(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년11월29일
(11) 등록번호 10-1802444
(24) 등록일자 2017년11월22일

(51) 국제특허분류(Int. Cl.)
G10L 15/20 (2006.01) *G10L 15/14* (2006.01)
G10L 19/02 (2006.01) *G10L 19/035* (2013.01)
G10L 21/0208 (2013.01)
(52) CPC특허분류
G10L 15/20 (2013.01)
G10L 15/14 (2013.01)
(21) 출원번호 10-2016-0089966
(22) 출원일자 2016년07월15일
심사청구일자 2016년07월15일
(56) 선행기술조사문헌
Ji-won Cho et al., 'Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition', Signal Processing, Vol.120, pp.200~208, March 2016*
Ji-won Cho et al., 'An efficient HMM-based feature enhancement method with filter estimation for reverberant speech recognition', IEEE Signal Processing Letters, Vol.20, No.12, December 2013.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
박형민
서울특별시 서초구 잠원로 150 5동 903호 (잠원동,잠원한신아파트)
조지원
서울특별시 마포구 독막로32안길 15 (신수동)
(74) 대리인
이지연

전체 청구항 수 : 총 9 항 심사관 : 정성윤

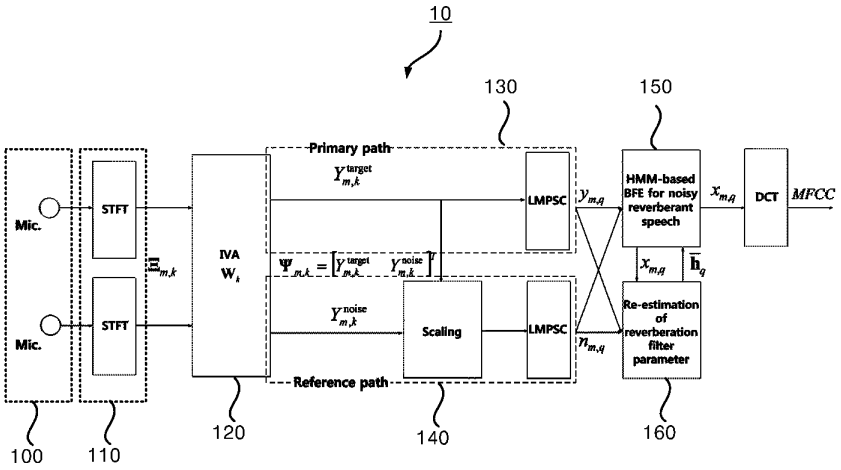
(54) 발명의 명칭 **독립 벡터 분석 및 반향 필터 재추정을 이용한 베이시안 특징 향상에 의한 강인한 음성 인식 장치 및 방법**

(57) 요 약

본 발명은 독립벡터분석 및 재추정된 반향필터파라메터를 이용한 베이시안 특징 향상시킨 음성 인식 장치 및 방법에 관한 것이다. 상기 음성 인식 방법은, (a) 외부로부터 입력된 복수 개의 음성 신호들을 단구간 푸리에 변환하여 각각 주파수 영역의 신호로 변환하여 출력하는 단계; (b) 상기 주파수 영역의 음성 신호들을 독립 벡터 분

(뒷면에 계속)

대 표 도



석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계; (c) 상기 독립벡터분석에 의해 추정된 IVA 타겟 음성 신호로부터 HMM-based BFE 하여 음성 특징을 추출하는 단계; (d) 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석에 의해 추정된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 단계; (e) 상기 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 HMM-based BFE 하여 음성 특징을 강화시켜 초기 음원 신호를 추정하는 단계; (f) 상기 노이즈 특징과 상기 추정된 초기 음원 신호를 이용하여 반향 필터 파라메터를 재추정하는 단계; (g) 상기 재추정된 반향 필터 파라메터를 이용하여 상기 음성 특징을 다시 강화시켜 음원 신호를 최종 추정하는 단계; 를 구비한다.

(52) CPC특허분류

G10L 19/02 (2013.01)
G10L 19/035 (2013.01)
G10L 21/0208 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711037581 (NRF-2014R1A2A2A01006581)
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 이공분야기초연구사업 중견연구자지원사업
연구과제명 시청각정보를 이용한 강인한 멀티모달 음성인식 기술 개발
기 여 율 1/1
주관기관 서강대학교 산학협력단
연구기간 2014.05.01 ~ 2017.04.30

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

외부로부터 음성 신호가 입력되는 음성 신호 입력부;

상기 음성 신호 입력부로 입력된 다수 개의 음성 신호들을 각각 주파수 영역의 신호로 변환하여 출력하는 푸리에 변환 모듈;

상기 푸리에 변환 모듈로부터 출력된 주파수 영역의 복수 개의 음성 신호들을 독립 벡터 분석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 독립벡터분석 모듈;

상기 독립벡터분석 모듈로부터 출력된 IVA 타겟 음성 신호로부터 음성 특징을 추출하는 목적 음성 강화 모듈;

상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석 모듈로부터 출력된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 목적 음성 제거 모듈;

상기 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터를 이용하여 음성 특징을 강화시켜 음원 신호를 추정하는 HMM 기반 특징 강화 모듈;

상기 목적 음성 제거 모듈로부터 제공된 노이즈 특징과 상기 HMM 기반 특징 강화 모듈로부터 제공된 추정된 음원 신호를 이용하여 반향 필터 파라메터를 재추정하여 상기 HMM 기반 특징 강화 모듈로 제공하는 반향 필터 재추정부;

를 구비하고, 상기 HMM 기반 특징 강화 모듈은 반향 필터 재추정부에 의해 재추정된 반향 필터 파라메터와 상기 강화된 음성 특징을 이용하여 음원 신호를 최종 추정하여 출력하는 것을 특징으로 하는 음성 인식 장치.

청구항 2

제1항에 있어서, 상기 HMM 기반 특징 강화 모듈은 HMM-based BFE 방법을 이용하여 음성 특징을 강화시키는 것을 특징으로 하는 음성 인식 장치.

