로봇 시스템에의 적용을 위한 음성 및 화자인식 알고리즘

Implementation of the Auditory Sense for the Smart Robot: Speaker/Speech Recognition

조현† ·김경호*·박영진**

Hyun Jo, Gyeongho Kim and Youngjin Park

Key Words : Isolated word recognition(고립단어인식), Dynamic Time Warping(DTW, 동적 시간 정합), Gamma distribution(감마분포)

ABSTRACT

We will introduce speech/speaker recognition algorithm for the isolated word. In general case of speaker verification, Gaussian Mixture Model (GMM) is used to model the feature vectors of reference speech signals. On the other hand, Dynamic Time Warping (DTW) based template matching technique was proposed for the isolated word recognition in several years ago. We combine these two different concepts in a single method and then implement in a real time speaker/speech recognition system. Using our proposed method, it is guaranteed that a small number of reference speeches (5 or 6 times training) are enough to make reference model to satisfy 90% of recognition performance.

1. 서 론

최근 로봇산업의 성장과 함께 사람과 로봇간의 의사소통을 위해서 다양한 화자/음성 인식 알고리즘이 로봇 시스템에 사용되고 있다. 특히 인간과의 의사소통이 필수적인 실버메이트(Sliver mate) 용 로봇의 경우, 로봇은 사용자의 음성을 분석하여 사용자가 누구인지 그리고 무엇을 말하는지를 파악해야 한다. 본 논문에서는 음성인식 분야 중에서도 고립단어인식에 대한 알고리즘을 소개한다. 이것은 로봇이 주인의 이름이나 명령어 등과 같은 단어 수준의 음성을 듣고 음성신호가 무엇인지, 누구에서부터의 것인지를 판단하는 것을 의미한다. 음성인식에 있어서 다양한 알고리즘이 적용되고 있으며 특히 동적 시간 정합(DTW, Dynamic Time Warping) 알고리즘은 그 간단함과 적은 계산 량으로 인해 고립단어인식에 널리 사용되고 있다.

또 다른 음성 인식 기법에는 은닉 마르코브 모델(HMM, Hidden Markov Model)이나 신경회로망(ANN, Artificial Neural Network) 등과 같은 알고리즘이 있다. 이 두 가지 알고리즘의 경우, 연속단어인식(Continuous word recognition)과 같은 고급 음성 인식에 사용되고 있다. 은닉 마르코브 모델이나 신경회로망과 같은 알고리즘도 고립단어인식에 사용될 수 있으나 간단한 계산을 필요로 하는 로봇 시스템에는 적용하기 힘들다.

음성 인식과 함께 로봇이 화자의 음성을 다른 것과 구별하기 위해서 화자 인식 알고리즘이 필요하다. 화자인식 알고리즘의 경우, 특정 화자의 특징 파라미터에 근거한 가우시안 혼합 모델(GMM, Gaussian Mixture Model)이 널리 사용되고 있다[1].

본 논문에서는 고립단어에 대한 화자/음성 인식 알고리즘을 확률모델 바탕 하에서 하나로 통합하는 방식에 대해 소개한다. 본 논문에서 제시하는 방식은 확률모델을 사용한다는 점에서 가우시안 혼합 모델과 유사하다. 하지만 벡터 형태의 특징 파라미터를 사용하는 가우시안 혼합 모델과는 달리 본 논문에서 제시하는 방식은 동적 시간 정합의 최종 누적 거리(Total accumulated distance)를 사용하여 확률 밀도 함수를 수립한다.

본 논문의 개요는 다음과 같다. 2 장에서는 화자/음성 인식에 사용된 전체 알고리즘에 대해 소개한다. 3 장에서는 고립단어인식에 사용된 두 가지 특징 파라미터에 대한 소개와 설명이 있을 것이다. 또한 이 장에서는 본 논문에서 사용된 동적 시간 정합 알고리즘에 대한 구체적인 언급이 포함될 것이다. 4 장에서는 최종 누적 거리로부터 확률 모델을 수립하는 과정에 대한 설명이 있다. 5 장에서는 본 논문에서 제안하는 방식을 바탕으로 실험했을 경우 어떠한 결과가 있는지 언급될 것이며, 마지막으로 6 장에서는 결론을 맺을 것이다.

