



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2014년08월05일
(11) 등록번호 10-1427047
(24) 등록일자 2014년07월30일

(51) 국제특허분류(Int. Cl.)
G10L 17/14 (2013.01) G10L 15/02 (2006.01)
(21) 출원번호 10-2013-0073808
(22) 출원일자 2013년06월26일
심사청구일자 2013년06월26일
(56) 선행기술조사문헌
KR1020110025356 A

(73) 특허권자
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
김무영
서울 강남구 선릉로 221, 210동 402호 (도곡동, 도곡텍슬아파트)
백승렬
서울 광진구 광나루로17길 24-7, (군자동)
(74) 대리인
특허법인태백

전체 청구항 수 : 총 8 항

심사관 : 이수철

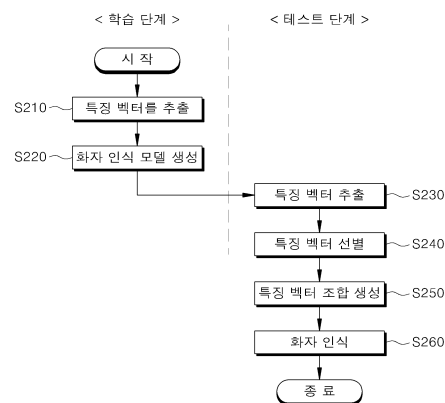
(54) 발명의 명칭 화자 인식 장치 및 그 방법

(57) 요약

본 발명은 화자 인식 장치 및 그 방법에 관한 것이다. 본 발명에 따른 화자 인식 장치는, 화자별 입력 문장에 대한 특징 벡터를 추출하여 화자 인식 모델을 생성하는 화자 모델 생성부; 상기 음성 신호에 대한 특징 벡터를 추출하는 특징 벡터 추출부; 상기 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 상기 잡음 신호를 마스크(mask)하여 상기 클린 신호에 대한 특징 벡터를 선별하는 특징 벡터 선별부; 상기 클린 신호에 대한 특징 벡터를 이용하여 복수의 특징 벡터 조합을 생성하는 특징 벡터 조합부; 상기 클린 신호에 대한 특징 벡터와 상기 특징 벡터 조합 중에서 적어도 하나와 상기 화자 인식 모델을 비교하여 화자를 인식하는 화자 인식부; 를 포함한다.

본 발명에 따르면, 잡음에 의한 영향을 최소화함으로써 화자 인식률을 향상시킬 수 있다. 또한, 화자 인식에 사용되는 특징 벡터를 구함에 있어 잡음의 영향을 덜 받음으로써 계산의 복잡성을 피해 연산량을 줄일 수 있다.

대표도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호 1345192901

부처명 교육과학기술부

연구사업명 일반연구자지원사업

연구과제명 Interactive Music Application을 위한 객체기반 오디오 콘텐츠 코딩, 신호분리 및 잡음제거 기법에 관한 기초 기술 연구

기 여 율 1/1

주관기관 세종대학교 산학협력단

연구기간 2012.05.01 ~ 2015.04.30

특허청구의 범위

청구항 1

잡음을 포함하는 음성 신호로부터 화자를 인식하는 장치에 있어서,
 화자별 입력 문장에 대한 특징 벡터를 추출하여 화자 인식 모델을 생성하는 화자 모델 생성부;
 상기 음성 신호에 대한 특징 벡터를 추출하는 특징 벡터 추출부;
 상기 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 상기 잡음 신호를 마스크(mask)하여 상기 클린 신호에 대한 특징 벡터를 선별하는 특징 벡터 선별부;
 상기 클린 신호에 대한 특징 벡터를 이용하여 복수의 특징 벡터 조합을 생성하는 특징 벡터 조합부;
 상기 클린 신호에 대한 특징 벡터와 상기 특징 벡터 조합 중에서 적어도 하나와 상기 화자 인식 모델을 비교하여 화자를 인식하는 화자 인식부; 를 포함하되,
 상기 특징 벡터 조합부는,
 상기 클린 신호에 대한 특징 벡터를 두 개 이상 조합하는 화자 인식 장치.

청구항 2

제1항에 있어서,
 상기 화자 모델 생성부는,
 가우시안 혼합 모델(Gaussian Mixture Model: GMM)을 이용하여 상기 화자 인식 모델을 생성하는 화자 인식 장치.

청구항 3

제1항에 있어서,
 상기 특징벡터 선별부는,
 다음의 수학적식과 같이 잡음대비 신호비(Signal-to-noise Ratio: SNR)를 기반으로 상기 클린 신호와 잡음 신호를 구별하는 화자 인식 장치:

$$SNR(b) = 10 \log_{10} \frac{\sum_{f \in b\text{-th band}} P_{EC}(f)}{\sum_{f \in b\text{-th band}} P_{EN}(f)}$$

상기 P_{EC} 와 P_{EN} 은 각각 예측된 클린 신호와 예측된 잡음 신호의 파워스펙트럼을 나타낸다.

