



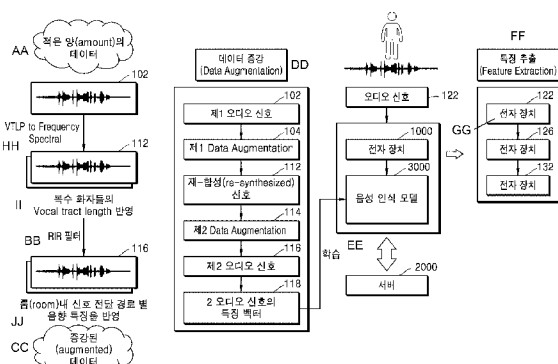
- (51) 국제특허분류:
G10L 25/75 (2013.01) G10L 15/06 (2006.01)
- (21) 국제출원번호: PCT/KR2019/016181
- (22) 국제출원일: 2019년 11월 22일 (22.11.2019)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
62/772,382 2018년 11월 28일 (28.11.2018) US
10-2019-0036376 2019년 3월 28일 (28.03.2019) KR
62/848,698 2019년 5월 16일 (16.05.2019) US
10-2019-0150494 2019년 11월 21일 (21.11.2019) KR
- (71) 출원인: 삼성전자 주식회사 (SAMSUNG ELECTRONICS CO., LTD.) [KR/KR]; 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).
- (72) 발명자: 김찬우 (KIM, Chanwoo); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 고다난자야 엔. (GOWDA, Dhananjaya N.); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 김성수 (KIM, Sung-soo); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 신민규 (SHIN, Minkyu); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 핵레리 폴

(HECK, Larry Paul); 94043 캘리포니아주 마운틴 뷰 클라이드 애비뉴 665, California (US). 가르그아비나브 (GARG, Abhinav); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 김광윤 (KIM, Kwangyun); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR). 쿠마르메홀 (KUMAR, Mehul); 16677 경기도 수원시 영통구 삼성로 129, Gyeonggi-do (KR).

- (74) 대리인: 리앤록 특허법인 (Y.P.LEE, MOCK & PARTNERS); 06292 서울시 강남구 언주로30길 13 대림아크로텔 12층, Seoul (KR).
- (81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE,

(54) Title: VOICE RECOGNITION DEVICE AND METHOD

(54) 발명의 명칭: 음성 인식 장치 및 방법



102 ... First audio signal
104 ... First Data Augmentation
112 ... Re-synthesized signal
114 ... Second Data Augmentation
116 ... Second audio signal
118 ... Feature vector of second audio signal
122 ... Audio signal
126, 132, 1000, GG ... Electronic device
2000 ... Server
3000 ... Voice recognition model
AA ... Small amount of data
BB ... RIR filter
CC ... Augmented data
DD ... Data Augmentation
EE ... Learning
FF ... Feature Extraction
HH ... VTLF to Frequency Spectral
JJ ... Reflect vocal track lengths of multiple speakers
JJ ... Reflect sound feature for each signal transfer path in room

(57) Abstract: The present disclosure relates to an electronic device for recognizing a user voice and a method for recognizing a user voice by the electronic device. According to an embodiment, a method for recognizing a user voice may comprise the steps of: acquiring an audio signal divided into a plurality of frame units; by applying a filter bank distributed according to a preconfigured scale to a frequency spectrum of the audio signal divided in units of the frames, determining an energy component for each filter bank; smoothing the determined energy component for each filter bank; extracting a feature vector of the audio signal on the basis of the smoothed energy component for each filter bank; and recognizing the user voice in the audio signal by inputting the extracted feature vector into a voice recognition model.

(57) 요약서: 본 개시는 사용자 음성을 인식하는 전자 장치 및 상기 전자 장치가 사용자 음성을 인식하는 방법에 관한 것이다. 일 실시 예에 의하면, 사용자의 음성을 인식하는 방법은 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하는 단계; 상기 결정된 필터 बैं크 별 에너지 성분을 평탄화(smoothing) 하는 단계; 상기 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하는 단계; 및 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 단계;를 포함할 수 있다.

LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 음성 인식 장치 및 방법

기술분야

- [1] 본 개시는 음성 인식 장치 및 방법에 관한 것이다. 보다 상세하게는, 음성 인식을 위한 오디오 데이터를 처리하는 방법에 관한 것이다.