† 조현; 한국과학기술원
E-mail : e.w.smagel@kaist.ac.kr
Tel : (042) 869-3075, Fax : (042) 869-8220

* 한국원자력연구소

** 한국과학기술원

2. 화자/음성 인식 시스템

고립단어를 대상으로 실시간 화자/음성 인식을 하기 위해서는 다음의 세가지 단계가 순차적으로 이루어져야 한다.

2.1 정적 모델 수립

이 과정에서는 배경소음을 측정함으로써 음성 신호의 끝점검출(Endpoint detection)에 사용될 문턱 값을 결정한다. 배경소음은 22.05KHz 의 표본주파수로 수집되었으며 수집된 신호로부터 신호의 에너지와 영 교차율(Zero-crossing rate)이 프레임 별로 계산된다. 이 두 가지 정보로부터 끝점검출의 문턱 값을 설정하였으며 이는 Rabiner 와 Sambur 가 제안한 방식에 근거하였다 [2]. 로봇이 다른 환경의 배경소음에 노출된다면 정확한 끝점검출을 위해 정적 모델을 다시 수립할 필요가 있다.

2.2 참조 모델 수립

이 과정에서는 화자가 화자/음성 인식에 사용될 단어를 여러 번 말함으로써 우수한 참조 모델을 수립한다. 신뢰성 있는 참조 모델을 위해서는 한 단어에 대해서 5 번 이상 말하는 것을 권장한다.

참조 모델 수립의 전체적인 순서는 다음과 같다.

첫째, 표본주파수 22.05KHz 로 화자로부터 음성이 포함된 데이터가 수집된다. 이후, Rabiner 와 Sambur 가 제안한 끝점검출 기법을 이용하여 수집된 신호 중 음성 부분만을 분리한다.

둘째, 분리된 음성 신호를 수십 ms 단위의 프레임으로 분할하여 개개 프레임 별로 특징 파라미터를 추출한다. 이 과정에서 사용된 특징 파라미터에 대해서는 추후에 언급하도록 하겠다.

마지막으로 동적 시간 정합 기법을 적용하여 단어 간의 특징 파라미터를 비교하고 서로간의 최종 누적 거리를 확보한다. 동적 시간 정합 알고리즘의 결과를 바탕으로 화자/음성 인식에 대한 문턱 값을 결정한다.

2.3 테스트 모델 개시

이제, 두 번째 단계에서 수립한 참조 모델을 바탕으로 시스템이 화자/음성 인식을 수행한다.

테스트 모델의 개시는 다음의 순서를 따른다.

우선, 아날로그 트리거링을 이용하여 음성이 포함된 데이터가 자동적으로 수집된다. 아날로그 트리거링에는 트리거링 이전 신호의 확보가 필요한데 이는 배경소음 또는 단어 전후의 무성음으로 인해 트리거링 이전에 음성 신호가 존재할 수 있기 때문이다. 트리거링에 사용되는 문턱 값과 트리거링 이전 신호의 길이 등은 시행착오를 바탕으로 설정될 수 있다. 표본 주파수는 22.05KHz 이며 수집되는 데이터 수는 참조 모델에 사용되는 단어의 길이에 따라 결정된다. 일반적인 고립단어의 경우, 22.05KHz 의 표본주파수를 사용한다면 1 만에서 2 만 사이의 데이터가 적당하다. 데이터 수는 짧을수록 실시간 화자/음성 인식에 유리하다.

