음성구간검출을 통한 화자식별 시스템의 성능개선에 관한 연구

신동성, 정영훈, 배명진
숭실대학교 정보통신공학과

A Study on the Improvement of Speaker Recognition System by Voice Activity Detection

DongSung Shin, YoungHoon Jung MyungJin BAE
Dept. of Information & Telecommunication Engineering, Soongsil Univ., Korea
mjbae@saint.ssu.ac.kr

Abstract

본 논문에서는 묵음구간을 제거하여 화자식별 시스템의 성능향상에 관한 실험을 하였다. 화자식별 방식은 등록된 음성과 테스트 음성을 비교하여 결정논리에 의하여서 화자를 식별하는 방식이다. 이러한 시스템에서 전처리(preprocessing)를 어떻게 해 주느냐에 따라서 인식률에 큰 영향을 미치게 된다. 본 논문에서는 전처리 과정 중에서 음성구간 검출에 대한 실험을 수행하여 성능을 비교하였다. 본 논문에서는 시간영역에서 안정구간(stationary region)과 전이구간(transition region)에서 Normalized AMDF를 적용하였을 때 피치점에서 골(valley)의 기울기가 크다는 점을 이용하여 유성을 검출하였다. 그리고 검출된 유성음 구간 앞뒤로 인접 샘플의 자기상관관계함수(Autocorrelation)의 비를 이용하여 무성음을 검출하였다. 결과적으로 처리시간은 비슷하였으나 전체 인식률은 약 2%정도 개선되었다.

1. Introduction

현대가 정보화 사회로 급속히 진행됨에 따라 대규모의 데이타베이스에 등록되어있는 개인이나 단체의 수많은 정보의 접근, 갱신, 수정이 빈번해지고 있다. 따라서 이에 따른 정보의 보안 문제가 심각해지고, 특정 지역의 출입 통제를 위한 보안 시스템이나 특정시스템을 사용할 때 사용자의 신분에 대한 확인 수단이 필수적이다. 그러나 종래의 개인 신분 확인 수단인 도장, 신분증, 카드등은 도난, 분실, 위조등의 위험을 수반한다. 또한 전화나 통신망을 이용해서 정보 접근을 할 경우에 개인 확인이 더욱 어려워진다. 이에 반해 음성을 이용한 화자 식별 시스템은 음성에 포함되어 있는 개개인 마다의 화자정보를 추출하여 개인을 확인하는 기술로서 사칭자에 대한 처리, 처리시간, 원격자 확인등 시스템 사용의 간편하고, 여러 가지 측면에서 가장 효과적인 기술이고 응용분야도 다양하다는 장점이 있다.[1][2] 그러나 기존의 DTW를 이용한 화자 식별 시스템에서는 많은 화자를 처리할 경우 처리량이 증가하여 인식결과를 얻기 위해서는 많은 시간이 소요된다는 단점을 수반하고 사칭자의 경우에 잘못된 인식을 수행한다는 단점을 수반하게 된다. 따라서 본 논문에서는 시간영역에서 안정구간과 전이구간에서 NAMDF를 적용하였을 때 피치점에서 valley의 기울기가 크다는 점을 이용하여 유성음을 검출하고 검출된 유성음 구간 앞뒤로 인접 샘플의 자기상관법의 비를 이용하여 무성음을 검출하는 방법을 제안하였다.

2. Speaker Recognition System

화자인식은 인식대상에 따라 화자식별(speaker identification)과 화자확인(speaker verification)으로 나눌 수 있다. 화자 식별은 입력된 미지의 음성이 이미 등록된

