(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년08월10일
(11) 등록번호 10-1647059
(24) 등록일자 2016년08월03일

(51) 국제특허분류(Int. Cl.)
G10L 21/0216 (2013.01) *G10L 15/14* (2006.01)
(52) CPC특허분류
G10L 21/0216 (2013.01)
G10L 15/142 (2013.01)
(21) 출원번호 10-2015-0037384
(22) 출원일자 2015년03월18일
심사청구일자 2015년03월18일
(56) 선행기술조사문헌
KR1020130125014 A*
Alexander Krugger et al., 'Model-based feature enhancement for reverberant speech recognition', IEEE Trans. on Audio,Speech and Language Processing, Vol.18, No.7, September 2010.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
박형민
서울특별시 강남구 삼성로 212 3동 1002호 (대치동, 은마아파트)
조지원
서울특별시 마포구 독막로32안길 15 옥탑 (신수동)
(74) 대리인
이지연

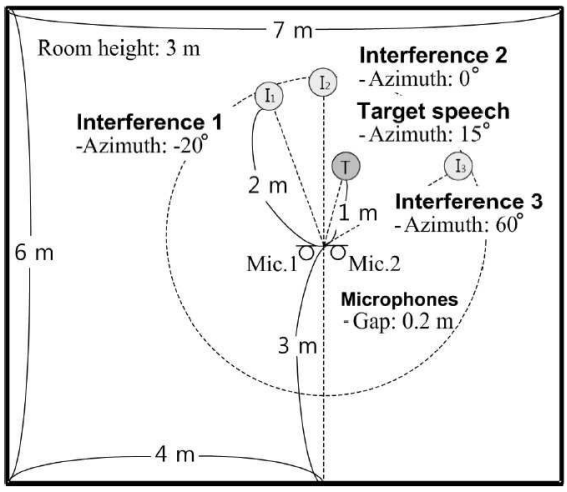
전체 청구항 수 : 총 3 항 심사관 : 정성윤

(54) 발명의 명칭 **독립 벡터 분석 및 모델 기반 특징 향상을 이용한 강인한 음성 인식 방법**

(57) 요 약

본 발명은 독립 벡터 분석 및 모델 기반 특징 향상 방법을 이용한 강인한 음성 인식 방법에 관한 것이다. 상기 음성 인식 방법은, (a) 마이크로폰으로 입력된 신호들에 대하여 독립 벡터 분석(Independent Vector Analysys)하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계; (b) 상기 추정된 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 각각 입력 신호 및 노이즈 신호로 설정하여, 모델 기반 특징 향상(Model-based Feature Enhancement;'MBFE') 방법을 적용하여 타겟 음성 신호를 추정하는 단계; 를 구비하여 음성 인식의 정확도를 향상시킨다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2014R1A2A2A01006581
부처명	미래창조과학부
연구관리전문기관	한국연구재단
연구사업명	이공분야기초연구사업 중견연구자지원산업
연구과제명	시청각정보를 이용한 강인한 멀티모달 음성인식 기술 개발
기 여 율	1/1
주관기관	서강대학교 산학협력단
연구기간	2014.05.01 ~ 2017.04.30

명 세 서

청구범위

청구항 1

(a) 마이크로폰으로 입력된 신호들에 대하여 독립 벡터 분석(Independent Vector Analysys)하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계;

(b) 상기 추정된 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 각각 입력 신호 및 노이즈 신호로 설정하여, 모델 기반 특징 향상(Model-based Feature Enhancement;'MBFE') 방법을 적용하여 타겟 음성 신호를 추정하는 단계;

를 구비하고, 상기 (a) 단계는

(a1) 타겟에 대한 방향을 추정하는 단계;

(a2) 마이크로폰으로 입력된 신호들에 대하여 독립 벡터 분석하여 소스 신호들을 출력하는 단계;

(a3) 상기 추정된 타겟에 대한 방향을 이용하여 상기 IVA 출력 신호 중 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 각각 선택하는 단계;

를 구비하여 음성 인식의 정확도를 향상시킨 것을 특징으로 하는 강인한 음성 인식 방법.

청구항 2

삭제

청구항 3

제1항에 있어서, 상기 (a3) 단계는,

분리 매트릭스를 이용하여 directions-of-Arrivals(DOA)를 계산하고, 상기 추정된 방향 정보와 상기 계산된 DOA를 비교하고,

(a2) 단계에서 출력된 소스 신호들 중에서, 상기 추정된 타겟에 대한 방향과 DOA의 차이가 최소인 출력은 IVA 타겟 음성 신호로 결정하고, 그 외의 다른 출력들은 IVA 노이즈 신호로 결정하는 것을 특징으로 하는 강인한 음성 인식 방법.

