



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년04월11일
(11) 등록번호 10-1720514
(24) 등록일자 2017년03월22일

- (51) 국제특허분류(Int. Cl.)
G10L 15/07 (2013.01) G10L 15/14 (2006.01)
G10L 15/20 (2006.01)
(52) CPC특허분류
G10L 15/07 (2013.01)
G10L 15/142 (2013.01)
(21) 출원번호 10-2016-0023335
(22) 출원일자 2016년02월26일
심사청구일자 2016년02월26일
(56) 선행기술조사문헌
Minook Kim et al., 'Efficient online target speech extraction using DOA-constrained independent component analysis of stereo data for robust speech recognition', Signal Processing, pp.126~137, 2015.

- (73) 특허권자
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
박형민
서울특별시 서초구 잠원로 150 5동 903호 (잠원동, 잠원한신아파트)
이호용
경기도 안양시 동안구 경수대로623번길 46 109동 1302호 (호계동, 럭키아파트)
(74) 대리인
(뒷면에 계속)
이지연

전체 청구항 수 : 총 8 항

심사관 : 정성윤

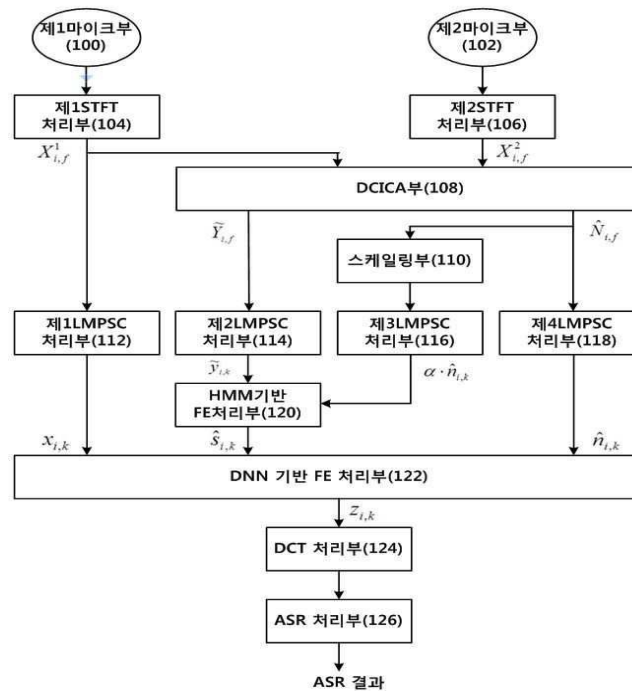
(54) 발명의 명칭 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치 및 방법

(57) 요약

본 발명에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치는, 다수의 마이크로로부터 입력되는 마이크 입력신호들을 각각 입력받아 STFT(Short-Time Fourier Transform) 처리하는 다수의 STFT; 상기 다수의 STFT 처리부가 출력하는 STFT 처리된 마이크 입력신호들을 입력받아 DCICA 처리를 이행하며, 타겟추정신호와 잡

(뒷면에 계속)

대표도 - 도1



음추정신호를 생성하는 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA)) 부; 상기 STFT 처리된 마이크 입력신호들 중 어느 하나인 제1마이크 입력신호와, 상기 DCICA부가 출력하는 타겟 추정신호와 잡음추정신호를 입력받아 LMPSC(logarithmic mel-frequency power spectral coefficient) 처리하여 스펙트럴 형태로 변환하여 출력하는 다수의 LMPSC 처리부; 상기 LMPSC 처리부가 출력하는 스펙트럴 형태로 변환되어 출력되는 제1마이크 입력신호와 타겟추정신호와 잡음추정신호를 입력받아 특징강화된 최종 타겟추정신호를 생성하는 DNN 기반 FE(DNN-Based Feature Enhancement) 처리부; 및 상기 최종 타겟추정신호를 제공받아 음성인식을 수행하는 ASR(Automatic Speech Recognition) 처리부;로 구성됨을 특징으로 한다.

또한 상기 방법은, 상기 DCICA부가 출력하는 잡음추정신호의 스케일을 가변하는 스케일링부; 상기 스케일링부가 출력하는 스케일링된 잡음추정신호를 입력받아 LMPSC 처리하여 스펙트럴 형태로 변환하여 출력하는 LMPSC; 및 상기 LMPSC가 출력하는 스펙트럴 형태로 변환된 스케일링된 잡음추정신호와 상기 DCICA가 출력하는 타겟추정신호를 입력받아 HMM(hidden markov model) 기반 FE 처리하여 특징 강화된 타겟추정신호를 생성하여 상기 DNN 기반 FE 처리부로 제공하는 HMM 기반 FE 처리부;를 더 구비하며, 상기 DNN 기반 FE 처리부는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟신호를 생성하며, 상기 DNN 기반 FE 처리부로 입력되는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호는 스펙트럴 형태임을 특징으로 한다.

(52) CPC특허분류

G10L 15/20 (2013.01)

G10L 25/51 (2013.01)

김민욱

경기도 고양시 일산동구 호수로 606, B동 904호 (장항2동, 코오롱레이크폴리스I)

(72) 발명자

조지원

서울특별시 마포구 독막로32안길 15 (신수동)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711023888

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 기초연구사업/중견연구자지원사업

연구과제명 시청각정보를 이용한 강인한 멀티모달 음성인식 기술 개발

기 여 율 1/1

주관기관 서강대학교 산학협력단

연구기간 2014.05.01 ~ 2017.04.30

명세서

청구범위

청구항 1

DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치에 있어서,

다수의 마이크로부터 입력되는 마이크 입력신호들을 각각 입력받아 STFT(Short-Time Fourier Transform) 처리하는 다수의 STFT;

상기 다수의 STFT 처리부가 출력하는 STFT 처리된 마이크 입력신호들을 입력받아 DCICA 처리를 이행하며, 타겟 추정신호와 잡음추정신호를 생성하는 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))부;

상기 STFT 처리된 마이크 입력신호들 중 어느 하나인 제1마이크 입력신호와, 상기 DCICA부가 출력하는 타겟추정신호와 잡음추정신호를 입력받아 LMPSC(logarithmic mel-frequency power spectral coefficient) 처리하여 스펙트럴 형태로 변환하여 출력하는 다수의 LMPSC 처리부;

상기 LMPSC 처리부가 출력하는 스펙트럴 형태로 변환되어 출력되는 제1마이크 입력신호와 타겟추정신호와 잡음추정신호를 입력받아 특징강화된 최종 타겟추정신호를 생성하는 DNN 기반 FE(DNN-Based Feature Enhancement) 처리부; 및

상기 최종 타겟추정신호를 제공받아 음성인식을 수행하는 ASR(Automatic Speech Recognition) 처리부;로 구성됨을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치.

