DTW 와 HMM의 상태별 파라미터 가중 기법을 이용한 문맥 종속형 화자인식

이 철 희, 정 성 환, 김 종 교
전북대학교 전자정보공학부
전화 : (0652) 272-1177/ 팩스 : (0652) 270-2400

Text-Dependent Speaker Recognition Using DTW and State-Dependent Parameter Weighting Method of HMM

Chul-Hee Lee, Sung-Hwan Chung, Chong-Kyo Kim
Division of Electronics and Information Eng., Chonbuk National University
E-mail : kissmeda@shinbiro.com

Abstract

In this paper, the speaker-recognition process based on both DTW and discrete HMM was performed using the method to evaluate state-dependent parameter weighting from training data so as the personal audio-characteristics are to be well reflected. In the suggested method below, we found the optimal state sequence using the Viterbi algorithm. The optimal path could be evaluated after comparing the sequence of base pattern which already have, with that of the other patterns. After that, the frame of which the pattern was matched with the base pattern in the same state are to be found so that the reference pattern can be gained by weighting on the numbers of matched frames.

1. 서 론

음성을 사용하여 인간과 컴퓨터 사이에 밀접하게 관련되어 있는 기술은 음성인식과 화자인식이다. 그 중에 화자인식은 화자의 음성신호에서 발생되는 개인의 발성 특성들을 추정하여 인식하는 방식이다. 따라서 화자의 개인의 발성 특성들을 잘 반영하도록 하는 것이 중요하다. 화자인식(speaker recognition)은 화자식별(speaker identification)과 화자확인(speaker verification)기법으로 나눌 수 있다. 화자식별은 등록된 화자들 중에서 가장 유사한 화자를 골라내는 것이다. 화자 확인은 핵심어 인식과 같이 승인(acceptance) 및 거절(rejection)과정을 거치게 된다. 화자인식은 실제로 어떤 형태로 구현 할 것인가의 관점에서 보면 문맥종속형(text dependent)과 문맥 독립형(text independent)로 나눌 수 있다. 화자인식의 근본적인 난점으로는 음성에서의 음운정보와 화자정보 분리의 어려움, 사칭자 거부와 시간변화에 따른 인식률 저하가 있을 수 있으므로 화자 개인의 특성이 잘 반영되도록 해야 한다.[1][2][4]

본 논문에서 제안한 상태별 파라미터 가중 기법은 학습 과정에서 HMM을 이용하여 생성된 모델에 각 상태별 파라미터 가중치를 구하는 과정이 추가되어 구성된다. 학습 과정의 이산HMM을 사용하도록 LBG 알고리듬으로 코드북을 생성한다.[1][5][6] 코드북으로부터 HMM에서 상태수는 각 단어의 음소 개수로 하고, Viterbi 알고리즘을 수행하여 최적의 상태열을 얻는다. DTW를 수행하여 얻은 대표패턴의 상태열을 기준으로 다른 데이터와 DTW를 수행하여 최적의 정합 열을 찾

는다.[1][2] 이때 대표패턴과 같은 상태내에서 정합이 된 프레임을 찾아 그 프레임에 정합된 개수와 각 상태의 프레임의 개수로부터 가중치를 구하고 화자인식을 수행한다. 인식과정은 참조패턴(reference pattern)과 시험패턴(test pattern)사이에 DTW로 화자인식을 수행한다.

본 논문의 구성은 2장에서는 화자인식 시스템의 구성에 대해 알아보고, 3장에서는 본 논문에서 제안한 상태별 파라미터 가중기법에 대해 설명하며, 4장에서는 실험 및 결과, 5장에서 결론을 맺는다.

2. 화자인식 시스템의 구성

전체적인 화자인식 시스템의 블록도는 다음 그림 2.1과 같다.