청구항 4

삭제

청구항 5

제1항에 있어서,
 상기 화자 인식부는,
 다음의 수학적식에 의한 특징 벡터와 상기 화자 인식 모델을 비교하여 화자를 인식하는 화자인식 장치:

$$\hat{s} = \arg \max_{s \in S} \sum_{t=1}^T \log p(Y_t | \lambda_s)$$

상기 수학식에서,

$$p(\lambda_s | Y_{sub}) = \frac{p(Y_{sub} | \lambda_s) p(\lambda_s)}{\sum_{s \in S} p(Y_{sub} | \lambda_s) p(\lambda_s)}$$

$$p(Y_{sub} | \lambda_s) = \sum_{m=1}^M w_{s,m} \prod_{k \in Y_{sub}} N_m(y_k | \lambda_{s,m})$$

이며, λ_s 는 s번째 화자의 GMM(mean, variance, weighting)을 나타내며, M개의 mixture를 갖는다. 또한, 가우스 분포 N_m 에서 s는 화자의 인덱스, m은 화자 모델의 인덱스, k는 특징 벡터의 인덱스를 각각 나타낸다. 또한, $w_{s,m}$ 은 s번째 화자, m번째 mixture의 weighting 값을 나타낸다.

청구항 6

잡음을 포함하는 음성 신호로부터 화자를 인식하는 방법에 있어서,

화자별 입력 문장에 대한 특징 벡터를 추출하여 화자 인식 모델을 생성하는 단계;

상기 음성 신호에 대한 특징 벡터를 추출하는 단계;

상기 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 상기 잡음 신호를 마스크(mask)하여 상기 클린 신호에 대한 특징 벡터를 선별하는 단계;

상기 클린 신호에 대한 특징 벡터를 이용하여 복수의 특징 벡터 조합을 생성하는 단계;

상기 클린 신호에 대한 특징 벡터와 상기 특징 벡터 조합 중에서 적어도 하나와 상기 화자 인식 모델을 비교하여 화자를 인식하는 단계; 를 포함하되,

특징 벡터 조합을 생성하는 단계에서,

상기 클린 신호에 대한 특징 벡터를 두 개 이상 조합하는 화자 인식 방법.

청구항 7

제6항에 있어서,

상기 화자 인식 모델을 생성하는 단계에서,

가우시안 혼합 모델(Gaussian Mixture Model: GMM)을 이용하여 상기 화자 인식 모델을 생성하는 화자 인식 방법.

청구항 8

제6항에 있어서,

상기 특징 벡터를 선별하는 단계에서,

다음의 수학식과 같이 잡음대비 신호비(Signal-to-noise Ratio: SNR)를 기반으로 상기 클린 신호와 잡음 신호를 구별하는 화자 인식 방법:

$$SNR(b) = 10 \log_{10} \frac{\sum_{f \in b\text{-th band}} P_{EC}(f)}{\sum_{f \in b\text{-th band}} P_{EN}(f)}$$

상기 P_{EC} 와 P_{EN} 은 각각 예측된 클린 신호와 예측된 잡음 신호의 파워스펙트럼을 나타낸다.

청구항 9

삭제

청구항 10

제6항에 있어서,

상기 화자를 인식하는 단계에서,

다음의 수학식에 의한 특징 벡터와 상기 화자 인식 모델을 비교하여 화자를 인식하는 화자 인식 방법:

$$\hat{s} = \arg \max_{s \in S} \sum_{t=1}^T \log p(Y_t | \lambda_s)$$

상기 수학식에서,

$$p(\lambda_s | Y_{sub}) = \frac{p(Y_{sub} | \lambda_s) p(\lambda_s)}{\sum_{s \in S} p(Y_{sub} | \lambda_s) p(\lambda_s)}$$

$$p(Y_{sub} | \lambda_s) = \sum_{m=1}^M w_{s,m} \prod_{k \in Y_{sub}} N_m(y_k | \lambda_{s,m})$$

이며, λ_s 는 s번째 화자의 GMM(mean, variance, weighting)을 나타내며, M개의 mixture를 갖는다. 또한, 가우스 분포 N_m 에서 s는 화자의 인덱스, m은 화자 모델의 인덱스, k는 특징 벡터의 인덱스를 각각 나타낸다. 또한, $w_{s,m}$ 은 s번째 화자, m번째 mixture의 weighting 값을 나타낸다.

명세서

기술분야

[0001] 본 발명은 화자 인식 장치 및 그 방법에 관한 것으로서, 더욱 상세하게는 잡음을 처리하기 위해 잡음대비 신호비(Signal-to-noise Ratio: SNR) 기반의 마스크 기법을 이용하는 화자 인식 장치 및 방법에 관한 것이다.

배경기술

[0002] 음성은 인간이 사용하는 가장 보편적이고 편리한 정보전달의 수단이라 할 수 있다. 음성에 의해 표현되는 말은 인간과 인간사이의 의사소통의 수단으로서뿐만 아니라 인간의 음성을 이용하여 기계 및 사용 장치를 동작시키는 수단으로서도 중요한 역할을 한다. 최근 컴퓨터 성능의 발전, 다양한 미디어의 개발, 신호 및 정보처리 기술의 발전으로 인해 Man Machine Interface의 필요성이 크게 대두되었으며, 음성 인식 기술은 이러한 인터페이스가 편리하고 자연스럽도록 개선시켜 줄 핵심기술 중의 하나이다.