청구항 2

제1항에 있어서,

상기 DCICA부가 출력하는 잡음추정신호의 스케일을 가변하는 스케일링부;

상기 스케일링부가 출력하는 스케일링된 잡음추정신호를 입력받아 LMPSC 처리하여 스펙트럴 형태로 변환하여 출력하는 LMPSC;

상기 LMPSC가 출력하는 스펙트럴 형태로 변환된 스케일링된 잡음추정신호와 상기 DCICA가 출력하는 타겟추정신호를 입력받아 HMM(hidden markov model) 기반 FE 처리하여 특징 강화된 타겟추정신호를 생성하여 상기 DNN 기반 FE 처리부로 제공하는 HMM 기반 FE 처리부;를 더 구비하며,

상기 DNN 기반 FE 처리부는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟신호를 생성하며,

상기 DNN 기반 FE 처리부로 입력되는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호는 스펙트럴 형태임을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치.

청구항 3

제1항에 있어서,

상기 특징 강화된 최종 타겟추정신호를 DCT 처리하여 상기 ASR 처리부로 제공하는 DCT 처리부를 더 구비함을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치.

청구항 4

제1항에 있어서,

상기 DNN 기반 FE 처리부는 다수의 인접 프레임으로부터의 상기 마이크 입력신호와 상기 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟추정신호를 생성함을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치.

청구항 5

DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식방법에 있어서,

다수의 마이크로부터 입력되는 마이크 입력신호들을 각각 입력받아 STFT 처리하여 STFT 처리된 마이크 입력신호들을 생성하는 단계;

상기 STFT 처리된 마이크 입력신호들을 DCICA 처리를 이행하며, 타겟추정신호와 잡음추정신호를 생성하는 단계;

상기 STFT 처리된 마이크 입력신호들 중 어느 하나인 제1마이크 입력신호와, DCICA부가 출력하는 타겟추정신호와 잡음추정신호에 대해 LMPSC 처리하여 스펙트럴 형태로 변환하는 단계;

상기 스펙트럴 형태로 변환되어 출력되는 제1마이크 입력신호와 타겟추정신호와 잡음추정신호에 대해 DNN 기반 FE 처리하여 특징강화된 최종 타겟추정신호를 생성하는 단계; 및

상기 최종 타겟추정신호로부터 음성인식을 수행하는 단계;를 구비함을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식방법.

청구항 6

제5항에 있어서,

상기 잡음추정신호의 스케일을 가변하는 단계;

상기 스케일링된 잡음추정신호에 대해 LMPSC 처리하여 스펙트럴 형태로 변환하는 단계;

상기 스펙트럴 형태로 변환된 스케일링된 잡음추정신호와 상기 타겟추정신호에 따라 HMM 기반 FE 처리하여 특징강화된 타겟추정신호를 생성하는 단계;를 더 구비하며,

상기 DNN 기반 FE 처리시에, 상기 제1마이크 입력신호와 상기 특징강화된 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟신호를 생성하며,

상기 DNN 기반 FE 처리시에, 상기 제1마이크 입력신호와 상기 특징강화된 타겟추정신호와 상기 잡음추정신호는 스펙트럴 형태임을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식방법.

청구항 7

제5항에 있어서,

상기 특징강화된 최종 타겟추정신호에 대한 음성인식 전에 DCT 처리하는 단계;를 더 구비함을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식방법.

청구항 8

제5항에 있어서,

상기 DNN 기반 FE 처리시에는 다수의 인접 프레임으로부터의 상기 마이크 입력신호와 상기 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟추정신호를 생성함을 특징으로 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식방법.

발명의 설명

기술 분야

본 발명은 음성 인식 기술에 관한 것으로, 더욱 상세하게는 DNN 기반 FE(DNN-Based Feature Enhancement) 처리시에 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))에 의해 생성된 타겟추정신호와 잡음추정신호, 마이크 입력신호를 이용하여 학습 환경과 실제 환경 사이가 불일치가 존재하거나 학습이 곤란한 환경에서 적응적으로 타겟추정신호의 특징을 향상시켜 음성인식 성능을 개선할 수 있게 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치 및 방법에 관한 것이다.

배경 기술

[0001]

- [0002] 음성 인식(Automatic Speech Recognition;ASR) 시스템에서 노이즈 강인성은 중요한 이슈로, 이는 음성 인식 시스템의 성능이 훈련 환경과 실제 환경 사이의 차이에 기인하여 심각하게 저하되기 때문이다.
- [0003] 이에 종래에는 훈련 환경과 실제 환경 사이의 부조화를 보상하기 위한 다양한 기술이 제안되었으나, 그러함에도 다양한 종류의 잡음이 존재하는 실제의 환경에서 높은 인식율을 얻을 수 없었다. 이는 T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Chichester, United Kingdom: John Wiley & Sons, Ltd., 2012., J. Droppo and A. Acero, "Environmental robustness," " in *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. Springer, 2008, pp. 653-680., B. Raj, V. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," " in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, Apr. 1997, pp. 851-854. 등에 개시되어 있다.
- [0004] 한편 딥 러닝(deep learning)은 상기한 문제를 해소하기 위한 방안으로 적용되기 시작하였으며, 이는 음성을 강화하거나 강인한 음성인식(ASR)을 위한 전처리를 위해 적용되고 있다. 또한 노이즈 제거를 위해서도 사용되며, 이는 잡음이 혼합된 입력으로부터 깨끗한 음성신호를 재구성하기 위해 딥 뉴럴 네트워크(deep neural network ; DNN)를 훈련하고, 변형된 입력신호를 인식 정밀성을 향상시킬 수 있는 깨끗한 신호로 매핑하는 훈련을 포함한다. 이는 X. Feng, Y. Zhang, and J. Grass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," " in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1778-1782. 및 F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," " in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 4623-4627. 및 K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," " *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982-992, 2015. 등에 개시되어 있다.
- [0005] 그러나 상기 DNN 기반 알고리즘은 노이즈 조건이 변화될 때에 성능이 저하되는 문제가 있으며, 이는 Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," " *IEEE Signal Processing Letters*, vol. 21, pp. 65-68, 2014. 및 Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," " *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7-19, 2015. 등에 개시되어 있다.
- [0006] 그리고 노이즈 강인성을 위한 노이즈 인식 훈련(noise-aware training ;NAT)은 상기 DNN 입력들에서의 추정된 노이즈 정보를 포함하며, 이는 M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," " in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7398-7402. 등에 개시되어 있다.
- [0007] 그리고 시간-주파수 도메인에서 음성 향상을 위한 DNN 기반 이진 마스크 추정(DNN-based binary mask estimation)은 음향 환경의 넓은 범위에서 훈련에 의해 다양한 노이즈 조건들에 대해 강인성을 보이고, 비율 마스크 추정(ratio mask estimation)은 강인한 ASR에 대해 고려되고 있으며, 이는 Y. X. Wang and D. Wang, "Towards scaling up classification-based speech separation," " *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 1381-1390, 2013., A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," " *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 826-835, 2014. 등에 개시되어 있다.
- [0008] 상기한 바와 같이 딥 러닝은 빅 데이터의 수집이 용이해짐에 따라 그 뛰어난 성능으로 각광받고 있었으나, 학습 환경과 실제 환경 사이의 불일치가 존재하거나 학습이 힘든 비정상적인 잡음 환경에서는 그 성능이 급락하는 문제가 있었다.

