(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년02월06일
(11) 등록번호 10-2496767
(24) 등록일자 2023년02월02일

(51) 국제특허분류(Int. Cl.)
G10L 21/0216 (2013.01) *G10L 25/18* (2013.01)
(52) CPC특허분류
G10L 21/0216 (2013.01)
G10L 25/18 (2013.01)
(21) 출원번호 10-2021-0159916
(22) 출원일자 2021년11월19일
심사청구일자 2021년11월19일
(56) 선행기술조사문헌
Jahn Heymann et al., 'BLSTM SUPPORTED GEV BEAMFORMER FRONT-END FOR THE 3RD CHIME CHALLENGE', ASRU, December 2015.*
Yu Takahashi et al., 'STRUCTURE SELECTION ALGORITHM FOR LESS MUSICAL-NOISE GENERATION IN INTEGRATION SYSTEMS OF BEAMFORMING AND SPECTRAL SUBTRA', IEEE/SP 15th Works. on Stat. Sig. Proc., 2009.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
충북대학교 산학협력단
충청북도 청주시 서원구 충대로 1 (개신동)
(72) 발명자
권오욱
세종특별자치시 남세종로 357, 109동 1105호
윤성욱
충청북도 청주시 서원구 모충로3번길 47, 304호(채림빌)
(74) 대리인
김정현

전체 청구항 수 : 총 1 항 심사관 : 정성윤

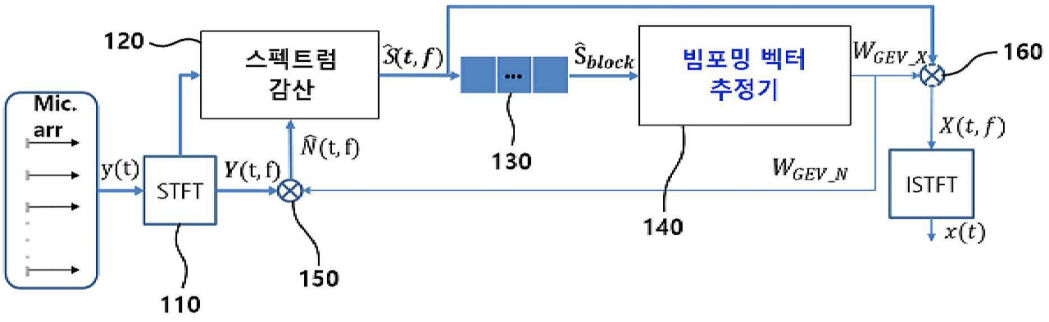
(54) 발명의 명칭 **BLSTM 기반 스펙트럼 감산 온라인 빔포밍 시스템**

(57) 요 약

본 발명은 스펙트럼 감산 온라인 빔포밍 시스템에 관한 것으로서, 관측 신호를 입력으로 하여, 음성 강화에 사용되는 음성 강화 빔포밍 벡터와 잡음 추정에 사용되는 잡음 강화 빔포밍 벡터를 추정하기 위한 빔포밍 벡터 추정기 및 상기 잡음 강화 빔포밍 벡터의 스펙트럼을 감산하기 위한 스펙트럼 감산부를 포함한다.

본 발명에 의하면, BLSTM 마스크 추정 값을 이용한 온라인 빔포밍 업데이트 알고리즘을 제안함으로써, 시시각각 변하는 음성, 잡음 및 발화자의 위치를 반영하여 적응된 빔포밍 벡터를 출력할 수 있는 효과가 있다.

대 표 도 - 도1



(52) CPC특허분류
G10L 2021/02166 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1415162327
과제번호 10080681
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 자동차산업핵심기술개발(R&D)
연구과제명 차량 주행 환경에서 90% 이상 대화음성인식이 가능한 음성인식 요소기술 개발 및 대화형 컴패니언 시스템 개발
기 여 율 1/1
과제수행기관명 주식회사 셀바스에이아이
연구기간 2017.09.01 ~ 2019.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

입력되는 관측 신호에 대해 단기 푸리에 변환(Short Time Fourier Transform)을 수행하기 위한 STFT;

관측 신호를 입력으로 하여, 음성 강화에 사용되는 음성 강화 빔포밍 벡터 W_{GEV_X}와 잡음 추정에 사용되는 잡음 강화 빔포밍 벡터 W_{GEV_N}를 추정하기 위한 빔포밍 벡터 추정기;

상기 STFT에서 출력된 신호 Y(t,f)와 상기 잡음 강화 빔포밍 벡터 W_{GEV_N}를 곱하기 위한 제1 곱셈부;

상기 제1 곱셈부에서 출력된 잡음 강화 빔포밍 벡터의 스펙트럼을 감산하고, 이렇게 스펙트럼이 감산된 신호인 향상 신호 스펙트럼 $\hat{S}(t,f)$ 을 출력하기 위한 스펙트럼 감산부;

상기 스펙트럼 감산부에서 출력된 향상 신호 스펙트럼 $\hat{S}(t,f)$ 을 입력받아 버퍼 길이에 따른 블록단위의 향상된 신호를 출력하여 상기 빔포밍 벡터 추정기에 전달하기 위한 링버퍼(ringbuffer);

상기 음성 강화 빔포밍 벡터 W_{GEV_X}와 상기 향상 신호 스펙트럼 $\hat{S}(t,f)$ 을 곱하여 X(t,f) 신호를 출력하기 위한 제2 곱셈부; 및

상기 제2 곱셈부에서 출력된 X(t,f) 신호에 대해 역 단기 푸리에 변환(Inverse Short Time Fourier Transform)을 수행하여 단일 채널의 빔포밍된 신호 x(t)를 출력하기 위한 ISTFT

