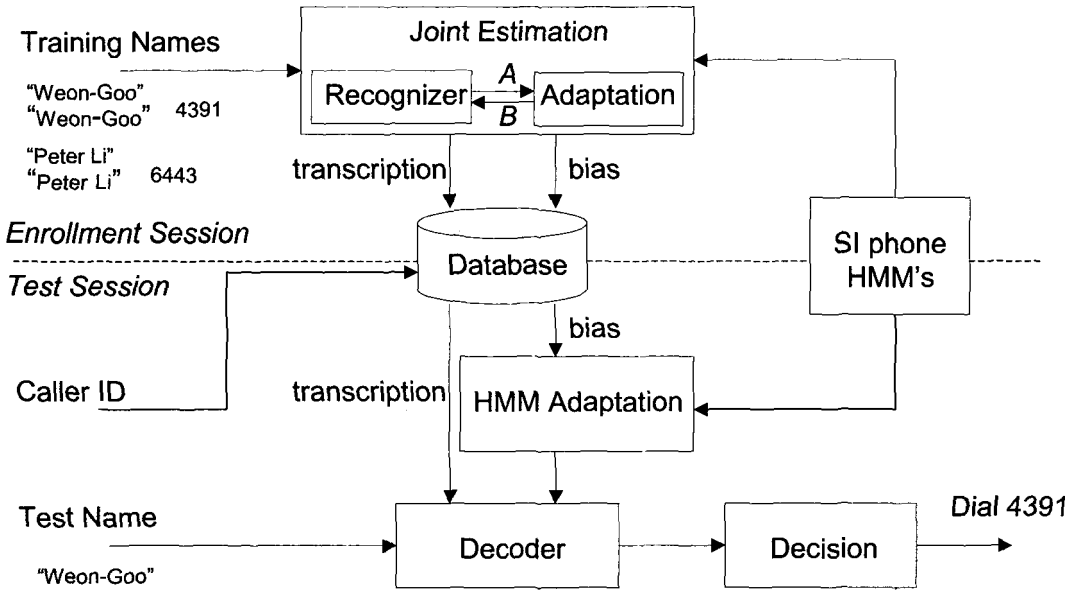

그림 2. 화자적응과 HMM을 이용한 개인용 음성 다이얼링 시스템의 개념도
Fig. 2. The concept of personal voice dialing system using HMM and speaker adaptation.

고델 변환 $\eta \leftarrow \eta'$을 $\mu_y = \mu_x + \mu_b$의 형태로 가정하면 변환 벡터 μ'_b는 최고 유사도 추정에 의하여 다음과 같이 구할 수 있다.

$$\mu'_{b_i} = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\frac{y_{t,i} - \mu_{n,m,i}}{\sigma^2_{n,m,i}}}{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)}, \quad i=1,\cdots,D \tag{3}$$

$$\gamma_t'(n,m) = \begin{cases} \dfrac{w_{n,m}M[y_t;\mu_{n,m},C_{n,m}]}{\sum_{j=1}^{M}w_{n,j}M[y_t;\mu_{n,j},C_{n,j}]}, & \text{if } \hat{s}=n \\ 0, & otherwise \end{cases} \tag{4}$$

여기서 N과 M은 HMM 상태 수와 분포 (mixture) 수를 나타내며 D는 특징벡터의 차수이다. $\mu_{n,m,i}$,와 $\sigma_{n,m,i}$는 각각 HMM에서 n번째 상태의 m번째 분포 (mixture)의 i번째 차수의 평균과 분산이며 $w_{n,m}$는 분포 (mixture) 가중이고 $M[\cdot]$은 가우시안 분포를 나타낸다. 또한 $\hat{s}$는 입력 벡터 열에 대한 상태 열 (state sequence)이다.

확률적 매칭을 위한 최고 유사도 방법을 적용한 음성 다이얼링 시스템의 학습 및 인식 과정은 다음과 같다.