이제 참조 모델 수립 과정에서와 같이 수집된 데이터를 바탕으로 끝점검출, 특징 파라미터 추출, 동적 시간 정합 기법 적용이 차례로 이루어진다. 이 경우의 동적 시간 정합 기법은 테스트 모델의 음성과 참조 모델에 존재하는 다수의 음성 신호간에 이루어진다. 결과적으로 최종 누적 거리는 참조 모델에서 수립한 음성의 수만큼 확보 가능하다.

마지막으로 화자/음성 인식이 수행된다. 계산된 최종 누적 거리가 참조 모델에서 수립한 문턱 값 이하이면 화자/음성 모두 인식되는 것이며 그렇지 않을 경우, 인식은 거부된다.

3. 특징 파라미터 추출 및 동적 시간 정합 알고리즘

3.1 피치 & MFCC

본 연구에서는 사람의 음성을 인식하기 위해 피치와 MFCC(Mel Frequency Cepstral Coefficient) 라는 두 가지 특징 파라미터를 사용하였다.

피치는 성대의 첫 번째 고유진동수를 나타내는 값이다. 피치의 추출은 Dubnowski, Schafer, 그리고 Rabiner 가 제안한 개선된 자기상관 방식(Modified autocorrelation method)에 근거하였다 [3]. 피치는 40ms 단위로 추출되었으며 이웃 프레임 간에 30ms 의 겹침을 주었다. 또한 각각의 프레임에는 Hanning 윈도우가 씌워졌다.

MFCC 는 성도의 특징을 대변하는 변수이다. 본 연구에서는 Dan Ellis 가 웹 상에 공표한 HTK 의 MFCC 를 사용하였다 [4].

3.2 동적 시간 정합 알고리즘

동적 시간 정합을 수행하기 앞서서 정합에 사용될 두 개의 특징 파라미터는 길이가 긴 쪽을 기준으로 보간 된다. 같은 단어에서 추출된 특징 파라미터라 할지라도 그 길이가 크게 차이가 날 경우 최종 누적 거리가 커지게 되는데 이 작업은 이를 방지하기 위함이다. 파라미터 보간 이후에 동적 시간 정합은 UE2-1 이 사용되었으며 delta 는

0 에서 7 사이로 선택되었다 [5]. Delta 는 끝점 추출에서 발생할 수 있는 에러를 보상하기 위해 설정한 값으로 특징 파라미터의 길이에 따라 그 값이 선택된다. $w(n)$ 을 최적행로(Optimal path)라고 할 때, 최종 누적 거리는 다음과 같이 정의된다.

$$D_T = \min_{\{w(n)\}} \sum_{n=1}^{N} D\big(R(n), T(w(n))\big) \quad (1)$$

여기서 $R(n)$ 과 $T(m)$ 은 비교되는 두 개의 특징 파라미터를 의미한다.

4. 확률 모델의 수립

4.1 최종 누적 거리에 근거한 확률모델 수립

참조 모델에서 화자가 하나의 단어에 대해 5 번 말을 반복한다면 피치와 MFCC 가 5 개씩 생기게 되며 각각의 특징 파라미터 간에 계산할 수 있는 최종 누적 거리는 총 10 개가 된다.

그림 1 은 화자가 같은 단어를 20 번 반복해서 말할 때, 피치와 MFCC 로부터 계산된 최종 누적 거리를 히스토그램으로 표현하였다. 화자는 10 번

