HMM을 이용한 음성인식 시스템의 전처리에 관한 연구

이윤주, 오세영, 이순규, 배명진(°)
숭실대학교 정보통신공학과

A Study of Preprocessing in the Speech Recognition System Using HMM Algorithm

Yoonju Lee, Seyoung Oh, Soonkyu Lee, Myungjin Bae(°)
Dept. of Telecommunication, Soongsil University
E-mail(°): mjbae@saint.soongsil.ac.kr

요약문

현대 사회의 컴퓨터 사용자 계층은 점점 그 범위와 수가 커지고 있다. 이러한 추세는 앞으로도 계속 증가할 것이다. 따라서 많은 사람들은 더 편리하고 익히기 쉬운 컴퓨터의 사용법을 원하고 생활속에서 더 많이 컴퓨터를 활용하기를 원한다. 그러므로 인간에게 가장 친숙한 음성을 이용함으로써 이런 사용자들의 필요를 충족시킬 수 있을 뿐 아니라 사용자가 쉽게 접할 수 있도록 할 수 있다. 그러므로 본 논문의 목적은 이러한 상황에서 인간과 기계와의 인터페이스를 인간의 기본적인 의사소통 수단인 음성을 이용하여 보다 빨리 작업 할 수 있게 하는 취지에 있다. 기존의 인식알고리즘은 그 복잡성이 높을수록 인식률은 증가하나 계산시간이 많이 걸린다는 단점이 있다. 이러한 계산시간의 증가는 윈도우환경의 컴퓨터 사용시 다른 프로그램의 실행에 지장을 줄 수 있다. 따라서 인식률은 증가시키면서 인식 시간은 감소시킬 수 있는 방법들이 필요하다. 본 논문에서는 컴퓨터 사용시 쓰이는 명령어를 기본으로 하여 보다 빠른 인식 처리를 수행하기 위해 기준 패턴의 후보자를 선정하는 방법을 제안한다.

1. 서론

인간의 목소리는 허파에서 압력에 의해 밖으로 나온 공기가 성문과 각종 조음기관을 거치면서 기본주파수와 공명주파수를 가지고 입을 통과해 나옴으로서 생성된다[1][2]. 이렇게 생성된 음성신호를 우리는 귀에서 입력받아 각각의 처리를 수행한 후 뇌가 인지하게 된다. 음성인식이란 사람뿐만 아니라 기계가 사람의 목소리를 알아들을 수 있도록 하는 것이다. 즉, HCI(Human Computer Interface)란 사람이 컴퓨터를 사용할 때 키보드나 마우스등의 입력장치뿐만 아니라 마이크를 이용하여 음성으로 명령을 할 수 있게 하는 것이다. 근래 들어 이러한 연구는 활발히 이루어지고 있으며 상용화된 제품도 있다.

이렇게 음성신호를 기계가 인지할 수 있게 하려면 많은 처리가 필요하다. 즉, 음성신호를 마이크를 이용하여 입력받은 뒤 음성구간과 배경잡음을 구별하고 음성 신호 고유의 특징을 추출한다. 이렇게 추출된 음성신호를 이용하여 기계를 학습시킨다. 학습방법으로는 여러 가지가 있으나 가장 널리 쓰이는 알고리즘은 벡터양자화법이다. 본 논문에서 사용하는 학습 데이타는 컴퓨터 명령어이다. 컴퓨터를 학습시키기 위해 사용하는 음성특징을 기준패턴이라 한다. 따라서 컴퓨터가 음성 데이타로 학습되었다면 우리는 학습되어 있는 명령어를 말로 하여 컴퓨터를 동작시킬 수 있다. 이러한 과정은 각종 음성인식 알고리즘을 적용함으로써 수행된다. 본 논문에서는 은닉마코브 모델(Hidden Makov Model, HMM)을 이용하여 음성인식을 수행하였다[2].

기존의 HMM을 이용한 음성인식은 컴퓨터에 등록되어 있는 명령어의 수가 많을수록 처리시간이 증가하여 윈도우환경의 컴퓨터를 사용할 경우 인식프로그램의 과다한 계산시간으로 인해 다중적인 프로그램을 수행하기에 어렵다는 단점이 있다. 따라서 이러한 문제점을 해결해야한다. 본 논문은 기준패턴의 후보자를 선정하는 방법을 사용하여 이러한 단점을 보완하였다. 기준패턴 선정시 사용하는 음성을 발성할 때 각 음절사이에 묵음구간을 두어 발성하면 초기 한 음절에 대한 각각의 특징파라메터 분포는 명령어에 따라 다르다. 따라서 이러한 특징을 이용한다면 인식알고리즘을 수행하기 전 미리 입력된 음성과 비교할 기준패턴의 후보를 정할 수 있을 것이다. 따라서 시간단축이 가능하다.

