Journal of the Korea Academia-Industrial cooperation Society
Vol. 12, No. 5 pp. 2402-2409, 2011

DOI: 10.5762/KAIS.2011.12.5.2402

동적 시간 신축 알고리즘을 이용한 화자 식별

정승도[1*]
[1]한양사이버대학교 정보통신공학과

Speaker Identification Using Dynamic Time Warping Algorithm

Seungdo Jeong[1*]
[1]Department of Information and Communication Engineering, Hanyang Cyber University

요 약 음성에는 전달하고자 하는 정보 이외에 화자 고유의 음향적 특징을 담고 있다. 화자간의 음향적 차이를 이용하여 말하고 있는 사람이 누구인지 판단하는 방법이 화자 인식이다. 화자 인식에는 화자 확인과 화자 식별로 구분되는데 화자 확인은 1명의 음성을 대상으로 본인인지 아닌지를 검증하는 방법이다. 반면, 화자 식별은 미리 등록된 다수의 종속 문장으로부터 가장 유사한 모델을 찾아 대상 의뢰인이 누군지 식별하는 방법이다. 본 논문에서는 MFCC(Mel Frequency Cepstral Coefficient) 계수를 추출하여 특징 벡터를 구성하였고, 특징 간 유사도 비교는 동적 시간 신축(Dynamic Time Warping) 알고리즘을 이용한다. 각 화자마다 두 개의 종속 문장을 훈련 데이터로 사용하여 음운성에 기반을 둔 공통적 특징을 기술하였고, 이를 통해 데이터베이스에 저장되어 있지 않은 단어를 사용하더라도 동일 화자임을 식별할 수 있도록 하였다.

Abstract The voice has distinguishable acoustic properties of speaker as well as transmitting information. The speaker recognition is the method to figures out who speaks the words through acoustic differences between speakers. The speaker recognition is roughly divided two kinds of categories: speaker verification and identification. The speaker verification is the method which verifies speaker himself based on only one's voice. Otherwise, the speaker identification is the method to find speaker by searching most similar model in the database previously consisted of multiple subordinate sentences. This paper composes feature vector from extracting MFCC coefficients and uses the dynamic time warping algorithm to compare the similarity between features. In order to describe common characteristic based on phonological features of spoken words, two subordinate sentences for each speaker are used as the training data. Thus, it is possible to identify the speaker who didn't say the same word which is previously stored in the database.

Key Words : Speaker Identification, MFCC, Dynamic Time Warping

1. 서론

음성에는 언어정보 외에 화자의 신원을 파악할 수 있는 정보도 포함되어 있다. 화자의 발성 습관 또는 화자의 성도의 구조적인 차이 등이 이러한 정보로 볼 수 있으며, 이를 이용하여 발성한 화자가 누구인지 파악할 수 있는 화자 식별(speaker identification)이 가능하다[2]. 화자 인식은 사용하고자하는 성격에 따라 화자 식별과 화자 확인 두 분야로 구분할 수 있다. 화자 식별은 신원 확인이 필요한 경우 미리 등록되어 있는 다수의 화자 중에 발성을 하는 당사자가 누구인지를 판단하는 것이다. 화자 확인에 관한 분야는 화자 한명을 대상으로 맞는지 아닌지를 검증하는 것이다. 즉, 화자 확인은 한 명의 화자를 대상으로 하는 작업인 반면에 화자 식별은 일반적으로 다수의 화자를 대상으로 이루어진다.

화자 식별은 특정 음성 입력에 대하여 미리 등록된 다

*교신저자 : 정승도(sdjeong@hycu.ac.kr)
접수일 11년 04월 20일 수정일 11년 05월 07일 게재확정일 11년 05월 12일

수의 화자 모델과 유사도를 조사하여 가장 유사한 모델을 찾음으로써 수행되며, 화자 식별에 기반한 응용분야는 매우 다양하다. 일례로 자동 녹음 음성 회의록과 같은 다수의 화자의 음원이 겹칠 수도 있는 상황에서 특정 화자에 대한 음성을 자동으로 분할하고자 할 때도 활용 될 수 있다.