청구항 3

제1항에 있어서, 상기 목적 음성 강화 모듈은 멜 스케일 필터 뱅크(Mel-scale filter bank)로 구성되며,

상기 목적 음성 강화 모듈에 의해 추출된 IVA 타겟 음성 신호에 대한 음성 특징은 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 하는 음성 인식 장치.

청구항 4

제1항에 있어서, 상기 목적 음성 제거 모듈에 의해 추출된 IVA 노이즈 신호에 대한 노이즈 특징은 IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 하는 음성 인식 장치.

청구항 5

제1항에 있어서, 상기 HMM 기반 특징 강화 모듈은 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 음성 특징을 강화시켜 초기 음원 신호를 추정하고, 추정된 초기 음원 신호를 상기 반향 필터 재추정 모듈로 제공한 후 상기 반향 필터 재추정 모듈로부터 재추정된 반향 필터 파라메터를 제공받고, 상기 재추정된 반향 필터 파라메터와 상기 강화된 음성 특징을 이용하여 음원 신호를 최종 추정하여 출력하는 것을 특징으로 하는 음성 인식 장치.

청구항 6

(a) 외부로부터 복수 개의 음성 신호들을 입력받는 단계;

(b) 상기 입력된 복수 개의 음성 신호들을 단구간 푸리에 변환하여 각각 주파수 영역의 신호로 변환하여 출력하는 단계;

(c) 상기 주파수 영역의 음성 신호들을 독립 벡터 분석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계;

(d) 상기 독립벡터분석에 의해 추정된 IVA 타겟 음성 신호로부터 음성 특징을 추출하는 단계;

(e) 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석에 의해 추정된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 단계;

(f) 상기 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 음성 특징을 강화시켜 초기 음원 신호를 추정하는 단계;

(g) 상기 노이즈 특징과 상기 추정된 초기 음원 신호를 이용하여 반향 필터 파라메터를 재추정하는 단계;

(h) 상기 재추정된 반향 필터 파라메터를 이용하여 상기 음성 특징을 다시 강화시켜 음원 신호를 최종 추정하는 단계;

를 구비하는 음성 인식 방법.

청구항 7

제6항에 있어서, 상기 (d) 단계에서 추출된 음성 특징은, 멜 스케일 필터 뱅크(Mel-scale filter bank)를 이용하여, 상기 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 하는 음성 인식 방법.

청구항 8

제6항에 있어서, 상기 (e) 단계에서 추출된 노이즈 특징은, 멜 스케일 필터 뱅크(Mel-scale filter bank)를 이용하여, IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 하는 음성 인식 방법.

청구항 9

제6항에 있어서, 상기 (f) 단계 및 (h) 단계는 HMM-based BFE 방법을 이용하여 음성 특징을 강화시키는 것을 특징으로 하는 음성 인식 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 강인한 음성 인식 장치 및 방법에 관한 것으로서, 더욱 구체적으로는 노이즈와 반향이 모두 존재하는 실제 환경에서 독립 벡터 분석과 반향 필터 파라메터 재추정을 이용하여 은닉 마르코프 모델(hidden Markov models ; HMM) 기반의 베이시안 특징 향상(BFE)시켜 인식 오류를 감소시킬 수 있도록 한 강인한 음성 인식 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 음성 인식 시스템에 있어서, 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] 또한, 음성 인식 시스템에 있어서 그 환경에 따라 반향도 많이 발생하게 되며, 이러한 반향에 의해 음성 인식

성능이 감소된다. 이와 같이, 음성 인식 시스템(ASR System)에 있어서, 실세계의 환경은 노이즈(additive noise) 뿐만 아니라 반향(reverberation)들이 함께 존재하므로 이러한 실세계의 환경에서 음성 인식 성능을 향상시키기 위해서는 노이즈 및 반향을 모두 고려하여야 할 뿐만 아니라 동시에 이들을 고려하여야 한다.

[0004] 하지만, 대부분의 음성 인식 시스템은 노이즈만을 고려하거나, 반향만을 주로 고려하거나, 노이즈와 반향을 순차적으로 고려하여 음성 인식하게 된다. 이 경우 음성 인식의 성능이 매우 낮아지고 그 결과 인식 오류가 발생하게 된다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국등록특허공보 제 10-1506547호

(특허문헌 0002) 한국등록특허공보 제 10-0329596호

(특허문헌 0003) 한국등록특허공보 제 10-1361034호

발명의 내용

해결하려는 과제

[0006] 전술한 문제점을 해결하기 위한 본 발명의 목적은 노이즈와 반향이 많은 환경에서 노이즈와 반향을 동시에 고려하기 위하여, 독립 벡터 분석과 반향 필터 파라메터 재추정을 이용하여 은닉 마르코프 모델 기반의 베이시안 특징 향상(HMM-based BFE)시킴으로써, 성능이 우수한 음성 인식 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0007] 전술한 기술적 과제를 달성하기 위한 본 발명의 제1 특징에 따른 음성 인식 장치는, 외부로부터 음성 신호가 입력되는 음성 신호 입력부; 상기 음성 신호 입력부로 입력된 다수 개의 음성 신호들을 각각 주파수 영역의 신호로 변환하여 출력하는 푸리에 변환 모듈; 상기 푸리에 변환 모듈로부터 출력된 주파수 영역의 복수 개의 음성 신호들을 독립 벡터 분석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 독립벡터분석 모듈; 상기 독립벡터분석 모듈로부터 출력된 IVA 타겟 음성 신호로부터 음성 특징을 추출하는 목적 음성 강화 모듈; 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석 모듈로부터 출력된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 목적 음성 제거 모듈; 상기 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터를 이용하여 음성 특징을 강화시켜 음원 신호를 추정하는 HMM 기반 특징 강화 모듈; 상기 목적 음성 제거 모듈로부터 제공된 노이즈 특징과 상기 HMM 기반 특징 강화 모듈로부터 제공된 추정된 음원 신호를 이용하여 반향 필터 파라메터를 재추정하여 상기 HMM 기반 특징 강화 모듈로 제공하는 반향 필터 재추정부; 를 구비하고, 상기 HMM 기반 특징 강화 모듈은 반향 필터 재추정부에 의해 재추정된 반향 필터 파라메터와 상기 강화된 음성 특징을 이용하여 음원 신호를 최종 추정하여 출력한다.

