음성의 묵음구간 검출을 통한 DTW의 성능개선에 관한 연구

A Study on the Improvement of DTW with Speech Silence Detection

김 종 국* · 조 왕 래* · 배 명 진*
Jong-Kuk Kim · Wang-rae Jo · Myung-Jin Bae

ABSTRACT

Speaker recognition is the technology that confirms the identification of speaker by using the characteristic of speech. Such technique is classified into speaker identification and speaker verification; The first method discriminates the speaker from the preregistered group and recognize the word, the second verifies the speaker who claims the identification. This method that extracts the information of speaker from the speech and confirms the individual identification becomes one of the most efficient technology as the service via telephone network is popularized. Some problems, however, must be solved for the real application as follows; The first thing is concerning that the safe method is necessary to reject the imposter because the recognition is not performed for the only preregistered customer. The second thing is about the fact that the characteristic of speech is changed as time goes by, So this fact causes the severe degradation of recognition rate and the inconvenience of users as the number of times to utter the text increases. The last thing is relating to the fact that the common characteristic among speakers causes the wrong recognition result. The silence parts being included the center of speech cause that identification rate is decreased. In this paper, to make improvement, We proposed identification rate can be improved by removing silence part before processing identification algorithm. The methods detecting speech area are zero crossing rate, energy of signal detect end point and starting point of the speech and process DTW algorithm by using two methods in this paper. As a result, the proposed method is obtained about 3% of improved recognition rate compare with the conventional methods.

Keywords: Speaker recognition, Speaker identification, DTW, ZCR, Energy

1. 서 론

현대가 정보화 사회로 급속히 진행됨에 따라 대규모의 데이터베이스에 등록되어 있는 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, 카드 등은 도난, 분실, 위조 등의 위험을 수반한다. 또한 전화나 통신망을 이용해서 정보 접근을 할 경우에 개인 확인이 더욱 어려워진다. 이에 반해 음성을 이용한 화자 식별 시

스템은 음성에 포함되어 있는 개개인마다의 화자정보를 추출하여 개인을 확인하는 기술로서 사칭자에 대한 처리, 처리시간, 원격자 확인 등 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. 화자 식별률은 화자수에 비례하여 정확도가 감소하므로 화자확인에 비해 어려우며, 실제 응용에서는 비협조적인 화자를 대상으로 하는 경우가 많으므로 화자의 정확한 판단에 어려움이 있다. 기존의 방법은 음성의 시작점과 끝점만을 검출하여 인식 알고리즘을 수행하는 방법을 택하고 있다. 이렇게 되면 비교할 음성데이터 중간에 포함 되어 있는 묵음구간이 인식률을 저하시키는 요인으로 작용하게 된다. 본 논문에서는 이를 개선하기 위하여 인식알고리즘을 수행하기 전에 묵음 구간을 제거함으로써 인식률을 개선하는 방법을 제안하였다.

2. 화자식별 시스템

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 4 가지로 구분할 수가 있다. 그 중 첫 번째는 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 수행하도록 하여 인식하는 신경회로망이 있다[9]. 그러나 이 방법은 새로운 화자의 추가 시 다시 학습시켜야 한다는 단점과 고도의 병렬계산 능력이 요구되기 때문에 실제 응용시에서는 적합치 않다. 세 번째 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[10].

일반적인 화자인식 시스템은 다음과 같은 3 가지로 크게 구분할 수가 있는데 그 전체적인 과정을 보면 다음과 같다. 우선 입력된 음성은 전처리 과정을 통해 디지털 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. 본 논문은 기존의 화자 인식보다 인식률을 향상시키는 방법에 관한 것으로 다음과 같은 특성을 이용한다. 일반적으로 화자가 한 문장을 발성을 할 때 호흡을 하기 위해서 중간에 발성

이 끊어지게 된다. 이 묵음 구간 때문에 화자식별 시스템이나 음성인식 시스템의 성능이 저하되게 된다. 기존의 DTW를 이용한 화자식별시스템에서는 이 묵음 구간을 제거하지 않고 인식 알고리즘을 수행하여 오인식률이 증가하는 문제점을 않고 있다. 따라서 본 논문에서는 화자식별에 사용되는 모든 음성 데이터에서 묵음 프레임을 제거함으로써 얻어지는 인식률에 대하여서 조사하였다.

