(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년10월29일
(11) 등록번호 10-1323061
(24) 등록일자 2013년10월22일

(51) 국제특허분류(Int. Cl.)
G10L 15/14 (2006.01)
(21) 출원번호 10-2008-7020272
(22) 출원일자(국제) 2007년02월13일
심사청구일자 2012년01월18일
(85) 번역문제출일자 2008년08월19일
(65) 공개번호 10-2008-0102373
(43) 공개일자 2008년11월25일
(86) 국제출원번호 PCT/US2007/004137
(87) 국제공개번호 WO 2007/098039
국제공개일자 2007년08월30일
(30) 우선권주장
11/358,302 2006년02월20일 미국(US)
(56) 선행기술조사문헌
Douglas A. Reynolds et al., 'Speaker verification using adapted gaussian mixture models', Digital Signal Processing, Vol.10, Nos.1-3, pp.19~41, July 2000*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
마이크로소프트 코포레이션
미국 워싱턴주 (우편번호 : 98052) 레드몬드 원 마이크로소프트 웨이
(72) 발명자
장, 정유
미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이
리우, 밍
미국 98052-6399 워싱톤주 레드몬드 원 마이크로소프트 웨이
(74) 대리인
제일특허법인

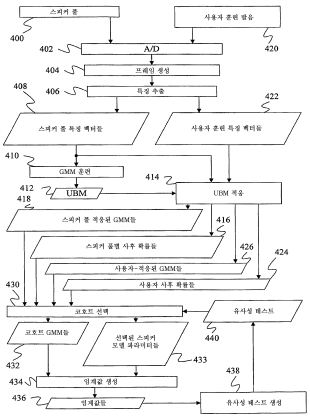
전체 청구항 수 : 총 10 항 심사관 : 정성윤

(54) 발명의 명칭 **스피커 인증 방법 및 이 방법을 수행하기 위한 컴퓨터 실행가능 명령어를 갖는 컴퓨터 판독가능 매체**

(57) 요 약

테스트 발음 및 저장된 훈련 발음에 대한 유사성 점수를 결정함으로써 스피커 인증이 수행된다. 유사성 점수를 산출하는 것은, 혼합 성분의 사후 확률과 적응된 평균과 백그라운드 평균 간의 차의 적을 포함하는 각 함수들 그룹의 합을 결정하는 단계를 포함한다. 적응된 평균은 백그라운드 평균과 테스트 발음에 기초하여 형성된다. 인증을 위해 사용자에 의해 제공되는 스피치 컨텐츠는 텍스트에 무관한 것(즉, 사용자들이 말하고자 하는 모든 컨텐츠)이거나 또는 텍스트에 좌우되는 것(즉, 훈련을 위해 사용되는 특정 구)일 수 있다.

대 표 도 - 도4



특허청구의 범위

청구항 1

스피커 인증을(speaker authentication) 위한 방법으로서,

스피치 신호(speech signal)를 수신하는 단계;

상기 수신된 스피치 신호에 기초하여 복수의 혼합 성분들(mixture components) 각각에 대한 백그라운드 평균들(background means)을 포함하는 백그라운드 모델(background model)을 적응시킴(adapt)으로써 상기 복수의 혼합 성분들 각각에 대해 적응된 평균들(adapted means)을 형성하는 단계;

사용자로부터의 훈련 스피치 신호에 기초하여 상기 백그라운드 모델을 적응시킴으로써 상기 복수의 혼합 성분들 각각에 대해 훈련 평균들(training means)을 형성하는 단계; 및

상기 복수의 혼합 성분에 대해 결정된 함수들의 합을 결정함으로써 유사성 점수(similarity score)를 결정하는 단계 - 각각의 함수는 상기 수신된 스피치 신호에 기초한 혼합 성분의 사후 확률(posterior probability)과, 적응된 평균과 백그라운드 평균 간의 차의 적(product)을 포함하고, 또한 상기 훈련 스피치 신호에 기초하는 혼합 성분의 사후 확률과, 훈련 평균과 백그라운드 평균 간의 차의 적을 더 포함함 - ;

를 포함하는

스피커 인증 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

명목상의 사용자 ID를 수신하고, 상기 명목상의 사용자 ID에 기초하여 상기 함수들에서 사용하기 위해 훈련 평균들을 선택하는 단계를 더 포함하는

스피커 인증 방법.

청구항 5

제1항에 있어서,

스피커 풀(speaker pool)의 복수의 스피커들 각각에 대해 스피커 풀 평균들을 형성하는 단계 - 스피커에 대한 상기 스피커 풀 평균들은 상기 스피커로부터의 스피치에 기초하여 상기 백그라운드 모델을 적응시킴으로써 형성됨 - 를 더 포함하는

스피커 인증 방법.

청구항 6

제5항에 있어서,

각각의 함수는 대응하는 임계값(corresponding threshold)을 더 포함하고,

각각의 임계값은 상기 스피커 풀의 스피커들의 서브집합에 대한 스피커 풀 평균들에 기초하는

스피커 인증 방법.

청구항 7

제6항에 있어서,

상기 스피커 풀 평균들과 상기 적응된 평균들로부터 결정된 유사성 점수에 기초하여 상기 스피커 풀 중에서 스피커들의 서브집합을 선택하는 단계를 더 포함하는

스피커 인증 방법.

청구항 8

제7항에 있어서,

사용자로부터의 훈련 스피치 신호에 기초하여 상기 백그라운드 모델을 적응시킴으로써 상기 복수의 혼합 성분들 각각에 대해 훈련 평균들을 형성하는 단계; 및

상기 스피커 풀의 스피커들의 제2 서브집합에 대한 스피커 풀 평균들에 기초하여 명목상의 사용자 임계값들을 결정하는 단계 - 상기 제2 서브집합은 상기 스피커 풀 평균들과 상기 훈련 평균들로부터 결정된 유사성 점수에 기초하여 상기 스피커 풀 중에서 선택됨 - 를 더 포함하는

스피커 인증 방법.

청구항 9

제8항에 있어서,

각각의 함수는 제2 임계값을 더 포함하는

스피커 인증 방법.

청구항 10

프로세서에 의해 실행될 때 상기 프로세서로 하여금 스피커 인증 방법을 수행하도록 하는 컴퓨터 실행가능 명령어들이 저장된 컴퓨터 판독가능 매체로서,

상기 스피커 인증 방법은,

제1 적응된 평균을 형성하기 위해 테스트 발음(test utterance)에 기초하여 백그라운드 평균을 포함하는 백그라운드 모델을 적응시키는 단계;

제2 적응된 평균을 형성하기 위해 저장된 사용자 발음에 기초하여 상기 백그라운드 모델을 적응시키는 단계;

유사성 점수들의 제1 집합을 형성하기 위해 상기 제1 적응된 평균에 기초하여 상기 테스트 발음과 훈련 발음들의 집합의 각각의 발음 간의 유사성 점수를 결정하는 단계;

상기 유사성 점수들의 제1 집합을 이용하여 상기 테스트 발음에 대한 코호트(cohorts)로서 훈련 발음들의 집합의 서브집합을 선택하는 단계;

유사성 점수들의 제2 집합을 형성하기 위해 상기 제2 적응된 평균에 기초하여 상기 저장된 사용자 발음과 훈련 발음들의 집합의 각각의 발음 간의 유사성 점수를 결정하는 단계;

상기 유사성 점수들의 제2 집합을 이용하여 상기 저장된 사용자 발음에 대한 코호트로서 상기 훈련 발음들의 집합의 서브 집합을 선택하는 단계;

상기 테스트 발음에 대한 코호트의 평균들을 이용하여 제1 임계값을 계산하는 단계;

상기 저장된 사용자 발음에 대한 코호트의 평균들을 이용하여 제2 임계값을 계산하는 단계;

상기 테스트 발음과 상기 저장된 사용자 발음 간의 인증 유사성 점수(authentication similarity score)의 계산에서 상기 제1 임계값, 상기 제2 임계값, 상기 제1 적응된 평균과 상기 백그라운드 평균 간의 차, 및 상기 제2 적응된 평균과 상기 백그라운드 평균 간의 차를 이용하는 단계; 및

상기 인증 유사성 점수를 이용하여 동일한 사용자가 상기 테스트 발음 및 상기 저장된 사용자 발음을 생성하였는지 여부를 판정하는 단계를 포함하는

컴퓨터 판독가능 매체.

청구항 11

제10항에 있어서,

상기 테스트 발음과 훈련 발음 간의 유사성 점수를 결정하는 단계는 상기 제1 적응된 평균과 상기 백그라운드 모델의 백그라운드 평균 간의 차를 결정하는 단계 및 상기 차를 이용하여 상기 유사성 점수를 결정하는 단계를 포함하는

컴퓨터 판독가능 매체.

청구항 12

제11항에 있어서,

상기 테스트 발음과 상기 훈련 발음 간의 유사성 점수를 결정하는 단계는 상기 테스트 발음에 기초하여 혼합 성분에 대한 확률을 결정하는 단계 및 상기 혼합 성분에 대한 확률과, 상기 제1 적응된 평균과 상기 백그라운드 평균 간의 차 및 상기 제2 적응된 평균과 상기 백그라운드 평균 간의 차의 적을 사용하여 상기 유사성 점수를 결정하는 단계를 더 포함하는

컴퓨터 판독가능 매체.

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

명 세 서

배 경 기 술

[0001] 스피커 인증(speaker authentication)은 스피치 신호(speech signal)에 기초하여, 승인을 구하는 스피커의 아이덴티티(the claimed identity of speaker)를 확인(verify)하는 프로세스이다. 인증은, 통상적으로, 그 시스템을 사용하는 각 사람에 대해 훈련되어온 스피치 모델을 이용하여 수행된다.

[0002] 일반적으로, 텍스트에 무관한 스피커 인증(text-independent speaker authentication)과 텍스트에 좌우되는 스피커 인증(text-dependent speaker authentication)의 두 가지 유형의 스피커 인증이 있다. 텍스트에 무관한 스피커 인증에서, 스피커는 자신이 제공하고자 하는 임의의 스피치 컨텐츠를 제공한다. 텍스트에 좌우되는 스

피커 인증에서는, 스피커는 모델 훈련 동안 그리고 인증 시스템의 사용 동안, 특정 구(phrase)를 읊는다(recite). 동일한 구를 반복함으로써, 음성 단위들(phonetic units) 및 이 음성 단위들 간의 천이의 강력한 모델이 텍스트에 좌우되는 스피커 인증 시스템에 대해 구축될(construct) 수 있다. 이것은 텍스트에 무관한 스피커 인증 시스템에서는 맞지(true) 않는데, 그 이유는 많은 음성 단위들과 이 음성 단위들 간의 많은 천이가 훈련 동안 관찰되지 않아, 모델로 잘 나타내어지지 않을 것이기 때문이다.

[0003] 상술된 내용은 단지 일반적인 배경 정보를 제공하기 위해서이며, 청구되는 내용의 범위를 결정하는 것을 돕는 것으로 사용되고자 의도되지 않는다.

발명의 상세한 설명

[0004] 스피커 인증은 테스트 발음(test utterance)과 저장된 훈련 발음(stored training utterance)에 대한 유사성 점수(similarity score)를 결정함으로써 수행된다. 유사성 점수를 산출하는 것은, 혼합 성분(mixture component)의 사후 확률(posterior probability)과 적응된 평균(adapted mean)과 백그라운드 평균(background mean) 간의 차의 적(product)을 포함하는 각각의 함수들의 그룹의 합(sum)을 결정하는 단계를 포함한다. 적응된 평균은 백그라운드 평균과 테스트 발음에 기초하여 형성된다.

[0005] 본 요약은 이하의 상세한 설명에서 더 설명될 일련의 개념을 단순화된 형태로 소개하기 위해 제공된다. 본 요약은 청구되는 내용의 주요 특징들 또는 핵심 특징들을 식별하고자 하는 것이 아니며, 또한 청구되는 내용의 범위를 결정하는 데 도움이 되는 것으로서 사용되고자 하는 것도 아니다. 청구되는 내용은 배경기술에서 언급된 임의의 단점 또는 모든 단점들을 해결하는 구현에 제한되지 않는다.

실 시 예

[0020] 도 1은 실시예들이 구현되기에 적합한 컴퓨팅 시스템 환경(100)의 일례를 도시하고 있다. 컴퓨팅 시스템 환경(100)은 적합한 컴퓨팅 환경의 일례에 불과하며, 본 발명의 용도 또는 기능성의 범위에 관해 어떤 제한을 암시하고자 하는 것이 아니다. 컴퓨팅 환경(100)이 예시적인 운영 환경(100)에 도시된 컴포넌트들 중 임의의 하나 또는 그 컴포넌트들의 임의의 조합과 관련하여 어떤 의존성 또는 요구사항을 갖는 것으로 해석되어서는 안 된다.