화자를 인식하는 시스템은 문맥 종속형 화자 인식 시스템과 문맥 독립형 시스템으로 나뉘는데 본 논문에서는 화자가 발성할 문장이 고정되어 있는 문맥 종속형 화자 인식 시스템을 다룬다. 이는 음운성에 기반을 둔 공통적 특징의 개인 차이를 평가하므로 음성인식과 거의 동일한 방법으로 인식기를 구성한다. 그리고 선정된 어휘나 문맥에 따라서 인식률에 많은 차이가 발생하기 때문에 화자의 특성이 잘 나타나는 모음, 비음 등의 음운이 균형을 이룬 어휘나 문맥을 선택하는 것이 중요하다.

본 논문에서는 화자 식별을 위해 MFCC를 사용하여 12차 계수를 추출하여 특징 벡터로 사용하고 동적 시간 신축(Dynamic Time Warping, DTW) 알고리즘을 사용하여 유사도를 계산한다. 같은 단어를 서로 다른 화자가 발성한 음성에는 길이 차이가 존재 하는데 이를 단순하게 표준 패턴과 비교할 수 없기 때문에 음성신호의 패턴과 입력된 음성 신호간의 거리를 동적 프로그래밍을 이용해 구하는 DTW 알고리즘을 사용한다[2].

DTW의 성능을 결정하는 요소들은 많이 있다. 그 중 중요한 요소로 끝점(end point), 국소 연속성 제한(local continuity constraints), 전역 경로 제한(global path constraints), 축 회전(axis orientation), 거리 측정(distance measure)을 들 수 있다[3]. 본 연구에서는 화자식별에 관한 인식 시스템에 관해서 다루고 있으며 5명의 음원 정보를 사용하여 입력 정보의 대상자가 맞는지 아닌지를 판단한다.

2. 관련 연구

화자 인식에 관한 연구는 크게 3가지 형태로 분류할 수 있다. 첫째는 인식률 향상을 목적으로 하는 것으로, 효율적인 인식단위의 설정에 관한 연구, 인식기의 사용목적에 가장 적합한 특징벡터의 추출에 관련된 연구, 학습 및 인식 알고리즘의 변형을 통한 인식률 향상에 관한 연구 등이 수행된다. 둘째, 환경에 강인한 인식기 구현에 관한 연구로는 이동통신, 전화와 같은 왜곡된 신호의 환경에서 강인한 인식기의 구현, 잡음 환경에서의 강인한 인식기를 구현하기 위한 방안, 그리고 소규모의 데이터베이스에 효과적이며, 우수한 성능을 보이는 인식기 구현에 관한 연구들이 진행이 되고 있다. 마지막으로 음성 인식에 기반한 응용분야로써 로봇 및 기계 제어를 위한 음성 인식, 사용자 인터페이스, 인간과 컴퓨터 상호 작용을 위한 응용 등에 널리 연구되고 있으며 원거리 인식 및 네트워크를 통한 인식 등에 관한 연구 등도 수행되고 있다[3].

3. DTW를 이용한 화자 식별

화자 식별을 하기 위한 첫 번째 단계는 화자가 발성한 음절의 음성에서 특징을 추출하는 것으로 인식에 유용하고 분별력을 갖춘 특징 성분을 음성 신호로부터 뽑아내는 과정이라 할 수 있다. 음성 인식에 주로 사용되는 특징은 LPC, MFCC, PLP, 필터뱅크 에너지 등이 있다[3]. 최근에는 잡음에 강하고 인간의 청각적 특성을 고려한 MFCC, PLP가 주로 사용된다. 본 논문에서도 역시 인간의 청각적 특성을 고려한 MFCC를 사용하여 특징 벡터를 추출하고 있으며, MFCC로 부터 뽑아낸 특징 벡터들을 이용해서 계수들에 의한 파형을 얻는다.

같은 단어를 발음한다고 하더라도 개인별로 음성길이의 차가 생기기 마련이며 음절 마다 길이도 모두 다르다. 따라서 이러한 차이가 존재하는 특징 벡터에 대해서 유사도를 계산하기 위한 방법으로 DTW 알고리즘을 활용하고자 한다.

3.1 MFCC

Mel Frequency Cepstral Coefficient (MFCC) 계수를 얻기 위해서는 Pre-emphasis, framing, windowing, FFT, Mel-filtering, Log 함수 적용, 그리고 DCT의 순서로 총 7단계를 거치게 된다. 이는 입력 음성으로부터 불필요한 부분들을 제거하고 시간 축을 주파수로 변환하는 과정을 거쳐 인간이 들을 수 있는 범위로 LOG화 시킨 후DCT을 거쳐서 최종적인 12차의 MFCC 계수를 얻는 것이다. 본 절에서는 MFCC 각각의 단계별 처리 과정을 간략히 살펴본다.