[0008] 전술한 제1 특징에 따른 음성 인식 장치에 있어서, 상기 HMM 기반 특징 강화 모듈은 HMM-based BFE 방법을 이용하여 음성 특징을 강화시키는 것이 바람직하다.

[0009] 전술한 제1 특징에 따른 음성 인식 장치에 있어서, 상기 목적 음성 강화 모듈은 멜-스케일 필터 뱅크(Mel-scale filter bank)로 구성되며, 상기 목적 음성 강화 모듈에 의해 추출된 IVA 타겟 음성 신호에 대한 음성 특징은 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0010] 전술한 제1 특징에 따른 음성 인식 장치에 있어서, 상기 목적 음성 제거 모듈에 의해 추출된 IVA 노이즈 신호에 대한 노이즈 특징은 IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0011] 전술한 제1 특징에 따른 음성 인식 장치에 있어서, 상기 HMM 기반 특징 강화 모듈은 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 음성 특징을 강화시켜 초기 음원 신호를 추정하고, 추정된 초기 음원 신호를 상기 반향 필터 재추정 모듈로 제공한 후 상기 반향 필터 재추정 모듈로부터

재추정된 반향 필터 파라메터를 제공받고, 상기 재추정된 반향 필터 파라메터와 상기 강화된 음성 특징을 이용하여 음원 신호를 최종 추정하여 출력하는 것이 바람직하다.

[0012] 본 발명의 제2 특징에 따른 음성 인식 방법은, (a) 외부로부터 복수 개의 음성 신호들을 입력받는 단계; (b) 상기 입력된 복수 개의 음성 신호들을 단구간 푸리에 변환하여 각각 주파수 영역의 신호로 변환하여 출력하는 단계; (c) 상기 주파수 영역의 음성 신호들을 독립 벡터 분석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계; (d) 상기 독립벡터분석에 의해 추정된 IVA 타겟 음성 신호로부터 음성 특징을 추출하는 단계; (e) 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석에 의해 추정된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 단계; (f) 상기 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 음성 특징을 강화시켜 초기 음원 신호를 추정하는 단계; (g) 상기 노이즈 특징과 상기 추정된 초기 음원 신호를 이용하여 반향 필터 파라메터를 재추정하는 단계; (h) 상기 재추정된 반향 필터 파라메터를 이용하여 상기 음성 특징을 다시 강화시켜 음원 신호를 최종 추정하는 단계; 를 구비한다.

[0013] 전술한 제2 특징에 따른 음성 인식 방법에 있어서, 상기 (d) 단계에서 추출된 음성 특징은, 멜-스케일 필터 뱅크(Mel-scale filter bank)를 이용하여, 상기 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0014] 전술한 제2 특징에 따른 음성 인식 방법에 있어서, 상기 (e) 단계에서 추출된 노이즈 특징은, 멜-스케일 필터 뱅크(Mel-scale filter bank)를 이용하여, IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0015] 전술한 제2 특징에 따른 음성 인식 방법에 있어서, 상기 (f) 단계 및 (h) 단계는 HMM-based BFE 방법을 이용하여 음성 특징을 강화시키는 것이 바람직하다.

발명의 효과

[0016] 본 발명에 따른 음성 인식 장치 및 방법은 강인한 음성 인식을 수행하게 된다.

[0017] 도 2는 본 발명에 따른 음성 인식 방법과 종래의 음성 인식 방법들의 성능을 분석하기 위하여, 구성된 소스와 마이크로폰들을 도시한 구성도이다. 도 3은 도 2의 시뮬레이션 환경에서, 간섭신호원이 1개인 경우, 2개인 경우, 3개인 경우에 대하여, 각각 종래방법 1(Baseline), 종래방법 2(BFE-SNS+CRP), 종래 방법 3(BFE-IVA+CRP), 종래 방법 4(BFE-SNS+RPR) 및 본 발명에 따른 방법(BFE-IVA+RPR)에서의 SNR을 측정하여 도시한 도표이다. 여기서, 종래방법 1은 BFE-IVA 및 반향필터 재추정을 하지 않은 기본적인 방법이며, 종래방법 2는 종래의 stationary noise statistics를 이용한 방법 및 종래의 성김 모델 기반의 반향필터 파라메터를 이용한 방법이며, 종래방법 3은 IVA 기반의 노이즈 추정을 이용한 BFE 방법과 종래의 성김 모델 기반의 반향필터 파라메터를 이용한 방법이며, 종래방법 4는 종래의 stationary noise statistics를 이용한 방법 및 반향 필터 파라메터 재추정을 이용한 방법이며, 본 발명은 IVA 기반의 노이즈 추정을 이용한 BFE 방법과 반향 필터 파라메터 재추정을 이용한 방법이다. 도 3을 참조하면, 본 발명에 따른 음성 인식 방법은 다른 종래의 방법들보다 데이터 인식에 대한 오류가 가장 작음을 알 수 있다.

[0018] 도 4는 도 2의 시뮬레이션 환경에서, 간섭신호원이 1개인 경우, 2개인 경우, 3개인 경우에 대하여, 각각 종래방법 1(Baseline), 종래방법 2(BFE-SNS+CRP), 종래 방법 5(BFE-IVA only), 종래 방법 6(BFE-IVA+BFE-RPR) 및 본 발명에 따른 방법(BFE-IVA+RPR)에서의 SNR을 측정하여 도시한 도표이다. 종래방법 5는 IVA 기반의 노이즈 추정을 이용한 BFE 방법만을 적용한 것이며, 종래방법 6은 IVA 기반의 노이즈 추정을 이용한 BFE 방법과 IVA 기반의 노이즈 추정을 이용한 BFE 방법을 순차적으로 적용한 것이다. 도 4를 참조하면, 본 발명에 따른 음성 인식 방법은 다른 종래의 방법들보다 데이터 인식에 대한 오류가 가장 작음을 알 수 있다. 특히, 종래 방법 6은 독립벡터분석 및 반향 필터 재추정을 순차적으로 수행한 것으로서, 이 경우보다도 본 발명에 따라 독립벡터분석 및 반향 필터 재추정을 동시에 수행한 음성 인식이 인식 오류가 감소됨을 쉽게 파악할 수 있다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 바람직한 실시예에 따른 음성 인식 장치를 전체적으로 도시한 블록도이다.