3. 제안한 알고리즘

3.1 음성특징 추출

LPC 켑스트럼 계수에서 귀의 특성을 고려한 멜 스케일로 왜곡(warping)하여 특징벡터로 사용하는 경우가 선형 주파수로 스케일된 LPC 켑스트럼 계수보다 성능이 좋기 때문에, 선형 주파수로 스케일된 LPC 켑스트럼 계수를 멜 스케일로 변환하는 것이 바람직하다. 이에 따라 Bilinear Transform을 사용하여 LPC 켑스트럼을 변화시키는데, 사용되는 Bilinear Transform은 전대역 필터(allpass filter)를 사용하여 주파수를 변화시키는 방법이다. 이에 대한 변환 식은 다음과 같다[3].

$$Z^{-1}_{new} = \frac{z^{-1} - a}{1 - az^{-1}} \qquad (-1 < a < 1)$$

$$w_{new} = w + 2\tan^{-1}\left(\frac{a \sin w}{1 - a \cos w}\right) \tag{1}$$

여기서 ω는 정규화된 샘플링 주파수이고, w_{new}는 변화되는 주파수를 나타내고, α는 주파수 왜곡 파라미터이다. 여기서 파라미터 α가 양수이면 낮은 주파수에서 더 높은 분석력을 가지게 할 수 있고, 일반적으로 0.4<α<0.8의 범위에서 멜 스케일로 변환할 수 있다[11].

본 논문에서 사용하는 특징벡터는 멜 켑스트럼으로 Bilinear Transform을 이용하는 방법을 사용하였고 그 처리과정은 그림 1과 같다. 이 방법을 사용한 이유는 일반적으로 음성의 영역변환에는 계산량이 많고, 실제 화자 인식 시스템에서 특징벡터 추출은 전처리 단계이기 때문에 계산량의 부하가 적은 방법으로 후자의 방법을 사용하였다. 우선 음성신호를 해밍 윈도우(hamming window)에 통과한다. 윈도우 처리된 음성 신호로부터 Durbin 알고리즘을 이용하여 p차의 선형 예측 계수를 구한 후 같은 p차의 켑스트럼을 구하였고, Bilinear Transform을 이용하여 멜 켑스트럼을 구하였다. 또한 특징벡터를 추출한 후 대역통과 리프트 함수를 통과시켜 인식을 수행하였다.

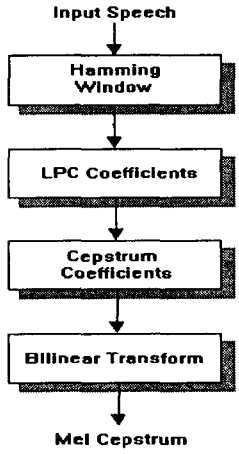

그림 1. 특징 벡터 추출 과정

3.2 에너지와 ZCR을 이용한 음성구간 검출

처음 5 개 프레임의 음성신호에 대한 평균 에너지를 구한다. 이후에 들어오는 음성신호에서 이 평균 에너지를 빼 줌으로써 PC(Personal Computer) 자체의 잡음과 마이크 입력시 생기는 DC잡음을 제거한다. 그리고 처음 입력된 5 프레임의 데이터를 잡음구간으로 가정하고 그 구간에서 에너지와 영교차율의 통계적 특성을 식 2부터 식 3을 이용하여 기준값을 정한다. 이때 에너지가 너무 커지는 것을 방지하기위해 문턱값의 최대값으로 10^5을 설정하였다.

$$E(fr) = \sum_{n=0}^{110} |data(n)|^2 \qquad fr = 1, 2, \ldots, 5$$

$$ThreshEL = \min(4 * mean(E(fr), 10^5) \tag{2}$$

$$ThreshEU = ThreshEL * 3$$

그리고 영교차율(zero crossing rate)은 원래 음성신호값이 0값을 교차하여 부호가 바뀌는 횟수를 말하지만 본 논문에서는 잡음의 영향을 줄이기 위하여 다음과 같이 정의한다. 먼저 ThreshEL에 비례하는 SilenceU와 SilenceL을 정하여 신호의 바이어스를 만들어 준 뒤, 이 영역을 교차하는 경우의 수를 사용하였다.

$$SilenceU = ThreshEL/10^5 + 1$$

$$SilenceL = -SilenceU$$

$$Z(fr) = \sum_{n=0}^{110} |sgn(data(110 \cdot fr + n)) - sgn(data(110 \cdot fr + n - 1))| \quad fr = 1, 2, \ldots, 5 \tag{3}$$

$$sgn(data(n)) = \begin{pmatrix} 1, & \text{for } data(n) \geq SilenceU \\ -1, & \text{for } data(n) \leq SilenceL \end{pmatrix}$$

$$ThreshZCR = \min(mean(zcr) + 2 \cdot \sigma_{zcr}, 25)$$

이와 같이 설정된 에너지와 영교차율의 임계치를 이용하여 음성의 끝점검출을 수행할 수 있는데 그 과정은 다음과 같다[4]. 우선 현재 프레임의 에너지를 구하고 임계값을 넘으면 확실한 음성 존재구간이라고 가정한다. 그리고 임계값이 넘지 않는 프레임에서 영교차율을 구하여 무성음과 묵음을 구별한다. 제안한 알고리즘에 대한 블럭도는 다음과 같다.

