(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년01월05일
(11) 등록번호 10-2198598
(24) 등록일자 2020년12월29일

(51) 국제특허분류(Int. Cl.)
G10L 13/02 (2006.01) *G10L 13/08* (2006.01)
G10L 19/06 (2006.01) *G10L 19/087* (2013.01)
G10L 25/30 (2013.01)
(52) CPC특허분류
G10L 13/02 (2013.01)
G10L 13/08 (2013.01)
(21) 출원번호 10-2019-0004085
(22) 출원일자 2019년01월11일
심사청구일자 2019년01월11일
(65) 공개번호 10-2020-0092501
(43) 공개일자 2020년08월04일
(56) 선행기술조사문헌
KR1020090129450 A*
KR1020170107683 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
네이버 주식회사
경기도 성남시 분당구 불정로 6, 그린팩토리 (정자동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
송은우
경기도 성남시 분당구 불정로 6(정자동, 그린팩토리)
김진섭
경기도 성남시 분당구 불정로 6(정자동, 그린팩토리)
(뒷면에 계속)
(74) 대리인
양성보

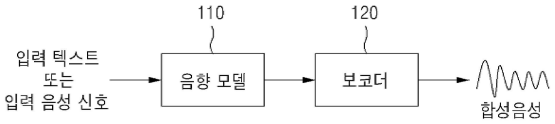
전체 청구항 수 : 총 15 항 심사관 : 정성윤

(54) 발명의 명칭 **합성 음성 신호 생성 방법, 뉴럴 보코더 및 뉴럴 보코더의 훈련 방법**

(57) 요 약

스펙트럼 관련 파라미터들 및 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하고, 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정하고, 추정된 여기 신호에 대해 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 뉴럴 보코더에 의한 음성 신호 생성 방법이 제공된다.

대 표 도 - 도1



(52) CPC특허분류
G10L 19/06 (2013.01)
G10L 19/087 (2013.01)
G10L 25/30 (2013.01)
(72) 발명자
강홍구
서울특별시 은평구 백련산로 38, 206동 301호(응암동, 백련산 힐스테이트 2차)

공지예외적용 : 있음

변경근
서울특별시 서대문구 성산로22길 2-1, 205호(창천동, 창천오피스텔)

명 세 서

청구범위

청구항 1

컴퓨터에 의해 구현되는 뉴럴 보코더(neural vocoder)에 의해 수행되는 음성 신호 생성 방법에 있어서,

스펙트럼 관련 파라미터들(spectral parameter) 및 여기(excitation)의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하는 단계;

상기 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정하는 단계; 및

상기 추정된 여기 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 단계

를 포함하고,

상기 뉴럴 보코더는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것이고,

상기 훈련은,

상기 입력된 음성 신호에 대해 선형 예측 분석 필터(Linear prediction analysis filter)를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 단계; 및

상기 분리된 여기 신호의 확률 분포를 모델링하는 단계

를 포함하고,

상기 여기 신호를 추정하는 단계는,

상기 모델링된 여기 신호의 확률 분포를 사용하여, 상기 복수의 음향 파라미터들에 대한 여기 신호를 추정하는, 음성 신호 생성 방법.

청구항 2

제1항에 있어서,

상기 여기 관련 파라미터들은 소정의 컷오프 주파수를 이하의 여기를 나타내는 제1 여기 파라미터 및 상기 컷오프 주파수 초과의 여기를 나타내는 제2 여기 파라미터를 포함하는, 음성 신호 생성 방법.

청구항 3

제2항에 있어서,

상기 제1 여기 파라미터는 상기 여기의 고조파 스펙트럼(harmonic spectrum)을 나타내고, 상기 제2 여기 파라미터는 상기 여기의 그 외의 부분을 나타내는, 음성 신호 생성 방법.

청구항 4

제1항에 있어서,

상기 스펙트럼 관련 파라미터들은,

음성 신호의 피치를 나타내는 주파수 파라미터, 음성 신호의 에너지를 나타내는 에너지 파라미터, 음성 신호의 유성음(voice) 또는 무성음(unvoice) 여부를 나타내는 파라미터 및 음성 신호의 라인 스펙트럼 주파수(Line Spectral Frequency; LSF)를 나타내는 파라미터를 포함하는, 음성 신호 생성 방법.

청구항 5

제4항에 있어서,

상기 타겟 음성 신호를 생성하는 단계는,

상기 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 단계; 및

상기 추정된 여기 신호에 대해 상기 변환된 LPC에 기반한 상기 선형 합성 필터를 적용하는 단계

를 포함하는, 음성 신호 생성 방법.

청구항 6

제1항에 있어서,

상기 복수의 음향 파라미터들은 입력된 텍스트 또는 입력된 음성 신호에 기반하여 음향 모델(acoustic model)에 의해 생성된 것인, 음성 신호 생성 방법.

청구항 7

삭제

청구항 8

제1항에 있어서,

상기 여기 신호를 분리하는 단계는,

상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 단계; 및

상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용하는 단계

를 포함하는, 음성 신호 생성 방법.

청구항 9

제1항에 있어서,

상기 분리된 여기 신호는 상기 입력된 음성 신호에 대한 잔차 성분(residual component)인, 음성 신호 생성 방법.

청구항 10

컴퓨터에 의해 구현되는 뉴럴 보코더의 훈련 방법에 있어서,

상기 뉴럴 보코더의 훈련을 위한 음성 신호를 입력 받는 단계;

상기 입력된 음성 신호로부터 스펙트럼 관련 파라미터들 및 여기의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 추출하는 단계;

상기 입력된 음성 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 예측 분석 필터를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 단계; 및

상기 분리된 여기 신호의 확률 분포를 모델링하는 단계

를 포함하고,

상기 모델링된 여기 신호의 확률 분포는 획득된 다른 스펙트럼 관련 파라미터들 및 다른 여기 관련 파라미터들을 포함하는 다른 음향 파라미터들에 대해 여기 신호를 추정하기 위해 사용되는, 뉴럴 보코더의 훈련 방법.

청구항 11

제10항에 있어서,

상기 여기 신호를 분리하는 단계는,

상기 스펙트럼 관련 파라미터들 중 상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 LPC로 변환하는 단계;

및

상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용하는 단계

를 포함하는, 뉴럴 보코더의 훈련 방법.

청구항 12

제10항에 있어서,

상기 여기 관련 파라미터들은 소정의 컷오프 주파수를 이하의 여기를 나타내는 제1 여기 파라미터 및 상기 컷오프 주파수 초과의 여기를 나타내는 제2 여기 파라미터를 포함하는, 뉴럴 보코더의 훈련 방법.

청구항 13

제1항 내지 제6항 및 제8항 내지 제12항 중 어느 한 항의 방법을 컴퓨터에 실행시키기 위한 프로그램이 기록되어 있는 비-일시적인 컴퓨터 판독가능한 기록 매체.

청구항 14

뉴럴 보코더에 있어서,

스펙트럼 관련 파라미터들(spectral parameter) 및 여기(excitation)의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하는 파라미터 획득부;

상기 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정하는 여기 신호 추정부; 및

상기 추정된 여기 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 음성 신호 생성부

를 포함하고,

상기 뉴럴 보코더는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것이고,

상기 입력된 음성 신호에 대해 선형 예측 분석 필터(linear prediction analysis filter)를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 여기 신호 분리부; 및

상기 분리된 여기 신호의 확률 분포를 모델링하는 모델링부

를 더 포함하고,

상기 여기 신호 추정부는, 상기 모델링된 여기 신호의 확률 분포를 사용하여, 상기 복수의 음향 파라미터들에 대한 여기 신호를 추정하는, 뉴럴 보코더.

청구항 15

제14항에 있어서,

상기 음성 신호 생성부는 상기 스펙트럼 관련 파라미터들 중 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부를 포함하고,

상기 추정된 여기 신호에 대해 상기 변환된 LPC에 기반한 상기 선형 합성 필터를 적용하는, 뉴럴 보코더.

청구항 16

삭제

청구항 17

제14항에 있어서,

상기 여기 신호 분리부는 상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부를 포함하고,

상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용하는, 뉴럴 보코더.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 뉴럴 보코더를 사용하는 합성 음성 신호 생성 방법과 뉴럴 보코더 및 뉴럴 보코더의 훈련 방법에 관한 것이다.

배 경 기 술

[0002] 음성 합성 기술은 입력된 데이터에 기반하여 사람의 음성과 유사한 합성음을 만들어내는 기술이다. 일례로, TTS(Text to speech)는 입력된 텍스트를 사람의 음성으로 변환하여 제공한다.

[0003] 이러한 합성 음성은 입력된 음향 파라미터에 기반하여 음성 신호를 생성하는 보코더에 의해 생성된다. 최근에는, 인공지능 및 딥러닝 기술의 발전에 따라 합성 음성의 생성에 있어서 인공 신경망을 활용하는 뉴럴 보코더가 제안되고 있다. 뉴럴 보코더는 화자들로부터의 음성 데이터를 사용하여 화자 독립적으로 또는 화자 종속적으로 훈련되고, 훈련의 결과를 사용하여 입력된 음향 파라미터들에 대해 합성 음성 신호를 생성할 수 있다.

[0004] 그러나, 음성 신호는 다이내믹한 특성을 가지므로 인공 신경망(예컨대, CNN)이 이러한 음성 신호의 다이내믹한 특성을 완전히 포착하기는 어렵다. 특히, 음성 신호의 고주파수 영역에 있어서는 스펙트럼 왜곡이 발생하기 쉬우며, 이는 합성 음성 신호의 품질 저하로 이어질 수 있다.

[0005] 따라서, 고주파수 영역에 있어서의 스펙트럼 왜곡을 저감하고 합성 음성 신호의 품질을 높일 수 있으면서, 음성 데이터를 훈련하는 과정 역시 간략화할 수 있는 뉴럴 보코더 시스템이 요구된다.

[0006] 한국공개특허 제10-2018-0113325호(공개일 2018년 10월 16일)는 음성 합성 장치가 음성 파형을 합성함에 있어서, 개발자나 사용자의 의도대로 합성음의 음성을 변조할 수 있도록 음성 합성기의 음성 모델을 부호화하고, 음성 모델 코드를 변환하고, 음성 모델을 복호화하여 변조된 음성 파형을 합성할 수 있도록 하는 기능을 제공하는 음성합성 장치 및 방법을 설명하고 있다.

[0007] 상기에서 설명된 정보는 단지 이해를 돕기 위한 것이며, 종래 기술의 일부를 형성하지 않는 내용을 포함할 수 있으며, 종래 기술이 통상의 기술자에게 제시할 수 있는 것을 포함하지 않을 수 있다.

발명의 내용

해결하려는 과제

[0008] 스펙트럼 관련 파라미터들 및 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하고, 복수의 음향 파라미터들에 기반하여 여기 신호를 추정하고, 추정된 여기 신호에 대해 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 뉴럴 보코더에 의한 음성 신호 생성 방법을 제공할 수 있다.

과제의 해결 수단

[0009] 일 측면에 있어서, 컴퓨터에 의해 구현되는 뉴럴 보코더(neural vocoder)에 의해 수행되는 음성 신호 생성 방법에 있어서, 스펙트럼 관련 파라미터들(spectral parameter) 및 여기(excitation)의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하는 단계, 상기 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정하는 단계 및 상기 추정된 여기 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 단계를 포함하는, 음성 신호 생성 방법이 제공된다.

[0010] 상기 여기 관련 파라미터들은 소정의 컷오프 주파수를 이하의 여기를 나타내는 제1 여기 파라미터 및 상기 컷오프 주파수 초과의 여기를 나타내는 제2 여기 파라미터를 포함할 수 있다.

[0011] 상기 제1 여기 파라미터는 상기 여기의 고조파 스펙트럼(harmonic spectrum)을 나타내고, 상기 제2 여기 파라미터는 상기 여기의 그 외의 부분을 나타낼 수 있다.

[0012] 상기 스펙트럼 관련 파라미터들은, 음성 신호의 피치를 나타내는 주파수 파라미터, 음성 신호의 에너지를 나타내는 에너지 파라미터, 음성 신호의 유성음(voice) 또는 무성음(unvoice) 여부를 나타내는 파라미터 및 음성 신호의 라인 스펙트럼 주파수(Line Spectral Frequency; LSF)를 나타내는 파라미터를 포함할 수 있다.

[0013] 상기 타겟 음성 신호를 생성하는 단계는, 상기 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 단계 및 상기 추정된 여기 신호에 대해 상기 변환된 LPC에 기반한 상기 선형 합성 필터를 적용하는 단계를 포함할 수 있다.

[0014] 상기 복수의 음향 파라미터들은 입력된 텍스트 또는 입력된 음성 신호에 기반하여 음향 모델(acoustic model)에 의해 생성된 것일 수 있다.

[0015] 상기 뉴럴 보코더는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것이고, 상기 훈련은, 상기 입력된 음성 신호에 대해 선형 예측 분석 필터(Linear prediction analysis filter)를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 단계 및 상기 분리된 여기 신호의 확률 분포를 모델링하는 단계를 포함하고, 상기 여기 신호를 추정하는 단계는, 상기 모델링된 여기 신호의 확률 분포를 사용하여, 상기 복수의 음향 파라미터들에 대한 여기 신호를 추정할 수 있다.

[0016] 상기 여기 신호를 분리하는 단계는, 상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 단계 및 상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용하는 단계를 포함할 수 있다.

[0017] 상기 분리된 여기 신호는 상기 입력된 음성 신호에 대한 잔차 성분(residual component)일 수 있다.

[0018] 다른 일 측면에 있어서, 컴퓨터에 의해 구현되는 뉴럴 보코더의 훈련 방법에 있어서, 음성 신호를 입력 받는 단계, 상기 입력된 음성 신호로부터 스펙트럼 관련 파라미터들 및 여기의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 추출하는 단계, 상기 입력된 음성 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 예측 분석 필터를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 단계 및 상기 분리된 여기 신호의 확률 분포를 모델링하는 단계를 포함하는, 뉴럴 보코더의 훈련 방법이 제공된다.

[0019] 상기 여기 신호를 분리하는 단계는, 상기 스펙트럼 관련 파라미터들 중 상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 LPC로 변환하는 단계 및 상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용하는 단계를 포함할 수 있다.

[0020] 상기 여기 관련 파라미터들은 소정의 컷오프 주파수를 이하의 여기를 나타내는 제1 여기 파라미터 및 상기 컷오프 주파수 초과의 여기를 나타내는 제2 여기 파라미터를 포함할 수 있다.

[0021] 또 다른 일 측면에 있어서, 뉴럴 보코더에 있어서, 스펙트럼 관련 파라미터들(spectral parameter) 및 여기(excitation)의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득하는 파라미터 획득부, 상기 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정하는 여기 신호 추정부 및 상기 추정된 여기 신호에 대해 상기 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 합성 필터를 적용함으로써 타겟 음성 신호를 생성하는 음성 신호 생성부를 포함하는, 뉴럴 보코더가 제공된다.