여러 명의 화자중 어떤 화자에 의해 발성된 음성인지를 판정하는 것을 말하고, 화자확인 방법은 신분 확인 및 음성인식 기술과 조합하여 본인 여부를 가려내는 것이다. 그리고 화자인식은 인식 방법에 따라서 다음과 같이 4가지로 구분할 수가 있다. 그 중 첫번째는 입력패턴을 미리 정해진 기준 패턴(reference pattern)과 비교하여 최적화된 유사성을 판단하는 방법으로 패턴정합법(Pattern Matching)인 동적 정합법(dynamic time warping, DTW), 각 화자별로 신경 회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하도록하여 인식하는 신경회로망이 있다. 그러나 이 방법은 새로운 화자의 추가 시 다시 학습시켜야 한다는 단점과 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에서는 적합치 않다. 세번째 방법인 벡터양자화방법은 입력 패턴과 양자화 코드북(codebook)사이의 거리로 유사성을 판단하는 방법이지만, 많은 학습자료가 필요하고, 화자간의 동적인 변화 특성을 이용하지 못하기 때문에 인식률에 한계가 있다. 마지막으로 HMM(hidden markov model)은 학습기능을 이용하여 화자내의 변이를 흡수 할 수 있으며, 입력패턴의 비선형 정합을 수행하는 특성이 있다. 또한 인식에 사용하는 문장의 종속 여부에 따라 정해진 어휘만을 발성해야하는 텍스트 종속형과 정해지지 않는 어휘로 인식을 수행하는 텍스트 독립형으로 나눈다.

일반적인 화자인식 시스템은 다음과 같은 3가지로 크게 구분할 수가 있는데 그 전체적인 과정을 보면 다음과 같다. 우선 입력된 음성은 전처리 과정을 통해 디지탈 신호로 변화되고, 이 변환된 신호는 음성 구간 검출과정을 거친 뒤 필요한 특징값을 추출하는데 사용되어진다. 이 추출된 음성 파라미터열은 DTW에 의해 패턴 정합을 수행함으로써 화자인식을 결정하게 된다.

2.1 Speech Feature Extraction

그림 2-1은 음성 특징 추출 과정을 보여주고 있다. 한 프레임에 해당하는 음성샘플은 윈도우(hamming window)를 이용한 뒤 고주파의 효과를 강조시키기 위해 식 2-1의 프리엠퍼시스(preemphasis) 필터를 거치게 된다.

$$h(z) = 1 - 0.98z^{-1} \qquad (2\text{-}1)$$

이 프리엠퍼시스된 음성신호로 부터 선형 예측 계수(linear prediction coefficient)를 구하고 이를 이용하여 cepstrum계수를 구하고, 귀의 특성을 고려한 mel-frequency scale로 왜곡하여 특징 파라미터인 mel-cepstrum을 구하였다

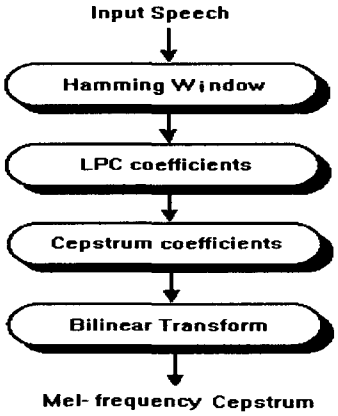

Fig 2-1. Speech feature extraction

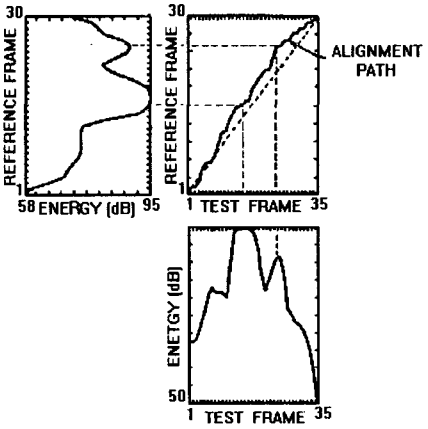

Fig 2-2 Pattern matching method using DP Algorithm

2.2 DTW를 이용한 패턴정합법

화자마다의 발성길이가 다르고 같은 화자가 동일한 어휘를 발성할 지라도 그 길이가 변하기 때문에 기준패턴과 테스트 패턴의 특징벡터(Feature vector)를 비교하기 위해서는 발성마다 기준패턴과 정합하기 위해서 비선형적으로 전개 또는 수축하면서 왜곡시키는 과정이 필요하게 되는데 이 과정을 dynamic time warping(DTW)라고 한다. 그림 2-2는 DTW 알고리즘을 이용한 패턴정합

를 보여주고 있다[3][4].