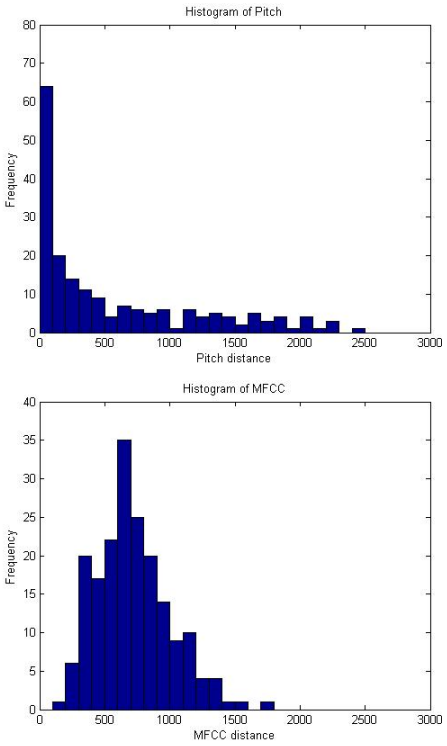

그림 1. “이리와”라고 말했을 때의 최종 누적 거리에 대한 히스토그램. 위의 그림은 피치에 대한 결과이고, 아래의 그림은 MFCC에 대한 결과이다. 실험자는 22세의 성인 남성이다.

을 무향실에서 말하였고 나머지 10 번을 일반 방에서 말하였다. 이 경우, 계산 가능한 최종 누적 거리는 총 190 개 이다. 피치로부터 계산한 최종 누적 거리는 대부분이 0 근처의 값을 가지며 전체적인 형상이 지수 분포를 따른다는 것을 확인할 수 있다. 반면, MFCC 로부터 계산한 최종 누적 거리는 Pitch 의 그것과는 달리 종 모양의 분포를 보인다. 이와 같은 최종 누적 거리에 대한 두 가지 경향은 시불변적이며 방의 음향 특성에 크게 영향 받지 않는다.

따라서 본 논문에서는 최종 누적 거리의 결과를 감마분포를 이용하여 모델링 하고자 하였다. 감마분포는 종 모양과 유사한 chi-square 분포와 지수 분포를 모두 표현할 수 있다. 감마 PDF 는 η 와 λ 라는 두 가지 변수에 의해 다음과 같이 정의된다.

$$y = f(x \mid \eta, \lambda) = \frac{\lambda^{\eta}}{\Gamma(\eta)} x^{\eta-1} e^{-\lambda x}, \quad x > 0. \quad (2)$$

위 식에서 Γ 는 감마함수이고 η 는 형상모수(Shape parameter) 이며 λ 는 역척도모수(Inverse scale parameter) 이다.

최종 누적 거리에 대한 감마분포 산출은 최대우도추정(MLE, Maximum Likelihood Estimate)에 의해 이루어졌으며 MLE 를 통한 형상모수와 역척도모수의 추정 값은 다음과 같이 표현된다 [6].

$$\hat{\lambda} = \frac{\bar{x}(n-1)}{\sum_{n=1}^{n}(x_i - \bar{x})^2} = \frac{\bar{x}}{s^2} \quad (3)$$

$$\hat{\eta} = \frac{\bar{x}^2(n-1)}{\sum_{n=1}^{n}(x_i - \bar{x})^2} = \hat{\lambda}\bar{x} \quad (4)$$

여기서 n 은 데이터의 수를 의미하고 $\bar{x}$ 과 s 는 각각 평균과 표준편차를 나타낸다.

그림 2 는 MLE 를 사용하여 피치와 MFCC 에 대한 최종 누적 거리를 감마분포로 추정한 결과이다. 실험은 무향실에서 한 단어를 10 번 말하도록 하였다. 그림 2 의 실선은 최종 누적 거리에 대한 피치 PDF 와 MFCC PDF 를 각각 나타내고 있다.

4.2 화자/음성 인식에 사용될 문턱 값 설정

최종 누적 거리를 바탕으로 수립된 두 가지 확률모델을 이용하면 화자/음성 인식 정확도를 백분율로 표현할 수 있다. 테스트 모델 개시 과정에서는 화자가 말하는 단어에 대해 참조 모델에서 수립한 음성의 개수만큼 최종 누적 거리가 얻어진다. 즉, 그림 2 의 예제로 참조 모델을 만들었을 경우

