# 특징벡터 기반 음성 인식 시스템 설계 및 구현

*박진혁, **김현성, ***이성운
*경일대학교 사이버보안학과
**경일대학교 사이버보안학과
***(교신저자)동명대학교 정보보호학과
e-mail : *vkqxkr@naver.com, **kim@kiu.ac.kr

# Design and Implementation on Feature Vector based Voice Recognition System

*Jinhyuck Park, **Hyunsung Kim, ***Sung Woon Lee
*Dept. of Cyber Security, Kyungil University
**Dept. of Cyber Security, Kyungil University
***(Corresponding author)Dept. of Information Security, Tongmyong Univ.

## Abstract

Due to the recent developing technology of voice recognition, it is tendency to be increased the recognition rate. This paper designs and implements a voice recognition system based on feature vector scheme. Proposed voice recognition system achieves a higher recognition rate than the other previous systems. It could be used varieties of applications on smart devices to help people to use their devices efficiently and user friendly.

## I. 서 론

음성 인식 기술은 구글의 Voice Search와 애플의 Siri 등의 성공적인 도출로 많은 관심을 받고 있는 기술이다. 특히, 다양한 스마트 기기의 음성 명령 장치 및 인터페이스 등에 활용될 수 있는 다양한 종류의 음성 인식 애플리케이션들이 상용화되고 있다[1-3]. 음성 인식 기술의 지속적인 발전으로 음성 인식을 이용한 다양한 서비스를 위한 기술들이 발전하고 있다. 이미 음성 인식 시스템은 다양한 기기에서 활용되고 있으며, 인터넷에서 데이터 검색을 지원하는 음식 인식 기술도 존재한다.

특히, 최근 스마트 기기를 위한 인터페이스를 위해 음성 인식 기술이 시도하고 있으나, 아직 만족할만한 결과를 제시하지 못하고 있다. 그 주요한 이유는 입력 음성의 음향적 특성을 왜곡시키는 다양한 요소들에 대한 효율적인 처리 기술의 미비로 인함 일 것이다.

일반적으로 특징 보상 기반의 전처리 기술은 음질 향상 기술에 비해 음성 인식 성능 향상에 효과적이다[4]. 음성 모델을 채용한 전처리 기술의 경우에는 음성 인식 시스템의 음성 음향 모델인 은닉 마르코프 모델(Hidden Markov Model, HMM)과 동일한 음향적 특성을 갖는 음성 모델을 채용해야 한다. 대부분의 상용 음성 인식 시스템은 처리방식이 공개되지 않기 때문에 독립적으로 개발된 특징 보상 기반의 전처리 기술을 그대로 적용하기가 어렵다[5].

본 논문은 특징벡터 기반 음성 인식 시스템을 설계하고 구현한다. 최적의 인식률을 가지는 음성 인식 시스템을 개발하기 위해서 기존의 다양한 음성 인식 알

고리즘들을 비교 분석하고, 특징벡터 알고리즘을 중심으로 시스템을 설계하고 구현한다. 구현결과 본 논문에서 개발한 음성 인식 시스템이 기존의 알고리즘보다 인식률이 좋음을 확인할 수 있다[6].

## II. 음성 인식 기술

본 장에서는 다양한 음성 인식 기술들의 개요를 살펴본다. 특히, 음성 인식을 위한 시스템의 구성요소 및 인식 방법에 대해서 기술한다.

### 2.1 은닉마르코프모델

은닉마르코프모델(Hidden Markov Model, HMM)은 모델을 구성하고 있는 상태들 간의 전이가 특정한 확률값을 통하여 이루어지는 통계적 모델 중 하나이다. 은닉마르코프모델은 시스템이 은닉된 상태와 관찰 가능한 결과의 두 가지 요소로 구성되었다고 가정한다. 관찰 가능한 결과를 야기하는 직접적인 원인은 관측될 수 없는 은닉 상태들이고 오직 그 상태들이 마르코프 과정을 통해 도출된 결과들만이 관찰될 수 있기 때문에 '은닉'이라는 수식어가 붙었다. 은닉마르코프모델은 동적 베이지안 네트워크로 간단하게 나타낼 수 있다. 이 모델에서 음성은 "특정한 문자열로부터 도출된 출력 변수"로 고려되며, 모델의 최적해를 찾는 과정은 "관찰된 출력 변수(음성)를 가장 잘 설명하는 은닉 상태(문자열)"를 찾는 과정으로 고려된다[7].