선행기술문헌

특허문헌

- [0009] (특허문헌 0001) 한국 특허등록 제10-1591626호

(특허문헌 0002) 한국 특허등록 제10-1561651호

(특허문헌 0003) PCT 특허공개 WO 2012/036934

(특허문헌 0004) 한국 특허공개 제10-2012-0102306호

발명의 내용

해결하려는 과제

[0010] 본 발명은 DNN 기반 FE(DNN-Based Feature Enhancement) 처리시에 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))에 의해 생성된 타겟추정신호와 잡음추정신호, 마이크 입력신호를 이용하여 학습 환경과 실제 환경 사이가 불일치가 존재하거나 학습이 곤란한 환경에서 적응적으로 타겟추정신호의 특징을 향상시켜 음성인식 성능을 개선할 수 있게 하는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치 및 방법을 제공하는 것을 그 목적으로 한다.

과제의 해결 수단

[0011] 상기한 목적을 달성하기 위한 본 발명에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치는, 다수의 마이크로부터 입력되는 마이크 입력신호들을 각각 입력받아 STFT(Short-Time Fourier Transform) 처리하는 다수의 STFT; 상기 다수의 STFT 처리부가 출력하는 STFT 처리된 마이크 입력신호들을 입력받아 DCICA 처리를 이행하며, 타겟추정신호와 잡음추정신호를 생성하는 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))부; 상기 STFT 처리된 마이크 입력신호들 중 어느 하나인 제1마이크 입력신호와, 상기 DCICA부가 출력하는 타겟추정신호와 잡음추정신호를 입력받아 LMPSC(logarithmic mel-frequency power spectral coefficient) 처리하여 스펙트럴 형태로 변환하여 출력하는 다수의 LMPSC 처리부; 상기 LMPSC 처리부가 출력하는 스펙트럴 형태로 변환되어 출력되는 제1마이크 입력신호와 타겟추정신호와 잡음추정신호를 입력받아 특징강화된 최종 타겟추정신호를 생성하는 DNN 기반 FE(DNN-Based Feature Enhancement) 처리부; 및 상기 최종 타겟추정신호를 제공받아 음성인식을 수행하는 ASR(Automatic Speech Recognition) 처리부;로 구성됨을 특징으로 한다.

[0012] 또한 상기 방법은, 상기 DCICA부가 출력하는 잡음추정신호의 스케일을 가변하는 스케일링부; 상기 스케일링부가 출력하는 스케일링된 잡음추정신호를 입력받아 LMPSC 처리하여 스펙트럴 형태로 변환하여 출력하는 LMPSC; 및 상기 LMPSC가 출력하는 스펙트럴 형태로 변환된 스케일링된 잡음추정신호와 상기 DCICA가 출력하는 타겟추정신호를 입력받아 HMM(hidden markov model) 기반 FE 처리하여 특징 강화된 타겟추정신호를 생성하여 상기 DNN 기반 FE 처리부로 제공하는 HMM 기반 FE 처리부;를 더 구비하며, 상기 DNN 기반 FE 처리부는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호를 입력받아 특징강화된 최종 타겟신호를 생성하며, 상기 DNN 기반 FE 처리부로 입력되는 상기 제1마이크 입력신호와 상기 특징 강화된 타겟추정신호와 상기 잡음추정신호는 스펙트럴 형태임을 특징으로 한다.

발명의 효과

[0013] 상기한 본 발명은 DNN 기반 FE(DNN-Based Feature Enhancement) 처리시에 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))에 의해 생성된 타겟추정신호와 잡음추정신호, 마이크 입력신호를 이용하여 학습 환경과 실제 환경 사이가 불일치가 존재하거나 학습이 곤란한 환경에서 적응적으로 타겟추정신호의 특징을 향상시켜 음성인식 성능을 개선할 수 있는 효과를 야기한다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 바람직한 실시예에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치의 구성도.

도 2는 본 발명의 바람직한 실시예에 따르는 DNN 기반 FE 처리부를 간략하게 도시한 도면.

도 3 내지 6은 본 발명의 바람직한 실시예에 따르는 테스트 결과를 예시한 도면.

