화자 적응을 이용한 대용량 음성 다이얼링

Large Scale Voice Dialling using Speaker Adaptation

김 원 구*
(Weon-Goo Kim)

Abstract: A new method that improves the performance of large scale voice dialling system is presented using speaker adaptation. Since SI (Speaker Independent) based speech recognition system with phoneme HMM uses only the phoneme string of the input sentence, the storage space could be reduced greatly. However, the performance of the system is worse than that of the speaker dependent system due to the mismatch between the input utterance and the SI models. A new method that estimates the phonetic string and adaptation vectors iteratively is presented to reduce the mismatch between the training utterances and a set of SI models using speaker adaptation techniques. For speaker adaptation the stochastic matching methods are used to estimate the adaptation vectors. The experiments performed over actual telephone line shows that proposed method shows better performance as compared to the conventional method. with the SI phonetic recognizer.

Keywords: speech recognition, speaker adaptation, phoneme string, phoneme HMM

I. 서론

일반적으로 음성 다이얼일 시스템은 화자 종속형의 시스템과 화자 독립형 시스템으로 나눌 수 있다. 화자 종속 시스템은 그 구조가 간단하고 화자종속의 형태를 갖기 때문에 인식 성능이 비교적 우수하지만 단어나 문장 단위로 모델을 저장해야 하기 때문에 많은 저장 공간이 필요하고 대상 단어수의 증가에 비례하여 필요한 저장도 증가하게 된다. 이러한 문제점은 핸드폰과 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 음성 인식을 수행하기 위한 데이터 저장 공간의 크기가 매우 중요한 문제가 된다.

이러한 문제를 해결하기 위한 방법 중의 하나로 화자독립 음소모델을 이용한 방법들이 제안되었다[1-4]. 이러한 방법들은 화자독립 음소모델을 사용하여 학습 데이터의 음소열을 구하여 저장하고, 입력 음성을 인식할 때 저장된 음소열과 화자독립 모델을 사용하는 것이다. 이러한 방법들의 장점은 저장 공간은 크게 줄일 수 있다는 것이다. 그러나 이러한 방법은 화자독립 음소 HMM을 사용하여 음소 인식을 수행할 때 많은 오차가 발생하는 것과 화자독립 모델을 음소 인식에 사용하기 입력 음성과 오차가 발생하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점이 있다.

본 논문에서는 대용량 음성 다이얼링 시스템의 성능을 개선하기 위하여 화자 독립 음소 모델과 입력 음성과의 불일치를 감소시키는 화자 적응 기법을 사용하여 인식 성능을 향상시키는 방법을 제안하였다. 사용된 화자 적응 방법은 학습 데이터를 사용하여 음소 열과 모델 변환 함수를 동시에 추정하는 방법을 제안하였다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭(stochastic matching) 방법을 이용하였으며 음소 열과 함께 반복적으로 추정되었다[5,6]. 이러한 변환 벡터는 크기가 작아서 작은 저장 공간을 사용하면서도 인식 성능을 화자종속 시스템에 근사하도록 향상시킬 수 있었다.

* 책임저자(Corresponding Author)
논문접수: 2010. 1. 10., 수정: 2010. 1. 30., 채택확정: 2010. 2. 7.
김원구: 군산대학교 전기공학과(wgkim@kunsan.ac.kr)
※ 본 논문은 2007년도 산학협동재단 학술연구비 지원에 의하여 이루어졌음.
※ 상기 논문은 제어 · 로봇 · 시스템학회 전북제주지부의 학술대회에서 초안이 발표되었습니다.

II. 화자 적응 알고리즘

본 논문에서는 대용량 음성 다이얼링 시스템의 성능을 개선하기 위하여 음소 열과 화자적응을 위한 화자독립 음소 모델의 변환함수를 동시에 추정하는 방법을 제안하였다. 제안된 시스템의 구조는 그림 1과 같다.

