제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

음성신호의 전이구간을 이용한 화자 인식의 성능향상에 관한 연구

오세영(*), 최성영(**), 배명진(*)
*숭실대학교 정보통신공학과, **서울시립기능대학

On a Performance Improvement of Speaker Recognition using the Transition Region of Speech Signal

Seyoung OH(*), Seongyoung CHOI(**) and Myungjin BAE(*)
*Dept. of Information and Telecommunication Engineering, Soongsil University
**Dept. of Electronic Technology, Polytechnic College of Seoul
E-mail(*) : mjbae@saint.soongsil.ac.kr

요약

기존의 DP(Dynamic Program)알고리즘을 이용하여 화자를 인식할 경우 시스템에 등록되어 있는 화자의 수가 증가할수록 처리해야할 데이터의 양이 많아진다. 그러므로 인식률이 저하되고 처리시간이 증가한다는 단점이 있다. 본 논문에서는 이러한 단점을 보완하기 위해 화자가 발성한 음성신호에서 안정구간내의 일정 파형을 삭제한 후 전이구간을 위주로 DP알고리즘을 적용하여 화자를 인식한다. 제안한 방법으로 실험한 결과 시스템의 전체 인식률은 기존의 DP알고리즘을 이용한 결과에 비해 1%의 향상을 보였고 처리시간은 21.6% 감소함을 볼 수 있었다.

1. 서 론

사회가 정보화, 고속화되면서 개인이나 특정 집단의 정보 교환 및 관리가 중요시 되고 있다. 따라서 정보의 보안 문제가 심각한 사회문제로 대두되고 이에 따른 대처방안을 고려하는 많은 방법들이 시도되고 있다. 종래에는 개인 확인 수단으로 도장, 신분증, 카드 등이 주로 사용되었다. 하지만 이러한 방법들은 도난, 분실, 위조 등의 위험을 안고 있다. 그리고 전화나 통신망을 이용해서 정보접근을 시도할 경우에는 개인 확인이 더욱 어려워진다. 이에 반해 개인의 음성을 이용한 신분 확인 방법은 개개인마다의 변별력 있는 화자정보를 추출하여 개인을 확인하는 기술이다. 이 방법은 종래의 개인확인 방법들에 비해 분실이나 도난 등과 같은 문제가 없으며 사칭자에 대한 처리, 처리시간, 원격자 확인 등의 경우에 효과적이다[1][2]. 이렇게 음성을 이용한 여러 가지 개인확인 방법 중 DP알고리즘을 이용한 방법이 있다. 하지만 이 방법은 처리해야할 데이터량이 증가함에 따라 인식률이 저하되고 처리시간이 증가한다는 단점이 있다.

본 논문에서는 음성신호를 양자화한 뒤 얻어진 오차 신호가 음성신호의 저역성분을 가진다는 특징을 이용하여 화자의 기본주파수를 정규화된 AMDF법을 이용하여 측정한 뒤 프레임간의 공통된 피치주기를 갖는 성분을 제거하였다. 이렇게 음성신호의 안정구간에서 일정한 음성파형을 제거시킨 음성시료를 패턴정합법을 사용하여 처리함으로써 인식율을 향상시킬 뿐만 아니라 데이터량을 감소시킴으로써 기존의 DP알고리즘의 단점인 계산시간을 줄일 수 있는 방법을 제안하고자한다.

2. 화자 인식 시스템

화자 인식은 일반적으로 처리 대상에 따라 다음과 같이 두 가지로 분류할 수 있다. 첫째로 화자식별(Speaker Identification)은 등록된 화자집단에 요청중인 화자의 발성이 등록되어 있는지를 결정하는 과정이고, 둘째로 화자확인(Speaker Verification)은 지금 발성중인 화자가 본인의 것

인지의 여부를 결정하는 과정이다.

또한 화자인식은 인식 방법에 따라서 다음과 같이 4가지로 구분할 수 있다. 첫째로 패턴정합법(Pattern Matching)에 의한 동적 정합(DP Algorithm)은 입력패턴을 미리 정해진 기준 패턴과 비교하여 최적화된 유사성을 판단하여 화자를 인식하는 방법이다. 둘째로 신경회로망은 각 화자별로 신경회로망을 구성하고 화자간의 변별력을 갖도록 학습을 수행하는 인식 방법이다. 그러나 이 방법은 새로운 화자의 추가시 다시 학습시켜야 하고 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에는 적합하지 않다는 단점이 있다. 세 번째 방법인 벡터양자화 방법은 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 화자인식 시스템은 인식에 사용하는 문장의 종속여부에 따라 정해지지 않는 어휘로 인식을 수행하는 텍스트 독립형(Text Independent)과 정해진 어휘만을 발성해야하는 텍스트 종속형(Text Dependant)으로 나눌 수 있다[6].