[0021] 실시예들은 많은 기타 범용 또는 특수 목적의 컴퓨팅 시스템 환경 또는 구성으로 동작할 수 있다. 각종 실시예들에서 사용하는 데 적합할 수 있는 잘 알려진 컴퓨팅 시스템, 환경 및/또는 구성의 예로는 퍼스널 컴퓨터, 서버 컴퓨터, 핸드-헬드 또는 랩톱 장치, 멀티프로세서 시스템, 마이크로프로세서 기반 시스템, 셋톱 박스, 프로그램가능한 가전제품, 네트워크 PC, 미니컴퓨터, 메인프레임 컴퓨터, 전화 시스템, 상기 시스템들이나 장치들 중 임의의 것을 포함하는 분산 컴퓨팅 환경, 기타 등등이 있지만 이에 제한되는 것은 아니다.

[0022] 실시예들은 일반적으로 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터 실행가능 명령어와 관련하여 기술될 것이다. 일반적으로, 프로그램 모듈은 특정 태스크를 수행하거나 특정 추상 데이터 유형을 구현하는 루틴, 프로그램, 개체, 컴포넌트, 데이터 구조 등을 포함한다. 본 발명은 또한 통신 네트워크를 통해 연결되어 있는 원격 처리 장치들에 의해 태스크가 수행되는 분산 컴퓨팅 환경에서 실시되도록 설계된다. 분산 컴퓨팅 환경에서, 프로그램 모듈은 메모리 저장 장치를 비롯한 로컬 및 원격 컴퓨터 저장 매체 둘 다에 위치할 수 있다.

[0023] 도 1과 관련하여, 일부 실시예들을 구현하는 예시적인 시스템은 컴퓨터(110) 형태의 범용 컴퓨팅 장치를 포함한다. 컴퓨터(110)의 컴포넌트들은 처리 장치(120), 시스템 메모리(130), 및 시스템 메모리를 비롯한 각종 시스템 컴포넌트들을 처리 장치(120)에 연결시키는 시스템 버스(121)를 포함하지만 이에 제한되는 것은 아니다. 시스템 버스(121)는 메모리 버스 또는 메모리 컨트롤러, 주변 장치 버스 및 각종 버스 아키텍처 중 임의의 것을 이용하는 로컬 버스를 비롯한 몇몇 유형의 버스 구조 중 어느 것이라도 될 수 있다. 예로서, 이러한 아키텍처는 ISA(industry standard architecture) 버스, MCA(micro channel architecture) 버스, EISA(Enhanced ISA) 버스, VESA(video electronics standard association) 로컬 버스, 그리고 메자닌 버스(mezzanine bus)로도 알려진 PCI(peripheral component interconnect) 버스 등을 포함하지만 이에 제한되는 것은 아니다.

[0024] 컴퓨터(110)는 통상적으로 각종 컴퓨터 판독가능 매체를 포함한다. 컴퓨터(110)에 의해 액세스 가능한 매체는 그 어떤 것이든지 컴퓨터 판독가능 매체가 될 수 있고, 이러한 컴퓨터 판독가능 매체는 휘발성 및 비휘발성 매체, 이동식 및 비이동식 매체를 포함한다. 예로서, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체 및 통신 매체를 포함하지만 이에 제한되는 것은 아니다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보를 저장하는 임의의 방법 또는 기술로 구현되는 휘발성 및 비휘발성, 이동식

및 비이동식 매체를 포함한다. 컴퓨터 저장 매체는 RAM, ROM, EEPROM, 플래시 메모리 또는 기타 메모리 기술, CD-ROM, DVD(digital versatile disk) 또는 기타 광 디스크 저장 장치, 자기 카세트, 자기 테이프, 자기 디스크 저장 장치 또는 기타 자기 저장 장치, 또는 컴퓨터(110)에 의해 액세스되고 원하는 정보를 저장할 수 있는 임의의 기타 매체를 포함하지만 이에 제한되는 것은 아니다. 통신 매체는 통상적으로 반송파(carrier wave) 또는 기타 전송 메커니즘(transport mechanism)과 같은 피변조 데이터 신호(modulated data signal)에 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터 등을 구현하고 모든 정보 전달 매체를 포함한다. "피변조 데이터 신호"라는 용어는, 신호 내에 정보를 인코딩하도록 그 신호의 특성들 중 하나 이상을 설정 또는 변경시킨 신호를 의미한다. 예로서, 통신 매체는 유선 네트워크 또는 직접 배선 접속(direct-wired connection)과 같은 유선 매체, 그리고 음향, RF, 적외선, 기타 무선 매체와 같은 무선 매체를 포함한다. 상술된 매체들의 모든 조합이 또한 컴퓨터 판독가능 매체의 영역 안에 포함되는 것으로 한다.

[0025] 시스템 메모리(130)는 판독 전용 메모리(ROM)(131) 및 랜덤 액세스 메모리(RAM)(132)와 같은 휘발성 및/또는 비휘발성 메모리 형태의 컴퓨터 저장 매체를 포함한다. 시동 중과 같은 때에, 컴퓨터(110) 내의 구성요소들 사이의 정보 전송을 돕는 기본 루틴을 포함하는 기본 입/출력 시스템(BIOS)(133)은 통상적으로 ROM(131)에 저장되어 있다. RAM(132)은 통상적으로 처리 장치(120)가 즉시 액세스 할 수 있고 및/또는 현재 동작시키고 있는 데이터 및/또는 프로그램 모듈을 포함한다. 예로서, 도 1은 운영 체제(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136) 및 프로그램 데이터(137)를 도시하고 있지만 이에 제한되는 것은 아니다.

[0026] 컴퓨터(110)는 또한 기타 이동식/비이동식, 휘발성/비휘발성 컴퓨터 저장매체를 포함한다. 단지 예로서, 도 1은 비이동식·비휘발성 자기 매체에 기록을 하거나 그로부터 판독을 하는 하드 디스크 드라이브(141), 이동식·비휘발성 자기 디스크(152)에 기록을 하거나 그로부터 판독을 하는 자기 디스크 드라이브(151), CD-ROM 또는 기타 광 매체 등의 이동식·비휘발성 광 디스크(156)에 기록을 하거나 그로부터 판독을 하는 광 디스크 드라이브(155)를 포함한다. 예시적인 운영 환경에서 사용될 수 있는 기타 이동식/비이동식, 휘발성/비휘발성 컴퓨터 기억 매체로는 자기 테이프 카세트, 플래시 메모리 카드, DVD, 디지털 비디오 테이프, 고상(solid state) RAM, 고상 ROM 등이 있지만 이에 제한되는 것은 아니다. 하드 디스크 드라이브(141)는 통상적으로 인터페이스(140)와 같은 비이동식 메모리 인터페이스를 통해 시스템 버스(121)에 접속되고, 자기 디스크 드라이브(151) 및 광 디스크 드라이브(155)는 통상적으로 인터페이스(150)와 같은 이동식 메모리 인터페이스에 의해 시스템 버스(121)에 접속된다.

[0027] 위에서 설명되고 도 1에 도시된 드라이브들 및 이들과 관련된 컴퓨터 저장 매체는, 컴퓨터(110)를 위해, 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 및 기타 데이터를 저장한다. 도 1에서, 예를 들어, 하드 디스크 드라이브(141)는 운영 체제(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146), 및 프로그램 데이터(147)를 저장하는 것으로 도시되어 있다. 여기서 주의할 점은 이들 컴포넌트가 운영 체제(134), 애플리케이션 프로그램(135), 기타 프로그램 모듈(136), 및 프로그램 데이터(137)와 동일하거나 그와 다를 수 있다는 것이다. 이에 관해, 운영 체제(144), 애플리케이션 프로그램(145), 기타 프로그램 모듈(146) 및 프로그램 데이터(147)에 다른 번호가 부여되어 있다는 것은 적어도 이들이 다른 사본(copy)이라는 것을 나타내기 위한 것이다.

[0028] 사용자는 키보드(162), 마이크(163) 및 마우스, 트랙볼(trackball) 또는 터치 패드와 같은 포인팅 장치(161) 등의 입력 장치를 통해 명령 및 정보를 컴퓨터(110)에 입력할 수 있다. 다른 입력 장치(도시 생략)로는 마이크, 조이스틱, 게임 패드, 위성 안테나, 스캐너 등을 포함할 수 있다. 이들 및 기타 입력 장치는 종종 시스템 버스에 결합된 사용자 입력 인터페이스(160)를 통해 처리 장치(120)에 접속되지만, 병렬 포트, 게임 포트 또는 USB(universal serial bus) 등의 다른 인터페이스 및 버스 구조에 의해 접속될 수도 있다. 모니터(191) 또는 다른 유형의 디스플레이 장치도 비디오 인터페이스(190) 등의 인터페이스를 통해 시스템 버스(121)에 접속될 수 있다. 모니터 외에, 컴퓨터는 스피커(197) 및 프린터(196) 등의 기타 주변 출력 장치를 포함할 수 있고, 이들은 출력 주변장치 인터페이스(195)를 통해 접속될 수 있다.

[0029] 컴퓨터(110)는 원격 컴퓨터(180)와 같은 하나 이상의 원격 컴퓨터로의 논리적 접속을 사용하여 네트워크화된 환경에서 동작할 수 있다. 원격 컴퓨터(180)는 또 하나의 퍼스널 컴퓨터, 핸드-헬드 장치, 서버, 라우터, 네트워크 PC, 피어 장치 또는 기타 통상의 네트워크 노드일 수 있고, 통상적으로 컴퓨터(110)와 관련하여 상술된 구성요소들의 대부분 또는 그 전부를 포함한다. 도 1에 도시된 논리적 접속으로는 LAN(171) 및 WAN(173)이 있지만, 기타 네트워크를 포함할 수도 있다. 이러한 네트워킹 환경은 사무실, 전사적 컴퓨터 네트워크(enterprise-wide computer network), 인트라넷, 및 인터넷에서 일반적인 것이다.

[0030] LAN 네트워킹 환경에서 사용될 때, 컴퓨터(110)는 네트워크 인터페이스 또는 어댑터(170)를 통해 LAN(171)에 접

속된다. WAN 네트워킹 환경에서 사용될 때, 컴퓨터(110)는 통상적으로 인터넷과 같은 WAN(173)을 통해 통신을 설정하기 위한 모뎀(172) 또는 기타 수단을 포함한다. 내장형 또는 외장형일 수 있는 모뎀(172)은 사용자 입력 인터페이스(160) 또는 기타 적절한 메커니즘을 통해 시스템 버스(121)에 접속된다. 네트워크화된 환경에서, 컴퓨터(110) 또는 그의 일부와 관련하여 기술된 프로그램 모듈은 원격 메모리 저장 장치에 저장될 수 있다. 예로서, 도 1은 원격 애플리케이션 프로그램(185)이 원격 컴퓨터(180)에 있는 것으로 도시하고 있지만 이에 제한되는 것은 아니다. 도시된 네트워크 접속은 예시적인 것이며 이 컴퓨터들 사이에 통신 링크를 설정하는 기타 수단이 사용될 수 있다는 것을 이해할 것이다.

[0031] 도 2는 예시적인 컴퓨팅 환경인 모바일 장치(200)의 블록도이다. 모바일 장치(200)는, 마이크로프로세서(202), 메모리(204), 입력/출력(I/O) 컴포넌트(206), 및 원격 컴퓨터들 또는 기타 모바일 장치들과의 통신을 위한 통신 인터페이스(208)를 포함한다. 한 실시예에서, 상술된 컴포넌트들은 적합한 버스(210)를 통해 서로 간의 통신을 위해 결합된다.

[0032] 메모리(204)는, 모바일 장치(200)로의 일반 전력이 셧다운 되었을 때 메모리(204)에 저장되어 있는 정보가 없어지지(lost) 않도록 배터리 백업 모듈(battery back-up module)을 지니는 RAM과 같은 비휘발성 전자 메모리로서 구현된다. 메모리(204)의 일부는 프로그램 실행을 위해 어드레스가능한 메모리로서 할당되는 것이 바람직한 반면, 메모리(204)의 또 다른 일부는 디스크 드라이브 상에 저장을 시뮬레이트(simulate)하는 것과 같이, 저장용으로 사용되는 것이 바람직하다.

[0033] 메모리(204)는 운영 체제(212), 애플리케이션 프로그램(214) 뿐만 아니라 객체 저장소(216)를 포함한다. 동작시, 운영 체제(212)는 메모리(204)로부터 프로세서(202)에 의해 실행되는 것이 바람직하다. 한 바람직한 실시예에서, 운영 체제(212)는 상업적으로 사용가능한 마이크로소프트 사의 WINDOWS® CE 브랜드 운영 체제이다. 운영 체제(212)는 모바일 장치용으로 설계되고, 노출된 애플리케이션 프로그래밍 인터페이스들과 메소드들의 집합을 통해 애플리케이션(214)에 의해 이용될 수 있는 데이터베이스 특징들을 구현하는 것이 바람직하다. 객체 저장소(216)의 객체들은, 노출된 애플리케이션 프로그래밍 인터페이스들과 메소드들로의 호출에 적어도 일부분 응답하여, 애플리케이션(214)과 운영 체제(212)에 의해 유지된다.