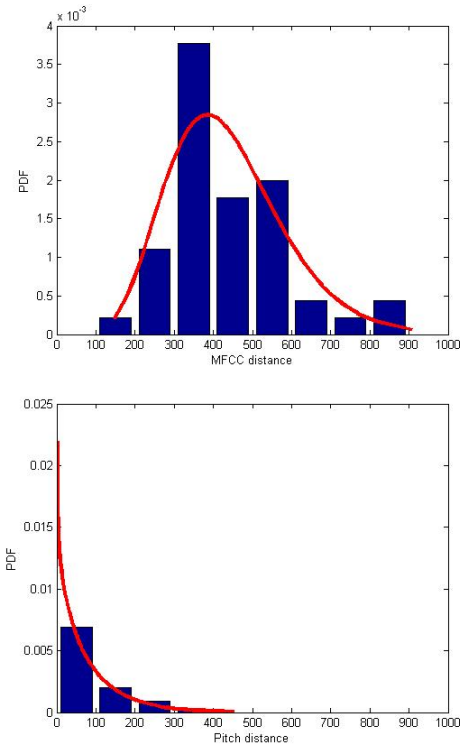

그림 2. “이리와”라고 말했을 때의 최종 누적 거리를 감마 PDF로 모델링한 결과. 위 그림은 MFCC에 대한 결과이며 아래 그림은 피치에 대한 결과이다. 실험자는 22세의 성인 남성이다.

에는 10 개의 최종 누적 거리가 피치와 MFCC 에 대해 얻어지게 된다. 얻어진 10 개의 최종 누적 거리에 대해 감마 CDF 를 이용하면 수립 모델과 테스트 모델간의 유사성을 정의할 수 있다. 우선, 감마 CDF 는 다음과 같다.

$$p = F(x \mid \eta, \lambda) = \frac{\lambda^{\eta}}{\Gamma(\eta)} \int_0^x t^{\eta-1} e^{-\lambda t} dt \,, \quad x > 0 \,. \tag{5}$$

수립 모델과 테스트 모델간의 유사성은 $1-p$ 로 정의된다. 최종 누적 거리가 짧을수록 $1-p$ 라는 값이 커지므로 유사성을 나타내는 의미 있는 값이다. 피치와 MFCC 에 대한 10 개씩의 유사성을 이용하여 화자 인식률과 음성 인식률을 다음과 같이 정의한다.

$$p_{Speaker\_recogntion} = \max\{1 - p_{Pitch}\}, \tag{6}$$

$$p_{Speech\_rocogntion} = \max\{1 - p_{MFCC}\}. \tag{7}$$

(6), (7) 식에서 최대값의 의미는 테스트 모델에서 사용한 음성신호가 참조 모델에서 사용된 음성신호 중 하나라도 같다면 화자 혹은 음성 인식의 성공으로 보겠다는 의미이다. 이것은 이상적인 상황을 고려한 것이며 그것은 바로 참조 모델의 음성신호가 테스트 모델의 음성신호와 완전히 같을 때, 인식률이 100%가 나와야 하기 때문이다.

마지막으로 최종 결과인 화자/음성 인식률 $p_{recognition}$ 을 다음과 같이 정의한다.

$$p_{recognition} = p_{Speaker\_rocognition} \times p_{Speech\_rocognition} \tag{8}$$

5. 실험 및 결과

본 논문에서 정의한 $p_{recognition}$ 이 어떠한 특성이 있는지 알아보기 위하여 실험을 수행하였다. 실험에는 3 명의 실험자가 참가하였다. (실험자 A – 22 세 남자, 실험자 B – 26 세 남자, 실험자 C – 27 세 남자) 실험에서 사용한 단어는 “안녕”, “이리와”, “저리가”, “청소해”, “그만”으로 총 다섯 단어이다. 실험자 A 는 각각의 단어를 다섯 번씩 말하여 참조 모델을 수립하였다. 이후 테스트 모델 개시에서는 실험자 A, B, C 가 각각의 단어를 5 번 말하도록 하여 화자/음성 인식률을 테스트 하였다.