● 학습 과정

1. 화자독립 음소 HMM Λ_X을 이용하여 학습데이터 Y에 대한 초기 음소 열 W를 추정한다.
2. 추정된 음소 열 W와 확률적 매칭 방법을 사용하여 변환 벡터 η'를 구한다.
3. 변환 벡터 η'를 이용하여 Λ_X를 변환된 음소모델 Λ_Y로 변환한다.
4. 변환된 음소모델 Λ_Y를 사용하여 음소 열 W를 다시 구한다.
5. 단계 2-4를 모델이 수렴될 때까지 반복한다.
6. 최종 변환 벡터 η와 최종 음소 열 W를 인식 과정을 위하여 저장한다.

● 인식 과정

1. 발신자 확인 (caller ID) 등에 의한 방법으로 입력 화자의 신원이 확인되면 화자독립 음소 HMM Λ_X를 입력 화자의 변환 벡터 η를 사용하여 변환시킨다.
2. 변환된 화자독립 음소 HMM Λ_Y과 저장된 음소 열 W를 사용하여 입력 음성을 인식한다.

III. 실험 및 결과

3.1. 데이터베이스 및 인식 시스템 구성

실험에 사용된 데이터 베이스는 남성 5명과 여성 5명의 총 10명으로 구성하였다[7]. 각 화자는 15개의 단어를 발음하였다. 데이터 녹음은 전화선을 통하여 이루어 졌으며 각 화자는 각기 다른 환경에서 가급적 다른 종류의 전화기를 사용하여 몇 주 간격을 두고 녹음하였다. 음성 신호는 6.67 kHz로 샘플링되었고 8 bit μ-law PCM으로 저장되었다. 학습에 사용된 데이터는 각 화자가 15개의 이름을 3회 반복한 것 (15개×3회=45개/명)으로 구성하였으며, 인식에 사용된 데이터는 각기 다른 날짜에 수행한 5회의 녹음에서 각 화자가 15개의 이름을 10회 반복한 데이터 (15개×10회=150개/명)로 구성하였다. 데이터 내용은 영어로 "Call office", "Call home", "Call mom" 등으로 구성되었다. 이 데이터 베이스는 모두 같은 단어로 시작되기 때문에 인식하기에 매우 어렵고 단어의 길이도 대부분 1초 이내로 매우 짧아 인식을 더욱 어렵게 한다.

실험에 사용된 특징벡터는 12차 LPC 켑스트럼, 1차 차분 켑스트럼, 2차 차분 켑스트럼, 에너지, 1차 차분 에너지, 2차 차분 에너지의 총 39차 벡터로 구성되었다. 켑스트럼 계수는 30 ms의 창 길이를 갖고 10 ms씩 이동하면서 구한 10차 LPC 계수로부터 구하였다.

화자독립 음소 HMM은 연속음성 인식을 위하여 전화선을 통하여 녹음된 데이터베이스를 사용하여 학습된 모델을 사용하였다. 따라서 본 실험에 참여한 화자와 중복된 경우는 없었다. 이러한 모델은 각 음소마다 3개 또는 5개의 상태 수를 갖는 LTR (left-to-right) 형태의 음소 모델 41개와 1개의 상태를 갖는 묵음 모델로 구성되었고 각각의 HMM은 연속밀도분포를 갖는 연속분포 HMM이다. 이러한 모델을 사용하여 입력 음성에 대한 음소 열을 추정하였다.

인식을 결정하기 위한 결정 법칙 (decision rule)은 K-NN (K-Nearest Neighbor) 법칙을 사용하였고 KNN은 2로 하였다.

3.2. 기준 시스템의 성능 평가

제안된 방법의 비교 평가를 위하여 기준 시스템을 구성하여 성능 평가를 수행하였다. 기준 시스템은 화자독립 음소 HMM을 사용한 화자 종속 음성 다이얼링 시스템으로 구현하였다. 구현된 시스템은 화자독립 음소모델을 사용하여 학습 데이터의 음소 열을 구하여 저장하고, 입력 음성을 인식할 때 저장된 음소 열과 화자독립 모델을

표 1. 음성 구간 검출 방법 사용에 따른 기준 시스템 성능 평가
Table 1. Performance comparison of the baseline system with or without endpoint detection.

