Dialog/4 보드를 이용한 전화음성 기반의 화자 인증 시스템의 구현

이순려, 박일구, 최홍섭
대진대학교 전자공학과

An Implementation of Telephone-based Speaker Verification System using Dialog/4 Board

Soon-Reyo Lee, Yil-Koo Park, Hong-Sub Choi
Dept. of Electronics Engineering, Daejin University
E-mail: mirage@daejin.ac.kr

요약

전화 음성용 화자인증 시스템 전화기에서 의뢰인의 ID와 음성을 입력받은 후 인증관련 처리를 PC에서 수행하여 그 인증 결과를 의뢰인의 전화로 알려주는 시스템으로 본 논문에서는 CTI(Computer Telephony Integration) 기술이 적용된 Dialog/4 보드를 이용하여시스템을 구현하였다.Dialog/4 보드를 통하여 시스템에 입력된 ID와 음성에 대하여 신호처리와 특징 추출을 거친 후 ID에 해당하는 화자모델과 배경화자 정보를 이용하여 유사도를 계산하여 의뢰인에 대한 인증 또는 거절의 결과를 알려준다. 실제의 전화음성을 이용한 화자인증 시스템이 성능평가에서 전화음성으로 실험을 할경우 99.6 %의 인증률을 보여주었다.

Ⅰ. 서론

화자 인증시스템은 음성신호에 포함되어 있는 화자간의 음향적 특징의 변이를 이용하여 발성한 사람의 ID를 확인하는 시스템으로 신원확인 및 출입통제 시스템에 이용되고 있다.

본 논문은 전화와 컴퓨터를 결합해주는 CTI 기술이 적용된 DSP 보드를 이용하여 전화망을 통하여 의뢰인의 ID와 음성을 입력받아 인증 관련 처리를 컴퓨터에서 수행한 후 의뢰인에게 전화로 그 인증 결과를 알려주는 전화용 기반의 화자 인증 시스템의 구현에 대하여 기술하였다. 논문의 중심은 기존 이론을 응용한 시스템의 구현에 두었으며, 실험에 사용한 DSP 보드는 Dialogic 사에서 만든 CTI 전용 DSP 보드인 Dialog/4를 이용하였다.

논문의 구성은 1장 서론, 2장에서 화자인증 시스템의 설계 및 구현을 위한 전반적 내용을 소개하였으며 3장에서 음성DB의 구성방법에 대하여 4장에서는 성능평가 및 결과를 그리고 5장에서는 결론 및 고찰을 기술하였다.

2. 화자인증 시스템의 설계 및 구현

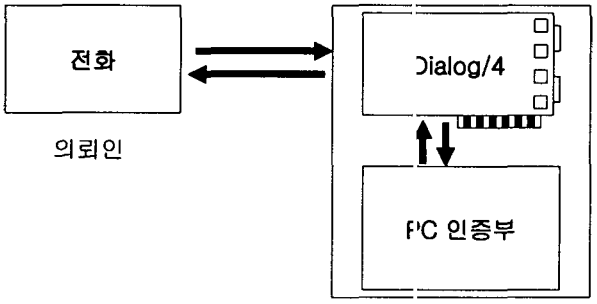
2.1 Dialog/4 보드를 이용한 화자인증 시스템의 개요

논문에서 구현한 전화 음성 기반의 화자인증 시스템은 Dialog/4보드를 이용해 의뢰인의 ID와 음성을 입력 받은 후 인증관련 처리를 PC상에서 수행하여 그 인증 결과를 다시 알려주는 시스템이다. <그림 1>는 시스템의 개략적인 구성도이다. <그림 2>는 프로그램의 실행화면을 보여준다. 프로그램이 실행되면 인증 시스템의 상태와 정보들이 표시창에 나타난다. <그림 3>는 화자인증 시스템의 처리 과정의 블록 다이어그램이다. 프로그램이 시작되면 dialog/4 보드를 초기화하고 콜 관련 메시지들을 처리한다. 프로그램은 보드를 제어하는 부분과 인증부가 있으며 인증부에서는

