韓國 OA學會 論文誌
第5卷 第3號, 2000. 9.

2000-5-3-3-5

음성을 이용한 화자 검증기 설계 및 구현

지 진 구*, 윤 성 일**

Design and Implementation of Speaker Verification System Using Voice

Jin-Koo Ji*, Sung-Il Yun**

요 약

본 논문은 음성을 이용하여 개인의 신원을 확인할 수 있는 화자 검증시스템을 설계, 구현하였다. 특징 파라메터로는 선형 예측 계수나 고속 후리에 변환보다 안정적이고 계산량이 적은 장점이 있는 필터뱅크(filterbank)를 사용했으며 추출된 파라메터들을 LBG알고리즘을 이용하여 각 개인의 코드북을 작성하였다. 작성된 코드북에 의해 특징 파라메터를 벡터양자화하여 얻어진 코드열로 화자 검증의 참조 패턴 및 입력 패턴을 생성, 이들을 동적시간 정합법을 이용하여 유사도를 측정하여 얻어진 유사도와 임계값을 비교하여 음성 의뢰자(client speaker)인지, 사칭자(impostor)인지 결정하는 화자 검증기를 설계, 구현 하였다.

Abstract

In this paper, we design implement the speaker verification system for verifing personal identification using voice. Filter bank magnitude was used as a feature parameter and code-book was made using LBG algorithm. The code book convert feature parameters into code sequence. The difference between reference pattern and input pattern measures using DTW(Dynamic Time Warping). The similarity measured using DTW and threshold value derived from deviation were used to discriminate impostor from client speaker

* 도우넷 웹 기술연구소 연구원
** 공주영상정보대학 컴퓨터정보과 조교수

Ⅰ. 서론

컴퓨터의 발달과 함께 사회는 점점 복잡해지고, 급속한 정보의 교류가 필요하게 되었다. 특히, 상거래에 있어서, 현재 신용카드를 이용하여 원격지에서 물건의 구매가 가능해지게 되었다. 그러나 이 경우, 익명으로 구입가능하기 때문에 신용카드 범죄에 의해 매년 수많은 피해가 발생 된다. 이러한 신용 범죄를 줄이기 비밀번호, 열쇠, 배지 등과 같은 개인의 독특한 소유물을 확인함으로써 이루어진다. 그러나 이러한 방법은 분실, 도난, 위조 등의 가능성이 있고, 이로 인하여 심각한 보안상의 문제가 야기될 수도 있다. 따라서 신체특성에 기반을 둔 개인확인 수단을 이용하여 보안에 대해 강화하고자 하는 노력이 일고 있다. 이러한 시도의 원시적인 형태는 개인의 글씨 특성인 서명(signature)을 이용하는 방법에서 찾아볼 수 있으며 최근에는 지문이나 손 모양(hand geometry), 망막검사 등을 이용하는 방법을 이용하여 물리적인 출입통제에는 적용 가능하지만, 전화 및 인터넷상에서는 쉽게 사용할 수 없어 자연스럽고 사용하기 쉬운 음성은 개인 확인 수단으로 사용할 수 있다. 특히 사람의 음성은 의사 소통의 가장 자연스러운 수단으로 여타의 다른 방법에 비해 비교적 거부감이나 부담스러움이 없는 방법이다. 특히 최근에는 디지털 신호처리 및 음성처리 기술과 컴퓨터 기술이 발달함에 따라 음성을 이용한 화자 인식(speaker identification)이나 화자 검증(speaker verification)방법을 적용할 수 있게 되었다.[2] 본 논문에서는 사람의 음성신호를 이용하여 신원을 확인할 수 있는 시스템을 설계, 구현하였다.

Ⅱ 관련 연구

1. 화자인식기

각 개인의 신원을 확인 하는 방법으로는 크게 화자 검증과 화자 인식로 나누어 진다. 화자 검증은 본 연구 내용에서와 같이 발성된 음성이 원하는 화자(의뢰인, client speaker)인지 아닌지(사칭자, impostor)를 구분해 내는 것으로 의뢰인에 대한 음성등록이 요구하게 된다.

