1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

시간축 스케일링에 의한 화자 식별 개선에 관한 연구

°정형교, 배재옥, 박영호, 배명진
숭실대학교 정보통신공학과

A Study on Improvement of Speaker Identification with Time axis Scaling

°Hyunggoue CHUNG, Jaeok BAE, Youngho PARK and Myungjin BAE
Dept. of Telecommunication, Soongsil University

요 약

기존의 DTW를 이용한 화자 인식 시스템은 DTW의 단점이라 할 수 있는 과다한 계산량을 갖는다는 문제점을 갖고 있다. 따라서 본 논문은 텍스트 종속 화자 인식 시스템에서 피치 분포도를 갖는 개별 화자의 DTW를 수행하기 전에 시간축 스케일링을 이용한 전처리로 인식시의 계산량을 감소시키는 과정을 미리 수행한 후 감소된 기준패턴들의 입력 신호에 대해서만 DTW를 수행하는 방법을 제안하고자 한다. 제안한 방법을 실험하였을 경우 87.5%의 평균 처리 시간이 감소하였고, 더불어 인식률 감소는 거의 없었다.

1. 서 론

현대 사회의 급속한 정보화는 대규모 데이타베이스에 등록되어 있는 개인과 단체의 수많은 정보의 접근, 갱신, 수정 작업을 빈번하게 하고 있다. 따라서 이에 따른 정보의 보안 문제가 심각해지고 있으며, 특정 지역의 출입 통제를 위한 보안 시스템이나 특정 시스템을 사용할 때 사용자의 신분에 대한 확인 작업이 필수적이게 되었다. 그러나 기존의 개인에 대한 신분 확인 수단인 도장, 신분증, 카드 등은 도난, 분실, 위조 등의 위험을 수반한다. 또한, 전화나 통신망을 이용한 정보 접근은 개인의 신분 확인을 더욱 어렵게 한다. 이에 반해 음성을 이용한 화자 식별 시스템은 음성에 포함되어 있는 개개인의 화자 정보를 추출하여 개인의 신분을 확인하는 기술로서 사칭자에 대한 거부, 처리시간, 원격자 확인등 시스템 사용이 간편하며 그에 따른 응용 분야도 다양하다[1][2].

DTW를 이용한 화자 식별 시스템에서는 다수의 화자를 처리할 경우 처리량이 증가하여 인식 결과를 얻기 위해서 많은 시간이 소요된다는 단점을 갖게 된다. 따라서, 본 논문에서는 피치 분포도에 의해 미리 후보자가 선정된 화자 개개인의 발성에 대해 시간축 스케일링을 적용한 후 DTW를 수행하여 계산량을 감소시키는 방법을 제안하고자 한다.

2. 화자인식 시스템

화자인식은 인식 대상에 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, 화자확인은 신분 확인 및 음성인식 기술과 혼합하여 본인 여부를 가려내는 것이다. 화자인식은 인식 방법에 따라서 다음과 같이 4가지로 구분할 수가 있다. 첫 번째는 입력패턴을 미리 정해진 기준 패턴(reference pattern)과 비교하여 최적화된 유사성을 판단하는 방법으로 패턴 정합법(pattern matching)인 동적 정합법(DTW:Dynamic Time Warping)이며 두 번째로는, 각 화자별로 신경 회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하도록 하여 인식하는 신경회로망(neural network)이 있다. 그러나 이 방법은 새로운 화자의 추가시 전체 네트워크를 재학습시켜야 한다는 단점과 고도의 병렬계산 능력이 필요하기 때문에 응용성이 높은 환경에서는 적합치 않다. 세 번째 방법인 벡터양자화 방법은 입력 패턴과 양자화 코드북(codebook)사이의 거리에 대한 유사성을 판단하는 방법이지만, 많은 학습 자료가 필요하고 화자간의 동적인 변화 특성을 이용하지 못하기 때문에 인식률에 한계가 있다는 문제점을 갖게 된다. 마지막으로 HMM(Hidden Markov Model)을 이용한 방법은, 학습기능을 이용하여 화자내의 변이를 흡수할 수 있으며 입력패턴의 비선형 정합을 수행하는 특성이 있다[7].

일반적인 화자인식의 과정은 다음과 같은 3단계로 분류할 수 있다. 우선 입력된 음성의 전처리 과정을 통한 입력 신호의 변환과 이 변환된 신호를 가지고 음성 구간 검출과정을 거친 뒤 필요한 특징 값을 추출하는 과정이 있다. 추출된 음성 파라미터 열은 DTW에 의해 패턴 정합을 수행함으로써 최종적인 화자인식을 결정하게 된다[9].

