Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

음성 인식을 위한 후처리에 관한 연구

A Study on the Post-processing for Speech Recognition

김 원 구
전북 군산시 군산대학교 전자정보공학부
E-mail: wgkim@kunsan.ac.kr

요 약

음성 다이얼링 시스템은 화자의 음성을 인식하여 원하는 전화번호로 자동으로 전화를 걸어주는 시스템으로 주로 이동 전화나 휴대형 통신 장비에 유용하게 사용된다. 개인 음성 다이얼링 시스템의 경우, 다이얼링에 사용되는 모든 구문은 사용자가 선택하고 사용자의 음성을 사용하여 학습되어 음성 인식을 위한 HMM을 생성한다. 이러한 시스템은 화자독립 시스템보다 매우 적은 메모리 공간과 계산량으로 구현이 가능하다. 그러나 이러한 시스템은 학습시 각 단어당 2-3개의 음성만을 사용하므로 음성인식 시스템의 성능을 개선하기 위한 각 상태에서의 상태지속분포을 추정하기는 매우 어렵다. 따라서 본 논문에서는 성능개선을 위한 후처리기를 제안하였다. 전화선을 통하여 구성된 데이터베이스를 이용한 실험에서 제안된 후처리기가 인식 시스템의 성능을 향상시킴을 확인하였다.

Key Words : 음성인식, 음성 다이얼링, 후처리기

1. 서 론

음성 다이얼링 시스템은 자동차 운전중, 업무중 또는 이동중과 같이 사용자가 눈과 손을 사용하고 있는 상황에서 책에서 전화번호를 찾아 손으로 전화를 거는 방법대신에 이름을 말하여 전화를 거는 방법으로 매우 편리하게 사용되어 지고 있다[1,2]. 이러한 시스템은 주로 사용자가 미리 각 전화번호에 해당되는 음성을 패턴 또는 모델로 저장하여 두었다가 입력되는 음성을 인식하는데 사용하는 화자종속 시스템의 구조를 갖는다. 이러한 구조는 저장하는 데이터량에 비례하여 메모리 사용량이 증가하고 사용자가 직접 학습시켜야 한다는 단점에도 불구하고 시스템의 구성이 간단하고 입력 음성의 형태에 제한이 없으며 인식 성능도 비교적 좋아 화자독립 형태의 음성 다이얼링 시스템에 비하여 많이 사용되고 있다.

HMM(Hidden Markov Model)[3] 을 이용한 음성 다이얼링 시스템의 구조는 학습과 인식의 두 단계로 이루어 진다. 학습단계(enrollment session)에서는 음성명령과 그것과 연결된 전화번호가 입력된 후 HMM 이 생성되어 저장된다. 이때 사용자는 입력 음성을 두세번 반복한다. 인식단계(test session)에서는 사용자는 저장된 음성중의 하나를 발음하면 음성 다이얼링 시스템이 입력 음성을 인식하여 그에 해당하는 전화번호로 전화를 건다. 이때 이동전화나 휴대용 통신장비인 경우에는 사용자 확인이 필요하지 않으나 일반전화인 경우에는 발신번호를 확인이나 사용자가 직접 입력한다.

이러한 시스템은 화자독립 시스템보다 매우 적은 메모리 공간과 계산량으로 구현이 가능하다. 그러나 이러한 시스템은 학습시 각 단어당 2-3개의 음성만을 사용하므로 HMM 파라메터를 충분히 학습시키기가 어렵다. 또한 음성인식 시스템의 성능을 개선하기 위한 각 상태에서의 상태지속분포을 추정하기는 더욱 어렵다. 따라서 본 논문에서는 성능개선을 위한 후처리기(post-processor)를 제안하였다. 이러한 후처리기는 입력 음성과 HMM 의 왜곡된 매칭을 감소시킴으로서 인식성능을 개선할 수 있을 뿐만 아니라 계산량도 매우 적다. 전화선을 통하여 수집된 음성 데이터베이스를 이용한 실험에서 제안된 후처리기가 인식 시스템의 성능을 향상시킴을 확인하였다.

2. 음성 인식 시스템

본 논문에서는 일반 전화선을 통한 개인용 음성 다이얼링 시스템 구현에 관하여 연구하였다. 전화선 환경은 여러가지 잡음과 채널의 영향이 많이 발생한다. 따라서 이에 대한 처리는 전처리 과정에서 수행하였다. 음성 다이얼링

시스템은 입력 단어의 수에 따라 메모리 사용량이 비례하여 증가하므로 메모리를 적게 사용하는 구조이어야 하며 학습과 인식에 필요한 계산량도 적어야 한다. 또한 학습시 단어당 2-3개의 음성만을 사용하므로 HMM 파라메터의 추정이 어려워 인식 오차가 발생하는 문제를 해결해야 한다. 본 논문에서는 후처리기를 제안하여 이러한 문제를 감소시켰다.

