파마메터 공간을 이용한 화자인식에 관한 연구

이용우*, 임동철, 이행세
아주대학교 전자공학과

A Study on Speaker Recognition Using MFCC Parameter Space

Yong-woo Lee*, Chol-dong Lim, Haing Sea Lee
Department of Electronic Engineering, Ajou University
E-mail : moriset@lycos.co.kr

Abstract

This paper reports on speaker-Recognition of context independence-speaker recognition in the field of the speech recognition. It is important to select the parameter reflecting the characteristic of each single person because speaker-recognition is to identify who speaks in the database. We used Mel Frequency Cesptrum Coefficient and Vector Quantization to identify in this paper. Specially, it considered to find characteristic-vector of the speaker in different from known method; this paper used the characteristic-vector which is selected in MFCC Parameter Space. Also, this paper compared the recognition rate according to size of codebook from this database and the time needed for operation with the existing one. The results is more improved 3~4% for recognition rate than established Vector Quantization Algorithm.

1. 서 론

최근 인터넷 관련 서비스의 급격한 증가와 더불어 음성을 이용한 고객인증의 필요성이 급증함에 따라 화자 인식을 이용한 고객 인증에 관한 관심이 증대되고 있다. 음성을 이용한 화자 인증은 카드 및 키 등과 같은 인공적인 수단보다는 매우 편리하며 음성은 분실 위험이나 도난위험이 전혀 없어 매우 안전하다. 화자인식 시스템은 화자검증(Speaker Verification) 및 화자식별(Speaker identification) 시스템으로 크게 나뉜다. 화자검증이란 검증을 요구하는 화자의 음성과 그 화자의 등록된 기준 패턴을 비교하여 미리 정해놓은 임계값(Thereshold)을 넘어서면 승인하고 반대의 경우엔 거절(Reject)하는 것이고, 화자 식별은 고립단어 인식 과정과 유사한 것으로 등록된 표준 패턴 중에 어떤 화자의 패턴과 입력음성의 패턴이 가장 유사한가를 비교하여 화자를 결정하는 것이다. 본 논문에서는 화자 인식을 시스템의 성능 개선을 위해 벡터 양자기 설계에 있어서 화자특성을 잘 반영하는 13차 MFCC계수들을 이용하여 한 음성에 대해 10번씩 발음한 5명의 화자들의 특성 벡터 중에서 동일음성에 대한 정보를 제거하고 화자 개개인의 특성만을 나타내는 벡터들만을 학습데이터로 사용한다.
화자 확인 및 인식은 VQ에 기반한 방법을 이용한다. 음성신호에 대한 벡터 양자화 실험을 한 결과 개선된 방법이 전체 특징벡터들을 적용한 결과보다 화자 인식률의 개선과 연산량이 감소함을 확인하였다. 본문의 구성은 다음과 같다. 2장에서는 화자인식기에 관해 설명하고 화자인식의 특징 파라메터들에 대해 기술한다. 3장에서는 기존의 코드북 생성 알고리즘과 제안알고리즘을 비교한다. 4장에서는 실험환경 및 기존의 방식과 제안된 방식의 성능평가결과를 보이고 마지막 5장에서는 결론 및 고찰에 대하여 기술한다.

2. 화자인식

본 화자 인식기의 블록구조를 아래의 그림2-1에 보인다.

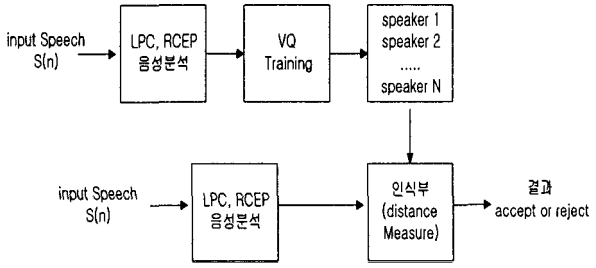

그림 2-1 본 화자인식 시스템의 블럭도

화자인식기는 그림 2-1과 같이 크게 음성분석, 훈련부, 인식부로 나눌 수 있다.