[0022] 상기 음성 신호 생성부는 상기 스펙트럼 관련 파라미터들 중 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부를 포함하고, 상기 추정된 여기 신호에 대해 상기 변환된 LPC에 기반한 상기 선형 합성 필터를 적용할 수 있다

[0023] 상기 뉴럴 보코더는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것이고, 상기 뉴럴 보코더는, 상기 입력된 음성 신호에 대해 선형 예측 분석 필터(linear prediction analysis filter)를 적용함으로써 상기 입력된 음성 신호로부터 여기 신호를 분리하는 여기 신호 분리부 및 상기 분리된 여기 신호의 확률 분포를 모델링하는 모델링부를 더 포함할 수 있고, 상기 여기 신호 추정부는, 상기 모델링된 여기 신호의 확률 분포를 사용하여, 상기 복수의 음향 파라미터들에 대한 여기 신호를 추정할 수 있다.

[0024] 상기 여기 신호 분리부는 상기 입력된 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부를 포함하고, 상기 입력된 음성 신호에 대해 상기 입력된 음성 신호의 변환된 LPC에 기반한 상기 선형 예측 분석 필터를 적용할 수 있다.

발명의 효과

[0025] 뉴럴 보코더가 여기 신호를 타겟으로 하여 추정을 수행하고, 추정된 여기 신호에 대해 선형 예측 필터를 적용하는 것을 통해 타겟 음성 신호가 생성됨으로써, 생성된 타겟 음성 신호의 품질을 높일 수 있으며, 특히, 음성 신호의 고주파수 영역에 있어서의 스펙트럼 왜곡을 저감할 수 있다.

도면의 간단한 설명

[0026] 도 1은 일 실시예에 따른 입력된 텍스트 또는 음성 신호에 기반하여 합성 음성 신호를 생성하는 방법을 나타낸다.

도 2는 일 실시예에 따른 뉴럴 보코더 시스템의 구조를 나타내는 블록도이다.

도 3은 일 실시예에 따른 뉴럴 보코더 시스템의 프로세서의 구조를 나타내는 블록도이다.

도 4는 일 실시예에 따른 음성 신호 생성 방법을 나타내는 흐름도이다.

도 5는 일 실시예에 따른 뉴럴 보코더를 훈련시키는 방법을 나타내는 흐름도이다.

도 6은 일 실시예에 따른 화자 적응형 모델을 구축하여 타겟 화자의 합성 음성을 생성하는 방법을 나타낸다.

도 7은 일 실시예에 따른 뉴럴 보코더의 프로세서의 구조를 나타내는 블록도이다.

도 8은 일 실시예에 따른 화자 적응형 모델을 구축하기 위한 뉴럴 보코더의 훈련 방법을 나타내는 흐름도이다.

도 9는 일 예에 따른 음성 신호 및 여기 신호와 그 관계를 나타낸다.

도 10a 내지 10c는 각각 다른 종류의 보코더를 사용하는 합성 음성 신호 생성을 위한 통계적 파라메트릭 음성 합성(Statistical Parametric Speech Synthesis; SPSS) 시스템을 나타낸다.

도 11 및 도 13은 일 실시예에 따른 훈련을 위해 입력된 음성 신호로부터 여기 신호를 분리함으로서 뉴럴 보코더를 훈련시키는 방법을 나타낸다.

도 12 및 도 14는 일 실시예에 따른 입력 텍스트에 기반하여 음향 모델에 의해 생성된 음향 파라미터들로부터 여기 신호를 추정하여 합성 음성 신호를 생성하는 방법을 나타낸다.

도 15는 일 예에 따른, 훈련 과정/합성 음성 신호의 생성 과정에서 획득된 음의 로그우도(Negative Log-Likelihood; NLL)의 음향 파라미터로서 여기의 주기성에 따라 구분되는 파라미터들의 사용 여부에 따른 차이를 나타내는 그래프이다.

도 16는 일 예에 따른 복수의 화자들로부터의 음성 신호에 대해, 음성 신호의 화자 종속적인 특징과 화자 독립적인 특징을 도식적으로 나타낸다.

도 17은 일 예에 따른 복수의 화자들로부터의 음성 데이터 세트들을 훈련시킴으로써 구축된 소스 모델과, 타겟 화자로부터의 음성 데이터 세트를 훈련시킴으로써 구축된 화자 적응형 모델을 사용하여 타겟 화자의 합성 음성을 생성하는 방법을 나타낸다.

도 18 및 도 19는 일 예에 따른 화자 적응(speaker adaptation) 알고리즘의 적용 여부에 따라 생성된 합성 음성 신호의 품질을 비교 평가한 결과를 나타낸다.

도 20은 일 예에 따른 ExcitNet 보코더와 타 보코더 간의 MOS(Mean Opinion Score)(MOS) 평가 결과를 나타낸다.

도 21은 일 예에 따른 F0 스케일링 팩터(scaling factor)를 상이하게 하는 경우에 있어서 화자 적응형 모델을 구축하는 뉴럴 보코더의 성능 변화를 나타낸다

발명을 실시하기 위한 구체적인 내용

[0027] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0029] 도 1은 일 실시예에 따른 입력된 텍스트 또는 음성 신호에 기반하여 합성 음성 신호를 생성하는 방법을 나타낸다.

[0030] 음성 신호는 음성을 나타낼 수 있고, 아래의 상세한 설명에서는 설명의 편의 상 '음성 신호'와 '음성'의 용어가 서로 혼용하여 사용될 수 있다.

[0031] 음향 모델(acoustic model)(110)은 합성 음성 신호의 생성을 위해 입력된 텍스트 또는 음성 신호로부터 음향 파라미터(들)를 생성할 수 있다. 음향 모델(110)은 딥러닝 기반의 통계적 파라메트릭 음성 합성(Statistical Parametric Speech Synthesis; SPSS) 시스템을 사용하여 설계된 것일 수 있다. 음향 모델(110)은 언어 입력과 음향 출력 파라미터 사이의 비선형 매핑 함수를 나타내기 위해 훈련된, 다중 피드 포워드 및 장단기 메모리 레이어들로 구성될 수 있다. 음향 모델(110)은 예컨대, DNN TTS 모듈일 수 있다. 음향 파라미터는 합성 음성 신호를 생성하기 위해 사용되는 피처이거나, 피처를 구성하기 위해 사용되는 파라미터일 수 있다.

[0032] 보코더(120)는 음향 모델(110)에 의해 생성된 음향 파라미터를 음성 신호로 변환함으로써 합성 음성 신호를 생성할 수 있다. 보코더(120)는 뉴럴 보코더일 수 있다. 뉴럴 보코더는 딥러닝 모델을 사용하여 훈련된 것일 수 있다. 뉴럴 보코더는 예컨대, WaveNet, SampleRNN, 또는 WaveRNN일 수 있다. 또한, 뉴럴 보코더는 이들에 제한되지 않는 일반적인 생성 모델(generative model)일 수 있다.

[0033] "뉴럴 보코더"는 (합성) 음성 신호의 생성을 위해 훈련된 모델(예컨대, WaveNet, SampleRNN, WaveRNN 또는 일반적인 모델) 및 각종 필터를 포함하는 장치를 나타내기 위해 사용될 수 있다.

[0034] 보코더(120)는 음향 모델(110)로부터 획득된 음향 파라미터들에 기반하여 음성 신호의 여기(excitation) 신호를 추정할 수 있다. 즉, 음성 신호의 여기 신호가 보코더(120)의 타겟이 될 수 있다.

[0035] 여기 신호는 음성 신호 중 음성의 떨림을 나타내는 성분으로서, 발화자의 입모양에 따라 변화되는 음성 신호의 변화를 나타내는 성분(스펙트럼 성분(spectral component))과는 구분될 수 있다. 여기 신호의 변화는 발화자의 성대의 움직임(vocal cord movement)에 의한 것으로만 제한될 수 있다. 여기 신호는 음성 신호의 잔차 신호(residual signal)일 수 있다.

[0036] 보코더(120)에 의해 추정된 여기 신호에 대해 음성 신호의 스펙트럼 성분을 나타내는 음향 파라미터에 기반하여 생성된 선형 예측(Linear Prediction) 필터가 적용됨에 따라 타겟 음성 신호(즉, 합성 음성 신호)가 생성될 수 있다.

[0037] 보코더(120)가 음성 신호가 아닌 여기 신호를 타겟으로 하고, 추정된 여기 신호에 대해 선형 예측 필터를 적용하는 것을 통해 타겟 음성 신호가 생성됨으로써, 생성된 타겟 음성 신호의 품질을 높일 수 있으며, 특히, 음성 신호의 고주파수 영역에 있어서의 스펙트럼 왜곡을 저감할 수 있다.

[0038] 여기 신호를 추정하는 것을 통해 타겟 음성 신호를 생성하는 보다 구체적인 방법과, 여기 신호를 추정하기 위해 뉴럴 보코더를 훈련시키는 보다 구체적인 방법에 대해서는 후술될 도 2 내지 5를 참조하여 더 자세하게 설명된다.

[0040] 도 2는 일 실시예에 따른 뉴럴 보코더 시스템의 구조를 나타내는 블록도이다.

[0041] 도 2를 참조하여 뉴럴 보코더 시스템(200)의 보다 상세한 구성에 대해 설명한다. 도시된 뉴럴 보코더 시스템(200)은 뉴럴 보코더를 포함하여 구성되는 컴퓨터(컴퓨터 시스템)을 나타낼 수 있다.

[0042] 뉴럴 보코더 시스템(200)은 컴퓨터 시스템으로 구현되는 고정형 단말이거나 이동형 단말로 구현될 수 있다. 예컨대, 뉴럴 보코더 시스템(200)은 AI 스피커, 스마트폰(smart phone), 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스 등으로 구현될 수 있다. 또는, 뉴럴 보코더 시스템(200)은 전술된 단말과 네트워크를 통해 통신하는 서버 또는 기타 컴퓨팅 장치로 구현될 수 있다.

[0043] 뉴럴 보코더 시스템(200)은 메모리(210), 프로세서(220), 통신 모듈(230) 및 입출력 인터페이스(240)를 포함할 수 있다. 메모리(210)는 비-일시적인 컴퓨터 판독 가능한 기록매체로서, RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 여기서 ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리(210)와는 구분되는 별도의 영구 저장 장치로서 뉴럴 보코더 시스템(200)에 포함될 수도 있다. 또한, 메모리(210)에는 운영체제와 적어도 하나의 프로그램 코드(일례로, 뉴럴 보코더 시스템(200)에 설치되어 구동되는 브라우저나 특정 서비스의 제공을 위해 뉴럴 보코더 시스템(200)에 설치된 어플리케이션 등을 위한 코드)가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 메모리(210)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독가

능한 기록매체를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독가능한 기록매체가 아닌 통신 모듈(230)을 통해 메모리(210)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 어플리케이션의 설치 파일을 배포하는 파일 배포 시스템(예컨대, 외부 서버)을 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(210)에 로딩될 수 있다.

[0044] 프로세서(220)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(210) 또는 통신 모듈(230)에 의해 프로세서(220)로 제공될 수 있다. 예를 들어 프로세서(220)는 메모리(210)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

[0045] 통신 모듈(230)은 네트워크를 통해 뉴럴 보코더 시스템(200)이 다른 전자기기 또는 서버와 서로 통신하기 위한 기능을 제공할 수 있다. 통신 모듈(230)은 뉴럴 보코더 시스템(200)의 네트워크 인터페이스 카드, 네트워크 인터페이스 칩 및 네트워킹 인터페이스 포트 등과 같은 하드웨어 모듈 또는 네트워크 디바이스 드라이버(driver) 또는 네트워킹 프로그램과 같은 소프트웨어 모듈일 수 있다.

[0046] 입출력 인터페이스(240)는 (도시되지 않은) 입출력 장치와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 입력 장치는 키보드, 마우스, 마이크로폰, 카메라 등의 장치를, 그리고 출력 장치는 디스플레이, 화자, 햅틱 피드백 디바이스(haptic feedback device) 등과 같은 장치를 포함할 수 있다. 다른 예로, 입출력 인터페이스(240)는 터치스크린과 같이 입력과 출력을 위한 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수도 있다. 입출력 장치는 뉴럴 보코더 시스템(200)의 구성일 수 있다. 뉴럴 보코더 시스템(200)이 서버로 구현되는 경우, 뉴럴 보코더 시스템(200)은 입출력 장치 및 입출력 인터페이스(240)를 포함하지 않을 수도 있다.

[0047] 또한, 다른 실시예에서 뉴럴 보코더 시스템(200)은 도시된 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요는 없는 바 이는 생략되었다.

[0048] 도 3을 참조하여서는, 프로세서(220)의 보다 세부적인 구성들을 중심으로 여기 신호를 추정하는 것을 통해 타겟 음성 신호를 생성하는 방법과, 여기 신호를 추정하기 위해 뉴럴 보코더를 훈련시키는 방법에 대해 설명한다.

[0049] 이상, 도 1을 참조하여 전술된 기술적 특징에 대한 설명은, 도 2에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0051] 도 3은 일 실시예에 따른 뉴럴 보코더 시스템의 프로세서의 구조를 나타내는 블록도이다.

[0052] 후술될 프로세서(220)의 구성들(310 내지 340)의 각각은 하나 이상의 소프트웨어 모듈 및/또는 하드웨어 모듈로 구현될 수 있다. 실시예에 따라 프로세서(220)의 구성요소들은 선택적으로 프로세서(220)에 포함되거나 제외될 수도 있다. 또한, 실시예에 따라 프로세서(220)의 구성요소들은 프로세서(220)의 기능의 표현을 위해 분리 또는 병합될 수도 있다.

[0053] 프로세서(220)의 구성요소들은 뉴럴 보코더 시스템(200)에 저장된 프로그램 코드가 제공하는 명령에 따라 프로세서(220)에 의해 수행되는 프로세서(220)의 서로 다른 기능들(different functions)의 표현들일 수 있다.

[0054] 프로세서(220)의 파라미터 획득부(310)는 스펙트럼 관련 파라미터들(spectral parameter) 및 여기(excitation)의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득할 수 있다. 파라미터 획득부(310)가 획득하는 복수의 음향 파라미터들은 사용자로부터 입력된 텍스트 또는 화자로부터 입력된 음성 신호에 기반하여 음향 모델(acoustic model)에 의해 생성된 것일 수 있다.

[0055] 프로세서(220)의 여기 신호 추정부(320)는 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정할 수 있다. 여기 신호 추정부(320)(뉴럴 보코더)는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것일 수 있다. 여기 신호 추정부(320)는 훈련을 통해 모델링된 여기 신호의 확률 분포를 사용하여, 복수의 음향 파라미터들에 대한 여기 신호를 추정할 수 있다.