제안된 방법은 학습과정에서 학습 데이터와 화자독립 음소 HMM을 사용하여 학습 데이터의 음소 열과 화자적응을 위한 변환 벡터를 동시에 추정한 후 음소 열과 함께 저장하고, 인식 단계에서 화자독립 음소 HMM을 각 화자의 변환 벡터를 사용하여 변환한 후 입력 음성에 대한 인식을

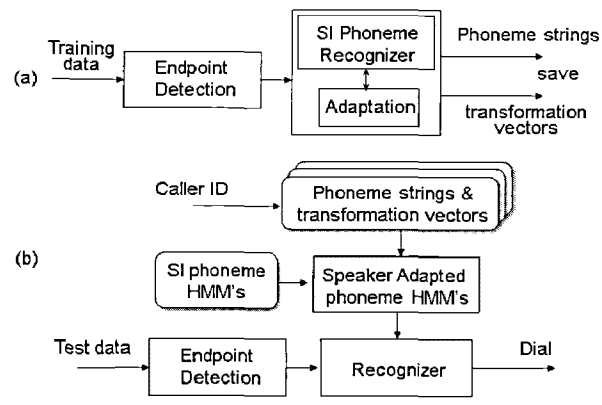


그림 1. 제안된 시스템 (a) 학습 (b)인식.
Fig. 1. The proposed system (a) training (b) recognition.

수행한다. 따라서 음성 다이얼링을 위해서는 화자의 신분을 먼저 확인한다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭 방법을 이용하였으며 음소 열과 함께 반복적으로 추정되었다.

화자 적응을 위한 확률적 매칭 방법은 다음과 같이 적용될 수 있다. 우선 일련의 특징 벡터 $Y = y_1, y_2, \cdots, y_T$ 와 화자독립 음소 HMM의 집합을 Λ_X 라고 할 때, 화자 적응되어 변형된 모델 Λ_Y 를 위한 모델 공간 변환은 $\Lambda_Y = G_\eta(\Lambda_X)$ 로 이루어진다. 여기서 $G_\eta(.)$ 는 모델 변환 함수이고 η 는 변환 파라메터이다. Y와 Λ_X 사이의 불일치를 줄이는 방법 중의 하나로 변환 파라메터 η 와 음소열 W의 결합 유사도를 최대화하는 η 와 W를 찾는 것이고 다음과 같이 정의된다.

$$\begin{aligned}(\eta', W') &= \arg\max_{(\eta, W)} p(Y, W | \eta, \Lambda_X) \\ &= \arg\max_{(\eta, W)} p(Y | W, \eta, \Lambda_X) P(W)\end{aligned} \tag{1}$$

식 (1)에서 변수 η 와 W 에 대한 결합 최대화는 먼저 η 를 고정시키고 W에 대하여 최대화하고, 그 다음에 W를 고정시키고 η 에 대하여 최대화하는 반복적인 방법으로 구현될 수 있다. 본 논문에서는 음소 군과 코드북 기반의 두 가지 종류의 확률적 매칭 방법을 사용하여 변환 벡터를 추정하였다. 음소 군 기반 확률적 매칭 방법에서는 음소 군의 개수에 비례하여 변환 벡터가 사용되었다. 코드북 기반의 확률적 매칭 방법에서는 모델 파라메터를 공통으로 사용하는 묶음의 개념을 사용하여 변환벡터의 개수를 결정하였다. 묶음의 정도에 따라 코드북의 크기와 그에 따른 변환벡터의 수가 조절되어진다. 코드북은 유클리디안(euclidean) 거리 측정 방법을 사용하는 Lloyd 알고리즘을 사용하여 화자독립 HMM의 모든 혼합 성분들의 평균 벡터를 군집화하여 구성되었다.

$\Phi_1, \cdots, \Phi_K$ 를 K개의 군집으로 가정하면, 음소군 또는 코드북 기반 확률적 매칭에서 K는 음소 군 또는 코드 워드의 개수가 된다. 변환 벡터 $\mu_{b_k} = \mu_{k_1}, \mu_{k_2}, \cdots, \mu_{K_D}$ 는 최대 유사도 추정(maximum likelihood estimation)에 의하여 다음과 같이 구해질 수 있다.

$$\mu'_{b_{k,i}} = \frac{\sum_{i=1}^{T} \sum_{(n,m)\in\Omega_k} \gamma_t(n,m) \frac{y_{t,i} - \mu_{n,m,i}}{\sigma^2_{n,m,i}}}{\sum_{i=1}^{T} \sum_{(n,m)\in\Omega_k} \frac{\gamma_t(n,m)}{\sigma^2_{n,m,i}}}, \tag{2}$$

$$i = 1, \cdots, D, \quad k = 1, \cdots, K$$

$$\gamma_t(n,m) = \begin{cases} \dfrac{w_{n,m}\, \mathcal{N}[y_t; \mu_{n,m}, C_{n,m}]}{\sum_{j=1}^{M} w_{n,j}\, \mathcal{N}[y_t; \mu_{n,j}, C_{n,j}]}, & if\ \hat{s} = n \\ 0, & otherwise \end{cases} \tag{3}$$