도 2는 본 발명에 따른 음성 인식 방법과 종래의 음성 인식 방법들의 성능을 분석하기 위하여, 구성된 소스와 마이크로폰들을 도시한 구성도이다.

도 3 및 도 4는 도 2의 시뮬레이션 환경에서, 간섭신호원이 1개인 경우, 2개인 경우, 3개인 경우에 대하여, 각

각 종래방법 1, 종래방법 2, 종래 방법 3, 종래 방법 4, 종래방법 5, 종래방법 6 및 본 발명에 따른 방법에서의 SNR을 측정하여 도시한 도표들이다.

도 5 및 도 6은 TIMIT 데이터베이스에 포함된 발음들에 의해 오류가 발생된 "five two one zero nine"의 발음에 대한 LMPSCs 이다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명에 따른 음성 인식 장치 및 방법은, 노이즈와 반향이 모두 존재하는 실제 환경에서 독립 벡터 분석과 반향 필터 파라메터 재추정을 이용하여 베이시안 특징 향상시켜 인식 오류를 감소시킬 수 있도록 한 것을 특징으로 한다.

[0021] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 강인한 음성 인식 장치 및 방법에 대하여 구체적으로 설명한다. 도 1은 본 발명의 바람직한 실시예에 따른 음성 인식 장치를 전체적으로 도시한 블록도이다. 도 1을 참조하면, 본 발명의 바람직한 실시예에 따른 음성 인식 장치(10)는 음성 신호 입력부(100), 푸리에 변환 모듈(110), 독립벡터분석 모듈(120), 목적 음성 강화 모듈(130), 목적 음성 제거 모듈(140), HMM 기반 특징 강화 모듈(150), 및 반향 필터 재추정 모듈(160)을 구비한다. 전술한 본 발명에 따른 음성 인식 장치는 노이즈와 반향이 많은 환경에서 노이즈와 반향을 동시에 고려하기 위하여, 독립 벡터 분석과 반향 필터 파라메터 재추정을 이용하여 HMM 기반으로 한 베이시안 특징 향상시킴으로써, 실세계에서 강인한 음성 인식을 수행할 수 있게 된다. 이하, 전술한 각 구성요소들에 대하여 구체적으로 설명한다.

[0022] 상기 음성 신호 입력부(100)는 외부로부터 음성 신호를 입력받는 복수 개의 음성 신호 입력 장치들, 예컨대 마이크 등으로 이루어지며, 입력된 음성 신호들은 푸리에 변환 모듈(110)로 제공된다.

[0023] 상기 푸리에 변환 모듈(110)은 상기 음성 신호 입력부의 각 음성 신호 입력 장치들로 입력된 음성 신호들을 각각 주파수 영역의 신호로 변환하여 출력하며, 특히 단구간 푸리에 변환(short-time Fourier transform ;'STFT')시키는 것이 바람직하다.

[0024] 상기 독립벡터분석 모듈(120)은 상기 푸리에 변환 모듈로부터 출력된 복수 개의 음성 신호들을 독립 벡터 분석(Independent Vector Analysis; 이하, 'IVA'라 한다)하여 IVA 타겟 음성 신호($Y_{m,k}^{\text{target}}$) 및 IVA 노이즈 신호($Y_{m,k}^{\text{noise}}$)를 추정하여 각각 목적 음성 강화 모듈(130) 및 목적 음성 제거 모듈(140)로 제공한다.

[0025] 여기서, 벡터 $\boldsymbol{\Xi}_{m,k}$ 는, 상호 독립된 N개의 미지의 음원 소스들의 음성이 혼합된 벡터로서, m번째 프레임이며 k 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

[0026] $$\boldsymbol{\Psi}_{m,k} = \mathbf{W}_k \boldsymbol{\Xi}_{m,k}$$

[0027] 여기서, $\boldsymbol{\Psi}_{m,k} = [Y_{m,k}^1, \cdots, Y_{m,k}^{N_S}]^T$ 은 추정된 소스 신호들의 시간-주파수 표현으로 구성된 벡터이며, $\mathbf{W}_k$ 는 k 번째 주파수 빈에서의 분리 매트릭스(separating matrix)이다.

[0028] 분리 매트릭스를 추정하기 위한 자연-경사도 IVA(natural-gradient IVA) 학습 규칙은 수학식 2에 의해 구할 수 있다.

수학식 2

$$\Delta \mathbf{W}_k \propto \{\mathbf{I} - E[\mathbf{\Phi}_k(\mathbf{Y}_m^{1:N_S})\mathbf{\Psi}_{m,k}^H]\}\mathbf{W}_k$$

[0029]

[0030] 여기서, $\mathbf{I}$ 는 아이덴티티 매트릭스(Identity matrix)이다. $\mathbf{Y}_m^n = [Y_{m,1}^n, \cdots, Y_{m,K}^n]^T$ 을 갖는 추정된 음원들에 대한 hypothesized pdf 모델들 $q(\mathbf{Y}_m^n) \propto \exp\left(-\sqrt{\sum_{k=1}^K |Y_{m,k}^n|^2}\right)$ 를 가정하면, $\mathbf{\Phi}_k(\mathbf{Y}_m^{1:N_S})$ 는 $\mathbf{\Phi}_k(\mathbf{Y}_m^{1:N_S}) = [\phi_k(\mathbf{Y}_m^1), \cdots, \phi_k(\mathbf{Y}_m^{N_S})]^T$ 를 나타내는 multivariate score function values이며, 여기서, $\phi_k(\mathbf{Y}_m^n) = -\partial \log q(\mathbf{Y}_m^n)/\partial Y_{m,k}^n = Y_{m,k}^n / \sqrt{\sum_{k'=1}^K |Y_{m,k'}^n|^2},\ 1 \le n \le N_S$ 이며, 주파수 빈들의 개수는 K이다.