발명을 실시하기 위한 구체적인 내용

- [0015] 본 발명은 DNN 기반 FE(DNN-Based Feature Enhancement) 처리시에 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))에 의해 생성된 타겟추정신호와 잡음추정신호, 마이크 입력신호를 이용하여 학습 환경과 실제 환경 사이가 불일치가 존재하거나 학습이 곤란한 환경에서 적응적으로 타겟추정신호의 특징을 향상시켜 음성인식 성능을 개선시킨다.
- [0016] <DNN 기반 FE를 수행하는 음성인식장치의 구성>
- [0017] 상기한 본 발명의 바람직한 실시예에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치의 구성도를 도시한 것이 도 1이다.
- [0018] 상기 도 1을 참조하면, 상기 음성인식장치는 제1 및 제2마이크부(100,102)와 제1 및 제2STFT(Short-Time Fourier Transform) 처리부(104,106)와 DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))부(108)와 스케일링부(110)와 제1 내지 제4LMPSC(logarithmic mel-frequency power spectral coefficient) 처리부(112 내지 118)와 HMM(hidden markov model) 기반 FE 처리부(120)와 DNN 기반 FE(DNN-Based Feature Enhancement) 처리부(122)와 DCT(Discrete Cosine Transform) 처리부(124)와 ASR(automatic speech recognition) 처리부(126)로 구성된다. 상기한 음성인식장치의 제1 및 제2STFT 처리부(104,106)와 DCICA부(108)와 스케일링부(110)와 제1 내지 제4LMPSC 처리부(112 내지 118)와 HMM 기반 FE 처리부(120)와 DNN 기반 FE 처리부(122)와 DCT 처리부(124)와 ASR 처리부(126)는 소프트웨어 모듈로 구성되어 음성인식장치에 설치된 프로세서 등에 의해 수행되거나 하드웨어 모듈로 구성될 수 있다.
- [0019] 상기 제1 및 제2마이크부(100,102) 각각은 서로 이격되어 설치되어 화자의 음성을 입력받아 그에 대응되는 마이크 입력신호를 생성하여 제1 및 제2STFT 처리부(104,106)로 제공한다.
- [0020] 상기 제1 및 제2STFT 처리부(104,106) 각각은 상기 제1 및 제2마이크부(100,102)가 출력하는 음성신호를 STFT 처리하여 출력하며, 각 출력을 $X_{i,f}^1$, $X_{i,f}^2$ 라 칭한다. 여기서, 상기 제1STFT 처리부(104)의 출력 $X_{i,f}^1$ 은 상기 제1LMPSC 처리부(112)와 상기 DCICA부(108)로 제공된다. 그리고 상기 제2STFT 처리부(106)의 출력 $X_{i,f}^2$ 은 상기 DCICA부(108)로 제공된다.
- [0021] 상기 DCICA부(108)는 제1 및 제2마이크부(100,102)로부터의 마이크 입력신호를 STFT 처리한 마이크 입력신호 $X_{i,f}^1$, $X_{i,f}^2$ 를 입력받아 DCICA 처리를 이행하고, 그에 따라 타겟추정신호 $\tilde{Y}_{i,f}$ 와 잡음추정신호 $\hat{N}_{i,f}$ 를 출력한다. 여기서, 상기 타겟추정신호 $\tilde{Y}_{i,f}$ 는 제2LMPSC 처리부(114)로 입력되고, 상기 잡음추정신호 $\hat{N}_{i,f}$ 는 스케일링부(110)와 제4LMPSC 처리부(118)로 입력된다.
- [0022] 상기 스케일링부(110)는 상기 DCICA부(108)에 의해 변경된 잡음추정신호 $\hat{N}_{i,f}$ 의 스케일을 타겟추정신호 $\tilde{Y}_{i,f}$ 에 대응되게 변경하기 위한 스케일링 조정값(α)을 곱하여 스케일 조정된 잡음추정신호를 출력한다. 상기 스케일링부(110)에 의해 출력된 스케일 조정된 잡음추정신호를 제3LMPSC 처리부(116)로 제공한다.
- [0023] 상기 제1 내지 제4LMPSC 처리부(112~118)는 입력된 신호, 즉 마이크 입력신호 $X_{i,f}^1$, 타겟추정신호 $\tilde{Y}_{i,f}$, 스케일 조정된 잡음추정신호, 잡음추정신호 $\hat{N}_{i,k}$ 를 입력받아 LMPSC 처리하여 스펙트럴 형태로 출력하며, 이는 LMPSC 처리된 마이크 입력신호 $x_{i,k}$, 타겟추정신호 $\tilde{y}_{i,k}$, 스케일 조정된 잡음추정신호 $\alpha \cdot \hat{n}_{i,k}$, 잡음추정신호 $\hat{n}_{i,k}$ 이다.
- [0024] 상기 LMPSC 처리된 타겟추정신호 $\tilde{y}_{i,k}$, 스케일 조정된 잡음추정신호 $\alpha \cdot \hat{n}_{i,k}$ 는 상기 HMM 기반 FE 처리부(120)에 입력된다. 상기 HMM 기반 FE 처리부(120)는 상기 LMPSC 처리된 타겟추정신호 $\tilde{y}_{i,k}$, 스케일 조정된 잡음추정신호 $\alpha \cdot \hat{n}_{i,k}$ 를 입력받아, 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$ 를 출력한다.
- [0025] 상기 마이크 입력신호 $x_{i,k}$, 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$, 잡음추정신호 $\hat{n}_{i,k}$ 는 DNN 기반 FE 처리부(122)로 입력된다.

[0026] 상기 DNN 기반 FE 처리부(122)는 상기 마이크 입력신호 $x_{i,k}$, 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$, 잡음추정신호 $\hat{n}_{i,k}$ 를 입력받아 DNN 기반 FE 처리하여 특징이 더욱 향상된 최종 타겟추정신호 $z_{i,k}$ 를 출력한다. 여기서, 상기 DNN 기반 FE 처리부(122)는 마이크 입력신호 $x_{i,k}$ 와 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$ 로부터 특징이 더욱 강화된 최종 타겟추정신호 $z_{i,k}$ 를 출력함에 있어 잡음추정신호 $\hat{n}_{i,k}$ 를 참조하여 특징강화성능을 향상시킨다. 이와달리 상기 DNN 기반 FE 처리부(122)는 마이크 입력신호 $x_{i,k}$ 와 잡음추정신호 $\hat{n}_{i,k}$ 를 토대로 더욱 특징 강화된 최종 타겟추정신호 $z_{i,k}$ 를 생성함에 있어 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$ 를 참조하여 특징강화성능을 향상시킬 수도 있다. 이러한 DNN 기반 FE 처리부(122)에 의한 특징 향상에 의한 음성인식성능의 향상은 본 발명과 종래 기술에 따른 테스트 결과를 통해 드러나며 이는 후술된다.

[0027] 상기 특징 향상된 타겟추정신호 $z_{i,k}$ 는 DCT 처리부(124)에 입력된다.

[0028] 상기 DCT 처리부(124)는 상기 특징 향상된 최종 타겟추정신호 $z_{i,k}$ 를 입력받아, DCT 처리하여 ASR 처리부(126)로 제공한다.

[0029] 상기 ASR 처리부(126)는 상기 DCT 처리된 최종 타겟추정신호를 제공받아 ASR을 수행하고 그에 따른 ASR 결과를 출력한다.

[0030] <DNN 기반 FE 처리부(122)의 초기화>

[0031] 상기한 바와 같이 본 발명은 DNN 기반 FE 처리부(122)에 상기 마이크 입력신호 $x_{i,k}$, 특징 강화된 타겟추정신호 $\hat{s}_{i,k}$, 잡음추정신호 $\hat{n}_{i,k}$ 를 입력받아 DNN 기반 FE 처리하여 더욱 특징 향상된 최종 타겟추정신호 $z_{i,k}$ 를 출력한다.

[0032] 이를 위한 상기 DNN 기반 FE 처리부(122)의 초기화 과정을 설명한다.

[0033] 상기 DNN 기반 FE 처리부(122)에 의해 수행되는 딥 러닝은 컴퓨팅 파워의 발달과 빅 데이터의 수집으로 인하여 뛰어난 성능을 보이고 있으나, 국소 최저값에 빠지거나 그래디언트가 사라지는 문제가 있다.

[0034] 이에 본 발명은 각 층별로 RBM(Restricted Boltzmann Machine)을 이용하여 선행 학습을 수행한다. 상기 RBM은 다음과 같은 확률 분포를 볼츠만(Boltzmann) 분포에 의해 정의한다.

수학식 1

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h}))$$

[0036] 여기서 Z와 E는 각각 다음과 같이 정의한다.

수학식 2

$$Z = \sum_{(\mathbf{v}, \mathbf{h})} \exp(-E(\mathbf{v}, \mathbf{h}))$$

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^n \sum_{j=1}^m \omega_{ij} h_i v_j - \sum_{j=1}^m a_j v_j - \sum_{i=1}^n b_i h_i,$$

[0038] 그리고, $(\mathbf{v}, \mathbf{h}) \in \{0,1\}^{m+n}$ 은 m개의 가시 유닛으로 이루어진 $\mathbf{V}=(V_1, \dots, V_m)$ 와 n개의 은닉 유닛으로 이루어진 $\mathbf{H}=(H_1, \dots, H_n)$ 로부터 얻어낸 랜덤변수이다. 이 확률 분포의 우도를 최대화하기 위하여 로그 우도를 그래디언트 상승 방법으로 최대화한다.