[0056] 프로세서(220)는 뉴럴 보코더의 훈련을 수행하기 위한 구성(340)을 포함할 수 있다. 프로세서(220)의 여기 신호 분리부(342)는 훈련을 위해 입력된 음성 신호에 대해 선형 예측 분석 필터(linear prediction analysis filter)를 적용함으로써 훈련을 위해 입력된 음성 신호로부터 여기 신호를 분리할 수 있다. 여기 신호 분리부(342)는 훈련을 위해 입력된 음성 신호의 라인 스펙트럼 주파수(Line Spectral Frequency; LSF)를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부(343)를 포함할 수 있다. 상기 선형 예측 분석 필터는 LSF를 나타내는 파라미터에 기반하는 것으로서 상기 변환된 LPC에 기반하여 생성되는 것

일 수 있다. 프로세서(220)의 모델링부(344)는 분리된 여기 신호의 확률 분포를 모델링할 수 있다.

[0057] 프로세서(220)의 음성 신호 생성부(330)는 여기 신호 추정부(320)에 의해 추정된 여기 신호에 대해 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 (예측) 합성 필터를 적용함으로써 타겟 음성 신호를 생성할 수 있다. 타겟 음성 신호는 함성된 음성 신호일 수 있다.

[0058] 음성 신호 생성부(330)는 획득된 스펙트럼 관련 파라미터들 중 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환하는 변환부(332)를 포함할 수 있다. 상기 선형 예측 합성 필터는 획득된 스펙트럼 관련 파라미터들 중 음성 신호의 LSF 를 나타내는 파라미터에 기반하는 것으로서 상기 변환된 LPC에 기반하여 생성되는 것일 수 있다. 말하자면, 음성 신호 생성부(330)는 추정된 여기 신호에 대해 변환된 LPC에 기반한 선형 예측 합성 필터를 적용함으로써 타겟 음성 신호를 생성할 수 있다.

[0059] 여기 신호를 추정하는 것을 통해 타겟 음성 신호를 생성하는 보다 구체적인 방법은 후술될 도 4의 흐름도를 참조하여 더 자세하게 설명되고, 여기 신호를 추정하기 위해 뉴럴 보코더를 훈련시키는 보다 구체적인 방법에 대해서는 후술될 도 5의 흐름도를 참조하여 더 자세하게 설명된다.

[0060] 이상, 도 1 및 도 2를 참조하여 전술된 기술적 특징에 대한 설명은, 도 3에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0062] 도 4는 일 실시예에 따른 음성 신호 생성 방법을 나타내는 흐름도이다.

[0063] 단계(410)에서, 파라미터 획득부(310)는 스펙트럼 관련 파라미터들 및 여기의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 획득할 수 있다. 파라미터 획득부(310)가 획득하는 복수의 음향 파라미터들은 사용자로부터 입력된 텍스트 또는 화자로부터 입력된 음성 신호에 기반하여 음향 모델에 의해 생성된 것일 수 있다. 즉, 파라미터 획득부(310)는 음향 모델로부터 상기 복수의 음향 파라미터들을 수신할 수 있다.

[0064] 스펙트럼 관련 파라미터들은 음성 신호를 구성하는 스펙트럼 성분 스펙트럼 성분(spectral component)을 나타내는 파라미터일 수 있다. 여기 관련 파라미터들은 음성 신호에서 스펙트럼 성분을 제외한 잔차 신호(여기 신호)에 해당하는 성분을 나타내는 파라미터일 수 있다. 스펙트럼 성분의 신호는 발화자의 입모양에 따라 변화되는 음성 신호의 부분을 나타낼 수 있다. 여기 신호는 음성 신호 중 음성의 떨림을 나타내는 음성 신호의 부분을 나타낼 수 있다. 여기 신호의 변화는 발화자의 성대의 움직임에 의한 것으로만 제한될 수 있다.

[0065] 스펙트럼 관련 파라미터들은 예컨대, 음성 신호의 피치를 나타내는 주파수 파라미터(F0), 음성 신호의 에너지를 나타내는 에너지 파라미터(일례로, 이득(gain)을 나타내는 파라미터), 음성 신호의 유성음(voice) 또는 무성음(unvoice) 여부를 나타내는 파라미터(v/uv) 및 음성 신호의 라인 스펙트럼 주파수(Line Spectral Frequency; LSF)를 나타내는 파라미터를 포함할 수 있다.

[0066] 여기 관련 파라미터들은 여기의 주기성에 따라 구분되는 파라미터들을 포함할 수 있다. 여기 관련 파라미터들은 예컨대, TFTE(Time-Frequency Trajectory Excitation) 파라미터들일 수 있다. TFTE는 주파수 축을 따르는 여기의 스펙트럼 모양과 시간 축을 따르는 이러한 모양의 전개(evolution)를 나타낼 수 있다. 여기 관련 파라미터들은 여기 신호 중 시간-주파수 축에서 더 천천히 변화하는 성분을 나타내는 제1 여기 파라미터(SEW(Slowly Evolving Waveform) 파라미터) 및 여기 신호 중 시간-주파수 축에서 더 빠르게 변화하는 성분을 나타내는 제2 여기 파라미터(REW(Rapidly Evolving Waveform) 파라미터)를 포함할 수 있다.

[0067] 제1 여기 파라미터는 소정의 컷오프 주파수를 이하의 여기를 나타낼 수 있고, 제2 여기 파라미터는 컷오프 주파수 초과의 여기를 나타낼 수 있다. 제1 여기 파라미터는 여기의 고조파 스펙트럼(harmonic spectrum)을 나타낼 수 있고, 제2 여기 파라미터는 여기의 그 외의 부분을 나타낼 수 있다. 예컨대, 고조파 여기 스펙트럼(harmonic excitation spectrum)에 해당하는 제1 여기 파라미터(SEW 파라미터)는 TFTE의 각 주파수 성분을 시간 영역 축을 따라 (소정의 컷 오프 주파수로) 저역 통과 필터링함으로써 획득될 수 있다. 소정의 컷 오프 주파수를 초과하는 잔류 잡음 스펙트럼은 제2 여기 파라미터(REW 파라미터)로서, TFTE로부터 SEW를 감산하는 것을 통해 획득될 수 있다. 제1 여기 파라미터(SEW 파라미터) 및 제2 여기 파라미터가 사용됨으로써, 여기의 주기성이 보다 효과적으로 표현될 수 있다. 제1 여기 파라미터 및 제2 여기 파라미터는 ITFTE(Improved Time-Frequency Trajectory Excitation) 파라미터에 해당할 수 있다.

[0068] 단계(420)에서, 여기 신호 추정부(320)는 복수의 음향 파라미터들에 기반하여 여기 신호(excitation signal)를 추정할 수 있다. 즉, 여기 신호 추정부(320)는 스펙트럼 관련 파라미터들 및 여기 관련 파라미터들을 입력으로

서 여기 신호를 추정할 수 있다. 추정되는 여기 신호는 여기 신호의 시간 시퀀스(time sequence)일 수 있다.

[0069] 여기 신호 추정부(320)는 훈련을 위해 입력된 음성 신호에 기반하여 훈련된 것으로서, 여기 신호 추정부(320)는 훈련에 따라 모델링된 여기 신호의 확률 분포를 사용하여, 획득된 복수의 음향 파라미터들에 대한 여기 신호를 추정할 수 있다. 여기 신호 추정부(320)를 포함하는 뉴럴 보코더의 훈련 방법에 대해서는 후술될 도 5를 참조하여 더 자세하게 설명된다.

[0070] 여기 신호 추정부(320)는, 예컨대, WaveNet, SampleRNN, 또는 WaveRNN으로 구현될 수 있다. 또한, 여기 신호 추정부(320)는 이들에 제한되지 않는 일반적인 생성 모델(generative model)로 구현될 수 있다.

[0071] 단계(430)에서, 음성 신호 생성부(330)는 여기 신호 추정부(320)에 의해 추정된 여기 신호에 대해 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 (예측) 합성 필터를 적용함으로써 타겟 음성 신호를 생성할 수 있다. 타겟 음성 신호는 함성된 음성 신호일 수 있다. 단계들(432 및 434)을 참조하여 단계(430)를 더 자세하게 설명한다.

[0072] 단계(432)에서, 변환부(332)는 획득된 스펙트럼 관련 파라미터들 중 음성 신호의 LSF를 나타내는 파라미터를 선형 예측 코딩(Linear Predictive Coding; LPC)으로 변환할 수 있다. 선형 예측 합성 필터는 획득된 스펙트럼 관련 파라미터들 중 음성 신호의 LSF 를 나타내는 파라미터에 기반하는 것으로서, 변환된 LPC에 기반하여 생성될 수 있다.

[0073] 단계(434)에서, 음성 신호 생성부(330)는 추정된 여기 신호에 대해 단계(432)에서 변환된 LPC에 기반한 선형 예측 합성 필터를 적용함으로써 타겟 음성 신호를 생성할 수 있다.

[0074] 단계들(410 내지 430)에 의해 생성된 타겟 음성 신호는, 여기 신호를 타겟으로서 추정하지 않고 음성 신호를 바로 추정하여 생성된 음성 신호에 비해 품질이 우수하며, 특히, 음성 신호의 고주파수 영역에 있어서의 스펙트럼 왜곡이 저감될 수 있다.

[0075] 이상, 도 1 내지 도 3을 참조하여 전술된 기술적 특징에 대한 설명은, 도 4에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0077] 도 5는 일 실시예에 따른 뉴럴 보코더를 훈련시키는 방법을 나타내는 흐름도이다.

[0078] 도 5를 참조하여, 획득된 음향 파라미터들에 기반하여 여기 신호를 추정할 수 있는 여기 신호의 확률 분포를 모델링하는 방법을 자세하게 설명한다.

[0079] 단계(510)에서, 뉴럴 보코더 시스템(200)은 훈련을 위한 음성 신호를 수신할 수 있다. 훈련을 위한 음성 신호는 화자로부터 직접 뉴럴 보코더 시스템(200)으로 입력되거나, 음성 신호를 포함하는 데이터가 음성 신호를 수신한 전자기기로부터 전송됨으로써 뉴럴 보코더 시스템(200)으로 입력될 수 있다.

[0080] 단계(520)에서, 뉴럴 보코더 시스템(200)은 입력된 음성 신호로부터 스펙트럼 관련 파라미터들 및 여기의 주기성에 따라 구분되는 여기 관련 파라미터들을 포함하는 복수의 음향 파라미터들을 추출할 수 있다. 뉴럴 보코더 시스템(200)은 음성 분석(speech analysis)을 통해 음성 신호로부터 복수의 음향 파라미터들을 추출할 수 있다. 예컨대, 뉴럴 보코더 시스템(200)은 그 내부 또는 외부에 존재하는 파라메트릭 보코더를 사용하여 음성 신호로부터 복수의 음향 파라미터들을 추출할 수 있다.

[0081] 스펙트럼 관련 파라미터들은 예컨대, 음성 신호의 피치를 나타내는 주파수 파라미터(F0), 음성 신호의 에너지를 나타내는 에너지 파라미터(일례로, 이득(gain)을 나타내는 파라미터), 음성 신호의 유성음(voice) 또는 무성음(unvoice) 여부를 나타내는 파라미터(v/uv) 및 음성 신호의 라인 스펙트럼 주파수(Line Spectral Frequency; LSF)를 나타내는 파라미터를 포함할 수 있다. 여기 관련 파라미터들은 여기의 주기성에 따라 구분되는 파라미터들을 포함할 수 있다. 여기 관련 파라미터들은 예컨대, TFTE(Time-Frequency Trajectory Excitation) 파라미터들일 수 있다. TFTE는 주파수 축을 따르는 여기의 스펙트럼 모양과 시간 축을 따르는 이러한 모양의 전개(evolution)를 나타낼 수 있다. 여기 관련 파라미터들은 여기 신호 중 시간-주파수 축에서 더 천천히 변화하는 성분을 나타내는 SEW 파라미터 및 여기 신호 중 시간-주파수 축에서 더 빠르게 변화하는 성분을 나타내는 REW 파라미터를 포함할 수 있다. SEW 파라미터는 소정의 컷오프 주파수를 이하의 여기를 나타낼 수 있고, REW 파라미터는 컷오프 주파수 초과의 여기를 나타낼 수 있다. SEW 파라미터는 여기의 고조파 스펙트럼(harmonic spectrum)을 나타낼 수 있고, REW 파라미터는 여기의 그 외의 부분을 나타낼 수 있다. 예컨대, 고조파 여기 스펙트럼(harmonic excitation spectrum)에 해당하는 SEW 파라미터는 TFTE의 각 주파수 성분을 시간 영역 축을 따라 (소정의 컷 오프 주파수로) 저역 통과 필터링함으로써 획득될 수 있다. 소정의 컷 오프 주파수를 초과하는

잔류 잡음 스펙트럼은 REW 파라미터로서, TFTE로부터 SEW를 감산하는 것을 통해 획득될 수 있다.

[0082] 전술된 단계들(510 및 520)은 후술될 단계들(530 및 540)과 마찬가지로 뉴럴 보코더 시스템(200)의 프로세서(220)에 의해 수행될 수 있다.

[0083] 단계(530)에서, 여기 신호 분리부(342)는 입력된 음성 신호에 대해 스펙트럼 관련 파라미터들 중 적어도 하나에 기반한 선형 예측 분석 필터(linear prediction analysis filter)를 적용함으로써 입력된 음성 신호로부터 여기 신호를 분리할 수 있다. 선형 예측 분석 필터는 음성 신호로부터 스펙트럼 포먼트(spectral formant) 구조를 분리하는 필터일 수 있다. 분리된 여기 신호는 입력된 음성 신호에 대한 잔차 성분(residual component)(즉, 잔차 신호)일 수 있다. 여기 신호는 정보량을 줄이기 위해 잔차 신호를 펄스 또는 잡음(PoN), 대역 비주기성(BAP), 성문 여기(glottal excitation), 및 시간-주파수 궤적 여기(TFTE) 모델 등과 같은 다양한 유형의 여기 모델들 중 적어도 하나에 의해 근사화된 것일 수 있다.

[0084] 단계(532 및 534)를 참조하여 음성 신호로부터 여기 신호를 분리하는 방법을 더 자세하게 설명한다.

[0085] 단계(532)에서, 여기 신호 분리부(342)의 변환부(343)는 스펙트럼 관련 파라미터들 중 입력된 음성 신호의 LSF를 나타내는 파라미터를 LPC로 변환할 수 있다. 선형 예측 분석 필터는 획득된 스펙트럼 관련 파라미터들 중 음성 신호의 LSF를 나타내는 파라미터에 기반하는 것으로서, 변환된 LPC에 기반하여 생성될 수 있다.

[0086] 단계(534)에서, 여기 신호 분리부(342)는 입력된 음성 신호에 대해 상기 LPC에 기반한 선형 예측 분석 필터를 적용함으로써 음성 신호로부터 여기 신호를 분리할 수 있다.

[0087] 단계(540)에서, 모델링부(344)는 분리된 여기 신호의 확률 분포를 모델링할 수 있다. 모델링부(344)는 예컨대, WaveNet, SampleRNN, 또는 WaveRNN으로 구현될 수 있다. 또한, 모델링부(344)는 이들에 제한되지 않는 일반적인 생성 모델(generative model)로 구현될 수 있다.