[0031] 상기 목적 음성 강화 모듈(130)은 상기 독립벡터분석 모듈로부터 출력된 IVA 타겟 음성 신호로부터 음성 특징($y_{m,q}$)을 추출하여 출력한다. 상기 목적 음성 강화 모듈은 멜-스케일 필터 뱅크(Mel-scale filter bank)로 구성되며, 상기 목적 음성 강화 모듈에 의해 추출된 IVA 타겟 음성 신호에 대한 음성 특징은 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 한다.

[0032] 여기서, 음성 특징($y_{m,q}$)은 수학식 3으로 표현될 수 있다.

수학식 3

$$\begin{aligned} y_{m,q} &= g(\bar{\mathbf{h}}_q, n_{m,q}, v_{m,q}) \\ &\approx g(\mathbf{b}_{\bar{\mathbf{h}}_q}, b_{n_{m,q}}, b_{v_{m,q}}) + \mathbf{d}_{\bar{\mathbf{h}}_q}^T(\bar{\mathbf{h}}_q - \mathbf{b}_{\bar{\mathbf{h}}_q}) \\ &\quad + d_{n_{m,q}}(n_{m,q} - b_{n_{m,q}}) + d_{v_{m,q}}(v_{m,q} - b_{v_{m,q}}) \end{aligned}$$

[0033]

[0034] 여기서, $\bar{\mathbf{h}}_q$ 및 $\mathbf{b}_{\bar{\mathbf{h}}_q}$ 는 각각 벡터 $[\bar{h}_{0,q}, \cdots, \bar{h}_{L_H-1,q}]^T$ 및 $[b_{\bar{h}_{0,q}}, \cdots, b_{\bar{h}_{L_H-1,q}}]^T$ 로 표현된다. 그리고 수학식 4를 나타낸다.

수학식 4

$$\mathbf{d}_{\bar{\mathbf{h}}_q} = \left[\frac{\partial f}{\partial \bar{h}_{0,q}}\Big|_{\bar{h}_{0,q}=b_{\bar{h}_{0,q}}}, \frac{\partial f}{\partial \bar{h}_{1,q}}\Big|_{\bar{h}_{1,q}=b_{\bar{h}_{1,q}}}, \cdots, \frac{\partial f}{\partial \bar{h}_{L_H-1,q}}\Big|_{\bar{h}_{L_H-1,q}=b_{\bar{h}_{L_H-1,q}}} \right]^T$$

$$= \begin{bmatrix} \frac{\exp(x_{m,q}+b_{\bar{h}_{0,q}})}{\exp(x_{m,q}+b_{\bar{h}_{0,q}})+\sum_{m'=1}^{L_H-1}\exp(\hat{x}_{m-m',q}+b_{\bar{h}_{m',q}})+\exp(n_{m,q})} \\ \frac{\exp(\hat{x}_{m-1,q}+b_{\bar{h}_{1,q}})}{\exp(x_{m,q}+b_{\bar{h}_{0,q}})+\sum_{m'=1}^{L_H-1}\exp(\hat{x}_{m-m',q}+b_{\bar{h}_{m',q}})+\exp(n_{m,q})} \\ \vdots \\ \frac{\exp(\hat{x}_{m-L_H+1,q}+b_{\bar{h}_{L_H-1,q}})}{\exp(x_{m,q}+b_{\bar{h}_{0,q}})+\sum_{m'=1}^{L_H-1}\exp(\hat{x}_{m-m',q}+b_{\bar{h}_{m',q}})+\exp(n_{m,q})} \end{bmatrix}$$

[0035]

[0036] $b_{n_{m,q}} = \mu_{n_{m,q}}$ 및 $b_{v_{m,q}} = \mu_{v_q}$ 일 때, $\bar{\mathbf{h}}_q$ 을 갖는 new posterior pdf는 수학식 5로 표현될 수 있다.

수학식 5

$$\begin{aligned} & p(\bar{\mathbf{h}}_q|\mathbf{y}_{1:m}) \\ & \propto \; p(y_{m,q}|\bar{\mathbf{h}}_q)\; p(\bar{\mathbf{h}}_q) \\ & \approx \; \mathcal{N}\left(y_{m,q}|g(\mathbf{b}_{\bar{\mathbf{h}}_q}, \mu_{n_{m,q}}, \mu_{v_q}) + \mathbf{d}_{\bar{\mathbf{h}}_q}^T(\bar{\mathbf{h}}_q - \mathbf{b}_{\bar{\mathbf{h}}_q}), d_{n_{m,q}}^2\sigma_{n_{m,q}}^2 + \sigma_{v_q}^2\right) \\ & \quad \cdot \mathcal{N}(\bar{\mathbf{h}}_q|\boldsymbol{\mu}_{\bar{\mathbf{h}}_q}, \boldsymbol{\Sigma}_{\bar{\mathbf{h}}_q}) \end{aligned}$$

[0037]

[0038] 여기서, prior distribution $p(\bar{\mathbf{h}}_q)$ 는 mean vector $\boldsymbol{\mu}_{\bar{\mathbf{h}}_q}$ 및 diagonal covariance matrix $\boldsymbol{\Sigma}_{\bar{\mathbf{h}}_q}$을 갖는 Gaussian 인 것으로 가정된다.

[0039] 따라서, $\bar{\mathbf{h}}_q$ 은 수학식 6으로 나타낼 수 있는 $\bar{\mathbf{h}}_q$ 에 대한 posterior pdf의 gradient of the logarithm을 사용하여 업데이트시킬 수 있다.