2.1 음성구간 검출

정확한 음성검출이 이루어지지 않는다면 화자인식의 인식률에 큰 영향을 미칠 수 있기 때문에 정확한 검출이 요구되며 실시간 시스템을 사용하기 위해서는 전체 계산량을 크게 증가시키지 않는 방법이여야 한다.

본 논문에서는 음성구간을 검출하기전 안정된 피치구간을 먼저 찾은 뒤 무성음구간을 포함하기 위해서 일정 범위내에서 입력된 음성을 모두 저장하게 된다. 이렇게 저장된 음성구간에 대해서만 단구간 에너지와 영교차율을 이용하여 음성구간을 검출한다. 그리고 음절사이의 묵음구간이 존재할 수 있기 때문에 끝점이 검출된 후에도 일정 프레임 동안 다시 음성의 시작점을 단구간 에너지를 이용하여 검출한다. 만일 또다시 시작점이 검출되면 묵음구간이 존재하는 음성으로 간주하고 다시 끝점을 검출하는 과정을 반복한다[2][6][8].

2.2 음성 특징 추출

음성신호에는 화자의 발성기관 특성과 발성습관에 따른 화자의 정보가 나타나게 된다. 즉, 성도의 변동특성이 스펙트럼상에서 공진주파수의 차이로 나타나게 된다. 이는 곧 화자간의 차이이다. 이러게 화자의 특성을 나타내는 여러가지 음성 특징 중 켑스트럼(Cepstrum) 계수가 인식에 유용하다고 알려져 있다. 본 논문에서는 켑스트럼 계수를 얻기 위해 다음과 같은 과정을 수행하였다. 먼저 음성신호를 윈도우(Hamming Window)함수를 사용하여 분할하였다. 그 후 고주파성분을 강조시키기 위해 분할된 음성신호를 프리엠퍼시스 (Preemphasis) 필터에 통과시킨다. 이렇게 고주파항이 강조된 음성신호로부터 선형 예측 계수(Linear Prediction Coefficient)를 구하게 되면 아래와 같은 식을 이용하여 켑스트럼 계수를 구할 수 있다.

$$c_m = a_m + \sum_{k=1}^{m-1}(\frac{k}{m})c_k a_{m-k} \quad 1 \le m \le p$$

$$c_m = \sum_{k=1}^{m-1}(\frac{k}{m})c_k a_{m-k} \quad p < m$$

여기서 p는 LPC계수의 차수, m은 켑스트럼 계수의 차수를 의미한다. 그림 2-1은 이러한 특징벡터를 추출하는 과정을 보여주고 있다[3].

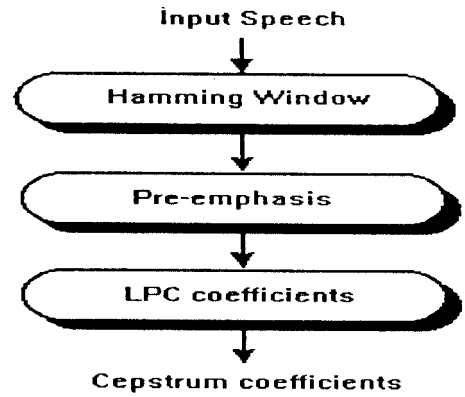

그림 2-1. 특징벡터 추출 과정

2.3 DP알고리즘을 이용한 패턴 정합법

동일한 화자가 같은 발성을 할 경우에도 발성의 길이차가 생기게 되므로 음성패턴에서 시간축의 비선형적 변동을 일으킨다. 따라서 패턴정합을 이용한 화자인식 시스템에서는 이러한 문제점을 고려하여야한다. 두 가지의 패턴들을 비교하는 패턴정합에는 다음과 같이 세 가지 방법이 있다. 첫 번째 방법은 하나의 패턴을 고정시킨 뒤 선형적인 정규화를 수행하는 방법이다. 그러나 이 방법은 발성의 시간적 변이가 항상 존재하기 때문에 모든 음성 정보의 고려가 어렵다는 단점이 있다. 두 번째 방법은 첫 번째 방법을 개선하기 위한 방법으로 하나의 패턴을 고정시킨 뒤 나머지 패턴을 고정된 패턴의 길이로 압축하는 방법이다. 그러나 이 방법은 음성패턴의 압축시에 중요한 파라미터 성분을 제거할 수 있기 때문에 인식율이 저하된다는 단점이 있다. 마지막으로 비선형 정합 방법이 있다. 이 방법은 비선형 왜곡함

수를 이용하여 두 신호의 시간적 변동을 압축하거나 확장한 뒤 두 패턴들 사이의 정규화되는 거리를 오차로 계산하여 패턴정합 하는 것이다[10]. 이중에서 마지막 방법인 DP알고리즘은 두 음성패턴의 시간 배열동안 발생하는 에러를 효과적으로 최소화할 수 있기 때문에 다른 패턴정합법에 비해 높은 인식율을 얻을 수 있다. 이를 위한 실제적인 알고리즘은 다음 식과 같이 정의된다[7].