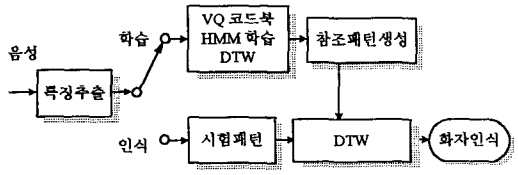


그림 2.1 화자인식 시스템의 블록도

2.1 특징 파라미터 추출

화자의 음성 신호로부터 음성 특징 파라미터 추출은 귀가 음성을 분석하는 방법을 이용한 청각(auditory)분석 방법으로 12차 멜 주파수 켑스트럼 계수(MFCC) 식 (1)과 1차 에너지를 이용한다. 그림 2.2에 특징추출의 블록도를 나타내었다.[3]

$$\widetilde{c_n} = \sum_{k=1}^{K} (\log \widetilde{S_k}) \cos\left(n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right) \quad (1)$$

$$\begin{cases} n = 1, 2, \cdots, L \\ L: \textit{cepstrum order} \\ \widetilde{S_k}: \textit{Mel-scaled power cepstrum} \end{cases}$$

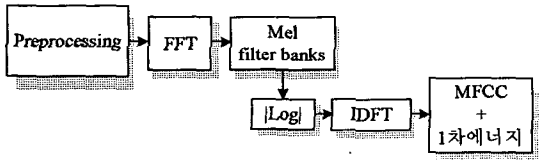


그림 2.2 특징 파라미터 추출 블록도

2.2 학습과정

화자의 음성특징을 추출하기 위해서 화자가 발성한 여러 음성들을 종합하여 분석해야한다. 따라서 특징 파라미터 계수를 이용하여 벡터 양자화(VQ)를 한다. 사용한 알고리즘은 LBG알고리즘을 이용하여 코드북을 생성한다. 생성된 코드북으로 HMM을 사용하여 상태열을 생성한다.

본 논문에서는 단순 left-to-right 모델을 적용한 DHMM(discrete hidden Markov model)을 사용하였으며 상태 수는 인식어휘 내의 음소 개수에 따라 상태수를 결정하였다. HMM을 구성하는 두 가지 요소로 상태천이 확률과 관측확률을 들 수 있는데 그림 2.3에 상태천이 확률을 결정하기 위한 단순 left-right 모델의 예를 나타내었다.

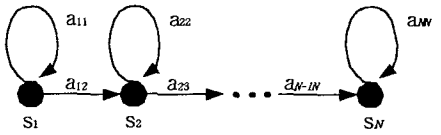


그림 2.3 단순 left-right 모델

HMM의 학습은 입력된 코드북과 특징 파라미터에서 Viterbi 알고리즘과 Baum-Welch 재추정 알고리즘에 이해 학습이 이루어지고 최적의 상태열을 얻는다. Viterbi 알고리즘은 관측 시퀀스와 모델 λ가 주어졌을 때 최적의 상태경로 $Q = q_1, q_2, \cdots q_N$을 찾는 방법이다. Baum-Welch 재추정 알고리즘은 주어진 모델에서 관측 확률을 최대화하기 위해 모델 λ에서 변수 (A, B, π)를 재조정 해준다. 다음은 Viterbi 알고리즘으로 최적 상태경로를 얻는 단계이다.

1 단계 : 초기화

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le n,$$

$$\Psi_1(i) = 0$$

2 단계 : 루프

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \quad 2 \le t \le T, \ 1 \le i \le N$$

$$\Psi_t(j) = \arg\max_{1 \le i \le N} [\sigma_{t-1}(i) a_{ij}] \quad 2 \le t \le T, \ 1 \le j \le N$$

3 단계 : 종료

$$P^* = \max_{1 \le i \le N} [\delta_T(i)],$$

$$q_T^* = \max_{1 \le i \le N} [\delta_T(i)]$$

4 단계 : 경로 역추적

$$q_t^* = \Psi_{t+1}(q_{t+1}^*), \ t = T-1, T-2, \cdots 1$$

DTW(dynamic time warping)알고리즘은 주어진 문

장 혹은 단어 인식에 있어서 참조 패턴과 입력 패턴사이의 거리를 계산할 때 DP(dynamic programming)정합을 이용하여 시간 축 상에 적당하게 나열하여 두 패턴의 차이 값을 최소화하도록 warping 함수를 찾는 것이다.