를 포함하여 이루어지고,

상기 빔포밍 벡터 추정기는 딥러닝 기반 마스크 추정을 수행하는 딥러닝 기반 마스크 추정기를 포함하여 이루어지고,

상기 딥러닝 기반 마스크 추정기는 BLSTM(Bidirectional Long Short-Term Memory) 기반 마스크 추정기이며,

상기 BLSTM 기반 마스크 추정기에 입력되는 신호 y_{block}을,

$$y_{block} = [y(t), \ldots, y(t-L+1)] \quad \text{(수학식 1)}$$

로 나타낼 수 있고, 여기서 y_{block}은 한 블록단위 입력 배치(batch)를 의미하는 신호이고, L은 한 블록 안에 포함된 프레임 수이고,

블록 단위로 추정된 마스크 값은 다음 수학식 2와 같이 계산되고,

$$M_v^l(f) = \sum_{t=1}^{L} M_v^l(t,f)/L \,, v \in \{X,N\} \quad \text{(수학식 2)}$$

여기서, t는 프레임 인덱스, f는 주파수 빈(bin)의 인덱스, l은 블록 인덱스, v는 잡음(N) 또는 음성(x) 클래스를 의미하고,

이전 블록에서 추정된 $M^l_y(t,f)$ 추정 마스크 값 및 입력 프레임을 STFT후, 매그니튜드(magnitude)를 취한 Y(t,f)를 입력으로 하여, 다음 수학식 3과 같이 l번째 블록에서 추정된 PSD(Power Spectral Density) 행렬 $\Phi^l_w(f)$를 계산하고,

$$\bar{\Phi}_w^l(f) = \sum_{t=1}^{L} M_v(t,f) Y(t,f) Y(t,f)^H \,, v \in \{X,N\} \quad \text{(수학식 3)}$$

여기서, H는 에르미트 연산자(Hermitian operator)이며, PSD 행렬의 차원은 F×C×C으로 다채널 마이크 사이의 음성과 잡음의 전력 분포를 의미하고,

추정된 $\phi^{l}_{w}(f)$는 다음 수학식 4와 같이 가중치 $\alpha^{l}_{w}(f)$를 이용해 누적 추정된 $\phi^{l-1}_{w}(f)$와 가중 합산되고, l번째 블록에서 누적 추정된 PSD $\phi^{l}_{w}(f)$가 얻어지고,

$$\Phi_{vv}^{l}(f)=\begin{cases}\sum_{t=1}^{L} M_v(t,f)Y(t,f)Y(t,f)^{H}, & l=1\\ \alpha_v^{l}(f)\hat{\Phi}_{vv}^{l}(f)+(1-\alpha_v^{l}(f))\Phi_{vv}^{l-1}(f), & \text{otherwise}\end{cases} \quad \text{(수학식 4)}$$

구해진 PSD 행렬 ϕ^{l}_{w}는 길이가 k인 링버퍼(ringbuffer)로 입력되며, 링버퍼가 차면 다음 수학식 5와 같이 계산되고,

$$\Phi_{vv}(f)=\sum_{i=0}^{K-1}\beta_i\Phi_{vv(i)}^{l-i}(f), \quad v\in\{X,N\} \quad \text{(수학식 5)}$$

여기서, β_i는 링버퍼의 i번째 자리의 PSD 행렬 가중치이고,

본 발명에서 GEV(Generalized Eigen Value) 빔포밍을 사용하는 빔포밍 알고리즘을 제안하고, 다음 수학식 6과 같이 Rayleigh coefficient에서 주파수 빈별 SNR을 최대화함으로써 구할 수 있고,

$$W_{GEV}=\arg\min_{W}\frac{W^{H}\Phi_{XX}W}{W^{H}\Phi_{NN}W} \quad \text{(수학식 6)},$$

이를 최적화하면,

$$W_{GEV}=P(\Phi_{NN}^{-1}\Phi_{XX}) \quad \text{(수학식 7)}$$

과 같이 나타낼 수 있고,

상기 수학식 6 및 수학식 7에서 음성 PSD 행렬 Φ_{XX}와 잡음 PSD 행렬 Φ_{NN}을 이용하여 잡음을 추정하는 빔포밍 벡터를 얻을 수 있는 것을 특징으로 하는 스펙트럼 감산 온라인 빔포밍 시스템.

청구항 2

삭제

청구항 3

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 BLSTM(Bidirectional Long Short-Term Memory) 기반 온라인 빔포밍과 스펙트럼 감산을 결합하는 기술에 관한 것이다.

배 경 기 술

[0002] 스펙트럼과 공간 정보를 활용한 다채널 음성 향상은 자동 음성인식(Automatic Speech Recognition, ASR)의 성능 향상에 효과적인 방법임이 입증되었다. 전통적인 다채널 음성 향상 방법으로 다채널 음수 미포함 행렬 분해

(Multichannel Nonnegative Matrix Factorization, MNMF), 다채널 위너 필터(Multichannel Wiener Filter, MWF), 빔포밍(beamforming)이 ASR 성능 향상을 위한 주요 기술로 사용되었다.

[0003] 이중 빔포밍은 성능 향상에 가장 중요한 기술로서 최소 분산 무왜곡 응답(Minimum-Variance Distortionless Response, MVDR), 일반화 고유값(Generalized Eigen Value, GEV), 일반 부엽 제거기(Generalized Sidelobe Canceller, GSC) 등이 활발하게 연구되었다.