	음성구간 검출을 사용하지 않은 경우	음성구간 검출을 사용한 경우
Error rate (%)	4.2	3.8

사용하였다.

이러한 방법은 저장공간은 크게 줄일 수 있으나 화자독립 음소 HMM을 사용한 음소 열 추정 결과에 많은 오차가 발생하는 문제점이 있다. 음소 열 추정에는 문법적인 정보를 전혀 사용하지 않기 때문에 전화선상에서 발생하는 많은 잡음으로 인하여 음성 구간의 전·후 및 잡음 구간에서 /s/와 같은 무성음 오차가 많이 발생한다. 이러한 음소의 인식 오차를 줄이는 방법으로 묵음 사이의 무성음은 묵음으로 처리하는 등의 간단한 논리를 사용하여 음소 인식 오차를 줄일 수 있다[1-4].

본 논문에서는 이러한 오차를 줄이는 방법으로 음성 구간 검출 방법을 사용하였다. 즉 에너지 파라메터를 사용한 음성 구간 검출을 수행하여 음성으로 판단된 음성 구간의 음소 열만을 입력 음성에 대한 음소 열로 저장하였다. 표 1은 음성 구간 검출을 사용하여 음소 열을 수정한 경우와 그렇지 않은 경우에 대한 성능 평가한 것이다. 표에서도 알 수 있듯이 음성 구간 검출을 사용한 경우는 잘못 인식된 음소 열을 제거하여 인식 오차가 4.2%에서 3.8%로 감소한 것을 알 수 있다.

본 논문에서는 음성구간 검출이 포함된 화자독립 음소 HMM을 사용한 화자 종속 음성 다이얼링 시스템을 기준 시스템으로 하여 제안된 확률적 매칭 방법을 통한 화자 적응을 수행한 방법과 성능을 비교 평가하였다.

3.3. 화자적응 알고리즘 성능평가

기존 화자독립 음소 HMM을 사용한 화자 종속 음성 다이얼링 시스템은 화자독립 모델을 음소 인식에 사용할 때 발생하는 오차로 인하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점이 있다. 본 논문에서는 이러한 문제점을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환함수를 동시에 추정하는 방법을 제안하였다. 위와 같은 데이터 베이스를 사용하고 화자적응 알고리즘을 사용한 음성 다이얼링 시스템의 성능은 그림 3과 같다.

그림 3에서 가로축은 학습 과정에서의 음소 열과 변환 벡터 추정과정의 반복 횟수 (iteration)를 나타낸다. 이 값이 0일 때는 음소 열만을 추정하고 변환 벡터는 추정하지 않은 기준 시스템의 인식 오차를 나타낸다. 또한 그래프의 각 선은 제안된 시스템에 사용된 변환 벡터의 수에 따른 시스템의 인식 오차를 나타낸다. 변환 벡터의 수는 음소의 형태에 따라서 1, 2, 3, 9, 14, 42, 180개의 총 7가지 경우를 사용하였다. 즉 1개인 경우는 모든 음소에 동일한 변환 벡터를 사용한 경우, 2개인 경우는 음성과 묵음에 각각 1개의 변환 벡터를 사용한 경우, 3개인 경우는 음성, 유성음과 무성음에 각각 1개의 변환 벡터를 사용한 경우, 9개인 경우는 묵음 모음, 이중모음, 반모음, 파열음 등에 각각 1개의 변환 벡터를 사용한 경우, 14개인 경우는 묵음, 전설모음, 중설모음, 후설모음 등에 각각 1개의 변환 벡터를 사용한 경우, 42개는 묵음과 모든 자음에 각각 1개의 변환 벡터를 사용한 경우이고 180개는 모든 음소 HMM의 각 상태마다 1개의 변환 벡터를 사용한 경우를 나타낸다. 그림에서 알 수 있듯이 변환 벡터가 1, 2, 3, 9의 경우에는 음소 열과 변환 벡터를 반복하여 추정하여도 인식 시스템의 성능이 기준 시스템보다 개선되지 않았다. 그러나 변환 벡터의 수를 14개 이상 사용하는 경우에는 시스템이 수렴하여 인식 오차가 감소하는 것을 알 수 있다. 여기서 변환 벡터의 수를 음소의 수와 같은 42개를 사용했을 때 가장 적은 인식 오차 (2.3%)로 수렴하는 것을 알 수 있다.

