포만트 정보를 이용한 음성인식률 개선에 관한 연구

신동성, 이윤주, 송종회 배명진
숭실대학교 정보통신학과
E-mail : mjbae@saint.soongsil.ac.kr

On a Study of the Improvement of Speech Recognition with Formant Information

Dongsung SHIN, Yoonjoo LEE, Jonghoi SONG Myungjin BAE
Dept. of Telecommunication, Soongsil University
E-mail : mjbae@saint.soongsil.ac.kr

요약문

개인용 컴퓨터가 멀티미디어 환경으로 변함에 따라서 인식률 향상과 처리시간 단축을 요구하고 있다.

본 논문은 기준패턴의 수가 증가함에 따라 발생하는 처리시간 증가 문제의 해결과 인식률 향상에 관한 것이다.

기준패턴의 수를 줄이기 위한 방법으로 각 모음별 포만트 정보를 구한 뒤 시험패턴과 비교할 후보자를 미리 정하여 인식률을 향상시키는 방법을 제안하고자 한다.

위와 같은 방법으로 모의 실험한 결과 전체 시스템 인식률이 기존의 방법에 비하여 0.5% 정도 향상되었고, 처리시간은 10%정도 감소하였다.

1. 서론

인간의 음성은 허파에서 나온 공기가 성문(vocal cord)과 성도(vocal tract)을 통과하면서 기본 주파수와 공명주파수의 특성을 가지게 된다[1][2]. 이러한 음성신호가 귀를 통하여 인지 할 수 있게 되는데 음성인식시스템은 사람이 기계를 사용함에 있어 보다 편리하게 사용하기 위한 것이다. 즉 HCI(Human Computer Interface)란 사람이 컴퓨터를 사용할 때 흔히 사용하는 입력 장치 뿐만 아니라 인간의 음성을 사용하여 명령을 할 수 있게 하는 것이다. 근래에 들어서 이러한 연구는 활발히 이루어지고 있고 상용화된 제품들도 나오고 있다.

이러한 멀티미디어 PC 환경 내에서 음성신호를 기계가 인지 할 수 있게 하기 위해서는 음성신호와 배경잡음을 구별하고 음성신호 고유의 특징을 추출해야 한다. 그리고 이를 비교하여 인식하는 것이다. 본 논문에서는 여러 가지 인식 알고리즘 중 음성신호를 확률적으로 모델링한 HMM(Hidden Markov Model)을 사용하였다. HMM 또한 미리 특징벡터를 추출해 놓은 기준패턴과 현재 들어온 음성 신호의 특징 벡터를 비교하는 것이다. 따라서 비교할 기준패턴의 수가 증가하면 할수록 많은 처리시간을 요구하게 된다.

본 논문에서는 이러한 문제점을 해결하기 위하여 초기 3 프레임 값들의 LPC 평균값과 표준편차 값을 구하여 기준패턴과 시험패턴을 비교한 후 비교할 패턴의 수를 줄이는 방법을 제안하고자 한다.

2. 인식시스템의 구성

기본적인 음성인식 과정은 그림 2-1과 같다[2]. 먼저 발성한 음성으로부터 특징벡터를 추출하여 기준패턴으로 삼는다. 그 후 사용자

가 발성을 한 음성에서 추출한 시험패턴과 기준패턴을 인식 알고리즘을 사용하여 비교한다.

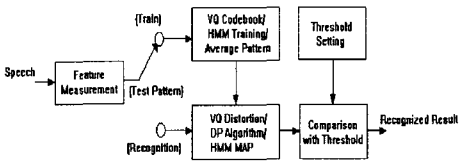

그림 2-1. 음성인식시스템의 개요도

2.1 선행처리

정확한 음성구간을 구별하지 못한다면 그에 따라 인식률이 상당히 낮아진다. 따라서 특징벡터를 추출하기 전에 배경잡음과 음성구간을 구분하여야 한다. 이를 음성구간검출이라고 한다. 음성구간검출에는 독립적 방법, 포함적 방법, 혼합적 방법이 있다. 독립적 방법은 인식과정과는 독립적으로 수행하는 방법이고 포함적인 방법은 음성구간검출을 인식과정과 동시에 수행하는 방법이다. 마지막으로 혼합적인 방법은 독립적 방법과 포함적 방법 두가지를 같이 사용하는 방법이다. 음성구간검출에서 시간이 많이 소요된다면 그만큼 전체처리시간이 길어지기 때문에 계산량이 적어야 한다. 따라서 독립적 방법이 구간검출성능에서는 포함적 방법에 못미치지만 빠르게 음성구간을 찾아낼 수 있기 때문에 본 논문에서는 이 방법을 사용하였다. 음성구간 검출과정은 그림2-2와 같다. 유성음이 배경 잡음보다 에너지 값이 큰점을 이용하여 유성음 구간을 검출하고 앞 뒤 수 프레임의 영교차율(ZCR: Zero Crossing Rate)을 이용하여 무성음구간을 검출한다[1].