그림 3 은 실험자 A 가 “안녕”이라는 참조 모델에 대해 화자/음성 인식률 $p_{recognition}$ 을 보여주고

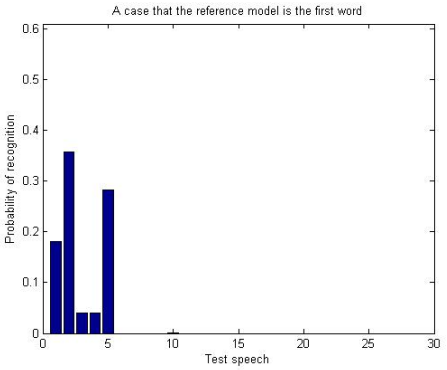

그림 3. 참조 모델이 실험자 A의 “안녕”이라는 단어로 만들어질 경우, 실험자 A의 음성인식률.

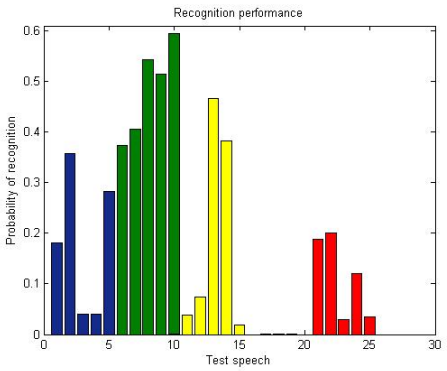

그림 4. 참조 모델과 테스트 모델에서 사용한 단어가 같을 경우의 화자/음성 인식률. (실험자 A의 경우)

있다. x 축은 테스트 모델에서 말한 음성신호를 의미하며, 1 부터 5 까지는 “안녕”, 6 부터 10 까지는 “이리와”를 나타낸다. 이와 마찬가지로 11 이후에도 저리가”, “청소해”, “그만” 이라는 말을 했을 때를 나타낸다. y 축은 해당 단어에 대한 확률 값을 보여준다.

그림 3 의 결과를 보면 테스트 모델에서 실험자 A 가 “안녕” 외의 단어를 말할 경우, 확률이 거의 0 이 나옴을 알 수 있다. $p_{recognition}$ 에 대한 문턱 값을 0.03 으로 설정한다면 “안녕”이라는 말에 대한 구분이 명확해진다.

그림 4 는 실험자 A 가 참조 모델에서 수립한 단어와 테스트 모델에서 말하게 되는 단어가 같을 경우의 결과이다. 그림에서 x, y 축의 의미는 그림 3 과 동일하다. 실험 결과에서 특별히 “청소해”라고 말한 경우는 확률이 거의 0 이 나오고 있는데 이는 참조 모델 수립 시, 화자가 같은 단어에 대해 너무 비슷하게 발음함으로써 최종 누적 거리가 분산되어 분포하지 못했기 때문이다. 본 논문에서 이를 참조 모델 수립 실패라고 명명할 것이며 이러한 문제는 참조 모델 기반의 화자/음성 인식의 문제로 알려져 있다 [7]. 이런 문제는 참조 모델 수립 시, 적절한 시간 간격을 두고 말하면 충분히 보상 가능하다.

그림 5 는 화자/음성 인식에서 오차가 발생한 부분을 보여주고 있다. 그림에서 y 축은 실험자 A 가 참조 모델에서 수립한 단어와 테스트 모델에서 말하게 되는 단어가 다를 경우에 발생할 수 있는 최대 확률 값을 표현하고 있다. 0.03 의 문턱 값에 대해 11 번째와 12 번째 단어인 “이리와”를 제외하고 오차가 발생하지 않았다.

그림 6 은 실험자 B, C 가 테스트 모델에서 말을 할 때의 인식률을 나타내고 있다. y 축은 그림 5 와 마찬가지로 발생할 수 있는 최대 오차를 표현하고 있다. 실험자 A 가 수립한 참조 모델의 단어에 대해 다른 화자가 말하게 되더라도 그 확률은 10 의 -4 승 차수로 낮게 나옴을 확인하였다. 이 결과로부터 본 논문에서 제시하는 방식을 이용하면 고립단어에 대한 음성인식 오차에 비해 화자인식 오차는 거의 발생하지 않는다는 결론을 내릴 수 있었다. 이 결과는 세 사람이 비슷한 연령의 남자라는 점을 생각해볼 때 놀라운 수치임을 알 수 있다.