수학식 6

$$\begin{aligned} \Delta\bar{\mathbf{h}}_q \propto & -\left(\frac{1}{d_{n_{m,q}}^2\sigma_{n_{m,q}}^2 + \sigma_{v_q}^2}\mathbf{d}_{\bar{\mathbf{h}}_q}\mathbf{d}_{\bar{\mathbf{h}}_q}^T + \boldsymbol{\Sigma}_{\bar{\mathbf{h}}_q}^{-1}\right)\bar{\mathbf{h}}_q \\ & + \frac{1}{d_{n_{m,q}}^2\sigma_{n_{m,q}}^2 + \sigma_{v_q}^2}\left(y_{m,q} - g(\mathbf{b}_{\bar{\mathbf{h}}_q}, \mu_{n_{m,q}}, \mu_{v_q}) + \mathbf{d}_{\bar{\mathbf{h}}_q}^T\mathbf{b}_{\bar{\mathbf{h}}_q}\right)\mathbf{d}_{\bar{\mathbf{h}}_q} \\ & + \boldsymbol{\Sigma}_{\bar{\mathbf{h}}_q}^{-1}\boldsymbol{\mu}_{\bar{\mathbf{h}}_q}. \end{aligned}$$

[0040]

[0041] 상기 업데이트는 convergence 가 될 때까지 반복된다. 업데이팅 알고리즘은 수학식 3의 the first-order Taylor series expansion을 사용한 linearization과 연관되어 있기 때문에, $\mathbf{b}_{\bar{\mathbf{h}}_q}$는 모든 프레임들에서 업데이트된 $\bar{\mathbf{h}}_q$ 에 의해 업데이트되어야만 한다.

[0042] 상기 목적 음성 제거 모듈(140)은 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석 모듈로부터 출력된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징($n_{m,q}$)을 추출하여 출력한다. 상기 목적 음성 제거 모듈에 의해 추출된 IVA 노이즈 신호에 대한 노이즈 특징은 IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것을 특징으로 한다.

[0043] 상기 HMM 기반 특징 강화 모듈(150)은 상기 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터를 이용하여 HMM 기반 베이시안 특징을 향상시킨 음원 신호($x_{m,q}$)를 추정하여 출력한다. 상기 HMM 기반 특징 강화 모듈은 먼저 목적 음성 강화 모듈로부터 제공된 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 HMM-based BFE하여 음성 특징을 강화시켜 초기 음원 신호를 추정하고, 상기 추정된 초기 음원 신호를 상기 반향 필터 재추정 모듈(160)로 제공한다. 다음, 상기 반향 필터 재추정 모듈로부터 재추정된 반향 필터 파라메터를 제공받고, 상기 재추정된 반향 필터 파라메터와 상기 음원 신호를 이용하여 상기 음성 특징을 다시 강화시키고, 상기 강화된 음성 특징을 이용하여 음원 신호($x_{m,q}$)를 최종 추정하여 출력한다.

[0044] 상기 반향 필터 재추정 모듈(160)은 상기 목적 음성 제거 모듈(140)로부터 제공된 노이즈 특징($n_{m,q}$)과 상기 HMM 기반 특징 강화 모듈(150)로부터 제공된 추정된 음원 신호($x_{m,q}$)를 이용하여 반향 필터 파라메터를 재추정하고, 재추정된 반향 필터 파라메터($\bar{\mathbf{h}}_q$)를 상기 HMM 기반 특징 강화 모듈(150)로 제공한다.

[0045] 도 5 및 도 6은 TIMIT 데이터베이스에 포함된 발음들에 의해 오류가 발생된 "five two one zero nine"의 발음에 대한 LMPSCs 이다. 도 5 및 도 6은 도 2의 Imterference 1 및 2에 배치된 2개의 간섭 음원을 포함하는 혼합 환경에 의해 생성된 noisy reverberant speech에 대한 것으로서, RT_{60}은 0.45s 이며, "Mic.1"에서의 Input SIR은 5dB이며, BFE 방법들은 noisy reverberant speech에 대한 observation model에 기반된다.

[0046] 도 5의 (a)는 "five two one zero nine"을 발음하는 clean speech의 LMPSC 이며, (b)는 "oh one"으로 잘못 인식된 "mic.1"에서의 noisy revererant speech의 LMPSC 이며, (c)는 "four eight one nine nine"으로 잘못 인식된 BFE-SNS+CRP에 의해 향상된 LMPSC 이며, (d)는 "five two nine zero nine"으로 잘못 인식된 BFE-IVA+CRP에 의해 향상된 LMPSC이며, (e) "five six one zero nine"으로 잘못 인식된 BFE-SNS+RPR에 의해 향상된 LMPSC이며, (f)"five two one zero nine"으로 정확하게 인식된 본 발명에 따른 BFE-IVA+RPR에 의해 향상된 LMPSC 이다.

[0047] 도 6의 (a)는 "five two one zero nine"을 발음하는 clean speech의 LMPSC 이며, (b)는 "oh one"으로 잘못 인식된 "mic.1"에서의 noisy revererant speech의 LMPSC 이며, (c)는 "four eight one nine nine"으로 잘못 인식된 BFE-SNS+CRP에 의해 향상된 LMPSC 이며, (d)는 "eight one zero nine"으로 잘못 인식된 BFE-IVA 만으로 향상된 LMPSC이며, (e) "five eight nine zero nine"으로 잘못 인식된 BFE-IVA+BFE-RPR에 의해 향상된 LMPSC이며, (f)"five two one zero nine"으로 정확하게 인식된 본 발명에 따른 BFE-IVA+RPR에 의해 향상된 LMPSC 이다.

[0048] 전술한 구성을 갖는 본 발명에 따른 음성 인식 장치에 의한 음성 인식 방법은, (a) 외부로부터 복수 개의 음성 신호들을 입력받는 단계; (b) 상기 입력된 복수 개의 음성 신호들을 단구간 푸리에 변환하여 각각 주파수 영역의 신호로 변환하여 출력하는 단계; (c) 상기 주파수 영역의 음성 신호들을 독립 벡터 분석하여 IVA 타겟 음성 신호와 IVA 노이즈 신호를 추정하는 단계; (d) 상기 독립벡터분석에 의해 추정된 IVA 타겟 음성 신호로부터 음성 특징을 추출하는 단계; (e) 상기 IVA 타겟 음성 신호를 이용하여 상기 독립벡터분석에 의해 추정된 IVA 노이즈 신호를 스케일링한 후 스케일링된 IVA 노이즈 신호로부터 노이즈 특징을 추출하는 단계; (f) 상기 음성 특징 및 반향 필터 파라메터의 초기 설정값을 이용하여 음성 특징을 강화시켜 초기 음원 신호를 추정하는 단계; (g) 상기 노이즈 특징과 상기 추정된 초기 음원 신호를 이용하여 반향 필터 파라메터를 재추정하는 단계; (h) 상기

재추정된 반향 필터 파라메터를 이용하여 상기 음성 특징을 다시 강화시켜 음원 신호를 최종 추정하는 단계; 를 구비한다.