[0034] 통신 인터페이스(208)는 모바일 장치(200)로 하여금 정보를 송수신할 수 있게 하는 수많은 장치와 기술을 나타낸다. 장치로는, 몇몇 예를 들자면, 유선 및 무선 모뎀, 위성 수신기 및 브로드캐스트 튜너 등이 있다. 모바일 장치(200)는 또한 컴퓨터와의 데이터 교환을 위해 컴퓨터에 직접 접속될 수 있다. 이러한 경우, 통신 인터페이스(208)는 적외선 송수신기, 또는 직렬 또는 병렬 통신 접속일 수 있으며, 이들 모두 스트리밍 정보를 전송할 수 있다.

[0035] 입력/출력 컴포넌트(206)는 터치-센서티브 스크린, 버튼, 롤러(roller) 및 마이크와 같은 각종 입력 장치 뿐만 아니라, 오디오 생성기, 진동 장치 및 디스플레이를 비롯한 각종 출력 장치를 포함한다. 상술된 장치들은 일례이며, 모바일 장치(200) 상에 모두 존재할 필요는 없다. 게다가, 기타 입력/출력 장치들이 모바일 장치(200)에 부착되거나, 또는 모바일 장치(200)에서 발견될 수 있다.

[0036] **텍스트에 무관한 스피커 확인(text-independent speaker verification)**

[0037] 본 발명의 한 실시예 하에서, 사용자의 스피치(speech)를 훈련시키도록 적응된 모델과 테스트 스피치 신호에 적응된 모델에 기초하는 유사성 측정치(similarity measure)를 형성함으로써 테스트 스피치 신호를 인증하는, 텍스트에 무관한 스피커 인증 시스템이 제공된다. 특히, 유사성 측정치는 두 개의 적응된 모델과 백그라운드 모델(background model) 간의 차를 이용한다.

[0038] 한 실시예에서, 백그라운드 모델은 이하와 같이 정의되는 가우시안 혼합 모델(Gaussian Mixture Model)이다.

수학식 1

[0039] $$P(x_t \mid \lambda_0) = \sum_{i=1}^{M} w_i P_i(x_t \mid \lambda_0) = \sum_{i=1}^{M} w_i N(x_t : m_i, \Sigma_i)$$

[0040] 여기서, M은 모델의 혼합 성분(mixture component)의 개수이며, w_i는 i번째 혼합 성분에 대한 가중치이며, m_i는 i번째 혼합 모델에 대한 평균이며, Σ_i는 i번째 성분의 공분산 행렬(covariance matrix)이다. λ_0라는 표기는 백그라운드 모델의 파라미터들의 집합(각 성분에 대한 가중치, 평균 및 공분산)을 나타낸다.

[0041] 백그라운드 모델은 이하의 수학식들을 이용하여 스피치를 훈련시키도록 적응된다:

수학식 2

[0042] $$\hat{\gamma}(i \mid \hat{x}_t) = \frac{w_i P_i(\hat{x}_t \mid \lambda_0)}{\sum_{j=1}^{M} w_j P_j(\hat{x}_t \mid \lambda_0)}$$

수학식 3

[0043] $$\hat{\gamma}(i) = \sum_{t=1}^{T} \hat{\gamma}(i \mid \hat{x}_t)$$

수학식 4

[0044] $$\tilde{m}_i = \frac{1}{\hat{\gamma}(i)} \sum_{t=1}^{T} \hat{\gamma}(i \mid \hat{x}_t) \hat{x}_t$$

수학식 5

[0045] $$\hat{m}_i = m_i + \frac{\hat{\gamma}(i)}{\hat{\gamma}(i) + \alpha} (\tilde{m}_i - m_i)$$

수학식 6

[0046] $$\hat{\Sigma}_i = \Sigma_i$$

[0047] 여기서, $\hat{x}_t$ 은 특정 스피커로부터의 훈련 특징 벡터(training feature vector)이고, $\hat{\gamma}(i \mid \hat{x}_t)$는 스피커로부터의 훈련 특징 벡터가 주어질 경우의 i번째 혼합 성분의 사후 확률이고, T는 특정 스피커로부터의 훈련 발음의 프레임 개수이고, $\hat{\gamma}(i)$ 는 특정 스피커로부터의 훈련 발음 전체에 걸쳐 i번째 혼합 성분에 속하는 프레임의 소프트 개수(soft count)이며, α 는 훈련 발음의 i번째 혼합 성분에 대해 관찰되는 프레임이 거의 없을 경우, 적응된 모델의 평균 $\hat{m}_i$ 로 하여금 백그라운드 모델을 채택하게 하는 평활 계수(smoothing factor)이다. 상술된 실시예에서, 적응된 모델에 대한 공분산이 백그라운드 모델에 대한 공분산과 동일함을 유의한다.

[0048] 한 실시예에서, 유사성 측정치는 이하와 같이 정의된다:

수학식 7

[0049] $$LLR(x_1^T) \le \frac{\sum_{i=1}^{M} \gamma(i) \frac{\hat{\gamma}(i)}{\hat{\gamma}(i) + \alpha} \hat{\delta}_i \Sigma_i^{-1} (\delta_i - \frac{\hat{\gamma}(i)}{\hat{\gamma}(i) + \alpha} \frac{\hat{\delta}_i}{2})}{\sum_{i=1}^{M} \gamma(i)}$$

[0050] 여기서,

수학식 8

[0051] $$\delta_i = \overline{m}_i - m_i$$

수학식 9

[0052] $$\hat{\delta}_i = \hat{m}_i - m_i$$

수학식 10

[0053] $$\gamma(i) = \sum_{t=1}^{T} \gamma(i \mid x_t)$$

[0054] 여기서, x_t 는 테스트 발음의 특징 벡터이고, T는 테스트 발음의 프레임 개수이며, $\overline{m}_i$는 이하의 수학식 11에 의해 정의되는 테스트 발음의 표본 평균이다.

수학식 11

[0055] $$\overline{m}_i = \frac{1}{\gamma(i)} \sum_{t=1}^{T} \gamma(i \mid x_t) x_t$$

[0056] 따라서, 수학식 7의 유사성 측정치에서, 적은, 테스트 발음에 대한 사후 확률 γ_i, 테스트 스피커에 대한 적응된 평균과 백그라운드 평균 간의 차 $\hat{\delta}_i$, 및 테스트 발음에 대한 표본 평균과 백그라운드 평균 간의 차 δ_i로부터 형성된다.

[0057] 한 실시예 하에서, 수학식 7의 유사성 측정치는 이하의 수학식 12로 단순화된다.

수학식 12

[0058] $$LLR_0 = \frac{\sum_{i=1}^{M} \gamma(i) \hat{\gamma}(i) \hat{\delta}_i \Sigma_i^{-1} \delta_i}{\sum_{i=1}^{M} \gamma(i) \hat{\gamma}(i)}$$

[0059] 추가의 실시예 하에서, 수학식 12의 LLR_0인 데이터 의존(data dependency)을 줄이기 위해, 신중하게 임계값을 선택함으로써 정규화(normalization)가 수행된다. 한 실시예 하에서, 우선, 다수의 스피커의 발음들로부터 적응되어 온 모델 파라미터들의 집합 또는 풀(pool) 중에서 적응된 모델 파라미터들의 서브집합을 선택함으로써, 임계값들이 구축된다(constructed). 적응된 모델 파라미터들의 한 서브집합은, 훈련 발음에 가장 유사한 파라미터들의 풀에 있는 파라미터들에 의해 나타내어지는 발음들을 식별함으로써 선택된다. 모델 파라미터들의 제2 서브집합은, 테스트 발음에 가장 유사한 파라미터들의 풀에 있는 모델 파라미터들에 의해 나타내어지는 발음들을 식별함으로써 선택된다. 한 실시예 하에서, 유사성 결정은 상술된 수학식 12를 이용하여 행해진다.

[0060] 예를 들면, 훈련 발음과 유사한 발음을 찾을(locate) 때, 모델 파라미터들의 풀에서 가져온 발음에 대한 모델 파라미터들은 수학식 12에서 테스트 발음의 모델 파라미터들로서 적용되는 반면, 훈련 발음에 대한 모델 파라미터들은 수학식 12에서 바로 사용된다. 테스트 발음과 유사한 발음을 찾을 때, 모델 파라미터들의 풀에서 가져온 발음에 대한 모델 파라미터들은 훈련 발음 모델 파라미터들로서 사용되고, 테스트 발음 모델 파라미터들은 수학식 12에서 바로 사용된다.

[0061] 훈련 발음과 테스트 발음 둘 다에 대해 유사한 발음의 서브집합(코호트 스피커 집합(cohort speaker set)으로 알려짐)이 일단 선택되면, 임계값은 이하와 같이 설정될 수 있다:

수학식 13

[0062] $$\hat{t}_i^0 = \frac{1}{N_{cohort}} \sum_{k=1}^{N_{cohort}} \hat{\delta}_i \Sigma_i^{-1} \delta_i^k$$

수학식 14

[0063] $$t_i^0 = \frac{1}{N_{cohort}} \sum_{s=1}^{N_{cohort}} \delta_i \Sigma_i^{-1} \delta_i^s$$

[0064] 여기서, $\hat{t}_i^0$는 i번째 혼합 성분에서의 훈련 발음에 대한 임계값이며, t_i^0는 i번째 혼합 성분에서의 테스트 발음에 대한 임계값이며, N_{cohort}는 임계값을 형성하기 위해 스피커 풀 중에서 선택된 적응된 모델의 개수이며, $\hat{\delta}_i$는 수학식 9에서 정의된 바와 같은 훈련 발음의 i번째 성분의 조정치(adjustment)이며, δ_i는 수학식 8에서 정의된 테스트 발음의 i번째 성분의 조정치이며, δ_i^k는 훈련 발음을 위해 선택된 코호트 스피커 k의 i번째 성분의 조정치이며, δ_i^s는 테스트 발음을 위해 선택된 코호트 스피커 s의 i번째 성분의 조정치이다.

수학식 15

[0065] $$\delta_i^k = m^k - m$$

수학식 16

[0066] $$\delta_i^s = m^s - m$$

[0067] 여기서, m^k 는 m번째 코호트 발음에 대한 평균이며, m^s 는 s번째 코호트 발음에 대한 평균이다.

[0068] 이들 임계값을 이용하며, 정규화된 LLR_0는 이하의 수학식 17이다:

수학식 17

[0069] $$LLR_1 = \frac{\Sigma_{i=1}^{M}\gamma(i)\hat{\gamma}(i)[\hat{\delta}_i\Sigma_i^{-1}\delta_i - (\hat{t}_i^0 + t_i^0)/2]}{\Sigma_{i=1}^{M}\gamma(i)\hat{\gamma}(i)}$$

[0070] 수학식 17의 유사성 측정치는 훈련 발음에 대해 테스트 발음을 인증하는 데에 바로 사용될 수 있다. 일부 실시예에서, 이 유사성 측정치는, 훈련 발음과 테스트 발음 둘 다에 대해 새로운 코호트 스피커 집합을 선택하기 위해 되풀이하여(iterative) 사용된다. 이후, 이 새로운 코호트 스피커 집합은 새 임계값을 확립하는 데에 사용된다. 수학식 17의 유사성 테스트가 수학식 12의 유사성 테스트와 상이하기 때문에, 수학식 17을 이용하여 선택된 코호트 집합들이 수학식 12를 이용하여 선택된 코호트 집합들과 상이할 것임을 유의한다. 새로운 코호트 집합들을 이용하여, 새 임계값은 이하와 같이 정의된다:

수학식 18

[0071] $$\hat{t}_i^1 = \frac{1}{N_{cohort}}\sum_{k=1}^{Ncohort}[\hat{\delta}_i\Sigma_i^{-1}\delta_i^k - (\hat{t}_i^0 + t_i^0)/2]$$

수학식 19

[0072] $$t_i^1 = \frac{1}{N_{cohort}}\sum_{s=1}^{Ncohort}[\hat{\delta}_i\Sigma_i^{-1}\delta_i^s - (\hat{t}_i^0 + t_i^0)/2]$$

[0073] 이후, 새로운 유사성 측정치는 이하와 같이 정의될 수 있다:

수학식 20

[0074] $$LLR_2 = \frac{\Sigma_{i=1}^{M}\gamma(i)\hat{\gamma}(i)[\hat{\delta}_i\Sigma_i^{-1}\delta_i - (\hat{t}_i^o + t_i^o)/2 - (\hat{t}_i^1 + t_i^1)/2]}{\Sigma_{i=1}^{M}\gamma(i)\hat{\gamma}(i)}$$

[0075] 코호트들이 유사성 테스트로부터 선택되고, 새로운 임계값들이 코호트들로부터 정의되고, 새로운 유사성 측정치가 새로운 임계값들로부터 정의되는 이러한 유형의 되풀이(iteration)는 필요한 만큼 여러 번 반복될 수 있고, 각각의 새로운 유사성 테스트는, 이전의 유사성 측정치의 분자(numerator)에 있는 이전의 임계값들의 평균으로부터 두 개의 새로운 임계값들의 평균을 뺌(substract)으로써 정의된다.

