(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2010-0098007
(43) 공개일자 2010년09월06일

(51) Int. Cl.
G10L 17/00 (2006.01) *G10L 15/18* (2006.01)
(21) 출원번호 10-2009-0016954
(22) 출원일자 2009년02월27일
심사청구일자 2009년02월27일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
육동석
서울 용산구 서빙고동 신동아아파트 1동 102호
조영규
서울특별시 마포구 공덕2동 공덕현대아파트 102-1305
(74) 대리인
특허법인충현

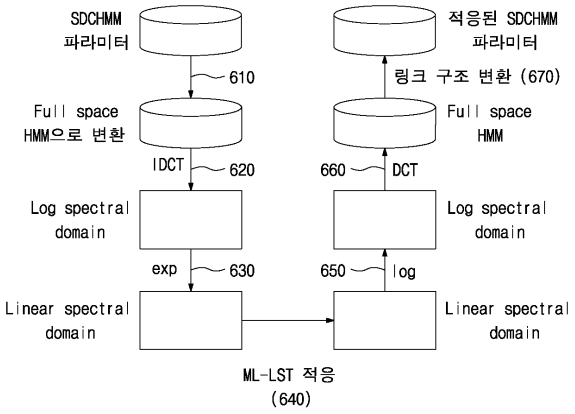
전체 청구항 수 : 총 12 항

(54) 고속 화자 인식 방법 및 장치, 고속 화자 인식을 위한 등록방법 및 장치

(57) 요 약

본 발명은 고속 화자 인식을 위한 등록 방법에 관한 것으로, 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 방법은 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하는 단계; 상기 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 단계; 및 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하는 단계를 포함한다. 본 방법에 의하면, 휴대폰과 같은 소형 이동 단말기에서 화자 모델을 적은 적응 데이터로도 빠르게 적응시킬 수 있을 뿐만 아니라, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있으며, 인식 과정에서의 계산량을 대폭 줄일 수 있어서 서버와 같은 다른 기기의 도움을 받을 필요 없이 소형 단말기 스스로 화자 인식을 처리 할 수 있는 효과가 있다.

대 표 도 - 도6



특허청구의 범위

청구항 1

부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하는 단계;

상기 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 단계; 및

최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하는 단계

를 포함하는, 고속 화자 인식을 위한 등록 방법.

청구항 2

제 1 항에 있어서,

상기 선형 스펙트럼 도메인으로 변환하는 단계는,

이산 코사인 역변환 (IDCT)을 이용하여 상기 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 단계; 및

상기 로그 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 단계를 포함하는, 고속 화자 인식을 위한 등록 방법.

청구항 3

제 1 항에 있어서,

상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM) 파라미터를 추출하는 단계를 더 포함하는, 고속 화자 인식을 위한 등록 방법.

청구항 4

제 3 항에 있어서,

상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM) 파라미터를 추출하는 단계는,

상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 단계; 및

상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)에 링크 구조 변환을 적용하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 추출하는 단계를 포함하는, 고속 화자 인식을 위한 등록 방법.

청구항 5

제 4 항에 있어서,

상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 단계는,

상기 적응된 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 단계; 및

상기 로그 스펙트럼 도메인의 상기 적응된 은닉 마코브 모델 (HMM)에 이산 코사인 변환 (DCT)을 적용하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 단계를 포함하는, 고속 화자 인식을 위한 등록 방법.

청구항 6

하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을 임의의 화자의 음성으로부터 추출된 가우시안의 부분 공간들과 비교하는 단계;

상기 비교 결과에 따라 상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 단계를 포함하고,

상기 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)은,

부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터로부터 변환된 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하고, 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 생성된 모델인 것을 특징으로 하는, 고속 화자 인식 방법.

청구항 7

제 6 항에 있어서,

상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 단계는,

상기 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들 중 상기 임의의 화자에 대해 최대 우도를 갖는 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)을 추출하는 단계를 포함하는 것을 특징으로 하는, 고속 화자 인식 방법.

청구항 8

제 1 항 내지 제 7 항 중 어느 한 항의 방법을 컴퓨터 시스템에서 실행하기 위한 프로그램이 기록된, 컴퓨터 시스템이 판독할 수 있는 기록매체.

청구항 9

부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하는 모델 변환부;

상기 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 선형 스펙트럼 도메인에서 화자에 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 도메인 변환부;

선형 스펙트럼 도메인으로 변환된 은닉 마코브 모델 (HMM)을 최대 우도 선형 스펙트럼 변환을 이용하여 화자에 적응하는 화자 적응부; 및

상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)에 링크 구조 변환을 적용하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 추출하는 링크 구조 변환부

를 포함하는, 고속 화자 인식을 위한 등록 장치.

청구항 10

제 9 항에 있어서,

상기 도메인 변환부는,

상기 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 이산 코사인 역변환부; 및

상기 로그 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 시간 도메인 변환부를 포함하는, 고속 화자 인식을 위한 등록 장치.

청구항 11

제 10 항에 있어서,

상기 도메인 변환부는,

상기 적응된 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 로그 스펙트럼 변환부; 및

상기 로그 스펙트럼 도메인의 상기 적응된 은닉 마코브 모델 (HMM)로부터 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 이산 코사인 변환부를 포함하는, 고속 화자 인식을 위한 등록 장치.

청구항 12

하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을 저장하는 메모리부;

임의의 화자의 음성을 입력받는 마이크부;

상기 메모리부에 저장된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을 상기 입력된 음성으로부터 추출된 가우시안의 부분 공간들과 비교하는 우도 연산부;

상기 비교 결과에 따라 상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 최대 우도 선택부를 포함하고,

상기 메모리부에 저장된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)은,

부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터로부터 변환된 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하고, 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 생성된 모델인 것을 특징으로 하는, 고속 화자 인식 장치.