[0049] 전술한 상기 (d) 단계에서 추출된 음성 특징은, 멜 스케일 필터 뱅크(Mel-scale filter bank)을 이용하여, 상기 IVA 타겟 음성 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0050] 전술한 상기 (e) 단계에서 추출된 노이즈 특징은, 멜 스케일 필터 뱅크(Mel-scale filter bank)을 이용하여, IVA 노이즈 신호에 대하여 로그 멜 주파수 전력 스펙트럼 도메인에서 추출된 LMPSCs(logarithmic mel-frequency power spectral coefficients)인 것이 바람직하다.

[0051] 전술한 상기 (f) 단계 및 (h) 단계는 HMM-based BFE 방법을 이용하여 음성 특징을 강화시키는 것이 바람직하다.

[0052] 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 특성을 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

[0053] 본 발명에 따른 장치 및 방법은 음성 인식 분야에 널리 사용될 수 있다.

부호의 설명

[0054] 10 : 음성 인식 장치

100 : 음성 신호 입력부

110 : 푸리에 변환 모듈

120 : 독립벡터분석 모듈

130 : 목적 음성 강화 모듈

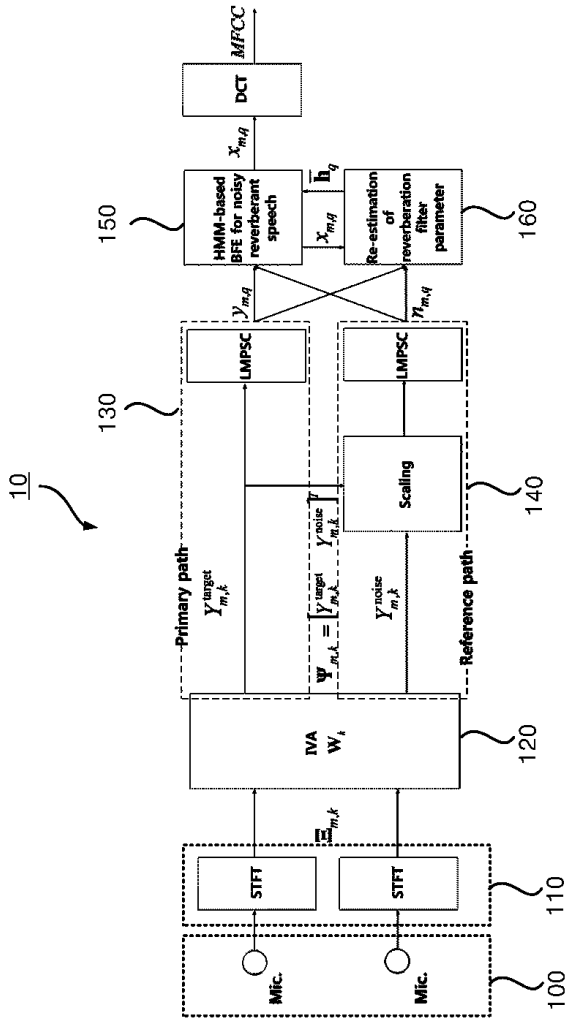
140 : 목적 음성 제거 모듈

150 : HMM 기반 특징 강화 모듈

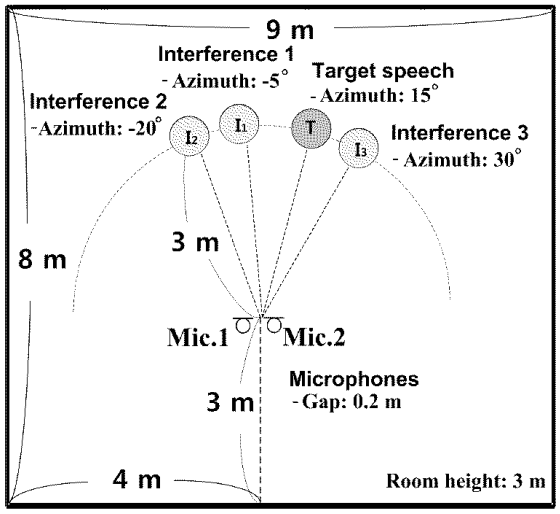
160 : 반향 필터 재추정 모듈

도면

도면1


MFCC
DCT
HMM-based BFE for noisy reverberant speech
Re-estimation of reverberation filter parameter
LMPSC
LMPSC
Scaling
Primary path
Reference path
IVA
STFT
STFT
Mic.
Mic.
10
100
110
120
130
140
150
160

도면2


9 m
Interference 1
- Azimuth: -5°
Target speech
- Azimuth: 15°
Interference 2
-Azimuth: -20°
Interference 3
- Azimuth: 30°
I2
I1
T
I3
8 m
3 m
Mic.1
Mic.2
Microphones
- Gap: 0.2 m
3 m
4 m
Room height: 3 m