[0004] 최근 심층신경망(Deep Neural Network, DNN)이 ASR에서 주목할 만한 성능 향상을 보였으며, 딥러닝을 이용한 최신의 빔포밍 기술로 시간-주파수(Time-Frequency, T-F) 마스크 추정 방법이 제안되었다. 많은 연구에서 딥러닝 기반 마스크 추정 빔포밍이 성공적으로 적용되었고, 전통적인 빔포밍의 성능을 뛰어넘었다. 이처럼 딥러닝이 성공적으로 빔포밍에 적용됨에 따라 실제 환경에서도 어느 정도 안정적인 성능을 보였다.

[0005] 다채널 마이크를 활용한 빔포밍 기술은 잡음제거 및 음성강화에 효과적이지만, 기존 빔포밍 알고리즘은 음성과 잡음이 완전히 겹쳐진 사전 발화를 대상으로 주로 연구되었다. 그러나 이는 실제 환경에 적용하기에 적합하지 않다는 문제점이 있다. 즉, 실제 사용 환경을 고려하면 잡음은 항상 존재하지만, 사용자 음성이 존재하는 구간이 희박한 연속된 잡음 및 잡음 음성 스트림을 처리하여야 한다. 이를 위해 시간에 따라 변하는 입력에 적응하는 온라인 빔포밍 알고리즘이 필요하고, 잡음만이 존재하는 프레임 입력은 온라인 빔포밍 알고리즘의 성능 열화로 이어지기 때문에 이에 대한 대책이 필요하다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 등록특허 10-2236471

발명의 내용

해결하려는 과제

[0007] 본 발명은 상기와 같은 문제점을 해결하기 위하여 안출된 것으로서, BLSTM(Bidirectional Long Short-Term Memory) 기반 마스크 추정 값을 활용하여 시간에 따라 변화하는 입력에 적응하는 빔포밍 벡터를 출력하고, 음성이 입력으로 들어오는 시점에서 빠르게 수렴하는 빔포밍 벡터 추정을 위한 시스템을 제안하는데 그 목적이 있다.

[0008] 또한, 본 발명은 낮은 신호 대 잡음비(Signal to Noise Ratio, SNR) 환경에서 전반적인 성능 향상을 위해 스펙트럼 감산을 빔포밍 기술과 결합하는 방법을 제공하는데 그 다른 목적이 있다.

[0009] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 또 다른 목적들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0011] 이와 같은 목적을 달성하기 위한 본 발명은 스펙트럼 감산 온라인 빔포밍 시스템에 관한 것으로서, 관측 신호를 입력으로 하여, 음성 강화에 사용되는 음성 강화 빔포밍 벡터와 잡음 추정에 사용되는 잡음 강화 빔포밍 벡터를 추정하기 위한 빔포밍 벡터 추정기 및 상기 잡음 강화 빔포밍 벡터의 스펙트럼을 감산하기 위한 스펙트럼 감산부를 포함한다.

[0012] 상기 빔포밍 벡터 추정기는 딥러닝 기반 마스크 추정을 수행하는 딥러닝 기반 마스크 추정기를 포함하여 이루어질 수 있다.

[0013] 상기 딥러닝 기반 마스크 추정기는 BLSTM(Bidirectional Long Short-Term Memory) 기반 마스크 추정기로 구현될 수 있다.

발명의 효과

[0014] 본 발명에 의하면, BLSTM 마스크 추정 값을 이용한 온라인 빔포밍 업데이트 알고리즘을 제안함으로써, 시시각각 변하는 음성, 잡음 및 발화자의 위치를 반영하여 적응된 빔포밍 벡터를 출력할 수 있는 효과가 있다.

[0015] 또한, 음성이 희박하고 잡음이 존재하는 환경에서 링버퍼를 사용해 안정적인 빔포밍 벡터 계산을 보장하는 동시에 실시간성을 확보할 수 있는 효과가 있다.

[0016] 또한, 블록 배치를 나누어 처리함으로써, PSD(Power Spectral Density) 행렬의 수렴을 빠르게 하고, 스펙트럼 감산을 GEV(Generalized Eigen Value) 빔포밍에 결합함으로써 낮은 SNR에서 성능을 향상시킬 수 있는 효과가 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 일 실시예에 따른 BLSTM 기반 스펙트럼 감산 온라인 빔포밍 시스템의 전체 블록도이다.

도 2는 본 발명의 일 실시예에 따른 빔포밍 벡터 추정기의 블록도이다.

도 3은 본 발명의 일 실시예에 따른 마스크 추정을 위한 신경망을 도시한 것이다.

도 4는 본 발명의 일 실시예에 따른 블록단위 PSD 행렬 업데이트 알고리즘을 도시한 것이다.

도 5는 본 발명의 일 실시예에 따른 PSD 행렬의 빠른 수렴을 위한 빔포밍 벡터 추정기의 병렬처리 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시 예를 가질 수 있는 바, 특정 실시 예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0019] 본 출원에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0020] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 갖는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0021] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 본 발명을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0022] 본 발명은 스펙트럼 감산 온라인 빔포밍 시스템에 관한 것이다.

[0023] 도 1은 본 발명의 일 실시예에 따른 BLSTM 기반 스펙트럼 감산 온라인 빔포밍 시스템의 전체 블록도이다.

[0024] 도 1을 참조하면, 본 발명의 스펙트럼 감산 온라인 빔포밍 시스템은 STFT(Short Time Fourier Transform)(110), 스펙트럼 감산부(120), 링버퍼(ringbuffer)(130), 빔포밍 벡터 추정기(140), 제1 곱셈부(150), 제2 곱셈부(160)를 포함한다.