6. 결 론

본 논문에서는 동적 시간 정합 기법의 응용으로써 최종 누적 거리를 바탕으로 화자/음성 인식

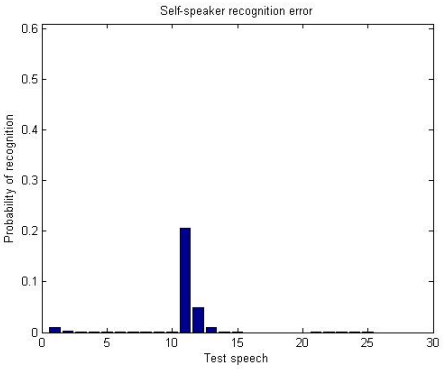

그림 5. 실험자 A의 화자/음성 인식 최대 오차

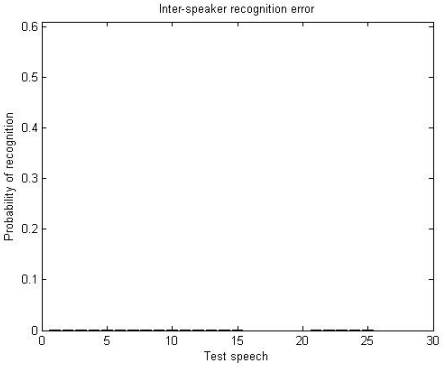

그림 6. 실험자 B, C의 화자 화자/음성 인식률

에 사용될 확률모델을 수립하였다. 확률모델은 감마분포를 사용하여 피치와 MFCC 의 최종 누적 거리를 잘 묘사하도록 하였다. 본 논문의 실험 및 결과 부분을 정리하면 다음과 같다. 첫째, 화자/음성 인식을 위해 5 번의 훈련으로 충분한 인식정확도를 확보할 수 있었으며 그것은 $p_{recognition}$ 에 대한 문턱 값을 10 의 -2 승 차수로 설정함으로써 가능하다. 둘째, 참조 모델 수립 실패가 일어나는 경우를 제외하면 같은 화자에 대해 음성 인식에서 오차가 나는 경우는 약 10% 정도이다. 셋째, 본 논문에서 제시하는 방식은 다른 화자의 말에 대해 매우 강인한 결과를 보여준다.

후 기

본 논문은 이 논문은 2007 년도 정부(과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업(M1050000 0112-05J0000-1121), 두뇌 한국 21 프로젝트와 전자부품연구원(KETI) 위탁과제로 공동 수행된 연구임.

참고문헌

[1] T. F. Quatieri, Discrete-Time Speech Signal Processing - Principles and Practice, Prentice Hall, NJ: 2002, pp. 709-725.

[2] Rabiner, L. R. and Sambur, M. R., “An Algorithm for Determining the Endpoints of Isolated Utterances,” The Bell System Technical Journal, vol. 54, no. 2, Feb. 1975, pp. 297-315.

[3] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, “Real-time digital hardware pitch detector,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, no. 1, pp. 2-8, Feb. 1976, pp. 2-8

[4] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/mfccs.html

[5] L. R. Rabiner, A. E. Rosenberg, S. E. Levinson, “Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, no. 6, Dec. 1978, pp.575-582.

[6] G. J. Hahn and S. S. Shapiro, Statistical Models in Engineering, John Wiley & Sons, Inc., NY·London·Sydney: 1967, pp. 87-88.

[7] W. H. Abdulla, D. Chow, G. Sin, “Cross-words Reference Template for DTW-based Speech Recognition Systems,” TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, 2003, vol. 4, pp. 1576-1579.