청구항 4

제1항에 있어서, 상기 (b) 단계에서의 상기 모델 기반 특징 향상 방법은 HMM(Hidden-Markov-model)을 기반으로 한 로그 멜-주파수 파워 스펙트럼 계수들(Logarithmic Mel-frequency Power Spectral Coefficients; 이하 'LMPSCs'라 한다)로부터 clean LMPSCs의 최소 평균 제곱 에러(a minimum mean square error ; 이하, 'MMSE'라 한다)의 추정을 유도해 내는 것을 특징으로 하는 강인한 음성 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 강인한 음성 인식 방법에 관한 것으로서, 더욱 구체적으로는, 독립 벡터 분석의 출력 신호들을 이용하여 모델-기반 특징 향상 방법을 적용함으로써 타겟 음성 신호를 추정하여 모든 노이즈에 강인한 음성 인식 방법에 관한 것이다.

배 경 기 술

[0002] 음성 인식 시스템(ASR System)에 있어서, 실제 상황은 대부분 노이즈가 많은 환경에 있기 때문에 노이즈에 강인한 특성(Noise robustness)을 갖는 것은 매우 중요하다. 음성 인식 시스템의 인식 성능의 감쇠는 주로 학습 환경과 실제 환경과의 차이로부터 기인하는 경우가 많으며, 캡스트럼 평균 정규화(Cepstral Mean normalization)

또는 스펙트럼 차감(spectral subtraction)과 같은 단순한 방법들을 사용하여 전술한 불일치를 보상하여 손실들을 회복하고자 하는 연구들이 많이 진행되고 있다. 이러한 접근들에 의하여 음성 인식의 정확성을 향상시키고는 있지만, 대부분은 다양하면서 안정화되지 못한 노이즈들을 갖는 실세계의 환경들에서는 여전히 인식 성능을 향상시키지 못하고 있는 실정이다.

[0003] 모델 기반 특징 향상(Model-based feature enhancement; 이하, 'MBFE'라 한다) 기법들은 관측된 noisy speech의 스펙트로그래픽 표현에서 노이즈로 인한 왜곡을 제거함으로써, Noise robustness을 향상시키는 것이다. 이러한 기법들은 주로 통계적인 공식에 기초한 것으로서, 입력된 Noisy speech에 대한 clean speech의 예상값, 입력 음성을 변질시키는 노이즈에 대한 일부 변수들의 추정 및 clean speech를 설명하는 a prior model을 제공한다.

[0004] MBFE의 성능은 corrupting noise를 설명하기 위하여 추정된 변수들에 주로 의존하게 된다. 하지만, 노이즈의 특성들은 매우 다양하고 모든 종류의 노이즈를 커버할 수 있는 노이즈 모델을 학습시키는 것은 불가능하다. 이 경우, 노이즈 모델에 관련된 변수들을 얻는 가장 쉬운 방법들 중 하나는, 단일의 음성 발성동안 정상적 노이즈(stationary noise) 및 low complex noise 모델을 가정하여, 묵음 기간동안의 평균(mean)과 같은 단순 통계를 추정하는 것이다.

[0005] 하지만, 배경음악이나 경쟁이 되는 다른 음성과 같은 corrupting noise들은 짧은 기간동안이라 할지라도 정상적(stationary)이지 않을 수 있다. 다수 개의 마이크로폰이 있다면, 노이즈의 일부 변수들은 음성 성분을 제거한 후의 speech period 동안에서도 추정될 수 있다. 많은 음성 인식 시스템에 있어서, 관심 화자는 마이크로폰과 상대적으로 가까운 위치에 배치되며, 그 결과 그 방향이 쉽게 추정되고 타겟 반향 성분들이 감소될 수 있는 점음원(point source)으로 취급될 수 있다.

[0006] 종래의 BSSA(Blind Spatial Subtraction Array)라 불리는 특징 향상 방법은, 타겟 음성 신호 출력을 제거한 후, projection-back 방법을 사용한 ICA 에 의해 추정된 노이즈 파워 스펙트럼(noise power spectrum)을 빼는 것을 특징으로 한다. 이러한 BSSA 방법은, ICA 의 타겟 음성 신호 출력이 여전히 노이즈를 포함하고 있으며, 노이즈 파워 스펙트럼 추정이 완벽할 수 없기 때문에, 음성 인식의 성능을 떨어뜨리게 되는 문제점이 있다.