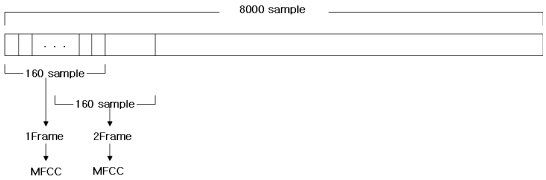
3.1.1 Pre-emphasis

음성 신호를 처리하기 위하여 가장 먼저 고대역 통과 특성을 갖는 디지털 Pre-emphasis 필터를 사용한다. 이 필터를 사용하는 이유는 다음과 같다. 첫째, 인간의 외의나 중이의 주파수 특성을 모델링 하기 위한 것으로 입술에서의 방사에 의하여 20dB/decade로 감쇄되는 것을 보상하게 되어 음성으로부터 성도 특성만을 얻게 된다. 둘째, 청각시스템이 1khz 이상의 스펙트럼 영역에 대해 민

감한 것을 어느 정도 보상하게 된다. 즉, 음성의 고주파를 강조시켜주는 과정으로 원하지 않는 소음을 제거 할 수 있다.

3.1.2 Framing

본 논문에서 8bits 모노 8000hz 샘플링으로 녹음한 음성 신호를 대상으로 한다. 본 스펙으로 음성 신호를 1초 녹음할 경우 샘플의 개수는 8000개가 된다. 이것을 한번에 FFT한다면 1초당 주파수 성분은 나타나지만 시간적인 정보는 나타나지 않게 되기 때문에 결과에 의미를 두기가 어렵다. 따라서 시간적인 정보를 고려한 짧은 구간의 음성을 취해 FFT를 하고 MFCC를 구해야 한다. 이를 위해 음성 구간을 나누는 것이 Framing이다. 프레임은 짧은 음성 구간으로 하나의 구간 내에서 특징 변화가 크게 발생하지 않고 고정된 특징을 갖출 수 있는 구간이어야 한다. 이에 본 논문에서는 20ms를 기준으로 음성 신호를 분할하였다. 20ms 구간의 샘플은 8000*20/1000(160)이 되고 이러한 구간 샘플 160개가 하나의 프레임이 된다. 프레임을 구분할 때 일정한 간격별로 구분을 할 경우 경계 영역에 걸쳐 손실되는 정보가 존재할 수 있기 때문에 이러한 누락 정보를 최소화하기 위해서 그림 1과 같이 절반이 겹치도록 프레임을 생성해야 한다[5].



[그림 1] MFCC 생성을 위한 프레임 구성

3.1.3 Windowing

프레임으로 음원을 잘랐을 때 불연속성을 보완하기 위해 윈도우를 사용한다. 윈도우는 해밍윈도우를 사용했으며 윈도우를 사용한 다는 것은 식 (1)과 같이 음성 샘플에 계수를 곱하는 것이다.

$$W(n) = H(n) \times (c_1 - c_2 \cos(\frac{2n\pi}{N-1})),\ 0 \le n \le N-1 \quad (1)$$

여기서 c_1, c_2는 해밍 윈도우 계수로 각각 0.54와 0.46을 사용하였다.

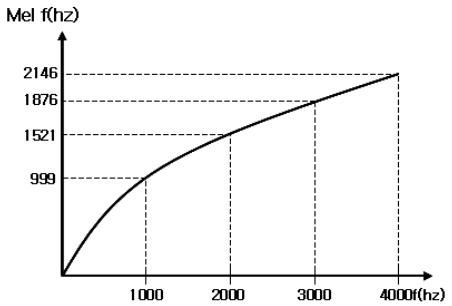
3.1.4 FFT

이전 과정에서 8khz로 샘플링 했을 때 160개의 샘플을 프레임으로 구성하였다. FFT를 하기 위해서는 샘플의 개수가 2의 제곱개가 되어야 하므로 160개의 샘플을 256개의 샘플로 확장한다. 256 샘플로 확장하기 위하여 zero padding을 수행하여 96개 샘플을 추가하였고, 총 256 샘플에 대해서 FFT를 구현하였다. FFT 결과는 대칭으로 나타나기 때문에 최종적으로는 주파수 변환된 길이의 절반만 사용한다.