[0076] 도 3은 본 발명의 한 실시예 하에서, 스피커 인증에서 사용되는 모델 파라미터들을 훈련시키는 방법의 흐름도를 제공한다. 도 4는 이 모델 파라미터들을 구축하는 데 사용되는 구성요소들의 블록도를 제공한다.

[0077] 단계(300)에서, 스피커 풀(400)에 있는 다수의 스피커들로부터의 발음들이 수신된다. 이 발음들은 아날로그-디지털 변환기(402)에 의해 디지털 값들의 시퀀스들로 변환되고, 프레임 생성기(frame constructor)(404)에 의해 프레임들로 그룹핑된다. 이후, 디지털 값들의 프레임들은 특징 추출기(feature extractor)(406)에 의해 특징 벡터들로 변환된다. 한 실시예 하에서, 특징 추출기는, 델타 계수(delta coefficients)로 MFCC(Mel-Frequency cepstral coefficient) 특징 벡터들을 형성하는 MFCC 특징 추출기이다. 이러한 MFCC 특징 추출 유닛은 당분야에 공지되어 있다. 이것은 특징 벡터들의 스피커 풀(408)을 생성한다.

[0078] 단계(302)에서, 스피커 풀 특징 벡터들은, 특징 벡터들을 사용하여, 한 실시예에서 가우시안 혼합 모델의 형태

를 취하는 범용 배경 모델(Universal Background Model; UBM)(412)을 정의하는 가우시안 혼합 모델 트레이너(Gaussian Mixture Model trainer)(410)에 적용된다. 이러한 훈련은, 특징 벡터들을 혼합 성분들로 그룹핑하는 단계와, 각각의 혼합 성분에 대해 가우시안 분포 파라미터(Gaussian distribution parameters)를 식별하는 단계를 포함한다. 특히, 각 혼합 성분에 대한 평균과 공분산 행렬이 결정된다.

[0079] 단계(304)에서, UBM 적응 유닛(UBM adaptation unit)(414)은, 상술된 수학식 2와 3을 이용하여, 스피커 풀(400)에 있는 각 스피커의 각 혼합 성분에 대한 스피커 풀 사후 확률(416)을 결정한다. 단계(306)에서, UBM 적응 유닛(414)은 이 사후 확률을 이용하여, 상술된 수학식 4 내지 6을 이용하여 스피커 풀(400)에 있는 각 스피커에 대한 스피커 풀 적응된 가우시안 혼합 모델을 결정한다. 수학식 2 내지 6에서, 특정 스피커에 대한 발음들이 조합되어 단일의 발음을 형성하고, 이것은 특징 벡터들의 시퀀스 $\hat{x}_1^T$를 형성하고, 여기서 T는 스피커의 발음들 모두에 걸친 프레임의 전체 개수이다.

[0080] 단계(308)에서, 시스템의 미래의 사용자로부터 훈련 발음(420)이 수신되고, 아날로그-디지털 변환기(402), 프레임 생성기(404) 및 특징 추출기(406)를 이용하여 사용자 훈련 특징 벡터들(422)로 변환된다. 단계(310)에서, UBM 적응 유닛(414)은 상술된 수학식 2 및 3을 이용하여 사용자 사후 확률(424)을 식별하고, 상술된 수학식 4 내지 6을 이용하여 사용자-적응된 가우시안 혼합 모델(426)을 형성한다. 단계들(308, 310 및 312)이 확인 시스템을 이용할 각 사람에 대해 반복됨을 유의한다.

[0081] 단계(314)에서, 유사성 임계값들이 훈련된다. 이들 임계값들을 훈련시키는 방법이 도 5의 흐름도에 도시되어 있다. 도 5에 도시된 방법은, 확인 시스템의 모든 사용자에 대해서뿐만 아니라, 스피커 풀의 모든 스피커에 대해 임계값들을 설정하는 되풀이되는 방법이다.

[0082] 도 5의 단계(500)에서, 스피커(스피커 풀로부터의 스피커 또는 시스템의 사용자 중 하나)가 선택된다. 단계(501)에서, 가우시안 혼합 모델 파라미터들과 선택된 스피커에 대한 사후 확률들이 선택된 스피커 모델 파라미터(433)로서 검색된다.

[0083] 단계(502)에서, 스피커 풀(400) 중에서 스피커들의 코호트를 선택하기 위해, 유사성 테스트(400)가 코호트 선택 유닛(430)에 의해 사용된다. 이 단계 동안, 스피커 풀에 있는 각 스피커와 관련된 모델 파라미터들 $(\gamma(i), m)$이, 현재 선택된 스피커에 대한 모델 파라미터들 $(\hat{\gamma}(i), \hat{m})$(433)과 함께, 유사성 테스트에 개별적으로 적용된다. 현재 선택된 스피커에 대해 가장 높은 유사성 측정치를 생성하는 스피커 풀로부터의 스피커들의 서브집합이 코호트로서 선택되고, 최종적으로 코호트 모델 파라미터들(432)이 된다. 한 실시예 하에서, 수학식 12의 유사성 테스트는, 최초의 되풀이 동안 유사성 테스트(440)로서 사용된다.

[0084] 단계(504)에서, 임계값 생성 유닛(434)은 코호트 모델 파라미터들(432)과 선택된 스피커 모델 파라미터들(433)을 이용하여, 선택된 스피커에 대한 임계값(436)을 생성한다. 한 실시예 하에서, 임계값을 산출하기 위해 수학식 13이 이용되고, 선택된 스피커 모델 파라미터들(433)로부터의 평균이 조정 값 $\hat{\delta}_i$을 정의하기 위해 사용되고, 코호트 모델 파라미터들(432)에 대한 평균이 각 코호트에 대한 δ_i^k을 정의하기 위해 사용된다.

[0085] 단계(506)에서, 도 5의 방법은, 스피커 풀에 또는 시스템 사용자들의 집합에 스피커가 더 있는지의 여부를 판정한다. 스피커가 더 있는 경우, 단계(500)로 리턴함으로써 다음 스피커가 선택되며, 새 스피커에 대한 코호트를 식별하기 위해 유사성 테스트(440)가 다시 사용된다. 이후, 새 스피커에 대한 임계값이 결정된다. 단계들(500, 502, 504 및 506)은, 스피커 풀의 모든 스피커 또는 시스템의 모든 사용자에 대해 임계값이 결정될 때까지 반복된다.

[0086] 더 이상 스피커가 없는 경우, 유사성 테스트 생성 유닛(438)은 단계(508)에서 새로운 유사성 테스트(440)를 생성한다. 한 실시예 하에서, 새로운 유사성 테스트는 상술된 수학식 17과 같이 정의된다.

[0087] 단계(510)에서, 본 방법은, 유사성 테스트가 수렴되었는지(converged)의 여부를 판정한다. 테스트가 수렴되지 않았다면, 프로세스는, 스피커가 스피커 풀에서 선택되거나 또는 시스템의 사용자들 집합에서 선택되는 단계(500)로 리턴한다. 이후, 코호트 스피커들을 선택하는 단계(502)가 사용되며, 이때 유사성 테스트 생성 유닛(438)에 의해 설정된 새로운 유사성 테스트(440)를 이용한다. 이후 단계(504)에서 새로이 선택된 코호트를 이용하여 새로운 임계값들(436)이 결정된다. 예를 들면, 일부 실시예 하에서, 두 번째 되풀이 동안 단계(504)에서 새로운 임계값들을 결정하기 위해 수학식 18이 이용된다. 단계들(500, 502, 504 및 506)은, 스피커 풀의 모

든 스피커 또는 시스템의 각 사용자에 대해 반복된다. 각 스피커에 대해 새로운 임계값들이 결정된 후에, 단계(508)에서 새로운 유사성 테스트가 정의된다. 예를 들면, 두 번째 되풀이 동안, 새로운 유사성 테스트는 수학식 20에서 확인되는 바와 같이 정의될 것이다.

[0088] 유사성 테스트를 이용하여 코호트들을 결정하는 것, 코호트들로부터 임계값들을 정의하는 것 및 새로운 임계값에 기초하여 유사성 테스트들을 재정의하는 것의 되풀이는, 유사성 테스트에서의 변경이 선택된 코호트 집합을 변경시키지 않도록, 단계(510)에서 유사성 테스트가 수렴할 때까지 되풀이되며 반복된다. 훈련 동안 임계값들을 설정하는 단계는 이후 단계(512)에서 종료된다.

[0089] 일단 모델들이 적응되었고, 그리고 스피커 풀의 모든 스피커와 시스템의 각 사용자에 대해 임계값들이 설정되었다면, 시스템은 사용자를 인증하기 위해 사용될 수 있다. 도 6의 흐름도 및 도 7의 블록도에 도시된 바와 같이, 테스트 발음에 대한 모델 파라미터들을 설정함으로써 인증이 시작된다. 도 6의 단계(600)에서, 도 7의 테스트 발음(700)이 수신된다. 테스트 발음은 아날로그-디지털 변환기(702)에 의해 디지털 값들의 시퀀스로 변환되고, 프레임 생성 유닛(704)에 의해 프레임들로 그룹핑된다. 디지털 값들의 프레임들이 특징 추출기(706)에 적용되고, 이 특징 추출기(706)는 도 4의 특징 추출기(406)와 동일한 특징 추출을 수행하여 테스트 발음 특징 벡터들(708)을 생성한다.

[0090] 단계(602)에서, 적응 유닛(710)은, 상술된 수학식 2 및 3을 이용하여, 범용 배경 모델(412)에 기초하여, 테스트 별(test-specific) 사후 확률들(712)을 형성한다. 단계(604)에서, 범용 배경 모델이 적응 유닛(710)에 의해 적응되어, 상술된 수학식 4 내지 6을 이용하여 테스트 적응된 GMM(Gaussian Mixture Model)(714)을 형성하며, 테스트 발음은 $\hat{x}_t$ 로서 사용된다.

[0091] 단계(606)에서, 테스트 발음에 대한 유사성 임계값들(724)이 결정된다. 유사성 임계값들을 결정하는 방법은 도 8의 흐름도에 더 상세하게 도시되어 있다.

[0092] 도 8의 단계(800)에서, 코호트 선택 유닛(718)에 의해 유사성 테스트(716)가 사용되어, 테스트 스피커와 가장 유사한, 스피커 풀에 있는 스피커들을 찾아낸다. 이 단계 동안, 스피커 풀의 각 스피커에 관련된 모델 파라미터들 $\{\gamma(i),m\}$ 은, 테스트 발음에 대한 모델 파라미터들 $\{\hat{\gamma}(i),\hat{m}\}$ (712, 714)과 함께, 유사성 테스트에 개별적으로 적용된다. 현재 선택된 스피커에 대해 가장 높은 유사성 측정치를 생성하는 스피커 풀로부터의 스피커들의 서브집합이 코호트로서 선택되고, 최종적으로 코호트 모델 파라미터(720)의 집합이 된다. 한 실시예 하에서, 수학식 12의 유사성 테스트는, 최초의 되풀이 동안 유사성 테스트(716)로서 사용된다.

[0093] 단계(802)에서, 임계값 생성 유닛(722)은 코호트 모델 파라미터들(720)과 테스트-적응된 GMM들(714)을 이용하여, 테스트 발음 임계값들(724)을 형성한다. 한 실시예 하에서, 임계값을 산출하기 위해 수학식 14가 이용되고, 테스트-적응된 GMM들(714)로부터의 평균들은 조정 값 δ_i 을 정의하기 위해 이용되고, 코호트 모델 파라미터들(720)에 대한 평균들은 각 코호트에 대한 δ_i^s 을 정의하기 위해 이용된다.

[0094] 단계(804)에서, 새로운 유사성 테스트(716)가, 단계(802)에서 설정된 테스트 발음 임계값들(724)과 도 5의 방법에서 설정된 스피커 풀 임계값들(436)을 이용하여, 유사성 테스트 생성 유닛(726)에 의해 형성된다. 한 실시예 하에서, 수학식 17의 유사성 테스트는 새로운 유사성 테스트(716)로서 사용된다. 단계(806)에서, 도 5의 흐름도에서 수행되었던 것과 동일한 회수의 되풀이에 도달했는지의 여부를 판정한다. 동일한 회수의 되풀이가 수행되지 않았다면, 단계(800)로 리턴함으로써 새로운 코호트들의 집합을 선택하기 위해 새로운 유사성 테스트가 사용된다. 새로운 코호트들(720)이 임계값 생성 유닛(722)에 의해 사용되어 새로운 테스트 발음 임계값들을 형성하며, 이 새로운 테스트 발음 임계값들은 테스트 스피커 임계값들(724)에 추가된다. 이 새로운 임계값들은, 단계(804)에서 유사성 테스트 생성 유닛(726)에 의해 사용되어, 수학식 20의 유사성 테스트와 같은 새로운 유사성 테스트를 형성한다. 단계들(800, 802, 804 및 806)은, 도 5의 방법에서 수행되었던 것과 동일한 회수의 되풀이가 수행될 때까지 도 8의 방법에서 반복되어, 최종적으로, 최종 유사성 테스트(716)가 도 5의 흐름도를 통해 형성된 최종 유사성 테스트(440)와 동일한 개수의 임계값들을 갖게 된다. 동일한 회수의 되풀이에 도달하면, 테스트 발음에 대한 유사성 임계값들을 산출하기 위한 프로세스는 단계(808)에서 종료한다.