[0088] 여기 신호 추정부(320)는 모델링부(344)에 의해 모델링된 여기 신호의 확률 분포를 사용하여, 전술된 단계(420)의 여기 신호의 추정을 수행할 수 있다.

[0089] 도 1 내지 도 4를 참조하여 전술된 실시예의 뉴럴 보코더는 여기 신호를 훈련하고, 여기 신호를 추정하여 합성 음성 신호를 생성한다는 점에서 ExcitNet 보코더로 명명될 수 있다.

[0090] 여기 신호의 변화는 발화자의 성대의 움직임에 의한 것으로만 제한될 수 있으므로, 여기 신호를 훈련하는 과정은 (음성 신호를 훈련하는 것에 비해) 훨씬 간단해질 수 있다. 또한, 여기 신호의 주기성의 정도를 효과적으로 나타내는 조건부 특징으로서 ITFTE 파라미터가 사용됨으로써 여기 신호의 확률 분포 모델링 정확도를 크게 향상시킬 수 있다.

[0091] 이상, 도 1 내지 도 4를 참조하여 전술된 기술적 특징에 대한 설명은, 도 5에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0093] 아래에서는, 도 6 내지 도 8을 참조하여, 타겟 화자로부터의 소량의(즉, 짧은 시간의) 음성 데이터만으로 높은 품질의 타겟 화자의 합성 음성을 생성할 수 있는 화자 적응형 모델을 구축하고, 타겟 화자의 합성 음성을 생성하는 방법에 대해 설명한다.

[0094] 도 6은 일 실시예에 따른 화자 적응형 모델을 구축하여 타겟 화자의 합성 음성을 생성하는 방법을 나타낸다.

[0095] 후술될 상세한 설명에서 음성 데이터 세트는 음성 신호 또는 음성 신호를 포함하는 데이터를 나타낼 수 있다. 예컨대, 음성 데이터 세트는 화자로부터 일정 시간 동안 녹음된 음성 신호를 나타낼 수 있다.

[0096] 소스 모델(610)은 복수의 화자들로부터의 음성 데이터 세트들에 대해 훈련된 음향 모델일 수 있다. 소스 모델(610)은 복수의 화자들에 대해 화자 독립적으로 훈련된 음향 모델일 수 있다. 예컨대, 소스 모델(610)은 10명의 화자들 각각으로부터의 1시간의 음성 데이터 세트를 사용하여, 화자 독립적으로 훈련된 음향 모델일 수 있다. 소스 모델(610)은 딥러닝 기반의 통계적 파라메트릭 음성 합성(Statistical Parametric Speech Synthesis; SPSS) 시스템을 사용하여 설계된 것일 수 있다. 음향 모델(110)은 예컨대, DNN TTS 모듈일 수 있다.

[0097] 복수의 화자들로부터의 음성 데이터 세트들을 통해 화자 독립적으로 훈련된 소스 모델(610)은 화자 적응형 모델(620)의 초기화기(initializer)로서 사용될 수 있다. 말하자면, 소스 모델(610)로부터의 가중치 값(weight)들은, 화자 적응형 모델(620)의 타겟 화자로부터의 음성 데이터 세트에 대한 훈련에 있어서 초기 값으로 설정될 수 있다. 소스 모델(610)로부터의 가중치 값들은 예컨대, 전술된 음향 파라미터에 대응할 수 있다.

[0098] 화자 적응형 모델(620)은 뉴럴 보코더를 사용하여 구현될 수 있다. 뉴럴 보코더는 딥러닝 모델을 사용하여 훈련된 것일 수 있다. 뉴럴 보코더는 예컨대, WaveNet, SampleRNN, ExcitNet 또는 WaveRNN일 수 있다. 또한, 뉴럴 보코더는 이들에 제한되지 않는 일반적인 생성 모델(generative model)일 수 있다.

[0099] 화자 적응형 모델(620)은 화자 적응(speaker adaptation) 알고리즘을 적용함으로써 특정 화자에 대해 종속적으로(speaker-dependent) 훈련될 수 있다. 예컨대, 화자 적응형 모델(620)은 특정한 타겟 화자(예컨대, 연예인, 유명인 등과 같은 셀럽)에 대해 화자 종속적으로 훈련될 수 있다. 화자 적응형 모델(620)은 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써 업데이트된 가중치 값(들)을 생성할 수 있다.

[0100] 화자 적응형 모델(620)은, 랜덤 값이 아닌, 화자 독립적으로 훈련된 소스 모델(610)로부터의 가중치 값들을 초기 값으로 사용하여 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써, 상대적으로 작은 크기(즉, 짧은 시간)의 음성 데이터 세트를 훈련하는 것만으로도 높은 품질의 타겟 화자의 합성 음성(합성 음성 신호)을 생성할 수 있다. 예컨대, 화자 적응형 모델(620)은 10분 내외의 타겟 화자의 음성 데이터 세트를 훈련하는 것만으로도 높은 품질의 타겟 화자의 합성 음성을 생성할 수 있다.

[0101] 실시예를 통해서는, 수시간 내지 수십 시간 이상의 음성 데이터 세트를 확보하기 어려운 셀럽에 대해 10분 내외의 음성 데이터 세트를 확보하여 이를 훈련 데이터로서 사용하는 것만으로도, 높은 품질의 타겟 화자의 합성 음성을 생성할 수 있는 화자 적응형 모델(620)을 구축할 수 있다.

[0102] 이상, 도 1 내지 도 5를 참조하여 전술된 기술적 특징에 대한 설명은, 도 6에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0104] 도 7은 일 실시예에 따른 뉴럴 보코더의 프로세서의 구조를 나타내는 블록도이다.

[0105] 도 7을 참조하여 설명되는 프로세서(220)는 도 3을 참조하여 전술된 프로세서(220)에 대응할 수 있다. 후술될 프로세서(220)의 구성들(710 내지 720)의 각각은 하나 이상의 소프트웨어 모듈 및/또는 하드웨어 모듈로 구현될 수 있다. 실시예에 따라 프로세서(220)의 구성요소들은 선택적으로 프로세서(220)에 포함되거나 제외될 수도 있다. 또한, 실시예에 따라 프로세서(220)의 구성요소들은 프로세서(220)의 기능의 표현을 위해 분리 또는 병합될 수도 있다. 구성들(710 내지 720)은 뉴럴 보코더 시스템(200)에 저장된 프로그램 코드가 제공하는 명령에 따라 프로세서(220)에 의해 수행되는 프로세서(220)의 서로 다른 기능들(different functions)의 표현들일 수 있다.

[0106] 프로세서(220)는 화자 적응형 모델 구축부(720)를 포함할 수 있다. 화자 적응형 모델 구축부(720)는 복수의 화자들로부터의 음성 데이터 세트들에 대해 화자 독립적으로 훈련된 소스 모델(610)로부터의 가중치 값(weight)들을 초기 값으로 설정할 수 있고, 설정된 초기 값에 대해, 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써 업데이트된 가중치 값들을 생성하는 화자 적응형 모델(620)을 구축할 수 있다. 화자 적응형 모델(620)을 통해 생성된 업데이트된 가중치 값들은 타겟 화자에 대응하는 합성 음성을 생성하기 위해 사용될 수 있다.

[0107] 프로세서(220)는 소스 모델 구축부(710)를 더 포함할 수 있다. 소스 모델 구축부(710)는 복수의 화자들로부터의 음성 데이터 세트들을 화자 독립적으로 훈련하는 소스 모델(610)을 구축할 수 있다. 구축된 소스 모델(610)은 타겟 화자로부터의 음성 데이터 세트를 훈련하기 위한 모델의 초기화기(initializer)로서 동작될 수 있다.

[0108] 소스 모델 구축부(710)는 프로세서(220)에 포함되지 않고, 뉴럴 보코더 시스템(200)과는 별개의 장치 내에서 구현될 수 있다. 화자 적응형 모델 구축부(720)는 이러한 별개의 장치 내에서 구현된 소스 모델 구축부(710)에 의해 구축된 소스 모델(610)로부터 가중치 값을 획득하여, 화자 적응형 모델(620)을 구축하기 위한 타겟 화자의 음성 데이터 세트를 훈련할 수 있다.

[0109] 이상, 도 1 내지 도 6을 참조하여 전술된 기술적 특징에 대한 설명은, 도 7에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0111] 도 8은 일 실시예에 따른 화자 적응형 모델을 구축하기 위한 뉴럴 보코더의 훈련 방법을 나타내는 흐름도이다.

[0112] 단계(810)에서, 소스 모델 구축부(710)는 복수의 화자들로부터의 음성 데이터 세트들을 화자 독립적으로 훈련하는 소스 모델(610)을 구축할 수 있다. 복수의 화자들은 소스 모델(610)을 훈련시키기 위한 음성 데이터 세트를 제공하는 임의의 사용자들일 수 있다.

[0113] 단계(820)에서, 화자 적응형 모델 구축부(720)는 소스 모델(610)로부터 가중치 값들을 획득할 수 있다. 소스 모델(610)로부터의 가중치 값들은 복수의 화자들로부터의 음성 데이터 세트들에 포함된 화자 별로 구분되지 않는 글로벌 특성을 나타내는 값을 나타낼 수 있다. 글로벌 특성은, 예컨대, 특정한 발음(일례로, '아(ah)' 또는 '이

(ee)' 등)에 대한 포먼트(formant) 특성 또는 진폭-주파수 특성(패턴)을 나타낼 수 있다. 말하자면, 소스 모델(610)은 복수의 화자들로부터의 음성 데이터 세트들을 사용하여 이와 같은 음성의 화자 독립적인 글로벌 특성을 훈련할 수 있다.

[0114] 단계(830)에서, 화자 적응형 모델 구축부(720)는 소스 모델(610)로부터 획득된 가중치 값들을 초기 값으로 설정할 수 있다. 말하자면, 소스 모델(610)은 화자 적응형 모델 구축부(720)에 의해 구축되는 화자 적응형 모델(620)의 초기화기로서 사용될 수 있다.

[0115] 단계(840)에서, 화자 적응형 모델 구축부(720)는 획득된 초기 값에 대해, 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써 업데이트된 가중치 값들을 생성할 수 있다. 말하자면, 화자 적응형 모델 구축부(720)는 소스 모델(610)로부터의 초기 값에 대해 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써 타겟 화자에 대해 적응된(즉, 타겟 화자에 대해 종속적인) 화자 적응형 모델(620)을 구축할 수 있다.

[0116] 화자 적응형 모델 구축부(720)는 소스 모델(610)로부터의 가중치 값들을 타겟 화자로부터의 음성 데이터 세트가 포함하는 타겟 화자의 고유한 특성이 반영되도록 조정함으로써, 업데이트된 가중치 값들을 생성할 수 있다. 예컨대, 화자 적응형 모델 구축부(720)는, 타겟 화자로부터의 음성 데이터 세트를 훈련하는 것을 통해, 소스 모델(610)로부터의 화자 별로 구분되지 않는 글로벌 특성을 나타내는 값을 타겟 화자의 고유한 특성을 포함하도록 미세 조정함으로써 업데이트된 가중치 값들을 생성할 수 있다.

[0117] 생성된 업데이트된 가중치 값들은 타겟 화자에 대응하는 합성 음성 신호를 생성하기 위해 사용될 수 있다. 타겟 화자에 대응하는 합성 음성 신호는 예컨대, 타겟 화자에 대응하는 셀럽의 합성 음성일 수 있다.

[0118] 소스 모델(610)을 훈련시키기 위한 복수의 화자들로부터의 음성 데이터 세트들의 각각의 크기(즉, 녹음된 음성 신호의 길이, 예컨대, 1시간 이상)는, 타겟 화자로부터의 음성 데이터 세트의 크기(즉, 녹음된 음성 신호의 길이, 예컨대, 10분)보다 더 클 수 있다.

[0119] 단계(830)에서 설명된 것과 같은 적응 프로세스의 미세 조정(fine-tuning) 메커니즘을 통해서는 타겟 화자의 음성 데이터 세트로부터 타겟 화자의 고유한 특성이 캡처될 수 있다. 따라서, 설명된 실시예의 방법을 통해서는 타겟 화자로부터의 훈련을 위한 음성 데이터 세트가 불충분하더라도 보코딩 성능을 향상시킬 수 있다.

[0120] 도 6 내지 도 8을 참조하여 전술된 뉴럴 보코더의 훈련 방법은 도 1 내지 도 4를 참조하여 전술된 실시예의 뉴럴 보코더의 훈련 방법과 합성 음성 신호를 생성 방법과 조합될 수 있다. 예컨대, 전술된 ExcitNet 보코더는 도 6 내지 도 8을 참조하여 전술된 실시예와 조합될 수 있다.

[0121] 일례로, 단계들(810 내지 840)을 수행함으로써 훈련된 뉴럴 보코더는 도 1 내지 도 4를 참조하여 전술된 뉴럴 보코더 시스템(200)에 대응할 수 있다. 단계(430)에서 생성된 타겟 음성 신호는 화자 적응형 모델(620)이 훈련한 타겟 화자에 대응하는 합성 음성 신호일 수 있다.

[0122] 도 6 내지 도 8을 참조하여 전술된 뉴럴 보코더의 훈련 방법과 도 1 내지 도 4를 참조하여 전술된 ExcitNet 모델의 기술적 특징을 조합함으로써, 타겟 화자에 대응하는 합성 음성의 품질을 높일 수 있다.

[0123] 이상, 도 1 내지 도 7을 참조하여 전술된 기술적 특징에 대한 설명은, 도 8에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0125] 도 9는 일 예에 따른 음성 신호 및 여기 신호와 그 관계를 나타낸다.

[0126] 도시된 것처럼, 음성 신호를 S(n)으로 가정하고, S(n)이 포함하는 여기 신호를 e(n)이라고 가정하면, S(n)과 e(n)의 관계는 아래 수학식 1과 같이 표현될 수 있다.

수학식 1

$$S(n) = e(n) * h(n)$$

[0127]

[0128] h(n)은 선형 예측 합성 필터를 나타낼 수 있다. h(n)은 S(n)의 e(n) 성분을 제외한 나머지 성분(즉, 스펙트럼 성분)을 나타낼 수 있다. h(n)은 S(n)의 LSF를 나타내는 파라미터에 기반하여 생성될 수 있다.

[0129] 수학식 1의 관계에 따라 도 4를 참조하여 전술된 단계(420)에 의해 추정된 여기 신호(즉, e(n))에 대해 선형 예측 합성 필터(즉, h(n))를 적용함으로써 타겟 음성 신호(S(n))가 생성될 수 있다. 선형 예측 합성 필터의 구체적인 예시에 대해서는 도 14를 참조하여 더 자세하게 설명된다.