단계 1)

$g_1(c(1)) = d(c(1)) \cdot w(1)$

단계 2)

$g_k(c(k)) = \min_{c(k-1)}[g_{k-1}(c(k-1)) + d(c(k)) \cdot w(k)]$

단계 3)

$D(A, B) = \frac{1}{N} g_K(c(K))$

2.4 F-ratio를 이용한 캡스트럼 가중치

F-ratio는 특징파라미터의 유용성 척도에 주로 사용되는 것으로 아래와 같이 정의할 수 있다.

$$F-ratio = \frac{variance\ of\ speaker\ means}{mean\ intraspeaker\ variance}$$

좋은 특징 파라미터는 화자간의 변이는 크고 화자내의 변이가 작은 것이다. 즉, F-ratio의 차수에 대한 분포가 일정하거나 변화가 작다면 개인성 정보는 캡스트럼 차수에 대해 거의 동일하게 분포되어 있다는 것을 의미한다. 이것은 화자인식이 어렵다는 것을 나타낸다. 따라서 차수에 따른 F-ratio 값의 변화가 심하게 나타난다면 특정 차수의 값이 화자인식에 매우 유용하게 적용될 수 있다[6][9]. 본 논문에서는 이러한 특성을 갖는 F-ratio를 적용하였다.

한편 기존의 DP 알고리즘을 이용한 화자인식 방법은 시간축의 변이가 커질수록 누적되는 오차가 커지기 때문에 인식률이 감소된다는 단점이 있다. 또한 처리시간이 증가한다는 문제점도 있다. 따라서 본 논문은 이러한 문제점을 해결하기 위해 음성신호의 안정된 구간에서 일정하게 반복되는 파형을 삭제한다. 이렇게 함으로써 인식률을 향상시킬 뿐만 아니라 데이터량 또한 줄일 수 있어 패턴정합시에 처리시간이 단축된다.

3. 안정구간 파형 삭제

동일 화자가 같은 단어를 발성할 경우라도 발성 시간은 변경된다. 발성시간 변경율이 클수록 화자인식에서의 인식율은 저하되게 된다. 그런데 같은 어휘의 발성시간이 늘어날수록 일정하게 반복되는 피치주기가 연속적으로 나타나게 되는 안정구간이 존재함을 알 수 있다. 따라서 이러한 안정구간에서 일정하게 반복되는 음성파형을 제거하면 불필요한 음성데이터를 줄일 수 있다. 일정하게 반복되는 파형을 제거하기 위해서 먼저 음성신호의 피치 주기를 검출해야한다. 본 논문에서는 음성신호의 양자화 오차를 구한 뒤 정규화된 AMDF법을 이용하여 피치주기를 검출하였다. 검출된 피치주기에서 동일한 피치주기값이 검출되면 삭제한 후 남은 새로운 음성파형을 추출하여 패턴정합에 사용한다.

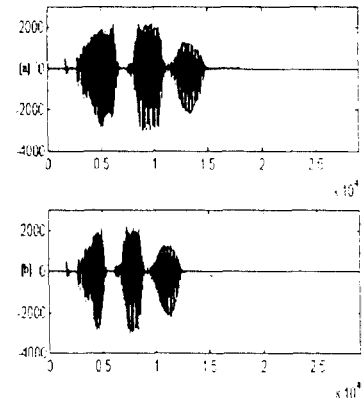

그림 3-1 음성신호의 안정구간 삭제 예
(음성 발성 : /배 재 옥/)
(a) 원음성
(b) 안정구간내 일정파형이 삭제된 음성파형

그림 3-1은 음성신호의 안정된 구간에서 일정파형을 삭제하고 남은 파형을 나타내고 있다. 그림에서 보는 바와 같이 원래 음성신호보다 데이터량이 감소했기 때문에 기존의 화자인식시스템에서 지속시간이 늘어남에 따른 처리시간 증가를 피할 수 있다.