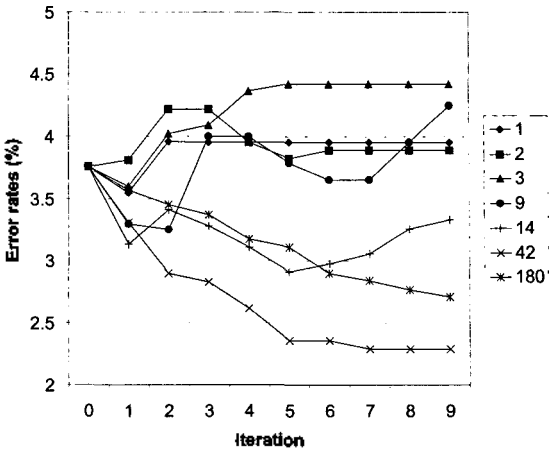

그림 3. 화자적응 알고리즘을 사용한 음성 다이얼링 시스템의 성능 평가 (변환 벡터의 개수: 1, 2, 3, 9, 14, 42, 180)
Fig. 3. Performance of the voice dialling system using speaker adaptation algorithm (number of transformation vector: 1, 2, 3, 9, 14, 42, 180).

본 실험에서는 제안된 방법의 성능을 기존의 방법과 비교하기 위하여 위에서 구현한 기준 시스템 이외에 다음과 같은 시스템을 구현하여 그 성능을 비교하였다.

A. 기준 시스템: 화자독립 음소 HMM과 음소 열을 이용한 경우

표 2. 제안된 화자적응 알고리즘과 기존 방법과의 성능비교
Table 2. Performance comparison of the proposed speaker adaptation algorithm with conventional methods.

시스템 형태	A	B	C	D
	기준 시스템	변환 벡터만 추정	변환 벡터와 음소열 동시 추정	화자종속
Error rate(%)	3.8	3.3	2.3	1.8

3. 기준 시스템에 변환 벡터만을 추정하여 화자 적응하는 경우
2. 기준 시스템에 변환 벡터와 음소 열을 동시에 추정하는 화자 적응 방법을 사용한 경우
1. 화자 종속 시스템을 사용한 경우

표 2에서 기준 시스템은 음소 HMM과 음성 구간 검출을 사용하여 얻어진 음소 열을 사용한 시스템의 성능으로 3.8%의 인식 오차를 나타내었다. 두 번째는 기준 시스템에 변환 벡터 추정을 추가한 시스템의 인식 성능을 평가하였다. 이것은 본 논문에서 제안한 음소 열과 변환 벡터를 순환적으로 추정하는 방법과 성능 비교를 하려는 것이다. 변환 벡터만을 추정한 경우에도 인식 오차는 3.3%로 감소하는 것을 알 수 있다. 다음은 제안된 방법으로 음소 열과 변환 벡터를 순환적으로 추정한 방법의 결과이다. 인식 오차는 2.3%로 기준 시스템의 인식 오차가 1.9% 감소되었다. 따라서 본 논문에서 제안한 화자적응 방법이 음소 열을 사용한 음성 다이얼링 시스템의 성능을 크게 향상시키는 것을 알 수 있다. 마지막 열은 제안된 방법과 비교를 위하여 화자종속 HMM을 사용한 단독음 인식 시스템의 성능을 나타내었다. 이 경우에 인식 성능은 1.8%로 가장 높게 나타나지만 각 단어마다 모델을 저장하여야 하기 때문에 많은 저장 공간이 필요하다.