[0003] 화자 인식 시스템을 이용하여 화자를 인식하는 경우, 사전에 화자의 음성, 즉 입력문장에 기초하여 화자 인식 모델을 생성한 후에, 입력되는 화자의 음성과 생성된 화자 인식 모델을 서로 비교하여 화자를 인식하게 된다. 그런데, 입력문장에 기초하여 화자 인식 모델을 생성하는 때와는 다르게 실제 인식 대상이 되는 화자의 음성이 입력되는 시점에서 화자의 음성 이외의 잡음이 입력되는 경우가 많다. 화자의 음성 이외의 잡음은 화자 인식에 장애 요소가 되며 화자 인식률을 떨어뜨리기도 하므로, 관련 기술 분야에서 잡음을 제거하려는 많은 노력이 기울여지고 있다.

[0004] 기존의 가우시안 혼합 모델(Gaussian Mixture Model: GMM) 기반의 화자 인식 시스템은 잡음이 없는 환경에서는 높은 인식률을 보여주지만, Non-stationary 잡음 환경에서는 그 성능이 매우 저하된다. 이 문제점을 극복하기 위해 여러 가지 기술들이 제안되었다. 그 중에서 일반적으로 사용되는 기법은 mask 기법이며, 잡음대비 신호비(Signal-to-noise Ratio: SNR) 기반의 mask가 일반적으로 사용되지만 성능 향상이 미비한 실정이다.

[0005] 본 발명의 배경이 되는 기술은 대한민국 공개특허공보 제10-2011-0024969호(2011.03.09)에 기재되어 있다. 상기 발명은 통계적 모델을 이용하여 잡음을 제거하는 기술을 개시하고 있으나 화자 인식에 있어서 화자의 음성과 섞여있는 잡음을 제거하고 효율적으로 화자를 인식함에는 적합하지 않다.

발명의 내용

해결하려는 과제

[0006] 따라서 본 발명이 해결하고자 하는 과제는, 잡음이 있는 환경에서 화자를 인식함에 있어서, 화자 인식률은 항상 시키고 계산의 복잡성은 줄일 수 있는 기법을 사용하여 Non-stationary 잡음 환경에서 강한 화자 인식 장치 및 그 방법을 제공하는 것이다.

과제의 해결 수단

[0007] 상기한 바와 같은 목적을 달성하기 위한 본 발명의 하나의 실시예에 따른 화자 인식 장치는, 화자별 입력 문장에 대한 특징 벡터를 추출하여 화자 인식 모델을 생성하는 화자 모델 생성부; 상기 음성 신호에 대한 특징 벡터를 추출하는 특징 벡터 추출부; 상기 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 상기 잡음 신호를 마스크(mask)하여 상기 클린 신호에 대한 특징 벡터를 선별하는 특징 벡터 선별부; 상기 클린 신호에 대한 특징 벡터를 이용하여 복수의 특징 벡터 조합을 생성하는 특징 벡터 조합부; 상기 클린 신호에 대한 특징 벡터와 상기 특징 벡터 조합 중에서 적어도 하나와 상기 화자 인식 모델을 비교하여 화자를 인식하는 화자 인식부; 를 포함한다.

[0008] 또한, 상기 화자 모델 생성부는, 가우시안 혼합 모델(Gaussian Mixture Model: GMM)을 이용하여 상기 화자 인식 모델을 생성할 수 있다.

[0009] 또한, 상기 특징벡터 선별부는, 다음의 수학적식과 같이 잡음대비 신호비(Signal-to-noise Ratio: SNR)를 기반으로 상기 클린 신호와 잡음 신호를 구별할 수 있다.

$$SNR(b) = 10 \log_{10} \frac{\sum_{f \in b\text{-th band}} P_{EC}(f)}{\sum_{f \in b\text{-th band}} P_{EN}(f)}$$

[0010]

[0011] 상기 P_{EC} 와 P_{EN} 은 각각 예측된 클린 신호와 예측된 잡음 신호의 파워스펙트럼을 나타낸다.

[0012] 또한, 상기 특징 벡터 조합부는, 상기 클린 신호에 대한 특징 벡터를 두 개 이상 조합할 수 있다.

[0013] 또한, 상기 화자 인식부는, 다음의 수학적식에 의한 특징 벡터와 상기 화자 인식 모델을 비교하여 화자를 인식할 수 있다.

$$\hat{s} = \arg \max_{s \in S} \sum_{t=1}^T \log p(Y_t | \lambda_s)$$

[0014]

[0015] 상기 수학적식에서,

$$p(\lambda_s | Y_{sub}) = \frac{p(Y_{sub} | \lambda_s) p(\lambda_s)}{\sum_{s \in S} p(Y_{sub} | \lambda_s) p(\lambda_s)}$$

[0016]

$$p(Y_{sub} | \lambda_s) = \sum_{m=1}^M w_{s,m} \prod_{k \in Y_{sub}} N_m(y_k | \lambda_{s,m})$$

[0017]

[0018] 이며, λ_s 는 s번째 화자의 GMM(mean, variance, weighting)을 나타내며, M개의 mixture를 갖는다. 또한, 가우스 분포 N_m 에서 s는 화자의 인덱스, m은 화자 모델의 인덱스, k는 특징 벡터의 인덱스를 각각 나타낸다. 또한,

$w_{s,m}$ 은 s번째 화자, m번째 mixture의 weighting 값을 나타낸다.