[0025] STFT(110)는 입력되는 관측 신호에 대해 단기 푸리에 변환(Short Time Fourier Transform)을 수행하는 역할을 한다. 본 발명의 일 실시예에서 관측 신호는 다채널 음성 신호일 수 있다.

[0026] 스펙트럼 감산부(120)는 추정된 잡음 스펙트럼 신호의 스펙트럼을 감산하는 역할을 한다.

[0027] 본 발명의 스펙트럼 감산부(120)에서 스펙트럼이 감산된 향상 신호 스펙트럼 $\hat{S}(t,f)$이 출력된다.

[0028] 본 발명에서 링버퍼(130)의 길이가 K라고 할 때, 링버퍼(130)에 향상 신호 스펙트럼이 입력되고, 블록단위의 향상된 신호 $\hat{S}_{block}$를 출력한다.

[0029] 빔포밍 벡터 추정기(Beamforming vector estimator)(140)는 관측 신호를 입력으로 하여, 음성 강화에 사용되는

음성 강화 빔포밍 벡터 W_{GEV_x}와 잡음 추정에 사용되는 잡음 강화 빔포밍 벡터 W_{GEV_N}를 추정한다.

[0030] 스펙트럼 감산부(Spectral subtraction)(120)는 잡음 강화 빔포밍 벡터의 스펙트럼을 감산하는 역할을 한다.

[0031] 제1 곱셈부(150)는 관측신호 스펙트럼과 잡음 강화 빔포밍 벡터를 곱하는 역할을 한다.

[0032] 제2 곱셈부(160)는 음성 강화 빔포밍 벡터와 향상 신호 스펙트럼을 곱하는 역할을 한다.

[0033] 제2 곱셈부(160)에서 출력된 신호 X(t,f)는 역 단기 푸리에 변환(Inverse Short Time Fourier Transform)을 수행하는 ISTFT를 거쳐 단일 채널의 빔포밍된 신호 x(t)가 된다.

[0034] 도 1에서 관측 신호 y(t)를 입력으로 하여, 음성 강화 빔포밍 벡터 W_{GEV_x}와 잡음 강화 빔포밍 벡터 W_{GEV_N}가 동시에 추정되며, W_{GEV_x}는 음성 강화에 사용되며, W_{GEV_N}는 잡음 추정에 사용된다. 그리고, 이 신호들은 스펙트럼 감산부(120) 및 빔포밍 벡터 추정기(140)를 거쳐 단일 채널의 빔포밍 된 신호 x(t)가 된다.

[0035] 잡음 강화 빔포밍 벡터 W_{GEV_N}는 다채널 관측 신호 Y(t,f)와 곱해져 잡음 추정에 사용된다.

[0036] 도 2는 본 발명의 일 실시예에 따른 빔포밍 벡터 추정기의 블록도이다.

[0037] 도 2를 참조하면, 빔포밍 벡터 추정기(140)는 딥러닝 기반 마스크 추정을 수행하는 딥러닝 기반 마스크 추정기를 포함하여 이루어질 수 있다.

[0038] 딥러닝 기반 마스크 추정기는 BLSTM(Bidirectional Long Short-Term Memory) 기반 마스크 추정기로 구현될 수 있다.

[0039] 도 2에서 y_{block}을 수학식으로 나타내면 다음과 같다.

[0040] (수학식 1)

[0041] $$y_{block} = [y(t), \ldots, y(t-L+1)]$$

[0042] 스펙트럼 감산을 적용하지 않을 경우, 수학식 1은 빔포밍 벡터 추정기(140)의 입력이다.

[0043] 수학식 1에서 y_{block}은 한 블록단위 입력 배치(batch)를 의미하며, L은 한 블록 안에 포함된 프레임 수이다.

[0044] y_{block}을 입력으로 하여 음성강화 빔포밍 벡터 W_{GEV_x}와 잡음 강화 빔포밍 벡터 W_{GEV_N}이 동시에 추정된다.

[0045] 도 2에서 BLSTM 기반 마스크 추정기가 도시되어 있으며, y_{block}은 STFT(Short Time Fourier Transform)된 후, 매그니튜드(magnitude)를 취하여 BLSTM 기반 마스크 추정에 사용된다. 블록 단위로 추정된 마스크 값 $M^{l}_{v}(t,f)$는 다음 수학식 2와 같이 계산된다.

[0046] (수학식 2)

[0047] $$M^{l}_{v}(f) = \sum_{t=1}^{L} M^{l}_{v}(t,f) / L \ , v \in \{X, N\}$$

[0048] 여기서, t는 프레임 인덱스, f는 주파수 빈(bin)의 인덱스이며, l은 블록 인덱스, v는 잡음(N) 또는 음성(x) 클래스를 의미한다.

[0050] 도 3은 본 발명의 일 실시예에 따른 마스크 추정을 위한 신경망을 도시한 것이다.

[0051] 도 3을 참조하면, 마스크 추정을 위한 신경망은 4층으로 구성된다.

[0052] 도 3의 실시예에서 잡음 음성 스트림을 16 kHz 샘플링 후 1,024 프레임 사이즈, 256 프레임 쉬프트 사이즈를 사용해 STFT(Short Time Fourier Transform)를 수행한다. STFT의 결과로부터 513개의 스펙트럼 크기를 취하여 마스크 추정을 위한 신경망의 입력으로 사용한다.

[0053] 첫 번째 층은 256 출력 유닛 BLSTM 층으로 이루어져 있으며, tanh 을 활성화 함수로 사용한다. BLSTM의 메모리 cell 개수는 1,024개이다.