수학식 3

$$\ln \mathcal{L}(\theta|\mathbf{v}) = \ln p(\mathbf{v}|\theta) = \ln \frac{1}{Z} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h})),$$

$$\frac{\partial \ln \mathcal{L}(\theta|\mathbf{v})}{\partial \omega_{ij}} = p(H_i = 1|\mathbf{v})v_j - \sum_{\mathbf{v}} p(\mathbf{v})p(H_i = 1|\mathbf{v})v_j.$$

[0039]

[0040]

[0041]

[0042]

[0043]

[0044]

이와 같은 방법으로 초기화한 네트워크를 최소 평균 제곱 에러 기법을 이용하여 재학습한다.

상기한 딥 러닝은 잡음추정신호로부터 클린신호, 즉 타겟추정신호를 맵핑시켜주는 회귀 함수로써 사용되었으며, 이러한 딥 러닝 기반 방법은 시간-주파수 정보를 모두 활용할 수 있다는 점에서 뛰어나지만 제한된 조건으로 학습됨에 따라 학습시의 잡음 환경에서 벗어나는 경우에는 그 성능이 급락하였다.

이에 본 발명은 마이크 입력신호로부터 추정된 잡음추정신호를 딥 러닝시에 반영함으로써 잡음 환경의 변화에도 유연하게 대응할 수 있게 한다.

즉, 본 발명은 독립 성분 분석 기술과 모델 기반 특징 강화 방법을 통해 얻어낸 특징 정보들을 딥러닝 학습에 활용함으로써, 변화하는 잡음 환경에서도 뛰어난 성능을 확보할 수 있게 한다.

상기한 학습을 위해 상기 네트워크로 입력되는 데이터는, 마이크로부터 얻어진 마이크 입력신호의 24차원 로그-멜 스펙트럼 계수와 독립 성분 분석 및 모델 기반 특징 강화 방법을 통해 얻어낸 향상된 타겟 추정 신호와 잡음 추정 신호의 로그-멜 스펙트럼 계수이다. 이때, 현재 프레임의 앞, 뒤로 각각 τ 개의 프레임 정보를 추가로 이용한다. 즉, 최종 입력 신호는 다음과 같다.

수학식 4

$$\tilde{\mathbf{v}}(i) = [\{\mathbf{x}(i-\tau)^T, \hat{\mathbf{s}}(i-\tau)^T, \hat{\mathbf{n}}(i-\tau)^T\}, \dots, \{\mathbf{x}(i)^T, \hat{\mathbf{s}}(i)^T, \hat{\mathbf{n}}(i)^T\}, \dots, \{\mathbf{x}(i+\tau)^T, \hat{\mathbf{s}}(i+\tau)^T, \hat{\mathbf{n}}(i+\tau)^T\}]^T,$$

[0045]

[0046]

[0047]

상기 $\mathbf{x}(i)$ 은 입력신호의 로그-멜 스펙트럼 계수이고, $\hat{\mathbf{s}}(i)$ 와 $\hat{\mathbf{n}}(i)$ 은 각각 음원 향상 신호와 추정 잡음 신호의 로그-멜 스펙트럼 계수이다.

본 발명에서는 상기 τ 는 3으로 설정하여, 총 7개의 프레임 정보를 이용하였다. 즉 네트워크의 입력 차원은 $\# \text{ of channels} \times 3 \times (2 \cdot \tau + 1) = 24 \times 3 \times 7 = 504$ 차원이다. 이를 위한 딥러닝 네트워크는 1024개의 은닉 유닛을 가진 3개의 은닉 층으로 이루어지며, 각 층의 활성화함수는 시그모이드를 사용하고, 마지막 출력 층에서는 선형 유닛을 사용한다.

[0048]

<DCICA 처리부>

[0049]

이제 DCICA 처리부(108)의 처리절차를 설명한다.

[0050]

잡음이 존재하는 실제 세상의 환경에서 타겟 음성은 상가성 잡음(additive noise)에 의해 종종 오염된다. 이 경우, i 번째 프레임과 j 번째 주파수 bin에서의 m 번째 마이크로부터의 잡음이 혼합된 음성 $X_{i,j}^m$ 에 대한 시간-주파수(t-f) 표현은 수학식 5에 나타난 바와 같다.

수학식 5

$$X_{i,j}^m = S_{i,j}^m + N_{i,j}^m,$$

[0051]

[0052] 상기 수학식 5에서, $S_{i,j}^m$ 와 $N_{i,j}^m$ 는 클린 음성과 잡음을 나타내며, 이는 각각 독립적으로 추정된다.

[0053] DCICA에서, 더미 출력(dummy outputs)은 타겟 음성 신호를 제거하는 것으로부터 잡음 추정을 시작한다. 타겟 스피커로부터 마이크로의 음향 경로가 직접 경로에 근접한다고 가정하면, 첫번째 그리고 m번째 마이크 신호에 대한 간단한 딜레이-서브트랙트 널 포머(delay and subtract nullformer)로부터 m번째 더미 출력 $U_{i,j}^m$ 을 얻을 수 있으며, 이는 수학식 6과 같다.

수학식 6

$$[0054] \quad U_{i,j}^m = X_{i,j}^m - (\Gamma_j)^{m-1} X_{i,j}^1, \quad m = 2, \dots, M,$$

[0055] 상기 수학식 6에서, $\Gamma_j = \exp(\sqrt{-1} \cdot \omega_j d \sin \theta_{\text{target}} / c)$ 으로 타겟 DOA θ_{target} 를 가진다.

[0056] 실제 타겟 음성 출력을 수학식 7과 같이 나타낸다고 가정한다.

수학식 7

$$[0057] \quad Y_{i,j} = \mathbf{w}_j \mathbf{x}_{i,j} = [W_j^1, \dots, W_j^M] [X_{i,j}^1, \dots, X_{i,j}^M]^T,$$

[0058] 상기 수학식 7에서 $\mathbf{w}_j$ 는 적응적 로(ROW) 벡터로, $Y_{i,j}$ 와 $U_{i,j}^m$ 사이의 종속성을 최소화하는 것에 의해 추정된다. 그리고 넌 하모닉 제약을 가지는 온라인 내추럴 그래디언트 알고리즘과 평탄화 파워추정에 의한 정규화에 의해 상기 적응적 로 벡터는 수학식 8에 나타낸 바와 같이 표현될 수 있다.