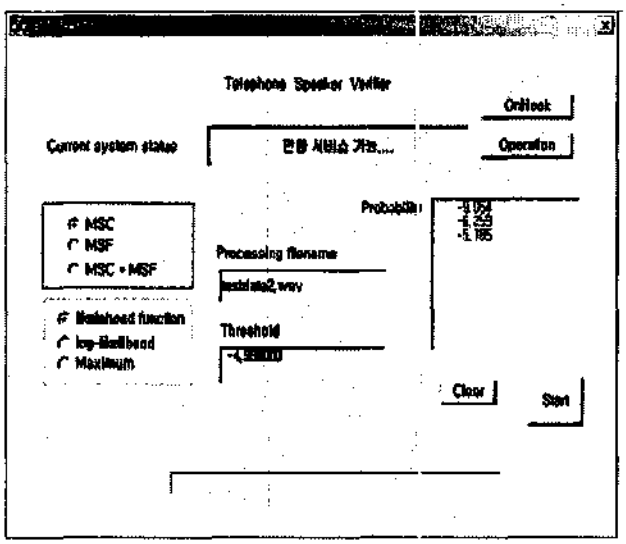
전화를 통해 입력된 ID와 음성 데이터를 저장한 후 특징벡터를 추출해 ID에 해당하는 화자모델과 배경화자 정보를 이용하여 유사도(likelihood ratio)를 계산하여 실시간으로 인증 또는 거절을 알려준다.

dialog/4 보드의 통제부는 전화벨을 감지하고, 상황에 알맞은 안내문을 읽어주고 인증 관련 처리가 끝나면 결과를 전송한 후 자동으로 전화를 끊는 처리 기능을 수행하며 전화를 통해 의뢰인의 ID와 음성을 입력을 받아 인증부로 넘겨준다. 의뢰인의 음성은 발성과 동시에 실시간으로 8kHz로 샘플링 되어 μ-law PCM 웨이브 파일

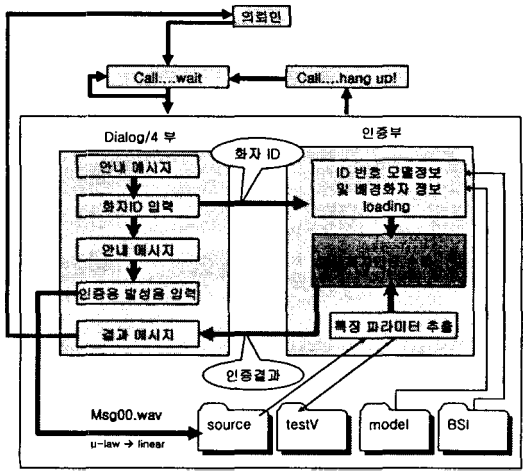
형태로 메모리에 저장한다. 인증부는 소스 폴더에 저장된 의뢰인의 음성파일을 읽어와 μ-law PCM으로 저장된 음성을 선형 PCM으로 바꾼 다음, 끝점 검출을 이용하여 음성구간을 찾은 후, 필터 $H(z) = 1-0.95z^{-1}$를 이용하여 프리엠퍼시스한다. 160샘플(20ms, 8kHz)을 한 프레임으로 하고 80샘플(10ms, 8kHz)마다 중첩하여 각 프레임에 해밍(hamming)창으로 가중치를 주어 매 10ms 마다 특징 파라미터인 12차 MFCC를 추출하여 이를 test V 폴더에 저장한다. 동시에 dialog/4에서 입력받은 ID에 해당하는등록화자의 모델이 메모리에 적재되며 화자 모델의 배경화자 정보를 참조하여 처리할 배경화자모델을 결정하는 작업이 수행된다. 두 과정이 완료되면 내부의 인증루틴에서 본인의 모델과 배경화자 모델들이 입력음성을 발생시킬 확률을 계산하여 수락할 것인지 거절할 것인지를 판단하여 의뢰인에게 알려준다.