[0130] 수학식 1의 관계는 도 5를 참조하여 전술된 단계(530)의 여기 신호(즉, e(n))의 분리에 대해서도 유사하게 적용될 수 있다. 말하자면, 훈련을 위해 입력된 음성 신호(S(n))에 대해 선형 예측 분석 필터가 적용됨으로써 음성 신호(S(n))로부터 여기 신호(e(n))가 분리될 수 있다. 선형 예측 분석 필터의 구체적인 예시에 대해서는 도 13을 참조하여 더 자세하게 설명된다.

[0131] 이상, 도 1 내지 도 8을 참조하여 전술된 기술적 특징에 대한 설명은, 도 9에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0133] 도 10a 내지 10c는 각각 다른 종류의 보코더를 사용하는 합성 음성 신호 생성을 위한 통계적 파라메트릭 음성 합성(Statistical Parametric Speech Synthesis; SPSS) 시스템을 나타낸다.

[0134] 도 10a는 음향 모델(1010)과, 음향 모델(1010)로부터의 음향 피처(음향 파라미터)를 LPC(Linear Predictive Coding) 합성함으로써 음성 신호를 생성하는 LPC 합성 모듈(1020)을 포함하는 음성 합성을 위한 프레임워크를 나타낸다. LPC 합성 모듈(1020)은 LPC 보코더로서, 예컨대, 전술된 선형 예측 합성 필터에 대응할 수 있다.

[0135] 도 10b는 음향 모델(1010)과 음향 모델(1010)로부터의 음향 피처(음향 파라미터)에 기반하여 음성 신호를 추정하는 뉴럴 보코더로서 WaveNet 보코더(1022)를 포함하는 음성 합성을 위한 프레임워크를 나타낸다.

[0136] 도 10c는, 도 1 내지 5를 참조하여 전술된 것과 같은, ExcitNet 보코더(1024)를 사용하는 음성 합성을 위한 프레임워크를 나타낼 수 있다. 도 10c에서 도시된 구조는 도 10a의 LPC 코더(1020) 및 도 10b의 WaveNet 보코더(1022)가 조합된 것일 수 있다.

[0137] 도 10c의 구조에서, ExcitNet 보코더(1022)는 음향 모델(1010)로부터의 음향 피처(음향 파라미터)에 기반하여 여기 신호를 추정할 수 있다. 추정된 여기 신호는 선형 예측 합성 필터(1030)를 통한 LPC(Linear Predictive Coding) 합성에 따라 타겟 음성 신호로서 변환될 수 있다.

[0138] 도 10c의 구조의 보다 상세한 예시는 도 12 및 도 14를 참조하여 더 자세하게 설명된다.

[0139] 이상, 도 1 내지 도 9를 참조하여 전술된 기술적 특징에 대한 설명은, 도 10a 내지 도 10c에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0141] 도 11 및 도 13은 일 실시예에 따른 훈련을 위해 입력된 음성 신호로부터 여기 신호를 분리함으로서 뉴럴 보코더를 훈련시키는 방법을 나타낸다.

[0142] 도 11에서 도시된 것처럼, 훈련을 위해 입력된 음성 신호에 대해, 파라메트릭 보코더(1110)는 음향 파라미터들을 추출할 수 있다. 입력된 음성 신호에 대해서는 추출된 음향 파라미터들 중 스펙트럼 관련 파라미터에 기반하여 생성된 선형 예측 분석 필터(1140)가 적용됨으로써, 입력된 음성 신호로부터 여기 신호가 분리될 수 있다.

[0143] WaveNet 보코더(1130)는 추출된 음향 파라미터들을 보조 피처(auxiliary feature)들로서 구성하여(1120) 수신할 수 있다. 보조 피처는 전술된 스펙트럼 관련 파라미터들 및 여기 관련 파라미터들을 포함할 수 있다. WaveNet 보코더(1130)는 보조 피처 및 분리된 여기 신호에 기반하여 여기 신호의 확률 분포를 모델링할 수 있다. WaveNet 보코더(1130)는 ExcitNet 보코더 또는 기타 일반적인 생성 모델(generative model)의 뉴럴 보코더로도 구현될 수 있다.

[0144] 도 13을 참조하여 도 11의 구조를 보다 상세하게 설명한다. 훈련을 위해 입력된 음성 신호는 음성 분석(1310)에 따라 음향 피처들(음향 파라미터들)이 추출될 수 있다. 음향 파라미터들 중 LSF를 나타내는 파라미터는 LPC로 변환될 수 있다(1320). 변환된 LPC에 기반하여 선형 예측 분석 필터(1340)가 구현될 수 있다. 입력된 음성 신호에 대해 선형 예측 분석 필터(1340)가 적용됨으로써 입력된 음성 신호로부터 여기 신호가 분리될 수 있다. 분리된 여기 신호는 ExcitNet 모델(즉, ExcitNet 보코더)(1350)로 입력될 수 있다. 한편, 음향 파라미터들을 보조 피처(auxiliary feature)들로서 구성될 수 있고(1330), 보조 피처들은 ExcitNet 모델((1350)로 입력될 수 있다. ExcitNet 모델(1350)은 입력된 보조 피처들(즉, 음향 파라미터들)과 분리된 여기 신호에 기반하여 여기

신호의 확률 분포를 모델링할 수 있다. 도시된 예시에서, e_n은 분리된 여기 신호에 대응할 수 있다.

[0145] 도 12 및 도 14는 일 실시예에 따른 입력 텍스트에 기반하여 음향 모델에 의해 생성된 음향 파라미터들로부터 여기 신호를 추정하여 합성 음성 신호를 생성하는 방법을 나타낸다.

[0146] 도 12에서 도시된 것처럼, 음향 모델(1150)은 수신된 언어 파라미터들에 기반하여 음향 파라미터들을 생성할 수 있다. WaveNet 보코더(1170)는 음향 파라미터들을 보조 피처들로서 구성하여(1160) 수신할 수 있다. 보조 피처는 전술된 스펙트럼 관련 파라미터들 및 여기 관련 파라미터들을 포함할 수 있다. WaveNet 보코더(1170)는 음향 파라미터들에 기반하여 여기 신호를 추정할 수 있다. WaveNet 보코더(1170)는 ExcitNet 보코더 또는 기타 일반적인 생성 모델(generative model)의 뉴럴 보코더로도 구현될 수 있다. 추정된 여기 신호에 대해서는 추출된 음향 파라미터들 중 스펙트럼 관련 파라미터에 기반하여 생성된 선형 예측 합성 필터(1180)가 적용됨으로써, 타겟 합성 음성이 생성될 수 있다.

[0147] 도 14을 참조하여 도 12의 구조를 보다 상세하게 설명한다. 합성 음성 신호의 생성을 위해 입력된 텍스트에 대해 텍스트 분석을 수행함으로써(1410) (전술된 언어 파라미터들에 대응하는) 언어 피처들이 추출될 수 있다. 언어 피처를 추출함에 있어서는 도시된 것처럼, 음소 듀레이션(phoneme duration)을 추정하는 듀레이션 모델(1420)이 더 사용될 수 있다. 음향 모델(1430)은 추출된 언어 피처들로부터 음향 피처들(음향 파라미터들)을 생성할 수 있다. 음향 파라미터들 중 LSF를 나타내는 파라미터는 LPC로 변환될 수 있다(1440). 변환된 LPC에 기반하여 선형 예측 합성 필터(1470)가 구현될 수 있다. 음향 파라미터들을 보조 피처(auxiliary feature)들로서 구성될 수 있고(1450), 보조 피처들은 ExcitNet 모델(즉, ExcitNet 보코더)(1460)로 입력될 수 있다. ExcitNet 모델(1460)은 입력된 보조 피처들(즉, 음향 파라미터들)에 기반하여 여기 신호를 추정할 수 있다. 추정된 여기 신호에 대해 변환된 LPC에 기반한 선형 예측 합성 필터(1470)가 적용됨으로써 타겟 음성 신호가 생성될 수 있다. 도시된 예시에서, $\tilde{x}_n$ 은 생성된 타겟 음성 신호에 대응할 수 있고, $\tilde{e}_n$은 추정된 여기 신호에 대응할 수 있다.

[0148] 이상, 도 1 내지 도 10c를 참조하여 전술된 기술적 특징에 대한 설명은, 도 11 내지 도 14에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0150] 도 15는 일 예에 따른, 훈련 과정/합성 음성 신호의 생성 과정에서 획득된 음의 로그우도(Negative Log-Likelihood; NLL)의 음향 파라미터로서 여기의 주기성에 따라 구분되는 파라미터들의 사용 여부에 따른 차이를 나타내는 그래프이다.

[0151] 훈련(training) 과정에 있어서 NLL이 낮을수록 모델링의 정확도가 높은 것으로 볼 수 있다. 도시된 그래프에서는 전술된 SEW 파라미터 및 REW 파라미터와 같은 ITFTE 파라미터를 사용한 경우에 있어서의 NLL이 그렇지 않은 경우보다 낮게 됨을 확인할 수 있다.

[0152] 또한, 검증(validation) 과정에 있어서도 NLL이 낮을수록 생성되는 합성 음성의 품질이 우수한 것으로 볼 수 있다. 도시된 그래프에서는 SEW 파라미터 및 REW 파라미터와 같은 ITFTE 파라미터를 사용한 경우에 있어서의 NLL이 그렇지 않은 경우보다 낮게 됨을 확인할 수 있다.

[0153] 말하자면, 도시된 그래프를 통해서는 뉴럴 보코더의 훈련에 있어서 ITFTE 파라미터를 사용하는 것을 통해 여기 신호의 확률 분포의 모델링의 오류를 크게 감소시킬 수 있으며, 합성 음성의 생성을 위한 여기 신호의 추정에 있어서 ITFTE 파라미터를 사용함으로써 합성 음성 신호의 생성에 있어서 오류를 크게 감소시킬 수 있다는 사실을 확인할 수 있다.

[0154] 이상, 도 1 내지 도 14를 참조하여 전술된 기술적 특징에 대한 설명은, 도 15에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0156] 도 16는 일 예에 따른 복수의 화자들로부터의 음성 신호에 대해, 음성 신호의 화자 종속적인 특징과 화자 독립적인 특징을 도식적으로 나타낸다. 도 17은 일 예에 따른 복수의 화자들로부터의 음성 데이터 세트들을 훈련시킴으로써 구축된 소스 모델과, 타겟 화자로부터의 음성 데이터 세트를 훈련시킴으로써 구축된 화자 적응형 모델을 사용하여 타겟 화자의 합성 음성을 생성하는 방법을 나타낸다.

[0157] 도 16에서 도시된 것처럼, 화자 독립적인 특징은 화자들(화자 1 내지 3)의 음성에 있어서 공통되는 특징일 수 있다. 말하자면, 화자 독립적인 특징은 화자 별로 구분되지 않는 글로벌 특성을 나타낼 수 있다. 화자 종속적인 특징은 화자 별 고유한 특성을 나타낼 수 있다.

[0158] 도 17에서 도시된 것처럼, 복수의 화자들로부터의 음성 데이터 세트들을 화자 독립적으로 훈련함으로써 소스 모델(610)이 구축될 수 있고, 이러한 소스 모델(610)로부터의 가중치 값에 기초하여 타겟 화자로부터의 음성 데이터 세트를 훈련함으로써 타겟 화자에 종속적인 화자 적응형 모델(620)이 구축될 수 있다. 소스 모델(610)로부터의 가중치 값은 화자 적응형 모델(620)에서 타겟 화자로부터의 음성 데이터 세트가 훈련됨에 따라 타겟 화자의 고유한 특성을 반영하도록 미세 조정될 수 있다. 도시된 것처럼, 소스 모델(610) 및 화자 적응형 모델(620)은 ExcitNet 모델을 사용하여 구현될 수 있다. 도시된 것처럼, 실시예를 통해서는 뉴럴 보코더에 대해 화자 적응(speaker adaptation) 알고리즘의 적용할 수 있다. 도시되지는 않았으나, 소스 모델(610)에 대응하는 음향 모델(예컨대, DNN TTS)에 대해서도 유사하게 화자 적응 알고리즘이 적용될 수 있다.

[0159] 이상, 도 1 내지 도 15를 참조하여 전술된 기술적 특징에 대한 설명은, 도 16 및 도 17에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0161] 도 18 및 도 19는 일 예에 따른 화자 적응(speaker adaptation) 알고리즘의 적용 여부에 따라 생성된 합성 음성 신호의 품질을 비교 평가한 결과를 나타낸다.

[0162] 도 18 및 도 19의 Score는 평가자들이 음성 신호를 청취하여 평가한 스코어의 평균을 나타낸다. RAW는 원본 음성 신호에 해당할 수 있다.

[0163] 도 18을 참조하면, WaveNet 모델 및 ExcitNet 모델을 모두에서 화자 적응 알고리즘을 적용한 경우의 합성 음성 신호에 대한 품질이 높게 평가되었음을 확인할 수 있다. 말하자면, 도 6 내지 도 8을 참조하여 전술한 바와 같이 화자 적응형 모델(620)을 구축하여 합성 음성 신호를 생성하는 경우(with SA)가 그렇지 않은 경우에 비해 우수한 성능을 나타냄을 확인할 수 있다.

[0164] 도 19는 보다 상세한 합성 음성 신호의 품질을 비교 평가한 결과를 나타낸다. 도 19에 대해서는 아래에서 보다 자세하게 설명한다.

[0165] 이상, 도 1 내지 도 17을 참조하여 전술된 기술적 특징에 대한 설명은, 도 18 및 도 19에 대해서도 그대로 적용될 수 있으므로 중복되는 설명은 생략한다.

[0167] 아래에서는, 도 1 내지 도 5를 참조하여 전술된 ExcitNet 모델에 대해 보다 상세하게 설명하고, 타 모델과의 비교 시험 결과에 대해 더 설명한다.

[0168] ExcitNet 모델(ExcitNet 보코더)은 통계적 파라메트릭 음성 합성(SPSS) 시스템을 위한 WaveNet 기반의 뉴럴 여기 모델일 수 있다. WaveNet 기반의 뉴럴 보코딩 시스템은 합성 음성 신호의 인식 품질을 크게 향상시키지만, 종종 음성 신호의 복잡한 시변 특성을 포착하지 못하는 경우가 있으므로 노이즈를 출력하는 경우가 있다. ExcitNet 기반의 뉴럴 보코딩 시스템은 음성 신호로부터 스펙트럼 성분을 분리하는 적응적 인버스 필터를 사용하여 (예컨대, WaveNet 프레임워크 내에서) 잔차 성분(즉, 여기 신호)을 분리하여 훈련할 수 있고, 합성 음성 신호를 생성함에 있어서 여기 신호를 타겟으로서 추정할 수 있다. 이러한 방식을 통해서는, 딥러닝 프레임워크에 의해 음성 신호의 스펙트럼 성분이 더 잘 표현될 수 있고 잔차 성분은 WaveNet 프레임워크에 의해 효율적으로 생성되므로, 합성된 음성 신호의 품질이 향상될 수 있다.

[0169] 아래의 실험 결과에서도 (화자 종속적 및 화자 독립적으로 훈련된) ExcitNet 기반의 뉴럴 보코딩 시스템이 기존의 선형 예측 보코더 및 WaveNet 보코더보다 뛰어난 성능을 보이는 것으로 나타났다.