명 세 서

발명의 상세한 설명

기 술 분 야

[0001] 본 발명은 화자인식에 관한 것으로, 특히, 고속 화자 인식을 위한 등록 방법 및 장치, 고속 화자 인식 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 음성 또는 소리의 검출 및 인식 시스템에 대한 다수의 방안들이 제안되어 왔고 어느 정도 성공적으로 구현되었다. 이러한 시스템들은 이론적으로 사용자의 발음(utterance)을 등록된 화자의 발음에 대해 매칭시켜 사용자 신원에 따라 장치 또는 시스템의 자원들에 대한 액세스를 허용 또는 거부하거나, 등록된 화자를 식별하거나 개별화된(customized) 커맨드 라이브러리들을 호출할 수 있다.

[0003] 화자 인식 기술은 발성한 화자가 등록된 화자들 중에 누구인지 또는 등록된 화자가 맞는지 아닌지를 결정하는 기술이다.

[0004] 대규모 자원들을 포함하는 대규모 시스템들은 많은 잠재적인 사용자들을 가질 수 있다. 따라서 등록된 화자의 수가 많아지면 화자를 인식하는데 상당한 양의 저장 및 프로세싱 오버헤드를 필요로 할 수 있다. 화자의 수가 증가하면, 상이한 화자들을 신속하게 구별하도록 설계된 간단하고 빠른 시스템들은 화자 인식 시스템의 성능 포화를 겪게될 것이다.

[0005] 화자 의존적(speaker-dependent) 시스템은 예를 들어, 상이한 화자에 적응된 은닉 마코브 모델(hidden Markov models)(HMM)과 같은 디코딩된 스크립트 모델상에서 발음의 디코딩 및 정렬을 수행하는 시스템을 말한다. 대부분의 화자 의존적 시스템의 성능은 화자의 수가 많은 경우 뿐만 아니라, 빠르고 간단한 시스템에서 화자의 수가 적은 경우에도 포화 및 성능 저하 경향이 발생할 수 있다.

[0006] 일례로, 화자 내지 화자 클래스(class) 식별에 대한 고속 매칭 기술로서 프레임 별 특징 클러스터링 및 분류(frame-by-frame feature clustering and classification)와 같은 텍스트 비의존적 시스템을 고려할 수 있다. 하지만, 허용 가능한 반응 시간 내에 실질적인 양의 프로세싱 오버헤드를 갖고 처리될 수 있는 화자 클래스의 수 및 각 클래스 내의 화자의 수는 제한되어 있다. 다시 말해, 프레임 별 분류기는 각각의 등록 화자에 대하여는 비교적 적은 양의 데이터를 요구하고 제한된 수의 화자에 대하여는 보다 적은 프로세싱 타임을 요구하는 반면, 모델 수의 증가로 인해 화자 모델들 간의 차이가 감소됨에 따라 그 구별 능력은 제한된다. 화자 발음에 관한 정보를 줄이려는 시도는 사용자 수가 많아질 때, 개별 등록 사용자를 식별할 수 있는 시스템 능력을 저하시킬 수 있다. 화자의 수가 상당히 많아지면, 화자 인식 시스템 또는 엔진은 더 이상 몇몇 화자들을 구별할 수 없게 된다. 이러한 상태를 포화라고 한다.

[0007] 반면에, 화자 인식을 제공하기 위하여 개별 화자에 적응된 화자 의존적 모델 기반 디코더를 사용하는 복잡한 시

스템들은 화자 인식을 달성하기 위해서 병렬적으로 또는 순차적으로 모델들을 실행해야 하므로 상당히 느리며, 대량의 메모리 및 프로세서 타임을 요구한다. 또한, 이러한 모델들은 전형적으로 모델을 형성하는데 대량의 데이터를 요구하기 때문에 학습시켜 적응시키기가 어렵다.

[0008] 템플릿 매칭 시스템(template matching system)에서는 저장 조건이 어느 정도 완화되는데, 이러한 시스템은 그 화자 인식 및/또는 인증 기능에 특유한 각 등록 화자의 특정한 발음에 의존함으로 화자 의존적일 뿐만 아니라 텍스트 의존적이기도 하다. 하지만, 이러한 시스템은 그 본질상, 사용자에게 투명(transparent)할 수 없는데, 왜냐하면 이런 시스템은 비교적 긴 등록 및 초기 인식 (예컨대, 로그온) 과정을 요구하고, 인증을 위해 종종 시스템의 사용을 주기적으로 중지시킬 것을 요구하기 때문이다. 더 중요하게는, 이러한 시스템은 각 화자의 노화, 피로, 병, 스트레스, 운율(prosody), 심리 상태 및 기타 상태들에 의해 발생할 수 있는 각 화자의 발음의 변이(화자 내부(intraspeaker) 변이)에 보다 민감하다.

[0009] 보다 구체적으로, 화자 의존적 음성 인식기는 작동 중 등록 단계 동안 각 화자에 대한 모델을 구축한다. 그 다음에, 화자 및 그 발음은 가장 큰 유사도 또는 가장 낮은 에러 레이트를 생성하는 모델에 의해 인식된다. 모든 발음들이 인식되도록 각 모델을 고유한 화자에 적응시키는 데에는 충분한 데이터가 필요하다. 이러한 이유로, 대부분의 화자 의존적 시스템은 또한 텍스트 의존적이며, 템플릿 매칭을 사용하여 각 모델 내에 저장될 데이터의 양을 줄인다. 이와 달리, 예컨대 HMM 또는 이와 유사한 통계적 모델을 사용하는 시스템은 대개 화자의 그룹에 기초한 유사 화자 모델(cohort model)을 생성하여 지나치게 가능성이 없는 화자는 거부한다.