### 2.2 특징 벡터 추출 알고리즘

특징 벡터 추출(Feature Vector Extraction, FVE)은 일반적인 음성 인식 시스템에서 가장 널리 사용되는 알고리즘이다. 인식을 위해 음성을 작은 구간으로 나누고 각 구간에서 그 구간을 대표할 수 있는 특징을 추출하여 사용한다. 각 구간에서 추출할 수 있는 특징 중 대표적인 두 가지는 MFCC (Mel-Frequency Cepstral Coefficients)와 LPC (Linear Prediction Coding)가 있다[6].

- MFCC : 단 구간 신호의 파워 스펙트럼을 표현하는 방법 중 하나로, 비선형적인 Mel스케일의 주파수 도메인에서 로그파워스펙트럼에 코사인변환을 취함으로써 특징을 추출할 수 있다. 일반적인 캡스트럼의 경우 주파수 밴드가 균등하게 나누어져 있는 반면 MFCC의 경우 주파수 밴드가 Mel-scale에서 균등하게 나누어진다. Mel-scale로의 주파수 워핑은 소리를 더욱 잘 표현할 수 있는 장점이 있다. 따라서 오디오 압축 등에 활용된다[8-10].
- LPC : 음성파형의 표본값을 과거의 인접하는 표본값 계열로부터 선형 예측 모델에 기반하여 음성을 분석한다. 이 방법은 스펙트럼 8~12개의 특징파라미터로 효율적으로 표현할 수 있고, 음성의 협대역 전송, 음성 응답 장치, 음성 인식 등의 분야에서 응용되고 있다. 주기적인 Pulse열과 White Gaussian Noise를 Excitation Source로 해서 LPC Filter를 걸쳐 생성되는 음성 생성 모델에서 비롯되었다. 이 벡터들은 음성학적 특성을 잘 나타내며 그 이외의 요소, 즉 배경 잡음, 화자 차이, 발음 태도 등에는 둔감해야 하며 이 과정을 거쳐 인식부에서는 순수하게 음성학적 특성에만 집중해 분석할 수 있게 된다[8].

### 2.3 동적시간 외곡 알고리즘

동적시간 외곡 알고리즘(Dynamic Time Warping, DTW)은 시간순서상에서 여러 개의 연속적인 데이터를 비교하여 그 데이터들 간의 유사도를 판별하는 기법이다[11-12].

### 2.4 인공신경망 알고리즘

인공신경망(Artificial Neural Network, ANN) 알고리즘은 Input layer, Hidden layer, Output layer로 그림 1과 같이 구성 된다. Input layer는 데이터의 입력을 위한 계층이고, Output layer는 데이터의 출력을 위한 계층이며, 이들 사이의 계층에 Hidden layer가 있다. 그림에서 보여주는 바와 같이 입력과 출력은 node로 표현된다. Hidden layer에서 인공신경망을 통해 학습을 실시하고, 이때 node수가 늘어날수록 좀 더 복잡한 데이터에 대한 정확한 처리가 가능하지만 계산량이 많아지는 단점이 발생할 수 있다. 따라서 입력 데이터의 크기를 적절히 설정해야 알고리즘의 성능을 보증할 수 있다[13].

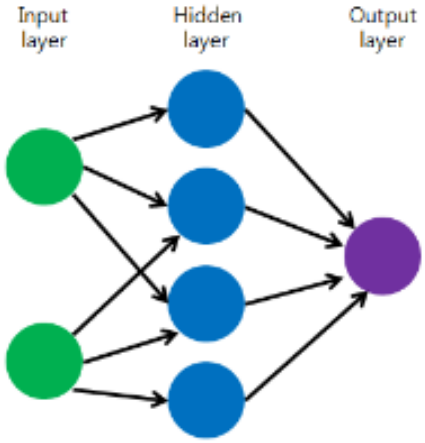


그림 1. 인공 신경망 처리 과정

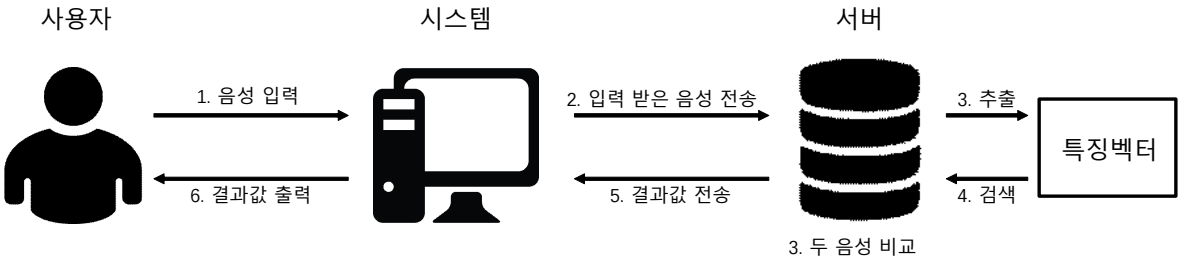


그림 2. 제안한 특징벡터 기반 음성 인식 시스템 개요도

## III. 특징벡터 기반 음성 인식 시스템

본 장에서는 특징벡터 기반 음성 인식 시스템을 설계하기 위해서 먼저 개발환경 구축에 대한 개요를 제시한다. 또한, 전체적인 시스템을 설계하기 위해서 입력모듈, 특징벡터 추출 및 매칭 모듈, 그리고 추출 모듈로 구성된 시스템의 구성 모듈에 대해 살펴본다.