[0019] 본 발명의 실시예에 따른 화자 인식 방법은, 화자별 입력 문장에 대한 특징 벡터를 추출하여 화자 인식 모델을 생성하는 단계; 상기 음성 신호에 대한 특징 벡터를 추출하는 단계; 상기 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 상기 잡음 신호를 마스크(mask)하여 상기 클린 신호에 대한 특징 벡터를 선별

하는 단계; 상기 클린 신호에 대한 특징 벡터를 이용하여 복수의 특징 벡터 조합을 생성하는 단계; 상기 클린 신호에 대한 특징 벡터와 상기 특징 벡터 조합 중에서 적어도 하나와 상기 화자 인식 모델을 비교하여 화자를 인식하는 단계; 를 포함한다.

발명의 효과

[0020] 본 발명의 화자 인식 장치 및 방법에 따르면, 잡음이 있는 환경에서 화자를 인식함에 있어서, 잡음에 의한 영향을 최소화함으로써 화자 인식률을 향상시킬 수 있다.

[0021] 또한, 화자 인식에 사용되는 특징 벡터를 구함에 있어 잡음의 영향을 덜 받음으로써 계산의 복잡성을 피해 연산량을 줄일 수 있다.

도면의 간단한 설명

[0022] 도 1은 본 발명의 하나의 실시예에 따른 화자 인식 장치에 관한 구성도이다.

도 2는 본 발명의 하나의 실시예에 따른 화자 인식 방법에 관한 개략적인 순서도이다.

도 3은 도 2에 관한 상세한 순서도이다.

도 4는 도 2에서 특징 벡터 추출 단계를 상세하게 도시한 순서도이다.

도 5는 GMM 및 Mask-GMM 시스템에서의 특징 벡터의 개념도이다.

도 6는 AMFT(Advanced Missing Feature Theory) 기법에서의 특징 벡터 조합의 개념도이다.

도 7은 본 발명의 하나의 실시예에 따른 Mask-AMFT(Mask-advanced Missing Feature Theory) 기법에서의 특징 벡터 조합의 개념도이다.

도 8은 본 발명의 하나의 실시예에 따른 마스크되는 특징 벡터에 대한 특징 벡터 조합수를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0024] 이하 첨부된 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다.

[0025] 먼저 본 발명의 화자 인식 장치에 대해서 설명한다.

[0026] 도 1은 본 발명의 하나의 실시예에 따른 화자 인식 장치에 관한 구성도이다.

[0027] 도 1을 참조하면, 본 발명의 실시예에 따른 화자 인식 장치는 학습(Training)을 위한 구성요소와 테스트(Test)를 위한 구성요소로 이루어지는데, 특징 벡터 추출부(110), 화자 모델 생성부(120), 특징 벡터 선별부(130), 특징 벡터 조합부(140) 및 화자 인식부(150)를 포함한다.

[0028] 이하 본 발명의 실시예에 따른 화자 인식 장치의 각각의 구성요소에 대해 설명한다.

[0029] 특징 벡터 추출부(110)는 화자별 입력문장으로부터 특징 벡터를 추출한다. 또한 특징 벡터 추출부(110)는 테스트(Test)를 위해서 화자의 음성 신호로부터 특징 벡터를 추출한다. 화자의 음성 신호로부터 추출된 특징 벡터는 상기 화자별 입력문장으로부터 추출된 특징 벡터와 구별된다. 화자별 입력문장으로부터 추출된 특징 벡터는 화자 모델을 생성하는데 사용되는 반면, 화자의 음성 신호로부터 추출된 특징 벡터는 상기 생성된 모델과 비교하여 화자를 인식하는데 사용된다. 특징 벡터 추출부(110)는 여러 단계를 거쳐 특징 벡터를 추출하는데 자세한 과정은 도 2 내지 도 7을 통한 화자 인식 방법에서 설명하기로 한다.

[0030] 화자 모델 생성부(120)는 추출된 특징 벡터를 이용하여 화자 모델을 생성한다. 본 발명의 하나의 실시예로서 화자 모델 생성부(120)는 가우시안 혼합 모델(Gaussian Mixture Model: GMM)을 이용하여 화자 모델을 생성한다.

