확률적 매칭을 사용한 음성 다이얼링 시스템

Voice Dialing system using Stochastic Matching

김 원 구
군산대학교 전자정보공학부

Weon-Goo Kim
School of Electronic and Information Eng., Kunsan National University
E-mail : wgkim@kunsan.ac.kr

요 약

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.

1. 서론

일반적으로 음성 다이얼일 시스템은 화자 종속형의 시스템을 사용하여 각 화자가 자동적으로 전화를 걸 때 사용할 명령이나 키워드를 포함하는 개인적인 목록을 사용한다. 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 방법중의 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 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. 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. 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.

2. 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 저장해야할 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 1과 같다.

제안된 방법은 등록 단계인 학습과정에서 학습 데이터와 화자독립 음소 HMM을 사용하여 학습 데이터의 음소 열과 화자적응을 위한 변환 벡터(bias)를 동시에 추정한 후 음소 열과 함께 저장하고, 인식 단계에서 화자독립 음소 HMM을 각 화자의 변환벡터를 사용하여 변환한 후 입력 음성에 대한 인식을 수행한다. 여기서 화자적응을 위한 변환 벡터는 확률적 매칭을 위한 최고 유사도 방법[5,6]을 이용하였으며 음소 열과 함께 반복적으로 추정되었다.

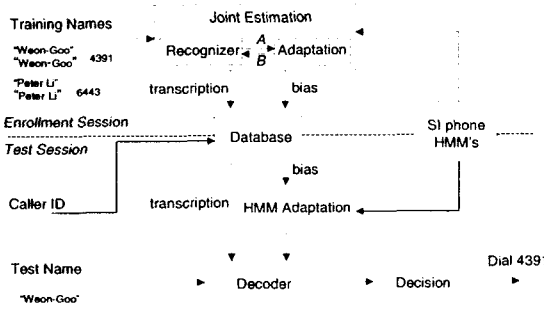

그림 1. 화자적응과 HMM을 이용한 개인용 음성 다이얼링 시스템의 개념도

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.

3. 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 6.67kHz로 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 날짜

에 수행한 5회의 녹음에서 각 화자가 15개의 이름을 10회 반복한 데이터(15개×10회=150개/명)로 구성하였다. 데이터 내용은 영어로 "Call office", "Call home", 등으로 구성되었다.

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. 이러한 모델을 사용하여 입력 음성에 대한 음소 열을 추정하였다.

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 사용하였다. 이러한 방법은 저장공간은 크게 줄일 수 있으나 화자독립 음소 HMM을 사용한 음소 열 추정 결과에 많은 오차가 발생하는 문제점이 있다. 이러한 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. 실험 결과, 음성 구간 검출을 사용한 경우는 잘못 인식된 음소 열을 제거하여 인식 오차가 4.2%에서 3.8%로 감소하였다.

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 2와 같다.

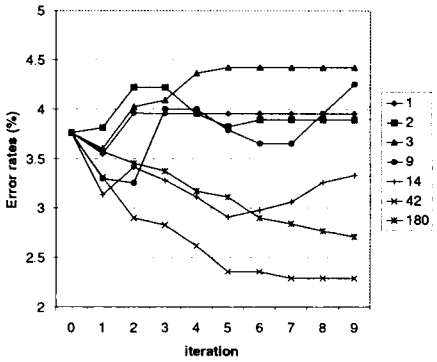

그림 2. 화자적응 알고리즘을 사용한 음성 다이얼링 시스템의 성능 평가

그림 2에서 가로축은 학습 과정에서의 음소 열과 변환벡터 추정과정의 반복 횟수(iteration)를 나타낸다. 변환 벡터의 수는 음소의 형태에 따라서 1, 2, 3, 9, 14, 42, 180 개 의 총 7가지 경우를 사용하였다. 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.

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
B. 기준 시스템에 변환 벡터만을 추정하여 화자적응하는 경우

C. 기준 시스템에 변환 벡터와 음소 열을 동시에 추정하는 화자 적응 방법을 사용한 경우
D. 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. 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.

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 1.5Kbyte 의 저장공간이 필요한 반면 화자종속 HMM을 이용한 경우에는 화자마다 평균 112Kbyte가 필요하였다.

표. 2 제안된 화자적응 시스템과 기준 시스템과의 성능비교

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

4. 결론

본 논문에서는 화자독립 음소모델을 사용한 음성 다이얼링 시스템의 성능을 개선하기 위하여 음소 열과 화자적응을 위한 모델 변환함수를 동시에 추정하는 방법을 제안하였다. 제안된 방법은 학습 과정에서 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.

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.

참고문헌

[1] N. Jain, R. Cole and E. Barnard, "Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing", in Proceedings of ICASSP96, pp. 881-884, 1996

[2] V. Fontaine and H. Bourlard, "Speaker-Dependent Speech Recognition Based on Phone-Like Units Models-Application to Voice Dialing", in Proceedings of ICASSP97, pp. 1527-1530, 1997

[3] B. Ramabhadran, L. R. Bahl, P. V. deSouza and M. Padmanabhan, "Acoustic-Only Based Automatic Phonetic Baseform Generation", in Proceedings of ICASSP98, pp. 2275-2278, 1998

[4] M. Shozakai, "Speech Interface for Car Applications", in Proceedings of ICASSP99, pp. 1386-1389, 1999

[5] G. Zavaliagkos, R. Schwartz and J. Makhoul, Batch, "Incremental and Instantaneous Adaptation Techniques for Speech Recognition", in Proceedings of ICASSP95, pp.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, vol. 4, pp. 190-202, May, 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*, Vol. 4, pp. 420-429, Nov. 1996