유성음의 경우 저차 포만트에서 고차 포만트로 갈수록 에너지 값이 감쇠한다는 특징이 있다. 스펙트럼 상에서의 동등비교를 위하여 음성신호를 Preemphasis Filter를 사용하여 고주파항의 영향을 높여준다.

2.2 음성 특징 추출

음성신호는 성도내의 조음기관의 모양에 의하여 음운학적 의미가 달라지게 된다[1]. 따라서 음성인식은 성도특성을 이용하게 된다.

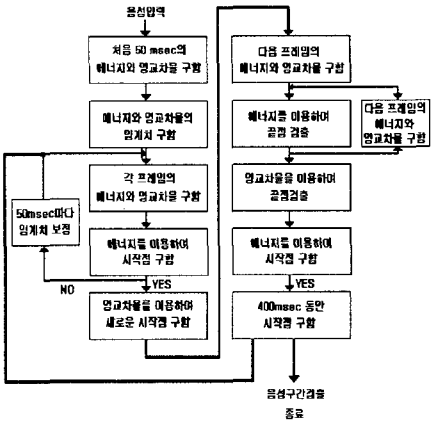

그림 2-2. 음성 구간 검출

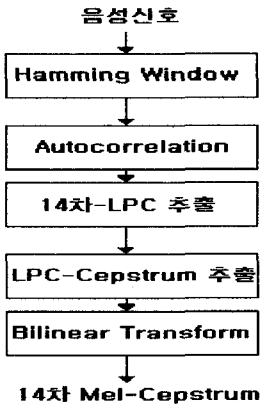

그림 2-3. 특징벡터 추출과정

음성신호는 단구간적으로 보면 안정구간이므로 예측이 가능하다. 따라서 음성신호를 선형예측필터를 사용하여 모델링 할 수 있으며 이때 추출되는 필터 계수를 선형예측(LPC : Linear Prediction Coding)계수라 한다. 선형예측계수는 성도의 특성인 각각의 포만트의 중심주파수와 대역폭을 나타낸다[3]. 그러므로 선형예측계수를 음성특징 파마미터로 사용할 수 있다. 이러한 선형예측계수를 켑스트럼(Cepstrum)으로 변환하고 인간의 귀의 특성을 반영한 멜 스케일(Mel Scale)이나 바크스케일(Bark scale)로 변환한다. 이 때의 계수를 멜켑스트럼 계수(MFCC : Mel Frequency Cepstrum Coefficient)라 한다. 본

논문에서는 음성특징벡터로 14차 멜켑스트럼 계수를 사용하였다[1][4][5]. 음성특징벡터 추출과정은 그림 2-3과 같다.

3. 인식 알고리즘

음성신호에 들어 있는 정보는 일반적으로 짧은 간격의 전력 스펙트럼에 들어 있고 시간이 지나감에 따라서 약간씩 변하게 된다, 그러므로 전력스펙트럼과 그것의 시간적 변화를 추정하는 방법을 필요로 하게 된다. 이와 같은 목적으로 성질이 비안정한 음성을 모델링해서 음성인식에 사용하는 것이 HMM이다[2].

HMM에는 몇 개의 상태가 있고 각 상태에는 각각의 불규칙함수에 의해서 하나의 출력을 내고 천이확률에 의해서 다음 상태로 넘어간다. 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 관측 심벌을 확률적으로 모델링한다면 그 모델링 파라미터는 다른 음성과는 다른 성질을 나타낼 것이다. 이러한 방법으로 각각의 음성을 구별하여 인식하는 방법이 HMM이다[2][5][6].

본 논문에서는 4상태 left to right HMM을 사용하여 음성인식을 수행하였다 이 모델은 그림3-1과 같다[2][5].

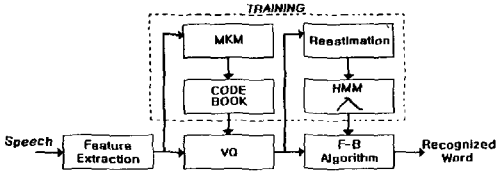

그림 3-1. 인식 알고리즘

4. 제안한 알고리즘

음성신호를 선형필터를 사용하여 모델링할 경우 이 필터의 계수를 선형예측계수라 하며 그 특성은 조음기관인 성도의 특성을 나타낸다. 이는 주파수축에서 포만트의 대역폭과 대역주파수를 나타낸다. 그림 4-1에서 보는 봐야 같이 선형예측계수는 유성음과 무성음에 따라서 각각 다른 평균값과 표준편차을 갖는 가우시안 분포를 따르게 된다[1]. 따라서 본 논문에서는 이 특징을 이용하여 시험패턴과 기준 패턴의 초성이 유성음인지 무성음지를 비교하여 기준패턴의 수를 감소시켰다. 즉, 각각의 기준패턴에 대하여 초기 3프레임의 평균값과 표준편차 값을 구하고 시험패턴의 초기 3프레임의 평균값과 표준편차 값과 비교하여 후보자 수를 줄이는 방법을 사용하였다.