### 3.1 음성 인식 개발 환경 구축

일반적으로 음성 인식 시스템을 개발하기 위해서는 비교되어야 할 음성 데이터의 큰 용량 때문에 음성 데이터를 저장할 서버가 필요하다. 그림 2는 본 논문에서 제안하는 특징벡터 기반 음성 인식 시스템의 개요도를 제시한다. 처리 과정은 먼저 음성을 입력 받고, 네트워크 연결을 통하여 입력된 데이터를 서버에 전송하여 인식 과정을 처리 한 후 그 결과 값을 다시 전송받아 인식 과정을 수행한다.

### 3.2 특징벡터 기반 음성 인식 시스템 모듈

제안된 특징벡터 기반 음성 인식 시스템은 입력모듈과 특징벡터 추출 및 매칭 모듈 그리고 추출모듈의 총 3개의 모듈로 구성된다. 특히, 시스템 구축을 위해 미리 학습되어 있는 데이터를 가지고 있는 매칭모듈을 활용한다고 가정한다.

(1) 입력모듈

입력모듈에서는 사용자의 음성을 입력 받는다. 그 후 입력 모듈은 음성 데이터를 특징벡터 출력 모듈로 보낸다. 특징벡터 추출 및 매칭모듈에서는 받은 음성 데이터의 특징을 추출한 다음에 그 특징과 가장 비슷한 데이터가 있는지 검색을 한다. 검색 후 추출된 특징과 가장 비슷한 값을 가지고 있는 데이터를 출력모듈로 보낸다. 출력 모듈에서는 최종 값을 사용자에게 보여준다.

인식을 위해 음성을 작은 구간으로 나누고 각 구간에서 그 구간을 대표할 수 있는 특징을 추출하여 사용하는데 본 논문에서의 구현은 MFCC 특징 벡터 추출 알고리즘을 사용하였다.

인간은 낮은 주파수 영역에서는 높은 Resolution을 갖는 주파수 영역에서는 낮은 해상도를 갖는데 이러한 것은 주파수 대역에서 Logarithm한 특성인데 이 단위가 바로 Mel이다. 일반적으로 이의 구현은 구간의 샘플을 FFT하여 얻어지는 스펙트럼의 크기나 Power를 구하고 이를 Mel-scale로 구성된 Filter-of-bank를 거쳐서 얻게 된다. MFCC는 일반적으로 다음과 같은 과정을 통해 구할 수 있다. 먼저, 음성 신호를 10 ms 마다 25 ms 구간으로 STFT(Short-Time Fourier Transform)를 수행한 뒤, 인간의 청각 모델을 모방한 Mel-scale Filter Bank를 통해서 각 대역의 에너지들을 얻는다. 이 에너지의 Log값에 DCT(Discrete Cosine Transform)을 수행하여, 최종적으로 MFCCs(Mel- Frequency Cepstral Coefficients)를 얻는다. 그림 3은 전체적인 MFCC 처리 과정을 보여준다[11].

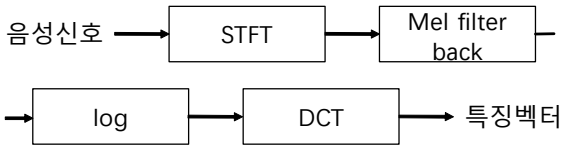


그림 3. MFCC 특징 벡터 추출 알고리즘 처리 과정

## IV. 구현

본 장에서는 3장에서 제안한 특징벡터 추출 알고리즘을 이용하여 인식률을 테스트한다.

### 4.1 개발 환경

시스템 구현을 위한 환경으로써 인식률 툴에 기반한 구현 개발 환경은 표 1과 같다. 클라이언트 시스템은

안드로이드 애플리케이션으로 개발되었고, 개발 플랫폼(Platform)은 Android Studio를 사용하였다.