- [0031] 특징 벡터 선별부(130)는 음성 신호에 대한 특징 벡터로부터 클린 신호와 잡음 신호를 구별하고, 잡음 신호를 마스크(mask)하여 클린 신호에 대한 특징 벡터를 선별한다. 여기서 클린 신호란 노이즈가 혼합되지 않은 화자의 음성 신호만을 포함하는 신호를 말한다. 본 발명의 하나의 실시예로서 특징 벡터 선별부(130)는 잡음대비 신호비(Signal-to-noise Ratio: SNR)를 기반으로 클린 신호와 잡음 신호를 구별하여 클린 신호에 대한 특징 벡터를 선별한다.
- [0032] 특징 벡터 조합부(140)는 추출된 클린 신호에 대한 특징 벡터를 이용하여 특징 벡터 조합을 생성한다. 추출된 단일의 특징 벡터와 함께 단일의 특징 벡터로 구성되는 특징 벡터 조합이 화자 인식에 사용되는데, 특징 벡터를 조합하여 사용함으로써 화자 인식률을 높일 수 있는 효과가 있다. 특징 벡터 조합부(140)는 단일의 특징 벡터를 두 개 이상 조합하는 과정을 거쳐 특징 벡터 조합을 생성한다. 여기서 특징 벡터 조합을 특징 벡터 서브셋(sub-set)으로 명명한다.
- [0033] 화자 인식부(150)는 생성된 화자 모델과 추출된 특징 벡터 및 특징 벡터 조합을 비교하여 화자를 인식한다. 화자 인식은 음성 신호에 대한 화자를 검출하는 것과, 입력된 음성 신호가 특정 화자에 해당하는지 여부를 판단하는 것을 포함한다. 이상으로 본 발명의 화자 인식 장치의 각각의 구성에 대해 설명하였다.
- [0034] 이하 본 발명의 실시예에 따른 화자 인식 방법에 대해서 설명한다.
- [0035] 도 2는 본 발명의 하나의 실시예에 따른 화자 인식 방법에 관한 개략적인 순서도이고, 도 3은 도 2에 관한 상세한 순서도이다.
- [0036] 먼저 본 발명의 실시예에 따른 화자 인식 모델을 생성하기 위한 학습 단계에 대하여 설명한다.
- [0037] 도 2 및 도 3을 살펴보면, 특징 벡터 추출부(110)는 입력 문장에서 특징 벡터를 추출한다(Feature Extraction)(S210). 즉, 도 3과 같이 학습(Training) 단계에서 특징 벡터 추출부(110)는 입력문장(Training Dataset)에서 특징 벡터를 추출한다.
- [0038] 도 4는 도 2 및 도 3에 나타낸 특징 벡터 추출 단계(S210)를 상세하게 도시한 순서도이다.
- [0039] 도 4에서 특징 벡터 추출부(110)는 특징 벡터로서 De-correlated Feature를 추출한다. 즉, 특징 벡터 추출부(110)는 Hamming Window를 사용하여 표본을 추출하고(Windowing), 그 표본을 푸리에 변환(Fourier Transform)함으로써 주파수 신호를 얻는다. 이 주파수 신호로부터 다음의 수학적 식 1을 사용하여 Log-amplitude를 생성한다.

수학적 식 1

$$L(b) = \log \left(\sum_{k=b-th \text{ low}}^{b-th \text{ high}} |X(f)| M(b, k) \right)$$

- [0040]
- [0041] 상기 수학적 식 1에서 M(b, k)는 b band, k번째 주파수 bin의 me-scale filter 값을 나타낸다. 또한, X(f)는 f번째 주파수 신호를 나타낸다. 또한, b-th low, b-th high는 각각 mel-scale filter bank low, high 경계점을 나타낸다.
- [0042] 본 발명의 실시예에 따른 화자 인식 방법에서는 21개의 mel-scale filter bank(b=1, ..., 21)를 사용한다.
- [0043] 이러한 과정을 거쳐서 만들어진 Log-amplitude는 High-pass filter, $H(z)=1-z^{-1}$ 를 통과한다. 결과적으로 De-correlate feature는, $\{D(1)=L(2)-L(1), \dots, D(20)=L(21)-L(20)\}$ 으로 Frame당 20개가 추출된다.
- [0044] 화자 모델 생성부(120)는 상기 추출된 특징 벡터로부터 EM 알고리즘(EM algorithm)을 사용하여 가우시안 혼합 모델(Gaussian Mixture Model: GMM)로 화자 모델(Speaker Model)을 생성한다(S220).
- [0045] 이하에서는 테스트 단계에서 상기 생성된 화자 모델과 비교함으로써 인식될 화자의 음성 신호로부터 특징 벡터를 추출(Feature Extraction)하는 방법에 대하여 설명한다.
- [0046] 먼저 특징 벡터 추출부(110)는 인식 대상이 되는 화자의 음성 신호로부터 특징 벡터를 추출한다(S230). 이 과정은 상기 화자 모델을 생성하기 위한 과정과 같이 Hamming Window를 사용하여 표본을 추출(Windowing), Log-amplitude 생성, Mel-scale filter, High-pass filter를 거쳐 특징 벡터가 추출된다.