[0054] 다음 2층과 3층은 513개의 유닛을 갖는 순방향(Feed Forward, FF) 층으로 이루어져 있으며, 정류 선형 유닛(Rectified Linear Unit, ReLU)을 활성화 함수로 사용하며, 입력의 513-포인트 스펙트럼의 크기를 고려하여 513 유닛을 사용한다.

[0055] 4층은 1026 유닛으로 구성되고, 2개의 부분으로 나누어지는데, 1 ~ 513 유닛은 Mx(t,f)를 추정하고, 514 ~ 1026 유닛은 MN(t,f)를 추정한다. 활성화 함수로는 sigmoid를 사용하여, 0 ~ 1 사이의 값을 추정하며, 각 시간-주파수 빈(bin)에서 음성, 잡음 마스크 추정 값의 합이 1이 되는 제약을 두지 않았다.

[0056] 도 4는 본 발명의 일 실시예에 따른 블록단위 PSD 행렬 업데이트 알고리즘을 도시한 것이다.

[0057] 도 4를 참조하면, 이전 블록에서 추정된 도 2의 $M^{l}_{y}(t,f)$ 추정 마스크 값 및 입력 프레임을 STFT후, 매그니튜드(magnitude)를 취한 Y(t,f)를 입력으로 하여, 다음 수학식 3과 같이 l번째 블록에서 추정된 PSD(Power Spectral Density) 행렬 $\Phi^{l}_{w}(f)$를 계산한다.

[0058] (수학식 3)

$$\bar{\Phi}^{l}_{vv}(f)=\sum_{i=1}^{L} M_{v}(t,f)Y(t,f)Y(t,f)^{H},\ v\in\{X,N\}$$

[0059]

[0060] 여기서, H는 에르미트 연산자(Hermitian operator)이며, PSD 행렬의 차원은 F×C×C으로 다채널 마이크 사이의 음성과 잡음의 전력 분포를 의미한다.

[0061] 추정된 $\Phi^{l}_{w}(f)$는 다음 수학식 4와 같이 가중치 $\alpha^{l}_{w}(f)$를 이용해 누적 추정된 $\Phi^{l-1}_{w}(f)$와 가중 합산되고, l번째 블록에서 누적 추정된 PSD $\Phi^{l}_{w}(f)$가 얻어진다.

[0062] (수학식 4)

$$\Phi^{l}_{vv}(f)=\begin{cases}\sum_{i=1}^{L} M_{v}(t,f)Y(t,f)Y(t,f)^{H}, & l=1\\ \alpha^{l}_{v}(f)\bar{\Phi}^{l}_{vv}(f)+(1-\alpha^{l}_{v}(f))\Phi^{l-1}_{vv}(f), & \text{otherwise}\end{cases}$$

[0063]

[0064] 구해진 PSD 행렬 Φ^{l}_{w} 는 길이가 k인 링버퍼(ringbuffer)로 입력되며, 링버퍼가 차면 계산되며, 최종적으로 다음 수학식 5와 같이 계산된다.

[0065] (수학식 5)

$$\Phi_{vv}(f)=\sum_{i=0}^{K-1}\beta_{i}\Phi^{l-i}_{w(i)}(f),\quad v\in\{X,N\}$$

[0066]

[0067] 여기서, β_{i}는 링버퍼의 i번째 자리의 PSD 행렬 가중치이며, 현재 블록 k-1번째에서 o번째로 가면서 작게 셋팅되는데, 현재 블록의 정보를 최대한 많이 반영하고, 과거 블록의 PSD 행렬 정보를 망각하기 위해서이다.

[0068] 블록 단위 업데이트 시 블록의 길이를 짧게 할수록 빔포밍 벡터가 계산되는 시점과 현재 입력 프레임 사이의 시간 지연이 블록 사이즈에 비례해서 줄어들지만, 빔포밍 벡터 계산을 위한 충분한 프레임을 모을 수 없는 경우, PSD 행렬 계산 시 특정 주파수 빈의 샘플 수가 부족하여 부정확한 빔포밍 값이 나올 수 있다.

[0069] 본 발명에서는 이러한 위험을 줄이며, 빔포밍 벡터가 적용되는 시점에서 현재 입력 프레임에서 시간적으로 가까운 잡음 및 음성 정보를 최대한 반영하여 빔포밍하기 위해 링버퍼를 사용한다.

[0070] 본 발명에서 제안하는 빔포밍 온라인 알고리즘에서는 GEV(Generalized Eigen Value) 빔포밍을 사용하며, 다음 수학식 6과 같이 Rayleigh coefficient에서 주파수 빈별 SNR을 최대화함으로써 구할 수 있다.

[0071] (수학식 6)

$$W_{GEV} = \arg\min_{W} \frac{W^H \Phi_{XX} W}{W^H \Phi_{NN} W}$$

[0072]

[0074] 최적화 해답은 다음 수학식 7과 같다.

[0075] (수학식 7)

$$W_{GEV} = P(\Phi_{NN}^{-1} \Phi_{XX})$$

[0076]

[0078] 수학식 6에서 음성 PSD 행렬 Φ_{XX}와 잡음 PSD 행렬 Φ_{NN}을 이용하여 잡음을 추정하는 빔포밍 벡터를 얻을 수 있다.

[0080] 도 5는 본 발명의 일 실시예에 따른 PSD 행렬의 빠른 수렴을 위한 빔포밍 벡터 추정기의 병렬처리 블록도이다.