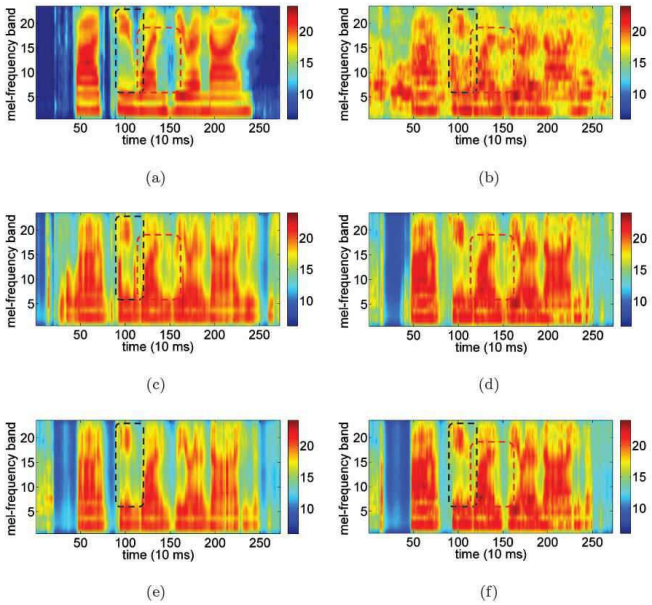
도면3

Number of Interfs.	Method	Input SIR (dB)							
		At an RT_{60} of 0.35 s				At an RT_{60} of 0.45 s			
		0	5	10	15	0	5	10	15
One	Baseline	78.19	62.90	45.45	29.42	82.74	70.76	57.28	44.01
	BFE-SNS+CRP	79.05	57.86	37.96	24.54	81.82	67.05	48.28	35.04
	BFE-IVA+CRP	30.01	23.77	18.58	18.58	40.05	32.92	26.78	23.96
	BFE-SNS+RPR	24.51	15.97	9.76	7.62	28.90	21.47	15.63	13.27
	BFE-IVA+RPR	**18.52**	**11.73**	**6.70**	**5.74**	**27.07**	**18.77**	**12.62**	**10.01**
Two	Baseline	79.18	62.78	45.27	29.64	83.57	71.96	56.70	43.55
	BFE-SNS+CRP	78.35	56.85	38.21	25.03	81.48	65.42	48.25	33.78
	BFE-IVA+CRP	52.18	30.25	21.38	19.59	61.82	38.67	29.08	23.77
	BFE-SNS+RPR	50.98	25.61	14.19	8.34	60.14	34.15	20.39	14.31
	BFE-IVA+RPR	**42.20**	**17.69**	**9.34**	**6.79**	**51.60**	**24.02**	**14.28**	**9.95**
Three	Baseline	80.93	65.97	46.65	29.94	85.78	74.48	57.86	43.95
	BFE-SNS+CRP	78.84	57.59	37.59	25.40	82.43	65.60	48.62	34.40
	BFE-IVA+CRP	58.05	35.47	24.42	21.31	66.46	43.49	30.95	24.48
	BFE-SNS+RPR	59.00	30.68	16.68	8.66	65.60	38.36	22.48	15.05
	BFE-IVA+RPR	**49.29**	**21.56**	**11.00**	**7.74**	**56.48**	**27.79**	**15.60**	**10.44**

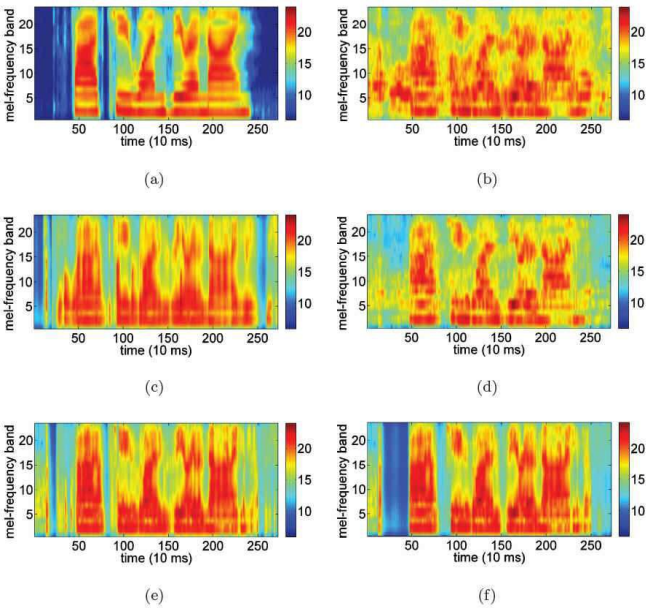
도면4

Number of Interfs.	Method	Input SIR (dB)							
		At an RT_{60} of 0.35 s				At an RT_{60} of 0.45 s			
		0	5	10	15	0	5	10	15
One	Baseline	78.19	62.90	45.45	29.42	82.74	70.76	57.28	44.01
	BFE-SNS+CRP	79.05	57.86	37.96	24.54	81.82	67.05	48.28	35.04
	BFE-IVA only	30.31	20.33	12.68	8.45	40.23	33.75	26.17	20.64
	BFE-IVA+BFE-RPR	19.87	13.79	8.02	6.79	28.62	20.85	15.42	12.13
	BFE-IVA+RPR	**18.52**	**11.73**	**6.70**	**5.74**	**27.07**	**18.77**	**12.62**	**10.01**
Two	Baseline	79.18	62.78	45.27	29.64	83.57	71.96	56.70	43.55
	BFE-SNS+CRP	78.35	56.85	38.21	25.03	81.48	65.42	48.25	33.78
	BFE-IVA only	50.34	26.87	14.93	9.43	64.22	41.49	29.39	22.14
	BFE-IVA+BFE-RPR	43.46	19.63	11.18	7.03	54.30	28.50	17.35	12.71
	BFE-IVA+RPR	**42.20**	**17.69**	**9.34**	**6.79**	**51.60**	**24.02**	**14.28**	**9.95**
Three	Baseline	80.93	65.97	46.65	29.94	85.78	74.48	57.86	43.95
	BFE-SNS+CRP	78.84	57.59	37.59	25.40	82.43	65.60	48.62	34.40
	BFE-IVA only	56.88	31.70	17.84	10.66	68.98	44.63	30.44	23.37
	BFE-IVA+BFE-RPR	51.84	24.17	12.56	8.23	59.06	33.23	19.84	13.39
	BFE-IVA+RPR	**49.29**	**21.56**	**11.00**	**7.74**	**56.48**	**27.79**	**15.60**	**10.44**

도면5


mel-frequency band
time (10 ms)
(a)
(b)
(c)
(d)
(e)
(f)

도면6


mel-frequency band
time (10 ms)
(a)
(b)
(c)
(d)
(e)
(f)