[0170] 시험을 위해서는, 음향 모델과 화자 종속적(SD) ExcitNet 보코더를 훈련시키기 위해 음성적으로 운율적으로 풍부한 두 개의 스피치 코퍼가 사용되었다. 각 코퍼스는 전문 한국어 여성(KRF)과 한국인 남성(KRM)이 녹음한 것이다. 음성 신호는 24 kHz에서 샘플링되었고, 각 샘플은 16 비트로 양자화되었다. 아래 표 1은 각 집합의 발화 수를 나타낸다. 화자 독립적(SI) ExcitNet 보코더를 훈련시키기 위해서는 한국인 여성 5 명과 한국인 남성 5 명이 녹음한 음성 코퍼가 사용되었다. 총 6,422 개(10 시간) 및 1,080개(1.7 시간)의 발화가 각각 훈련 및 검증(validation)에 사용되었다. SI 데이터 세트에 포함되지 않은 동일한 KRF 및 KRM 화자에 의해 녹음된 음성 샘플이 시험을 위해 사용되었다.

표 1

[0171]

화자	훈련	검증	시험
KRF	3,826(7h)	270(30min)	270(30min)
KRM	2,294(7h)	160(30min)	160(30min)

[0172] 아래 표 2 및 표 3은 객관적인 시험의 결과로서 원본 음성과 생성된 음성 간의 왜곡을 LSD(Log-Spectral Distance)(dB)와 F0 RMSE(Root Mean Square Error)(Hz)로 각각 나타낸 것이다. WN은 WaveNet 보코더를 나타내고, WN-NS는 WaveNet 보코더에 노이스-쉐이핑 방법을 적용한 것을 나타내며, ExcitNet은 ExcitNet 보코더를 나타낸다. 가장 작은 오류가 나타난 부분은 볼드체로 표시하였다. 표 2 및 표 3은 유성음에 대해 측정된 결과일 수 있다.

표 2

[0173]

화자	시스템	SD				SI
		1h	3h	5h	7h	
KRF	WN	4.21	4.19	4.18	4.13	4.18
	WN-NS	4.15	4.12	4.07	**4.01**	4.06
	ExcitNet	**4.11**	**4.09**	**4.05**	**3.99**	**4.04**
KRM	WN	3.73	3.72	3.69	3.67	3.70
	WN-NS	3.54	**3.46**	**3.41**	3.41	3.46
	ExcitNet	**3.53**	**3.46**	**3.41**	**3.40**	**3.45**

표 3

[0174]

화자	시스템	SD				SI
		1h	3h	5h	7h	
KRF	WN	32.61	31.69	31.56	31.49	32.30
	WN-NS	31.96	31.75	31.52	31.38	32.23
	ExcitNet	**31.44**	**31.43**	**31.37**	**31.29**	**31.88**
KRM	WN	12.60	12.39	12.05	12.05	13.96
	WN-NS	**12.32**	12.16	11.97	11.97	13.34
	ExcitNet	12.72	**12.10**	**11.93**	**11.93**	**12.96**

[0175] 표 2 및 3에서 표시된 것처럼, SD 및 SI의 대부분의 경우에 있어서 ExcitNet 보코더의 경우가 원본 음성과 생성된 음성 간의 왜곡이 가장 적게 나타남을 확인할 수 있다. 아래 표 4는 무성음 및 트랜지션 영역(transition regions)에 대해 측정된 LSD(dB)를 나타낸다.

표 4

[0176]

화자	시스템	SD				SI
		1h	3h	5h	7h	
KRF	WN	4.19	4.16	4.15	4.09	4.18
	WN-NS	4.26	4.18	4.12	4.03	4.06
	ExcitNet	**4.15**	**4.10**	**4.06**	**3.98**	**4.04**
KRM	WN	3.95	3.96	3.92	3.92	3.70
	WN-NS	4.41	3.95	3.88	3.88	3.46
	ExcitNet	**3.91**	**3.83**	**3.76**	**3.76**	**3.45**

[0177] 표 4에서 표시된 것처럼, SD 및 SI의 모든 경우에 있어서 ExcitNet 보코더의 경우가 원본 음성과 생성된 음성 간의 왜곡이 가장 적게 나타남을 확인할 수 있다. 아래 표 5 및 6은 주관적인 시험의 결과로서 선호도 테스트의 결과(%)를 나타낸다. 청취자로부터 높은 선호도가 나타난 부분을 표에서 볼드체로 표시하였다. 나머지들에 비해 ExcitNet 보코더의 경우 합성 음성의 인식 품질이 현저하게 우수함을 확인할 수 있다. 평가자는 12인의 한국어를 모국어로 사용하는 청취자들이며, 20개의 랜덤으로 선택된 발화에 대해 시험이 이루어졌다.

표 5

[0178]

KRF	WN	WN-NS	ExcitNet	Neutral	p-value
SD	6.8	**64.1**	-	29.1	$<10^{-30}$
	7.3	-	**83.6**	9.1	$<10^{-49}$
	-	12.7	**58.2**	29.1	$<10^{-17}$
SI	12.7	66.8	-	20.5	$<10^{-22}$
	8.6	-	**73.6**	27.7	$<10^{-35}$
	-	19.5	**38.6**	41.8	$<10^{-3}$

표 6

[0179]

KRF	WN	WN-NS	ExcitNet	Neutral	p-value
SD	11.8	**60.5**	-	27.7	$<10^{-19}$
	17.3	-	**77.7**	5.0	$<10^{-24}$
	-	16.4	**73.6**	10.0	$<10^{-22}$
SI	27.3	**48.6**	-	24.1	$<10^{-3}$
	13.6	-	**75.5**	10.9	$<10^{-27}$
	-	17.3	**63.6**	19.1	$<10^{-15}$

[0180] 도 20은 일 예에 따른 ExcitNet 보코더와 타 보코더 간의 MOS(Mean Opinion Score)(MOS) 평가 결과를 나타낸다. 녹음된 음성으로부터 음향 피처가 추출되는 경우인 분석/합성(A/S)의 결과에 대한 평가 및 음향 모델로부터 음향 피처가 생성되는 경우인 SPSS에 있어서의 결과의 평가가 이루어졌다.

[0181] S/A에 있어서, SI-ExcitNet 보코더는 ITFTE 보코더와 유사한 성능을 나타냈으나 WORLD 시스템보다는 훨씬 뛰어난 것으로 나타났다. 모든 경우에 있어서 SD-ExcitNet 보코더는 최고의 인식 품질(KRF 및 KRM 화자에 대해 각각 4.35 및 4.47 MOS를 나타냄)을 나타냈다. 고음의 여성 음성을 표현하는 것이 어렵기 때문에 KRF 화자에 대한 MOS 결과는 SI 보코더들(WORLD, ITFTE 및 SI-ExcitNet)에 있어서 KRM 화자의 경우보다 좋지 않게 나타났다. 반면, SD-ExcitNet의 KRF 화자에 대한 결과는 KRM 화자에 대한 결과와 비슷하다는 점에서, 고음의 목소리를 효과적으로 표현하기 위해서는 화자 별 특성을 모델링되어야 함을 나타낸다. SPSS의 측면에서는, SD 및 SI-ExcitNet 보코더 모두 파라메트릭 ITFTE 보코더보다 훨씬 우수한 인식 품질을 나타냈다. 음향 모델이 지나치게 평탄한 음성 매개 변수를 생성했지만, ExcitNet 보코더는 시간 영역 여기 신호를 직접 추정함으로써 평활화 효과를 완화할 수 있었다. 결과적으로, SD-ExcitNet 보코더를 사용하는 SPSS 시스템은 각각 KRF 및 KRM 화자에 대해 3.78 및 3.85 MOS를 달성했다. SI-ExcitNet 보코더는 KRF 및 KRM 화자에 대해 각각 2.91 및 2.89 MOS를 달성했다.

[0183] 아래에서는, 도 6 내지 도 8을 참조하여 전술된 화자 적응형 모델(620)을 구축하는 뉴럴 보코더에 대해 보다 상세하게 설명하고, 타 모델과의 비교 시험 결과에 대해 더 설명한다. 실시예의 뉴럴 보코더는 단 10 분의 음성 데이터 세트와 같이, 타겟 화자로부터의 훈련 데이터가 불충분한 경우에도 고품질의 음성 합성 시스템을 구축할 수 있다.

[0184] 실시예의 뉴럴 보코더는 타겟 화자에 대한 제한된 훈련 데이터로 인해 발생하는 타겟 화자 관련 정보의 부족 문제를 해결하기 위해, 복수의 화자들에 대해 보편적인 특성을 추출할 수 있는 화자 독립적으로 훈련된 소스 모델(610)로부터의 가중치 값을 활용한다. 이러한 소스 모델(610)로부터의 가중치 값은 화자 적응형 모델(620)의 훈련을 초기화하기 위해 사용되며, 타겟 화자의 고유한 특성을 나타내기 위해 미세 조정될 수 있다. 이러한 적응 과정에 따라, 딥 뉴럴 네트워크가 타겟 화자의 특성을 포착할 수 있으므로, 화자 독립적인 모델에서 발생하는 불연속성 문제를 감소 시킬 수 있다. 후술될 실험 결과 역시 실시예의 뉴럴 보코더가 종래의 방법에 비해 합성된 음성의 인식 품질을 현저히 향상시킨다는 것을 보여준다.

[0185] SD는 (소스 모델(610)로부터의 가중치를 초기 값으로 하지 않고) 화자 종속적으로 훈련된 모델을 나타내고, SI는 화자 독립적으로 훈련된 모델을 나타내고, SA는 도 6 내지 도 8을 참조하여 전술한 바와 같은 화자 적응형으로 훈련된 모델(즉, 소스 모델(610)로부터의 가중치를 초기 값으로 하여 화자 종속적으로 훈련된 모델)을 나타낸다.

[0186] SD 및 SA 모델에 있어서, 한국 여성 화자가 녹음한 음성 코퍼가 사용되었다. 음성 신호는 24 kHz에서 샘플링되었고, 각 샘플은 16 비트로 양자화되었다. 훈련, 검증 및 시험에는 총 90개(10 분), 40개 (5 분), 130개 (15 분)의 발화가 사용되었다. SI 모델을 훈련시키기 위해 SD와 SA 모델 훈련에는 포함되지 않은 5 명의 한국 남성 화자 및 5 명의 한국 여성 화자가 녹음한 음성 데이터가 사용되었다. 이를 위해 훈련 및 검증에 각각 6,422개(10 시간) 및 1,080개(1.7 시간)의 발화가 사용되었다. SD 및 SA 모델의 테스트 세트는 SI 모델을 평가하기 위해서도 또한 사용되었다.

[0187] 아래 표 7 및 표 8은 객관적인 시험의 결과로서 원본 음성과 생성된 음성 간의 왜곡을 LSD(Log-Spectral Distance)(dB)와 F0 RMSE(Root Mean Square Error)(Hz)로 각각 나타낸 것이다. 표 7은 녹음된 음성으로부터 추출된 음향 피처들이 보조 피처를 구성하기 위해 직접적으로 사용되는 경우의 분석/합성의 결과에 대한 평가(A/S)를 나타낸다. 표 8은 SPSS에 있어서의 결과의 평가를 나타낸다. 가장 작은 오류가 나타난 부분은 볼드체로 표시하였다.

표 7

[0188]

시스템	WaveNet		ExcitNet	
	LSD	F0 RMSE	LSD	F0 RMSE
SD	3.65	38.27	2.10	19.83
SI	2.00	10.64	1.32	10.48
SA	**1.79**	**10.70**	**1.12**	**9.66**

표 8

[0189]

시스템	WaveNet		ExcitNet	
	LSD	F0 RMSE	LSD	F0 RMSE
SD	4.78	48.75	4.50	39.14
SI	4.51	35.53	4.42	36.28
SA	**4.45**	**35.45**	**4.36**	**35.47**

[0190] 표 7 및 표 8에서, WaveNet 보코더 및 ExcitNet 보코더 모두에 있어서 SA의 경우가 원본 음성과 생성된 음성 간의 왜곡이 가장 적게 나타남을 확인할 수 있다.도 21은 일 예에 따른 F0 스케일링 팩터(scaling factor)를 상이하게 하는 경우에 있어서 화자 적응형 모델을 구축하는 뉴럴 보코더의 성능 변화를 나타낸다.

[0191] 실시예의 SA를 적용한 훈련 방법의 유효성을 검증하기 위해 F0을 수동으로 변경했을 때의 뉴럴 보코더의 성능 변화를 조사하였다. SI 모델은 피치를 수정한 합성 음성을 생성함에 있어서 효과적이라는 것이 밝혀져 있다. SA 모델 역시 SI 모델을 활용하는 것인 바 SD 접근법에 비해서는 높은 성능을 보일 것으로 기대할 수 있다.

[0192] 시험에 있어서, F0 궤적은 SPSS 프레임 워크에 의해 생성된 후 보조 피처 벡터를 수정하기 위해 스케일링 팩터(0:6, 0:8, 1:0 및 1:2)가 곱해졌다. 음성 신호는 뉴럴 보코드 시스템을 사용하여 합성되었다.

[0193] 도 21은 상이한 F0 스케일링 팩터에 대한 F0 RMSE (Hz) 시험 결과를 나타낸다. 도 21을 통해, SA 모델이 기존의 SD 모델에 비해 훨씬 작은 수정 오류(modification error)를 포함하고 있음을 확인할 수 있다. SI 모델과 비교해서는 SA-ExcitNet 보코더는 모든 가중치가 타겟 화자의 특성에 맞게 최적화되었음에도 동등한 품질이 유지되고 있음을 확인할 수 있다.

[0194] 또한, ExcitNet 보코더는 WaveNet 보코더보다 훨씬 우수한 성능을 나타냄을 확인할 수 있다. ExcitNet 보코더는 성대 움직임의 변화(여기 신호의 변화)를 훈련하므로, WaveNet 기반의 접근 방식보다 유연하게 F0 수정된 음성 세그먼트를 재구성 할 수 있는 것으로 볼 수 있다.

[0195] 도 19는 주관적인 시험 결과로서, SD, SI 및 SA의 보코더들 간의 MOS 평가 결과를 나타낸다. 녹음된 음성으로부터 음향 피처가 추출되는 경우인 분석/합성(A/S)의 결과에 대한 평가 및 음향 모델로부터 음향 피처가 생성되는

경우인 SPSS에 있어서의 결과의 평가가 이루어졌다.