수학식 8

$$[0059] \quad \Delta \mathbf{w}_j \propto \frac{\phi(Y_{i,j})}{\sqrt{\xi_{i,j}}} \left[\sum_{m=2}^M (\Gamma_j)^{m-1} (U_{i,j}^m)^*, -(U_{i,j}^2)^*, \dots, -(U_{i,j}^M)^* \right]$$

[0060] 상기 수학식 8에서 $p(Y_{i,j}) = \exp\{\sqrt{-1} \cdot \arg(Y_{i,j})\}$ 이고, 평탄화 파워 추정 $\xi_{i,j} = \beta \xi_{i-1,j} + (1 - \beta) \|\mathbf{x}_{i,j}\|^2 / M$ 으로 팩터 β 를 가진다. 최소 왜곡 원리는 $Y_{i,j}$ 에 의해 분리된 출력 신호의 스케일링 불확정성을 해소하기 위해 사용되며, 이는 $\sum_{m=1}^M (\Gamma_j)^{m-1} W_j^m$ 와 같다.

[0061] 더미 출력이 잡음 파라미터의 추정을 위해 사용될지라도, 타겟 음성 출력은 그 특징을 더욱 향상하기 위해 베이지안 FE에서 수행되는 잡음이 있는 음성을 이용한다.

[0062] 각 주파수 bin에서의 DCICA의 출력신호를 얻기 위한 혼합신호들의 웨이트드 썸(weighted sum)은 신호들의 변형을 수반하며, 타겟 음성과 잡음 신호들 사이의 상대적 변형은 J.-W. Cho and H.-M. Park, Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition, *Signal Processing*, vol. 120, pp. 200~208, 2016.에 개시된 방법에 따라 보상될 수 있다.

[0063] 상기 베이지안 FE에 $|Y_{i,j}|^2$ 에 k번째 밴드 멜 스케일 필터(k-th band mel-scale filter)와 LMPSC(the logarithmic mel-frequency power spectral coefficient) $y_{i,k}$ 를 적용하면, 상기 베이지안 FE는 수학식 9와 같이 나타낼 수 있다.

수학식 9

$$[0064] \quad y_{i,k} = \log[\exp(s_{i,k}) + \exp(n_{i,k})]$$

[0065] 상기 수학적 식 9에서 $s_{i,k}$ 와 $n_{i,k}$ 는 클린 음성과 노이즈 성분의 LMPSCs이다.

[0066] 상기 $y_{i,k}$ 의 베이지안 FE는 깨끗한 LMPSC 벡터의 MMSE 추정을 수행한다.

수학적 식 10

[0067] $\hat{s}_i = E[s_i | y_{1:i}]$,

[0068] 상기 수학적 식 10에서 s_i 는 $[s_{i,1}, \dots, s_{i,K}]^T$ 의 특징 벡터를 나타내고, $y_{1:i}$ 는 관찰된 노이즈 음성 $(y_1, \dots, y_i)$ 의 특징벡터들의 시퀀스들을 나타낸다.

[0069] 클린 음성이 싱글 가우시안 관찰 분포를 가지는 에르고딕(ergodic) HMM에 의해 표현되어 추정되는 동안에, 상기 $n_{i,k}$ 는 보상후 더미 출력으로부터 평균 $\mu_{n_{i,k}}$ 과 변화량 $\sigma_{n_{i,k}}^2$ 을 추정하는 가우시안 랜덤 프로세스에 의해 추정되며, 이는 수학적 식 11에 따른다.

수학적 식 11

[0070] $p(s_{i,k} | \gamma_i = l) = \mathcal{N}(s_{i,k} | \mu_{s_{i,k} | \gamma_i = l}, \sigma_{s_{i,k} | \gamma_i = l}^2)$, $l = 1, \dots, N_\gamma$,

[0071] 상기 수학적 식 11에서 γ_i 과 N_γ 는 i번째 프레임에서의 히든 스테이트 인덱스와 상태의 번호를 나타낸다.

[0072] 이 경우 강화된 LMPSCs는 수학적 식 12에 의해 정의된다.

수학적 식 12

[0073] $\hat{s}_{i,k} = \sum_{l=1}^{N_\gamma} P(\gamma_i = l | y_{1:i}) \hat{\mu}_{s_{i,k} | \gamma_i = l}$,

[0074] $\hat{\mu}_{s_{i,k} | \gamma_i = l} = \mu_{s_{i,k} | \gamma_i = l} + \frac{d_{s_{i,k} | \gamma_i = l} \sigma_{s_{i,k} | \gamma_i = l}^2 \{y_{i,k} - \log[\exp(\mu_{s_{i,k} | \gamma_i = l}) + \exp(\mu_{n_{i,k}})]\}}{d_{s_{i,k} | \gamma_i = l}^2 \sigma_{s_{i,k} | \gamma_i = l}^2 + (1 - d_{s_{i,k} | \gamma_i = l})^2 \sigma_{n_{i,k}}^2}$,

[0075] 상기 수학적 식 12에서 $d_{s_{i,k} | \gamma_i = l} = \exp(\mu_{s_{i,k} | \gamma_i = l}) / [\exp(\mu_{s_{i,k} | \gamma_i = l}) + \exp(\mu_{n_{i,k}})]$ 이다. 그리고 $P(\gamma_i = l | y_{1:i})$ 는 수학적 식 13에 따른 베이지안 물에 의해 규정될 수 있는 $y_{1:i}$ 에 대한 1번째 상태의 분산으로 계산된다.

수학적 식 13

[0076] $P(\gamma_i = l | y_{1:i}) = \frac{P(\gamma_i = l | y_{1:i-1}) p(y_i | y_{1:i-1}, \gamma_i = l)}{\sum_{l'=1}^{N_\gamma} P(\gamma_i = l' | y_{1:i-1}) p(y_i | y_{1:i-1}, \gamma_i = l')}$

[0077] 상태전이 확률 $P(\gamma_i = l | \gamma_{i-1} = l')$ 을 사용하여, $P(\gamma_i = l | y_{1:i-1})$ 는 수학적 식 14로 표현될 수 있다.

수학적 식 14

[0078] $P(\gamma_i = l | y_{1:i-1}) = \sum_{l'=1}^{N_\gamma} P(\gamma_i = l | \gamma_{i-1} = l') P(\gamma_{i-1} = l' | y_{1:i-1})$

[0079] 그리고 $p(y_i | y_{1:i-1}, \gamma_i = l)$ 는 $p(y_i | \gamma_i = l)$ 로 근사화되며, 이는 수학적 식 15에 따른다.

수학적식 15

$$p(y_{i,k}|\gamma_i = l) = \mathcal{N}(y_{i,k}|\mu_{y_{i,k}|\gamma_i=l}, \sigma_{y_{i,k}|\gamma_i=l}^2)$$

with $\mu_{y_{i,k}|\gamma_i=l} = \log[\exp(\mu_{s_{i,k}|\gamma_i=l}) + \exp(\mu_{n_{i,k}})]$ and $\sigma_{y_{i,k}|\gamma_i=l}^2 = d_{s_{i,k}|\gamma_i=l}^2 \sigma_{s_{i,k}|\gamma_i=l}^2 + (1 - d_{s_{i,k}|\gamma_i=l})^2 \sigma_{n_{i,k}}^2$.