이와같은 화자식별 시스템은 모든 기준패턴과 DTW를 수행하여 비교하여야 하기 때문에 과다한 계산량을 요구하게 되고, 사칭자가 발성한 경우에도 잘못 인식하게 되는 결과를 수반하게 된다.

3. 음성구간검출

3.1 NAMDF를 이용한 유성음검출

윈도우의 영향에 무관하게 현재의 프레임이 어떤 상태에 존재하는 지를 측정하는 새로운 방법으로는 다음과 같이 규준화된 AMDF(Average Magnitude Difference Function) 를 정의해서 사용할 수 있다.[8][9][10]

$$NAMDF(d)=\frac{\sum_{n=1}^{N}|s(n)-s(n-d)|}{\sum_{n=1}^{N}|s(n)|+|s(n-d)|} \quad (3\text{-}1)$$

N은 AMDF를 구하려는 윈도우 구간이다. 지연인자를 점차 증가시키면서 이 AMDF를 구해보면, 지연인자가 프레임내 음성피치에 정수배가 될 때마다 NAMDF는 거의 영이 된다.

식 3-1의 AMDF값은 지연인자 d간격을 갖는 N개 샘플간의 진폭에 대한 평균 차이 값이 되기 때문에 음성파형의 d구간사이의 유사도를 나타내는 규준 화된 거리 값으로 적용할 수 도있다. 식 3-1의 규준화된 (Normalized) AMDF는 d간격의 두 음성파형 블록에 대해, 평균진폭 차이 값을 나타내지만 음성신호가 갖는 피치주기의 변화는 배제하지 않았다. 따라서 두 음성파형 블록에 대한 피치주기의 영향을 제거하려면 시간을 고려하는 지연인자 d의 값이 음성피치에 일치하였을 때의 규준화된 AMDF값을 두 파형블럭의 유사도 값으로 사용할 수 있게된다.

규준화된 AMDF값이 영에 근접하면 d간격을 유지하는 두 음성파형 N개 블록간에는 유사성이 최대가 되고 이때의 점에서 양의 기울기를 측정하면 유사성이 최대가 아닐 때와 비교하여 큰 기울기를 갖게된다. 따라서 각 프레임(frame)에서 기울기 값이 큰 값이 계속 유지가 되면 안정구간에 놓여있게 되고 기울기 값이 큰 값에서 작은 값 또는 작은 값에서 큰 값으로 변하게 되면 전이구간에 놓여 있게된다.

3.2 무성음 구간 검출

단구간 자기상관 함수는 다음 식으로 표현가능하다.

$$\phi_n(i,j)=\sum_{m=0}^{N-1-(j-i)} s_n(m)s_n(m+j-i),\ 1\le i\le p,\ 0\le j\le p \quad (3\text{-}2)$$

여기서, $R_{n(j)}=\sum_{m=0}^{N-1-j} S_{n(m)}S_{n(m+j)}$

$$\sum_{i=0}^{p}\alpha_i\phi_n(i,j)=\phi_n(i,0),\ \text{for } i=1,\ldots\ldots,p \quad (3\text{-}3)$$

자기상관법을 이용하여 위 식을 풀면 다음과 같다.

$$\begin{vmatrix} R_{n(0)} & R_{n(1)} & \cdot & R_{n(p-1)} \\ R_{n(1)} & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot \\ R_{n(p-1)} & \cdot & \cdot & R_{n(0)} \end{vmatrix} \begin{vmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \alpha_p \end{vmatrix} = \begin{vmatrix} R_{n(1)} \\ R_{n(2)} \\ \cdot \\ R_{n(p)} \end{vmatrix}$$

p=1에 대하여 위의 식을 정리하면 다음과 같은 식으로 표한 가능하다.