[0095] 스피커 인증은, 도 10의 블록도의 구성요소들을 이용하여 도 9에 도시된 프로세스로 계속된다. 단계(900)에서, 명목상의 사용자 ID(nominal user identification)(1000)가 수신된다. 이 명목상의 사용자 ID를 이용하여, 단계(902)에서, 명목상의 사용자에 대한 적응된 가우시안 혼합 모델(1002), 사후 확률(1004) 및 임계값(1006)이

수신된다. 이들 파라미터들은, 도 3의 흐름도의 명목상의 사용자의 훈련 발음들로부터 결정되었다.

[0096] 단계(904)에서, 도 7의 테스트 발음 적응된 가우시안 혼합 모델(714), 테스트 발음 사후 확률(712) 및 테스트 발음 임계값(724)이 검색된다.

[0097] 단계(906)에서, 유사성 스코어링 모듈(similarity scoring module)(1010)에 의해 최종 유사성 테스트(716)가 사용되어, 테스트 발음 모델 파라미터들(712, 714, 724)과 명목상의 사용자 모델 파라미터들(1002, 1004, 1006) 간의 유사성 점수(1012)를 형성한다. 한 실시예 하에서, 최종 유사성 테스트(716)는 수학식 20의 유사성 테스트이다. 단계(908)에서, 스피커 인증 유닛(1014)에 의해 유사성 점수(1012)가 사용되어, 테스트 발음이 명목상의 사용자 ID(1000)에 의해 식별되는 사용자로부터의 것인지의 여부에 관해 결정을 내린다.

[0098] **텍스트에 좌우되는 스피커 인증(text-dependent speaker authentication)**

[0099] 본 발명의 추가의 실시예 하에서, 은닉 마르코프 모델(Hidden Markov Model)이 구축되어 스피커 인증을 수행하기 위해 사용되는, 텍스트에 종속되는 스피커 인증 시스템이 제공된다. 도 11은 이러한 은닉 마르코프 모델을 훈련시키는 방법을 제공하고, 도 12는 이 은닉 마르코프 모델을 훈련시키는 데에 사용되는 구성요소들의 블록도를 제공한다.

[0100] 도 11의 단계(1100)에서, 텍스트에 무관한 범용 배경 모델이 훈련된다. 한 실시예 하에서, 범용 배경 모델은 스피커 풀(1200)의 서로 다른 많은 스피커들로부터 텍스트에 무관한 스피치를 수집함으로써 훈련되는 가우시안 혼합 모델이다. 스피치 풀(1200)의 각 발음은 아날로그-디지털 변환기(1202)에 의해 디지털 값들의 시퀀스로 변환되고, 디지털 값들은 프레임 구성 유닛(1204)에 의해 프레임들로 그룹핑된다. 각 프레임에 대해, 특징 추출 유닛(1206)은 특징 벡터를 추출하며, 이것은 한 실시예에서 델터 벡터를 갖는, MFCC이다. 추출된 특징 벡터들(1208)은 가우시안 혼합 모델 트레이너(1210)에 적용되어, 범용 배경 모델(1212)을 형성한다. 가우시안 혼합 모델 트레이너는 당분야에 공지되어 있으며, 특징 벡터들을 혼합 성분들로 그룹핑하고, 각 성분에 할당된 특징 벡터들의 분포를 기술하는 가우시안 파라미터들을 식별함으로써 가우시한 혼합 모델을 형성한다.

[0101] 단계(1101)에서, 훈련 발음들(1216)이 수신되며, 아날로그-디지털 변환기(1218)에 의해 디지털 값들로 변환되며, 프레임 생성 유닛(1220)에 의해 프레임들로 그룹핑된다. 각 프레임에 대해, 특징 추출 유닛(1222)은 특징 벡터를 추출하여 훈련 특징 벡터들(1224)을 형성하며, 이것은 스피커 풀 특징 벡터들(1208)과 동일한 유형의 벡터들이다. 한 실시예 하에서, 훈련 발음들(1216)은 한 단어 또는 한 구를 반복하는 단일의 스피커에 의해 형성된다.

[0102] 단계(1102)에서, 베이스라인 은닉 마르코프 모델 상태 확률 파라미터들(baseline Hidden Markov Model state probability parameters)(1213)를 정의하는 데 범용 배경 모델(1212)이 사용된다. 한 실시예 하에서, 이것은 각 혼합 성분의 평균과 공분산을 대응하는 은닉 마르코프 모델 상태의 평균 및 공분산으로 설정함으로써 수행된다.

[0103] 단계(1103)에서, 범용 배경 모델(1212)이 적응 유닛(1226)에 의해 특정 스피커에 적응되며, HMM(Hidden Markov Model) 상태 확률 파라미터들(1214)로 변환된다. 특히, 훈련 특징 벡터들(1224)이 가우시안 혼합 모델 적응 유닛(1226)에 제공되며, 가우시안 혼합 모델 적응 유닛(1226)은 또한 범용 배경 모델(1212)을 수신한다. 가우시안 혼합 모델 적응 유닛(1226)은 상술된 수학식 2 내지 6을 사용하여 범용 배경 모델을 적응시키면서, 훈련 특징 벡터들을 x_t로서 사용한다. 각각의 혼합 성분에 대한 최종적인 평균과 공분산이, 대응하는 HMM 상태 확률 분포에 대한 모델 매개변수들로서 저장된다. 따라서, 각 혼합 성분은 개별의 HMM 상태를 나타낸다.

[0104] 단계(1104)에서, 훈련 특징 벡터들(1224)은 특징 벡터들의 시퀀스를 디코딩하는 은닉 마르코프 모델 디코더(1228)에 적용되어, 특징 벡터들(1224)의 시퀀스가 주어질 경우, 가장 확실한(most probable) HMM 상태들의 시퀀스(1230)를 식별한다. 이 디코딩을 수행하기 위해, HMM 디코더(1228)는 HMM 상태 확률 파라미터들(1214)과 HMM 천이(transition) 확률 파라미터들(1232)의 초기 집합을 이용한다. 한 실시예 하에서, HMM 천이 확률은 당초에, 두 상태 간의 천이의 확률이 모든 상태에 대해 동일하도록 균일한 값으로 설정된다.

[0105] 단계(1106)에서, 디코딩된 상태 시퀀스(1230)가 천이 확률 산출기(transition probability calculator)(1234)에 의해 사용되어, HMM 천이 확률 파라미터들(1232)을 훈련시킨다. 이 산출은, 각종 상태 간의 천이의 횟수를 세는 단계와 그 횟수에 기초하여 각 천이에 확률을 할당하는 단계를 포함한다. 단계(1108)에서, 훈련 특징 벡터들(1224)은 HMM 디코더(1228)에 의해 다시 한 번 디코딩되며, 이때, 새로운 HMM 천이 확률 매개변수들(1232)과 HMM 상태 확률 매개변수들(1214)을 사용한다. 이것은 새로운 디코딩된 상태 시퀀스(1230)를 형성한다. 단

계(1110)에서, 본 방법은, 디코딩된 상태 시퀀스가 수렴되었는지의 여부를 판정한다. 수렴되지 않았다면, 단계(1106)로 리턴함으로써 HMM 천이 확률 파라미터들(1232)을 재훈련시키는 데에 새로운 상태 시퀀스가 사용된다. 단계(1108)에서, 훈련 특징 벡터들(1224)은 새로운 천이 확률 파라미터들을 이용하여 다시 디코딩된다. 단계들(1106, 1108 및 1101)은, 출력 HMM 상태 시퀀스가 안정될 때까지 반복되며, HMM 훈련은 단계(1112)에서 완료된다.

[0106] 일단 은닉 마르코프 모델이 훈련되면, 이것은 도 13의 흐름도와 도 14의 블록도에 도시된 바와 같이 스피커 인증을 수행하는 데에 사용될 수 있다. 도 13의 단계(1300)에서, 명목상의 사용자 ID(nominal user identification)(1400)가 수신되고, 이것은 단계(1302)에서, 은닉 마르코프 모델 상태 확률 파라미터들(1404)과 은닉 마르코프 모델 천이 확률 파라미터들(1406)을 선택하기 위해, HMM 검색 유닛(1402)에 의해 이용된다.

[0107] 단계(1304)에서, 테스트 발음(1408)이 수신된다. 테스트 발음은 아날로그-디지털 변환기(1410)에 의해 디지털 값들의 시퀀스로 변환되고, 이 디지털 값들의 시퀀스는 프레임 생성 유닛(1412)에 의해 프레임들로 그룹핑된다. 각 프레임에 대해, 특징 추출기(1414)는 특징 벡터들(1416)의 시퀀스를 형성하는 특징 벡터를 추출한다.

[0108] 단계(1306)에서, 테스트 발음 특징 벡터들(1416)이 은닉 마르코프 모델 디코더(1418)에 적용되며, 이 은닉 마르코프 모델 디코더(1418)는, 범용 배경 모델(1420)로부터 생성되는 베이스라인 은닉 마르코프 모델 상태 확률 파라미터들(1213)과, 도 11의 방법을 이용하여 훈련된 HMM 천이 확률 파라미터들(1406)로 구성된 베이스라인 은닉 마르코프 모델을 이용하여 특징 벡터들을 디코딩한다. HMM 디코더(1418)는, 베이스라인 HMM 상태 확률 파라미터들(1213)과 HMM 천이 확률 파라미터들(1406)이 주어지면, 가장 확실한 상태 시퀀스에 대한 베이스라인 확률(1422)을 생성한다.

[0109] 단계(1308)에서, HMM 디코더(1418)는 명목상의 사용자 ID로부터 식별되는 은닉 마르코프 모델 상태 확률 파라미터들(1404)과 HMM 천이 확률 파라미터들(1406)을 이용하여 특징 벡터들(1416)을 디코딩한다. 이 디코딩은 최종적으로 명목상의 사용자 확률(1424)이 되며, 이것은 확률 파라미터들(1404)과 HMM 천이 확률 파라미터들(1406)이 주어질 경우, 식별되는 HMM 상태들의 가장 확실한 시퀀스에 대한 확률을 제공한다.

[0110] 단계(1310)에서, 명목상의 사용자 확률(1424)과 베이스라인 확률(1422)의 비(ratio)가 스코어링 모듈(1428)에 의해 로그 함수에 적용되어, 로그 우도비 점수(log likelihood ratio score)(1426)를 결정한다. 단계(1312)에서, 이 점수는 인증 모듈(1430)에 의해 임계값과 비교되어, 테스트 발음이 명목상의 사용자 ID에 의해 식별되는 스피커로부터의 것인지의 여부를 판정한다.

[0111] 본 내용이 구조적인 특징들 및/또는 방법론적인 액트에 특정한 언어로 기술되었지만, 첨부된 청구항에 정의된 본 내용이 상술된 특정 특징들 또는 액트에 제한될 필요는 없음을 이해할 것이다. 오히려, 상술된 특정 특징들 및 액트들은 청구항을 구현하는 예시적인 형태로서 개시된다.

도면의 간단한 설명

[0006] 도 1은 일부 실시예들이 실시될 수 있는 한 컴퓨팅 환경의 블록도.

[0007] 도 2는 일부 실시예들이 실시될 수 있는 대안의 컴퓨팅 환경의 블록도.

[0008] 도 3은 텍스트에 무관한 인증 시스템을 훈련시키는 방법의 흐름도.

[0009] 도 4는 텍스트에 무관한 인증 시스템을 훈련시키는 데 사용되는 구성요소들의 블록도.

[0010] 도 5는 훈련 동안 임계값을 설정하기 위한 방법의 흐름도.

[0011] 도 6은 테스트 발음에 대한 모델 파라미터를 식별하는 방법의 흐름도.

[0012] 도 7은 도 6 및 도 8의 방법에서 사용되는 구성요소들의 블록도.

[0013] 도 8은 테스트 발음에 대한 임계값을 결정하기 위한 방법의 흐름도.

[0014] 도 9는 테스트 발음을 인증하는 방법의 흐름도.

[0015] 도 10은 테스트 발음을 인증하는 데 사용되는 구성요소들의 블록도.

[0016] 도 11은 텍스트에 좌우되는 인증 시스템을 위해 은닉 마르코프 모델(Hidden Markov Model)을 훈련시키는 방법의 흐름도.

