



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년10월02일
(11) 등록번호 10-2711817
(24) 등록일자 2024년09월25일

(51) 국제특허분류(Int. Cl.)
G10L 13/047 (2013.01) G10L 19/04 (2006.01)
G10L 25/18 (2013.01)
(52) CPC특허분류
G10L 13/047 (2013.01)
G10L 19/04 (2013.01)
(21) 출원번호 10-2022-0109743
(22) 출원일자 2022년08월31일
심사청구일자 2022년08월31일
(65) 공개번호 10-2022-0127190
(43) 공개일자 2022년09월19일
(30) 우선권주장
202111138464.5 2021년09월27일 중국(CN)
(56) 선행기술조사문헌
KR1020200092501 A*
Lauri Juvela et al., 'GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-spectrogram', arXiv:1904.03976v3 [eess.AS], 26 Jun 2019.*
Takuma Okamoto et al., 'AN INVESTIGATION OF SUBBAND WAVENET VOCODER COVERING ENTIRE AUDIBLE FREQUENCY RANGE WITH LIMITED ACOUSTIC FEATURES', ICASSP 2018, April 2018.*
Xin Wa et al., 'Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis', IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 28, 2020.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
베이징 바이두 넷컴 사이언스 테크놀로지 컴퍼니 리미티드
중국 베이징 하이디안 디스트릭트 샹디 10번가 넘버 10, 바이두 캠퍼스 2층
(72) 발명자
호우, 지안캉
중국 베이징 100085 하이디안 디스트릭트 샹디 10번가 넘버 10, 바이두 캠퍼스 2층
순, 타오
중국 베이징 100085 하이디안 디스트릭트 샹디 10번가 넘버 10, 바이두 캠퍼스 2층
(뒷면에 계속)
(74) 대리인
특허법인성암

전체 청구항 수 : 총 9 항

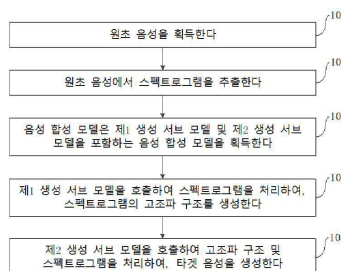
심사관 : 정성윤

(54) 발명의 명칭 음성 처리 방법, 장치, 전자 기기 및 저장 매체

(57) 요약

본 발명은 음성 처리 방법, 장치, 전자 기기 및 저장 매체를 제공하는 바, 음성 기술 및 딥 러닝 등 인공지능 기술 분야에 관한 것이고, 구체적인 구현 수단은, 원초 음성을 획득하는 단계; 상기 원초 음성에서 스펙트로그램을 추출하는 단계; 음성 합성 모델은 제1 생성 서브 모델 및 제2 생성 서브 모델을 포함하는 음성 합성 모델을 획득한다; 제1 생성 서브 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성한다; 제2 생성 서브 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성한다
(뒷면에 계속)

대표도 - 도1



하는 단계; 제1 생성 서브 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하는 단계; 및 제2 생성 서브 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성하는 단계를 포함한다. 이로하여, 당해 방법은 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특징을 구비한다.

(52) CPC특허분류

G10L 25/18 (2013.01)

(72) 발명자

니, 지평

중국 베이징 100085 하이디안 디스트릭트 샹디 10
번가 넘버 10, 바이두 캠퍼스 2층

창, 리치양

중국 베이징 100085 하이디안 디스트릭트 샹디 10
번가 넘버 10, 바이두 캠퍼스 2층

지아, 레이

중국 베이징 100085 하이디안 디스트릭트 샹디 10
번가 넘버 10, 바이두 캠퍼스 2층

왕, 하이펑

중국 베이징 100085 하이디안 디스트릭트 샹디 10
번가 넘버 10, 바이두 캠퍼스 2층

명세서

청구범위

청구항 1

음성 처리 방법에 있어서,

원초 음성을 획득하는 단계;

상기 원초 음성에서 스펙트로그램을 추출하는 단계;

제1 생성 서브 모델 및 제2 생성 서브 모델을 포함하는 음성 합성 모델을 획득하는 단계;

상기 스펙트로그램을 상기 음성 합성 모델에 입력하고, 상기 제1 생성 서브 모델을 호출하여, 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 스펙트로그램을 처리하여, 상기 스펙트로그램의 고조파 구조를 생성하고, 상기 고조파 구조를 조건 정보로서 상기 제2 생성 서브 모델에 입력하는 단계; 및

상기 제2 생성 서브 모델을 호출하여, 내장된 제2 순수 컨볼루션 네트워크를 통해 상기 고조파 구조를 기반으로 상기 스펙트로그램을 처리하여 복수의 서브 밴드 음성을 획득하고, 내장된 다중 서브 밴드 합성기를 통해 상기 복수의 서브 밴드 음성을 합성시켜, 타겟 음성을 생성하는 단계; 를 포함하고,

상기 제1 생성 서브 모델은 하기의 방식을 통해 생성하고,

상기 방식은,

샘플 음성을 획득하고, 상기 샘플 음성에서 샘플 스펙트로그램을 추출하는 단계;

상기 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 샘플 스펙트로그램을 처리하여, 필터 बैं크 및 다중 세트의 샘플 사인과 파라미터를 생성하는 단계;

상기 필터 बैं크를 통해 화이트 노이즈를 필터링하여, 상기 샘플 스펙트로그램의 비주기 신호를 생성하는 단계;

상기 다중 세트의 샘플 사인과 파라미터의 사인파를 각각 생성하는 단계;

상기 다중 세트의 샘플 사인과 파라미터의 사인파를 중첩하여, 상기 샘플 스펙트로그램의 샘플 고조파 구조를 생성하는 단계;

상기 샘플 고조파 구조와 상기 비주기 신호를 중첩하여, 예측된 제1 합성 음성을 획득하는 단계; 및

상기 제1 합성 음성 및 상기 샘플 음성에 따라 제1 손실 값을 생성하고, 상기 제1 손실 값에 따라 상기 제1 생성 서브 모델을 훈련하는 단계; 를 포함하는,

것을 특징으로 하는 음성 처리 방법.

청구항 2

제1항에 있어서,

상기 제1 생성 서브 모델을 호출하여, 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 스펙트로그램을 처리하여, 상기 스펙트로그램의 고조파 구조를 생성하는 단계는,

상기 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 스펙트로그램을 처리하여, 다중 세트의 사인과 파라미터를 생성하는 단계 - 각 세트의 상기 사인과 파라미터는 진폭 및 주파수를 포함함 - ;

상기 다중 세트의 사인과 파라미터의 사인파를 각각 생성하는 단계; 및

상기 다중 세트의 사인과 파라미터의 사인파를 중첩하여, 상기 고조파 구조를 생성하는 단계; 를 포함하는,

것을 특징으로 하는 음성 처리 방법.

청구항 3

제1항에 있어서,

상기 제2 생성 서버 모델은 하기의 방식을 통해 생성하고, 상기 방식은,

상기 샘플 스펙트로그램 및 상기 샘플 고조파 구조를 상기 제2 생성 서버 모델에 입력하고, 상기 내장된 제2 순수 컨볼루션 네트워크를 통해 상기 샘플 고조파 구조를 기반으로 상기 샘플 스펙트로그램을 처리하여 복수의 샘플 서버 밴드 음성을 획득하고, 상기 내장된 다중 서버 밴드 합성기를 통해 상기 복수의 샘플 서버 밴드 음성을 합성시켜, 예측된 제2 합성 음성을 획득하는 단계;

판별기에 따라 상기 샘플 음성 및 상기 제2 합성 음성을 판별하여, 제2 손실 값을 생성하는 단계; 및

상기 제2 손실 값에 따라 상기 제2 생성 서버 모델을 훈련하는 단계; 를 포함하는,

것을 특징으로 하는 음성 처리 방법.

청구항 4

음성 처리 장치에 있어서,

원초 음성을 획득하는데 사용되는 제1 획득 모듈;

상기 원초 음성에서 스펙트로그램을 추출하는데 사용되는 추출 모듈;

제1 생성 서버 모델 및 제2 생성 서버 모델을 포함하는 음성 합성 모듈을 획득하는데 사용되는 제2 획득 모듈;

상기 제1 생성 서버 모델을 호출하여, 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 스펙트로그램을 처리하여, 상기 스펙트로그램의 고조파 구조를 생성하는데 사용되는 제1 생성 모듈;

상기 제2 생성 서버 모델을 호출하여, 내장된 제2 순수 컨볼루션 네트워크를 통해 상기 고조파 구조를 기반으로 상기 스펙트로그램을 처리하여 복수의 서버 밴드 음성을 획득하고, 내장된 다중 서버 밴드 합성기를 통해 상기 복수의 서버 밴드 음성을 합성시켜, 타겟 음성을 생성하는데 사용되는 제2 생성 모듈; 및

제1 훈련 모듈; 을 포함하고,

상기 제1 훈련 모듈은 하기의 방식을 통해 상기 제1 생성 서버 모델을 생성하고,

상기 방식은,

샘플 음성을 획득하고, 상기 샘플 음성에서 샘플 스펙트로그램을 추출하고;

상기 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 샘플 스펙트로그램을 처리하여, 필터 बैं크 및 다중 세트의 샘플 사인과 파라미터를 생성하고;

상기 필터 बैं크를 통해 화이트 노이즈를 필터링하여, 상기 샘플 스펙트로그램의 비주기 신호를 생성하고;

상기 다중 세트의 샘플 사인과 파라미터의 사인파를 각각 생성하고;

상기 다중 세트의 샘플 사인과 파라미터의 사인파를 중첩하여, 상기 샘플 스펙트로그램의 샘플 고조파 구조를 생성하고;

상기 샘플 고조파 구조와 상기 비주기 신호를 중첩하여, 예측된 제1 합성 음성을 획득하고;

상기 제1 합성 음성 및 상기 샘플 음성에 따라 제1 손실 값을 생성하고, 상기 제1 손실 값에 따라 상기 제1 생성 서버 모델을 훈련하는,

것을 특징으로 하는 음성 처리 장치.