따라서 화자 검증시스템의 중요 핵심 부분은 화자의 음성이 의뢰인의 음성인지를 검증하는 부분이다. 이러한 검증부분에 대해서는 많은 연구가 있었으며, 본 논문 또한, 이러한 연구중 한 분야의 내용이다.

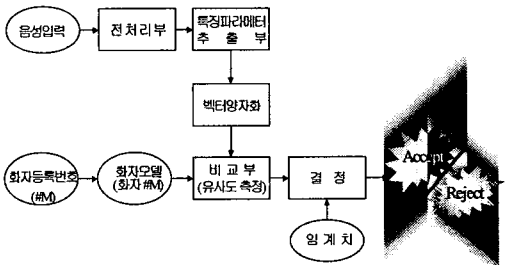

그림 1. 화자 검증 시스템의 구성

화자 인식은 화자 검증과는 달리 등록된 N명의 사람들 중 유사도가 가장 높은 사람을 찾아내는 과정이다. 따라서 화자인식 집합의 크기에 따라 오류율이 좌우된다. 또한 화자 인식 시스템은 현재 입력 받음 음성에 대한 유사도 측정이므로 등록되어져 있는 모든 화자에 대한 음성 데이터와의 비교과정이 있으므로 인식의 시간적인 효율성에서 다소 시스템상의 성능에도 영향을 많이 미친다. 이러한 화자 인식 시스템은 폐쇄형 시스템(closed set system)과 개방형 시스템(open set system)으로 나뉘어 진다.[그림 2]

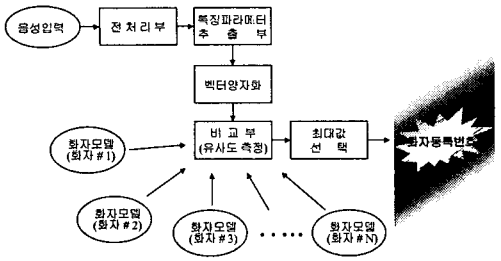

그림 2. 화자 인식 시스템의 구성

2 한국어 음성의 특징

본 장에서는 음성의 음향학적인 특성을 기반으로 한국어 음성을 분석한다. 음성을 분류하는 기준은 조음적인 측면과 음향학적인 측면의 두 가지 방법이 있는데, 파열음이나 마찰음 등으로 분류하는 것은 조음적인 기준이 적용된 것이고, 포르만트(formant)나 공명도(sonority)등을 통하여 분류하는 것은 음향학적인 기준이 적용된 것이다

2.1 자음의 특성

조음적 기준에 따르면 자음과 모음은 구강 내의 기류 장애 여부에 의해 구분된다. 이때, 허파에서 올라오는 공기가 성도(vocal tract)의 어딘가에서 장애(obstacle)를 받으면 자음(consonant)이 된다.

자음은 혀가 성도의 어떤 부분에 좁은 장소를 만들어 폐로부터 올라오는 공기류를 난기류로 바꿔 생기는 잡음적인 음성과, 혀나 입술로 성도를 차단함으로서 공기류를 일시적으로 멈추게 했다가 이를 급히 개방해서 생기는 임펄스(impulse)적인 음성 및 구강을 닫고 비강으로부터 방사하는 음성등으로 나눌 수 있다. 이들을 파열음, 마찰음, 파열음, 파찰음, 설측음, 비음이라고 부른다.

허파에서 올라오는 공기를 차단한 후에 공기의 압력이 높아진 상태에서 구강안의 압축된 공기를 급속히 밖으로 빠져 나가게 함으로써 발생되는 폭팔음을 파열음이라 한다. 즉 파열음은 성도의 어떤 부분을 폐쇄함으로서 발생되는데 이 음의 가장 큰 특징은 그림 3의 (a)에서 보는 바와같이 스펙트럼상에서 40ms에서 120ms사이의 길이를 갖는 공백으로 나타난다는 것이다.