3.1.5 Mel-filtering

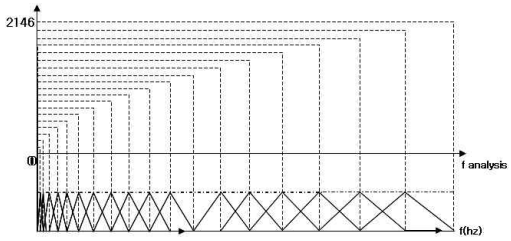
사람의 귀에서 수행되는 음성 주파수 성분에 대한 인식은 선형 스케일을 따르는 것이 아니라 Mel-frequency 스케일을 따르기 때문에 이에 대한 처리가 필요하다. 즉, FFT를 통해 복소수 형태의 주파수로 변환된 신호를 선형 스케일이 아닌 Mel-frequency 스케일로 해석하여야 한다. 예를 들어, 1khz이하는 선형적 그래프를, 1khz 이상은 식 (2)로부터 그림 2와 같은 로그 형태의 그래프를 얻을 수 있다.

$$F_{mel} = 2591\log(1 + \frac{f_{Hz}}{700}) \quad (2)$$



[그림 2] Mel-Filtering 그래프

그래프를 토대로 보면, 사람의 귀는 1khz이하의 주파수가 들어올 때는 선형적으로 해석하고 그 이하일 때는 로그 스케일로 해석하게 된다. 이러한 이유 때문에 FFT를 통해 해석 주파수 영역으로 변환된 음성 신호는 사람의 귀의 특성에 맞는 Mel-frequency 스케일로 분석해야 하는 것이다. 먼저 Mel 주파수 영역을 19개로 분할한다. 여기서의 19개는 필터 뱅크의 수가 된다. 분할 지점에 해당하는 주파수는 중심 주파수라고 한다. 다음으로 그림 3과 같이 19개로 나누어진 중심 주파수를 x축으로 대응시켜보면 FFT에서 해석주파수와 정합된다.



[그림 3] 중심주파수와 FFT 해석주파수와 매칭

그 다음으로 Mel-frequency 주파수 12개의 특징을 구해야 한다. 이는 매칭 된 해석 주파수의 크기 값(magnitude)에 대하여 그림 3과 같은 삼각형 모양의 가중치가 부여된 총 합으로서 하나의 필터 뱅크 에너지가 되고 이를 특징으로 사용한다. 정합된 중심주파수는 가중치 1이 곱해지고 그 주위로는 1보다 작은 값이 곱해져 각각이 더해지게 된다. 이렇게 12개의 필터 뱅크 에너지가 구해지게 되고 이는 19개로 분석한 Mel-frequency의 특징이 된다.

3.1.6 LOG

이전 과정에서 구해진 19개의 필터뱅크의 출력 에너지를 그냥 사용하는 것이 아니라 로그를 취하게 된다. 그 이유는 우리의 귀가 소리의 크기에 대해 로그 함수로 느끼기 때문이다.

3.1.7 Discrete Cosine Transform

MFCC를 얻기 위한 마지막 단계는 Discrete Cosine Transform(DCT)이다. DCT의 역할은 첫째, 필터 뱅크의 출력간의 상관관계를 없애주고 파라미터의 특징을 모아준다. 둘째, DCT는 결과 값이 실수로 이루어질 뿐만 아니라 DCT에 의한 결과 벡터 값들은 상호 독립적이기 때문에 계산상 효율성도 가지고 있다. 신호를 공간 혹을 시간 도메인에서 주파수 도메인으로 변환시켜 줄 때 신호정보들은 저주파에 집중된다. 그리하여 신호정보에 기여도가 적은 고주파 성분은 버리고 대부분의 신호 정보들을 담고 있는 저주파 성분만을 이용하여 신호의 특성을 표현 한다. 본 논문에서도 역시 저주파 성분인 MFCC 계수 12개를 사용하였다. DCT 이후에 얻어지는 12개 MFCC 계수는 식 (3)과 같이 표현된다.

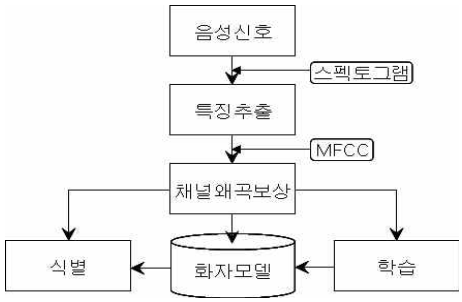
$$C_n = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right) \frac{\pi}{m_p} \right] \right\}, n = 1, \ldots, 12 \quad (3)$$

여기서 $\log(S_k)$는 k번째 필터 뱅크 에너지를 의미하고, m_p는 필터 뱅크의 총 수를 나타낸다.