[0196] A/S 의 결과에 있어서, SD-WaveNet 보코더는 제한적인 훈련 데이터로는 타겟 화자의 특성을 훈련할 수 없었는 바 가장 나쁜 결과가 나타났다. SI-WaveNet 보코더는 ITFTE 보코더와 유사한 성능을 나타냈으며 WORLD 시스템보다는 뛰어난 성능을 보이는 것으로 나타났다. 모든 WaveNet 보코더에 있어서의 SA의 활용은 우수한 성능을 나타낸다는 것이 확인되었다. ExcitNet 보코더에 대한 결과는 WaveNet 보코더의 경우와 비슷한 경향을 보였으나 ExcitNet 보코더는 LP 인버스 필터를 통해 음성 신호의 포먼트 구성 요소를 분리함으로써 나머지 신호의 모델링 정확도를 향상시키는 바 전체적으로 훨씬 우수한 성능을 보이는 것으로 나타났다. 결과적으로, SA-ExcitNet 보코더는 A/S 결과에 있어서 4.40 MOS를 달성했다.

[0197] SPSS의 결과에 있어서, SI-WaveNet 보코더와 SI-ExcitNet 보코더는 모두 파라메트릭 ITFTE 보코더보다 우수한 인식 품질을 제공하는 것으로 나타났다. 결과적으로, 실시예의 SA 훈련 모델은 기존의 화자 의존적인 방법과 화자 독립적인 방법에 비해 합성 음성 신호의 품질을 크게 향상 시킨다는 것을 확인할 수 있었다. A/S 결과에서와 마찬가지로 ExcitNet 보코더는 SPSS 결과에 있어서 WaveNet 보코더보다 우수한 성능을 나타냈다. 음향 모델이 지나치게 평탄한 음성 매개 변수를 생성했지만, ExcitNet 보코더는 시간 영역 여기 신호를 직접 추정함으로써 평활화 효과를 완화할 수 있었다. 결과적으로 SA-ExcitNet 보코더가 있는 SPSS 시스템은 3.77 MOS를 달성했다.

[0199] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0200] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0201] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 이때, 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수 개의 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 어플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

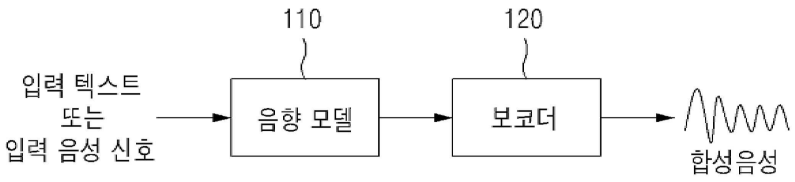
[0202] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0203] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한

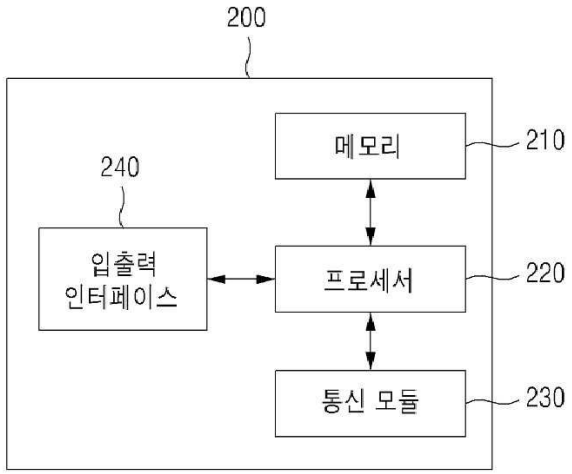
다.

도면

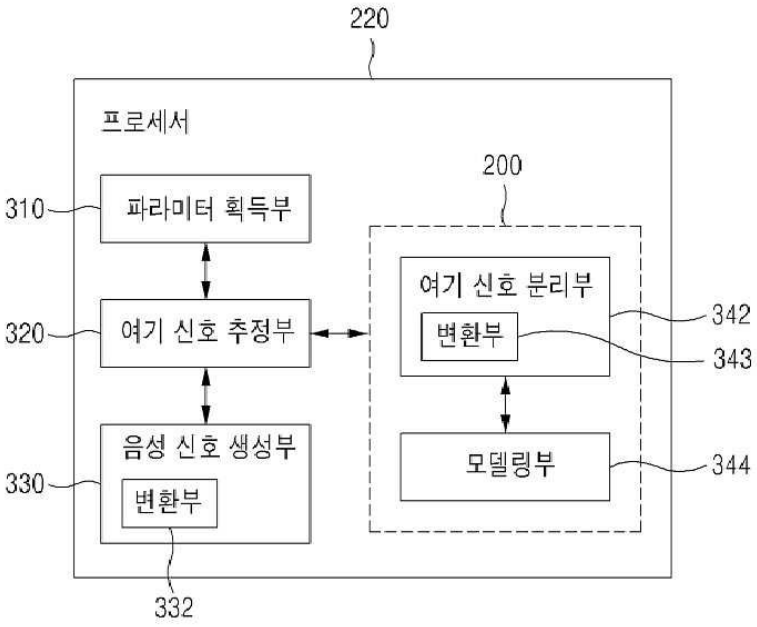
도면1



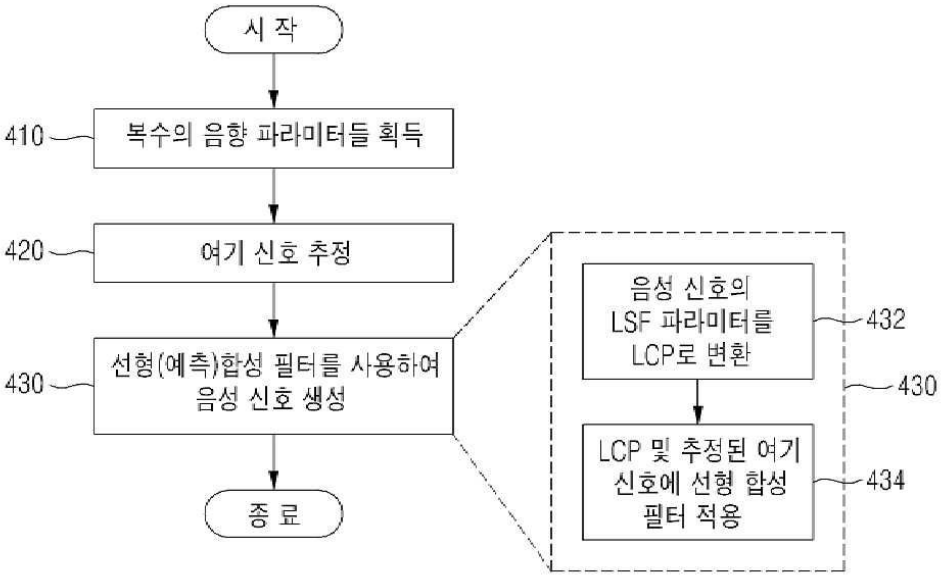
도면2



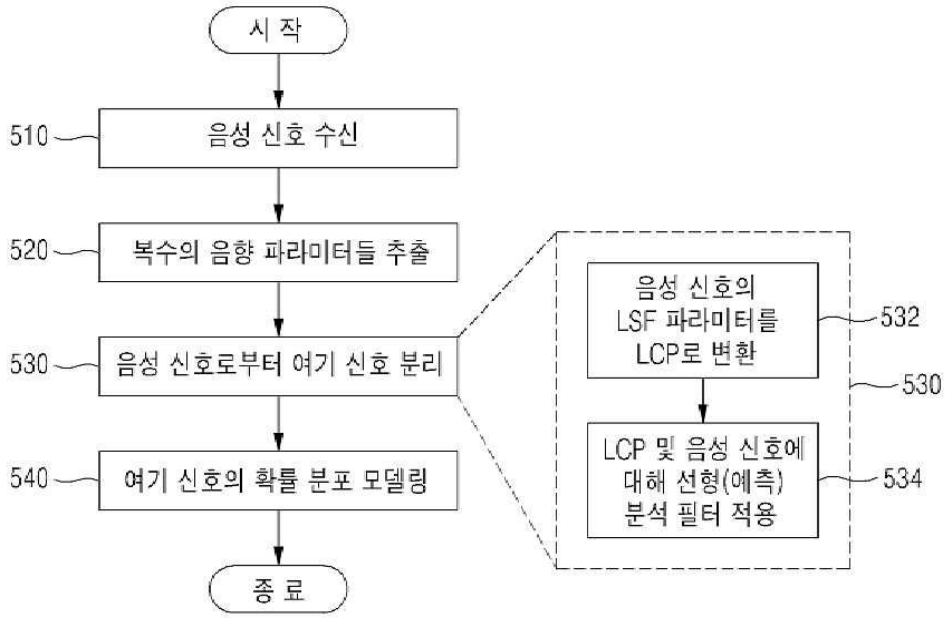
도면3



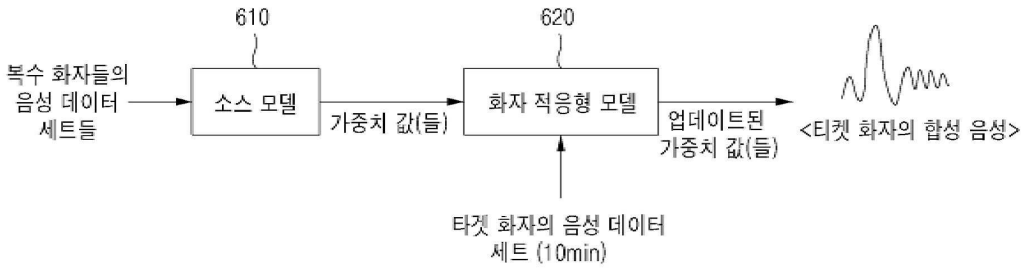
도면4


시 작
410
복수의 음향 파라미터들 획득
420
여기 신호 추정
430
선형(예측)합성 필터를 사용하여 음성 신호 생성
종 료
음성 신호의 LSF 파라미터를 LCP로 변환
432
430
LCP 및 추정된 여기 신호에 선형 합성 필터 적용
434

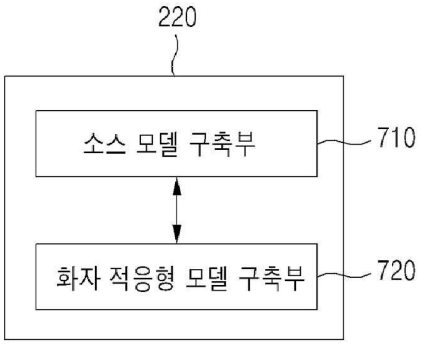
도면5


시 작
510
음성 신호 수신
520
복수의 음향 파라미터들 추출
530
음성 신호로부터 여기 신호 분리
540
여기 신호의 확률 분포 모델링
종 료
음성 신호의 LSF 파라미터를 LCP로 변환
532
530
LCP 및 음성 신호에 대해 선형(예측) 분석 필터 적용
534

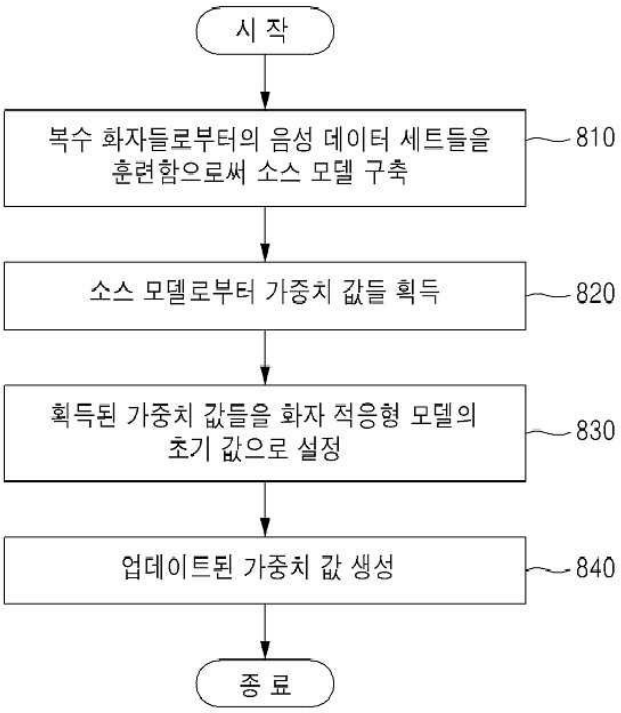
도면6


610
620
복수 화자들의 음성 데이터 세트들
소스 모델
가중치 값(들)
화자 적응형 모델
업데이트된 가중치 값(들)
<티켓 화자의 합성 음성>
타겟 화자의 음성 데이터 세트 (10min)

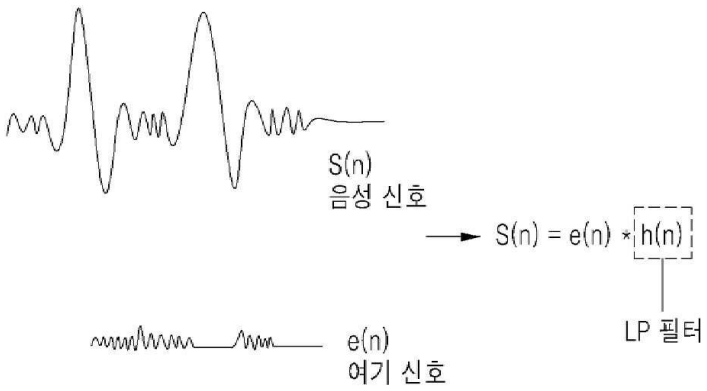
도면7


220
소스 모델 구축부
710
화자 적응형 모델 구축부
720

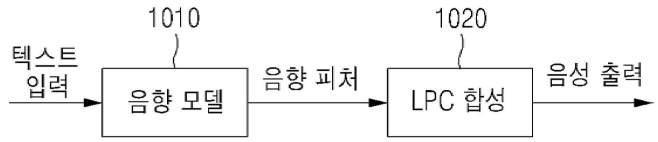
도면8


시 작
복수 화자들로부터의 음성 데이터 세트들을 훈련함으로써 소스 모델 구축
810
소스 모델로부터 가중치 값들 획득
820
획득된 가중치 값들을 화자 적응형 모델의 초기 값으로 설정
830
업데이트된 가중치 값 생성
840
종 료