[0047] 다음으로 특징 벡터 선별부(130)에 의한 특징 벡터를 선별하는 단계(Sub-band mask Generation)(S240)를 거쳐서 최종적으로 특징 벡터 조합부(140)에 의해 특징 벡터 조합(Masked Feature set)을 생성한다(S250). 이하 특징 벡터를 선별하는 단계(S240)와 특징 벡터를 조합하는 단계(S250)에 대해서 설명한다.

[0048] 먼저 S240단계에서 특징 벡터 선별부(130)는 아래 수학적 2의 Missing Feature Theory(MFT)를 이용하여 특징 벡터를 선별한다. MFT는 잡음대비 신호비(Signal-to-noise Ratio: SNR)를 이용하여 잡음을 포함하는 음성 신호에서 클린 신호와 잡음 신호를 구분하는데 사용되는 기법이다. 여기서 특징 벡터 선별부(130)는 클린 신호를 Reliable feature로, 잡음 신호를 Unreliable feature로 구분하여 특징 벡터를 선별한다. 이후 Reliable feature만이 화자 인식에 사용된다.

수학적 2

$$SNR(b) = 10 \log_{10} \frac{\sum_{f \in b\text{-th band}} P_{EC}(f)}{\sum_{f \in b\text{-th band}} P_{EN}(f)}$$

[0049]

[0050] 상기 수학적 2에서 P_{EC} 및 P_{EN} 은 각각 예측 음성신호와 예측 잡음신호의 파워 스펙트럼(Power Spectrum)을 나타내며, 특징 벡터 선별부(130)는 SNR을 mel-scale band별로 계산한다. 결과적으로 특징 벡터 선별부(130)에서는 21개의 band별 SNR이 측정되며, De-correlated feature에 적용하기 위해서 다음 수학적 3에 의한 연산 과정을 거친다.

수학적 3

$$I(k) = \begin{cases} 1, & \text{if } SNR(b) > \delta \text{ and } SNR(b+1) > \delta \\ 0, & \text{otherwise.} \end{cases}$$

[0051]

[0052] 여기서 $I(k)$ 는 특징 벡터 중에서 Reliable featur와 Unreliable feature를 판별하는 판별식이다. $I(k)$ 의 값이 1인 경우 De-correlate feature를 화자 인식에 사용하고, 0인 경우에는 De-correlate feature를 화자 인식에 사용하지 않는다.

[0053] 다음으로 S250 단계에서는, 도 2 및 도 3에서 특징 벡터 조합부(140)는 선별된 특징 벡터를 이용하여 특징 벡터 조합(Masked Feature set)을 생성하는데,

[0054] 도 5 내지 도 7을 통하여 본 발명의 실시예에 따른 특징 벡터에 대하여 설명한다.

[0055] 도 5는 GMM 및 Mask-GMM 시스템에서의 특징 벡터의 개념도이고, 도 6은 AMFT(Advanced Missing Feature Theory) 기법에서의 특징 벡터 조합의 개념도이다. 도 7은 본 발명의 하나의 실시예에 따른 Mask-AMFT(Mask-advanced Missing Feature Theory) 기법에서의 특징 벡터 조합의 개념도이다.

[0056] 도 5 내지 도 7에서, 특징 벡터는 De-correlated feature가 두 개씩 포함된 De-correlated sub-band feature(DSF)를 사용한다.

[0057] 도 5에서는 마스크 기법을 사용하지 않는 일반적인 GMM의 특징 벡터의 개념도와 잡음의 사전정보를 이용하는 Mask-MFT의 특징 벡터의 개념도를 나타내었다.

[0058] 도 6에서는 잡음의 사전정보를 이용하지 않고 MFT를 진행하는 기법의 특징 벡터의 개념도를 나타내었다. 구체적인 과정을 살펴보면, AMFT 기법은 특징 벡터의 모든 sub-set을 이용하고, 각각의 Likelihood 값을 비교한다. 즉, AMFT에서는 사전정보 없이 각각의 featur를 Reliable, Unreliable로 가정하고 모든 경우의 수를 이용한다.

[0059] AMFT에서 특징 벡터는 다음과 같이 $Y(1)=(D(1), D(2)), \dots, Y(10)=(D(19), D(20))$ 으로 총 10개의 DSF로 구성된

다. 위와 같은 DSF로 feature-set의 모든 조합을 고려하게 되는데, 총 $\sum_{c=1}^{10} C_c^{10}$ 개의 경우의 수가 존재한다.

[0060] 이하에서는 도 7을 통하여 DSF1 내지 DSF3로 구성되는 Mask-AMFT(Mask-advanced Missing Feature Theory) 기법에서의 특징 벡터 조합에 대하여 예를 들어 설명한다.

[0061] 예를 들어서, 도 7에서 추출된 특징 벡터의 DSF가 DSF1 내지 DSF3이라고 할 때, 특징 벡터 선별부(130)는 DSF2를 마스크(mask)하여 특징 벡터 DSF1 및 DSF3을 선별하고, 특징 벡터 조합부(140)는 특징 벡터 조합을 통하여 DSF1 및 DSF3 이외에 (DSF1, DSF3) 조합을 추가적으로 생성한다.