2.1 전처리 과정

일반 전화선 환경에서는 채널의 영향뿐 아니라 사용자 환경에서 발생하는 여러 가지 잡음에 음성이 첨가된다. 따라서 이러한 영향을 최소화 할 수 있는 전처리 과정의 선택이 필요하다. 전처리과정에서 음성 신호로부터 LPC 켑스트럼 계수를 추출하고 매치 필터(match filter)를 사용한 음성구간 검출 알고리즘[4]을 사용하여 음성 구간을 검출한 후 켑스트럽 평균 차감법(cepstrum mean subtraction: CMS)[5,6]을 사용하여 채널에 포함된 바이어스(bias)를 제거하였다.

2.2 음성 다이얼링 시스템 구성

학습과정에서는 각 명령어 또는 이름을 2-3회 정도 발음한 후에 각 이름당 한 개의 HMM을 segmental K-means 알고리즘[7]을 사용하여 구성한다. 이 모델은 이와 관련된 전화번호와 함께 사용자 정보로 저장된다. 인식과정은 입력 음성 O에 대하여 프레임당 평균 대수 유사도(log-likelihood score) L_i를 각 모델에 대하여 다음과 같이 구한다.

$$L_i(O,\lambda_i) = \frac{1}{N} logP(O|\lambda_i) \quad (1)$$

여기서 N은 입력 음성의 총 특징 벡터수이고 $P(O|\lambda_i)$는 i번째 모델과 입력 음성간의 유사도 합이다. 모든 모델에 대한 유사도를 비교한 후, k번째 이름은 다음과 같이 선택된다.

$$k = \arg\max_{1 \le i \le N} L_i(O,\lambda_i) \quad (2)$$

이때 입력음성이 저장된 단어에 없는 경우를 위하여 선택된 유사도는 문턱치 Θ와 비교하여 승낙과 거절을 결정한다.

$$\begin{cases} \text{승낙: } L_k \ge \theta, \\ \text{거절: } L_k < \theta \end{cases} \quad (3)$$

2.3 후처리기

위와 같은 음성 다이얼링 시스템에서 발생하는 오인식의 원인중의 한가지는 모델과 입력 데이터의 왜곡된 매칭에 기인한다. 이러한 왜곡된 매칭을 감소시켜 인식 성능을 향상시키는 방법중의 한가지가 상태지속분포를 이용하는 것이다. 그러나 화자 종속형 음성 다이얼링 시스템의 경우에는 각 단어당 두세번 정도 발성하기 때문에 이러한 분포를 추정하는 것이 매우 어렵다. 따라서 본 논문에서는 왜곡된 매칭을 감소시켜 성능개선을 하기 위하여 다음과 같은 후처리기를 제안하였다.

$$\hat{L}_i(O,\lambda_i) = \frac{1}{N_i} logP(O|\lambda_i) + \alpha \frac{1}{S}\sum_{j=1}^{S} \frac{1}{T_j} logP(O_j|\lambda_i) \quad (4)$$

여기서 a는 가중함수이고 S는 총 상태수이다. O_j 는 입력 특징벡터열 O가 비터비 디코딩을 통하여 분할되어 상태 j에 해당되는 특징벡터 열이고, T_j 는 O_j 의 프fp임 수이다. (4)의 맨 오른쪽 항은 후처리 가중함수로서 각 상태에서의 유사도를 상태수로 정규화하여 왜곡된 매칭이 클수록 유사도가 작아지는 특성을 갖도록 하였다. 이러한 후처리기는 입력 음성과 HMM 사이의 왜곡된 매칭을 감소시킴으로서 인식성능을 개선할 수 있을 뿐만 아니라 유사도 계산을 위한 계산량도 매우 적다.

3. 실험 및 결과 고찰

3.1 인식 시스템 구성 및 데이터베이스

실험에 사용된 특징벡터는 12차 켑스트럼과 12차 차분 켑스트럼으로 구성되었다. 켑스트럼 계수는 30ms의 창길이를 갖고 10ms씩 이동하면서 구한 10차 LPC 계수로부터 구하였다.