<그림 1> 화자 인증 시스템의 구성도



<그림 2> 프로그램 실행 화면



<그림 3> 화자 인증 시스템의 처리과정

2.2 Dialog/4 보드의 특성

Dialog/4 보드는 Intel® Dialogic®의 제품으로 PC와 인터페이스 된다. 아날로그 라인을 직접적으로 연결할 수 있는 2개의 RJ-11포트를 가지고 있으며 포트마다 2개의 채널을 사용할 수 있도록 되어 있다. 하나의 PC에 16개의 보드까지 확장 가능하며 MS-DOS, Windows 95, Windows NT/2000, OS/2 및 UNIX 에서 운영될 수 있는 C언어 응용 프로그램 인터페이스(API)를 제공한다. 음성데이터을 저장할 때는 각 채널당 24kb/s에서 64kb/s 까지의 데이터 속도를 선택할 수 있으며, 음성 포맷으로 6kHz와 8kHz의 샘플링율로 ADPCM과 μ-law PCM 포맷을 지원한다. 그리고 dialog/4 보드는 신호처리를 수행하는 Motorola 56001(33MHz) DSP 프로세서와 의사결정과 데이터 전송 기능을 수행하는 제어용 마이크로프로세서인 80C188이 결합된 이중 프로세서 구조를 사용한다.

2.3 GMM(Gaussian Mixture Model) 화자 모델

GMM은 통계적 방법으로 화자의 음성을 모델링 하는 방법으로 특정 화자가 인증을 위해 발성한 음성신호로부터 얻어낸 특징 벡터들이 지정된 화자의 GMM모델에 의해 발생될 확률을 계산하여 이를 문턱 값과 비교하여 화자인증을 하는 방식이다.[3][4]

화자 s 에 대한 GMM 화자모델 λ_s는 다음 식(1)으로 표현된다.

$$\lambda_s = \{w_i^s, \mu_i^s, \Sigma_i^s\}, i = 1, ..., M \quad (1)$$

여기서, M은 가우시안 확률분포의 개수, 즉 믹스처의 크기를 의미하고, w_i은 i번째 가우시안 믹스처의 가중치가 된다. 또, μ_i^s과 Σ_i^s은 각각 가우시안의 평균과 분산이다.

2.4 배경화자(background speaker) 선정방법

배경화자의 필요성은 본인과 비슷하지만 본인이 아닌 사람들의 정보를 반영함으로써 화자인증 시스템의 성능을 향상시키기 위함이다. 배경화자 선정 방법에는 화자와 가장 유사도가 높은 화자모델들을 배경화자 모델로 선정하는 MSC(Maximally Spread Close) 방법과 유사도가 작은 화자모델들을 배경화자모델에 포함하는 MSF(Maximally Spread Far) 방법이 있다.[2]
본 논문에서는 MSC(Maximally Spread Close) 방법을 사용하였다.

2.5 화자인증 시스템의 판정기준

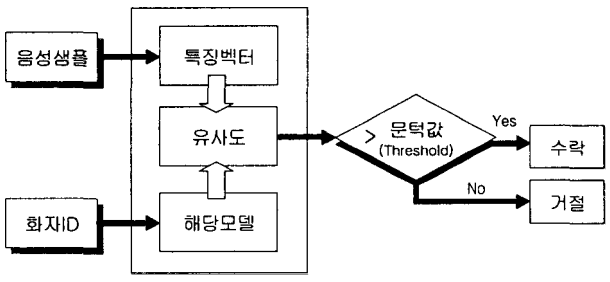
<그림 4>은 GMM을 이용한 화자인증 시스템의 구성도이다. 발성 음과 의뢰인 모델 사이의 유사 도는 식(2)과 같이 나타낼 수 있다.[2]

$$L(X) = \frac{P(\lambda_c | X)}{P(\lambda_{\bar{c}} | X)} \qquad (2)$$

식(2)을 Bayes 정리를 이용해 로그를 취하면 식(3)이 된다.