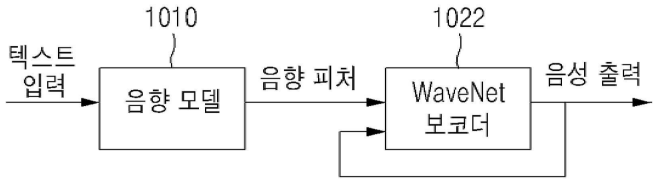
도면9


S(n)
음성 신호
S(n) = e(n) * h(n)
LP 필터
e(n)
여기 신호

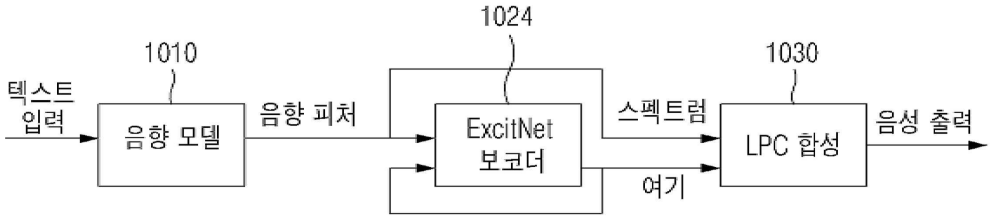
도면10a



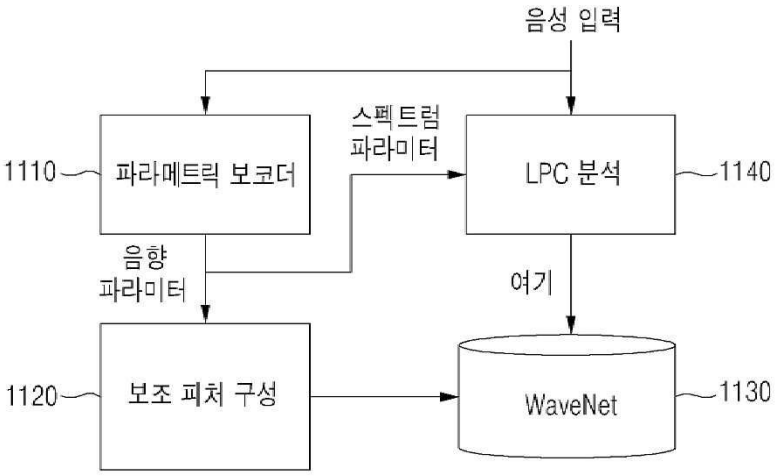
도면10b



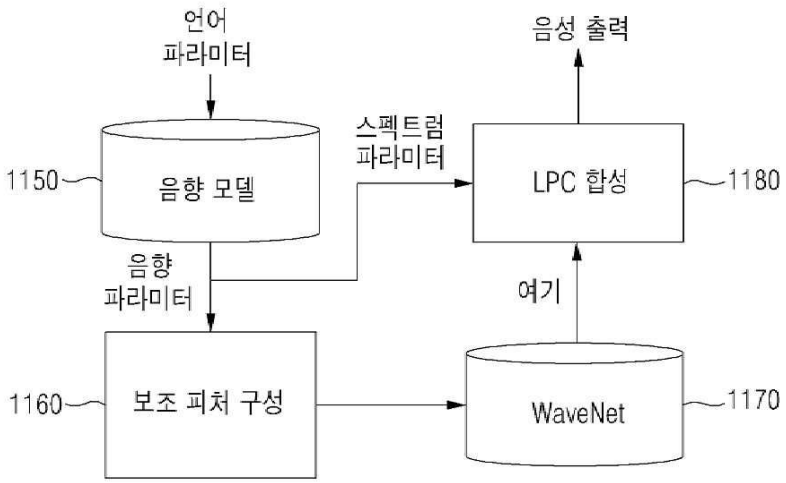
도면10c



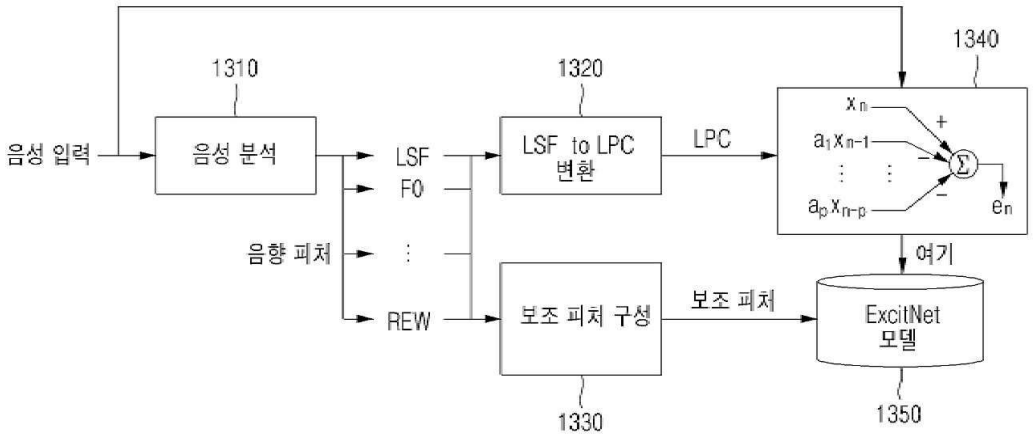
도면11



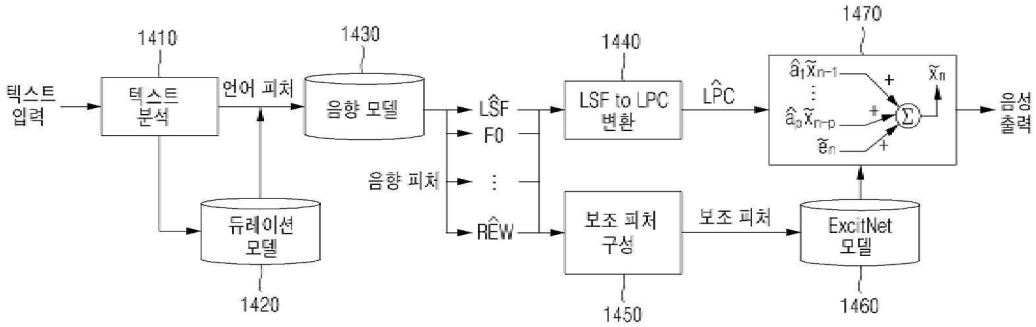
도면12



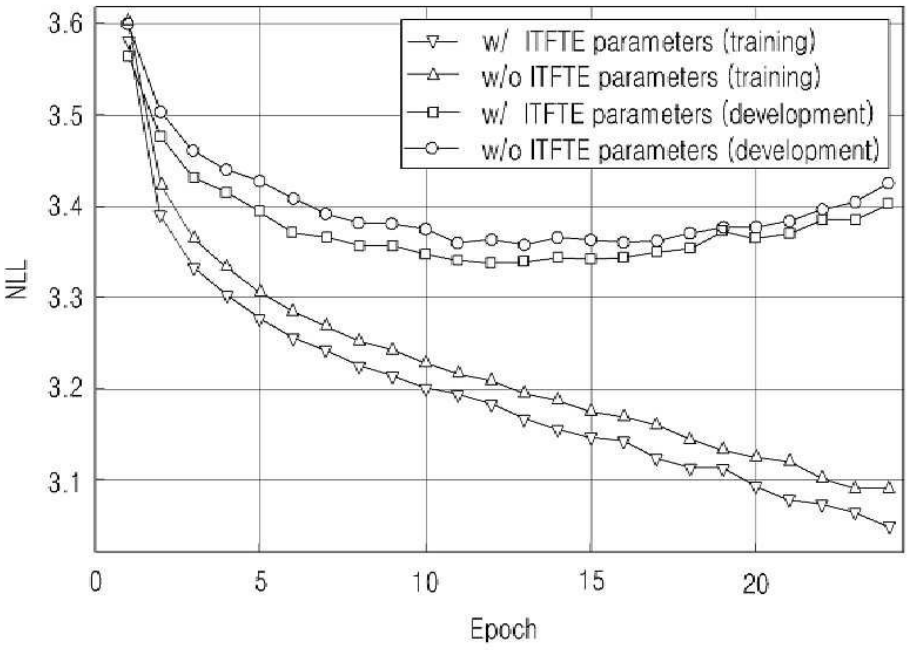
도면13



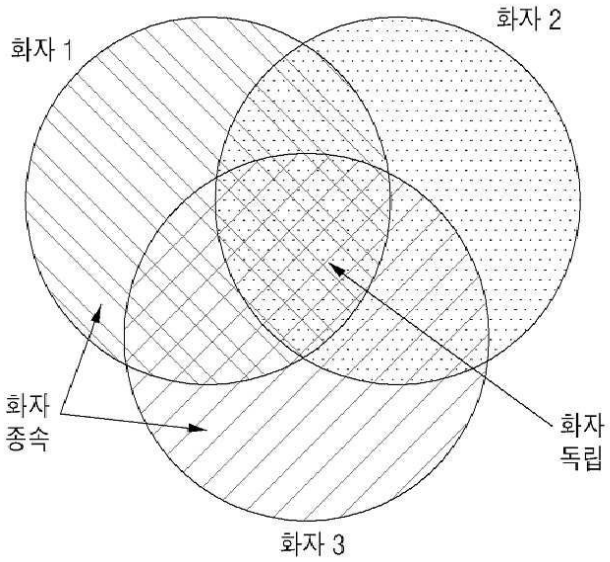
도면14



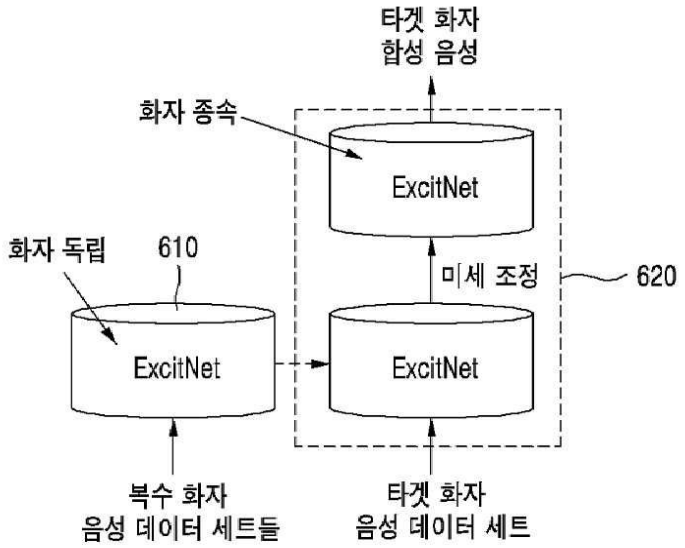
도면15


w/ ITFTE parameters (training)
w/o ITFTE parameters (training)
w/ ITFTE parameters (development)
w/o ITFTE parameters (development)
NLL
Epoch
3.6
3.5
3.4
3.3
3.2
3.1
0
5
10
15
20

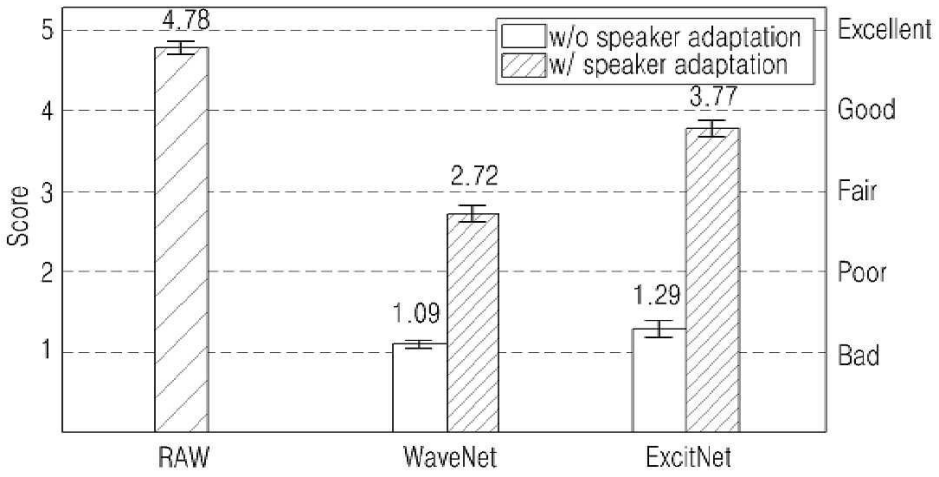
도면16


화자 1
화자 2
화자 3
화자 종속
화자 독립

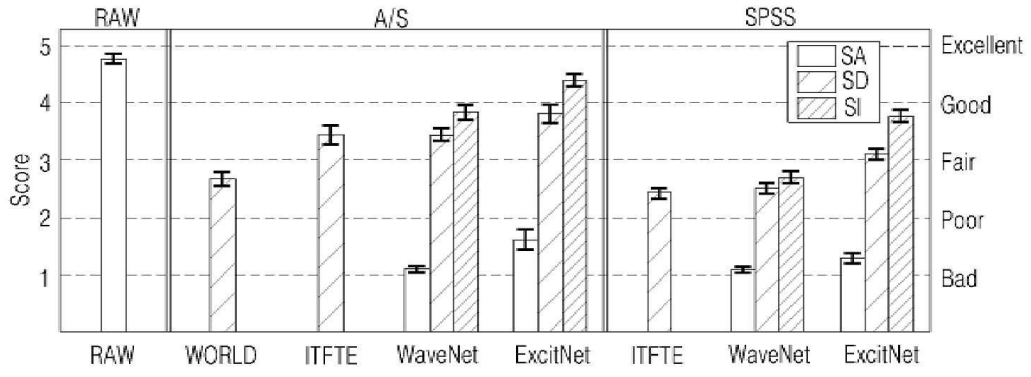
도면17



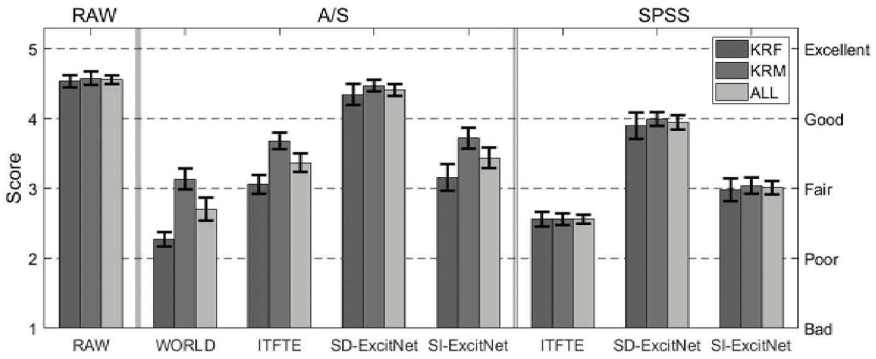
도면18



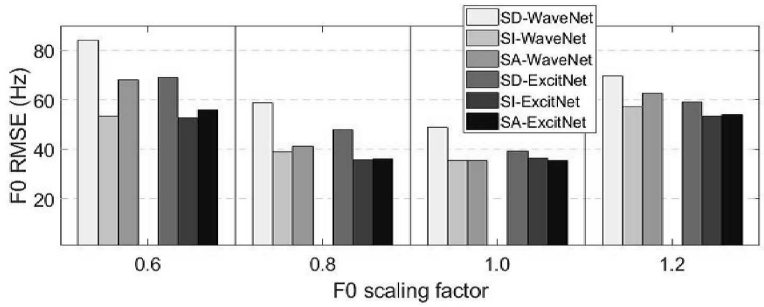
도면19



도면20


RAW
A/S
SPSS
KRF
KRM
ALL
Excellent
Good
Fair
Poor
Bad
Score
5
4
3
2
1
RAW
WORLD
ITFTE
SD-ExcitNet
SI-ExcitNet
ITFTE
SD-ExcitNet
SI-ExcitNet

도면21


SD-WaveNet
SI-WaveNet
SA-WaveNet
SD-ExcitNet
SI-ExcitNet
SA-ExcitNet
F0 RMSE (Hz)
80
60
40
20
0.6
0.8
1.0
1.2
F0 scaling factor