표 1. 개발환경

| Term | Property |
|---|---|
| Processor | Intel Core i7 |
| Memory | 8GB |
| OS | Windows 10 |
| Language | JAVA |
| Platform | Android Studio |
| S/W version | Android 4.3 |

### 4.2 실험결과

그림 4는 본 논문에서 제안된 특징벡터 추출 알고리즘을 사용한 음성 인식 시스템의 클라이언트를 보여준다. 클라이언트 프로그램은 사용전에 네트워크 연결되어있어야 한다. 사용자가 Start to Speak버튼을 누르고 클라이언트 프로그램을 통하여 음성을 입력한다. 음성 입력 후 클라이언트는 이 음성 데이터를 서버에 전송하면 서버는 특징벡터 알고리즘을 기반으로 음성 데이터의 특징점을 추출하고, 추출된 특징점에 기반한 결과를 클라이언트에 보낸다. 클라이언트는 결과의 매칭도가 가장 높은 순으로 단어 또는 문장을 출력한다. 만약에 매칭도가 낮을 경우 결과 값은 출력이 되지 않는다. Stop버튼을 누르면 음성입력을 중지한다.


(1) 시작화면 (2) 인식화면

그림 4. 음성 인식 클라이언트 실행 화면

특히, 시스템 구현결과 기존의 DTW음성인식 시스템은 평균 56%의 인식률을 제시하였고, 본 논문에서 제안한 시스템은 인식률이 93%로서 기존의 시스템보다 우수한 인식률을 갖는 것을 확인할 수 있었다.

## V. 결론 및 향후 연구 방향

본 논문에서 제안한 특정벡터 추출 알고리즘을 이용한 인식률 툴의 정확도는 주변 환경과 영향이 있다. 특히, 주변 환경으로부터 소음으로 인해 음성을 입력하고 인식할 때 인식오류가 발생할 수도 있다. 주변 소음이 있는 환경에서 테스트를 했을 경우 10번의 경우 3~4번 정도가 오류가 생기는 것을 확인 하였다. 하지만 주변 소음이 없을 경우 뛰어난 인식률을 확인 할 수 있었다.

## 사사(Acknowledgement)

본 연구는 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 연구임(한국연구재단2010-0021575). 또한 2011년 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 연구임(한국연구재단 2011-0008890).

## 참고문헌

[1] 김수훈, 안종영, “모바일기반 음성인터페이스에 관한 연구”, 한국인터넷방송통신학회논문지, Vol. 13, No. 1, pp. 199-204, 2013.

[2] 이유나, “애플, 새 운영체제 IOS9정식배포...시리 기능 향상”, 머니투데이방송, 2015.

[3] 손경호, “구글, 보이스액세스 공개...앱 음성 제어”, ZDNet Korea, 2015.

[4] 김우일, “이기종 음성 인식 시스템에 독립적으로 적용 가능한 특징 보상 기반의 음성 향상 기법”, 한국정보통신학회논문지, Vol. 18, No. 10, pp. 2367-2374, 2014.

[5] 조영임, “효과적 잡음 제어에 의한 음성 인식 시스템의 성능 개선”, 한국지능시스템학회2010년도춘계학술대회학술발표논문집, Vol. 20, No. 1, pp. 361-362, 2010.

[6] 최재승, “음성의 특징벡터를 사용한 정규화 인식수법”, 한국해양정보통신학회2011년추계학술대회논문집, pp. 616-618, 2011.

[7] 조성정, “Introduction to Hidden Markov Model and Its Application”, 한국정보과학회, pp. 1-72, 2005.

[8] 오상엽, ”MFCC와 LPC 특징 추출 방법을 이용한 음성 인식 오류 보정“, 디지털융복합연구, Vol. 11, No. 6, pp. 137-142, 2013.

[9] 이광석, 김현주, “주성분 해석기법에 의한 음성 특징벡터 추출 및 음성 성능 평가”, 한국지식정보기술학회논문지, Vol. 5, No. 6, pp. 239-245, 2010.
[10] 정다해, 배민경, 김윤경, 이의철, 정진우, “MFCC 특징을 이용한 녹취 파일의 화자 구분”, 한국통신학회종합학술발표회논문집, pp. 866-867, 2014.
[11] 위키백과, DTW, 2015.
[12] 안종영, 김성수, 김수훈, 고시영, 허강인, “잡음환경에서의 Noise Cancel DTW를 이용한 음성인식에 관한 연구”, 한국인터넷방송통신학회논문지 vol. 11, No. 4, pp. 181-186, 2011.
[13] 네이버지식백과, 인공 신경 회로망, 2015.
[14] 박현신, 김신웅, 유창동, “최신 기계학습 기반 음성인식 기술 동향”, 전자공학회지, Vol. 41, No. 3, pp. 18-27, 2014.