[0007] 이와 같이, 종래의 모델 기반 특징 향상 기법은 stationary noise에 대해서는 noise robustness를 가지나, 배경음악이나 경쟁이 되는 다른 음성 신호들과 같은 nonstationary noise에 대해서는 추정할 수 없게 된다. 따라서, 이러한 종래의 모델 기반 특징 향상 기법을 사용하여 음성 인식을 수행하는 경우, 노이즈가 없는 타겟 음성 신호를 정확하게 추정하기 어렵기 때문에 음성 인식의 정확도가 떨어지게 된다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국등록특허 제 10-0329596호

(특허문헌 0002) 한국등록특허 제 10-1361034호

발명의 내용

해결하려는 과제

[0009] 전술한 문제점을 해결하기 위한 본 발명의 목적은 관심 화자인 타겟에 대한 방향 정보를 이용하여 독립 벡터 분석에 의해 얻은 출력 신호들로부터 IVA 타겟 음성 신호와 IVA 노이즈 신호를 1차로 추정하고, 1차로 추정된 신호들을 이용하여 모델 기반 특징 향상 방법을 적용함으로써, 정상적 노이즈(stationary noise) 뿐만 아니라 비정상적 노이즈(non-stationary noise)들에도 강인한 타겟 음성 신호를 추정할 수 있는 강인한 음성 인식 방법을 제공하는 것이다.

과제의 해결 수단

[0010] 전술한 기술적 과제를 달성하기 위한 본 발명의 특징에 따른 강인한 음성 인식 방법은, (a) 마이크로폰으로 입력된 신호들에 대하여 독립 벡터 분석(Independent Vector Analysys)하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계; (b) 상기 추정된 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 각각 입력 신호 및 노이즈 신호로 설정하여, 모델 기반 특징 향상(Model-based Feature Enhancement;'MBFE') 방법을 적용하여 타겟 음성 신호를 추정하는 단계; 를 구비하여 음성 인식의 정확도를 향상시킨다.

[0011] 전술한 특징에 따른 강인한 음성 인식 방법에 있어서, 상기 (a) 단계는 (a1) 타겟에 대한 방향을 추정하는 단계; (a2) 마이크로폰으로 입력된 신호들에 대하여 독립 벡터 분석하여 소스 신호들을 출력하는 단계; (a3) 상기 추정된 타겟에 대한 방향을 이용하여 상기 IVA 출력 신호 중 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 각각 선택하는 단계;를 구비하는 것이 바람직하다.

[0012] 전술한 특징에 따른 강인한 음성 인식 방법에 있어서, 상기 (a3) 단계는, 분리 매트릭스를 이용하여 directions-of-Arrivals(DOA)를 계산하고, 상기 추정된 방향 정보와 상기 계산된 DOA를 비교하고, (a2) 단계에서 출력된 소스 신호들 중에서, 상기 추정된 타겟에 대한 방향과 DOA의 차이가 최소인 출력은 IVA 타겟 음성 신호로 결정하고, 그 외의 다른 출력들은 IVA 노이즈 신호로 결정하는 것이 바람직하다.

[0013] 전술한 특징에 따른 강인한 음성 인식 방법에 있어서, 상기 (b) 단계에서의 상기 모델 기반 특징 향상 방법은 HMM(Hidden-Markov-model)을 기반으로 한 로그 멜-주파수 파워 스펙트럼 계수들(Logarithmic Mel-frequency Power Spectral Coefficients; 이하 'LMPSCs'라 한다)로부터 clean LMPSCs의 최소 평균 제곱 에러(a minimum mean square error ; 이하, 'MMSE'라 한다)의 추정을 유도해 내는 것이 바람직하다.

발명의 효과

[0014] 본 발명에 따른 강인한 음성 인식 방법은 정상적 노이즈(stationary noise) 뿐만 아니라, 비정상적 노이즈(nonstationary noise)들에도 강인한 음성 인식을 수행할 수 있게 된다.

도면의 간단한 설명

[0015] 도 1은 본 발명에 따른 음성 인식 방법과 종래의 음성 인식 방법들의 성능을 분석하기 위하여, 구성된 소스와 마이크로폰들을 도시한 구성도이다.

도 2는 도 1의 시뮬레이션 환경에서, 간섭신호원이 1개인 경우, 2개인 경우, 3개인 경우에 대하여, 각각 종래방법 1(Baseline), 종래방법 2(FE), 종래 방법 3(IVA), 종래 방법 4(BSSA) 및 본 발명에 따른 방법(IVA-FE)에서의 SNR을 측정하여 도시한 도표이다.

발명을 실시하기 위한 구체적인 내용

[0016] 본 발명에 따른 강인한 음성 인식 방법은, 독립 벡터 분석(Independent Vector Analysis; 이하, 'IVA'라 한다)을 이용하여 마이크로폰들의 입력 신호들로부터 타겟 음성 신호 및 노이즈 신호를 1차로 추정하고, 1차 추정된 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 이용하여 모델 기반 특징 향상(Model-based Feature Enhancement) 방법을 적용하여, 정상적 노이즈 (stationary noise) 뿐만 아니라 비정상적 노이즈(nonstationary noise)도 제거된 타겟 음성 신호를 추정하는 것을 특징으로 한다.