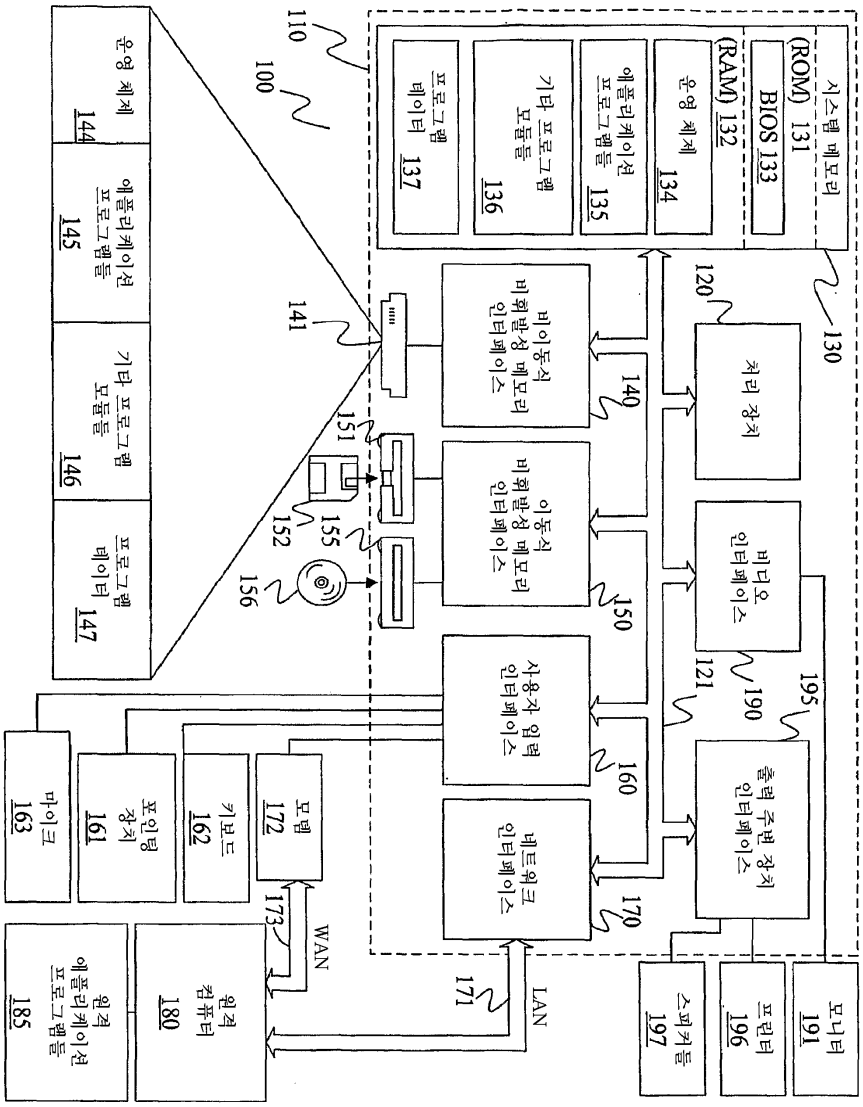
[0017] 도 12는 은닉 마르코프 모델을 훈련시키는 데 사용되는 구성요소들의 블록도.

[0018] 도 13은 은닉 마르코프 모델을 이용하여 테스트 발음을 인증하는 방법의 흐름도.

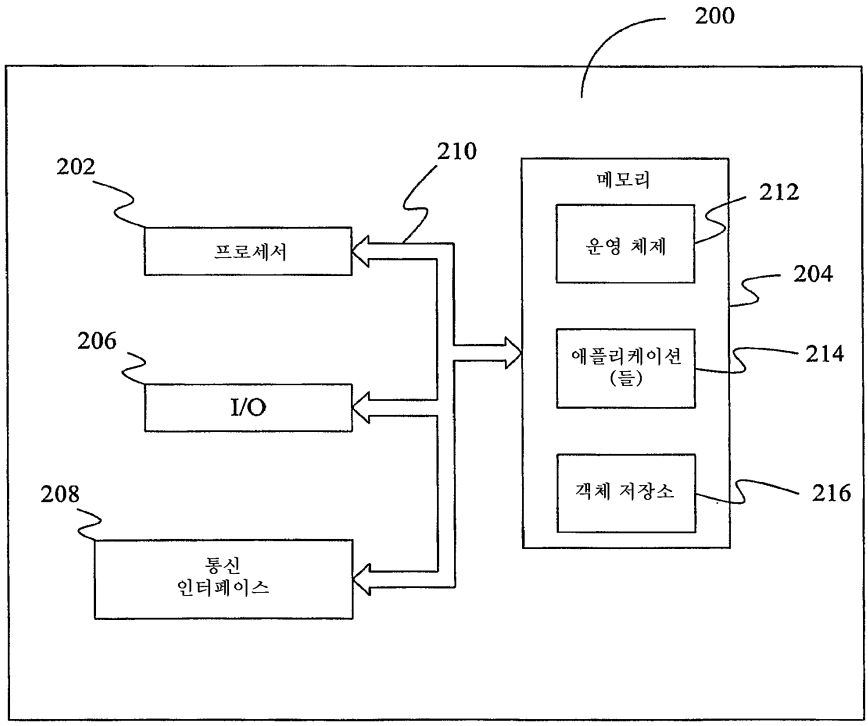
[0019] 도 14는 은닉 마르코프 모델을 이용하여 테스트 발음을 인증하는 데 사용되는 구성요소들의 블록도.

도면

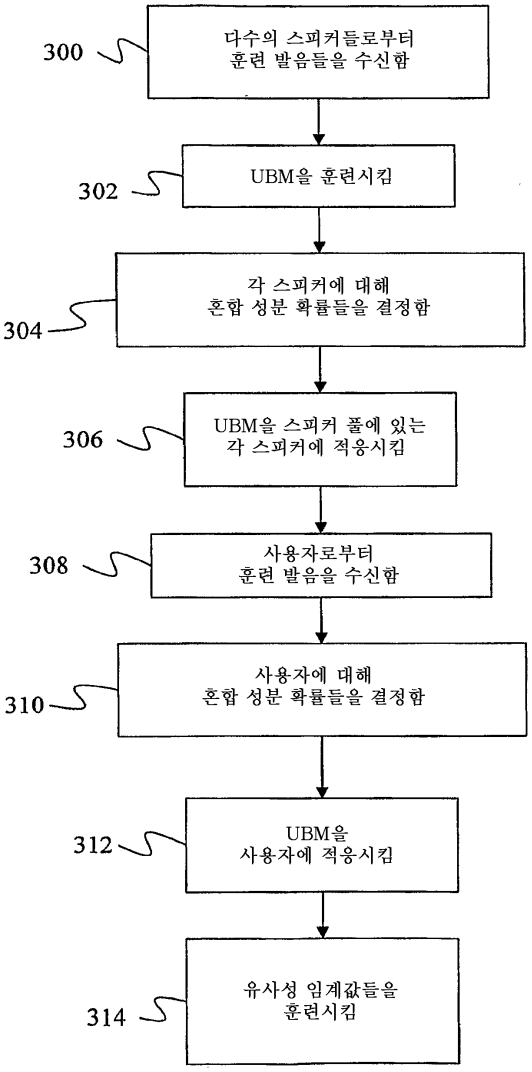
도면1



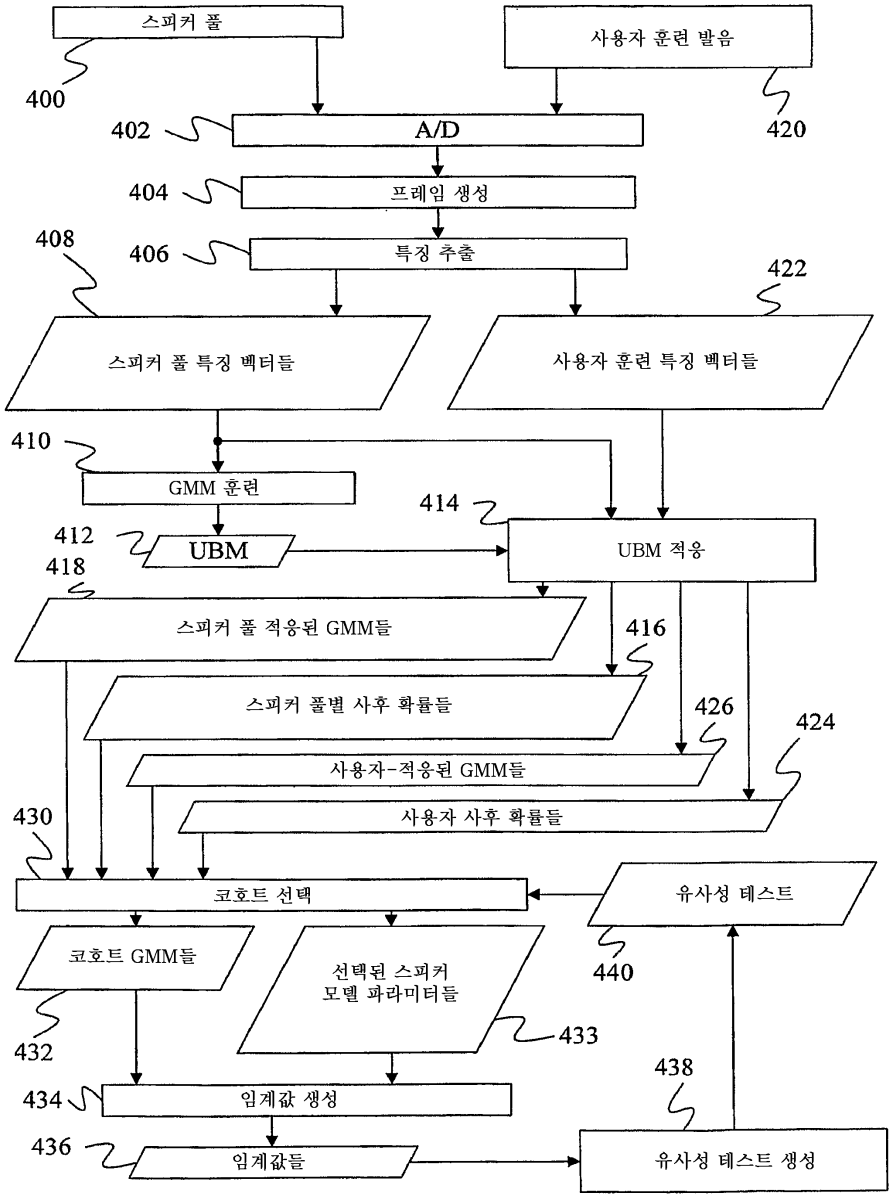
도면2


200
202
210
메모리
212
프로세서
운영 체제
204
206
I/O
애플리케이션
(들)
214
208
객체 저장소
216
통신
인터페이스

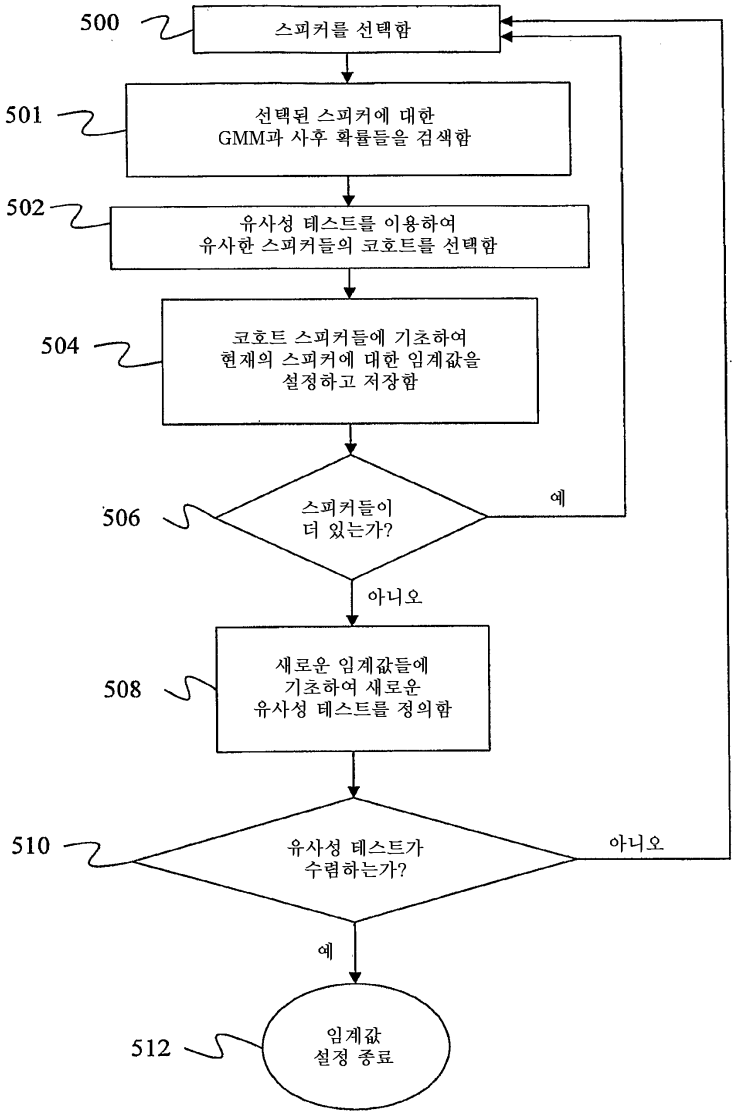
도면3


300
다수의 스피커들로부터
훈련 발음들을 수신함
302
UBM을 훈련시킴
304
각 스피커에 대해
혼합 성분 확률들을 결정함
306
UBM을 스피커 풀에 있는
각 스피커에 적응시킴
308
사용자로부터
훈련 발음을 수신함
310
사용자에 대해
혼합 성분 확률들을 결정함
312
UBM을
사용자에 적응시킴
314
유사성 임계값들을
훈련시킴