청구항 5

제4항에 있어서,

상기 제1 생성 모듈은,

상기 내장된 제1 순수 컨볼루션 네트워크를 통해 상기 스펙트로그램을 처리하여, 다중 세트의 사인과 파라미터를 생성하고, 각 세트의 상기 사인과 파라미터는 진폭 및 주파수를 포함하고;

상기 다중 세트의 사인파 파라미터의 사인파를 각각 생성하고;
 상기 다중 세트의 사인파 파라미터의 사인파를 중첩하여, 상기 고조파 구조를 생성하는데 사용되는,
 것을 특징으로 하는 음성 처리 장치.

청구항 6

제4항에 있어서,
 상기 장치는 제2 훈련 모듈을 포함하고,
 상기 제2 훈련 모듈은 하기의 방식을 통해 상기 제2 생성 서브 모델을 생성하고, 상기 방식은,
 상기 샘플 스펙트로그램 및 상기 샘플 고조파 구조를 상기 제2 생성 서브 모델에 입력하고, 상기 내장된 제2 순
 수 컨볼루션 네트워크를 통해 상기 샘플 고조파 구조를 기반으로 상기 샘플 스펙트로그램을 처리하여 복수의 샘
 플 서브 밴드 음성을 획득하고, 상기 내장된 다중 서브 밴드 합성기를 통해 상기 복수의 샘플 서브 밴드 음성을
 합성시켜, 예측된 제2 합성 음성을 획득하고;
 판별기에 따라 상기 샘플 음성 및 상기 제2 합성 음성을 판별하여, 제2 손실 값을 생성하고;
 상기 제2 손실 값에 따라 상기 제2 생성 서브 모델을 훈련하는,
 것을 특징으로 하는 음성 처리 장치.

청구항 7

전자 기기에 있어서,
 적어도 하나의 프로세서; 및
 상기 적어도 하나의 프로세서에 통신 가능하게 연결되는 메모리; 를 포함하고,
 상기 메모리에는 상기 적어도 하나의 프로세서에 의해 실행 가능한 명령이 저장되어 있고, 상기 명령이 상기 적
 어도 하나의 프로세서에 의해 실행될 경우, 상기 적어도 하나의 프로세서가 제1항 내지 제3항 중 어느 한 항의
 음성 처리 방법을 수행하는,
 것을 특징으로 하는 전자 기기.

청구항 8

컴퓨터 명령이 저장되어 있는 비일시적 컴퓨터 판독 가능 저장 매체에 있어서,
 상기 컴퓨터 명령은 컴퓨터가 제1항 내지 제3항 중 어느 한 항의 음성 처리 방법을 수행하도록 하는,
 것을 특징으로 하는 비일시적 컴퓨터 판독 가능 저장 매체.

청구항 9

컴퓨터 판독 가능 저장 매체에 저장되어 있는 컴퓨터 프로그램에 있어서,
 상기 컴퓨터 프로그램이 프로세서에 의해 수행될 경우 제1항 내지 제3항 중 어느 한 항의 음성 처리 방법이 구
 현되는,
 것을 특징으로 하는 컴퓨터 프로그램.

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

발명의 설명

기술 분야

[0001] 본 발명은 컴퓨터 기술 분야에 관한 것으로, 구체적으로 음성 기술 및 딥 러닝 등 인공지능 기술 분야에 관한 것이고, 특히 음성 처리 방법, 장치, 전자 기기 및 저장 매체에 관한 것이다.

배경 기술

[0002] 보코더 기술은 음향 특징을 음성 신호의 한 향으로 전환하는 기술이다. 보코더는 음성 합성 링크의 중요한 구성 부분이고, 직접적으로 합성 오디오의 안정성, 음질 및 표현력을 결정한다.

[0003] 음성 합성 기술의 발전은 현재까지 다양한 보코더 기술을 형성하고, 특히 최근 몇 년간 딥 러닝 기술의 점차적인 성숙에 따라, 합성 품질이 비교적 좋은 신경 보코더가 많이 나타나고 있다. 당해 보코더는 딥 러닝 기술을 사용할지 여부에 따라, 전통 보코더 및 신경망 보코더로 나눌 수 있고, 딥 러닝 기술을 응용하지 않는 보코더는 전통 보코더라고 통칭하고, 다른 유형은 딥 러닝 기술을 사용한 보코더이고, 신경망 보코더라고 한다.

발명의 내용

해결하려는 과제

[0004] 본 발명은 음성 처리 방법, 장치, 전자 기기 및 저장 매체를 제공한다.

과제의 해결 수단

- [0005] 본 발명의 일 측면에 따르면, 음성 처리 방법을 제공하고, 상기 방법은,
- [0006] 원초 음성을 획득하는 단계;
- [0007] 상기 원초 음성에서 스펙트로그램을 추출하는 단계;
- [0008] 제1 생성 서브 모델 및 제2 생성 서브 모델을 포함하는 음성 합성 모델을 획득하는 단계;
- [0009] 상기 제1 생성 서브 모델을 호출하여 상기 스펙트로그램을 처리하여, 상기 스펙트로그램의 고조파 구조를 생성하는 단계; 및
- [0010] 상기 제2 생성 서브 모델을 호출하여 상기 고조파 구조 및 상기 스펙트로그램을 처리하여, 타겟 음성을 생성하는 단계; 를 포함한다.
- [0011] 본 발명의 다른 측면에 따르면, 음성 처리 장치를 제공하고, 상기 장치는,
- [0012] 원초 음성을 획득하는데 사용되는 제1 획득 모듈;
- [0013] 상기 원초 음성에서 스펙트로그램을 추출하는데 사용되는 추출 모듈;
- [0014] 제1 생성 서브 모델 및 제2 생성 서브 모델을 포함하는 음성 합성 모델을 획득하는데 사용되는 제2 획득 모듈;
- [0015] 상기 제1 생성 서브 모델을 호출하여 상기 스펙트로그램을 처리하여, 상기 스펙트로그램의 고조파 구조를 생성하는데 사용되는 제1 생성 모듈; 및
- [0016] 상기 제2 생성 서브 모델을 호출하여 상기 고조파 구조 및 상기 스펙트로그램을 처리하여, 타겟 음성을 생성하는데 사용되는 제2 생성 모듈; 을 포함한다.
- [0017] 본 발명의 다른 측면에 따르면, 전자 기기를 제공하고, 상기 전자 기기는,
- [0018] 적어도 하나의 프로세서; 및
- [0019] 상기 적어도 하나의 프로세서에 통신 가능하게 연결되는 메모리; 를 포함하고,

- [0020] 상기 메모리에는 상기 적어도 하나의 프로세서에 의해 실행 가능한 명령이 저장되어 있고, 상기 명령이 상기 적어도 하나의 프로세서에 의해 실행될 경우, 상기 적어도 하나의 프로세서가 상기 일 측면 실시예의 음성 처리 방법을 수행한다.
- [0021] 본 발명의 다른 측면에 따르면, 컴퓨터 명령이 저장되어 있는 비일시적 컴퓨터 판독 가능 저장 매체를 제공하고, 상기 컴퓨터 명령은 컴퓨터가 상기 일 측면 실시예의 음성 처리 방법을 수행하도록 한다.
- [0022] 본 발명의 다른 측면에 따르면, 컴퓨터 판독 가능 저장 매체에 저장되어 있는 컴퓨터 프로그램을 제공하고, 상기 컴퓨터 프로그램이 프로세서에 의해 수행될 경우 상기 일 측면 실시예의 음성 처리 방법이 구현된다.
- [0023] 이해해야 할 것은, 본 발명의 내용 부분에서 설명하는 내용은 본 발명 실시예의 관건 또는 중요한 특징을 식별하기 위한 것이 아니고, 본 발명의 범위를 한정하기 위한 것도 아니다. 본 발명의 기타 특징은 이하의 명세서를 통해 용이하게 이해된다.

도면의 간단한 설명

- [0024] 도면은 본 기술적 수단을 더 잘 이해하는데 사용되고, 본 발명을 한정하려는 것은 아니다.
 도1은 본 발명의 실시예에 제공된 음성 처리 방법의 흐름도이다.
 도2는 본 발명의 실시예에 제공된 다른 음성 처리 방법의 흐름도이다.
 도3은 본 발명의 실시예에 제공된 다른 음성 처리 방법의 흐름도이다.
 도4는 본 발명의 실시예에 제공된 다른 음성 처리 방법의 흐름도이다.
 도5는 본 발명의 실시예에 제공된 다른 음성 처리 방법의 흐름도이다.
 도6은 본 발명의 실시예에 제공된 음성 처리 장치의 구조 개략도이다.
 도7은 본 발명 실시예의 음성 처리 방법의 전자 기기 블록도이다.