(a)바다 (b)소리 (c)추위 (d)라디오 (e)동그라미

그림 3. 자음 발성

성도의 어떤 부분을 협착시켜 그곳을 공기가 무리하게 빠져나오면서 나는 소음이다. 즉, 좁은 통로로 공기를 분출시켜 공기를 윗니의 모서리에 부딪쳐서 소음을 얻는데 그림 3의 (b)에서 보는 바와 같이 스펙트럼상에서는 3000 ~ 5000Hz 사이에서 소음의 흔적을 볼 수 있다.

이는 파열음과 마찰음의 중간적인 성격을 갖는 발음이며 비교적 마찰음에 가까운 형태로서 스펙트럼상에서는 공백이 계속 지속된다는 특징을 갖는다. 즉, 그림 3의 (c)에서 보는 바와 같이 70ms에서 140ms가량 계속되는 잡음이 그 특징으로 나타난다.

혀끝을 윗니 바로 뒤에 두고 혀의 좌우 혹은 어느 한쪽과 입천장 사이에 불안전한 폐쇄를 형성한 뒤 공기가 혀의 좌우로 빠져나가게 함으로써 이루어지는데, 이때 마찰음 보다는 적은 소음이 발생한다.

입술을 다물어 허파에서 올라오는 공기를 완전 차단한 뒤 비강으로 공기를 내보내며 내는 소리를 말한다. 'ㅂ'의 경우에는 축적된 공기를 입밖으로 파열시켜 내보내는데 비하여 'ㅁ'의 경우에는 연구개를 내려 공기가 인강으로부터 비강을 통해 코에서 밖으로 나가도록 한다.

2.2. 모음의 특성

모음의 음질(vocal quality)은 그림 4에서 보는 바와 같이 제 1 포르만트, 제 2 포르만트에 의해 거의 결정되며, 제 3 포르만트 이상은 진동수가 지나치게 높기 때문에 소리의 높이를 정의할 변별적 기준으로 사용하기 어렵다. 특히, 제 4 포르만트 이상은 화자 개인의 음질적 특성을 반영하는 경우에 사용된다. 모음을 낼때에는 발성기관들이 서로 가까이 접촉하는 일이 없으며 기류의 이동이 비교적 자유롭다. 모음은 공명실의 모양을

조절하는 혀의 위치와 입술 모양 즉, 혀의 전후위치와 혓몸의 고저, 그리고 입술의 원순성 여부에 의하여 분류된다.

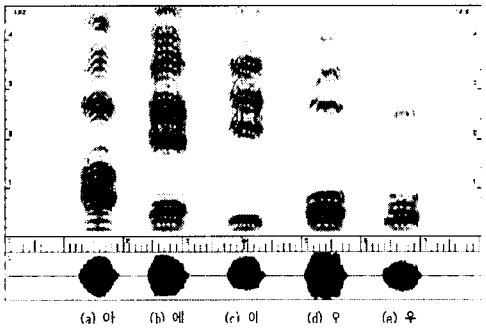

그림 4. 모음 발성

모음에 대한 전통적인 조음 기술은 포르만트의 진동수들과 관련을 가지고 있는데 일반적으로 원순화음이 될수록 포르만트의 진동수는 줄어든다. 예로서, 전설모음 '이'에서는 제 3포르만트에서, 후설모음 '우'에서는 제 2포르만트에서 각각 특징이 더 크게 나타난다. 따라서, 모음의 피치는 혀의 높이 보다는 제 1 포르만트의 진동수로서 더 자세하게 구할 수 있다. 그리고, 전설, 후설의 구분은 혀의 실제 위치에 대한 측정보다는 제 1, 2 포르만트의 진동수들간의 차이로 표현하는 것이 보다 바람직하다.

Ⅲ. 화자 검증기 구성

본 논문의 화자 검증기는 그림 3과 같이 음성 패스워드 등록과 검증부분으로 구분되어 설계하였다.

첫 번째 음성 등록부분은 입력 받은 음성에 대한 전처리부분, 특징 파라메타 추출 및 코드북 생성부분, 코드양자화, 임계값 결정부분으로 구성되었다.

두 번째 음성 검증부분은 음성 등록과 동일한 방법으로 음성에 대한 특징을 추출하여, 동적시간 정합법을 이용하여 임계값과 비교하여 승인 여부를 결정한다.