[0081] 도 5를 참조하면, 블록단위 PSD 행렬 업데이트 시, PSD 행렬의 빠른 수렴을 위해 $\hat{S}_{block}$을 절반으로 나누어 병렬처리 한다. 나누어진 배치 각각은 잡음 및 음성 PSD 행렬 계산에 사용되며, 각 배치로 계산된 PSD 행렬에 평균을 취하여 최종적인 Φ_{XX}와 Φ_{NN}을 추정한다. 후에는 기존과 같은 절차로 GEV 빔포밍 벡터를 계산한다. 블록을 나누어 처리하는 이유는 연속된 음성을 처리하는 온라인 빔포밍 태스크에서 PSD 행렬의 빠른 수렴이 성능 향상에 중요하기 때문에, 블록을 나누어 각각의 PSD 행렬을 업데이트 후 평균을 취하는 것이 전체 블록을 이용해 PSD 행렬을 한번 업데이트하는 것 보다 빠른 수렴으로 이어지는 효과를 기대할 수 있기 때문이다.

[0082] (수학식 8)

$$\left|\hat{S}(t,f)\right| = \left|Y(t,f)\right| - \left|\hat{N}(t,f)\right| * \eta(t,f)$$

[0083]

[0084] 수학식 8과 같이, 각 채널에서 관측신호 스펙트럼 크기 $|Y(t,f)|$에서 추정된 잡음 스펙트럼 크기 $|\hat{N}(t,f)|$를 감산하여 향상 신호 스펙트럼 크기 $|\hat{S}(t,f)|$를 추정한다.

[0085] 감산 가중치 η(t,f)는 관측 신호 스펙트럼 크기와 추정 잡음 스펙트럼 크기에 따라 수학식 9와 같이 정의한다.

[0086] (수학식 9)

$$\eta(t,f) = \begin{cases} |Y(t,f)| / |\hat{N}(t,f)| * \lambda_N(f), & |Y(t,f)| < |\hat{N}(t,f)| \\ \left(|Y(t,f)| - |\hat{N}(t,f)|\right) / |\hat{N}(t,f)| * \lambda_N(f), & \text{otherwise} \end{cases}$$

[0087]

[0088] (수학식 10)

$$\lambda_N(f) = \sum_{t=L/2+1}^{L} M_N(t,f) / (L/2)$$

[0089]

[0090] 수학식 10과 같이, $\lambda_N(f)$는 잡음 마스크 추정 값을 사용하여 정의되며, 블록단위 빔포밍 벡터 업데이트에서 사용된 L개의 프레임 중 시간적으로 현재에 가까운 L/2 개의 프레임만을 사용한다. 그 이유는 현재 시점의 프레임에 포함된 잡음 주파수 빈 별 분포를 최대한 고려해서 감산하기 위해서이다. 만약 L 개의 프레임을 사용하면 적용 시점의 잡음 분포와 유사성이 떨어져 성능 저하로 이어진다.

[0091] 추정된 향상 신호 스펙트럼 복원을 위해서, 위상 정보가 필요하다. 짧은 구간의 위상 정보는 상대적으로 주요하지 않기 때문에, 관측 신호 스펙트럼의 위상 정보를 사용하여 다음 수학식 11과 같이 복원한다.

[0092] (수학식 11)

$$\hat{S}(t,f)=\left|\hat{S}(t,f)\right|e^{j\angle Y(t,f)}$$

[0093]

[0094] 이렇게 스펙트럼 감산으로 추정된 향상 신호 스펙트럼 $\hat{S}(t,f)$ 는 길이가 L인 링버퍼로 들어가고, 링버퍼가 차면 수학식 12의 블록 단위의 향상된 신호 $\hat{S}_{block}$ 가 빔포밍 벡터 추정기의 입력으로 들어간다.

[0095] (수학식 12)

$$\hat{S}_{block}=[\hat{S}(t,f),\ldots,\hat{S}(t-L+1,f)]$$

[0096]

[0098] 이상에서 기술한 바와 같이, 온라인 빔포밍에서는 시시각각 변하는 음성 및 잡음, 발화자의 위치를 반영하여 적응된 빔포밍 벡터를 출력하는 것이 중요하다. 이를 위해, 본 발명에서는 BLSTM 마스크 추정 값을 이용한 온라인 빔포밍 업데이트 알고리즘을 제안한다.

[0099] 또한 실제 사용 환경을 고려할 때, 잡음은 항상 존재하며 음성이 희박한 환경에서 음성이 들어오는 시점으로부터 빠르게 안정적인 빔포밍 벡터를 계산하는 것이 성능에 중요하다. 그러나 빔포밍의 수학 수식상 관측신호의 길이가 충분히 길지 않으면 안정적인 빔포밍 벡터를 계산하기 어렵다. 이를 위해 본 발명에서는 링버퍼를 사용해 안정적인 빔포밍 벡터 계산을 보장하는 동시에 실시간성을 확보한다.

[0100] 추가적으로 PSD 행렬의 빠른 수렴을 위해서 블록 배치를 나누어 처리하고, 마지막으로 낮은 SNR에서 성능 향상을 위해 스펙트럼 감산을 GEV 빔포밍에 결합한다.

[0102] 이상 본 발명을 몇 가지 바람직한 실시 예를 사용하여 설명하였으나, 이들 실시 예는 예시적인 것이며 한정적인 것이 아니다. 본 발명이 속하는 기술분야에서 통상의 지식을 지닌 자라면 본 발명의 사상과 첨부된 특허청구범위에 제시된 권리범위에서 벗어나지 않으면서 다양한 변화와 수정을 가할 수 있음을 이해할 것이다.