[0062] 이와 같이 본 발명의 실시예에 따르면 S240 단계에서 특징 선별부(130)에서 클린 신호에 대응하는 특징 벡터를 선별하고, S250 단계에서 특징 벡터 조합부(140)는 상기 선별된 특징 벡터를 이용하여 특징 벡터 조합(Feature sub-set)을 생성한다.

[0063] 다음으로 화자 인식부(150)는 조합된 각 sub-set에 대해서 아래 수학적 식 4 내지 수학적 식 5을 이용하여 Likelihood를 연산하는 과정을 거쳐 최종적으로 가장 높은 확률값에 해당하는 화자의 인덱스 값을 수학적 식 6의 의해 산출함으로써 화자를 인식한다(S260).

수학적 식 4

$$p(\lambda_s | Y_{sub}) = \frac{p(Y_{sub} | \lambda_s) p(\lambda_s)}{\sum_{s \in S} p(Y_{sub} | \lambda_s) p(\lambda_s)}$$

[0064]

수학적 식 5

$$p(Y_{sub} | \lambda_s) = \sum_{m=1}^M w_{s,m} \prod_{k \in Y_{sub}} N_m(y_k | \lambda_{s,m})$$

[0065]

[0066] 상기 수학적 식 5에서, λ_s 는 s번째 화자의 GMM(mean, variance, weighting)을 나타내며, M개의 mixture를 갖는다. 또한, 가우스 분포 N_m 에서 s는 화자의 인덱스, m은 화자 모델의 인덱스, k는 특징 벡터의 인덱스를 각각 나타낸다. 또한, $w_{s,m}$ 은 s번째 화자, m번째 mixture의 weighting 값을 나타낸다.

수학적 식 6

$$\hat{s} = \arg \max_{s \in S} \sum_{t=1}^T \log p(Y_t | \lambda_s)$$

[0067]

[0068] 상기 수학적 식 6에서, 생성된 화자 모델에 대한 최대의 Likelihood 값을 산출하게 된다. 즉, 수학적 식 6은 입력된 음성 신호의 특징 벡터(Y_t)와 화자 모델의 특징 벡터(λ_s)를 비교하여 산출된 Likelihood 값 중에서 최대값을 반환하게 된다.

[0069] 도 8은 본 발명의 하나의 실시예에 따른 마스크되는 특징 벡터에 대한 특징 벡터 조합수를 나타내는 그래프이다.

[0070] 이와 같이 본 발명의 실시예에 따른 화자 인식 장치 및 방법에 따르면, 잡음이 있는 환경에서 화자를 인식함에 있어서, 잡음에 의한 영향을 최소화함으로써 화자 인식률을 향상시킬 수 있다.

[0071] 또한, 화자 인식에 사용되는 특징 벡터를 구함에 있어, 연산량은 가중시키면서 인식률을 저하시키는, 잡음 신호에 대응하는 특징 벡터를 마스크함으로써 특징 벡터의 조합 수를 줄일 수 있게 되어 연산량을 줄일 수 있고 인식률을 높일 수 있다.

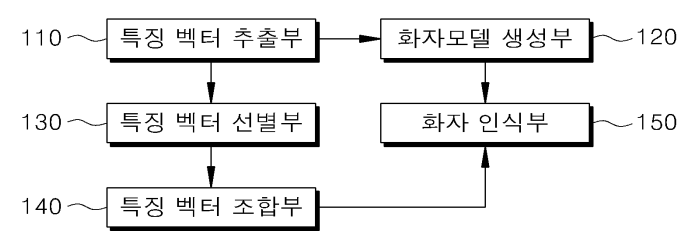
[0072] 이제까지 본 발명에 대하여 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 따라서 본 발명의 범위는 전술한 실시예에 한정되지 않고 특허청구범위에 기재된 내용 및 그와 동등한 범위 내에 있는 다양한 실시 형태가 포함되도록 해석되어야 할 것이다.

부호의 설명

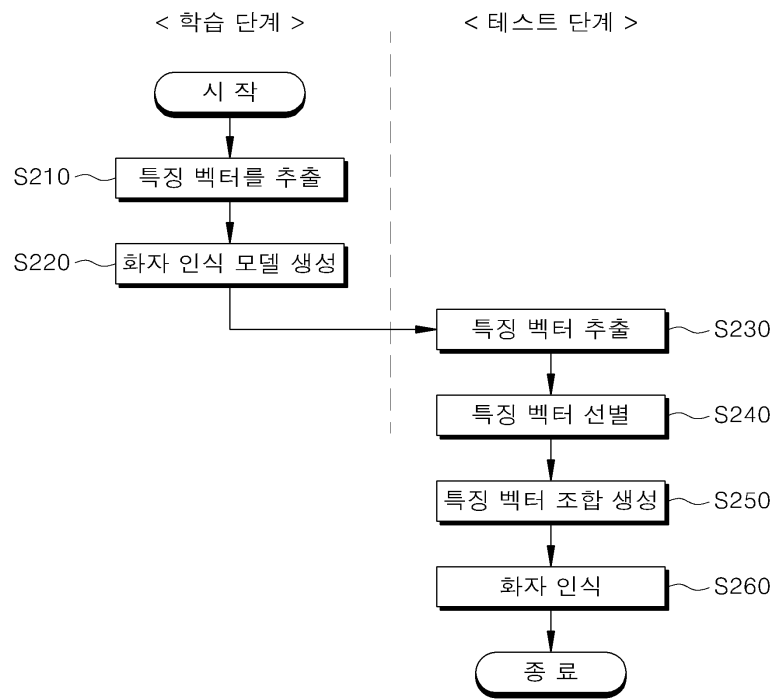
- [0073] 110: 특징 벡터 추출부 120: 화자모델 생성부
130: 특징 벡터 선별부 140: 특징 벡터 조합부
150: 화자 인식부