여기서 D는 특징 벡터의 차수이고 $\mu_{n,m}$, $C_{n,m}$은 상태 n에서 m번째 혼합성분에 해당하는 평균과 분산 벡터이고, $w_{n,m}$은 상태 n에서 m번째 복합성분의 확률이고, N은 정규분포(normal distribution)이고 s는 입력 음성에 해당하는 상태 열이다. 만일 $\mu_{n,m}$이 k번째 군 또는 코드워드에 군집되면 $(n,m) \in \Phi_k$ 이다.

변환 벡터에 의하여 변환된 화자독립 HMM에서 각각의 혼합성분의 평균 $\hat{\mu}_{n,m}$은 화자독립 HMM의 혼합성분의 평균 $\mu_{n,m}$에 변환 벡터를 더하여 다음과 같이 구해진다.

$$\hat{\mu}_{n,m} = \mu_{n,m} + \sum_{k=1}^{K} \mu_{b_k} \mathrm{I}_{\Phi_k}(n,m), \quad \forall (n,m) \tag{4}$$

여기서 $\mathrm{I}_{\Phi_k}(\cdot)$ 는 집합 Φ_k를 나타내는 기호이다.

제안된 방법의 학습 과정에서 네 단계로 구현된다.

(1) 화자독립 음소 HMM을 이용하여 다수의 학습 음성에 대하여 초기 음소 열과 상태 분할 정보를 추정한다.
(2) 식 (2)와 (3)을 이용하여 변환 벡터를 추정된다.
(3) 식 (4)를 사용하여 화자독립 음소 모델들을 변환한다.
(4) 모델이 수렴될 때까지 단계 (2), (3)을 반복하고 최종 변환 벡터와 음소 열을 인식 과정을 위하여 저장한다.

인식 과정에서 사용자는 먼저 신분을 확인한다. 전화망에서는 발신자 번호 또는 사용자의 입력에 의하여 신원이 확인된다. 화자독립 음소 HMM이 해당 사용자의 변환 벡터를 사용하여 변환된 후에 인식기는 음소 열과 변환된 화자독립 음소 HMM을 사용하여 입력 음성을 인식한다.

III. 실험 및 결과

1. 음성 다이얼링 시스템 구성

실험에 사용된 데이터 베이스는 남성 5명과 여성 5명의 총 10명으로 구성하였다[6]. 각 화자는 15개의 단어를 발음하였다. 음성 신호는 6.67kHz로 샘플링되었고 8bit 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 녹음은 전화선을 통하여 이루어 졌으며 각 화자는 각기 다른 환경에서 가급적 다른 종류의 전화기를 사용하여 몇 주 간격을 두고 녹음하였다.

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 30ms의 창 길이를 갖고 10ms씩 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 left-to-right 형태의 음소 모델 41개와 1개의 상태를 갖는 묵음 모델로 구성되었고 각각의 HMM은 연속밀도분포를 갖는 연속분포 HMM이다. 입력 음성에 대한 음소열은 이러한 모델을 사용하여 인식하였다.

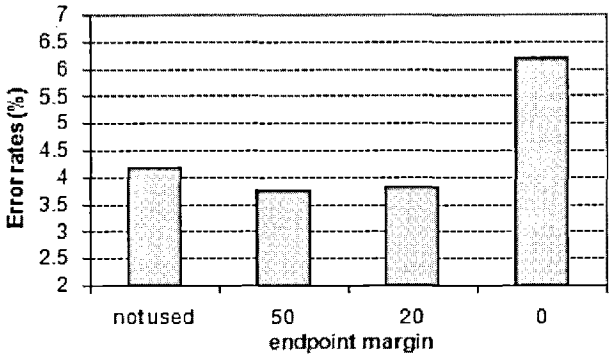


그림 2. 음성 구간 검출을 이용한 인식 성능.

Fig. 2. Performance of the speech recognition system using endpoint detection.

2. 기준 시스템의 성능 향상

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 사용하였다. 이러한 방법은 저장 공간은 크게 줄일 수 있으나 화자독립 음소 HMM을 사용한 음소 열 추정 결과에 많은 오차가 발생하는 문제점이 있다.