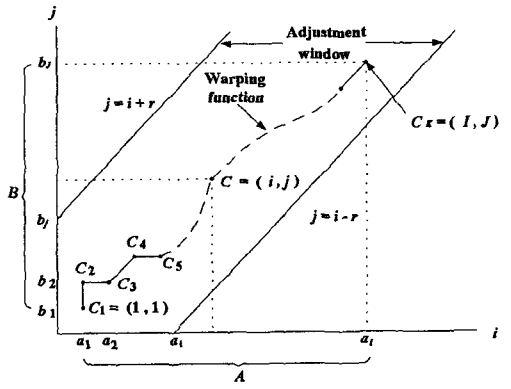
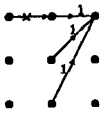


그림 2.4 두 패턴 A와 B 사이의 warping 함수

본 논문에서는 국부경로제한 조건으로 표 1과 같은 Itakura방식의 동적 프로그램 식을 사용하였다. DTW를 이용하여 모델이 되는 입력 파라미터들 중 최소 누적거리에 있는 패턴을 대표 패턴으로 이용한다.

표 1. 국부 경로 제한 따른 동적 프로그래밍식

Type	국부경로제한	동적 프로그램 식
ITAKURA		$\min\begin{cases}D(i_x-1,j)g(k)+d(i_x,j_y)\\D(i_x-1,j_y-1)+d(i_x,j_y)\\D(i_x-1,j_y-2)+d(i_x,j_y)\end{cases}$ with $g(k)=\begin{cases}1 & \phi(k-1)\neq\phi_y(k-2)\\ \infty & \phi(k-1)=\phi_y(k-2)\end{cases}$

2.3 인식과정

입력된 음성 데이터를 학습과정에서 생성된 참조 패턴과 시험 패턴을 DTW를 수행하여 문턱값을 설정하고 거리값을 비교한다. 거리값이 문턱값보다 작으면 "허가", 반대의 경우"거부"판단을 한다.

$$D(R_k, T) < \theta_k \quad \text{허가(acceptance)}$$
$$D(R_k, T) \geq \theta_k \quad \text{거부(rejection)}$$

문턱값을 크게 설정하면 사용자 거부율(false rejection)이 낮아지나 사칭자 허용율(false acceptance)은 높아진다. 따라서 필요에 따라 사용자 거부율을 낮추어 사용할 수도 있고, 사칭자 허용율을 낮추어 사용할 수도 있다.

3. 제안한 상태별 파라미터 가중기법

학습과정에서 학습 파라미터를 DTW에 의한 대표패턴을 생성하고 입력된 코드북과 특징 파라미터로 HMM의 Viterbi 알고리즘을 수행하면 최적의 상태 열을 얻을 수 있다. 대표패턴을 기준으로 나머지 패턴과 정합(matching)을 수행하여 최적 경로를 찾는다.

여기에서 찾은 각각의 상태 열(Q_i)들과 최적의 경로열(P_j)들을 이용하여 같은 상태내의 각 프레임에 정합되는 프레임 개수 $f_c(k)$와 한 상태내의 전체 프레임의 개수 $s_c(i)$로 식(2), 식(3), 식(4)과 같이 가중치 $w(k)$를 구한다.

$$f_c(k) = \sum_n p_{nk} \tag{2}$$
$$s_c(i) = \sum f_{ci}(k), \quad k = 1, 2, \cdots, K \tag{3}$$
$$w(k) = \frac{f_c(k)}{s_c(i)} + 1 \tag{4}$$

여기에서 n은 시험 패턴의 개수, k는 대표 패턴의 프레임 번호, i는 상태를 나타낸다. 과정을 그림 2.5에 보였다. 가중치 값 $w(k)$을 대표 패턴 $y(k)$에 곱하여 참조 패턴을 얻는다.

$$r(k) = y(k) \times w(k) \tag{5}$$

여기서, $r(k)$는 참조패턴 값, $y(k)$는 대표패턴 값이다.