$$\Lambda(X) = logP(X|\lambda_c) - logP(X|\lambda_{\bar{c}}) \qquad (3)$$

$P(X|\lambda_c)$는 의뢰인 모델에서 발성 음이 생성될 확률이고, $P(X|\lambda_{\bar{c}})$는 배경화자모델에서 발성음이 생성될 확률이다. 논문에서는 판정기준인 문턱 값으로 T_{EER} 값을 사용한다. 식(3)을 계산하여 유사도 $\Lambda(X)$을 구한 후 그 값을 시스템의 문턱 값 T_{EER}과 비교하여 $\Lambda(X)>T_{EER}$ 이면 수락(acceptance) 메시지를, $\Lambda(X)<T_{EER}$이면 거절(rejection) 메시지를 의뢰인에게 알려준다.



<그림 4> 화자 인증 시스템의 구성

3. 화자인증시스템의 음성 DB 구성

음성DB는 20명의 화자에 대하여 PC기반에서 고급 마이크를 사용하여 녹음된 깨끗한 음성과 직접 시스템에 전화를 걸어 입력받은 전화음성 2가지로 구성하였다. 음성DB의 수집환경을 <표 1>에 정리하였다.

<표 1> 음성DB 수집환경

발화내용	발성기간	발성횟수	녹음환경	녹음방법
고정문장 (단 문장)	15초	10	연구실	고성능마이크
				전화
				휴대전화
비고정문장 (장문장)	150초	1	연구실	고성능마이크
				전화
				휴대전화

사용한 고성능 마이크는 SHURE사의 VR250BT이며 발화내용은 15초 정도의 길이의 단문장과 150초정도의 장문장으로 선정하였다. 단문장의 경우 같은 내용을 10번 발음함으로써 10개의 샘플을 취하였고, 장문장의 경우 임의의 문장을 150초간 1회 발성하여 1개의 샘플을 취하였다. 발화 내용은 <표2>에 정리하였다

<표 2> 음성 DB 발화내용

단 문장
안녕하십니까? 여기는 대전대학교 전자공학과 멀티미디어 통신연구실 입니다. 화자 인증기의 성능평가를 위한 검증용 음성을 녹음하는 중입니다. 이 문장은 약 15초간 녹음됩니다.

장 문장
코르시카 섬,아작 시오의 골목길, 바다내음, 소나무와 유향나무, 소귀나무와 도금양나무의향기...이 모든 유년기의 세계가 이제 내밀한 추억으로 멀어질 것이다 아버지가 형과 그를 오행의 학교에 남겨두고 떠날 때 ,그는 이를 악물려 야했다. 그보다 한살위인 형 조제프는 교회로 ,나폴레오네는 군대로 향할 운명에 접어 들었다. 오랭에서, 1월1일부터 4월21일 까지 석 달 남짓 그는 불어를 배웠다. 고향마을 아작 시오의 거리를 점령한 프랑스 군인들이 떠들어 대던 언어. 아버지는 불어를 할줄 알았지만 어머니는 한 마디도하지 못했다..... [나폴레옹에서 각150초간씩 발취]

4. 성능 평가 및 결과

화자인증 시스템의 성능은 20명의 화자를 대상으로 수집한 음성DB를 이용하였으며 기본 인증 성능과 전화 기반에서의 성능으로 구분하여 평가하였다. 음성의 특징벡터로 12차 MFCC를 사용하였으며, 화자 모델을 위해 사용된 GMM의 차수는 32차이다. 배경화자의 수는 6명으로 하였으며, 배경화자 구성 방법으로는 MSC만 사용 하였다.

4.1. 인증부 자체 성능

인증부 실험은 화자인증 시스템의 기본 성능을 확인하기 위한 실험으로 PC에서 고성능 마이크를 사용하여 깨끗한 음성으로 화자모델을 구한 후 같은 환경에서 녹음된 테스트 음성으로 화자인증을 하는 경우의 성능평가이다. <표 3>은 발성한 각각의 고정문장(단문장)의

샘플을 모두 더한 약 150초 길이의 음성을 이용하여, 테스트 음성으로는 15초 길이의 단 문장 샘플 10개와 150초 길이를 15초씩 분할한 장문장의 샘플 10개를 사용하여 실험한 결과이다.