부호의 설명

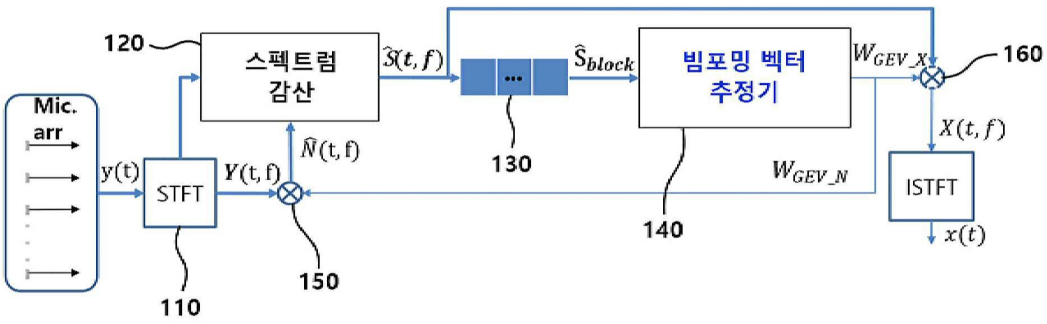
[0104] 110 STFT 120 스펙트럼 감산부

130 링버퍼 140 빔포밍 벡터 측정기

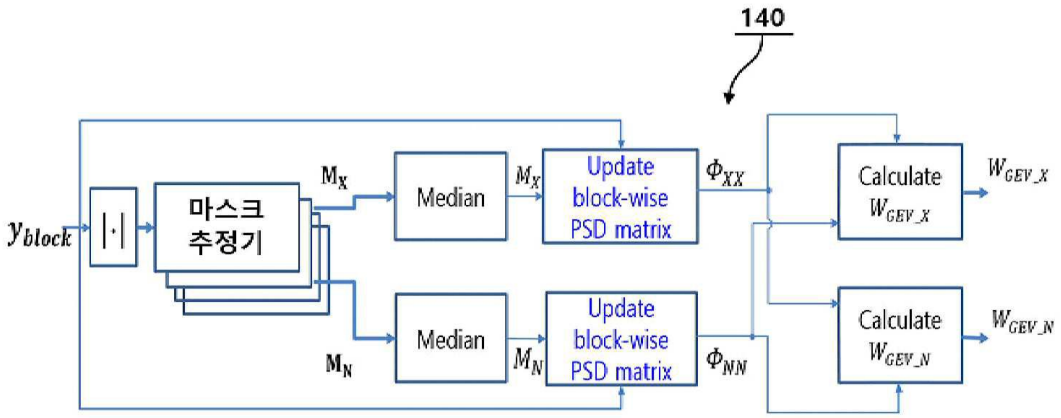
150 제1 곱셈부 160 제2 곱셈부

도면

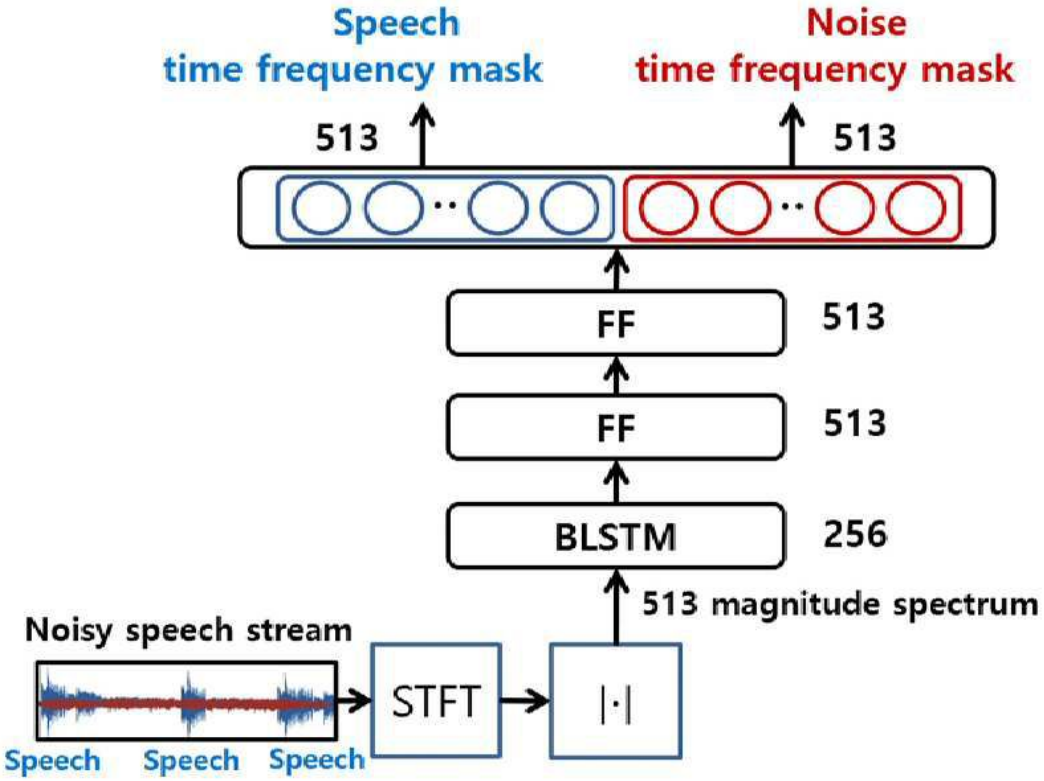
도면1



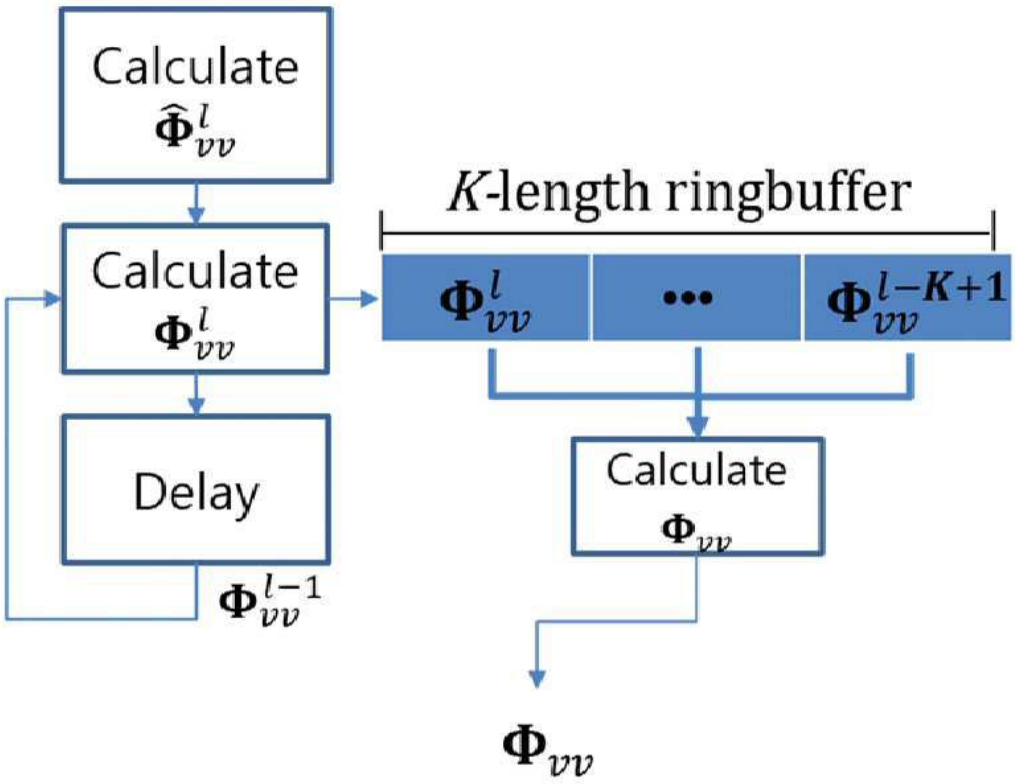
도면2


140
y_{block}
|·|
마스크 추정기
M_X
Median
M_X
Update block-wise PSD matrix
Φ_{XX}
Calculate W_{GEV_X}
W_{GEV_X}
M_N
Median
M_N
Update block-wise PSD matrix
Φ_{NN}
Calculate W_{GEV_N}
W_{GEV_N}

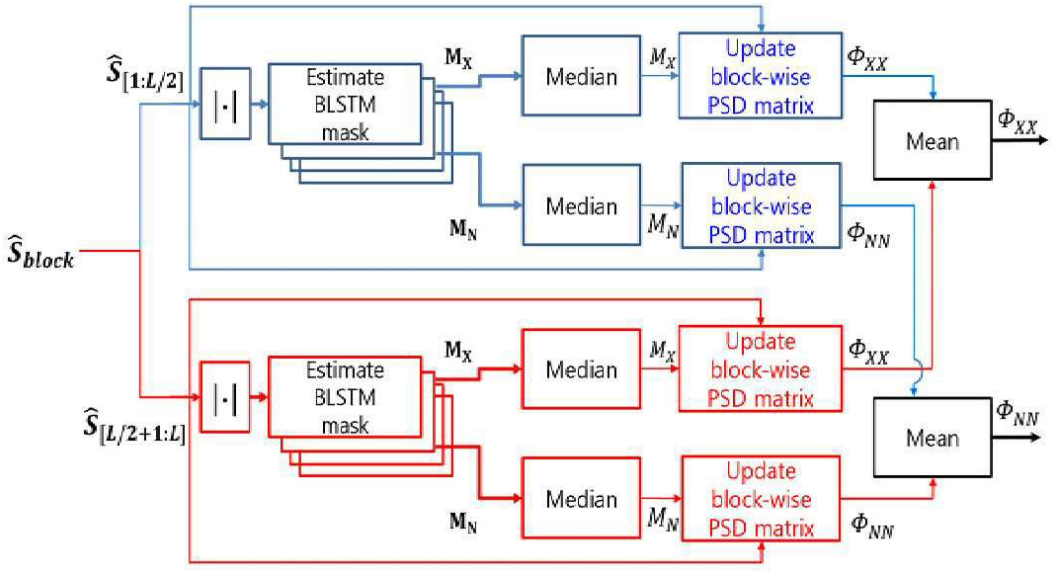
도면3


Speech time frequency mask
Noise time frequency mask
513
513
FF
513
FF
513
BLSTM
256
513 magnitude spectrum
Noisy speech stream
STFT
|·|
Speech
Speech
Speech

도면4


Calculate
$\hat{\Phi}_{vv}^{l}$
Calculate
Φ_{vv}^{l}
Delay
Φ_{vv}^{l-1}
K-length ringbuffer
Φ_{vv}^{l}
•••
Φ_{vv}^{l-K+1}
Calculate
Φ_{vv}
Φ_{vv}

도면5


$\hat{S}_{[1:L/2]}$
$\hat{S}_{block}$
$\hat{S}_{[L/2+1:L]}$
|·|
Estimate BLSTM mask
M_X
M_N
Median
Update block-wise PSD matrix
Φ_{XX}
Φ_{NN}
Mean