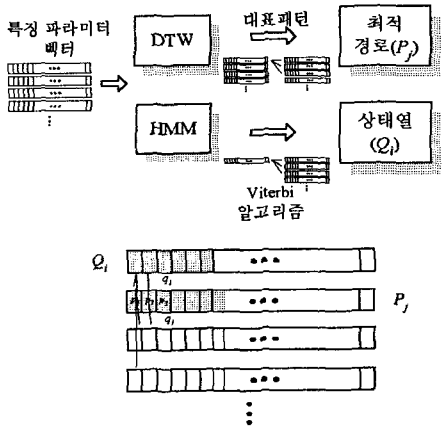


그림 2.5 제안한 상태별 파라미터 가중기법

4. 실험 및 결과

본 논문에서는 화자 인식 실험을 위해 5명의 화자가 약 3개월 동안 3음절, 4음절로 구성된 단어 5개를 각각 20회씩 발성한 데이터와 그 외 14명이 각각 2회씩 발성한 총 780개의 데이터를 수집하였다. 녹음환경은 주변잡음이 존재하는 일반적인 실험실이며 IBM PC에서 마이크를 통한 음성신호를 사운드 카드로 8kHz, 16bit linear PCM으로 A/D 변환하여 수집하였다. 음성의 특징 파라미터 추출을 위한 분석 프레임의 크기는 30ms, 이동 프레임의 크기는 15ms로 하였다. 특징 파라미터 추출은 12차 멜 주파수 캡스트럼 계수와 1차 에너지를 이용하였다. 벡터 양자화 기법으로는 *LBG* 알고리즘을 이용한 32개 코드북을 작성하였다. 상태수는 각 단어의 음소의 개수로 정하였다. 인식 실험을 위한 참조 패턴 생성은 5명의 화자에 대해서 각각 4개의 데이터를 사용하였고, 나머지 데이터를 시험 패턴으로 사용하였다.

화자 확인을 위한 문턱값 결정은 사칭자 허용율(FA)과 사용자 거부율(FR)이 같아지는 동일 오류율(EER ; equal error rate)을 기준으로 사칭자 허용율이 최소로 되는 값을 문턱값으로 결정하였다.

실험 결과 제안한 방법이 기존의 방법에 비해 사칭자 허용율이 0.75%, 사용자 거부율이 1.25% 향상되었다.

표 2. 실험 결과(%)

인식율 / 방법	FA	FR
기존의 방법	4.67	6.25
제안한 방법	3.92	5.0

5. 결 론

현재 화자인식 시스템에 관한 연구가 활발히 진행되고 있고, 앞으로 네트워크를 통한 인터넷과 PC에서의 이용도가 높아질 것으로 예상된다. 따라서 화자의 음성 특징을 잘 반영할 수 있는 여러 가지 파라미터 추출방법 등이 제안되고 있다. 본 논문은 DTW와 이산 HMM을 이용하여 화자인식을 수행하였다. 각 화자의 참조패턴을 생성하기 위해 화자 내의 변이를 수용할 수 있는 특징 파라미터 추출을 제안하였다. 제안한 상태별 파라미터 가중 기법은 DTW로 얻은 대표패턴과 HMM에서의 상태열을 이용한 가중치를 구하여 각각의 대표패턴에 곱하여 새로운 참조패턴을 이용하여 화자인식을 수행했다. 실험결과 사칭자 허용율이 0.75%, 사용자 거부율이 1.25% 향상 되었다.

앞으로 전화망이나 여러 잡음 환경에서 잡음의 영향을 최소화하면서 화자의 특성을 잘 반영할 수 있는 파라미터에 관한 연구와 문턱값 결정에 관한 연구가 필요하다.

참고문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition," Prentice-*Hall*, 1993.

[2] 서광석, "DTW를 이용한 향상된 문맥제시형 화자인식", 전북대학교 석사학위논문, 1999.

[3] John R. Deller, Jr., John G. Proakis, John H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan, 1993.

[4] 김세현, 장길진, 오영환, "문장 종속형 화자확인에서의 관측확률 가중기법", 한국음향학회 학술대회 논문집, pp. 28-31, 1999.

[5] X. D. Huang, Y. Ariki, M. A. Jack, "Hidden Markov moldels for Speech Recognition," Edinburgh Univ. Press, 1990.

[6] Y. Linde, A. Buzo, R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, 1980.