[0017] 특히, 본 발명은 강인한 음성 인식을 위하여 사용되는 특징 향상 기법은, HMM(Hidden Markov model)을 기반으로 한 로그 멜-주파수 파워 스펙트럼 계수(Logarithmic Mel-frequency Power Spectral Coeffecients; 이하, 'LMPSCs'라 한다.)로부터 clean LMPSCs의 최소 평균 제곱 에러(a minimum mean square error ; 이하, 'MMSE'라 한다)의 추정을 유도해 내는 것이다.

[0018] 특징 향상을 위한 노이즈 모델과 관련된 변수들은, 단일 음성 발성 동안 정상적 노이즈(stationary noise) 및 단일 노이즈 모델을 가정하여, 단일 음성 발성이 없는 기간 동안의 평균이나 편차와 같은 기본 통계에 의해 추정할 수도 있지만, 노이즈는 단일의 음성 발성내에 짧은 간격에서조차도 stationary하지 않을 수 있다. 따라서, 본 발명에 따른 방법은 모델 기반 특징 향상을 위하여, 독립 벡터 분석(Independent Vector Analysis; 'IVA')를 채용하여, 사용된 음성 동안 연속적으로 변수들을 추정하고 향상된 타겟 음성 신호를 획득한다.

[0019] 본 발명에 따른 방법은 실험적 결과로부터 경쟁되는 음성이 있는 경우라도 음성 인식에서의 노이즈 강인함(Noise robustness)을 더욱 향상시킬 수 있음을 확인할 수 있다.

[0020] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 강인한 음성 인식 방법에 대하여 구체적으로 설명한다.

[0021] 먼저, 본 발명에 다른 강인한 음성 인식 방법은 마이크로폰으로부터 입력된 입력 신호들에 대하여 독립 벡터 분석하고, 음원 소스들 중 관심 화자인 타겟의 방향에 대한 정보를 사전에 제공받고, 상기 타겟의 방향에 대한 정보를 이용하여 독립 벡터 분석의 출력들 중 타겟 음성 신호 및 노이즈 신호를 1차적으로 추정한다. 이하, 독립 벡터 분석하여 IVA 타겟 음성 신호 및 IVA 노이즈 신호를 1차적으로 추정하는 과정을 구체적으로 설명한다.

[0022] 먼저, 벡터 $\boldsymbol{\Xi}_{i,j}$는, 상호 독립된 N개의 미지의 음원 소스들의 음성이 혼합된 벡터로서, i번째 프레임이며 j 번째 주파수 빈에서의 M 개의 observations의 시간-주파수 표현들로 구성된다. 소스 신호들을 복구하기 위한 observations에 대한 선형 변환은 수학식 1로 표현될 수 있다.

수학식 1

[0023] $$\boldsymbol{\Psi}_{i,j} = \mathbf{W}_j \boldsymbol{\Xi}_{i,j}$$

[0024] 여기서, $\boldsymbol{\Psi}_{i,j} = [Y_{i,j}^1, \cdots, Y_{i,j}^N]^T$ 은 추정된 소스 신호들의 시간-주파수 표현으로 구성된 벡터이며, $\boldsymbol{W}_j$는 j번째 주파수 빈에서의 분리 매트릭스(separating matrix)이다.

[0025] 분리 매트릭스를 추정하기 위한 자연-경사도 IVA(natural-gradient IVA) 학습 규칙은 수학식 2에 의해 구할 수 있다.

수학식 2

[0026] $$\Delta \mathbf{W}_j \propto \{\mathbf{I} - E[\boldsymbol{\Phi}_j(\mathbf{Y}_i^{1:N}) \boldsymbol{\Psi}_{i,j}^H]\} \mathbf{W}_j$$

[0027] 여기서, I 는 아이덴티티 매트릭스(identity matrix)이다. $[Y_{i,1}^n, \cdots, Y_{i,J}^n]^T$ 를 나타내는 $\boldsymbol{Y}_i^n$를 갖는 추정된 소스들의 hypothesized pdf model $q(\mathbf{Y}_i^n) \propto \exp\left(-\sqrt{\sum_{j=1}^J |Y_{i,j}^n|^2}\right)$ 을 가정하면, multivariate score function은 $\boldsymbol{\Phi}_j(\mathbf{Y}_i^{1:N}) = [\phi_j(\mathbf{Y}_i^1), \cdots, \phi_j(\mathbf{Y}_i^N)]^T$ 및 $\phi_j(\mathbf{Y}_i^n) = -\partial \log q(\mathbf{Y}_i^n)/\partial Y_{i,j}^n = Y_{i,j}^n / \sqrt{\sum_{j'=1}^J |Y_{i,j'}^n|^2}$, $1 \le n \le N$ 에 의해 주어지며, 여기서, $\boldsymbol{J}$는 주파수 빈의 개수를 나타낸다.