<표 3> 인증부실험결과

실험내용	훈련	실험	결과	
			인증률(%)	ERR(%)
문장 종속	마이크단문장	마이크단문장	98.2	1.8
문장 독립	마이크장문장	마이크장문장(15초로 분할)	95.9	4.1

인증부 자체 성능 평가에서 인증률은 각각 98.2%와 95.9%를 보여주었다. 화자 등록(훈련)에 사용한 음성발성 내용과 테스트에 사용한 음성발성 내용이 같은 경우인 문장 종속 화자인증 일 때가 문장 독립인 경우보다 인증률이 향상되었음을 보였다.

4.2. 전화 기반에서의 성능

이 실험은 전화 기반에서의 성능을 알아보기 위한 실험으로 실제 전화음성으로 화자인증을 하는 경우의 성능평가이다.

<표 4>은 발성한 각각의 고정문장(단독-장)의 샘플을 모두 더한 약 150초 길이의 음성을 이용하여 화자 모델링하고, 테스트 음성으로는 15초 길이의 단 문장 샘플 10개와 150초의길이를 15초씩 분할한 장문장의 샘플 10개를 사용하여 실험한 결과이다

<표 4> 전화 기반에서의 성능결과

실험내용	훈련	실험	결과	
			인 증률(%)	ERR(%)
문장 종속	전화단문장	전화단문장	99.6	0.4
문장 독립	전화장문장	전화장문장(15초로분할)	97.8	2.2
녹음환경	마이크단문장	전화단문장	76.8	23.2
	마이크장문장	전화장문장(15초로분할)	63.3	36.7

실제의 전화음성을 이용한 화자인증 시스템의 성능평가에서는, 마이크로 입력받은 음성을 사용하여 화자를 모델링한 경우 76.8%, 전화음성으로 화자를 모델링 하였을 경우는 99.6%의 인증률을 보여주었다. 이는 화자 훈련 및 실험에서 서로 다른 환경에서 녹음된 음성 데이터를 사용한 결과로 생각된다.

5. 결론 및 고찰

본 논문은 Dialog/4 보드를 이용한 전화음성 기반의 화자 인증시스템 구현과 그 성능 평가에 대하여 서술하였다. 성능평가에서 문장종속인 경우 문장 독립인 경우보다 인식성능이 좋아짐을 알 수 있으며 실제의 전화음성을 이용한 성능 평가에서는 마이크로 입력을 받은 음성을 사용하여 화자를 모델링한 경우가 전화음성으로 화자를 모델링한 경우보다 인증률이 낮았다. 이러한 화자모델의 인증률 차이는 화자모델에 사용한 음성과 실험에 사용한 음성데이터의 녹음환경의 차이에 따른 것으로 일반적으로 훈련과 실험시동일한 환경일수록 인증성능이 좋아진다는 사실을 보여주고있다.차후 켑스트럼 평균차감법(CMS:Cepstral Mean Subtraction)과 같은 채널보상을 위한 방법들이 추가적으로 구현된다면 더 나은 성능을 보여 질 수 있을 것이다.

< 참고문헌 >

[1] L. R. Rabiner & R. W. Schafer (1978), Digital processing of speech signals, Prentice Hall.

[2] D. A. Reynolds (1995), "Speaker identification and verification using gaussian mixture speaker models", Speech Communication, Vol. 17 pp.91~108.

[3] D. A. Reynolds, R. C. Rose (1995), "Robust text-independent speaker identification using gaussian mixture speaker models", IEEE. Trans. On Speech and Audio Processing, Vol. 3, No. 1, pp.72-83.

[4] G. H, S. M (1994), "Text-independent speaker identification", IEEE Signal Processing Magazine, pp.18-32.