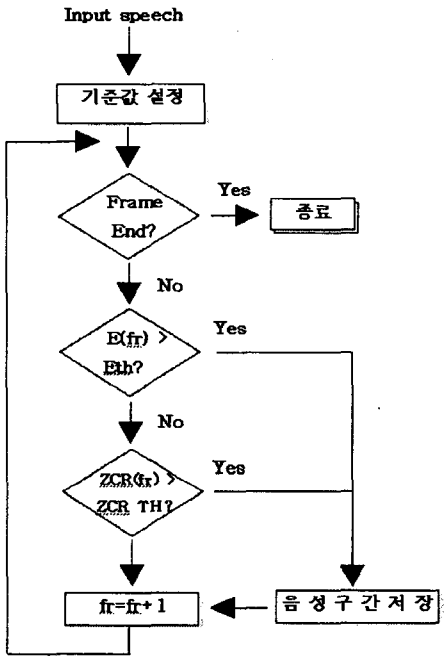

그림 2. 제안한 알고리즘의 블록도

4. 실험 및 결과

본 논문의 알고리즘을 실험하기 위해 IBM PC에 마이크가 장치된 16 비트 A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 20 명의 남녀 화자가 각각 본인의 이름을 발성한 음성 시료를 11 khz로 샘플링하고 16 비트로 양자화하여 사용하였다. 한 프레임의 길이는 512 샘플이며, 256 샘플씩 오버랩(Overlab)시켜 특징벡터를 추출하였다. 인식을 위한 특징벡터로는 14 차 Mel-Cepsturm을 사용하였다. 논문에서 사용되는 텍스트 종속용 데이터베이스는 5초 정도의 한 문장이며 등록자의 수는 55 명이다. 사칭자의 효과를 알아보기 위해서 테스트 화자의 데이터베이스는 각 1 번씩 2 회(110 개), 등록된 화자 중 6 명 1 주일 동안 2 번씩 5 회(60 개) 발성하였다. 본 논문에서 제안한 방법은 동일 화자 동일 음절이라 하더라도 묵음이 존재하는 범위가 다르기 때문에 인식률이 저하되는 것을 막기 위한 것이다. 먼저 음성을 입력받고 독립적인 음성 검출방법을 이용하여 묵음 구간을 제거하였다. 이를 가지고 특징 벡터추출을 한 후 인식 알고리즘을 수행하여 결과를 얻었다. 제안한 알고리즘의 흐름도는 그림 3과 같다.

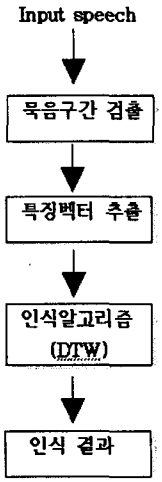

그림 3. 제안한 알고리즘의 흐름도

본 논문에서 제안한 방법의 효과를 알아보기 위해서 기존의 음성구간을 검출하지 않고 하는 방법과 제안한 알고리즘을 적용했을 때의 전체 인식률을 비교해 보았다. 인식률은 170 개 중 20 개(약 12%)가 오인식이 되었고, 제안한 알고리즘은 170 개 중 15(약 9%) 정도가 오인식 되었다. 실험 결과는 아래 표 1과 같으며 처리시간 31 명을 기준패턴으로 놓고 인식을 수행하였을 때 약 10%정도 감소하였다. 처리시간은 표 2와 같다.