[0028] 한편, 본 발명에 따른 강인한 음성 인식 방법에 있어서, 음성 소스들 중 관심 화자인 타겟의 도달 방향(Direction-of-Arrival; 이하, 'DOA')에 대한 정보는 사전에 입력되거나 image-based source detection 방법들에 의해 추정되어 제공될 수 있다. 전술한 방법에 의하여 타겟 소스에 대한 추정된 방향은 분리 매트릭스 $\boldsymbol{W}_j$로부터 계산된 DOA와 비교함으로써, IVA 출력 신호들 중에서 타겟 음성 신호 및 노이즈 신호를 결정할 수 있다.

[0029] 균일하게 선형적인 마이크로폰 어레이를 사용하는 경우, j 번째 주파수 빈에서 n 번째 출력에 대응되는 사운드

소스의 DOA인 θ_j^n 는 m 번째 및 m'번째 마이크로폰 쌍으로부터 수학식 3에 의해 구해질 수 있다.

수학식 3

[0030]
$$\theta_j^n = \arcsin\left\{\frac{\arg\left(\frac{[\mathbf{W}_j^{-1}]^{mn}}{[\mathbf{W}_j^{-1}]^{m'n}}\right)}{2\pi(j-1)/J \cdot f_s \cdot d(m-m')/c}\right\}$$

[0031] 여기서, $[\cdot]^{nm}$ 은 매트릭스의 m 번째 열과 n 번째 행의 성분을 나타내며, f_s는 샘플링 주파수이며, d는 인접한 마이크로폰들 사이의 거리이며, c는 음속이다.

[0032] permutation 문제들이 IVA에서 발생되지 않기 때문에, DOA는 spatial aliasing 없이 주파수 빈들에 대한 θ_j^n 의 median value에 의해 구할 수 있다. 따라서, IVA 출력들에 있어서, 추정된 타겟 소스 방향으로부터 DOA 차이가 최소인 출력은 타겟 음성 출력으로 결정될 수 있으며, 그 외의 다른 출력들은 노이즈 출력으로 결정될 수 있다.

[0033] 다음, 본 발명에 따른 음성 인식 방법은 인식 성능을 향상시키기 위하여, 전술한 IVA에 의해 추정된 타겟 음성 출력(Y_i^n) 및 노이즈 출력을 이용하여 특징 향상(the feature enhancement)을 수행한다.

[0034] 특징 향상을 수행함에 있어서, noisy speech로서 IVA 타겟 음성 출력(Y_i^n) 을 사용하고, noise로는 IVA 노이즈 출력을 사용한다. 이하, HMM 모델 기반의 특징 향상의 수행 과정에 대하여 구체적으로 설명한다.

[0035] i 번째 프레임 및 j 번째 주파수 빈에서의 노이즈 음성의 시간-주파수 표현인 $Y_{i,j}$는 수학식 4에 의해 나타낼 수 있다.

수학식 4

[0036]
$$Y_{i,j} = S_{i,j} + N_{i,j}$$

[0037] 여기서, $S_{i,j}$ $N_{i,j}$ 는 각각 clean speech 및 noise를 표시한다.

[0038] K-채널 멜-스케일 필터들(K-channel mel-scale filters) Λ_k, k=1,...,K on $|Y_{i,j}|^2$ 를 적용하면, noisy speech에 대한 LMPSC 인 $y_{i,k}$는 수학식 5에 의해 얻을 수 있다.

수학식 5

$$y_{i,k} = \log\left\{\sum_{j=N_l(k)}^{N_u(k)} |Y_{i,j}|^2 \Lambda_k(j)\right\}$$

[0039]

[0040] 여기서, $N_l(k)$ 및 $N_u(k)$는 각각 k 번째 필터의 non-zero weights의 하위 및 상위 bounds 들이다.

[0041] 음성신호와 노이즈 신호가 서로 상관되어 있지 않다(uncorrelated)고 가정하면, LMPSC $y_{i,k}$는 수학식 6으로 표현될 수 있다.

수학식 6

$$y_{i,k} = \log[\exp(s_{i,k}) + \exp(n_{i,k})]$$

[0042]

[0043] 여기서, $s_{i,k}$ 및 $n_{i,k}$는 , 각각 clean speech $S_{i,j}$ 와 노이즈 $N_{i,j}$의 LMPSCs 들이다. noisy LMPSCs 의 향상을 수행하기 위하여, 수학식 6은 $(b_{s_{i,k}}, b_{n_{i,k}})$에서 the first-order Taylor series expansion을 사용하여 선형화시킴에 의해 근사적으로 수학식 7로 표현할 수 있게 된다.