도면

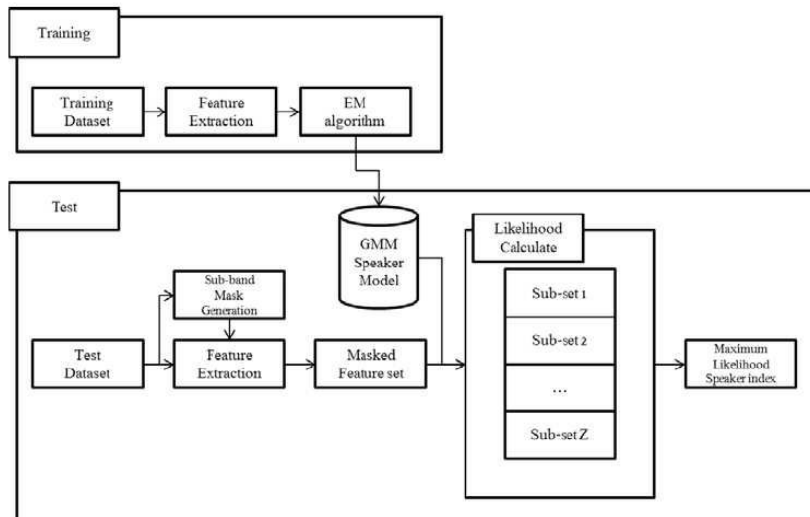
도면1



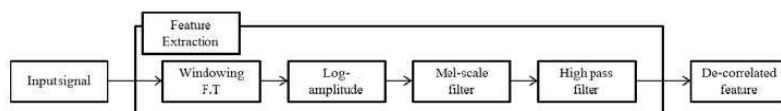
도면2



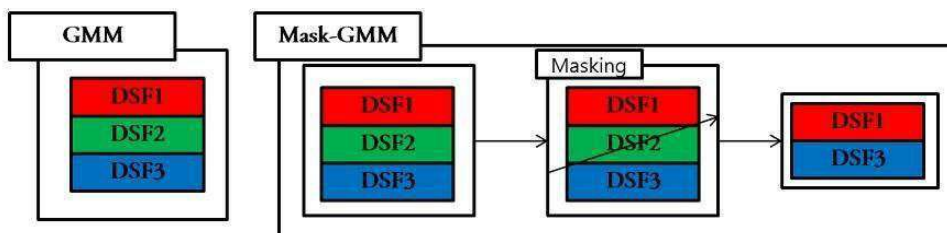
도면3



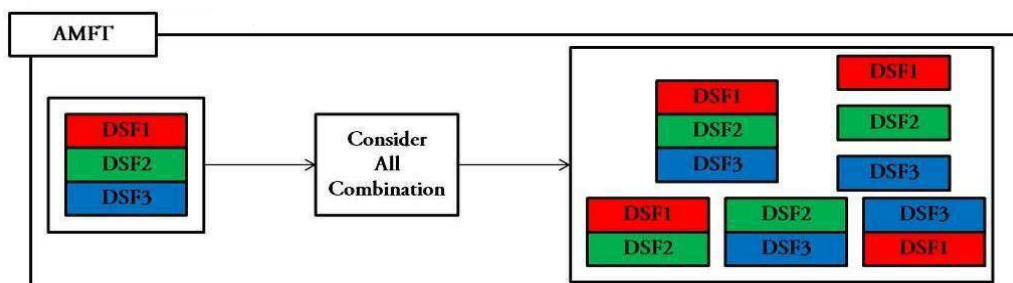
도면4



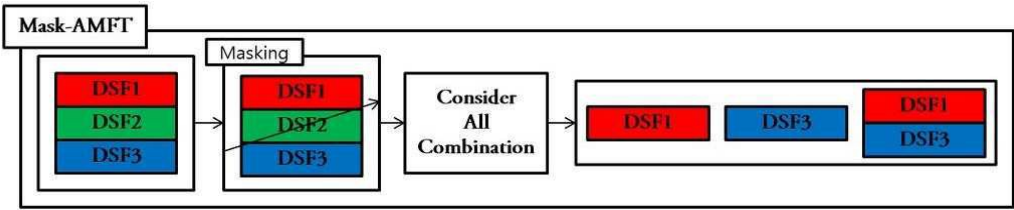
도면5



도면6



도면7



도면8