본 논문에서는 이러한 오차를 줄이는 방법으로 음성 구간 검출 방법을 사용하였고 음소 열 인식과정에서 음소의 수를 가능하면 적게 만들도록 인식 시스템을 설정하였다. 즉 에너지 파라메터를 사용한 음성 구간 검출을 수행하여 음성으로 판단된 음성 구간의 음소 열만을 입력 음성에 대한 음소 열로 저장하였다. 이때 음성 구간 검출의 오차를 고려하여 검출된 구간보다 조금 크게 음성구간을 설정한 경우에 더 우수한 성능을 나타내었다(그림 2). 이러한 방법은 음성 앞뒤에 발생하는 오인식된 음소를 제거하는데 효과적이다. 또한 음소 인식기가 인식한 음소의 수가 적은 것이 많은 것보다 우수한 성능을 나타내었다. 이러한 방법을 사용한 경우 잘못 인식된 음소 열을 제거하여 인식 오차가 4.2%에서 3.8%로 감소하였다.

3. 제안된 시스템의 성능 평가

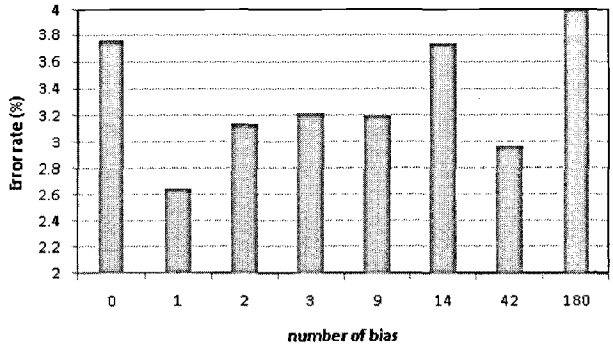
본 논문에서는 기존 화자독립 음소 HMM을 사용한 화자 종속 음성 다이얼링 시스템은 화자독립 모델을 음소 인식에 사용할 때 발생하는 오차로 인하여 화자종속 모델을 사용하는 방법보다는 인식 성능이 저하되는 문제점을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환함수를 동시에 추정하는 방법을 제안하였다. 여기에서 화자 적응을 위해서 음소 군과 코드북 기반의 두 가지 방법의 확률적 매칭 방법을 사용하였다.

위와 같은 데이터 베이스를 사용하고 화자적응 알고리즘을 사용한 음성 다이얼링 시스템의 성능은 그림 3과 같다. 그림 3(a)에서 가로축은 변환 벡터의 수를 나타낸다. 변환 벡터의 수는 음소의 형태에 따라서 1, 2, 3, 9, 14, 42, 180 개 의 총 7가지 경우를 사용하였다(표 1). 그림에서 알 수

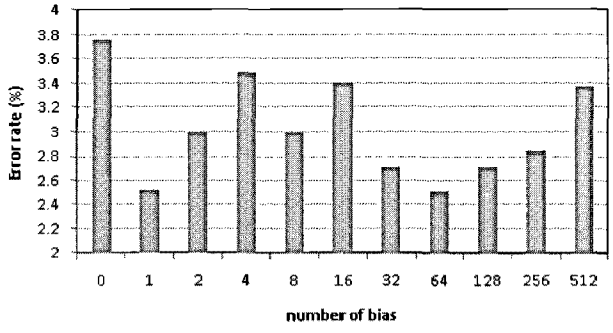
표 1. 음소군 기반 확률적 매칭 방법에서 음소 군에 따른 변환 벡터의 개수.

Table 1. The number of transformation vectors according to the phonetic class based stochastic matching.

변환 벡터 개수	음소 군
1	no phonetic class
2	silence and speech
3	silence, vowel and consonants
9	silence, vowel, diphthongs, semivowels, stops,...
14	silence, front, mid, back} vowel, diphthongs, liquids, glides, {voiced, unvoiced} stops,...
42	silence and all phones
180	all SI HMM states to the acoustic phonetic class



(a)



(b)

그림 3. 음소군(a)와 코드북(b)에 따른 제안된 학습 알고리즘을 사용한 음성 인식 시스템의 성능 평가.

Fig. 3. Performance of the speech recognition system using proposed algorithm according to the phonetic class(a) and codebook(b).