1. 음성 등록 부분

등록과정은 동일한 패스워드를 3회 반복 발화하여 각각의 발음된 음성에 대해 전처리 및 특징 추출과정과 벡터 양자화를 거쳐 생성된 서로다른 3개의 데이터를 신간측 정합을 통해 일정한 길이의 참조패턴을 생성하게 되고, 최종적으로 생성된 참조 패턴, 코드북, 임계값은 보조기억장치에 대응되는 개인의 ID와 함께 저장이 된다.

1.1 끝점 검출

음성인식 시스템은 우선 미지의 음성 파형이 입력되면 이를 음성과 비음성으로 구분하기 위해 시작점과 끝점을 찾아 내야 한다. 본 연구에서는 에너지를 이용하는 수정된 Rabiner & Sambur방식[5]을 사용하였다. 이는 다음의 그림에서 보인 바와 같이 음성의 특성상 영교차율을 끝점 검출에 사용할 수 없어, 에너지와 영교차율을 같이 사용하는 일반적인 Rabiner & Sambur 방식을 적용할 경우 정확한 끝점 검출이 이루어지지 못하기 때문이다.

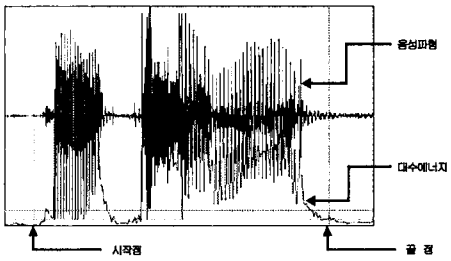

그림 5. 음성파형

그림 5는 파형에 대한 대수에너지를 구한 것이다. 대수에너지의 경우 음성이 있는 구간에서 에너지가 커지는 것을 알 수 있으며 비음성 구간에서는 거의 나타나지 않는 것을 알 수 있다. 따라서 본 연구에서는 끝점 검출을 위한 특징으로 대수 에너지를 사용하였다. 끝점 검출에 대수 에너지만을 사용하므로 정확한 시작점을 찾지 못할 수도 있으며 사람의 발음 특성상 시작 부분과 끝부분의 불명확함을 고려 실제 시작점과 끝점을 계산된 것보다 각각 5프레임씩 연장하여 끝점을 결정하였다.

1.2 특징 추출부분

아날로그 음성 신호로부터 특징 벡터를 추출하기 위하여 본 연구에서는 전처리 단계로 저대역 통과기, A/D 변환, 해밍윈도우(hamming window)를 적용하여, 특징 파라메터로는 필터뱅크를 사용 하였다.

필터뱅크 계수 또는 선형 예측 계수를 사용한 경우 양쪽 모두 높은 인식률을 보였으나 필터뱅크 계수가 비교적 간단한 계산 방식으로 되어 있어 계산량이 적고 선형 예측 계수보다 상대적으로 안정적인 특징을 보이기 때문에 본 연구에서는 최종적으로 필터 뱅크를 특징 파라메터로 선정하여 구현하였다.

필터 뱅크란 주파수 대역 위에 여러 개의 대역 제한 필터를 두어 이 필터를 통과시켜 출력된 값을 파라메터로 사용하는 방법이다. 필터뱅크는 선형 예측 계수나 고속 후리에 변환보다 안정적이고 계산량이 적은 장점이 있다. 필요한 대역에 여러 개의 필터를 두기 위해 기본 주파수를 다음공식에 의해 구한다.

$$f[i] = f_i \times 2^{(\frac{i}{q})} \qquad (1)$$

여기서 f_1은 초기 중심 주파수이며 q는 옥타브를 지칭하는 것으로, 초기 중심 주파수에서 q옥타브 간격으로 필터뱅크 계수를 구한다. 선형 예측 계수는 화자 인식 및 검증에 일반적으로 쓰이는 특징 파라메터이다.

1.3 코드북 생성

코드북을 구성하는 방법에는 단어마다 각각의 코드북을 두는 방법과 전체 단어를 하나의 코드북으로 두는 방법이 있다. 본 연구에서는 등록인 각 개인마다 코드북을 작성했다.