2. 인식 시스템의 구성

음성 신호는 허파에서 밀려나온 공기가 성문의 개폐 주기에 의해 신호주기(피치-Pitch)가 결정되며 성도내의 조음기간의 모양에 의해 음운학적 의미가 달라지게 된다. 성문의 개폐주기에 의해 결정되는 피치는 사람마다 그 변화특성이 크며 같은 사람인 경우에도 그 변화정도가 감정상태, 건강상태, 시간변화에 따라 매우 심하다. 따라서 음성인식에서는 이렇게 변화가 심한 성문의 특성보다는 변화의 정도가 작은 성문의 특성을 이용한다[1][3].

음성신호는 단구간적으로 정적(Stationary)이므로 예측이 가능하다. 따라서 음성신호를 선형예측필터를 사용하여 모델링할 수 있으며 이때 추출되는 필터 계수를 선형예측계수(LPC-Linear Prediction Coefficients)라 한다. LPC는 성도의 특성인 각각의 포만트(Formant)의 중심주파수와 대역폭을 나타낸다. 그러므로 LPC를 음성특징 파라메터로 사용할 수 있다. 이러한 LPC를 LPC-켑스트럼(Cepstrum)으로 변환한 뒤 인간의 귀의 특성을 반영하기 위해 멜스케일(Mel scale)이나 바크스케일(Bark scale)로 변환한다. 이때 이 계수를 멜켑스트럼계수(MFCC-Mel Frequency Cepstrum Coefficients)라 한다. 본 논문에서는 음성특징으로 14차의 MFCC를 사용하였다[2]. 아래 그림2-1은 위와 같은 음성특징을 이용하여 음성을 인식하는 일반적인 시스템에 대한 것이다.

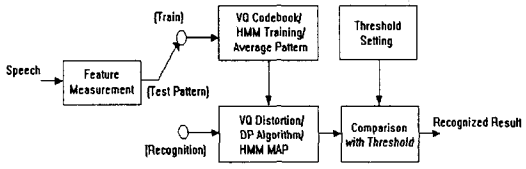

그림 2-1. 일반적인 음성인식 시스템

그림에서 보는 바와 같이 시스템에 입력된 음성의 특징을 추출한 후 등록된 음성 특징과 각각의 알고리즘을 이용하여 비교함으로써 음성인식을 수행한다. 이 과정을 보다 자세히 살펴보면 다음과 같다.

2.1 선행처리

음성의 특징을 추출하기 전에 수행할 과정들을 선행처리라 하며 이에는 음성구간검출, Preemphasis 등이 있다. 음성구간 검출이란 음성과 배경잡음을 구별하는 것으로 이는 인식률에 큰 영향을 미칠 수 있으므로 정확한 검출이 요구된다. 또한 실시간으로 동작하는 음성시스템을 위해서는 계산량이 적어야 한다.

본 논문에서는 그림 2-2와 같은 방법을 이용하여 음성구간 검출을 수행하였다. 유성음이 배경잡음보다 에너지가 크다는 성질을 이용하여 단구간 에너지 값으로 유성음 구간을 검출한다. 그리고 무성음 구간을 고려하기 위해 영교차율(ZCR-Zero Crossing Rate)을 이용한다. 또한 음절사이의 묵음을 고려하기 위해 끝점이 검출된 후에도 에너지 값을 이용하여 수십 msec동안 시작점을 찾는다. 만일 이 구간 내에 시작점이 나타나지 않으면 음성구간 검출 과정을 종료한다[1].

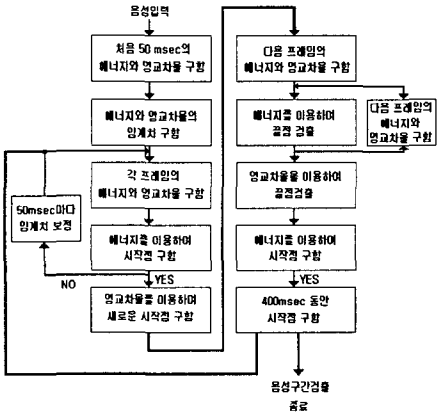

그림 2-2 음성구간 검출

다음으로 음성신호는 유성음의 경우 저차 포만트(Formant)의 에너지가 고차 포만트의 에너지에 비해 크다는 특성이 있다. 이는 스펙트럼 거리 비교시 거리 값이 저차 포만트들의 작은 차이에 의해 고차포만트의 큰 차이가 상대적으로 줄어든다는 단점을 초래한다. 따라서 이러한 문제점을 해결하기 위해 음성신호를 Preemphasis 필터를 이용하여 고주파항의 영향을 높여야 한다. 이렇게 함으로써 스펙트럼이 평탄화 되어 올바르게 두 패턴의 비교가 수행될 수 있다. 따라서 인식률 또한 증가하게 된다.