발명을 실시하기 위한 구체적인 내용

- [0025] 이하 도면과 결합하여 본 발명의 예시적인 실시예를 설명한다. 여기에는 이해를 돕기 위해 본 발명의 실시예의 다양한 세부 사항을 포함하고, 실시예들은 단지 예시적인 것으로 간주되어야 한다. 때문에 본 발명에 속하는 기술 분야의 통상의 기술자는 본 발명의 범위 및 사상을 벗어나지 않고 실시예에 여러가지 변경과 수정을 할 수 있다는 것을 인식해야 한다. 동시에 정확성과 간결성을 위해 하기의 설명에서 공지 기능과 구조에 대한 설명은 생략한다.
- [0026] 아래는 도면을 참조하여 본 발명 실시예의 음성 처리 방법, 장치, 전자 기기 및 저장 매체를 설명한다.
- [0027] 인공지능, 컴퓨터로 사람의 일부 사고 프로세스 및 지능 행위(예를 들면, 학습, 추리, 사고, 계획등)를 시뮬레이션하는 학과이고, 하드웨어 차원의 기술이 있고, 소프트웨어 차원의 기술도 있다. 인공지능 하드웨어 기술은 통상적으로, 센서, 전용 인공지능 칩, 클라우드 컴퓨팅, 분산 메모리 및 빅데이터 처리 등 기술을 포함하고; 인공지능 소프트웨어 기술은 주로, 컴퓨터 비전 기술, 음성 인식 기술, 자연 언어 처리 기술 및 기계 학습/딥 러닝, 빅데이터 프로세싱 기술, 지식 그래프 기술 등 몇 가지 방향을 포함한다.
- [0028] 음성 기술은 자동 음성 인식 기술(Automatic Speech Recognition, ASR) 및 음성 합성 기술(Text to Speech, TTS)을 포함하는 컴퓨터 분야의 핵심 기술이다. 컴퓨터가 듣고, 보고, 말하고, 느낄 수 있도록 하는 것은 미래 인간 - 기계 인터페이스의 발전 방향이고, 음성은 미래에 가장 유망한 인간 - 기계 인터페이스 방법이 될 것이고, 음성은 기타 인터페이스 방식에 대비해 더 많은 우세를 구비하고; 최초의 음성 기술은 "자동 번역 전화" 계획에 의해 시작된 것이고, 음성 인식, 자연어 이해 및 음성 합성의 세 가지 매우 중요한 기술을 포함한다. 음성 인식의 연구 작업은 20세기 50년대 AT&T 벨 실험실의 Audry 시스템으로 거슬러 올라갈 수 있고, 이 후 연구자들은 큰 어휘량, 연속 음성 및 비특정인의 3대 장애물을 점차 돌파하였고; 컴퓨터가 말하게 하려면 음성 합성 기술을 사용해야 하고, 당해 핵심은 텍스트 음성 변환 기술(Text to Speech)이고, 음성 합성은 심지어 자동차의 정보 시스템에도 적용되고 있고, 차주는 컴퓨터에 다운로드된 텍스트 파일, 이메일, 인터넷 뉴스 및 소설을 음성으로 변환하고 차에서 들을 수 있다.
- [0029] 딥러닝은 기계학습 분야의 새로운 연구 방향이다. 딥 러닝은 샘플 데이터의 내적 법칙과 표현 계층을 학습하는

것이고, 당해 학습 과정에서 획득된 정보는 텍스트, 이미지, 사운드 등의 데이터 해석에 큰 도움이 된다. 당해 최종 목표는 기계가 인간처럼 분석하고 학습하는 능력을 구비하고 문자, 이미지 및 소리와 등 데이터를 인식할 수 있도록 하는 것이다. 딥 러닝은 복잡한 기계학습 알고리즘이고, 음성 및 이미지 인식 측면에서 획득된 효과는 이전의 관련 기술을 훨씬 초과한다.

- [0030] 본 발명의 실시예에 제공된 음성 처리 방법은, 전자 기기에 의해 수행될 수 있고, 당해 전자 기기는 PC(Personal Computer, 개인용 컴퓨터) 컴퓨터, 태블릿 PC, 팜톱 컴퓨터, 휴대폰 또는 서버 등일 수 있고, 여기서 어떠한 한정도 하지 않는다.
- [0031] 본 발명의 실시예에서, 전자 기기에는 처리 어셈블리, 저장 어셈블리 및 구동 어셈블리가 설치되어 있을 수 있다. 선택적으로, 당해 구동 어셈블리 및 처리 어셈블리는 통합 설치될 수 있고, 당해 저장 어셈블리는 동작 시스템, 응용 프로그램 또는 기타 프로그램 모듈을 저장할 수 있고, 당해 처리 어셈블리는 저장 어셈블리에 저장된 응용 프로그램을 수행하여 본 발명의 실시예에 제공된 음성 처리 방법을 구현한다.
- [0032] 도1은 본 발명의 실시예에 제공된 음성 처리 방법의 흐름도이다.
- [0033] 본 발명 실시예의 음성 처리 방법은, 본 발명의 실시예에 제공된 음성 처리 장치에 의해 수행될 수도 있고, 당해 장치는 전자 기기에 구성되어, 획득된 원초 음성에서 스펙트로그램을 추출하도록 구현할 수 있고, 음성 합성 모델의 제1 생성 서브 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하고, 음성 합성 모델의 제2 생성 서브 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성함으로써, 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특점을 구비한다.
- [0034] 일 가능한 상황으로서, 본 발명 실시예의 음성 처리 방법은 서버단에서 수행될 수도 있고, 서버는 클라우드 서버일 수 있고, 클라우드에서 당해 음성 처리 방법을 수행할 수 있다.
- [0035] 본 발명의 실시예에서, 본 발명의 당해 실시예의 음성 처리 방법은 일부 음성 방송(예를 들면, 지도 네비게이션 음성, 차량 사물 통신 음성 인터랙션, 사전 펜 방송 등)의 APP(Application, 응용 프로그램)에 응용될 수 있고, 당해 APP는 PC 컴퓨터, 태블릿 PC, 팜톱 컴퓨터, 휴대폰 등 전자 기기에 장착될 수 있다.
- [0036] 도1에 도시된 바와 같이, 당해 음성 처리 방법은 단계101 내지 단계105를 포함한다.
- [0037] 단계101에서, 원초 음성을 획득한다. 원초 음성은 다수일 수 있고, 당해 원초 음성은 각 언어 표현을 사용한 음성일 수 있다. 예를 들면, 중국어 음성, 영어 음성, 러시아어 음성, 말레이어 음성 및 중국어와 영어의 혼용 음성 등일 수 있고, 여기서 어떠한 한정도 하지 않는다. 원초 음성은 1개의 문자, 1개의 단락 또는 1개의 편장을 포함할 수 있다. 예를 들면, 보도 자료 등을 포함할 수 있다.
- [0038] 본 발명의 실시예에서, 상기 원초 음성은 사용자가 음성 인식을 통해 입력한 음성 정보를 포함할 수 있고, 사용자가 네트워크에서 다운 받은 일부 영화, 음악, 드라마 등의 오디오 정보(음성 정보) 및 사용자가 음성 제공된 기기에서 획득한 음성 정보를 포함할 수 있고, 음성 제공 기기는 MP3(음악 파일을 재생할 수 있는 플레이어), VCD 플레이어, 서버, 모바일 단말 및 스마트 하드 디스크를 포함할 수 있다. 여기서 어떠한 한정도 하지 않는다.
- [0039] 설명해야 할 것은, 사용자는 수집(획득)된 음성 정보를 전자 기기의 저장 공간에 입력하여, 저장할 수 있어 후속의 사용에 편리하다. 당해 저장 공간은 엔티티를 기반으로 하는 저장 공간에 한정되지 않는다. 예를 들면, 하드 디스크일 수 있고, 당해 저장 공간은 전자 기기를 연결하는 웹 하드의 저장 공간(클라우드 저장 공간)일 수도 있다.
- [0040] 구체적으로, 전자 기기(예를 들면, 휴대폰) 자체의 저장 공간에서 원초 음성을 획득하고, 또는 자체의 음성 기능에 의해 녹음하여 원초 음성을 획득하고, 또는 음성 제공 기기에서 원초 음성을 획득할 수 있다.
- [0041] 단계102에서, 원초 음성에서 스펙트로그램을 추출한다. 당해 스펙트로그램은 Mel(멜) 스펙트로그램일 수 있다.
- [0042] 본 발명의 실시예에서, 미리 설정된 추출 알고리즘에 따라 원초 음성에서 스펙트로그램을 추출할 수 있다. 미리 설정된 추출 알고리즘은 실제 상황에 따라 설정될 수 있다.
- [0043] 구체적으로, 전자 기기는 원초 음성을 획득한 후, 미리 설정된 추출 알고리즘에 따라 원초 음성에서 스펙트로그램을 추출할 수 있다.
- [0044] 일 가능한 상황으로서, 추출 모델에 따라 원초 음성을 처리하여, 원초 음성에서 스펙트로그램을 추출할 수도 있

다. 설명해야 할 것은, 당해 실시예에서 설명된 추출 모델은 미리 훈련된 것일 수 있고, 전자 기기의 저장 공간에 미리 저장하여, 선택하고 응용하는데 편리하다.