표 1. 전체 인식률

	기존의 방법	제안한 방법
오인식률	12%	9%

표 2. 처리시간

	기존의 방법	제안한 방법	감소율
오인식률	1.21	1.09	10%

5. 결 론

화자 인식은 음성의 특성을 이용해서 화자의 신원을 확인하는 기술이다. 이러한 기술은 등록된 화자집단 중 화자를 식별하는 화자 식별(speaker identification)과 지금 발성한 화자만을 비교하여 확인하는 화자 확인(speaker verification)이 있다. 두 방법의 가장 확실한 차이는 화자 확인에서 사용되는 문턱값의 차이이다. 그러나 실제 경우에 있어서 높은 화자 식별률이 높은 화자 확인율을 얻게 하고, 그 반대의 경우도 마찬가지이기 때문에 이러한 차이점은 모호해진다. 이러한 화자 인식은 음성에 내재되어 있는 화자정보를 추출하여 개인을 확인하는 기술

로 전화망을 통한 서비스가 확산되어 가고 있는 현대사회에 가장 효과적인 기술 중 하나이다. 화자 식별은 입력된 미지의 음성이 이미 등록된 여러 명의 화자 중 어떤 화자에 의해 발성된 음성인지를 판정하는 방법이고, 화자 확인은 입력된 음성이 본인의 것인지의 여부를 판정함을 말한다. 따라서 화자 식별율은 화자수에 비례하여 정확도가 감소하므로 화자 확인에 비해 어려우며, 실제 응용에서는 비협조적인 화자를 대상으로 하는 경우가 많으므로 화자의 정확한 판단에 어려움이 많다. 또한 비교할 음성 데이터 중간에 포함되어 있는 묵음 구간 또한 인식률을 저하시키는 요인으로 작용하게 된다.

본 논문에서는 이를 개선하기 위하여 인식 알고리즘을 수행하기 전에 묵음 구간을 제거함으로써 인식률을 개선하는 것을 제안하였다. 음성구간을 검출하는 방법에는 영교차율, 신호의 에너지, 1차 자기 상관 계수, 선형 예측 계수, 예측 오차 에너지, LSP 분포도, AMDF의 기울기 등 다양하게 존재하며 본 논문에서는 신호의 에너지와 영교차율을 이용하여 음성구간을 검출하였다. 기존의 방법에 비하여 약 3%정도 개선이 되었다. 실험결과로 보아 음성데이터 중간에 있는 묵음구간이 인식률에 많은 영향을 미치는 것을 알 수 있다. 현재 실제 환경에 적용하기 위해서는 잡음에 강한 음성구간 검출 방법에 대하여서 연구를 진행하고 있는 중이다.

참 고 문 헌

[1] 정종순, 배재옥, 배명진. "윈도우 환경에서 음성을 이용한 사용자 확인에 관한 연구." 한국음향학회지, Vol.17, No.5. 1998.

[2] Furui. S., *Digital Speech Processing, Synthesis and Recognition.* Marcel Dedder. Inc., 1992.

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

[4] 구명완 외. "실시간 음성 끝점 검출 알고리즘." 제 5회 신호처리 합동학술회 논문집, 제 5권 1호, pp.11-14. 1992.

[5] 배재옥, 오세영, 배명진. "F1/F0율을 이용한 화자인식의 성능 향상에 관한 연구." 한국음향학회 학술발표대회 논문집 Vol. 16. No. 2(s), pp.137-140. 1997년 11월.

[6] Furui, Sadaoki. "Cepstral Analysis Technique for Automatic Speaker Verification." IEEE Trans. on ASSP, vol.29, No.2, pp.254-272. Apr. 1981.

[7] Mammone, Richard J., Xiaoyu Zhang, and Ravi P. Ramachandran. "Robust Speaker Recognition." IEEE Sognal Processing Magazine, pp.58-71. Sep. 1996.

[8] Soong, Frank K., Rosenberg, Aaron E.. "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition." IEEE Trans. on ASSP, vol.36, No.6, pp.871-879. Jun. 1988.

[9] 김순협. "음성인식 기술 현황." 한국통신학회지, 제 11권 제 9호. 1994년.

[10] Rabiner, L. R. and Juang, B. H. *Fundamentals Of Speech Recognition.* Prentice-Hall, AT&T, U.S.A. 1993.

[11] Furui, S. and Sondhi. *Advances in Speech Signal Processing.* Dekker.

접수일자: 2003. 11. 3.
게재결정: 2003. 12. 15.

▲ 김종국
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 정보통신공학과 음성통신연구실
Tel: +82-2-824-0906
E-mail: kokjk@hanmail.net

▲ 조왕래
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 정보통신공학과 음성통신연구실
Tel: +82-2-824-0906
E-mail: wrjo@unitel.co.kr

▲ 배명진
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 정보통신공학과 음성통신연구실
Tel: +82-2-820-0902
E-mail: mjbae@ssu.ac.kr