2.1 전처리 과정

음성 신호처리에서 전처리 과정은 그 과정을 어떻게 하느냐에 따라 코딩에서의 압축률과 인식부에서의 인식률에 큰 영향을 미치게 되는 중요한 과정이다.
본 논문에서는 음성 샘플은 SB-64 사운드카드로 모노 채널, 11kHz 샘플링 주파수, 16bit 양자화 레벨로 남성 5명이 10번씩 발음한 음성샘플을 채집하였다. 8번 발음한 음성은 벡터양자화의 학습벡터로 이용하며, 2번 발음한 것은 Testing에 이용한다.[5]
본 논문에서는 잡음신호 제거를 위해 첫 프레임의 에너지 평균을 전체 신호에서 제거한 후, 화자들의 음성 신호의 끝점추출을 아래에 정의된 단구간 에너지 검출과 영점 교차율을 이용하였다.[1]

$$E_s(m) = \sum_{n=m-N+1}^{m} \{ s(n)\, w(m-n) \}^2 \quad (1)$$

$$Z(m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \frac{|sgn\{s(n)\} - sgn\{s(n-1)\}|}{2} w(m-n) \quad (2)$$

$$sgn\{s(n)\} = \begin{cases} +1, & s(n) \geq 0 \\ -1, & s(n) < 0 \end{cases}$$

$s(n)$ = 샘플링값 $w(m-n)$ = 창함수

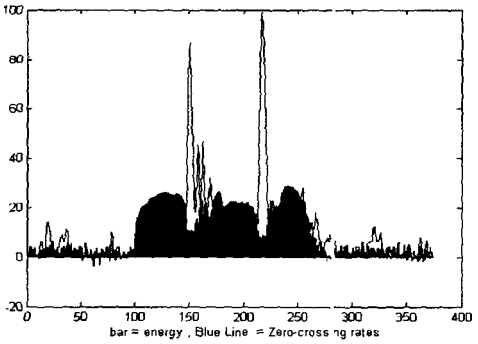

그림 2-4. "이삼사"발음의 ZCR 과 Energy

그림 3-2 "이삼사"발음의 끝점추출전

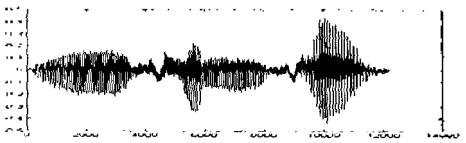

그림 2-4 "이삼사"발음의 끝점추출후

2.2 음성 분석

음성 분석부에서는 동일 음성 신호간의 일관성을 높임과 동시에 다른 화자의 음성간에 변별력을 높이기 위한 파라메터를 추출하는 부분으로 음성 인식이나 화자인식에서는 주로 LPC계수, MFCC켑스트럼, PARCOR계수 등이 사용되어진다. LPC계수는 음성의 발성기관을 하나의 필터로 가정하고 그 필터의 계수를 음성의 특징 파라메터로 사용하는 것이다. MFCC는 음성의 스펙트럼에 기초한 선형예측계수를 lifting이나 weighting 과정을 통해 완화시켜 스펙트럼의 변이성을 제거하여 음성의 정적특성을 강조하는 것이다. 본 논문에서는 현재 화자인식에 대표적으로 쓰이는 MFCC (멜 주파수 켑스트럼계수)를 이용하였다. 멜켑스트럼 20차계수를 추출하는 수식을 아래 수식(3)에 나타내었다.[4]

$$C_m(t) = \sum_{k=1}^{20} X_k \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{20} \right], \quad i = 1,2,\cdots,M \quad (3)$$

아래의 그림 2-5에는 멜켑스트럼 추출 과정의 블럭도이다.

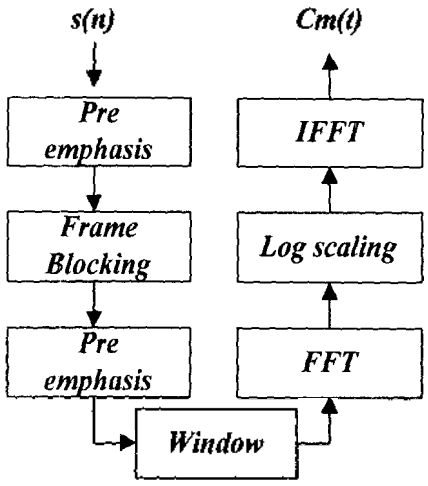

그림 2-5 MFCC 추출과정

멜 켑스트럼은 일정한 중심간격과 대역폭을 가지는 Critical Band 필터를 사용하여 구한다. 본 논문에서는 24개의 필터 뱅크를 적용하였다.