또한, 12개의 특징과 별도로 19개의 필터 뱅크의 로그 에너지의 합을 구하고 이를 추가적으로 사용하여 화자 인식을 위한 특징 벡터로 사용하였다.

3.2 동적 시간 신축(DTW)

음성 인식처럼 음성 신호가 입력되면 실제 화자가 발성한 실 음성 부분을 분리하고, 화자 식별에 이용될 특징 벡터를 추출한다. 그 후 사용되는 환경에 따른 채널 왜곡을 보상해 주어야 하는데, 대부분의 화자 식별기의 응용분야는 네트워크나 전화망을 통하여 음성 신호가 전송되므로 이에 맞는 왜곡 보상이 반드시 필요하다. 화자 식별을 구현하기 위한 전체 과정은 그림 4와 같다.



[그림 4] 화자 인식 과정

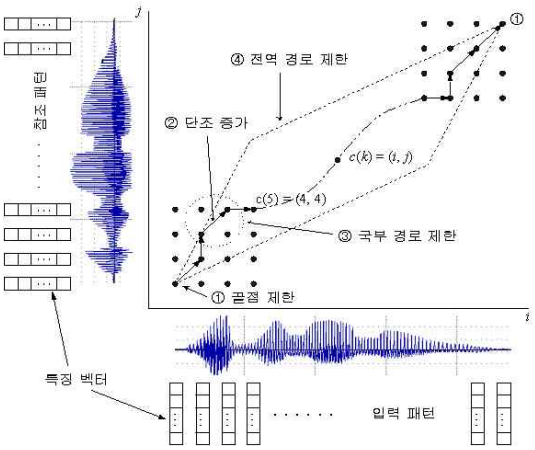
동일인이 동일한 단어를 발성해도 발성할 때 마다 단어의 시간적 길이가 다르다. 이를 표준 패턴과 단순히 비교하면 시간축이 고르지 않기 때문에 오인식의 원인이 된다. 이러한 문제를 해결하기 위해서는 시간 축에 대하여 정규화가 필요하다.

기존에 사용되던 시간 축 정규화 방법은 선형 신축 방법으로 두 패턴 길이를 동일하게 맞춰주는 정규화 방법이다. 그러나 음성은 시간에 따라 신축정도가 비선형적이고 모음과 자음의 신축 정도가 다르기 때문에 선형 신축으로는 정확한 비교가 어렵다. 이를 해결하기 위한 방법이 시간 축의 비선형 신축에 의한 정합법인 동적 시간 신축(DTW) 알고리즘이다. 그림 5는 하나의 입력 패턴과 참조 패턴이 정합되어지는 비선형 함수를 나타낸다. 동적 시간 정합 법은 서로 다른 두 개의 자료에서 최적의 정합 경로를 찾아 두 자료를 비교 할 수 있는 방법이다.

3.2.1 DTW 알고리즘

음성은 특징 벡터 열로서 특징을 추출하여 표현 할 수 있는데 길이가 I, J 인 음성신호 A, B의 특징 벡터는 식

(4)와 같이 벡터 열로 나타낼 수 있다[3].



[그림 5] 입력 패턴과 참조 패턴의 정합 과정

$$A = a_1, a_2, a_3, \ldots, a_I \quad (4)$$
$$B = b_1, b_2, b_3, \ldots, b_J$$

이 열들은 패턴 A의 시간 축으로부터 패턴 B로 정합을 하기 위한 하나의 함수로써 표현할 수 있다. 이것을 warping 함수라 한다. 두 음성 패턴 A, B를 각각 i, j 축에 놓을 때 서로 정합시켜 주는 점을 c(k)라 하면, warping 함수 F는 식 (5)와 같고 그 과정을 그림 5에서 보였다.

$$F = c(1), c(2), c(3), \ldots, c(k), \ldots, c(K) \quad (5)$$

여기서 c(k)는 i, j에서 두 패턴 간의 차이로서, c(k)=(i(k), j(k))와 같이 표시된다. 이들 패턴 간의 시간차가 없을 경우 warping 함수는 대각선 i-j에 일치하고 시간차는 이 대각선으로부터 유도된다. 두 특징벡터 a, b 사이의 거리는 식 (6)과 같이 구할 수 있다.