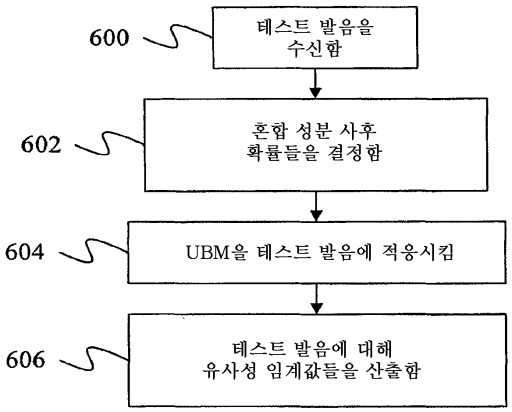
도면4


스피커 풀
사용자 훈련 발음
400
402
A/D
420
404
프레임 생성
408
406
특징 추출
422
스피커 풀 특징 벡터들
사용자 훈련 특징 벡터들
410
GMM 훈련
414
412
UBM
UBM 적응
418
스피커 풀 적응된 GMM들
416
스피커 풀별 사후 확률들
426
424
사용자-적응된 GMM들
사용자 사후 확률들
430
코호트 선택
유사성 테스트
코호트 GMM들
선택된 스피커
모델 파라미터들
440
432
433
438
434
임계값 생성
436
임계값들
유사성 테스트 생성

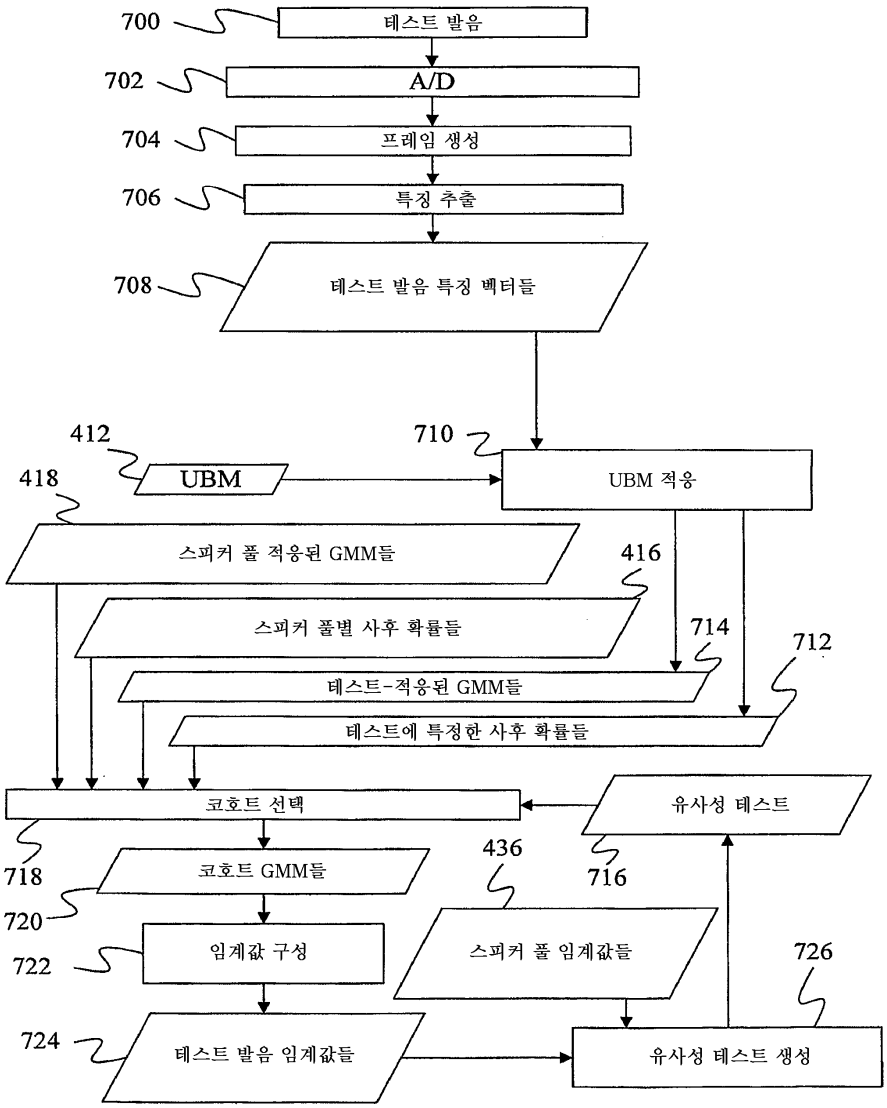
도면5


500
스피커를 선택함
501
선택된 스피커에 대한
GMM과 사후 확률들을 검색함
502
유사성 테스트를 이용하여
유사한 스피커들의 코호트를 선택함
504
코호트 스피커들에 기초하여
현재의 스피커에 대한 임계값을
설정하고 저장함
506
스피커들이
더 있는가?
예
아니오
508
새로운 임계값들에
기초하여 새로운
유사성 테스트를 정의함
510
유사성 테스트가
수렴하는가?
아니오
예
512
임계값
설정 종료

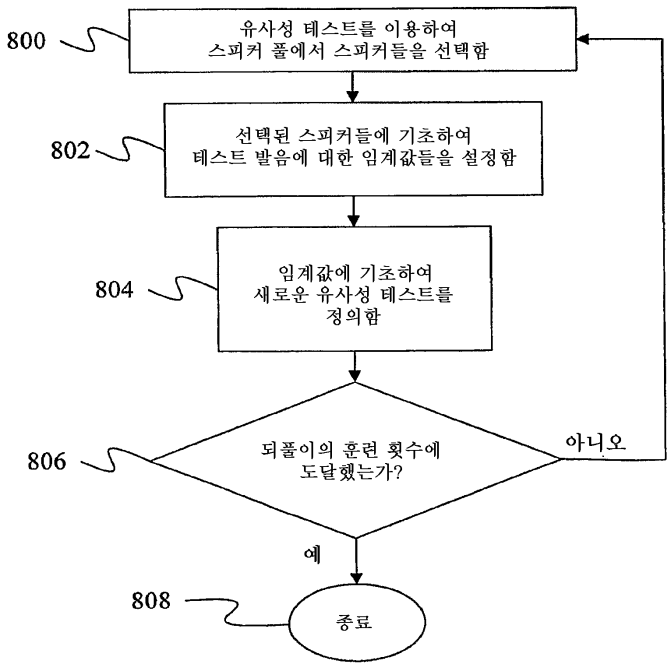
도면6


600
테스트 발음을
수신함
602
혼합 성분 사후
확률들을 결정함
604
UBM을 테스트 발음에 적응시킴
606
테스트 발음에 대해
유사성 임계값들을 산출함

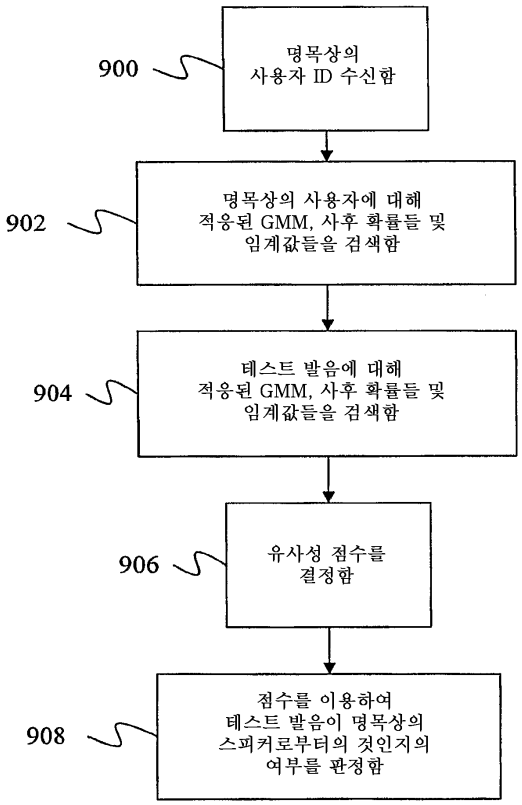
도면7


700
테스트 발음
702
A/D
704
프레임 생성
706
특징 추출
708
테스트 발음 특징 벡터들
412
UBM
710
UBM 적응
418
스피커 풀 적응된 GMM들
416
스피커 풀별 사후 확률들
714
712
테스트-적응된 GMM들
테스트에 특정한 사후 확률들
코호트 선택
유사성 테스트
718
코호트 GMM들
436
716
720
722
임계값 구성
스피커 풀 임계값들
726
724
테스트 발음 임계값들
유사성 테스트 생성

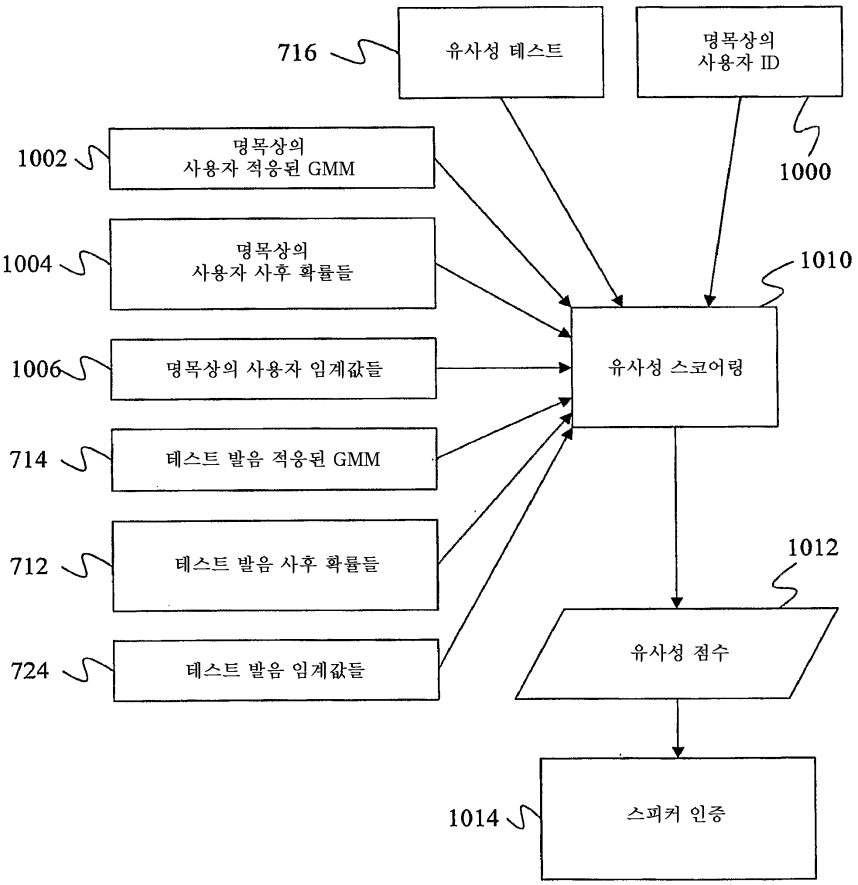
도면8


800
유사성 테스트를 이용하여
스피커 풀에서 스피커들을 선택함
802
선택된 스피커들에 기초하여
테스트 발음에 대한 임계값들을 설정함
804
임계값에 기초하여
새로운 유사성 테스트를
정의함
806
되풀이의 훈련 횟수에
도달했는가?
아니오
예
808
종료