[0010] 일반적으로, 모델 기반 방법이 가장 일반적인데 가우시안 혼합 모델 (gaussian mixture model; GMM) 또는 HMM이 사용된다. 이를 적용한 화자 인식 과정은 크게 두 가지로 구분되는데 첫 번째가 화자 등록 과정이고, 두 번째가 인식 과정이다. 등록 과정은 등록하고자 하는 화자의 목소리를 받아 들여 그 화자의 모델을 만드는 과정이다. 이 과정은 일반적으로 화자 독립 모델을 적응 기법을 이용하여 화자 종속 모델로 만드는 것이다. 적응을 위해서는 최대 우도 선형 회귀 (maximum likelihood linear regression; MLLR)나 MAP (maximum a posterior) 알고리즘이 주로 사용된다. 이 적응 알고리즘들은 등록 화자의 음성이 많으면 많을수록 더 좋은 성능을 보인다. 하지만 환경이 자주 변화하는 모바일 환경에서의 화자 인식을 위해서는 화자 등록을 위해 수 십초에서 수 분 동안의 적은 적응 데이터를 이용해서 화자의 음성을 빠르게 적응하는 방법이 필요하다. 인식 과정은 적응된 모델들 각각에서의 인식하고자 하는 화자 음성의 우도 (likelihood)를 계산하여 가장 높은 우도 값을 보이는 모델의 화자를 선택하는 것이다. 하지만 화자 인식기에 등록된 화자 종속 모델이 너무 많은 경우 테스트 음성이 들어 왔을 때 비교 검사해야 하는 모델이 너무 많기 때문에 계산량이 많아지게 된다. 특히 모바일 기기와 같이 연산 능력이 떨어지는 기기에서는 이러한 문제가 굉장히 중요한 이슈로 부각된다.

[0011] 기존의 적응 알고리즘들은 높은 성능을 보이기 위해서 등록하고자 하는 화자의 목소리를 몇 분 또는 몇 십분 정도의 많은 양을 필요로 한다. 또한, 기존의 화자 적응 방법은 새로운 잡음환경에서 효과적이지 못하다. 한편, 화자 인식에 있어서도 등록된 화자가 많은 경우 결정 과정에서 비교되어야 할 모델의 수가 늘어나기 때문에 계산량이 너무 많다는 문제점이 있다.

발명의 내용

해결 하고자하는 과제

[0012] 따라서, 본 발명이 이루고자 하는 첫 번째 기술적 과제는 아주 적은 양의 적응 데이터로도 강인한 적응이 가능하고, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있는 고속 화자 인식을 위한 등록 방법을 제공하는 데 있다.

[0013] 본 발명이 이루고자 하는 두 번째 기술적 과제는 등록 화자의 수가 증가하더라도 인식 과정에서의 계산량이 크게 증가하지 않으며, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있는 고속 화자 인식 방법을 제공하는 데 있다.

[0014] 본 발명이 이루고자 하는 세 번째 기술적 과제는 아주 적은 양의 적응 데이터로도 강인한 적응이 가능하고, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있는 고속 화자 인식을 위한 등록 장치를 제공하는 데 있다.

[0015] 본 발명이 이루고자 하는 네 번째 기술적 과제는 등록 화자의 수가 증가하더라도 인식 과정에서의 계산량이 크게 증가하지 않으며, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있는 고속 화자 인식 장치를 제공하는 데 있다.

과제 해결수단

[0016] 상기의 첫 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 방법은 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하는 단계; 상기 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 단계; 및 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하는 단계를 포함한다.

[0017] 바람직하게는, 상기 선형 스펙트럼 도메인으로 변환하는 단계에서, 이산 코사인 역변환 (IDCT)을 이용하여 상기 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하고, 상기 로그 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환할 수 있다.

[0018] 바람직하게는, 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 방법은 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM) 파라미터를 추출하는 단계를 더 포함할 수 있다. 이 경우, 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM) 파라미터를 추출하는 단계에서, 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하고, 상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)에 링크 구조 변환을 적용하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 추출할 수 있다. 보다 구체적으로, 상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 단계는 상기 적응된 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 단계와 상기 로그 스펙트럼 도메인의 상기 적응된 은닉 마코브 모델 (HMM)에 이산 코사인 변환 (DCT)을 적용하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 단계로 구성될 수 있다.

[0019] 상기의 두 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 고속 화자 인식 방법은 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을 임의의 화자의 음성으로부터 추출된 가우시안의 부분 공간들과 비교하는 단계; 상기 비교 결과에 따라 상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 단계를 포함한다. 여기서, 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)은, 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터로부터 변환된 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하고, 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 생성된 모델이다.

[0020] 바람직하게는, 상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 단계에서, 상기 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들 중 상기 임의의 화자에 대해 최대 우도를 갖는 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)을 추출할 수 있다.

[0021] 상기의 세 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 장치는 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 전체 공간에 대한 은닉 마코브 모델 (HMM)로 변환하는 모델 변환부; 상기 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 선형 스펙트럼 도메인에서 화자에 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 도메인 변환부; 선형 스펙트럼 도메인으로 변환된 은닉 마코브 모델 (HMM)을 최대 우도 선형 스펙트럼 변환을 이용하여 화자에 적응하는 화자 적응부; 및 상기 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)에 링크 구조 변환을 적용하여 상기 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터를 추출하는 링크 구조 변환부를 포함한다.

[0022] 바람직하게는, 상기 도메인 변환부는 상기 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 이산 코사인 역변환부; 및 상기 로그 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하는 시간 도메인 변환부를 포함할 수 있다.

[0023] 바람직하게는, 상기 도메인 변환부는 상기 적응된 은닉 마코브 모델 (HMM)을 로그 스펙트럼 도메인으로 변환하는 로그 스펙트럼 변환부; 및 상기 로그 스펙트럼 도메인의 상기 적응된 은닉 마코브 모델 (HMM)로부터 전체 공간에 대해 적응된 은닉 마코브 모델 (HMM)을 구하는 이산 코사인 변환부를 포함할 수 있다.