$$d(c) = d(i,j) = \|a_i - b_j\| \quad (6)$$

warping 함수 F상에서의 가중치 합의 거리는 식 (7)과 같이 구할 수 있다. 여기서 K는 warping 함수 F상에서의 점들의 수를 나타낸다.

$$E(F) = \sum_{k=1}^{K} d(c(k))w(k) \quad (7)$$

벡터 열 A, B를 정합시키는 것은 두 패턴의 차이 값을 최소화 하도록 warping 함수를 찾는 것이다. A, B에서 시간 정규화된 거리는 식 (8)과 같이 나타낼 수 있다. w(k)는 가중치 계수이고 E(F)의 탄력성 있는 특성을 유도하는데 도입되며 적절한 warping 함수 F를 찾는데도 이용한다. warping 함수 F는 시간차를 최적인 상태로 맞춤으로써 구할 수 있다.

$$D(A,B) = \min_{F}\left[\frac{\sum_{k=1}^{K} d(c(k))w(k)}{\sum_{k=1}^{K} w(k)}\right] \quad (8)$$

여기서, 분모는 warping 함수 F에서 점의 개수 k에 의한 영향을 보상하기 위해 사용된다. 이렇게 시간 축 상에 정렬된 두 벡터 열은 서로 비슷한 음소끼리 정합되어진다고 볼 수 있다. 하지만 음성 신호의 경우 기울기의 제한 없이 지나치게 차이가 나는 두 패턴을 정합시키면 실제적으로 맞지 않는 warping 될 가능성이 있기 때문에 warping 함수의 기울기는 어느 정도의 제한을 두는 것이 바람직하다. 음성 신호에 있어서 같은 발음에 대해 지나치게 많은 차이가 나지 않기 때문에 기울기에 제한을 두는 것이 인식률을 높일 수 있다. 그림 6에서 DTW 알고리즘 개요를 보였다.

3.2.2 채널 왜곡 보상

채널 왜곡을 보상하는 방법으로는 긴 구간의 음성에서 그 왜곡을 일정한 값으로 가정하고 그 값을 차감하는 캡스트럼 평균 차감법, 최대 우도법(Maximum Likelihood)으로 채널 왜곡을 추정하는 신호 편차 제거법, 채널 왜곡을 추정하는 캡스트럼 선형 변환법 등이 있다[5]. 캡스트럼은 음성의 특징을 나타내는 성도와 성대의 특징 함수들이 선형적인 합으로 나타나므로 화자 정보와 언어 정보를 구분하기가 쉽다. 또한 캡스트럼을 차분, 차분 값의 차분하여 얻는 파라미터와 멜 캡스트럼과 같은 인간의 청각 특성을 고려한 비 균일 대역 캡스트럼은 잡음 환경에 강인한 것으로 알려져 있다. 본 논문에서는 채널 왜곡 보상을 위하여 캡스트럼 평균 차감법을 사용했으며 특징 벡터로는 여러 개의 연속 음절 인식을 동시에 수행하기 위하여 멜 주파수 캡스트럼 계수와 그 차분 값과 차분 값의 차이를 사용하였다[4].