2.1 음성구간 검출

음성 구간의 검출은 화자인식 시스템에 큰 영향을 줄 수 있기 때문에 정확한 검출이 요구될 뿐만 아니라 실시간 시스템을 사용하기 위해서 전체 계산량을 크게 증가시키지 않는 효율적인 방법이어야 한다. 본 논문에서는 음성구간을 검출하기전 안정된 피치구간을 먼저 찾는다. 이렇게 찾은 피치구간에서 무성음 구간을 포함하기 위해서 일정 범위 내에서 입력된 음성을 모두 저장하게 된다. 이렇게 저장된 음성구간에 대해서만 단구간 에너지와 영교차율을 이용한 음성구간 검출을 사용하였다. 그리고 음절사이의 묵음구간이 존재 할 수 있기 때문에 나타날 수 있는 오류를 개선하기 위해서 끝점이 검출된 후에도 40프레임간 다시 음성의 시작점을 검출하는 과정을 수행하게 되고, 다시 음성의 시작점이 검출되면 묵음구간이 존재하는 음성으로 간주하고 다시 끝점을 검출하는 과정을 반복한다[2][7][10].

2.2 음성 특징 추출

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

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

이 프리엠퍼시스된 음성 신호로부터 선형 예측 계수(linear prediction coefficient)를 얻어 cepstrum 계수를 구하고, 귀의 특성을 고려한 mel-frequency scale로 왜곡하여 특징 파라미터인 mel-cepstrum을 구하였다[3]

2.3. DTW를 이용한 패턴 정합법

화자들은 각각의 발성 길이가 다르고 같은 화자가 동일 어휘를 발성하더라도 그 길이가 변하기 때문에 기준 패턴과 테스트 패턴의 특징벡터를 비교하기 위한 과정이 필요하게 된다. 이때, 발성마다 기준패턴과 정합하기 위해 비선형적으로 전개 또는 수축하면서 왜곡시키는 과정이 적용되는데 이 과정을 Dynamic Time Warping(DTW)이라고 한다. 그림 2-2는 DTW 알고리즘을 이용한 패턴정합을 보여주고 있다[3][8].

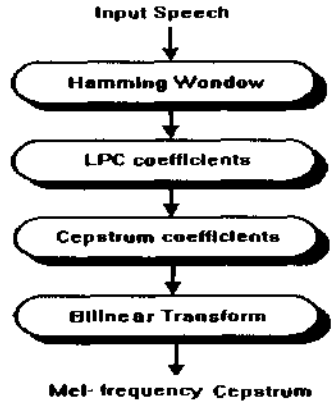

그림 2-1 음성 특징 추출

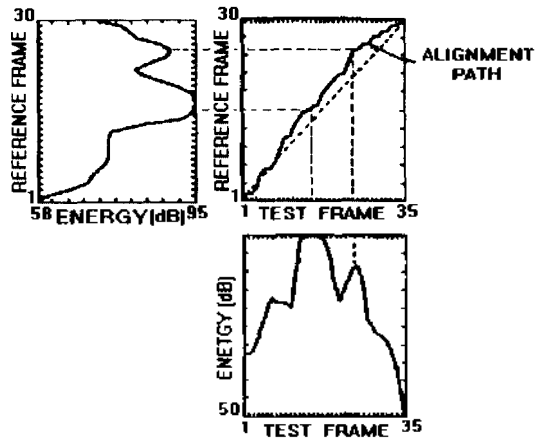

그림 2-2 DP 알고리즘을 이용한 패턴정합법

이와 같은 화자식별 시스템은 모든 기준 패턴과 DTW를 수행하여 비교하여야 하기 때문에 과다한 계산량을 요구하게 되고, 사칭자가 발성한 경우에도 잘못 인식하게 되는 결과를 수반하게 된다.

3. 양자화 오차를 이용한 피치 분포도

M비트로 선형 양자화된 음성신호 s(n)은 식 3-1과 같이 나타낼 수 있다. 여기서 Q_L은 음성신호를 (M-N)비트로 부호화할 때 발생하는 양자화 오차이다. 양자화 오차는 진폭 변화의 범위가 2^N-1로 정규화된 특성을 갖기 때문에 시간 영역에서 음소 천이에 따른 파형의 변화의 영향을 줄일 수 있게 된다. 그리고, 유성음 파형의 경우에 에너지가 우세한 기본주파수가 Q_L의 최대진폭을 유지하게 되고 에너지가 낮은 고차의 포만트들은 Q_L의 진폭범위 내에서 파형의 빠른 변화를 이루게 된다.