- [0045] 당해 추출 모델의 훈련과 생성은 모두 관련 훈련 서버에 의해 수행될 수 있고, 당해 훈련 서버는 클라우드 서버일 수 있고, 컴퓨터의 호스트일 수도 있고, 당해 훈련 서버와 본 발명의 실시예에 제공된 음성 처리 방법을 수행할 수 있는 전자 기기 사이는, 통신 연결이 구축되어 있고, 당해 통신 연결은 무선 네트워크 및 유선 네트워크 연결 중의 적어도 하나일 수 있다. 당해 훈련 서버는 훈련된 추출 모델을 당해 전자 기기에 송신하여, 당해 전자 기기가 호출해야 할 경우, 당해 전자 기기의 계산 압력을 크게 감소할 수 있다.
- [0046] 구체적으로, 전자 기기는 원초 음성을 획득한 후, 자체의 저장 공간에서 추출 모델을 선택하고, 당해 원초 음성을 당해 추출 모델에 입력함으로써, 당해 추출 모델을 통해 당해 원초 음성에서 스펙트로그램을 추출하여, 당해 추출 모델에 의해 출력된 스펙트로그램을 획득할 수 있다.
- [0047] 다른 가능한 상황으로서, 전자 기기는 추출 도구(예를 들면, 플러그 인)를 사용하여, 원초 음성에서 스펙트로그램을 추출할 수도 있다.
- [0048] 단계103에서, 음성 합성 모델은 제1 생성 서버 모델 및 제2 생성 서버 모델을 포함하는 음성 합성 모델을 획득한다.
- [0049] 본 발명의 실시예에서, 상기 음성 합성 모델은 보코더일 수 있다.
- [0050] 설명해야 할 것은, 당해 실시예에서 설명된 음성 합성 모델은 미리 훈련된 것일 수 있고, 전자 기기의 저장 공간에 미리 저장하여, 선택하고 응용하는데 편리하다. 당해 음성 합성 모델은 순수 컨볼루션 구조일 수 있으므로, 일정한 정도에서 네트워크의 훈련 및 예측 속도를 가속화시킬 수 있다.
- [0051] 단계104에서, 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성한다.
- [0052] 설명해야 할 것은, 당해 실시예에서 설명된 고조파 구조는 주기 신호를 포함할 수 있다.
- [0053] 단계105에서, 제2 생성 서버 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성한다.
- [0054] 구체적으로, 전자 기기는 원초 음성에서 추출된 스펙트로그램을 획득한 후, 먼저 자체의 저장 공간에서 음성 합성 모델을 선택(획득)하고, 당해 스펙트로그램을 당해 음성 합성 모델에 입력함으로써, 당해 음성 합성 모델을 통해 당해 스펙트로그램을 처리하여, 타겟 음성을 생성한다. 당해 음성 합성 모델은 당해 스펙트로그램을 수신한 후, 먼저 내장된 제1 생성 서버 모델을 통해 스펙트로그램을 처리하여, 당해 제1 생성 서버 모델에 의해 출력된 당해 스펙트로그램의 고조파 구조를 획득하고, 내장된 제2 생성 서버 모델을 통해 당해 고조파 구조 및 당해 스펙트로그램을 처리하여, 당해 제2 생성 서버 모델에 의해 출력된 타겟 음성을 획득한다. 즉, 당해 음성 합성 모델에 의해 출력된 타겟 음성을 획득한다. 이로하여, 음질 및 음색이 원초 음성에 더 접근된 타겟 음성을 생성할 수 있고, 떨림 및 음소거 상황이 나타나지 않는다.
- [0055] 본 발명의 실시예에서, 먼저 원초 음성을 획득하고, 원초 음성에서 스펙트로그램을 획득하고, 음성 합성 모델을 획득한다. 음성 합성 모델은 제1 생성 서버 모델 및 제2 생성 서버 모델을 포함하고, 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하고, 제2 생성 서버 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성한다. 이로하여, 당해 방법은 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특징을 구비한다.
- [0056] 상기 실시예를 명확하게 설명하기 위해, 본 발명의 실시예에서, 도2에 도시된 바와 같이, 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하는 단계는 단계201 내지 단계203을 포함한다.
- [0057] 단계201에서, 스펙트로그램을 처리하여, 다중 세트의 사인과 파라미터를 생성하고, 각 세트의 사인과 파라미터는 진폭 및 주파수를 포함한다. 설명해야 할 것은, 당해 실시예의 다중 세트의 사인과 파라미터 중 각 세트의 사인과 파라미터는 모두 고조파 관계를 구성할 수 있다.
- [0058] 본 발명의 실시예에서, 상기 제1 생성 서버 모델은 제1 순수 컨볼루션 네트워크를 포함할 수 있고, 당해 제1 순수 컨볼루션 네트워크는 업 샘플링 컨볼루션 및 잔여 네트워크로 구성될 수 있다.
- [0059] 단계202에서, 다중 세트의 사인과 파라미터의 사인파를 각각 생성한다.

- [0060] 본 발명의 실시예에서, 사인과 생성 전략을 기반으로 다중 세트의 사인과 파라미터의 사인파를 각각 생성할 수 있고, 사인과 생성 전략은 실제 상황에 따라 설정될 수 있다.
- [0061] 일 가능한 상황으로서, 사인과 생성 모델에 따라 다중 세트의 사인과 파라미터의 사인파를 각각 생성할 수 있고, 다중 세트의 사인과 파라미터 중 각 세트의 사인과 파라미터를 사인과 생성 모델에 순차대로 입력함으로써, 당해 사인과 생성 모델을 통해 각 세트 사인과 파라미터의 진폭 및 주파수를 각각 처리하여, 다중 세트의 사인과 파라미터의 사인파를 각각 생성할 수 있다.
- [0062] 단계203에서, 다중 세트의 사인과 파라미터의 사인파를 중첩하여, 고조파 구조를 생성한다.
- [0063] 구체적으로, 제1 생성 서브 모델은 상기 스펙트로그램을 수신한 후, 내장된 제1 순수 컨볼루션 네트워크를 통해 당해 스펙트로그램을 처리하여, 다중 세트의 사인과 파라미터를 생성(예측)하고, 각 세트 사인과 파라미터의 진폭 및 주파수에 따라, 다중 세트의 사인과 파라미터의 사인파를 각각 생성하고, 당해 다중 세트의 사인과 파라미터의 사인파를 중첩하여, 고조파 구조를 생성할 수 있다.
- [0064] 나아가, 제1 생성 서브 모델은 내장된 제1 순수 컨볼루션 네트워크를 통해 당해 고조파 구조를 조건 정보로서 상기 제2 생성 서브 모델에 입력하여, 제2 생성 서브 모델의 생성 프로세스를 지도한다.
- [0065] 이로하여, 제1 순수 컨볼루션 네트워크를 통해 극소의 계산량으로 비교적 뚜렷하고 원초 오디오에 가까운 고조파 구조를 획득할 수 있고, 당해 고조파 구조는 또한 강한 조건 정보로서 제2 생성 서브 모델의 학습을 지도할 수 있고, 제2 생성 서브 모델의 모델링 난이도를 대폭 저하한다. 또한, 당해 고조파 구조를 통해 제2 생성 서브 모델의 생성 프로세스를 지도하여 생성된 타겟 음성의 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 우수한 특점을 구비하도록 한다.
- [0066] 나아가, 본 발명의 실시예에서, 도3에 도시된 바와 같이, 제2 생성 서브 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성하는 단계는 단계301 내지 단계302를 포함한다.
- [0067] 단계301에서, 고조파 구조 및 스펙트로그램에 따라 복수의 서브 밴드 음성을 생성한다.
- [0068] 본 발명의 실시예에서, 상기 제2 생성 서브 모델은 제2 순수 컨볼루션 네트워크 및 다중 서브 밴드 합성기를 포함할 수 있고, 당해 제1 순수 컨볼루션 네트워크는 업 샘플링 컨볼루션 및 잔여 네트워크로 구성될 수 있다.
- [0069] 단계302에서, 복수의 서브 밴드 음성을 합성하여, 타겟 음성을 생성한다.
- [0070] 구체적으로, 제2 생성 서브 모델은 상기 스펙트로그램 및 고조파 구조를 수신한 후, 먼저, 내장된 제2 순수 컨볼루션 네트워크를 통해 당해 고조파 구조를 기반으로 당해 스펙트로그램에 대해 서브 밴드 음성 길이의 업 샘플링을 하고, 당해 고조파 구조를 기반으로 당해 업 샘플링 결과에 대해 잔여 처리를 하여, 복수의 서브 밴드 음성을 획득할 수 있고, 당해 제2 순수 컨볼루션 네트워크는 전치 컨볼루션 알고리즘을 사용하여 당해 스펙트로그램에 대해 서브 밴드 음성 길이의 업 샘플링을 하고, 깊이 분리 가능한 컨볼루션 알고리즘 및 1차원 컨볼루션 알고리즘을 사용하여 당해 업 샘플링의 결과에 대해 여러 번의 잔여 처리를 할 수 있다. 그리고 당해 제2 생성 서브 모델은 내장된 다중 서브 밴드 합성기를 통해 복수의 서브 밴드 음성을 합성시켜, 타겟 음성을 생성할 수 있다. 전치 컨볼루션 알고리즘, 깊이 분리 가능한 컨볼루션 알고리즘 및 1차원 컨볼루션 알고리즘은 실제 상황에 따라 설정될 수 있다.
- [0071] 이로하여, 타겟 음성을 생성할 경우 고조파 구조의 조건 정보를 추가하여, 제2 생성 서브 모델 계산량을 대폭으로 저하시킬 수 있는 동시에, 생성된 오디오 발음의 안정성을 보장할 수 있으므로, 음성 합성 모델이 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적고, 모든 단축 장면에 응용 가능한 우수한 특점을 구비하도록 한다.
- [0072] 본 발명의 실시예에서, 도4에 도시된 바와 같이, 상기 제1 생성 서브 모델은 하기의 단계를 통해 생성하고, 상기 단계는 단계401 내지 단계407을 포함한다.
- [0073] 단계401에서, 샘플 음성을 획득하고, 샘플 음성에서 샘플 스펙트로그램을 추출한다. 샘플 음성은 다수일 수 있고, 당해 샘플 음성은 중국어 음성, 영어 음성 또는 독일어 음성 등일 수 있고, 여기서 어떠한 한정도 하지 않는다.
- [0074] 본 발명의 실시예에서, 샘플 음성을 획득하는 수단은 다수일 수 있고, 동시통역할 경우 동시통역 장치의 음성을 수집하여, 샘플 음성을 획득하고; 인위적이고 자발적으로 샘플 음성을 창조할 수 있다. 예를 들면, 관련 인원의 수요에 따라, 관련 녹음 기기를 통해 샘플 음성의 녹음을 하고; 행인의 음성을 자발적으로 수집하여, 샘플 음성

을 획득할 수 있고; 음성 제공 기기에서 샘플 음성을 획득할 수도 있고, 여기서 어떠한 한정도 하지 않는다.