[0024] 상기의 네 번째 기술적 과제를 이루기 위하여, 본 발명의 일 실시 예에 따른 고속 화자 인식 장치는 하나 이상의 화자에 적응된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을 저장하는 메모리부; 임의의 화자의 음성을 입력받는 마이크부; 상기 메모리부에 저장된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)들을

상기 입력된 음성으로부터 추출된 가우시안의 부분 공간들과 비교하는 우도 연산부; 상기 비교 결과에 따라 상기 임의의 화자가 등록된 화자들 중 누구에 해당하는지를 출력하는 최대 우도 선택부를 포함한다. 여기서, 메모리부에 저장된 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)은 부분공간 분배 클러스터링 은닉 마코브 모델 (SDCHMM)의 파라미터로부터 변환된 은닉 마코브 모델 (HMM)을 선형 스펙트럼 도메인으로 변환하고, 최대 우도 선형 스펙트럼 변환을 이용하여 상기 선형 스펙트럼 도메인의 은닉 마코브 모델 (HMM)을 화자에 적응하고, 상기 적응된 은닉 마코브 모델 (HMM)을 켑스트럼 도메인으로 변환하여 생성된 모델이다.

효 과

[0025] 본 방법에 의하면, 휴대폰과 같은 소형 이동 단말기에서 화자 모델을 적은 적응 데이터로도 빠르게 적응시킬 수 있을 뿐만 아니라, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있으며, 인식 과정에서의 계산량을 대폭 줄일 수 있어서 서버와 같은 다른 기기의 도움을 받을 필요 없이 소형 단말기 스스로 화자 인식을 처리 할 수 있는 효과가 있다.

발명의 실시를 위한 구체적인 내용

[0026] 이하에서는 도면을 참조하여 본 발명의 바람직한 실시 예를 설명하기로 한다. 그러나, 다음에 예시하는 본 발명의 실시 예는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 다음에 상술하는 실시 예에 한정되는 것은 아니다.

[0027] 본 발명에서는 모바일 환경에 사용되는 화자 인식기를 위해 부분공간 분배 클러스터링 은닉 마코브 모델 (Subspace Distribution Clustering Hidden Markov Model; SDCHMM)을 사용한다. SDCHMM은 전체 다변수 가우시안 (multivariate Gaussian)들을 부분공간 (subspace)이라는 특징 공간 (feature space)들로 나눈 뒤, 각 부분공간에 속한 분포들을 양자화하여 양자화된 부분공간 분포 프로토타입 즉, 코드워드 (codeword)들을 조합하여 원래의 전체 공간 (full space) 분포를 표현한다.

[0028] 다시 말하면 원래의 전체 공간 분포와 가장 유사한 부분공간 프로토타입들의 조합을 찾아서 그것들을 링크한다. 이와 같이 SDCHMM은 적은 수의 부분공간 코드워드 (subspace codeword)를 이용하여 원래의 전체 공간 분포를 표현한다.

[0029] 도 1은 13개의 부분공간으로 구성된 SDCHMM의 예시도이다.

[0030] 13개의 부분공간으로 구성된 SDCHMM은 13개의 부분공간 분포 프로토타입에서 가장 유사한 부분공간 분포들의 결합으로 하나의 전체 공간 분포를 표현한다. 예를 들어, 도 1에 도시된 두 그룹의 선분들 중 좌측의 분포는 Sp13의 첫 번째 컬럼을 포함하는 13개의 연결선인데, 이는 부분공간 분포의 결합으로 표현된다. 또한, 도 1에 도시된 두 그룹의 선분들 중 우측의 분포는 Sp13의 세 번째 컬럼을 포함하는 13개의 연결선인데, 이는 또 다른 부분공간 분포의 결합으로 표현된다.

[0031] SDCHMM은 HMM의 모든 분포들을 적은 수의 부분공간 분포 프로토타입으로 표현할 수 있기 때문에 메모리와 계산량을 줄일 수 있어 모바일 디바이스에 적합한 모델이다.

[0032] SDCHMM을 적응 시키는 방법에는 프로토타입을 적응 변화시키는 방법이 있고 프로토타입을 변화 시키지 않고 링크 구조만 적응 변화시키는 방법이 있다.

[0033] 도 2는 SDCHMM을 위한 링크 구조 적응 방법의 예시도이다.

[0034] SDCHMM을 링크 구조 적응 방법을 이용하여 적응시키면 링크의 변화만으로도 각각 다른 화자의 모델을 만들 수 있게 되고, 이는 적응 처리에 강력할 수 있다. 즉, 기존 HMM을 적응시키는 경우처럼 수많은 서로 다른 다변수 가우시안 분포들을 적응시키는 것 보다 적은 수의 프로토타입을 링크하는 구조를 적응 시키는 것이 더 적은 데이터로도 강인하게 적응시킬 수 있다. 또한 SDCHMM을 링크 구조 적응 방법을 이용하여 적응시키면 인식 과정에서 우도를 계산할 때도 단 한번만 프로토타입들의 우도를 계산해 두면 모든 화자 모델들은 그 계산된 우도들을 모두 공유하여 덧셈만 하면 되므로 계산량을 줄일 수 있다. 이는 각각의 화자 모델들이 하나의 프로토타입들을 공유하면서 링크하고 있기 때문이다.

[0035] 음성 인식 및 화자 인식에서는 모델을 만들어서 입력된 음성과 모델과의 유사도가 얼마나 높은지를 측정하는 방법을 사용한다. HMM 기반의 모델을 사용하는 경우 유사도를 우도라고 부른다. 우도는 수학식 1과 같이 표현된다.