1. 초기화 조건

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

2. 동적 프로그래밍

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

3. 시간 축으로 정규화된 거리

$$D(A,B) = \frac{1}{N} g_k(c(k))$$

$$where\, c \quad c(0) \;\; means \;\; c(0,0)$$

[그림 6] DTW 알고리즘 개요

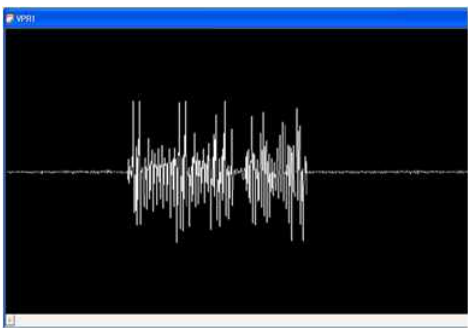
4. 성능 평가

4.1 실험 환경

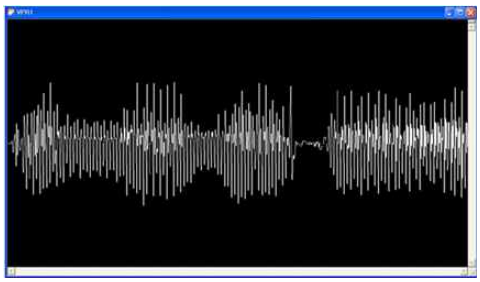
총 5명의 실험자를 대상으로 "안녕하세요.", "테스트 중입니다." 두 가지 문구를 빠르기 3단계, 음정 3단계로 각각 9가지 샘플을 만들었다. 각 실험자마다 "아" 발음을 길게 하여 음 높이를 고·중·저로 하여 3개씩 두어 총 샘플 수는 화자 당 11개이다. 데이터베이스에 사용한 데이터와 테스트에 사용한 데이터 모두 PC용 Mic를 이용하여 8bits 모노 웨이브 파일로 녹음하고 사용하였다.

4.2 MFCC

그림 7은 특정 화자의 "안녕하세요" 음성을 마이크로 녹음한 파형이다. 해당 파형을 보면, 무음이 포함된 것이므로 데이터베이스에 있는 음성과 비교를 쉽게 하기 위해서 그림 8과 같이 무음부분을 제거하고 음성신호만 추출한다.

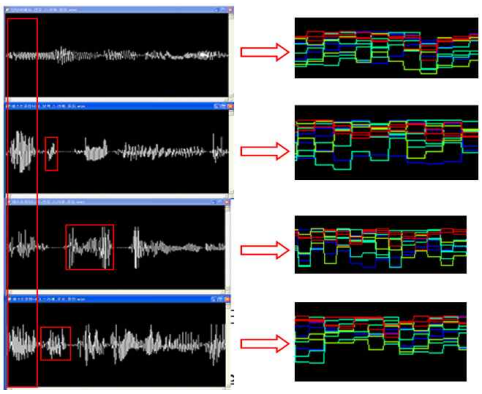


[그림 7] "안녕하세요" 음성 입력 신호



[그림 8] 무음을 제거한 음성 입력 신호

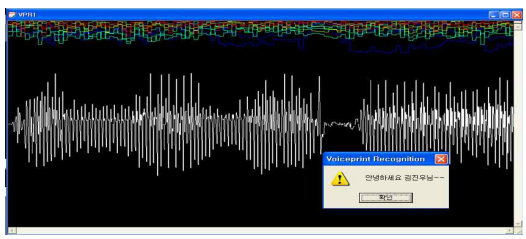
추출된 음성신호에서 MFCC 계수를 뽑아내어 한 화자의 특징으로 사용한다. 즉, 0차부터 11차까지 계수를 뽑아 특징으로 사용하는 것으로 그림 9에서 서로 다른 4명에 대한 MFCC 계수를 보였다.



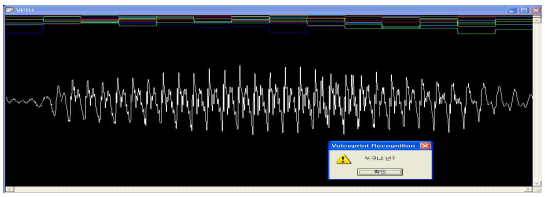
[그림 9] 서로 다른 4명의 "안녕하세요"의 MFCC

4.3 화자 식별 결과

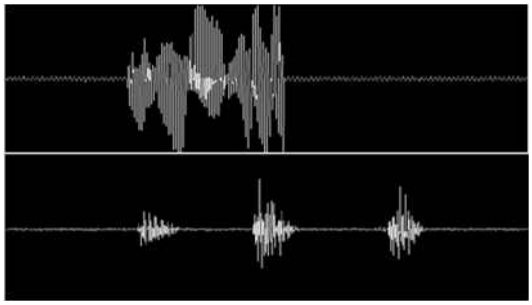
음성 입력이 있을 경우, 해당 입력에 대한 12차 MFCC 특성 계수를 추출하고, 데이터베이스에 저장되어 있는 5명의 특징 벡터와 비교를 통해 거리값이 가장 작게 나오는 화자로 인식이 이루어진다. 그림 10은 올바르게 화자를 찾은 것을 보여주는 결과이고, 그림 11은 5명이 아닌 이외의 사람의 음성이 들어왔을 때 데이터베이스에 일치하는 화자가 없음을 나타내는 결과이다.