따라서 본 논문에서 제안한 방법은 3.8%의 인식 오차를 나타내는 화자독립 HMM과 음소열을 사용하는 기존 음성 다이얼링 시스템의 성능을 확률적 매칭 방법을 사용하여 화자 적응시킨 결과, 인식 오차를 2.3%로 감소시킬 수 있었다. 이러한 것은 화자 종속 시스템의 인식 성능인 1.8%에 근접하는 것으로 제안된 방법에 의하여 화자독립 HMM과 음소열이 각 화자에 적응되어 인식 성능이 개선되었기 때문이다.

총 10명의 화자가 각 화자마다 15개의 이름을 3회 반복한 것 (15개×3회=45개/명)을 학습에 사용한 경우, 음소 열과 화자독립 HMM을 이용한 시스템은 화자마다 평균 1.5 Kbyte 의 저장공간이 필요한 반면 화자종속 HMM을 이용한 경우에는 화자마다 평균 112 Kbyte가 필요였다.

IV. 결 론

본 논문에서는 화자독립 음소모델을 사용한 음성 다이얼링 시스템의 성능을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환함수를 동시에 추정하는 방법을 제안하였다. 제안된 방법은 학습과정에서 학습 데이터의 음소 열과 화자적응을 위한 변환 벡터를 동시에 추정한 후 음소 열과 함께 저장하고, 인식 시에 화자독립 음소 HMM을 각 화자의 변환벡터를 사용하여 변환한 후 인식을 수행하였다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭을 위한 최고 유사도 방법을 이용하였으며 음소 열과 함께 반복적으로 추정되었다. 이러한 변환 벡터는 크기가 작아서 적은 저장공간을 사용하면서도 인식 성능을 화자종속 시스템에 근사하도록 향상시킬 수 있었다.

전화선을 통하여 구성된 데이터 베이스를 사용한 인식 실험에서 기존 시스템의 인식오차 3.8%가 제안된 화자적응 방법을 사용하여 2.3%로 감소하여 1.5%정도의 인식 시스템 성능이 향상되는 것을 확인하였다. 본 논문에서 사용한 데이터베이스는 유사한 내용이 많아 매우 인식하기 어려운 것으로 실제 상황에서 유사성이 적은 데이터가 사용되는 경우에는 보다 높은 인식 성능을 기대할 수 있을 것이다.

감사의 글

본 연구는 정보통신부에서 지원하는 대학기초연구지원사업으로 수행되었습니다.

참 고 문 헌

1. N. Jain, R. Cole and E. Barnard, "Creating speaker-specific phonetic templates with a speaker-independent phonetic recognizer: Implications for voice dialing," *Proceedings of ICASSP96*, 881-884, 1996.
2. V. Fontaine and H. Bourlard, "Speaker-dependent speech recognition based on phone-like units models-application to voice dialing," *Proceedings of ICASSP97*, 1527-1530,

1997.
3. B. Ramabhadran, L. R. Bahl, P. V. deSouza and M. Padmanabhan, "Acoustic-only based automatic phonetic baseform generation," *Proceedings of ICASSP98*, 2275-2278, 1998.
4. M. Shozakai, "Speech interface for car applications," *Proceedings of ICASSP99*, 1386-1389, 1999.
5. G. Zavaliagkos, R. Schwartz and J. Makhoul, Batch, "Incremental and instantaneous adaptation techniques for speech recognition," *Proceedings of ICASSP95*, 676-679, 1995.
6. A. Sankar and C. H. Lee, "A Maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, 4, 190-202, 1996.
7. R. A. Sukkar and C. H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Processing*, 4, 420-429, 1996.

저자 약력

● **김 원 구 (Weon-Goo Kim)**

1983년 3월~1987년 2월: 연세대학교 전자공학과 학사
1987년 9월~1989년 8월: 연세대학교 전자공학과 석사
1989년 9월~1994년 2월: 연세대학교 전자공학과 박사
1998년 9월~1999년 9월 Bell Lab, Lucent Technologies(USA) 객원연구원
1994년 9월~현재: 군산대학교 전자정보공학부 부교수

※ 주관심분야: 음성 신호처리, 음성인식, 음성변환, 감성인식 등

● **Chin-Hui Lee**

1986년~현재: Member of Technical Staff and Head of Dialog System Research Departmenat at Bell-Laboratories, Lucent Technologies