벡터양자화를 이용한 음성 인식 시스템에서 코드북은 학습용 데이터의 특성이 잘 나타나도록 코드북을 만들어야 한다. 본 연구에서는 LBG Algorithm을 이용하여 filterbank를 바탕으로 하여 코드북을 만들었다

1.4 벡터 양자화

벡터양자화란 연속이거나 이산벡터들의 계열을 통신이나 디지털 채널에 저장하기에 적당한 디지털 계열과 매핑하기 위한 코딩 방법이다. 이 벡터양자화의 가장 큰 목적은 데이터 압축으로 대표 패턴이 저장된 코드북으로부터 이에 대응되는 양자화 값으로 차원수를 줄이거나 범위를 줄이는 방법이다.[3] 데이터의 신뢰성을 잃지 않으며, 최대 한도로 전송 bit rate를 줄이는데 있다.

2. 음성 검증 부분

음성 검증부분은 자신의 ID를 입력한 후 등록시킨 자신의 패스워드를 발음하게 되면 입력 음성은 등록과정과 동일한 방법으로 전처리 및 특징 추출과 벡터양자화 과정을 거쳐 1차원의 특징으로 변환된다. 이 특징은 곧바로 ID에 해당하는 참조 패턴과의 검증율 측정이 이루어 진다. 이때 얻어진 데이터가 임계값 범위내에 있으면 발성한 화자가 입력된 ID와 동일인임을 승인하게 된다.

2.1 동적시간 정합법

동적 프로그래밍을 기초로 하는 동적 시간 정합 음성 판별 시스템은 구현이 쉬우면서도 비교적 높은 정확도를 얻을 수 있다는 특징 때문에 널리 사용되고 있다. 입력패턴과 참조패턴의 매칭 방법은 음성이 시간 길이에 비선형적이며, 모음과 자음의 신축 정도가 서로 다른 이유 등으로 인해 올바른 매칭을 하기 위해서는 시간에 대한 비선형적인 정렬(alignment)이 필요하게 된다. 또한, 효율적으로 시간정렬을 하기 위해서는 늘이기와 압축에 대해 제약을 주어야 한다. 이와 같은 제약 조건에는 Itakura나 Sakoe & Chiba 등에 의해 소개가 된 전역 제약(global constraint), 지역 제약(local constraint) 또는 단조성(monotonicity)등이 있는데 이러한 제약 조건을 따르는 경우 극히 일부분의 처리만으로 전체적인 평가를 마칠 수 있게 된다

본 연구에서는 이러한 벡터 거리를 모두 더한 값과 판별 논리기의 임계값을 비교해 화자의 사용 승인 여부를 결정하게 된다[8].

2.2 검증과정

입력된 음성은 ID에 해당하는 화자 모델 참조패턴 3개와 서로 유사도를 비교하여 각 임계값 1과 임계값 2값과 비교하여 이 두조건을 모두 만족할 때 의뢰자를 ID와 동일한 사람으로 승인하게 된다. 아래 그림6는 임계값에 의한 화자 검증과정이다.

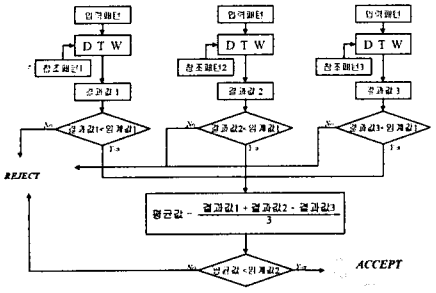

그림 6. 임계값에 의한 검증과정

Ⅳ. 실험 및 결과

1. 실험 환경

본 시스템은 IBM호환기종상에서 MS Windows 95 및 DOS를 운영체제로 하여 개발 되었다. 개발 언어는 C++언어를 사용하였으며 Borland C++ 4.0컴파일러를 사용하였다. 연구에서 사용된 음성데이타는 일반 PC에서 사운드 카드를 이용하여 음성을 입력 받았다. Mono형식으로, 11025Hz로 샘플링 하였다. 화자 구성은 남,녀 각 5명씩 10단어를 10회씩 조용한 사무실에서 발성하였다. 화자 검증은 음성인식과 달리 개인의 음향적 특성을 중요시 하므로 각 개인의 음성 등록은 자신들이 평소에 자연스럽게 발화할 수 있는 단어(전화번호, 주민등록번호)를 선정 하여 발화 하였다.