있듯이 변환 벡터의 개수가 42개까지는 음소 열과 변환 벡터를 반복하여 추정하면 음성 인식 시스템의 성능이 기준 시스템보다 개선되었다. 전체적으로 변환 벡터의 개수가 1개일 때 가장 우수한 성능을 나타내고 변환 벡터의 개수가 증가하면 성능 개선 정도가 작아진다. 여기서 변환 벡터의 개수를 1개로 하였을 때 2.7%의 가장 작은 인식 오차로 수렴하였다.

그림 3(b)에서는 변환 벡터를 모든 화자독립 음소 HMM의 평균 벡터를 집단화하여 적용하였다. 이때 사용된 코드북의 개수는 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 개의 총

표 2. 제안된 학습 알고리즘과 기존 방법의 성능 비교.

Table 2. Performance comparison of the proposed training algorithm with convention methods.

시스템 형태	인식 오차(%)
SI 음소 인식기	4.2
SI 음소 인식기와 음성구간 검출	3.8
SI 음소 인식기와 변환벡터 추정	3.3
제안된 시스템	2.5
제안된 시스템에 음소 열을 알 경우	2.3
화자 종속 인식기	1.8

10가지 경우를 사용하였다. 그림에서 알 수 있듯이 음소 열과 변환 벡터를 반복하여 추정하면 인식 시스템의 성능이 기준 시스템보다 개선되는 것을 알 수 있다. 여기서도 변환 벡터의 개수가 1개일 때 가장 우수한 2.5%의 인식 오차를 나타내었다. 그러나 변환 벡터의 개수를 증가시켜도 크게 개선되지는 않았다. 이것은 음소 별로 변환 벡터를 사용한 이전 실험의 경우의 최소 오차보다 약간 우수한 결과를 나타낸다.

본 실험에서는 제안된 방법의 성능을 기존의 방법과 비교하기 위하여 위에서 구현한 기준 시스템 이외에 다음과 같은 시스템을 구현하여 그 성능을 비교하였다. 표 2에서 기준 시스템은 음소 HMM과 음성 구간 검출 등을 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 2.5%로 기준 시스템의 인식 오차가 1.7% 감소되었다. 음소 열을 알 경우는 입력 음성의 음소 열을 알고 있다고 가정하고 시스템을 학습한 경우이다. 결과를 보면 음소 열을 알 경우와 제안된 시스템의 성능이 같다. 이러한 것은 제안된 시스템이 음소열과 변환 벡터를 잘 추정하여 수렴했다는 것을 나타낸다. 마지막 열은 화자종속 HMM을 사용한 단독음 인식 시스템의 성능을 나타내었다.

VI. 결론

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 매칭 방법을 이용하였으며 음소 열과 함께 반복적으로 추정되었다. 본 논문에서는 음소 군과 코드북 기반의 두 가지 종류의 확률적 매칭 방법을 사용하여 변환 벡터를 추정하였다.

전화선을 통하여 구성된 데이터 베이스를 사용한 인식 실험에서 기존 시스템의 인식오차 4.2%가 제안된 화자적응 방법을 사용하여 2.5%로 감소하는 것을 확인하였다.

참고문헌

[1] N. Jain, R. Cole, and E. Barnard, "Creating speaker specific phonetic templates with a speaker-independent phonetic recognizer: implications for voice dialing," *Proc. of ICASSP*, pp. 881-884, 1996.

[2] V. Fontaine and H. Bourlard, "Speaker-dependent speech recognition based on phone-like units models-application to voice dialing," *Proc. of ICASSP*, pp. 1527-1530, 1997.

[3] B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, "Acoustic-only based automatic phonetic baseform generation," *Proc. of ICASSP*, pp. 2275-2278, 1998.

[4] S. Deligne and L. Mangu, "On the use of lattices for automatic generation of pronunciations," *Proc. of ICASSP*, pp. 204-207, 2003.

[5] A. Sankar and C. H, Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 190-202, 1996.

[6] R. A. Sukkar and C. H. Lee, "Vocabulary Independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, pp. 420-429, 1996.

김 원 구

1987년 연세대 전자공학과 졸업. 1989년 동 대학원 석사. 1994년 동 대학 박사. 1994년~현재 군산대학교 전기공학과 교수. 1998년~1999년 Bell lab, Lucent Technologies (USA) 객원 연구원. 2008년~2009년 호주 Griffith 대학 교환교수. 관심분야는 음성신호처리, 음성인식, 음성변환, 감정인식, 화자인식.