배경기술

- [2] 인공지능(Artificial Intelligence, AI) 시스템은 기존 Rule 기반 스마트 시스템과 달리 기계가 스스로 학습하고 판단하며 똑똑해지는 시스템이다. 인공지능 시스템은 사용할수록 인식률이 향상되고 사용자 취향을 보다 정확하게 이해할 수 있게 되어, 기존 Rule 기반 스마트 시스템은 점차 딥러닝 기반 인공지능 시스템으로 대체되고 있다.
- [3] 인공지능 기술은 기계학습(딥러닝) 및 기계학습을 활용한 요소 기술들로 구성된다. 기계학습은 입력 데이터들의 특징을 스스로 분류/학습하는 알고리즘 기술이며, 요소기술은 딥러닝 등의 기계학습 알고리즘을 활용하는 기술로서, 언어적 이해, 시각적 이해, 추론/예측, 지식 표현, 동작 제어 등의 기술 분야로 구성된다.
- [4] 인공지능 기술이 응용되는 다양한 분야는 다음과 같다. 언어적 이해는 인간의 언어/문자를 인식하고 응용/처리하는 기술로서, 자연어 처리, 기계 번역, 대화시스템, 질의 응답, 음성 인식/합성 등을 포함한다. 시각적 이해는 사물을 인간의 시각처럼 인식하여 처리하는 기술로서, 객체 인식, 객체 추적, 영상 검색, 사람 인식, 장면 이해, 공간 이해, 영상 개선 등을 포함한다. 추론 예측은 정보를 판단하여 논리적으로 추론하고 예측하는 기술로서, 지식/확률 기반 추론, 최적화 예측, 선호 기반 계획, 추천 등을 포함한다. 지식 표현은 인간의 경험정보를 지식데이터로 자동화 처리하는 기술로서, 지식 구축(데이터 생성/분류), 지식 관리(데이터 활용) 등을 포함한다. 동작 제어는 차량의 자율 주행, 로봇의 움직임 제어를 하는 기술로서, 움직임 제어(항법, 충돌, 주행), 조작 제어(행동 제어) 등을 포함한다.
- [5] 최근 음성 인식 분야에서 사용자의 음성을 인식하기 위한 다양한 기술들이 연구되고 있고, 종단 간(End to End) 학습 방식을 사용하는 인공 신경망 기반의 음성 인식 방법이 활발히 연구되고 있다. 특히, 음성 인식 분야에서 음성 인식의 정확도(accuracy)를 향상시키기 위한 기술 개발이 요구되고 있다.

발명의 상세한 설명

기술적 과제

- [6] 일 실시 예에 따르면, 음성 인식 모델을 이용하여 사용자의 음성을 인식하는 음성 인식 장치 및 방법이 제공될 수 있다.
- [7] 또한, 일 실시 예에 따르면, 오디오 신호에 기초하여 음성 인식 모델을 학습하기

위한 장치 및 방법이 제공될 수 있다. 보다 상세하게는, 음성 인식 정확도를 향상하기 위한 오디오 데이터 처리 방법이 제공될 수 있다.

과제 해결 수단

- [8] 일 실시 예에 의하면, 전자 장치가 사용자의 음성을 인식하는 방법에 있어서 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하는 단계; 상기 결정된 필터 बैं크 별 에너지 성분을 평탄화(smoothing) 하는 단계; 상기 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하는 단계; 및 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 단계; 를 포함하는 방법이 제공될 수 있다.
- [9] 또한, 일 실시 예에 의하면, 사용자 음성을 인식하는 전자 장치에 있어서, 하나 이상의 인스트럭션을 저장하는 메모리; 및 상기 하나 이상의 인스트럭션을 실행하는 프로세서; 를 포함하고, 상기 프로세서는 상기 하나 이상의 인스트럭션을 실행함으로써, 복수의 프레임 단위로 구분되는 오디오 신호를 획득하고, 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하고, 상기 결정된 필터 बैं크 별 에너지 성분을 평탄화(smoothing) 하고, 상기 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하고, 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 전자 장치가 제공될 수 있다.
- [10] 또한, 일 실시 예에 의하면, 전자 장치가 음성 인식 모델을 학습하기 위한 오디오 신호 처리 방법에 있어서, 상기 음성 인식 모델을 학습하기 위해 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 제1 오디오 신호의 주파수 스펙트럼을 획득하는 단계; 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 단계; 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성하는 단계; 및 상기 생성된 제2 오디오 신호의 특징 벡터를 추출하는 단계; 를 포함하는, 방법이 제공될 수 있다.
- [11] 상기 방법은 소정의 룸(Room) 내에서 상기 제2 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room Impulse) 필터를 상기 제2 오디오 신호에 적용하는 단계; 및 상기 룸 임펄스 필터가 적용된 상기 제2 오디오 신호로부터 특징 벡터를 추출하는 단계; 를 더 포함할 수 있다.
- [12] 상기 주파수 축을 변환하는 단계는 상기 프레임 별로 랜덤하게 생성되는

워핑(warping) 계수를 결정하는 단계; 상기 결정된 워핑 계수에 기초하여, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하기 위한 워핑 함수를 결정하는 단계; 상기 결정된 워핑 함수를 이용하여 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 단계; 를 포함할 수 있다.