수학식 1

[0036] $$P(o|\lambda)$$

[0037] 이는 모델 λ에서 데이터 o가 발생할 확률을 의미한다.

[0038] 모델을 만들 때에는 학습 데이터(음성)의 우도가 최대가 되도록 한다. 이때 모델을 만드는 것을 학습이라고 한다. 우도가 최대가 되도록 학습시킬 때, 최대 우도 (ML)라는 개념이 사용된다. 학습을 시킬 때에는 충분히 많은 데이터를 이용하여야 한다. 그렇지 않으면 예외 상황 또는 학습 때 발견되지 않았던 데이터가 테스트 상황에 발생하여 정확한 인식을 방해하기 때문이다.

[0039] 잘 학습된 모델을 적은 데이터를 이용하여 변환시키는 것을 적응이라고 한다. 적응도 학습과 마찬가지로 적응 데이터의 우도가 최대가 되도록 모델을 변환시킨다. MLLR이나 ML-LST 모두 ML 기법을 이용하여 적응시킨다.

[0040] 하지만 우도를 직접적으로 최대화 시킬 수 있는 방법은 존재하지 않는다. 수학식 2는 우도를 최대화하는 데 사용되는 Q함수(Q-function)를 나타낸다.

수학식 2

[0041] $$Q(\lambda',\lambda)=P(o|\lambda')\log P(o|\lambda)$$

[0042] 수학식 2의 λ'와 λ는 각각 이전 모델과 재추정된 모델을 의미하며, 수학식 2가 최대가 되면 수학식 3이 만족된다.

수학식 3

[0043] $$P(o|\lambda)>P(o|\lambda')$$

[0044] λ'와 λ를 각각 적응되기 전의 모델과 적응 후의 모델이라고 가정하고 수학식 2가 최대가 되도록 하면 적응 전의 모델보다 적응 후의 모델에서의 적응 데이터 o의 우도가 크게 된다. 이와 같이 수학식 3을 만족시키는 과정을 반복하면 점점 우도가 커지게 되고 최대 우도에 근접하게 된다.

[0045] 도 3은 시간 도메인 (Time domain)에서 가산 잡음과 채널 잡음의 결합 과정을 도시한 것이다.

[0046] 사람이 발성한 깨끗한 음성은 주변 잡음들로 인해 발생하는 가산 잡음(Additive noise)과 마이크 등의 입력 장비간의 채널 차이 등으로 발생하는 채널 잡음(Channel noise)에 의해 오염되어 입력되게 된다.

[0047] 화자 인식이 일반적으로 켑스트럼 도메인 (cepstral domain) 에서 동작하기 때문에 대부분의 적응 알고리즘들이 켑스트럼 도메인에서 수행되도록 설계된다. 하지만 켑스트럼 도메인에서는 가산 잡음이나 채널 잡음을 직접적으로 표현하기가 어렵다.

[0048] 최대 우도 선형 스펙트럼 변환 (Maximum likelihood linear spectral transform; ML-LST)을 이용하면 시간 도메인에서 발생하는 잡음들을 명확하게 표현할 수 있는 스펙트럼 도메인 (spectral domain)에서 적응할 수 있다.

[0049] 도 4는 ML-LST를 위한 음성 데이터의 도메인 변환 과정을 도시한 것이다.

[0050] 화자 인식은 켑스트럼 도메인에서 처리되고 ML-LST를 이용한 적응은 스펙트럼 도메인에서 적용되기 때문에 적응 전에 도메인 변환 과정이 필요하다. 모델을 선형 스펙트럼 도메인으로 변환하기 위해서 이산 코사인 역변환 (Inverse Discrete Cosine Transformation; IDCT)과 지수 연산(exponential operation)이 순차적으로 적용된다. 선형 스펙트럼 도메인 (Linear spectral domain)에서 적응이 된 데이터는 인식을 하기 위해 다시 켑스트럼 도메인으로 되돌려 진다.

[0051] 도 5는 본 발명에 따른 화자 등록 과정을 도시한 것이다.

[0052] 본 발명에서는 ML-LST와 링크 구조 적응을 이용하여 화자 독립 모델을 화자 종속 모델로 변환한다. ML-LST에서는 새로운 모델을 적응되기 전의 모델을 선형 스펙트럼 도메인으로 변환하고 잡음을 적용시킨 후에 다시 켑스트

럼 도메인으로 되돌아온다. ML-LST는 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있는 스펙트럼 도메인에서 적응하기 때문에 기존의 MLLR이나 MAP 보다 새로운 잡음환경에서 더욱 효과적이다.

[0053] 켑스트럼 도메인에서 잡음 평균 벡터 μ를 식으로 나타낸 것이 수학식 4이다.

수학식 4

[0054] $$\boldsymbol{\mu} = \mathbf{C} \cdot \log(\mathbf{A}\hat{\boldsymbol{\mu}} + \mathbf{b})$$

[0055] 여기서, C는 N*M DCT (Discrete Cosine Transformation) 행렬을 의미한다. M은 필터 뱅크(filterbank)의 수, N은 μ의 차원 (dimensionality)을 의미한다. A, b는 잡음을 나타낸다.

[0056] 도 6은 구체적인 도메인 변화 과정을 도시한 것이다.

[0057] SDCHMM의 파라미터를 전체 공간에 대한 HMM으로 변환(610)한 후, 켑스트럼 도메인에서 IDCT (620)와 지수 연산(630)을 거치면 선형 스펙트럼 도메인으로 변환된다. ML-LST 적응(640)과 잡음의 적용은 선형 스펙트럼 도메인에서 이루어진다.