2. 실험 결과

실험의 결과는 type Ⅰ에러율과 type Ⅱ에러율로 나타낸다. type Ⅰ에러는 의뢰자의 음성을 사칭자로 오인식할경우이며, type Ⅱ에러는 사칭자의 음성을 의뢰자로 오인식할 경우다.

실험 결과, 코드북생성 방법과 특징 파라메터의 종류에따라 검증율이 각각 다르게 나타났다.

먼저, 코드북의 종류를 다음과 같이 세가지 방법으로 테스트했다. 코드북Ⅰ은 단독화자가 "0"에서 부터 "9"까지 10개의 단어로 한 개의 코드북을 작성하여 각 화자마다 한 개의 코드북을 작성하였다. 코드북Ⅱ는 10인의 화자가 발성한 10개의 음성패스워드로 한 개의 통합된 코드북을 작성하여 시스템에 단 한 개의 코드북만을 작성하였다. 코드북Ⅲ은 음성 패스워드 등록자가 자신의 등록용 음성만으로 구성된 코드북으로 각각 개인마다 한 개의 코드북을 갖는다.

그림 7과 그림 8은 각 코드북별 type Ⅰ,type Ⅱ 에러율을 보여주고 있다.

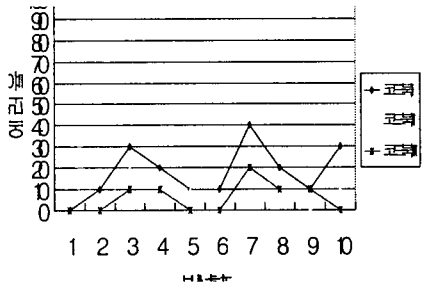

그림 7. 코드북별 Type Ⅰ 에러율

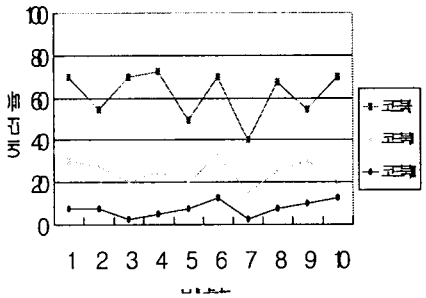

그림 8. 코드북별 typeⅡ 에러율

실험결과 그림 9와 같이 코드북Ⅰ은 등록 단어에 대한 단어적인 제약을 얻을수 있지만 각 개인의 음향적 성질에는 타당하지 않음을 볼 수 있다. 코드북Ⅱ는 음성 등록과정에서 각 개인마다 따로 코드북을 작성하지 않으므로 시간적인 효율성은 높으나 typeⅡ에러율에서와 같이 10인의 통합된 코드북이므로 각 개인의 음향적 특징을 나타내기 부적합함을 볼수 있다. 따라서 본 시스템에서는 상대적으로 높은 검증율을 보여주는 코드북Ⅲ을 이용하여 구현하였다.

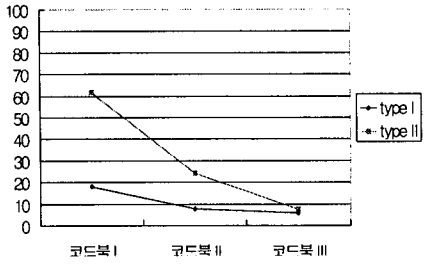

그림 9. 코드북별 에러율

또한, 어떠한 특징 파라메터가 가장 좋은 검증율을 나타내는지를 알아보기 위하여 필터뱅크, 선형예측계수를 이용하여 테스트 했다. 표1은 특징파라메라별 에러율이다.