2.2 음성 특징 추출

앞에서 언급한 MFCC의 음성특징을 추출하기 위해 본 논문에서는 그림 2-3과 같은 과정을 수행하였다. 전처리가 수행된 음성신호를 해밍윈도우를 이용하여 단구간으로 나눈다.

다음으로 LPC값을 추출한 후 LPC 켑스트럼 변환식을 이용하여 LPC-Cepsturm을 구한다. 그리고 음성

특징 파라메터가 귀의 특성을 반영할 수 있도록 하기 위해 Bilinear 변환을 이용하여 최종적으로 14차 MFCC 파라메터를 추출한다[2][3][4].

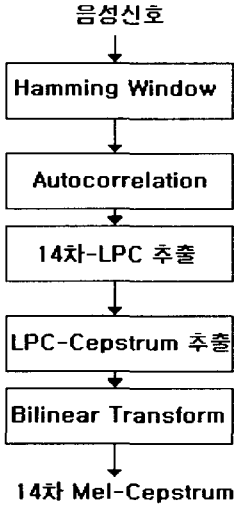

그림 2-3. 특징 벡터 추출 과정

3. 인식 알고리즘

음성신호에 들어 있는 정보는 일반적으로 짧은 간격의 전력 스펙트럼에 들어있고 시간이 지나감에 따라 약간씩 변하게 된다. 그러므로 전력 스펙트럼과 그것의 시간적 변화를 추정하는 방법을 필요로 하게 된다. 이와 같은 목적으로 성질이 non-stationary한 음성을 모델링해서 음성인식에 사용하는 것이 HMM 알고리즘이다[2].

HMM에는 몇 개의 상태가 있고 각 상태에는 각각의 불규칙함수에 의해서 하나의 출력을 내고 천이 확률에 의해서 다음 상태로 넘어간다. 우리는 이 출력은 볼 수 있지만 상태가 어떻게 변화하는 지는 알 수 없다. 음성신호가 Markov process에 의해서 발생한다고 생각해 보면 성도가 몇 개의 상태로 나뉘어져 있고 각 상태에서 짧은 시간 간격의 이 신호는 한정된 수의 기준 spectra의 어느 하나로 치환할 수 있다고 생각할 수 있다. 그러므로 어떤 짧은 시간의 전력 스펙트럼은 그 한 상태에 의해서만 결정된다고 볼 수 있다. 그리고 스펙트럼의 시간적 변화는 상태천이에 의해서 설명될 수 있다. 따라서 어떠한 음성에 대한 상태천이 확률과 상태내의 관측심볼을 확률적으로 모델링한다면 그 모델링 파라메터는 다른 음성과는 다른 성질을 나타낼 것이다. 이러한 방법으로 각각의 음성을 구별하여 인식하는 방법이 HMM이다.

본 논문에서는 left to right HMM 모델을 사용하여 음성인식을 수행하였다. 이 모델은 다음과 같이 표현할 수 있다[2].

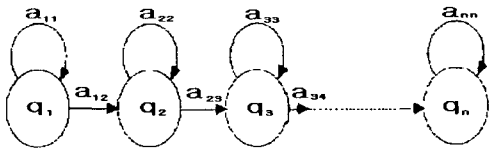

그림 3-1. left to right HMM

본 논문에서 구현한 전체적인 인식알고리즘은 다음과 같다.

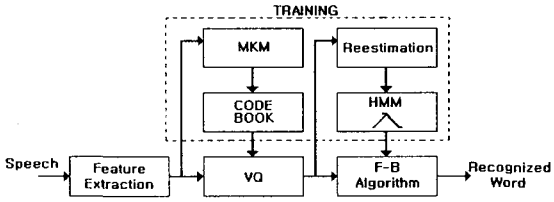

그림 3-2. 인식 알고리즘

4. 후보자 선택 방법

기준패턴 선정시 본 논문에서는 컴퓨터 명령어를 고립어로 발성하였고 또한 저장된 기준패턴은 문장종속 어휘를 사용하였다. 따라서 각각의 기준패턴의 경우 초기 음절이 같은 음일 경우가 있는 반면 대부분 다른 음일 경우가 많다. 예를 들어 "잘라내기"라는 명령과 "읽고 쓰기"라는 명령어는 초기 "잘"과 "읽"이라는 음절이 다르다. 특히 이런 경우 "ㅈ"과 "ㅇ"음소는 무성음과 유성음으로 구별되며 이런 경우 음성특징 파라메터의 차이가 크다. 따라서 본 논문에서는 음성의 시작점에서 수 msec에 대한 MFCC 파라메터의 각각의 차수에 대해 그 분포도를 작성하여 평균값을 추출한다. 이렇게 구해진 평균값을 이용하여 기준패턴 선정시 이 값을 함께 저장한다. 시험패턴이 인식 시스템에 들어온 경우 위와 같은 방법으로 초기 수 msec동안 해당하는 MFCC 파라메터에 대해 각각의 차수값을 구한 후 기준패턴에 저장되어 있는 평균값과 거리를 측정한다. 거리 측정 결과 거리가 가장 작은 순으로 정렬하여 n개의 기준패턴 후보자를 선정한다. 여기서 n이라는 기준패턴의 수가 증가할수록 그 수를 증가시킨다.