[0058] 수학식 4가 적응 이후의 모델이기 때문에 이 모델의 우도가 최대가 되도록 한다. 수학식 4의 우도가 최대가 되도록 하기 위해서 수학식 4를 수학식 2의 Q함수에 적용하고, Q함수가 최대가 되도록 하는 조건을 구한다. 이 과정을 반복하면 선형 스펙트럼 도메인에서 우도가 최대가 되는 ML-LST 적응이 이루어진다.

[0059] 이후, 로그 연산(650)과 DCT(660)를 거치면 켑스트럼 도메인으로 변환된다. 수학식 4에서 잡음 A와 b가 선형 스펙트럼 도메인에서 적용된 후에 log와 DCT를 거쳐 다시 켑스트럼 도메인으로 변환되었음을을 알 수 있다.

[0060] 마지막으로, 켑스트럼 도메인의 HMM을 링크 구조 변환하여 적응된 SDCHMM 파라미터를 생성한다.

[0061] 도 7은 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 장치의 블록도이다.

[0062] 모델 변환부(710)는 SDCHMM의 파라미터를 전체 공간에 대한 HMM으로 변환한다.

[0063] 도메인 변환부(720)는 HMM을 선형 스펙트럼 도메인으로 변환한다.

[0064] 또한, 도메인 변환부(720)는 화자 적응부(730)가 화자에 적응시킨 선형 스펙트럼 도메인의 HMM을 켑스트럼 도메인으로 변환하여 전체 공간에 대해 적응된 HMM을 구한다.

[0065] 화자 적응부(730)는 등록하고자 하는 화자의 음성에 HMM을 적응시킨다. 즉, 화자 적응부(730)는 선형 스펙트럼 도메인으로 변환된 HMM을 ML-LST를 이용하여 화자에 적응한다.

[0066] 링크 구조 변환부(740)는 도메인 변환부(720)에서 생성된 스켑트럼 도메인의 HMM 즉, 전체 공간에 대해 적응된 HMM에 링크 구조 변환을 적용하여 화자에 적응된 SDCHMM의 파라미터를 추출한다. 이러한 파라미터들은 화자 종속 모델을 구성한다.

[0067] 도 8은 본 발명에 따른 화자 인식 과정을 도시한 것이다.

[0068] 본 발명에서는 위와 같이 구성된 화자 종속 모델과 입력되는 음성의 가우시안 부분 공간 사이의 우도를 연산하고, 이 중에서 최대 우도를 갖는 모델을 선택하는 방식으로 화자를 인식한다.

[0069] 도 9는 본 발명의 일 실시 예에 따른 고속 화자 인식 장치의 블록도이다.

[0070] 메모리부(910)는 하나 이상의 화자에 적응된 SDCHMM들을 저장한다. 메모리부(910)에 저장된 SDCHMM은 상술한 바와 같이, SDCHMM의 파라미터로부터 변환된 HMM을 선형 스펙트럼 도메인으로 변환하고, 최대 우도 선형 스펙트럼 변환을 이용하여 선형 스펙트럼 도메인의 HMM을 화자에 적응하고, 적응된 HMM을 켑스트럼 도메인으로 변환하여 생성된 모델이다. 메모리부(910)는 휘발성 메모리 소자나 비휘발성 메모리 소자 등 물리적인 메모리 블록으로 구성된다.

[0071] 마이크부(920)는 입력되는 화자의 음성을 전기적인 오디오 신호로 변환한다.

[0072] 우도 연산부(930)는 메모리부(910)에 저장된 SDCHMM들을 마이크부(920)에 입력된 음성으로부터 추출된 가우시안의 부분 공간들과 비교한다.

[0073] 최대 우도 선택부(940)는 우도 연산부(930)의 비교 결과에 따라 음성을 발화한 화자가 미리 등록된 화자들 중

누구에 해당하는지를 출력한다.

[0074] 본 발명은 소프트웨어를 통해 실행될 수 있다. 바람직하게는, 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 방법이나 고속 화자 인식 방법을 컴퓨터에서 실행시키기 위한 프로그램을 컴퓨터로 읽을 수 있는 기록매체에 기록하여 제공할 수 있다. 소프트웨어로 실행될 때, 본 발명의 구성 수단들은 필요한 작업을 실행하는 코드 세그먼트들이다. 프로그램 또는 코드 세그먼트들은 프로세서 판독 가능 매체에 저장되거나 전송 매체 또는 통신망에서 반송파와 결합된 컴퓨터 데이터 신호에 의하여 전송될 수 있다.

[0075] 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 장치의 예로는 ROM, RAM, CD-ROM, DVD±ROM, DVD-RAM, 자기 테이프, 플로피 디스크, 하드 디스크(hard disk), 광데이터 저장장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 장치에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[0076] 본 발명은 도면에 도시된 일 실시 예를 참고로 하여 설명하였으나 이는 예시적인 것에 불과하며 당해 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 실시 예의 변형이 가능하다는 점을 이해할 것이다. 그리고, 이와 같은 변형은 본 발명의 기술적 보호범위 내에 있다고 보아야 한다. 따라서, 본 발명의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해서 정해져야 할 것이다.

산업이용 가능성

[0077] 본 발명은 화자 모델을 적은 적응 데이터로도 빠르게 적응시킬 수 있을 뿐만 아니라, 가산 잡음이나 채널 잡음을 효과적으로 처리할 수 있으며, 인식 과정에서의 계산량을 대폭 줄일 수 있는 화자 등록 및 인식 기술에 관한 것으로, 휴대폰과 같은 소형 이동 단말기나 자원의 제약이 많은 화자 인식 장치 등 화자를 등록하기 위해 충분한 데이터의 수집이 어렵고 적은 계산으로도 빠르게 인식을 해야 하는 모든 환경에 적용될 수 있다.

도면의 간단한 설명

[0078] 도 1은 13개의 부분공간으로 구성된 SDCHMM의 예시도이다.

[0079] 도 2는 SDCHMM을 위한 링크 구조 적응 방법의 예시도이다.