수학식 7

$$\begin{aligned} y_{i,k} \approx\ & \log[\exp(b_{s_{i,k}}) + \exp(b_{n_{i,k}})] \\ & + d_{s_{i,k}}(s_{i,k} - b_{s_{i,k}}) + d_{n_{i,k}}(n_{i,k} - b_{n_{i,k}}) \end{aligned}$$

[0044]

[0045] 여기서, $d_{s_i,k}$ 및 $d_{n_i,k}$는 수학식 8 및 9이다.

수학식 8

$$d_{s_{i,k}} = \frac{\partial y_{i,k}}{\partial s_{i,k}}|_{s_{i,k}=b_{s_{i,k}}} = \frac{\exp(b_{s_{i,k}})}{\exp(b_{s_{i,k}}) + \exp(b_{n_{i,k}})}$$

[0046]

수학식 9

$$\begin{aligned} d_{n_{i,k}} &= \frac{\partial y_{i,k}}{\partial n_{i,k}}|_{n_{i,k}=b_{n_{i,k}}} = \frac{\exp(b_{n_{i,k}})}{\exp(b_{s_{i,k}}) + \exp(b_{n_{i,k}})} \\ &= 1 - d_{s_{i,k}} \end{aligned}$$

[0047]

[0048] 대부분의 실질적인 상황에 있어서, 수학식 6에서의 $n_{i,k}$의 정확한 값은 알지 못하며, random process로서 간주된다. 따라서, $s_{i,k}$ 및 $y_{i,k}$사이의 관계는 deterministic transformation으로서 설명될 수 없으며, probabilistic framework로 설명될 수 있다.

[0049] noise statistics를 설명하는 일부 변수들이 noisy signal 로부터 추정될 수 있다면, noisy speech $y_{i,k}$및 추정된 변수들의 값이 주어진다면 noisy LMPSCs의 향상은 수학식 10에 의해 나타내어지는 clean LMPSC vector의 MMSE 추정을 계산함에 의해 얻을 수 있게 된다.

수학식 10

[0050] $$\hat{\mathbf{s}}_i = E[\mathbf{s}_i|\mathbf{y}_{1:i}]$$

[0051] 여기서, s_i는 $[s_{i,1},\cdots,s_{i,K}]^T$의 특징 벡터를 나타내며, $y_{1:i}$는 특징 벡터들의 시퀀스인 (y_1, ... , y_i)를 나타낸다. Bayesian framework를 사용하면, MMSE 추정은 posterior 확률밀도함수(pdf)인 $p(\mathbf{s}_i|\mathbf{y}_{1:i})$로부터 얻을 수 있다.

[0052] prior distribution을 수식화하기 위하여, clean speech는 Single-Gaussian observation distribution을 갖는 HMM에 의해 설명될 수 있다고 가정하고, 수학식 11을 유도한다.

수학식 11

[0053] $$p(\mathbf{s}_i|\gamma_i = l) = \mathcal{N}(\mathbf{s}_i|\boldsymbol{\mu}_{\mathbf{s}_i|\gamma_i=l}, \boldsymbol{\Sigma}_{\mathbf{s}_i|\gamma_i=l}),\ \ l = 1,\cdots,N_\gamma$$

[0054] 여기서, γ_i N_γ는 각각 프레임 i에서의 hidden state index 및 state들의 개수를 나타낸다.

[0055] 또한, covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s}_i|\gamma_i=l}$ 이 $\sigma^2_{s_{i,k}|\gamma_i=l}$의 k-th diagonal element와 diagonal하다고 가정하고, 노이즈 $n_{i,k}$ 는 mean과 variance 가 각각 $\mu_{n_i,k}$, $\sigma^2_{n_i,k}$인 Gaussian random process라고 가정한다.

[0056] 그 다음, posterior pdf는 수학식 12로 수식화할 수 있다.

수학식 12

[0057] $$p(\mathbf{s}_i|\mathbf{y}_{1:i}) = \sum_{l=1}^{N_\gamma} P(\gamma_i = l|\mathbf{y}_{1:i})p(\mathbf{s}_i|\mathbf{y}_i, \gamma_i = l)$$

[0058] 여기서, k 번째 멜 주파수 밴드의 오른쪽의 second term은 Bayesian rule에 의해 수학식 13과 같이 표현될 수 있다.

수학식 13

$$p(s_{i,k}|y_{i,k},\gamma_i=l) \propto p(y_{i,k}|s_{i,k},\gamma_i=l)p(s_{i,k}|\gamma_i=l)$$

[0059]

[0060] 수학식 7에서 $b_{s_{i,k}}=\mu_{s_{i,k}|\gamma_i=l}$ 및 $b_{n_{i,k}}=\mu_{n_{i,k}}$ 일 때, 수학식 13의 오른쪽의 first term은 수학식 14와 같이 근사화할 수 있다.