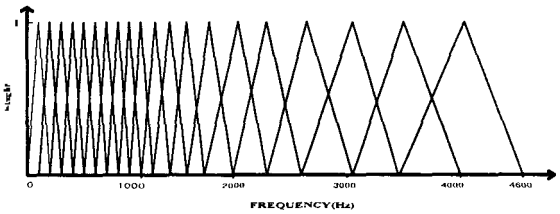

그림 2-6 멜-켑스트럼 계수를 생성하기 위한 필터

3. 개선된 백터양자화

3.1 기존 알고리즘

기존의 백터 양자화에 기반한 화자인식 알고리즘은 음성 분석부에서 생성한 LPC나 MFCC나 PARCOR 계수 등을 특징 파라메터로 이용하여 K-means와 같은 알고리즘으로 각 화자별 코드북을 생성하여 입력된 음성 데이터와 만들어둔 각 화자별 코드북과의 스펙트럼거리가 가장 가까운 것을 선택하여 그 코드북에 해당하는 화자를 인식으로 결정한다. 아래 그림 2-7에는 백터양자화 훈련과정을 보인 것이다.

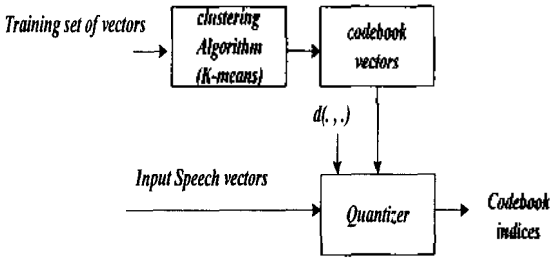

그림 2-7 VQ Training과 분류 구조 블럭도

3.2 제안된 방법

기존의 알고리즘의 문제점은 각 화자들의 특징벡터들을 학습할 때 생성된 클러스터들은 상당부분 겹치게 되어 있고 화자 확인시 변별력을 떨어뜨리게 한다. 본 논문에서 제안한 알고리즘은 코드북을 생성하는데 있어 새로운 방법을 제안한다. 즉, 여러 화자들의 백터들의 공통된 백터 군집을 제거한 후, 나머지 백터들에 한해서 K-means 알고리즘을 이용하여 코드북을 생성한다. 본 논문에서는 이 나머지 백터군을 화자 고유의 특징으로 가정한다. 다음 그림 2-8에 두 화자가 동일 음성을 10번 발음한 음성에 대한 특징 벡터를 2개의 파라메터를 이용 2차원 으로 보인 것이다.

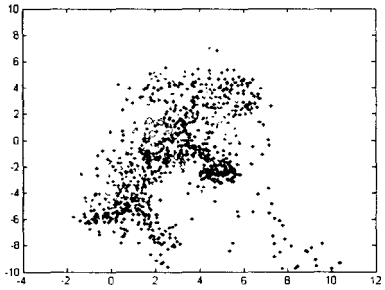

그림 2-8 화자A, B의 백터군집

즉 입력된 모든 화자들의 특징백터들의 평균값을 구한 후, 그 중점벡터를 기준으로 가장 멀리 떨어진 백터의 수가 전체 백터의 40%정도 되도록 계산한후, 그 40%의 백터들을 가지고 K-means알고리즘을 적용하여 발음한 음성에 대한 개인별 화자 코드북을 생성한다. 아래의 그림2-9는 전체 발음음성에 대한 공통부분을 제거한 후의 벡터들을 2차원으로 나타내었다.

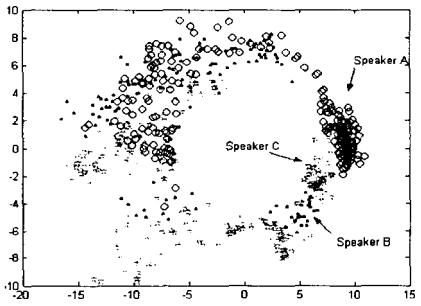

그림 2-9 화자 A,B,C의 공통된 영역 벡터 제거 후

K-means 의 클러스터링은 다음과 같은 알고리즘[4]을 갖는다. 먼저 공통백터를 제거한후의 백터를 K개의 초기 cluster로 나누거나 또는 임의로 K개의 Seed로 선택한후, 각 클러스터의 중심(Centroid)를 구한다. 다음 임의의 한 데이터를 선택하여 각 클러스터 중심 까지의 거리를 계산한다. 만일 그 데이터와 클러스터 중심까지의 거리중 가장 가까운 것이 자신이 속한 클러스터라면 그대로 보존하고 그렇지 않으면 중심과의 거리가 가장 가까운 클러스터에 재할당한다. 위의 과정을 모든 데이터에 대해 재 할당이 없을 때 까지 반복한다. 화자인식은 입력된 음성 데이터를 훈련과정을 통해 13차 MFCC를 프레임별로 추출하여 전체 데이터에서 구한 평균값과 분산값을 이용하여 버려진 공통인 구역경계에 포함되는 데이터인가를 확인한후, 그 구역에 속하면 그 값을 무시하고 나머지 data 만을 가지고 정해진 수의 대표값 벡터를 구한다.