- [0075] 구체적으로, 샘플 음성을 획득한 후, 상기 미리 설정된 추출 알고리즘에 따라 당해 샘플 음성에서 샘플 스펙트로그램을 추출할 수 있다.
- [0076] 단계402에서, 샘플 스펙트로그램을 처리하여, 필터 बैं크 및 다중 세트의 샘플 사인과 파라미터를 생성한다.
- [0077] 설명해야 할 것은, 당해 실시예에서 설명된 필터 बैं크는 비주기 신호의 필터 बैं크일 수 있다.
- [0078] 단계403에서, 필터 बैं크를 통해 화이트 노이즈를 필터링하여, 샘플 스펙트로그램의 비주기 신호를 생성한다. 설명해야 할 것은, 당해 실시예에서 설명된 화이트 노이즈는 전자 기기의 저장 공간에 미리 저장될 수 있어, 선택하고 응용하는데 편리하고, 또는 화이트 노이즈 생성기를 통해 직접 생성될 수도 있다.
- [0079] 본 발명의 실시예에서, 상기 비주기 신호는 1개 세그먼트의 화이트 노이즈를 통해 획득될 수 있고, 주기 신호는 일련의 사인과 중첩을 통해 획득될 수 있다. 1개의 세그먼트의 주기 신호에서, 기본파 신호 및 고조파 신호를 포함하고, 당해 주기 신호 주기와 같은 사인과 컴포넌트를 기본파 컴포넌트라 하고, 당해 기본파 컴포넌트의 주파수를 기본 주파수라 하고, 주파수가 기본 주파수의 정배수와 같은 사인과 컴포넌트를 고조파라 하고, 기본 주파수와 고조파를 더하면 1개의 주기 신호를 획득할 수 있고, 당해 주기 신호는 상기 실시예의 고조파 구조이다.
- [0080] 단계404에서, 다중 세트의 샘플 사인과 파라미터의 사인파를 각각 생성한다.
- [0081] 단계405에서, 다중 세트의 샘플 사인과 파라미터의 사인파를 중첩하여, 샘플 스펙트로그램의 샘플 고조파 구조를 생성한다.
- [0082] 구체적으로, 훈련할 경우 제1 생성 서버 모델의 제1 순수 컨볼루션 네트워크는, 당해 샘플 스펙트로그램의 입력에 따라, 먼저, 1개 세트의 비주기 신호의 필터 बैं크를 예측하고, 당해 필터 बैं크를 통해 화이트 노이즈에 대해 필터링하여, 샘플 스펙트로그램의 비주기 신호를 획득하고 동시에, 당해 제1 순수 컨볼루션 네트워크는 다중 세트의 고조파 관계를 구성하는 사인과 파라미터를 예측하고, 다중 세트의 고조파 관계를 구성하는 사인과 파라미터의 사인파를 각각 생성할 수 있다. 그리고 당해 고조파 관계를 구성하는 사인파를 중첩하여 샘플 음성의 샘플 고조파 구조를 획득한다.
- [0083] 단계406에서, 샘플 고조파 구조와 비주기 신호를 중첩하여, 예측된 제1 합성 음성을 획득한다.
- [0084] 단계407에서, 제1 합성 음성 및 샘플 음성에 따라 제1 손실 값을 생성하고, 제1 손실 값에 따라 제1 생성 서버 모델을 훈련한다.
- [0085] 본 발명의 실시예에서, 제1 합성 음성과 샘플 음성을 대비(판별)하여, 당해 제1 합성 음성과 당해 샘플 음성 사이의 차이를 획득하고, 당해 차이를 제1 손실 값이라고 할 수 있다. 판별기를 통해 제1 합성 음성 및 샘플 음성을 판별하여 제1 손실 값을 생성하고, 또는 미리 설정된 판별 알고리즘을 통해 제1 합성 음성과 샘플 음성을 판별하여 제1 손실 값을 생성할 수 있고, 여기서 어떠한 한정도 하지 않는다. 당해 판별기는 순수 컨볼루션 구조일 수 있고, 당해 판별기는 전자 기기의 저장 공간에 미리 저장될 수 있고, 선택하고 응용하는데 편리하다. 설명해야 할 것은, 당해 실시예에서 설명된 미리 설정된 판별 알고리즘은 실제 상황에 따라 설정될 수 있다.
- [0086] 구체적으로, 훈련할 경우 제1 생성 서버 모델의 제1 순수 컨볼루션 네트워크는 샘플 음성의 샘플 고조파 구조를 획득한 후, 본 고조파 구조(즉, 주기 신호)와 비주기 신호를 중첩하여 예측된 제1 합성 음성을 획득하고, 당해 제1 합성 음성과 당해 샘플 음성을 대비(판별)하여, 당해 제1 합성 음성과 당해 샘플 음성 사이의 차이를 획득하고, 당해 차이를 제1 손실 값이라고 할 수 있다. 마지막으로 당해 제1 손실 값에 따라 제1 생성 서버 모델을 훈련함으로써, 제1 생성 서버 모델을 최적화하고, 생성의 정확도를 향상시킨다.
- [0087] 본 발명의 실시예에서, 도5에 도시된 바와 같이, 상기 제2 생성 서버 모델은 하기의 단계를 통해 생성하고, 상기 단계는 단계501 내지 단계503을 포함한다.
- [0088] 단계501에서, 샘플 스펙트로그램 및 샘플 고조파 구조를 제2 생성 서버 모델에 입력하여, 예측된 제2 합성 음성을 획득한다.
- [0089] 구체적으로, 훈련할 경우 제2 생성 서버 모델의 제2 순수 컨볼루션 네트워크는 상기 샘플 고조파 구조를 수신한 후, 당해 샘플 고조파 구조를 기반으로 당해 샘플 스펙트로그램에 대해 서버 밴드 음성 길이의 업 샘플링을 하고, 당해 샘플 고조파 구조를 기반으로 당해 업 샘플링의 결과에 대해 잔여 처리를 하여, 복수의 샘플 서버 밴드 음성을 획득할 수 있다. 그리고, 당해 제2 생성 서버 모델의 다중 서버 밴드 합성기는 복수의 서버 밴드 음

성을 합성하여 예측된 제2 합성 음성을 획득할 수 있다.

- [0090] 단계502에서, 판별기에 따라 샘플 음성 및 제2 합성 음성을 판별하여, 제2 손실 값을 생성한다.
- [0091] 단계503에서, 제2 손실 값에 따라 제2 생성 서버 모델을 훈련한다.
- [0092] 구체적으로, 훈련할 경우 상기 다중 서버 밴드 합성기에 의해 예측된 제2 합성 음성 및 상기 샘플 음성을 판별기에 입력함으로써, 당해 판별기를 통해 당해 샘플 음성 및 당해 제2 합성 음성을 판별하여, 당해 판별기에 의해 출력된 제2 손실 값을 획득할 수 있다. 그리고 당해 제2 손실 값에 따라 제2 생성 서버 모델을 훈련함으로써, 제2 생성 서버 모델을 최적화하고, 생성의 정확도를 더 향상시킨다.
- [0093] 일 가능한 상황으로서, 상기 미리 설정된 판별 알고리즘을 통해 샘플 음성 및 제2 합성 음성을 판별하여, 제2 손실 값을 생성할 수 있다.
- [0094] 본 발명의 실시예에서, 상기 음성 합성 모델의 합성 음성의 음질 및 음색은 모두 원초 오디오에 매우 접근되고, 떨림 및 음소거 문제가 나타날 수 없고, 더 중요한 것은 당해 음성 합성 모델의 합성 실시간율은 전통 보코더와 같고, 당해 합성 음성의 품질은 혼한 신경 보코더와 같을 수 있다.
- [0095] 도6은 본 발명의 실시예에 제공된 음성 처리 장치의 구조 개략도이다.
- [0096] 본 발명 실시예의 음성 처리 장치는, 전자 기기에 구성되어, 획득된 원초 음성에서 스펙트로그램을 추출하도록 구현할 수 있고, 음성 합성 모델의 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하고, 음성 합성 모델의 제2 생성 서버 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성함으로써, 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특징을 구비한다.
- [0097] 본 발명의 실시예에서, 본 발명의 당해 실시예의 음성 처리 장치는 일부 음성 방송(예를 들면, 지도 네비게이션 음성, 차량 사물 통신 음성 인터랙션, 사전 펜 방송 등)의 APP(Application, 응용 프로그램)에 설치(장착)될 수 있고, 당해 APP는 PC 컴퓨터, 태블릿 PC, 팜톱 컴퓨터, 휴대폰 등 전자 기기에 장착될 수 있다.
- [0098] 도6에 도시된 바와 같이, 당해 음성 처리 장치(600)는 제1 획득 모듈(610), 추출 모듈(620), 제2 획득 모듈(630), 제1 생성 모듈(640) 및 제2 생성 모듈(650)을 포함한다.
- [0099] 제1 획득 모듈(610)은, 원초 음성을 획득하는데 사용된다. 원초 음성은 다수일 수 있고, 당해 원초 음성은 각 언어 표현을 사용한 음성일 수 있다. 예를 들면, 중국어 음성, 영어 음성, 러시아어 음성, 말레이어 음성 및 중국어와 영어의 혼용 음성 등일 수 있고, 여기서 어떠한 한정도 하지 않는다. 원초 음성은 1개의 문자, 1개의 단락 또는 1개의 편장을 포함할 수 있다. 예를 들면, 보도 자료 등을 포함할 수 있다.
- [0100] 본 발명의 실시예에서, 상기 원초 음성은 사용자가 음성 인식을 통해 입력한 음성 정보를 포함할 수 있고, 사용자가 네트워크에서 다운 받은 일부 영화, 음악, 드라마 등의 오디오 정보(음성 정보) 및 사용자가 음성 제공된 기기에서 획득한 음성 정보를 포함할 수 있고, 음성 제공 기기는 MP3, VCD 플레이어, 서버, 모바일 단말 및 스마트 하드 디스크를 포함할 수 있다. 여기서 어떠한 한정도 하지 않는다.
- [0101] 설명해야 할 것은, 사용자는 수집(획득)된 음성 정보를 전자 기기의 저장 공간에 입력하여, 저장할 수 있어 후속의 사용에 편리하다. 당해 저장 공간은 엔티티를 기반으로 하는 저장 공간에 한정되지 않는다. 예를 들면, 하드 디스크일 수 있고, 당해 저장 공간은 전자 기기를 연결하는 웹 하드의 저장 공간(클라우드 저장 공간)일 수도 있다.
- [0102] 구체적으로, 제1 획득 모듈(610)은 전자 기기의 저장 공간에서 원초 음성을 획득하고, 또는 자체의 음성 기능에 의해 녹음하여 원초 음성을 획득하고, 또는 음성 제공 기기에서 원초 음성을 획득할 수 있다.
- [0103] 추출 모듈(620)은, 원초 음성에서 스펙트로그램을 추출하는데 사용된다. 당해 스펙트로그램은 Mel(멜) 스펙트로그램일 수 있다.
- [0104] 본 발명의 실시예에서, 추출 모듈(620)은 미리 설정된 추출 알고리즘에 따라 원초 음성에서 스펙트로그램을 추출할 수 있다. 미리 설정된 추출 알고리즘은 실제 상황에 따라 설정될 수 있다.
- [0105] 구체적으로, 제1 획득 모듈(610)이 원초 음성을 획득 한 후, 추출 모듈(620)은 미리 설정된 추출 알고리즘에 따라 원초 음성에서 스펙트로그램을 추출할 수 있다.
- [0106] 일 가능한 상황으로서, 추출 모듈(620)은 추출 모델에 따라 원초 음성을 처리하여, 원초 음성에서 스펙트로그램