4. 실험 및 결과

제안한 방법을 시뮬레이션하기 위해 IBM-PC MMX/200에 마이크가 장치된 16-bit A/D변환기를 사용하였다. 실험은 일반 실험실 환경에서 실행하였다. 음성 시료는 11kHz로 샘플링하고 16bit로 양자화하였다. 화자인식을 위한 특징 벡터로 14차 LPC-캡스트럼계수를 사용하였고 이를 추출하기 위해 한 프레임의 길이는 256샘플, overlap은 128샘플로 하였다. 특징벡터들의 비교방법으로는 DP 알고리즘을 사용하였다. 기준 패턴으로는 20대 남, 여 10명이 발성한 음성으로 구성하였고, 비교 패턴으로는 동일한 화자 10명이 각각 본인의 이름을 10번씩 발성한 것으로 하였다. 또한 사칭자에 대한 에러율을 측정하기 위해 5명의 사칭자가 10번씩 다른 사람의 이름을 발성하였다. 위와 같은 방법

을 C언어로 구현하여 실험하였다.

그림 4-1은 본 실험에서 사용한 화자인식의 전체 블럭도이다. 우선 입력된 음성 데이터에 대해 음성구간을 검출하고 검출된 구간에서 정규화된 양자화 오차신호를 구한다. 그리고 이 신호로부터 정규화된 AMDF법을 적용하여 피치주기를 검출한 뒤 안정된 음성구간에서 일정하게 반복되는 음성파형을 제거하고 남은 음성신호에서 인식을 수행하기 위한 특징벡터를 추출하였다. 그리고 비교패턴과 미리 저장된 기준패턴들과 DP알고리즘을 사용하여 인식을 수행하였다. 본 논문에서 제안한 방법의 성능을 측정하기 위해 기존의 인식방법과 제안한 방법의 실험결과를 표4-1과 표4-2에 나타내었고 표4-3은 처리시간을 나타내었다. 실험결과 제안한 방법의 인식률이 기존의 방법보다 1% 향상되었고, 사칭자의 거부능력이 1.5% 향상되었다. 또한 처리시간도 기존의 방법보다 21.6% 감소되었다.

표4-1 에러율(%)

	FA	FR
기존의 방법	4.5	3
제안한 방법	3	2.5

표4-2 전체 인식률(%)

	전체 인식률
기존의 방법	92.5
제안한 방법	93.5

표4-3 처리시간(sec)

	처리시간
기존의 방법	1.25
제안한 방법	0.98

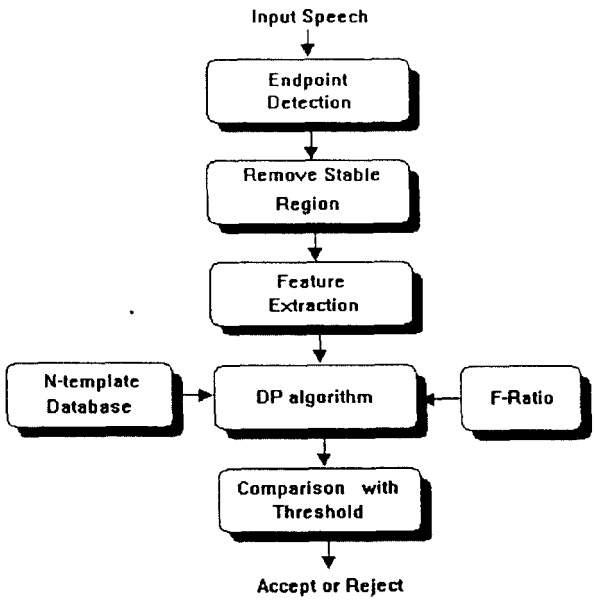

그림 4-1 본 논문에서 제안한 방법

5. 결 론

본 논문은 기존에 사용한 DP 알고리즘의 단점인 화자수 증가에 따른 처리시간의 증가를 보완하기 위한 것이다. 음성신호의 양자화 오차신호에서 추출된 피치주기를 이용해 안정된 음성구간에서 일정하게 반복되는 음성파형을 제거한 후 인식을 수행한다. 제안한 방법의 전체 인식율은 기존의 방법보다 1% 향상되었고, 처리시간은 21.6% 감소되었다.

6. 참 고 문 헌

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

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

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

[7] 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.

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

[9] H.Ney and R.Gierloff, "Speaker Recognition Using a Feature Weighting Technique," in Proc ICASSP'82, pp1645-1648, 1982

[10] X.D.Hunag, Y.Ariki, and M.A.Jack, *Hidden Markov Models For Speech Recognition*, Edinburgh University Press