수학식 14

$$p(y_{i,k}|s_{i,k},\gamma_i=l) \approx \mathcal{N}(y_{i,k}|\tilde{\mu}_{y_{i,k}|\gamma_i=l}, d^2_{n_{i,k}}\sigma^2_{n_{i,k}})$$

[0061]

[0062] 여기서, $\tilde{\mu}_{y_{i,k}|\gamma_i=l}$ 은 수학식 15 이다.

수학식 15

$$\begin{aligned}\tilde{\mu}_{y_{i,k}|\gamma_i=l} &= \log[\exp(\mu_{s_{i,k}|\gamma_i=l})+\exp(\mu_{n_{i,k}})] \\ &\quad + d_{s_{i,k}}(s_{i,k}-\mu_{s_{i,k}|\gamma_i=l})\end{aligned}$$

[0063]

[0064] second term은 수학식 11로부터 수학식 16과 같이 나타낼 수 있다.

수학식 16

$$p(s_{i,k}|\gamma_i=l)=\mathcal{N}(s_{i,k}|\mu_{s_{i,k}|\gamma_i=l},\sigma^2_{s_{i,k}|\gamma_i=l})$$

[0065]

[0066] 따라서, $p(s_{i,k}|y_{i,k},\gamma_i=l)$ 은 또한 수학식 17과 같이 나타낼 수 있다.

수학식 17

$$p(s_{i,k}|y_{i,k},\gamma_i=l) \approx \mathcal{N}(s_{i,k}|\hat{\mu}_{s_{i,k}|\gamma_i=l},\hat{\sigma}^2_{s_{i,k}|\gamma_i=l})$$

[0067]

수학식 18

$$\begin{aligned}\hat{\mu}_{s_{i,k}|\gamma_i=l} &= \mu_{s_{i,k}|\gamma_i=l} \\ &+\frac{d_{s_{i,k}}\sigma^2_{s_{i,k}|\gamma_i=l}\{y_{i,k}-\log[\exp(\mu_{s_{i,k}|\gamma_i=l})+\exp(\mu_{n_{i,k}})]\}}{d^2_{s_{i,k}}\sigma^2_{s_{i,k}|\gamma_i=l}+d^2_{n_{i,k}}\sigma^2_{n_{i,k}}}\end{aligned}$$

[0068]

수학식 19

$$\hat{\sigma}^2_{s_{i,k}|\gamma_i=l} = \frac{\sigma^2_{s_{i,k}|\gamma_i=l} d^2_{n_{i,k}} \sigma^2_{n_{i,k}}}{d^2_{s_{i,k}} \sigma^2_{s_{i,k}|\gamma_i=l} + d^2_{n_{i,k}} \sigma^2_{n_{i,k}}}$$

[0069]

[0070] 수학식 12의 오른쪽의 first term은 수학식 20과 같이 나타낼 수 있다.

수학식 20

$$P(\gamma_i = l|\mathbf{y}_{1:i}) = \frac{P(\gamma_i = l|\mathbf{y}_{1:i-1}) p(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \gamma_i = l)}{\sum_{l'=1}^{N_\gamma} P(\gamma_i = l'|\mathbf{y}_{1:i-1}) p(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \gamma_i = l')}$$

[0071]

[0072] state transition probability $P(\gamma_i = l|\gamma_{i-1} = l')$을 사용하면, $P(\gamma_i = l|\mathbf{y}_{1:i-1})$는 수학식 21과 같이 나타낼 수 있다.

수학식 21

$$P(\gamma_i = l|\mathbf{y}_{1:i-1}) = \sum_{l'=1}^{N_\gamma} P(\gamma_i = l|\gamma_{i-1} = l') P(\gamma_{i-1} = l'|\mathbf{y}_{1:i-1})$$

[0073]

[0074] 그리고, $p(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \gamma_i = l)$은 HMM의 observation distribution $p(\mathbf{y}_i|\gamma_i = l)$에 의해 획득될 수 있다.

[0075] 수학식 12를 사용하면, 향상된 LMPSCs는 수학식 22에 의해 결정될 수 있다.

수학식 22

$$\hat{s}_{i,k} = \sum_{l=1}^{N_\gamma} P(\gamma_i = l|\mathbf{y}_{1:i}) \hat{\mu}_{s_{i,k}|\gamma_i=l}$$

[0076]

[0077] 이하, 전술한 구성을 갖는 본 발명에 따른 강인한 음성 인식 방법에 대하여 성능을 살펴보기 위하여 임의의 조건에서 시뮬레이션하여, 종래의 다른 방법들과 비교분석하였다.

[0078] 도 1은 본 발명에 따른 음성 인식 방법과 종래의 음성 인식 방법들의 성능을 분석하기 위하여, 구성된 소스와 마이크로폰들을 도시한 구성도이다.