5. 실험 및 결과

제안한 방법을 시뮬레이션하기 위해 IBM-PC MMX/220에 마이크가 장치된 16-bit A/D변환기를 인터페이스 시켰다. 실험은 일반 실험실 환경에서 수행하였다. 음성 시료는 11025Hz로 샘플링하고 16bit로 양자화하였다. 음성의 특징파라메터 추출을 위해 길이가 30msec인 해밍윈도우를 사용하였고 이를 15msec씩 오버랩하였다. 인식알고리즘은 3개의 상태를 갖는 left to right HMM을 이용하였고 개선된 K-means 알고리즘을 이용하여 벡터양자화를 수행하였다. 코드북의 크기는 512로 하였다. 시스템의 등록단어는 20대 남,여 10명이 각각 10번씩 발성한 10개의 윈도우 환경의 컴퓨터 명령어를 사용하였다.

초기 MFCC 파라메터의 추출을 위해 60msec동안의 특징을 사용하였으며 후보자의 수는 3개로 하였다. 본 논문에서 구현한 음성인식시스템의 전체적인 구성은 아래 그림과 같다.

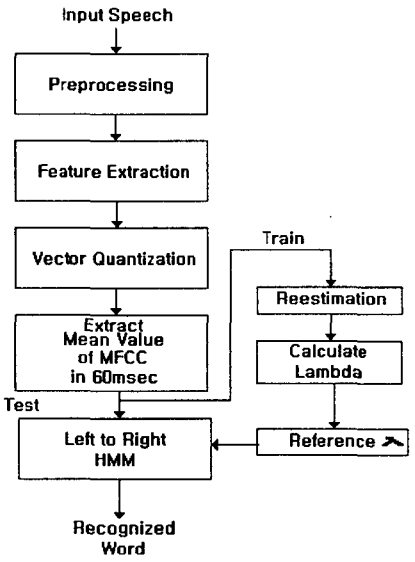

그림 5-1. 음성인식 시스템의 구성도

제안한 방법의 성능 평가를 위해 기준패턴의 후보자를 선택하지 않은 기존의 방법과 제안한 방법의 처리 시간과 인식률을 비교하였다. 실험결과 제안한 방법의 처리 시간이 기존의 방법에 비해 약 38% 감소하였다. 그러나 약간의 인식률 저하가 있었으며 이는 음성구간 검출의 문제점 때문이었다. 앞으로 다른 파라메터를 이용하여 후보자 선택을 수행 할 예정이며 초기 구간의 변화를 주어 이 문제점을 해결할 것이다.

모의 실험 결과는 아래 표와 같다.

표5-1 처리 시간(sec)

	Test
기존의 방법	1.52
제안한 방법	0.94

표5-2 전체 인식률(%)

	전체 인식률
기존의 방법	97.5
제안한 방법	96.3

6. 결론

HCI를 위한 음성인식의 연구는 현재 활발히 이루어지고 있다. 이에 대한 연구의 두 가지 큰 목적은 인식률 향상과 처리 시간 단축이다. 이러한 연구방향은 앞으로 인간이 보다 편리하고 쉽게 컴퓨터를 사용하고자 하는 희망에서 나온 것이다. 지금까지 개발된 음성인식기나 연구된 음성인식 기술은 사용화 단계에 접어들었으며 연속음 인식과 대용량 어휘인식을 위한 노력이 계속적으로 진행되고 있다. 이러한 현실에서 인식을 위한 처리시간 단축은 반드시 이루어져야 한다.

본 논문은 고립단어 어휘종속 컴퓨터 명령어를 사용하여 음성인식을 수행할 경우 음성 파라메터의 분포특성을 이용하여 기준패턴의 후보자를 선정함으로써 처리 시간 단축을 수행하였다.

6. 참 고 문 헌

[1] L. R. Rabiner & R.W.Schafer, *"Digital Processing of Speech Signal."*, Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978

[2] L. R. Rabiner & Biing-Hwang Juang, *"Fundamentals Of Speech Recognition."*, Prentice -Hall, AT&T, U.S.A, 1993

[3] Sadaoki Furui., "Digital Speech Processing, Synthesis, and Recognition., Marcel Dekker INC., 1992.

[4] Sadaoki Furui, Sondhi., "Advances in Speech Signal Processing.", Marcel Dekker INC., 1992.