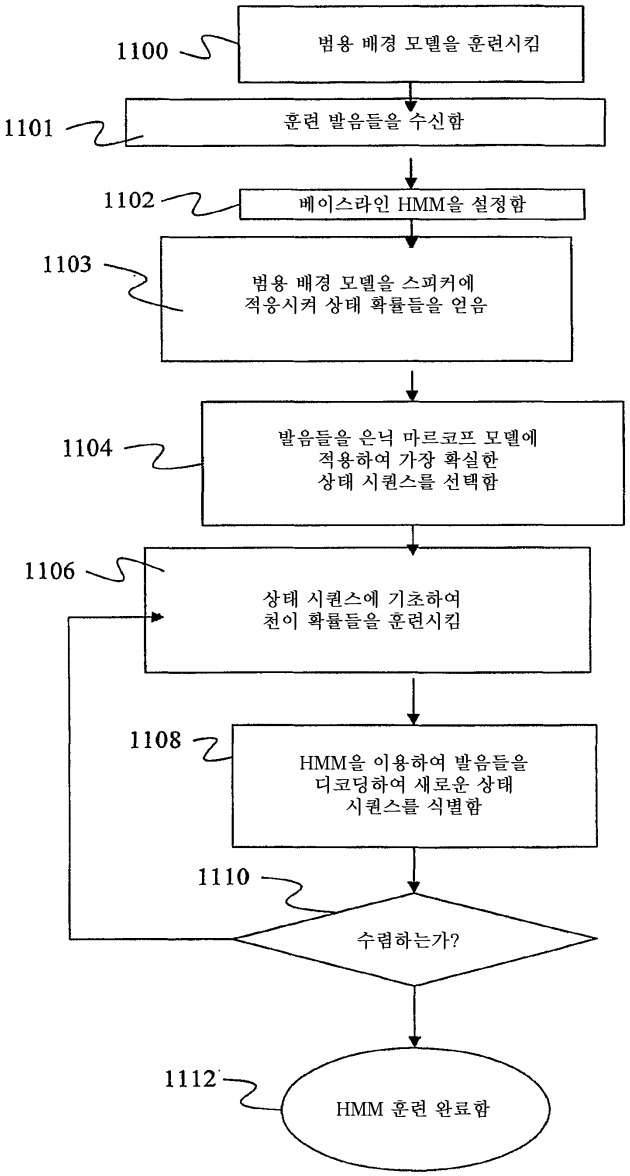
도면9


900
명목상의
사용자 ID 수신함
902
명목상의 사용자에 대해
적응된 GMM, 사후 확률들 및
임계값들을 검색함
904
테스트 발음에 대해
적응된 GMM, 사후 확률들 및
임계값들을 검색함
906
유사성 점수를
결정함
908
점수를 이용하여
테스트 발음이 명목상의
스피커로부터의 것인지의
여부를 판정함

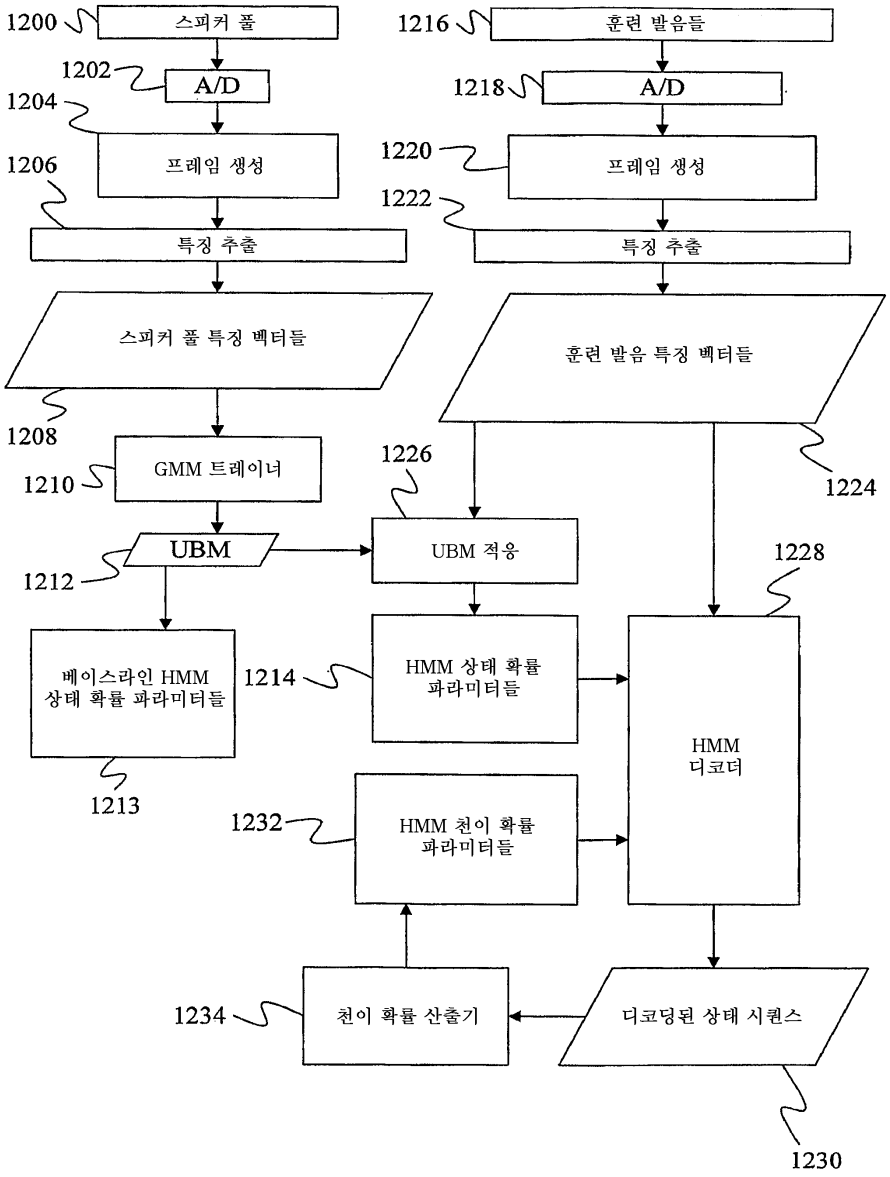
도면10


716
유사성 테스트
명목상의
사용자 ID
1000
1002
명목상의
사용자 적응된 GMM
1004
명목상의
사용자 사후 확률들
1010
1006
명목상의 사용자 임계값들
유사성 스코어링
714
테스트 발음 적응된 GMM
1012
712
테스트 발음 사후 확률들
유사성 점수
724
테스트 발음 임계값들
1014
스피커 인증

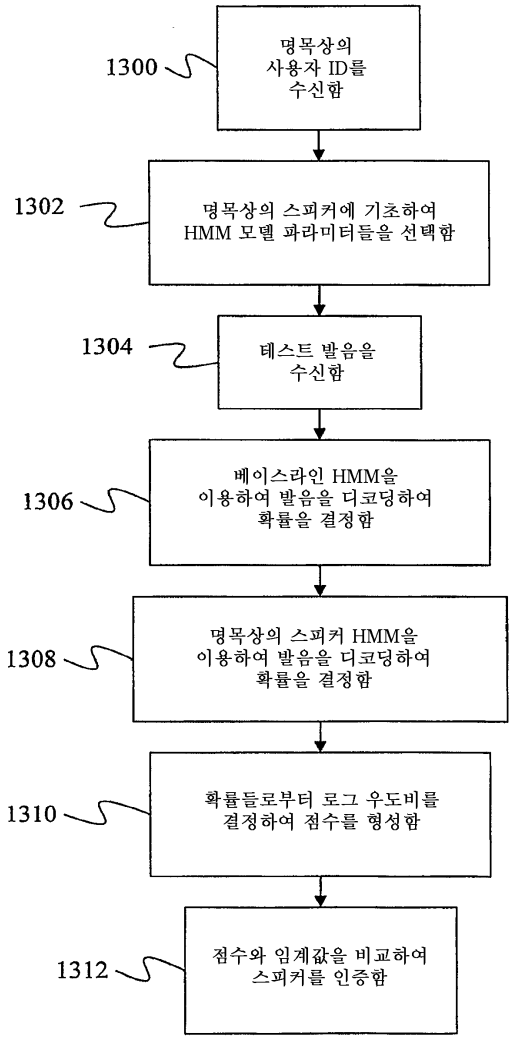
도면11


1100
범용 배경 모델을 훈련시킴
1101
훈련 발음들을 수신함
1102
베이스라인 HMM을 설정함
1103
범용 배경 모델을 스피커에
적응시켜 상태 확률들을 얻음
1104
발음들을 은닉 마르코프 모델에
적용하여 가장 확실한
상태 시퀀스를 선택함
1106
상태 시퀀스에 기초하여
천이 확률들을 훈련시킴
1108
HMM을 이용하여 발음들을
디코딩하여 새로운 상태
시퀀스를 식별함
1110
수렴하는가?
1112
HMM 훈련 완료함

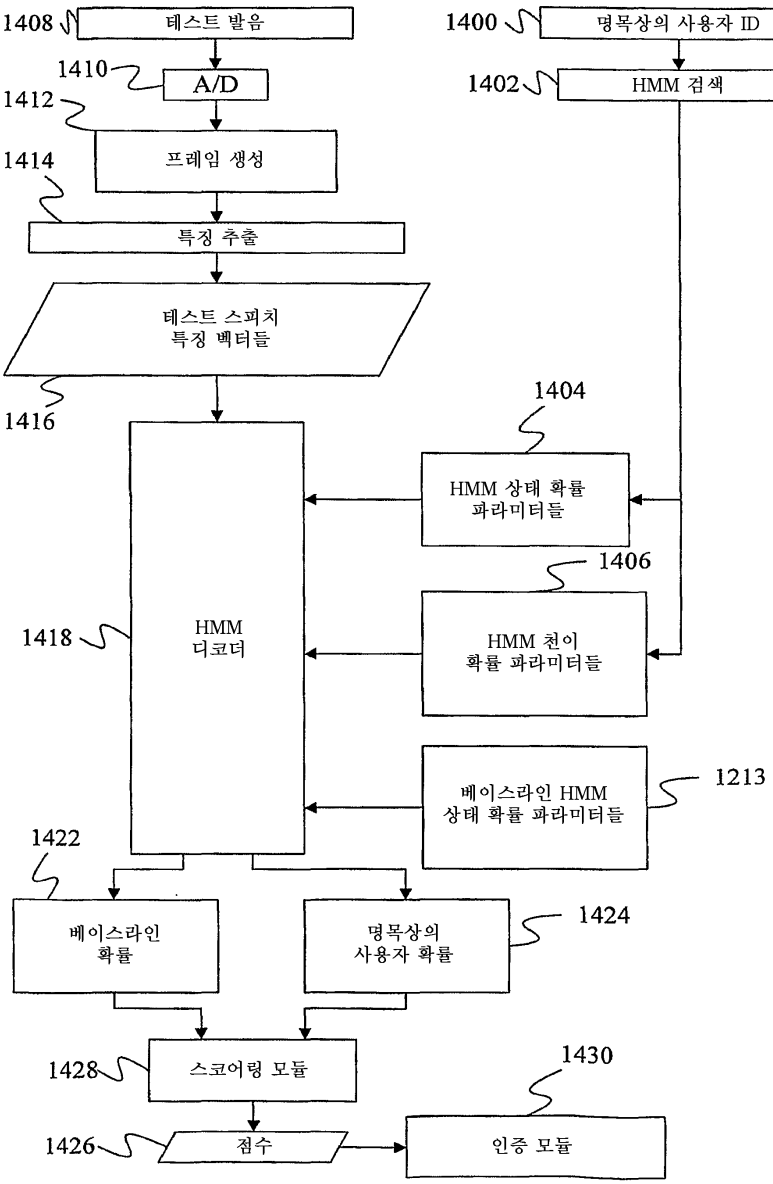
도면12


1200
스피커 풀
1202
A/D
1204
프레임 생성
1206
특징 추출
스피커 풀 특징 벡터들
1208
1210
GMM 트레이너
1212
UBM
베이스라인 HMM
상태 확률 파라미터들
1213
1214
HMM 상태 확률
파라미터들
1216
훈련 발음들
1218
A/D
1220
프레임 생성
1222
특징 추출
훈련 발음 특징 벡터들
1224
1226
UBM 적응
1228
HMM
디코더
1232
HMM 천이 확률
파라미터들
1234
천이 확률 산출기
디코딩된 상태 시퀀스
1230

도면13


1300
명목상의 사용자 ID를 수신함
1302
명목상의 스피커에 기초하여 HMM 모델 파라미터들을 선택함
1304
테스트 발음을 수신함
1306
베이스라인 HMM을 이용하여 발음을 디코딩하여 확률을 결정함
1308
명목상의 스피커 HMM을 이용하여 발음을 디코딩하여 확률을 결정함
1310
확률들로부터 로그 우도비를 결정하여 점수를 형성함
1312
점수와 임계값을 비교하여 스피커를 인증함

도면14


1408
테스트 발음
1400
명목상의 사용자 ID
1410
A/D
1402
HMM 검색
1412
1414
프레임 생성
특징 추출
테스트 스피치
특징 벡터들
1416
1404
HMM 상태 확률
파라미터들
1406
HMM 천이
확률 파라미터들
1418
HMM
디코더
베이스라인 HMM
상태 확률 파라미터들
1213
1422
베이스라인
확률
명목상의
사용자 확률
1424
1428
스코어링 모듈
1430
1426
점수
인증 모듈