실험에 사용된 데이터베이스는 남성 18명과 여성 20명의 총 38명으로 구성하였다[8]. 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. 학습에 사용된 데이터는 각 화자가 15개의 이름을 3회 반복한 것으로 구성하였으며, 인식에 사용된 데이터는 각기 다른 날짜에 수행한 5회의 녹음에서 각 화자가 15개의 이름을 10회 반복한 데이터로 구성하였다. 데이터는 8bit μ-law PCM으로 저장되었다. 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.

각 이름에 대한 HMM은 left-to-right 형태의 모델을 사용하였으며 상태수는 학습데이터의 길이에 비례하도록 10 프레임당 1개의 상태를 갖게 설정하였다. HMM은 연속밀도분포를 사용하는 연속분포 HMM을 사용하였고 각 상태마다 4개의 가우시안 분포를 사용하였다. 또한 학습에 사용되는 데이터의 양이 적으므로 각 상태에서 가우시안 분포의 분산 추정이 어려우므로 공통 분산(global variance)을 모든 가우시안 분포에 사용하였다.

3.2 후처리기 성능평가

학습과정에는 각 화자당 15개의 이름을 3회 반복한 음성이 사용되었고 인식과정에는 각 화자마다 15개의 단어를 5회에 걸쳐 10회 반복(15×10)한 음성이 사용되었다. 이와 같은 데이터 베이스를 사용한 화자종속 음성 다이얼링 시스템의 성능은 표 1과 같다.

표. 1. 화자종속 음성다이얼링 시스템의 성능

인식오차(%)	L	$\hat{L}$
남성(18명)	3.0	2.5
여성(20명)	3.5	3.0
평균	3.3	2.8

표 1에서 두 번째 칸은 유사도 (1)을 사용한 인식 오차이다. 남성과 여성의 인식 오차는 각각 3.0%와 3.5%이고 평균 오차는 3.3%이다. 세 번째 칸은 상태 지속에 가중을 두는 후처리기를 사용한 유사도 (2)에 대한 결과이다. 남성과 여성의 인식 오차는 각각 2.5%와 3.0%이고 평균 오차는 2.8%이다. 즉 후처리기를 사용한 경우에 약 15%정도의 인식 오차가 감소되었다. 후처리기를 사용한 유사도 (2)의 가중파라메터 α와 인식 오차의 관계는 α값이 0인 경우는 기존 유사도인 (1) 과 같고 α 값이 증가함에 따라 인식 오차가 감소하여 α값이 1근처일 때 가장 우수한 성능을 나타내었다.

4. 결 론

본 논문에서는 음성 다이얼링 시스템의 성능 개선을 위한 후처리기를 제안하였다. HMM을 사용하는 화자종속형 음성인식 시스템인 경우, 학습시 각 단어당 2-3개의 음성만을 사용하므로 음성인식 시스템의 성능을 개선하기 위한 상태지속분포를 추정하는 것은 매우 어렵다. 제안된 후처리기는 입력 음성과 HMM사이의 왜곡된 매칭을 감소시킴으로서 인식성능을 개선할 수 있을 뿐만 아니라 계산량도 매우 적다.

전화선을 통하여 구성된 데이터베이스를 이용한 실험에서 기존 시스템의 인식 오차 3.3%가 제안된 후처리기를 사용하여 2.8%로 감소하여 15% 정도 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

이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(R01-2007-000-20989-0(2007))

참 고 문 헌

[1] C. S. Ramalingam, Y. Gong, L. P. Netsch, W. W. Anderson, J. J. Godfrey, and Y. H. Kao, "Sspeaker-Dependent Name Dialing in a Car Environment With Out-Of- Vocabulary Rejection", *in Proceedings of ICASSP'99* (Phoenix, USA), vol. 3, pp. 1780-1783, May, 1999

[2] D. van Compernolle, "Speech Recognition in the Car: From Phone Dialing to Car Navigation," *in Proceedings of EUROSPEECH'97* (Rhodes, Greece), vol. 5, pp. 2431-2434, Sept. 1997

[3] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, NJ, Prentice-Hall, 1993

[4] Q. Li and A. Tsai, "A matched filter approach to endpoint detection for robust speaker verification," in *Workshop of Automatic Identification*(Summit, NJ, USA) Oct. 1999

[5] A. E. Rosenberg, C. H. Lee, F. K Soong, "Cepstral Channel Normalization Techniques for HMM Based Speaker Verification," *in Proceedings of ICSLP'94*, pp. 1835-1838, 1994

[6] M. G. Rahim, B. H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 4, No. 1, pp. 19-30, 1996

[7] L. R. Rabiner, J. G. Wilpon and B. H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Technical Journal*, Vol. 65, pp.21-31, May/June, 1986

[8] 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