[0079] 도 1을 참조하면, Target speech, 간섭신호원 (Interference) 1, 2, 3 및 2개의 마이크로 폰이 설치되어 있다. 도 1의 시뮬레이션 환경에서, 간섭신호원이 1개인 경우, 2개인 경우, 3개인 경우에 대하여, 각각 종래방법 1(Baseline), 종래방법 2(FE), 종래 방법 3(IVA), 종래 방법 4(BSSA) 및 본 발명에 따른 방법(IVA-FE)에서의 SNR을 측정하였으며, 그 결과를 도 2에 도시하였다. 여기서, 종래 방법 1은 마이크로폰으로부터 입력된 신호로부터 단어를 인식하는 방법이며, 종래 방법 2는 종래의 특징 향상(Feature Enhancement) 방법을 이용하여 음성 인식하는 것이며, 종래 방법 3은 독립 벡터 분석하여 음성 인식하는 것이며, 종래 방법 4는 BSSA 에 의해 음성

인식하는 것이다.

[0080] 도 2를 참조하면, 모든 경우에서 본 발명에 따른 음성 인식 방법이 종래의 다른 방법들보다 SNR이 월등히 우수함을 쉽게 알 수 있다.

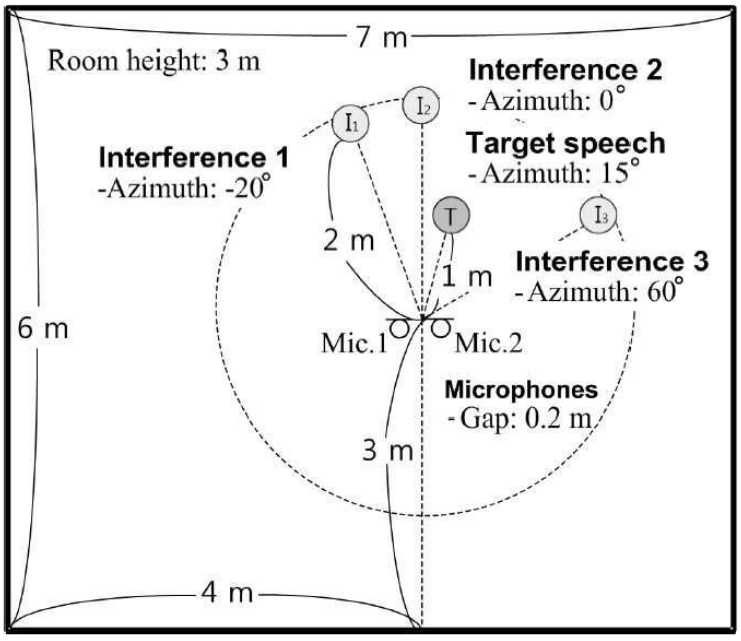
[0081] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특징을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

산업상 이용가능성

[0082] 본 발명에 따른 방법은 음성 인식 시스템에 널리 사용될 수 있다.

도면

도면1



도면2

Number of Interfs.	Method	Input SNR (dB)							
		At a RT_{60} of 0.35 s				At a RT_{60} of 0.45 s			
		0	5	10	15	0	5	10	15
One	Baseline	30.41	45.73	62.93	75.34	28.32	44.04	58.66	71.44
	Conv. FE	31.60	49.63	67.54	79.45	29.61	46.96	63.82	75.37
	IVA	68.83	80.62	87.07	90.29	55.80	73.22	80.22	85.26
	BSSA	35.84	53.99	66.09	74.02	28.59	46.22	57.77	66.74
	IVA-FE	**81.08**	**90.29**	**93.61**	**94.78**	**73.83**	**87.56**	**90.76**	**91.58**
Two	Baseline	24.66	42.51	60.90	76.66	23.96	39.71	56.57	71.10
	Conv. FE	26.14	47.91	68.34	80.37	24.36	46.04	63.64	76.44
	IVA	54.12	72.91	83.63	88.70	45.64	65.66	77.46	84.61
	BSSA	29.08	46.35	61.86	71.71	25.43	41.19	54.61	68.40
	IVA-FE	**70.55**	**86.61**	**92.75**	**94.41**	**67.54**	**83.23**	**89.53**	**91.77**
Three	Baseline	24.08	43.77	63.45	78.56	21.84	38.97	56.91	71.90
	Conv. FE	22.76	44.87	65.08	79.76	21.87	42.51	62.29	75.00
	IVA	26.60	46.28	70.79	85.87	22.94	48.80	70.64	83.14
	BSSA	19.81	42.75	60.63	73.13	17.84	37.78	57.46	69.32
	IVA-FE	**29.36**	**55.62**	**78.19**	**92.14**	**29.76**	**64.50**	**83.60**	**91.55**