을 추출할 수도 있다. 설명해야 할 것은, 당해 실시예에서 설명된 추출 모델은 미리 훈련된 것일 수 있고, 전자 기기의 저장 공간에 미리 저장하여, 선택하고 응용하는데 편리하다.

- [0107] 당해 추출 모델의 훈련과 생성은 모두 관련 훈련 서버에 의해 수행될 수 있고, 당해 훈련 서버는 클라우드 서버 일 수 있고, 컴퓨터의 호스트일 수도 있고, 당해 훈련 서버와 본 발명의 실시예에 제공된 음성 처리 장치를 구성할 수 있는 전자 기기 사이는, 통신 연결이 구축되어 있고, 당해 통신 연결은 무선 네트워크 및 유선 네트워크 연결 중의 적어도 하나일 수 있다. 당해 훈련 서버는 훈련된 추출 모델을 당해 전자 기기에 송신하여, 당해 전자 기기가 호출해야 할 경우, 당해 전자 기기의 계산 압력을 크게 감소할 수 있다.
- [0108] 구체적으로, 제1 획득 모듈(610)이 원초 음성을 획득한 후, 추출 모듈(620)은 전자 기기의 저장 공간에서 추출 모델을 선택하고, 당해 원초 음성을 당해 추출 모델에 입력함으로써, 당해 추출 모델을 통해 당해 원초 음성에서 스펙트로그램을 추출하여, 당해 추출 모델에 의해 출력된 스펙트로그램을 획득할 수 있다.
- [0109] 다른 가능한 상황에서, 추출 모듈(620)은 추출 도구(예를 들면, 플러그 인)를 사용하여, 원초 음성에서 스펙트로그램을 추출할 수도 있다.
- [0110] 제2 획득 모듈(630)은, 음성 합성 모델은 제1 생성 서버 모델 및 제2 생성 서버 모델을 포함하는 음성 합성 모델을 획득하는데 사용된다.
- [0111] 본 발명의 실시예에서, 상기 음성 합성 모델은 보코더일 수 있다.
- [0112] 설명해야 할 것은, 당해 실시예에서 설명된 음성 합성 모델은 미리 훈련된 것일 수 있고, 전자 기기의 저장 공간에 미리 저장하여, 선택하고 응용하는데 편리하다. 당해 음성 합성 모델은 순수 컨볼루션 구조일 수 있으므로, 일정한 정도에서 네트워크의 훈련 및 예측 속도를 가속화시킬 수 있다.
- [0113] 제1 생성 모듈(640)은, 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하는데 사용된다.
- [0114] 설명해야 할 것은, 당해 실시예에서 설명된 고조파 구조는 주기 신호를 포함할 수 있다.
- [0115] 제2 생성 모듈(650)은, 제2 생성 서버 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성하는데 사용된다.
- [0116] 본 발명의 실시예에서, 먼저, 제1 획득 모듈을 통해 원초 음성을 획득하고, 추출 모듈을 통해 원초 음성에서 스펙트로그램을 추출하고, 제2 획득 모듈을 통해 음성 합성 모델을 획득하며, 음성 합성 모델은 제1 생성 서버 모델 및 제2 생성 서버 모델을 포함한다. 그리고 제1 생성 모듈을 통해 제1 생성 서버 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하고, 제2 생성 모듈을 통해 제2 생성 서버 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성한다. 이로하여, 당해 장치는 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특징을 구비한다.
- [0117] 본 발명의 실시예에서, 제1 생성 모듈(640)은 구체적으로, 스펙트로그램을 처리하여, 다중 세트의 사인과 파라미터를 생성하고, 각 세트의 사인과 파라미터는 진폭 및 주파수를 포함하고; 다중 세트의 사인과 파라미터의 사인파를 각각 생성하고; 다중 세트의 사인과 파라미터의 사인파를 중첩하여, 고조파 구조를 생성하는데 사용된다.
- [0118] 본 발명의 실시예에서, 제2 생성 모듈(650)은 구체적으로, 고조파 구조 및 스펙트로그램에 따라 복수의 서브 밴드 음성을 생성하고; 복수의 서브 밴드 음성을 합성하여, 타겟 음성을 생성하는데 사용된다.
- [0119] 본 발명의 실시예에서, 도6에 도시된 바와 같이, 당해 음성 처리 장치(600)는 제1 훈련 모듈(660)을 더 포함하고, 제1 훈련 모듈(660)은 하기의 방식을 통해 제1 생성 서버 모델을 생성하고, 상기 방식은 샘플 음성을 획득하고, 샘플 음성에서 샘플 스펙트로그램을 추출하고; 샘플 스펙트로그램을 처리하여, 필터 बैं크 및 다중 세트의 샘플 사인과 파라미터를 생성하고; 필터 बैं크를 통해 화이트 노이즈를 필터링하여, 샘플 스펙트로그램의 비주기 신호를 생성하고; 다중 세트의 샘플 사인과 파라미터의 사인파를 각각 생성하고; 다중 세트의 샘플 사인과 파라미터의 사인파를 중첩하여, 샘플 스펙트로그램의 샘플 고조파 구조를 생성하고; 샘플 고조파 구조와 비주기 신호를 중첩하여, 예측된 제1 합성 음성을 획득하고; 및 제1 합성 음성 및 샘플 음성에 따라 제1 손실 값을 생성하고, 제1 손실 값에 따라 제1 생성 서버 모델을 훈련한다.
- [0120] 본 발명의 실시예에서, 도6에 도시된 바와 같이, 당해 음성 처리 장치(600)는 제2 훈련 모듈(670)을 더 포함할 수 있고, 제2 훈련 모듈(670)은 하기의 방식을 통해 제2 생성 서버 모델을 생성하고, 상기 방식은 샘플 스펙트

로그램 및 샘플 고조파 구조를 제2 생성 서브 모델에 입력하여, 예측된 제2 합성 음성을 획득하고; 판별기에 따라 샘플 음성 및 제2 합성 음성을 판별하여, 제2 손실 값을 생성하고; 및 제2 손실 값에 따라 제2 생성 서브 모델을 훈련한다.