[0080] 도 3은 시간 도메인에서 가산 잡음과 채널 잡음의 결합 과정을 도시한 것이다.

[0081] 도 4는 ML-LST를 위한 음성 데이터의 도메인 변환 과정을 도시한 것이다.

[0082] 도 5는 본 발명에 따른 화자 등록 과정을 도시한 것이다.

[0083] 도 6은 구체적인 도메인 변화 과정을 도시한 것이다.

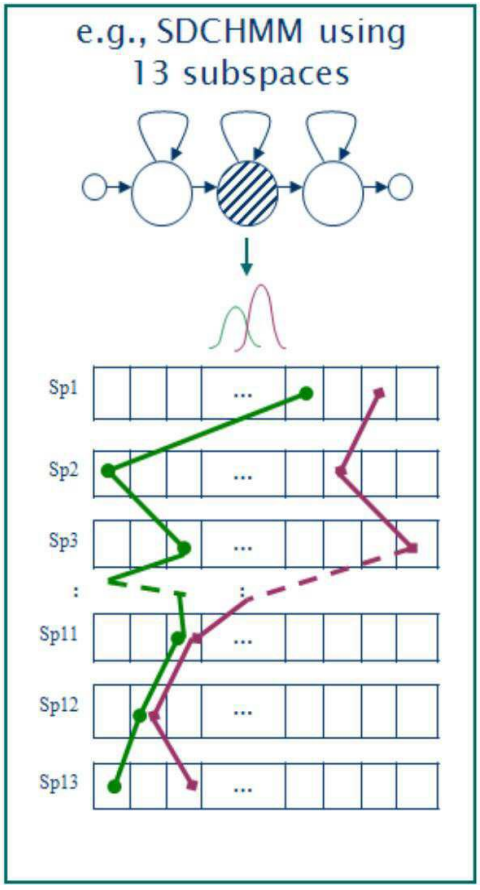
[0084] 도 7은 본 발명의 일 실시 예에 따른 고속 화자 인식을 위한 등록 장치의 블록도이다.

[0085] 도 8은 본 발명에 따른 화자 인식 과정을 도시한 것이다.

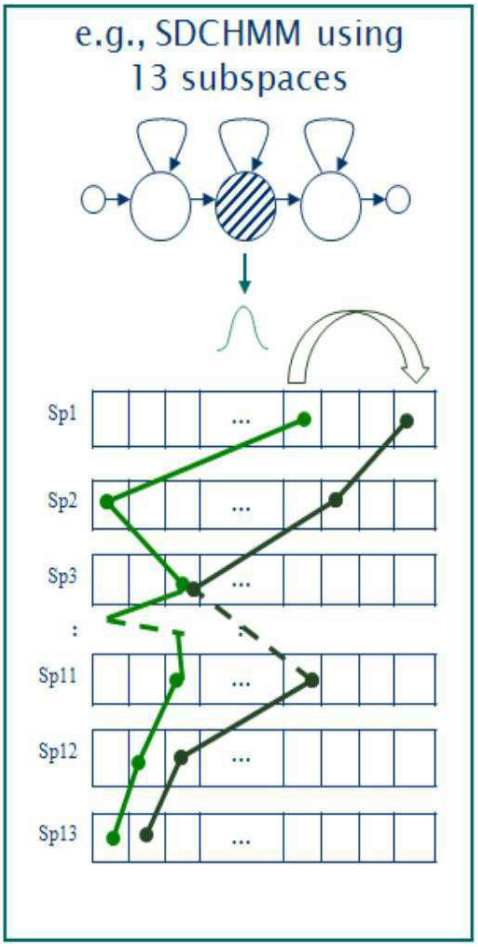
[0086] 도 9는 본 발명의 일 실시 예에 따른 고속 화자 인식 장치의 블록도이다.

도면

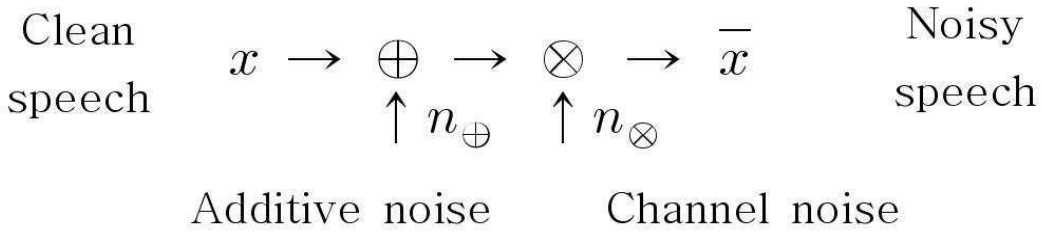
도면1



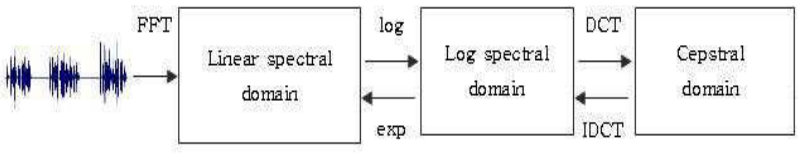
도면2


e.g., SDCHMM using
13 subspaces
Sp1
Sp2
Sp3
Sp11
Sp12
Sp13

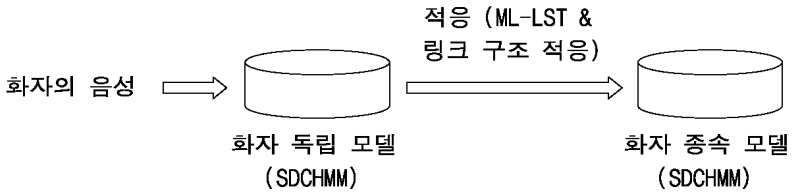
도면3


Clean
speech
x
$\oplus$
$n_{\oplus}$
$\otimes$
$n_{\otimes}$
$\overline{x}$
Noisy
speech
Additive noise
Channel noise