[0080]

[0081] <DNN 기반 FE 처리부>

[0082] 상기 DNN은 다양한 패턴의 인식 문제에 성공적으로 적용됨에 따라 머신 러닝 및 신호 프로세싱 필드에서 최신 기술로서 주목을 받았다. 상기 DNN의 뛰어난 성능의 장점에 의해, FE를 위해 잡음이 있는 음성 특징을 클린 신호로 맵핑하는 기능으로 사용된다. 이러한 상기 DNN 기반 FE 방법은 특징 벡터의 시퀀스를 사용하여 시간 또는 주파수 축에 따르는 신호를 동시에 사용하기 때문에 매우 유용하다.

[0083] 그러나 상기 DNN 기반 FE는 훈련시에 제공되지 않았던 노이즈 환경에서는 그 성능이 저하되었다. 왜냐하면, 잡음 스펙트럴 특징은 음성이 없는 구간에서 추정된 정상성을 가지는 잡음에서의 평균 입력 특징을 토대로 추정되기 때문이다.

[0084] 한편 DCICA는 음성이 있는 구간에서도 잡음 스펙트럴 특징을 추정하므로, 상기 DCICA에 의해 검출된 잡음 스펙트럴 특징을 채용하는 경우에 음성인식 성능을 높일 수 있게 된다.

[0085] 이에 본 발명은 DCICA-FE의 결과를 활용하는 DNN 기반 FE 처리를 이행한다.

[0086] 이를 위해 먼저 마이크 입력신호에 대한 스펙트럴 특징정보와, DCICA에 의해 검출된 타겟추정신호와 잡음추정신호를 토대로 HMM 기반 FE 처리된 타겟추정신호에 대한 스펙트럴 특징정보와, 상기 잡음추정신호에 대한 스펙트럴 특징정보를 일시적인 변동에 대응하기 위해 인접한 다수의 프레임 단위로 입력받는다.

[0087] 예를들어, 상기 DNN의 입력의 차원은 프레임 당 24 멜 주파수 밴드를 가지는 일곱 프레임의 시퀀스를 사용하여 504가 될 수 있고, 출력의 수는 24가 될 수 있다. 이 경우 상기 DNN은 도 2에 도시한 바와 같이 세 개의 히든 레이어를 가지며, 각 히든 레이어는 1,024 유닛을 가진다. 여기서, 상기 히든의 동적 기능과 출력 유닛은 시그모이드와 선형 평선들이다.

[0088] <음성인식 성능 평가>

[0089] 본 출원인은 본 발명에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치의 성능 평가를 위해 DARPA Resource Management(RM) database를 사용하였으며 Kaldi toolkit을 이용하여 음성 인식률을 산출하였다.

[0090] 그리고 음성 인식 과정에는 13차 MFCC와 그 delta, acceleration 정보와 HMM을 이용하였다. 타겟 신호는 NoiseX-92 corpus에서 얻은 babble 잡음을 섞어 잡음 환경을 구축하였으며 이때 반향은 0.3초로 설정하였다. 또한 딥러닝 네트워크 학습 시 테스트에서 쓰일 babble 잡음을 이용하여 학습한 경우(matched)와 테스트시 쓰이지 않는 다른 car, F16, factory, operation rooms 잡음을 이용하여 학습한 경우(mismatched)로 나누어 실험을 진행하였다.

[0091] 상기 실험시에는 본 발명에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치에 대해 DNN없이 DCICA-FE의 결과, 기존 DNN기반 방법, noise-aware training(NAT)와 비교하였다. NAT는 입력 신호의 앞 6프레임 평균값을 이용하여 잡음 정보를 추정하여 딥러닝 학습에 이용하는 방법이다. 또한, 본 발명과 유사하게 DCICA-FE가 아닌 Ideal DUET이나 BF-Softmask를 이용하여 얻어낸 특징 정보를 딥러닝 학습에 활용하는 방법과도 그 결과를 비교하였다.

[0092] 도 3 내지 도 6은 타겟 음원 외에 방해 요소로 다른 발화가 추가된 경우에 대한 실험 결과를 나타낸 것이다. 상기 도 3 내지 도 6을 참조하면, 타겟과 잡음의 위치에 대한 Ideal한 파라미터 값을 사용하는 Ideal DUET 방법을 제외하면 제안하는 방법이 기존 방법보다 매우 뛰어난 성능을 보임을 확인할 수 있다.

[0093] 이와 같이 딥러닝 기반 방법으로는 학습이 매우 힘든 비정상적 잡음 환경에서는 본 발명에 따르는 DCICA를 이용한 DNN 기반 특징향상을 수행하는 음성인식장치의 성능은 월등하게 높고, 이는 실험을 통해 증명된다.

부호의 설명

[0094]

100, 102 : 제1 및 제2마이크부

104,106 : 제1 및 제2STFT(Short-Time Fourier Transform) 처리부

108 : DCICA(Direction of arrival(DOA) Constrained Independent Component Analysis(ICA))부

110 : 스케일링부

112 내지 118 : 제1 내지 제4LMPSC(logarithmic mel-frequency power spectral coefficient) 처리부

120 : HMM((hidden markov model) 기반 FE 처리부

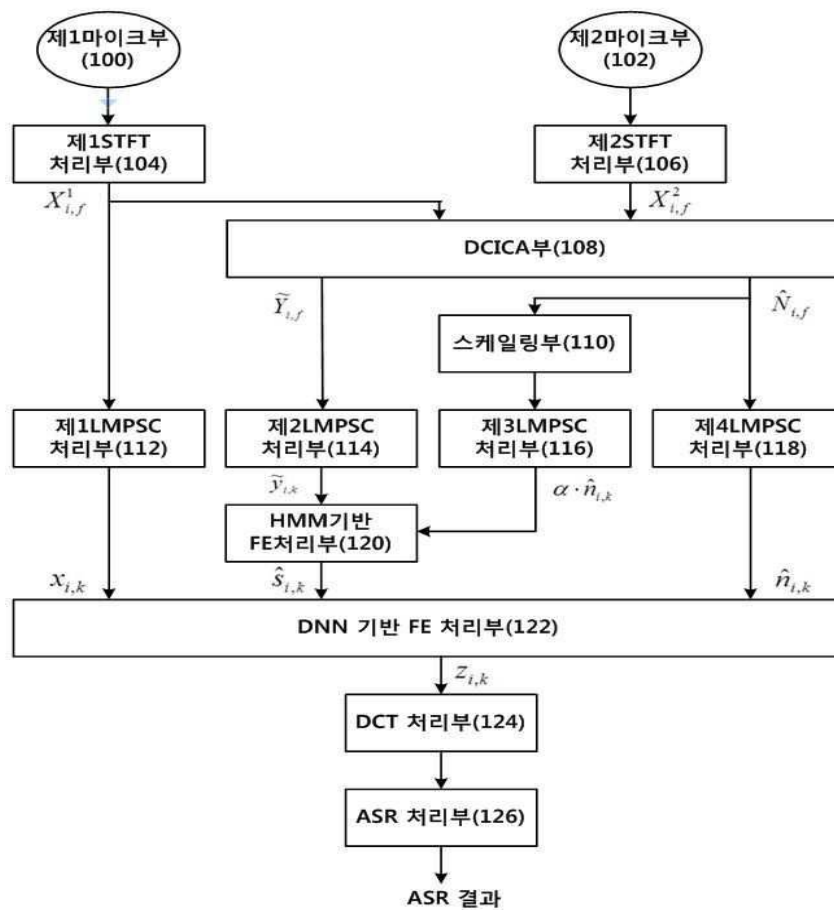
122 : DNN 기반 FE(DNN-Based Feature Enhancement) 처리부