- [0121] 설명해야 할 것은, 상기 음성 처리 방법 실시예에 대한 해석 설명은 당해 실시예의 음성 처리 장치에도 적용되고, 여기서 더는 설명하지 않는다.
- [0122] 본 발명 실시예의 음성 처리 장치는, 먼저, 제1 획득 모듈을 통해 원초 음성을 획득하고, 추출 모듈을 통해 원초 음성에서 스펙트로그램을 추출하고, 제2 획득 모듈을 통해 음성 합성 모델을 획득하며, 음성 합성 모델은 제1 생성 서브 모델 및 제2 생성 서브 모델을 포함한다. 그리고 제1 생성 모듈을 통해 제1 생성 서브 모델을 호출하여 스펙트로그램을 처리하여, 스펙트로그램의 고조파 구조를 생성하고, 제2 생성 모듈을 통해 제2 생성 서브 모델을 호출하여 고조파 구조 및 스펙트로그램을 처리하여, 타겟 음성을 생성한다. 이로하여, 당해 장치는 합성 음질이 비교적 좋고, 음색 환원도가 높고, 발음이 안정적이며 계산량이 적은 특징을 구비한다.
- [0124] 본 발명의 기술 수단에서, 언급된 사용자의 개인정보의 수집, 저장, 사용, 가공, 전송, 제공 및 공개등 처리는 관련 법규의 규정에 해당되고, 공서양속을 위반하지 않는다.
- [0125] 본 발명의 실시예에 따르면, 본 발명은 전자 기기, 관독 가능 저장 매체 및 컴퓨터 프로그램을 더 제공한다.
- [0126] 도7에 도시된 바와 같이, 도7은 본 발명 실시예를 구현하는데 사용되는 전자 기기(700)의 개략적인 블록도이다. 전자 기기는 다양한 형식의 디지털 컴퓨터를 표시한다. 예를 들면, 랩톱 컴퓨터, 데스크톱 컴퓨터, 워크스테이션, 개인 정보 단말(PAD), 서버, 블레이드 서버, 메인 프레임 및 기타 적합한 컴퓨터일 수 있다. 전자 기기는 다양한 형식의 모바일 장치를 표시한다. 예를 들면 개인 정보 단말(PAD), 셀룰러 폰, 스마트 폰, 웨어러블 기기 및 기타 유사한 컴퓨팅 장치일 수 있다. 본 발명에 개시된 컴포넌트, 이들의 연결과 관계, 및 기능은 단지 예시적인 것 뿐이며, 본 발명에서 설명 및/또는 요구한 본 발명의 구현을 한정하려는 것은 아니다.
- [0127] 도7에 도시한 바와 같이, 전자 기기(700)는 컴퓨팅 유닛(701)을 포함하고, 읽기 전용 메모리(ROM)(702)에 저장된 컴퓨터 프로그램 또는 저장 유닛(708)에서 랜덤 액세스 메모리(RAM)(703)에 로딩된 컴퓨터 프로그램에 따라, 각 적당한 동작 및 처리를 수행한다. RAM(703)에서, 전자 기기(700) 동작에 수요되는 각 프로그램 및 데이터를 저장할 수도 있다. 컴퓨팅 유닛(701), ROM(702) 및 RAM(703)은 버스(704)를 통해 서로 연결된다. 입력/출력 I/O 인터페이스(705)도 버스(704)에 연결된다.
- [0128] 전자 기기(700)의 복수의 컴포넌트는 I/O 인터페이스(705)에 연결되고, 복수의 컴포넌트는, 키보드, 마우스 등과 같은 입력 유닛(706); 다양한 유형의 모니터, 스피커 등과 같은 출력 유닛(707); 자기 디스크, 광 디스크 등과 같은 저장 유닛(708); 및 네트워크 카드, 모뎀 또는 무선 통신 송수신기 등과 같은 통신 유닛(709)을 포함한다. 통신 유닛(709)은 전자 기기(700)가 인터넷과 같은 컴퓨터 네트워크 및/또는 각 전신 네트워크를 통해 기타 기기와 정보/데이터를 교환할 수 있도록 허용한다.
- [0129] 컴퓨팅 유닛(701)은 각 처리 및 계산 기능을 구비한 범용/전용 처리 컴포넌트일 수 있다. 컴퓨팅 유닛(701)의 일부 예시는 중앙 처리 장치(CPU), 그래프 처리 장치(GPU), 각 전용 인공지능 계산 칩, 각 기계 학습 모델 알고리즘을 운영하는 컴퓨팅 유닛, 디지털 신호 처리 장치(DSP), 임의의 적합한 프로세서, 제어기 및 마이크로 제어기 등을 포함하나 이에 한정되지 않는다. 컴퓨팅 유닛(701)은 상기 설명한 각 방법 및 처리를 수행한다. 예를 들면 음성 처리 방법을 수행한다. 예를 들면, 일 실시예에서, 음성 처리 방법은 컴퓨터 소프트웨어 프로그램으로 구현될 수 있고, 유형적으로 저장 유닛(705)과 같은 기계 관독 가능 매체에 포함된다. 일 실시예에서, 컴퓨터 프로그램의 일부 또는 전부는 ROM(702) 및/또는 통신 유닛(709)에 의해 전자 기기(700)에 로딩 및/또는 설치될 수 있다. 컴퓨터 프로그램이 RAM(703)에 로딩되고 컴퓨팅 유닛(701)에 의해 수행될 경우, 상기 설명한 음성 처리 방법의 하나 또는 복수의 단계를 수행할 수 있다. 대안적으로, 기타 실시예에서, 컴퓨팅 유닛(701)은 기타 임의의 적합한 방식(예를 들면, 펌웨어)으로 본 발명의 실시예에 따른 음성 처리 방법을 수행할 수 있도록 구성된다.
- [0130] 여기서 설명하는 시스템과 기술의 여러 가지 실시형태는 디지털 전자회로 시스템, 집적회로 시스템, 프로그래밍 가능 게이트 어레이(FPGA), 주문형 직접 회로(ASIC), 전용 표준 제품(ASSP), 칩상 시스템(SOC), 복합 프로그래머블 논리 소자(CPLD), 컴퓨터 하드웨어, 펌웨어, 소프트웨어 및/또는 이들의 조합에서 실현될 수 있다. 이러한 여러 가지 실시형태는 하나 또는 복수의 컴퓨터 프로그램에서 실시되는 것을 포함할 수 있고, 당해 하나 또는 복수의 컴퓨터 프로그램은 적어도 하나의 프로그래밍 가능 프로세서를 포함하는 프로그래밍 가능 시스템에서 실행 및/또는 해석되며, 당해 프로그래밍 가능 프로세서는 전용 또는 일반 프로그래밍 가능 프로세서일 수

있으며, 저장 시스템, 적어도 하나의 입력 장치 및 적어도 하나의 출력 장치에서 데이터와 명령을 수신할 수 있고, 데이터와 명령을 당해 저장 시스템, 당해 적어도 하나의 입력 장치 및 당해 적어도 하나의 출력 장치에 전송할 수 있다.

[0131] 본 발명의 방법을 수행하는데 사용되는 프로그램 코드는 하나 또는 복수의 프로그래밍 언어의 임의의 조합으로 작성될 수 있다. 당해 프로그램 코드는 범용 컴퓨터, 전용 컴퓨터 또는 기타 프로그래밍 가능한 데이터 처리 장치의 프로세서 또는 제어기에 제공하여, 프로그램 코드가 프로세서 또는 제어기에 의해 수행될 경우 흐름도 및/또는 블록도에서 규정한 기능/동작을 실시하게 된다. 프로그램 코드는 완전히 또는 부분적으로 기계에서 수행되고, 독립 소프트웨어 패키지로서 부분적으로 기계에서 수행하고 부분적으로 또는 완전히 원격 기계 또는 서버에서 수행된다.

[0132] 본 발명의 콘텍스트에서, 기계 판독 가능 매체는 유형적인 매체일 수 있고, 명령 수행 시스템, 장치 또는 기기가 사용하거나 명령 수행 시스템, 장치 또는 기기와 결합하여 사용하도록 제공하는 프로그램을 포함 또는 저장할 수 있다. 기계 판독 가능 매체는 기계 판독 가능 신호 매체 또는 기계 판독 가능 저장 매체일 수 있다. 기계 판독 가능 매체는 전자, 자기, 광학, 전자기, 적외선 또는 반도체 시스템, 장치 및 기기, 또는 상기 내용의 임의의 적합한 조합을 포함하나 이에 한정되지 않는다. 기계 판독 가능 저장 매체의 더 구체적인 예시는 하나 또는 복수의 선을 기반으로 하는 전기 연결, 휴대용 컴퓨터 디스크, 하드 디스크, 랜덤 액세스 메모리(RAM), 읽기 전용 메모리(ROM), 지울 수 있는 프로그래밍 가능한 읽기 전용 메모리(EPROM 또는 플래시 메모리), 광섬유, 시디롬(CD-ROM), 광학 저장 기기, 자기 저장 기기, 또는 상기 내용의 임의의 적합한 조합을 포함할 수 있다.

[0133] 사용자와의 대화를 제공하기 위해, 여기서 설명된 시스템 및 기술은 컴퓨터에서 구현할 수 있으며, 당해 컴퓨터는 사용자에게 정보를 디스플레이하는 디스플레이 장치(예를 들면, CRT음극선관) 또는 LCD(액정 디스플레이) 모니터; 및 키보드와 지향 장치(예를 들면, 마우스 또는 트랙볼)를 구비하고, 사용자는 당해 키보드와 당해 지향 장치를 통해 컴퓨터에 입력을 제공할 수 있다. 기타 유형의 장치도 사용자와의 대화에 사용될 수 있는 바, 예를 들면 사용자에게 제공된 피드백은 임의의 형식의 감각 피드백(예를 들면, 시각적 피드백, 청각적 피드백 또는 촉각적 피드백)일 수 있고, 임의의 형식(음향 입력, 음성 입력 또는 촉각 입력)에 의해 사용자로부터의 입력을 수신할 수 있다.

[0134] 여기서 설명한 시스템과 기술을, 백그라운드 컴포넌트를 포함하는 컴퓨팅 시스템(예를 들면 데이터 서버), 또는 미들웨어 컴포넌트를 포함하는 컴퓨팅 시스템(예를 들면, 애플리케이션 서버), 또는 프론트 엔드 컴포넌트를 포함하는 컴퓨팅 시스템(예를 들면, 그래픽 사용자 인터페이스 또는 네트워크 브라우저를 구비한 사용자 컴퓨터에서 실시될 수 있고, 사용자는 당해 그래픽 사용자 인터페이스 또는 당해 네트워크 브라우저를 통해 여기서 설명한 시스템과 기술의 실시형태와 대화할 수 있다), 또는 이러한 백그라운드 컴포넌트, 미들웨어 컴포넌트 또는 프론트 엔드 컴포넌트의 임의의 조합을 포함하는 컴퓨팅 시스템에서 실시될 수 있다. 임의의 형태 또는 매체의 디지털 데이터 통신(예를 들면, 통신 네트워크)을 통해 시스템의 컴포넌트를 서로 연결할 수 있다. 통신 네트워크의 예시는 근거리 통신망(LAN), 광역 통신망(WAN), 인터넷 및 블록 체인 네트워크를 포함한다.