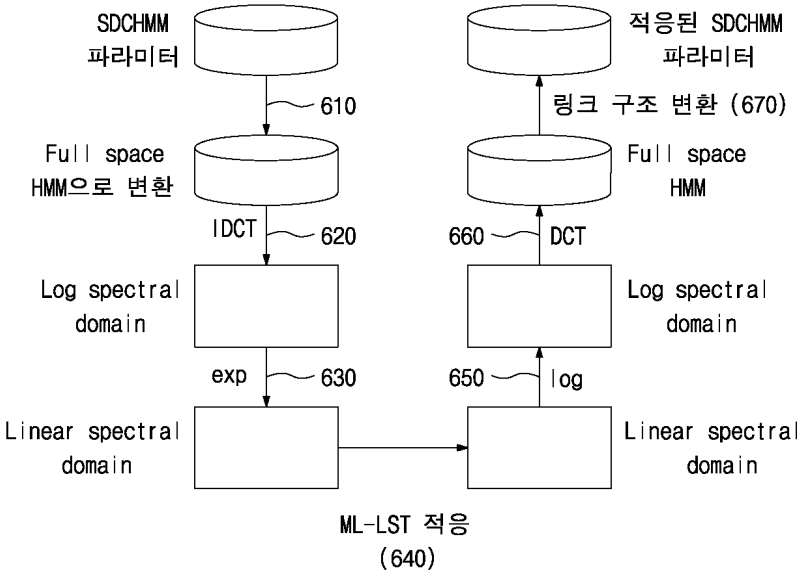
도면4


FFT
Linear spectral
domain
log
exp
Log spectral
domain
DCT
IDCT
Cepstral
domain

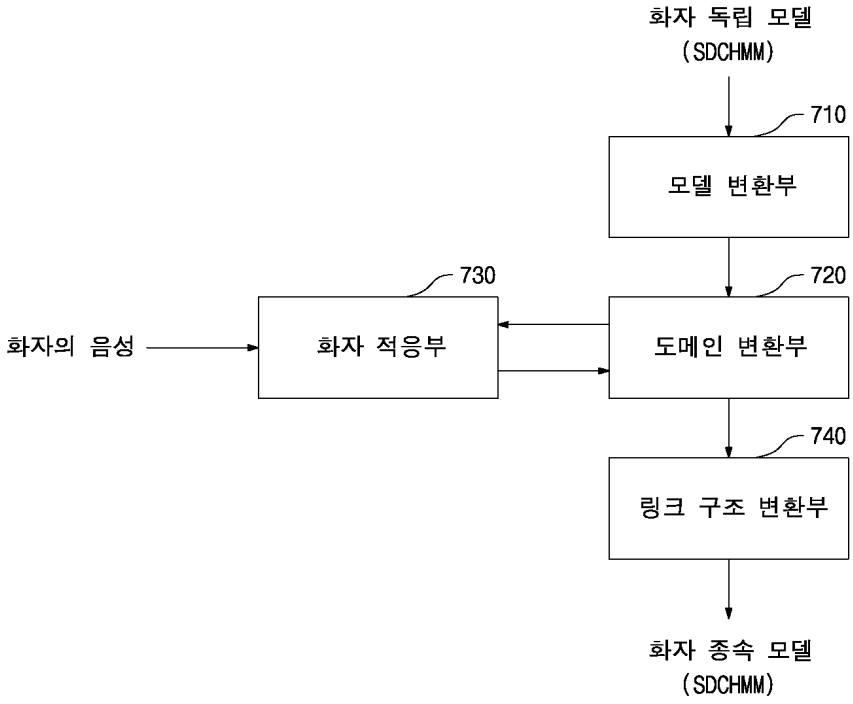
도면5


화자의 음성
화자 독립 모델
(SDCHMM)
적응 (ML-LST &
링크 구조 적응)
화자 종속 모델
(SDCHMM)

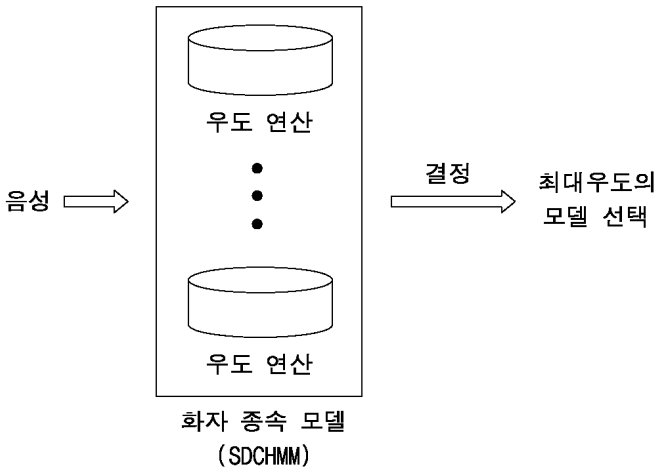
도면6


SDCHMM
파라미터
610
Full space
HMM으로 변환
IDCT
620
Log spectral
domain
exp
630
Linear spectral
domain
적응된 SDCHMM
파라미터
링크 구조 변환 (670)
Full space
HMM
660
DCT
Log spectral
domain
650
log
Linear spectral
domain
ML-LST 적응
(640)

도면7


화자 독립 모델
(SDCHMM)
710
모델 변환부
730
720
화자의 음성
화자 적응부
도메인 변환부
740
링크 구조 변환부
화자 종속 모델
(SDCHMM)

도면8


우도 연산
음성
결정
최대우도의
모델 선택
우도 연산
화자 종속 모델
(SDCHMM)

도면9