$$\alpha_1=\frac{R_n(1)}{R_n(0)} \quad (3\text{-}4)$$

위 식은 무성음일 때 음수값을 갖게 되는 특징이 있다. 이를 이용하여 무성음구간을 검출한다.[11]

4. Experimental Result

본 논문의 알고리즘을 모의실험하기 위해 IBM PC에 마이크가 장치된 16비트 A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 20명의 남녀 화자가 각각 본인의 이름을 발성한 음성 시료를 11khz로 샘플링하고 16비트로 양자화하여 사용하였다. 한 프레임의 길이는 512샘플이며, 256샘플씩 오버랩(Overlab)시켜 특징벡터를 추출하였다. 인식을 위한 특징벡터로는 14차 Mel-Cepsturm를 사용하였다. 기준패턴으로는 20대 남녀 50명이 2번 발성하여 일주일 동안 발성된 음성을 사용하였다. 그리고 사칭자의 효과를 알아보기 위해서 10명으로 하여금 등록된 화자의 음성을 일주일 동안 4번씩 발성하게 하였다. 실험결과 제안한 방법이 기존의 방법에 비해 인식시간은 비슷하였고 인식률은 2%정도 증가하였다.

Table 4-1. The experimental result

	conventional method	Proposed method
Recognition rate	93%	95%

5. 결 론

현대가 정보화 사회로 변모되어 가면서 많은 정보의 보안문제가 중요한 사회문제로 대두되고 있다. 이러한 해결책으로 음성을 이용한다면 도난, 분실, 위조등의 위험을 수반하는 종래의 개인 신분확인 수단을 사용하지 않고도 효과적으로 사용할 수 있다. 본 논문은 DP알고리즘을 사용한 문장 종속 화자인식시스템에서 위의 문제로 인하여서 발생하는 오인식률 증가의 단점를 보완하기 위해 규준화된 AMDF의 기울기와 인접 샘플의 자기상관법의 비를 이용해서 음성구간을 정확히 검출하여 인식알고리즘을 수행하는 것를 제안하였다..

실험결과 기존의 방법에 비해 계산 시간은 비슷하였으나 전체 인식률은 2%정도 개선되었다.

6.Reference

[1] S. Funui, Digital Speech Processing, Synthesis and Recognition, Marcel Dedder, Inc., 1992

[2] L. R. Rabiner & R.W.Schafer, Digital Processing of Speech Signal, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978

[3] L. R. Rabiner & Biing-Hwang Juang, Funda- mentals Of Speech Recognition, Prentice-Hall AT&T, U.S.A, 1993

[4] Hiroaki Sakoe & Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, vol.26, No.1, pp.43-49, Feb. 1978.

[5] C.J. Weinstein, S.S. McCandless, L.F. Mondshein, and V.W. Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans. on ASSP, Vol, ASSP-23, No.1, pp.54-67, Feb. 1975.

[6] W.F. Ganong, and R.j. Zatorre, "Measuring Phoneme Boundaries Four Ways", J. Acoust. Soc. Am. Vol. 68, No.2, pp.431-439, Aug. 1980.

[7] S.J Kim, "A Segmentation Algorithm of the Connected Word Speech by Statistical Method", IEEK, Vol.26, No.4, pp.151-162, Apr., 1989.

[8] R. Mori, P. Laface, and E. Piccoo, "Automatic Detection and Description of Syllabic Features in Continuous Speech", IEEE Trans. On ASSP, Vol.ASSP-24, No.2, pp.880-883,Oct., 1976.

[9] M.J. Bae, "On Detecting the Steady State Segments of Speech Waveform by using the Normalized AMDF", IEEK, Vol.14, No.1, pp.600-603, Jun., 1991.

[10] 손상목, 홍성훈, 정형교, 배명진, "규준화된 AMDF에 의한 음성신호의 전이구간검출," 한국통신학회 하계종합학술발표회 논문집, Vol.15, No.1, pp.128-131, 1996년 7월.

[11] 민소연, 신동성, 배명진, "성문특성 측정을 통한 유/무성음 결정에 관한 연구", 대한전자공학회 하계학술발표대회, No.4, pp.281-284, 2001 6월