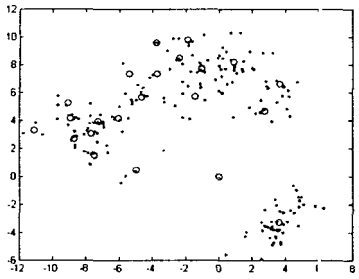

그림 2-9 화자A의 고유특성과 Centroid

다음 등록된 화자의 코드북과 비교되는데 본 논문에서의 판단 척도로서는 두 패턴간의 유사도(거리값)가[1,2] 사용되며 가장 거리값이 작은 코드북의 화자가 승인된다. 그림 2-10에는 한 화자에 대한 중복된 특징 벡터를 제거한 다음 고유특성에 대한 centroid를 구한 것이다.

4. 실험 환경 및 결과

본 논문에서 사용된 음성 데이타는 샘플링 주파수 11.025KHz이고, 512Point의 음성 샘플을 한 프레임으로 하여 128Point씩 Shift하여 1/4씩 중첩 되도록 하였고, 창함수(window)는 Hamming Window를 취한 후, 12차의 멜 켑스트럼을 구하였다[4]. 사용된 음성 데이터는 조용한 연구실 환경에서 구성된 5명의 남성 화자가 동일음성 10번씩 발음한 8개의 총 400개의 음성 데이터를 대상으로 실험을 하였다. 8번 발음한 음성은 코드북 생성에 입력될 Training Vector로 쓰이고 나머지 2번 발음한 음성은 화자인식 테스트에 이용된다. 표 1에 LPCC와 기존의 벡터양자화를 사용한 코드북 사이즈에 따른 화자들의 인식률(FA) 및 오류율을 나타내었다. 표 1 과 2에 나타난 숫자는 100번의 사칭 실험과 인증 실험을 하여 구한 오류횟수를 나타낸 것이다.

화자 / 코드북	A		B		C		D		E		연산 시간
	FR	FA	FR	FA	FR	FA	FR	FA	FR	FA	
32	14	10	13	9	13	12	11	12	11	9	13.9
64	9	8	10	8	10	9	10	10	8	9	29.3
128	8	7	7	5	6	5	7	6	6	4	60.5

표 1 LPCC 와 기존의 벡터양자화를 사용한 화자인식오류율

화자 / 코드북	A		B		C		D		E		연산 시간
	FR	FA	FR	FA	FR	FA	FR	FA	FR	FA	
32	9	9	10	8	7	8	9	8	8	8	7.74
64	7	9	9	8	6	7	7	6	7	8	14.3
128	5	4	4	3	4	3	4	3	3	5	24.2

표 2 MFCC 와 제안된 벡터양자화를 사용한 화자인식오류율

5. 결론

본 논문에서는 화자들간의 중복된 특징벡터로 인한 벡터양자화의 문제점을 해결하기 위하여 각 화자들의 변별력 있는 화자 고유특징 벡터를 벡터양자화에 적용하는 방법을 제시한다. 중복된 특징벡터들을 제거한 후, 벡터 양자화를 이용한 화자인식에 적용한 결과 연산량이 1.8배정도 크게 감소하였다. 또한 인식률에 있어서도 기존의 방법보다 성능이 3~4%개선되었지만 실험 결과에서 나타난 인식률은 화자의 수에 따라 변화가 있으리라 생각된다. 32개, 64개인 코드북 사이즈의 일때의 인식률은 크게 차이가 나지 않음을 확인했다. 본 논문에서 제안한 방법은 화자 고유 특성을 갖는 적은 양의 훈련 데이터를 가지고도 개선된 성능을 보임을 알 수 있다.

참 고 문 헌

[1] Lawrence Rabiner, Bling-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall International inc,

[2] Shikano, K.(1985) 'Text-independent Speaker Recognition experiment using codebooks in vector quantization' J. Acoust. Soc Am.

[3] Gish and M. Schmidt, 'Text-Indepedent Speaker Indentification' IEEE Signal Processing Magazine. October, 1994.

[4] John R. Deller, Jr. & John G. Proakis 'Discrete Time Processing of Speech Signals'

[5] V. Bjorn and S. Roweis. Speaker Recognition System. Technical Report CNS/EE248, Califonia Institude of Technology, 1995