양자화 오차 Q_L을 사용하여 저역 특성이 강한 음성신호의 정규화된 파형을 추출하여 정규화된 주기성을 강하게 만들어 피치 주기 검색에 사용한다.

이와 같이 양자화 오차 신호에서 주어진 프레임에서 문턱값을 입상(rising) 교차점이 시작하는 점 N_S

$$\begin{aligned} s(n) &= \sum_{i=0}^{M-1} a_i 2^i \\ &= \sum_{i=0}^{N-1} a_i 2^i + \sum_{i=N}^{M-1} a_i 2^i \\ &= Q_L + Q_H \end{aligned} \qquad (3\text{-}1)$$

과 끝나는 점 N_E 사이의 간격과 그 사이의 입상교차율(RTCR: Rising Threshold level Crossing Rate)로 나누어 다음과 같이 그 프레임의 평균주기를 검출한다.

$$PITCH(fr) = \frac{(N_E - N_S)}{RTCR} \qquad (3\text{-}3)$$

식(3-3)에 의해 fr번째 프레임에서 검출된 PITCH(fr) 값은 피치의 존재 영역인 2.5-25ms이내에 있어야 하며 검출된 값이 그 프레임내의 개별 입상교차점간의 간격에 비해 10%이내에 존재하면 올바른 피치 주기로 판정한다. 이러한 조건이 만족되지 않는 경우에는 무성마찰음, 무성파열음, 묵음 등의 구간으로 처리한다[5][6].

이와 같이 구해진 피치로 만든 피치분포도를 이용하여 기준패턴과 비교하게 되어 정해진 문턱 값보다 높은 범위의 피치를 갖는 기준패턴들만을 후보자로 선정하게 된다.

4. 시간축 스케일링에 의한 처리 시간 감소

피치 분포도를 작성하여 후보자를 결정하면 전체 인식 대상의 범위는 감소하게 된다. 그러나 DTW로 인한 계산량은 아직 상당 부분 남게 되므로 DTW 자체의 계산량을 줄이기 위해 패턴들의 시간축 스케일링을 수행하게 된다.

유성음의 스펙트럼은 제 1포만트와 제 2포만트로 구성되는 협대역 신호와 제 3포만트, 제 4포만트 이상의 포만드로 구성되는 광대역 신호로 구성된다. 만약 음성 신호가 시간 영역에서 일정 비율로 스케일링된다면 광대역 신호와 협대역 신호가 똑같이 스케일링되기 때문에 원래 신호가 가지고 있는 신호 특성의 변화는 크지 않게 된다. 이와 같은 시간축 스케일링 알고리즘을 화자 식별의 전처리 과정에 적용하면 인식률이 유지되면서 전체 처리 시간을 감소시킬 수 있게 된다.

본 논문에서 사용되는 시간축 스케일링의 효과를 나타내기 위해 전체 기준패턴의 스케일링 비율을 2, 3, 4로 하였다. 표 4-1은 그 결과를 나타내는 것으로, 비율이 2인 경우 기존의 DTW에 비해서 약 67%의 계산량 감소를 발생시킨다는 것을 알 수 있다.

표 4.1 시간축 스케일링을 적용한 계산 시간
(평균 처리 clock)

	기존의 방법	스케일링 비율 2	3	4
계산시간	12	4	5.3	6.7
	93	92.5	92.5	93

5. 실험 및 결과

본 논문을 시뮬레이션하기 위해 사용된 실험 장비는 IBM-PC 586에 마이크가 장치된 16-bit A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 10명의 남녀 화자가 발성한 음성 시료를 11kHz로 샘플링하고 16bit로 양자화하여 사용하였다. 한프레임의 길이는 300샘플로, 150샘플씩 overlap시켜 특징벡터를 추출하였다. 인식을 위한 특징벡터로는 14차 mel-cepstrum을 사용하였다. 기준 패턴으로는 20대 남녀 10명이 20번씩 발성한 음성으로 구성하였다. 실험을 위한 본 논문의 구현은 C언어로 작성하였다.

그림 5-1은 본 실험에서 사용한 화자인식 시스템의 전체 블럭도이다. 우선 입력된 음성 데이터의 양자화 오차를 구한다. 그리고, 이 신호에 대한 안정된 분포를 갖는 범위를 찾아서 허용 범위 내에서 음성을 저장한 뒤 피치를 검출한다. 이렇게 검출된 음성에서 검출한 피치를 이용하여 테스트 패턴의 피치분포도를 구하여 기준 패턴과의 비교를 통해 문턱치보다 높은 값에 있는 기준 패턴들을 후보자로 선정하였다. 그리고 비교패턴을 형성하기 전에 DTW의 계산량을 줄이기 위해 두샘플마다 시간축 스케일링을 수행하였다. 이와 같이 시간축 스케일링된 비교 패턴에 대해 DTW를 수행하여 화자 인식을 하게 된다.