[그림 10] 올바른 화자 식별 결과



[그림 11] 데이터베이스에 없는 화자의 결과



[그림 12] 입력 음성의 길이 차가 2배 이상인 경우

화자 식별이 제대로 되지 않는 경우는 한명의 화자가 아닌 두 사람이상의 목소리가 섞였을 경우와 입력으로 들어온 화자의 목소리가 데이터베이스에 있는 목소리 길이 보다 2배 이상 차이가 나는 경우 등이 있었다. 또한 입력된 화자의 목소리가 감기 등의 요인으로 인하여 변형되거나 입력 음성의 소리가 너무 작을 때에도 역시 오인식 되는 결과를 보였다.

그림 12는 음성의 길이가 비교하고자하는 데이터베이스에 있는 음성과 2배 이상 차이가 났을 때, MFCC를 제대로 구하지 못한 경우의 음성 파형의 예를 보여준 것이다.

4.4 성능 평가

본 연구에서는 두 가지에 대하여 성능 평가를 하였다. 첫 번째는 같은 단어를 발성하는 경우 사람의 음성을 인식하므로 다른 단어를 말하더라도 화자를 인식하는 지 여부를 평가하는 것이다. 표 1은 여러 화자에 대하여 데이터베이스에 있는 단어와 같은 단어를 발성하게 하였을 경우와 다른 단어를 발성하게 하였을 경우에 대한 인식률을 나타내는 표이다. 표 1의 A와 B 화자의 경우는 음정이 한 사람에 대해 변화가 없을 경우이며 비교적 높은 성능을 보인다. C, D 화자인 경우는 감기가 걸렸거나 음색이 달라서 상대적으로 낮은 성능을 보였다.

전반적으로 60%~72%의 인식률을 보이고 있으며 이는 높은 인식률이 아니기 때문에 성능개선을 위한 방법이 요구된다.

[표 1] 같은 단어를 발성하는 경우와 다른 단어를 발성하는 경우의 화자 인식률

음정 / 화자	같은 단어를 발성하는 경우	다른 단어를 발성하는 경우
A	80%	75%
B	80%	60%
C	70%	55%
D	75%	65%
E	55%	45%
Total	72%	60%

4. 결 론

본 논문에서는 입력 음성 신호로부터 MFCC를 뽑아내고 DTW 알고리즘을 이용해서 데이터베이스에 저장된 5명의 화자샘플과 비교해서 입력 음성의 신호가 어떤 화자인지 보여주는 화자 식별을 구현하였다. 실험 결과 전반적으로 높은 성능을 보이지 못했지만 데이터베이스의 음성과 비슷하게 발성하도록 하는 경우와 같이 제한적인 환경에서 화자식별 가능성을 제시했다. 보다 높은 성능을 도출해 내기 위해서는 더 나은 정합 알고리즘이 필요하다. 또한 다양한 화자의 목소리가 섞여있는 경우와 다양한 잡음이 포함된 경우에도 인식이 가능한 방법에 관한 연구가 필요하다.

참고문헌

[1] L. R. Rabiner and B.-H. Juang, “Fundamentals of Speech Recognition,” A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] 김현구, “인식점수의 궤환을 통한 음성 및 화자인식 시스템의 구현에 관한 연구”, 석사 학위 논문, 한국과학기술연구원, 2005.

[3] 이기엄, 배철수, 최갑석 “Speaker Recognition using Statistical process and DTW,” 제 1회 신호처리 합동 Workshop 논문집 제 1권 1호, 1988.

[4] Øystein Birkenes, “Automatic Speech Recognition-Plug_In MAP, Kernel Methods, and Hybrid Systems,” march 27, 2005.

[5] X. Shao and B. Milner, "Clean Speech Reconstruction from Noisy Mel-Frequency Cepstral Coefficients using A Sinusoidal Model," IEEE International Conference on Acoustics, Speech and

Signal Processing, pp. 704-707, 6-10 April 2003.

정 승 도(Seungdo Jeong) [종신회원]

- 1999년 2월 : 한양대학교 전자전자통신전파공학과 (공학사)
- 2001년 2월 : 한양대학교 전자통신전파공학과 (공학석사)
- 2007년 8월 : 한양대학교 전자통신전파공학과 (공학박사)
- 2009년 7월 ~ 현재 : 한양사이버대학교 정보통신공학과 전임강사

<관심분야>
멀티미디어 정보검색, 증강현실, 영상처리, 텐서 응용