124 : DCT(Discrete Cosine Transform) 처리부

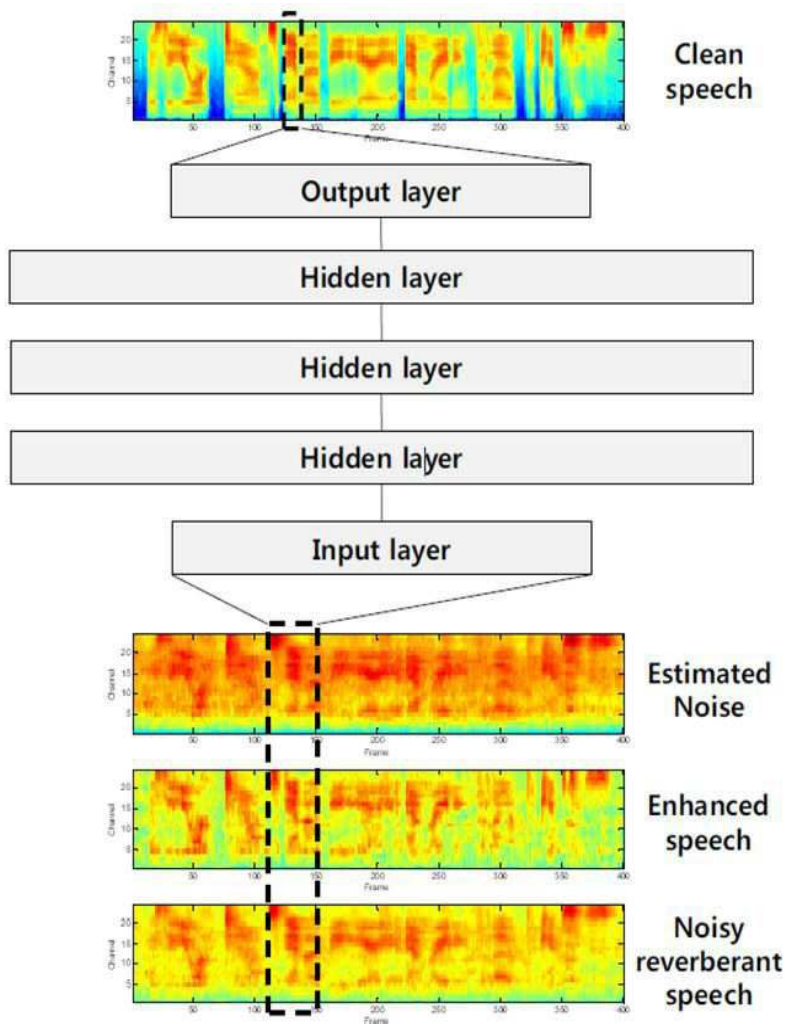
126 : ASR(automatic speech recognition) 처리부

도면

도면1



도면2



도면3

Number of interfering sources : 1				
Word Accuracy [%]	Input SNR			
	0dB	5dB	10dB	15dB
Baseline	16.99	38.73	58.61	71.18
DCICA-MBFE	44.59	58.81	67.98	70.95
Conv. DNN (matched)	71.61	84.46	88.91	91.25
DNN+ NAT (matched)	72.2	85.94	89.81	92.03
DNN+ IDUET (matched)	87.39	91.8	93.4	94.14
DNN+ BF-Softmask (matched)	81.45	89.65	93.05	93.83
Proposed method (matched)	89.14	93.13	94.69	95
Conv. DNN (mismatched)	48.34	74.23	87	91.68
DNN+ NAT (mismatched)	58.96	79.07	88.99	92.27
DNN+ IDUET (mismatched)	85.04	92.15	94.42	94.49
DNN+ BF-Softmask (mismatched)	62.83	81.22	91.06	93.6
Proposed method (mismatched)	78.8	88.56	93.64	95.43

도면4

Number of interfering sources : 2				
Word Accuracy [%]	Input SNR			
	0dB	5dB	10dB	15dB
Baseline	13.74	38.34	59.16	70.4
DCICA-MBFE	30.96	51.15	61.23	65.68
Conv. DNN (matched)	74.78	85.67	89.85	91.92
DNN+ NAT (matched)	75.4	86.68	91.64	92.89
DNN+ IDUET (matched)	81.76	89.61	92.42	92.85
DNN+ BF-Softmask (matched)	80.24	89.38	92.35	93.75
Proposed method (matched)	84.5	91.25	93.79	94.69
Conv. DNN (mismatched)	47.25	73.14	87.15	91.6
DNN+ NAT (mismatched)	56.03	78.06	89.18	92.82
DNN+ IDUET (mismatched)	63.22	82.04	90.78	93.36
DNN+ BF-Softmask (mismatched)	53.1	76.69	89.54	92.82
Proposed method (mismatched)	61.23	80.05	89.65	93.71

도면5

Number of interfering sources : 1				
Word Accuracy [%]	Input SNR			
	0dB	5dB	10dB	15dB
Baseline	1.95	15.23	40.84	59.82
DCICA-MBFE	22.34	46.08	61.58	69.54
Conv. DNN	5.78	38.77	59.16	68.37
DNN+ NAT	5.15	40.34	66.69	75.91
DNN+ IDUET	80.79	88.4	92.7	93.09
DNN+ BF-Softmask	65.72	81.92	87.78	92.19
Proposed method	82.12	89.3	91.92	93.44

도면6

Number of interfering sources : 2				
Word Accuracy [%]	Input SNR			
	0dB	5dB	10dB	15dB
Baseline	3.28	13.39	39.05	62.28
DCICA-MBFE	11.32	30.57	54.78	66.81
Conv. DNN	-0.82	36.63	60.25	70.44
DNN+ NAT	1.29	39.2	66.22	75.52
DNN+ IDUET	63.49	79.7	88.44	90.9
DNN+ BF-Softmask	42.95	72.71	86.02	89.93
Proposed method	67.08	84.81	91.1	93.4

【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 5의 7번째 줄

【변경전】

상기 DCICA부가 출력하는

【변경후】

DCICA부가 출력하는