표 1. 특징 파라메터별 에러율

	정규에너지		필터뱅크		선형예측계수	
구 분	type I	typeII	type I	typeII	type I	typeII
데이터	100	400	100	400	100	400
오인식	46	248	6	30	10	22
에러율	46%	62%	6%	7.5%	9%	5.5%

세가지 방법 모두 코드북III의 방식으로 벡터 양자화 하여 검증율은 양쪽 모두 결과와 같이 비슷하였으나, 필터뱅크 계수가 비교적 간단한 계산 방식으로 되어 있어 계산량이 적고 선형 예측 계수보다 상대적으로 안정적이기 때문에 본 연구에서는 최종적으로 필터 뱅크를 특징 파라메터로 선정하여 구현하였다.

V 결론

본 논문에서는 자연스럽고 사용하기 쉬운 각 개인의 음성을 이용하여 개인의 신원을 확인 할 수 있는 화자 검증 시스템을 구현하였다.

본 연구에서 구현한 시스템은 상대적인 대수 에너지만을 사용하여 개략적인 끝점 검출을 실행하였고, 검출된 음성부분에 대하여 몇 가지 특징 파라메터에 대한 검증율 실험의 결과를 바탕으로 특징 파라메터로는 필터 뱅크 계수를 사용하였다. 또한, 자신이 등록한 음성 패스워드로 각각 개인의 코드북을 따로 구성하는 방법을 사용하여 화자 검증율을 높였다. 벡터양자화에 대응하여 동작하는 상부 인식기로는 소규모 고립단어 인식에 적합한 동적 시간 정합법을 채택하였다.

본 논문에서는 효과적인 화자검증기를 구현하기 위한 특징파라메터 및 코드북 구성에 따른 상대적인 검증율을 실험을 통하여 측정하여, 음성을 등록한 사람에 대해 각각의 코드북을 생성하여 화자 검증기를 구현하였다.

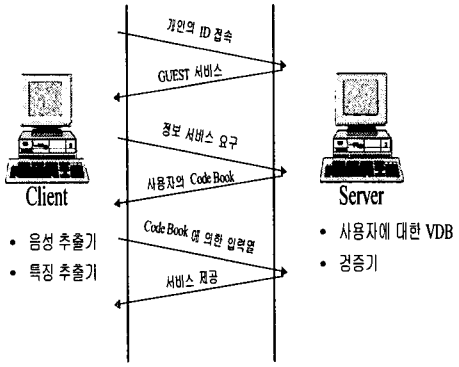

그림 10. 음성 패스워드를 이용한 전자서명

향후 연구과제로 본 시스템은 잡음에 대해서 고려하지 않았지 때문에 잡음 환경에서도 사용자의 실제 음성을 이용하여 화자 검증 할 수 있도록 개선이 있어야 할 것이며, 전자 상거래에서의 전자 서명과 같이 음성을 이용하여 신원을 확인하는 응용분야에 대한 연구가 필요하다.

참고문헌

[1] H. Gish and M. Schmidt, "Text-independent speaker identification," IEEE Signal Processing Magazine, no. 10, pp. 18-32, 1994.

[2] J.M. Naik, "Speaker verification : a tutorial," IEEE Communications Magazine, no. 1, pp. 42-48, 1990.

[3] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol, COM-28, pp. 84-95, 1980.
[4] B.S. Atal, "Automatic recognition of speaker from their voices," Proceeding of IEEE, vol. 64, no. 4, pp. 460-475. 1976.
[5] J.R. Dellek and J.G. Proakis, Discrete-time processing of speech signals, McMillan Publishing Company, 1993.
[6] R.M. Gray, "Vector quantization technique," IEEE ASSP Magazine, no. 1, pp. 4-29, 1984.
[7] H.F. Silverman and D.P. Morgan, "The application of dynamic programming to connected speech recognition," IEEE ASSP Magzine, no. 7, pp. 6-25, 1990.
[8] J. R. Deller and J. G. Proakis, Discrete-time processing of speech signals, MacMillan Publishing Company, 1993.

저 자 소 개

지진구
현재 : 도우넷 '웹 기술연구소 연구원
관심분야 : 인공지능, 멀티미디어

윤성일
현재 : 공주영상정보대학 컴퓨터정보과 조교수
관심분야 : 시스템소프트웨어, 컴퓨터통신