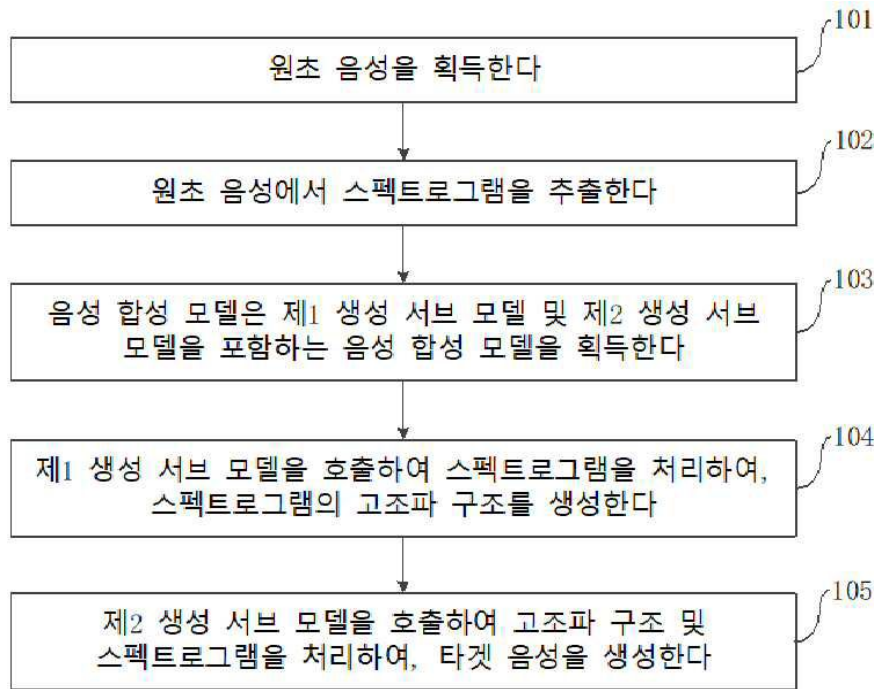
[0135] 컴퓨터 시스템은 클라이언트와 서버를 포함할 수 있다. 클라이언트와 서버는 일반적으로 서로 떨어져 있으며, 통신 네트워크를 통해 서로 대화한다. 대응하는 컴퓨터에서 운행되고 서로 클라이언트-서버 관계를 가지는 컴퓨터 프로그램에 의해 클라이언트와 서버의 관계를 생성한다. 서버는 클라우드 서버일 수 있고, 클라우드 계산 서버라고도 하고, 클라우드 계산 서비스 시스템의 호스트 제품이고, 전통적인 물리 호스트와 VPS서비스("Virtual Private Server", 또는 "VPS")에서, 관리가 어렵고, 업무 확장성이 약한 결함을 해결한다. 서버는 분산식 시스템의 서버 또는 블록 체인을 결합한 서버일 수도 있다.

[0136] 이해해야 할 것은, 상기 복수 형식의 흐름에 의해, 단계를 재정렬, 추가 또는 삭제할 수 있다. 예를 들면, 본 발명에 기재한 각 단계는 병행하여 또는 순차적으로 실행할 수도 있고, 서로 다른 순서로 실행할 수도 있다. 본 발명에서 개시한 기술적 수단이 원하는 결과만 구현할 수 있으면 본 발명에서는 이에 대해 한정하지 않는다.

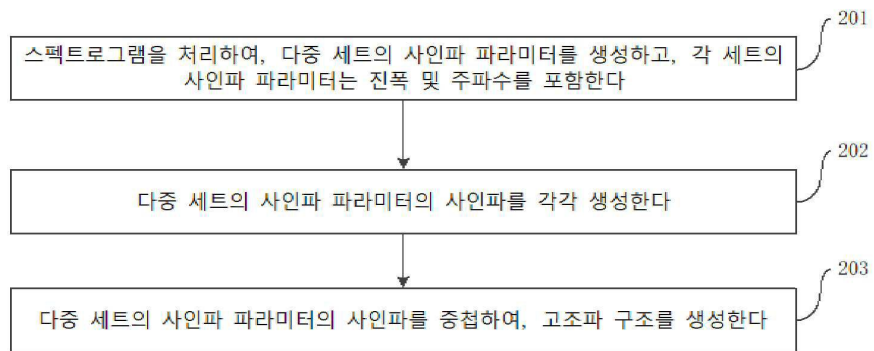
[0137] 상기 구체적인 실시 방식은 본 발명의 보호 범위를 한정하지 않는다. 본 발명이 속하는 기술 분야의 통상의 기술자는 설계 요구 및 기타 요소에 의해 여러가지 수정, 조합, 서브 조합 및 대체가 이루어질 수 있음을 이해해야 한다. 본 발명의 정신과 원칙 내에서 이루어진 모든 수정, 동등한 대체 및 개선은 모두 본 발명 보호 범위에 포함된다.

도면

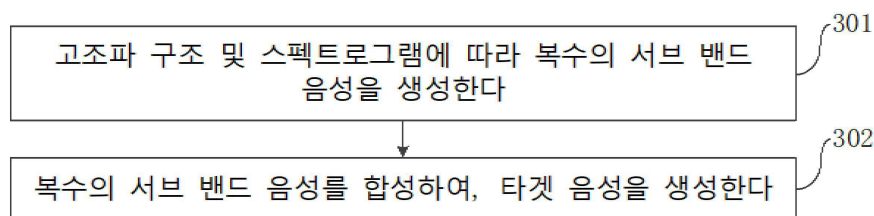
도면1



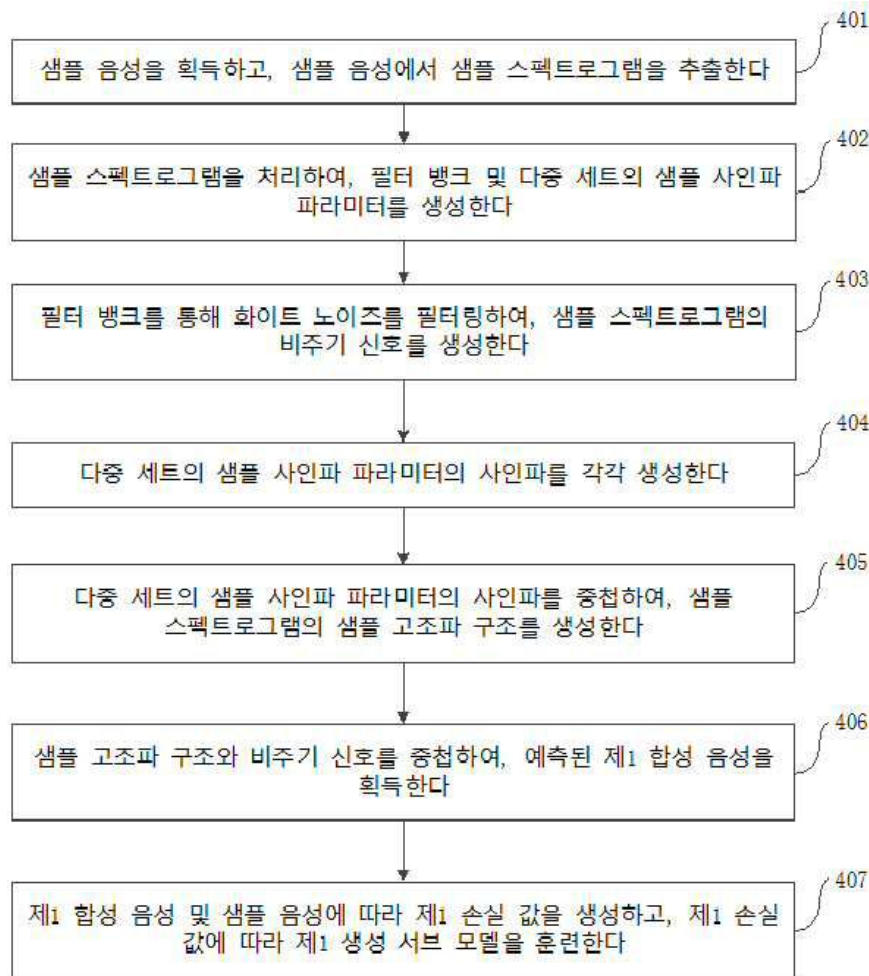
도면2



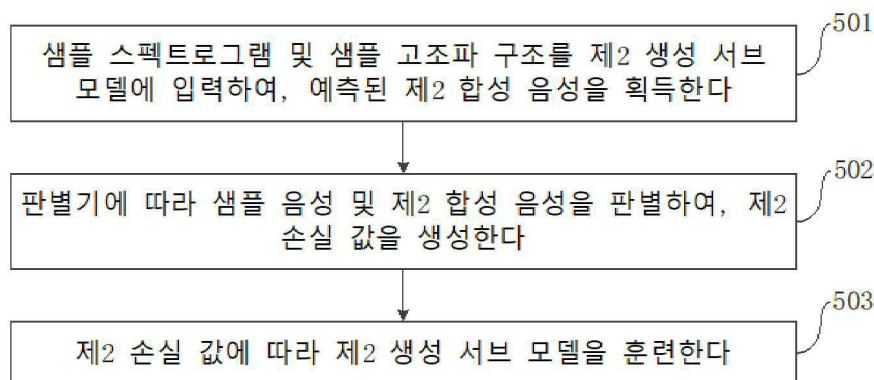
도면3



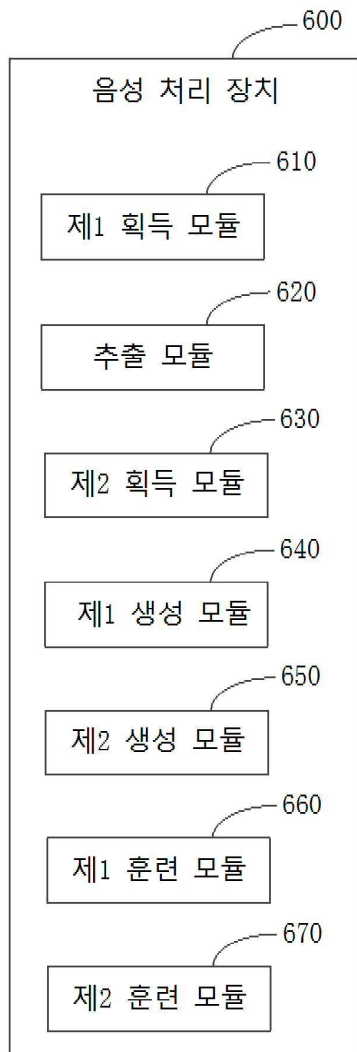
도면4



도면5



도면6



도면7