표 5-1과 표 5-2는 데시메이션 비율을 2로 하고, 피치분포도를 이용하여 선택된 후보자에 대해서만 비교를 수행한 결과를 보여주고 있다. 시뮬레이션상에서 소모되는 클럭수를 비교해볼 때 제안한 인식 시스템이 DTW만 사용하는 기존의 방법보다 87.5% 정도의 처리시간이 감소되었고, 전체 인식률은 기존 방법보다 평균 0.5% 감소하였다.

6. 결 론

현대 정보화 사회의 급속한 대두는 많은 정보의 보안문제를 중요한 사회 문제로 불러일으키고 있다.

이러한 문제의 해결책으로 화자 개인의 음성 신호를 이용한다면 도난, 분실, 위조 등의 위험을 수반하는 종래의 개인 신분 확인 수단을 사용하지 않고도 효과적으로 신분 확인이 가능하다.

그러나, 기존의 DTW를 이용한 화자 식별 시스템에서는 많은 화자를 처리할 경우 처리량이 증가하여 인식결과를 얻기 위해서는 많은 시간이 소요된다는 단점을 수반한다. 따라서 본 논문에서는 피치 분포도를 이용하여 식별해야할 후보자들을 미리 선정한 후, 시간축 스케일링된 비교패턴을 가지고 DTW를 처리해서 계산량을 줄이는 방법을 제안하였다. 본 논문에서 제안한 방법을 실험한 결과 전체처리시간은 87.5% 정도 단축시킬 수가 있었고 전체 인식률은 평균 0.5% 감소하였다.

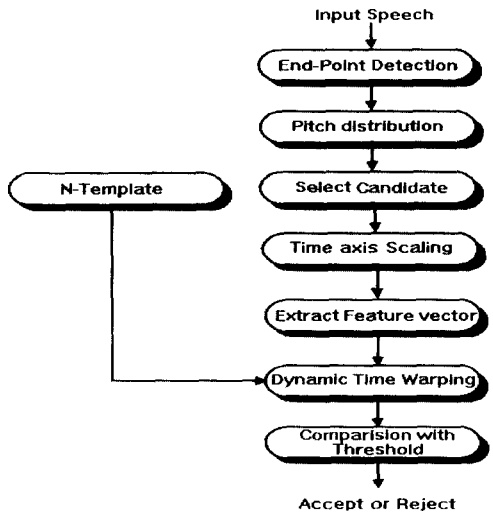

그림 5.1 본 논문에서 제안한 방법.

표 5.1 전체 처리 시간(평균 처리 clock)

	12
	1.5

표 5.2 전체 인식률(%)

	93
	92.5

6. 참 고 문 헌

[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, Fundamentals Of Speech Recognition, Prentice-Hall AT&T, U.S.A, 1993

[4] 정종순, 경연정, 최승호, 이황수, "대표 평균치 패턴과 가중 켑스트럼을 이용한 화자 인식의 성능 향상에 관한 연구", 제 12회 음성 통신 및 신호처리워크샵논문집, pp.179-183,June,1995.

[5] 정형교, 홍성훈, 배명진, 변경진, 유하영, "양자화 오차를 이용한 음성신호의 피치검출" 한국 음향학회, 제 9회 신호처리합동학술대회 논문집, Vol.9, No.1, pp.467-470, 1996년 10월.

[6] 함명규, 이동기, 배명진, "음성선형 PCM파형에서 양자화 오차를 이용한 F1/F0을 검출" 한국음향학회, 제 10회 신호처리합동학술대회논문집, Vol.10, No.1, pp.261-264, 1997년 9월.

[7] 정종순, "대표 평균패턴과 가중 켑스트럼을 이용한 화자인식의 성능 향상에 관한 연구", 석사 학위 논문, 한국과학기술원, 1996년.

[8] 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.

[9] S.Furui, "Cepstral Analysis Techniques for Automatic Speaker Verification", IEEE Trans. on ASSP, vol.29, pp.254-272, April 1981.

[10] 배명진, 이인섭, 안수길, "음성신호를 표본화할 동안 효율적인 실시간 저장기법" 한국음향학회, 학술발표대회, pp.66-74, 1986년 11월.