이를 위하여 식 4.1을 사용하였다.

$$DIFF_{AVE} = \sum_{i=1}^{14} |A_T(i) - A_R(i)| \quad \text{(식 4.1)}$$

여기서 A_T는 시험패턴에서 구한 각 계수들의 평균값이고 A_R는 기준패턴에서 구한 각 계수들의 평균값이다.

본 논문에서는 전체 기준패턴에서 60%의 값을 위와 같은 방법으로 비교할 기준패턴으로 선정하고 인식알고리즘에 적용하였다.

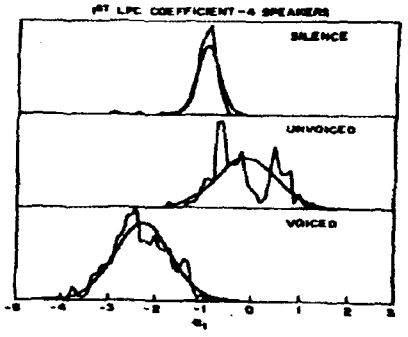

그림 4-1. 유무성음의 확률적 분포

5. 실험 및 결과

본 논문에서 구현한 음성인식 시스템의 전체적인 구성은 그림 5-1과 같다.

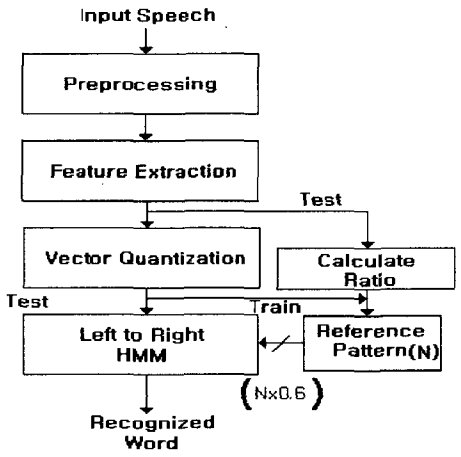

그림 5-1. 제안한 알고리즘의 구성도

제안한 방법을 실험하기 위하여 IBM-PC pentium Ⅲ 450Mhz에 16-bit A/D변환기를 사용하였다. 음성시료는 11,025Khz로 샘플링하였고 16bit로 양자화하였다. 30msec의 Hamming Window를 사용하였고 감쇠되는 부분을 보상해 주기 위하여 15msec씩 오버랩시켰다. 인식 알고리즘은 4상태 left to right HMM을 사용하였고 modified K-Means 알고리즘을 이용하여 벡터양자화를 수행하였다. 코드북의 크기는 512로 하였다. 시스템의 등록단어는 20대 남녀 화자 10명이 각각 10번씩 발성한 10개의 컴퓨터 명령어를 사용하였다.
모의 실험한 결과는 다음과 같다.

표 5-1 처리시간(초)

	Test
기존의 방법	1.52
제안한 방법	1.35

표 5-2. 전체인식률(%)

	전체 인식률
기존의 방법	97.5
제안한 방법	98

전체인식률은 0.5% 증가하였고, 처리시간도 10%정도 감소하였다.

6. 결론

개인용 컴퓨터가 멀티미디어 환경으로 변함에 따라서 사용자들을 충족 시키기 위하여 높은 인식률과 적은 처리시간이 필요하다. 이를 위하여 잡음 환경내에서의 인식률 향상에 관한 연구와 대용량 어휘인식에 관한 연구가 이루어지고 있다. 본 논문에서는 컴퓨터 명령어를 이용한 고립단어 음성인식에서 인식률 향상과 처리시간을 감소시키기 위하여 포만트 정보를 이용하였다. 실험결과 기존방법에 비하여 0.5%의 인식률 향상과 10%정도 처리시간을 단축 시킬수 있었다.

7. 참고문헌

[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]Shuzo Saito &Kazuo Nakata "Fundam-entals of Speech Signal Process-ing" Academic Press, 1985
[4] Sadaoki Furui., "Digital Speech Processing, Synthesis, and Recognition., Marcel Dekker INC., 1992.
[5] Sadaoki Furui, Sondhi., "Advances in Speech Signal Processing.",Marcel Dekker INC., 1992.
[6] 신동성, 오세영, 이윤주, 배명진, "음성인식 시스템의 처리시간 단축에 관한 연구," 한국통신학회, 한국통신학회 추계학술발표회 논문집(하), Vol.18, No.2, p.1765-1768, 1999년