- [13] 상기 제2 오디오 신호를 생성하는 단계는 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환(Inverse Fast Fourier Transform)하는 단계; 상기 역 고속 푸리에 변환된 제1 오디오 신호의 주파수 스펙트럼을 시간 축 상에서 중첩(overlap)함으로써 상기 제2 오디오 신호를 생성하는 단계; 를 포함할 수 있다.

- [14] 상기 제2 오디오 신호의 특징 벡터를 추출하는 단계는 상기 프레임 단위로 구분되는 상기 제2 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 결정하는 단계; 상기 결정된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하는 단계; 상기 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하는 단계; 를 포함할 수 있다.

- [15] 또한, 일 실시 예에 의하면, 전자 장치가 음성 인식 모델을 학습하는 방법에 있어서, 상기 음성 인식 모델을 학습하기 위해 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 제1 오디오 신호의 주파수 스펙트럼을 획득하는 단계; 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 단계; 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성하는 단계; 상기 생성된 제2 오디오 신호의 특징 벡터를 추출하는 단계; 및 상기 추출된 특징 벡터를 이용하여 상기 음성 인식 모델을 학습하는 단계; 를 포함하는, 방법이 제공될 수 있다.

- [16] 또한, 일 실시 예에 의하면, 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 동작; 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 결정하는 동작; 상기 결정된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하는 동작; 상기 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하는 동작; 및 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 동작; 을 포함하는 사용자의 음성을 인식하는 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체가 제공될 수 있다.

발명의 효과

- [17] 본 개시에 따른 음성 인식 장치 및 방법에 따르면, 적은 양의 데이터를 이용하여

효율적으로 사용자의 음성을 인식할 수 있고, 오디오 신호의 평탄화된 필터뱅크별 에너지 성분에 기초하여 사용자 음성을 정확하게 인식할 수 있다.

도면의 간단한 설명

- [18] 도 1은 일 실시 예에 따른, 전자 장치가 사용자의 음성을 인식하는 방법 및 음성 인식을 위한 오디오 신호 처리 방법을 개략적으로 나타내는 도면이다.
- [19] 도 2는 일 실시 예에 따라 전자 장치가 사용자의 음성을 인식하는 방법을 나타내는 흐름도이다.
- [20] 도 3은 일 실시 예에 따라 전자 장치가 필터뱅크별 에너지 성분을 결정하는 방법을 구체적으로 나타내는 흐름도이다.
- [21] 도 4는 일 실시 예에 따라 전자 장치가 미리 학습되는 평탄화 계수를 이용하여 필터뱅크별 에너지 성분을 평탄화(smoothing)하는 방법을 나타내는 흐름도이다.
- [22] 도 5는 필터뱅크 채널 별로 학습되는 평탄화 계수를 설명하기 위한 도면이다.
- [23] 도 6은 일 실시 예에 따라 전자 장치가 타겟 히스토그램을 매핑하기 위한 매핑 함수를 이용하여 필터뱅크별 에너지 성분을 평탄화(smoothing)하는 방법을 나타내는 흐름도이다.
- [24] 도 7은 일 실시 예에 따른 특정 채널 인덱스에 해당하는 필터뱅크의 에너지 성분을 나타내는 도면이다.
- [25] 도 8은 일 실시 예에 따른 필터뱅크 에너지에 대한 비선형 출력(nonlinearity output)을 설명하기 위한 도면이다.
- [26] 도 9는 일 실시 예에 따른 전자 장치가 음성 인식 모델 학습을 위한 오디오 신호 처리 방법을 나타내는 흐름도이다.
- [27] 도 10은 일 실시 예에 따른 전자 장치가 음성 인식 모델 학습을 위한 오디오 신호를 처리 과정을 개략적으로 설명하기 위한 도면이다.
- [28] 도 11은 일 실시 예에 따른 전자 장치가 주파수 축을 변환하는 방법을 구체적으로 나타내는 흐름도이다.
- [29] 도 12는 워핑 계수에 따라 입력 주파수가 워핑(warped)되는 과정을 설명하기 위한 도면이다.
- [30] 도 13은 일 실시 예에 따라 전자 장치가 주파수 축이 변환된 제1 오디오 신호를 이용하여 제2 오디오 신호를 재합성하는 방법을 나타내는 도면이다.
- [31] 도 14는 일 실시 예에 따른 전자 장치의 블록도이다.
- [32] 도 15는 또 다른 실시 예에 따른 전자 장치의 블록도이다.
- [33] 도 16은 일 실시 예에 따른 음성 인식 모델을 학습시키는 전자 장치의 구조를 설명하기 위한 도면이다.
- [34] 도 17은 일 실시 예에 따른 전자 장치가 음성 인식 모델을 학습하는 과정을 나타내는 도면이다.
- [35] 도 18은 일 실시 예에 따른 전자 장치가 필터뱅크별 에너지 성분이 평탄화된

특징 벡터를 음성 인식 모델에 입력함으로써 음성 인식을 수행하는 과정을 나타내는 도면이다.

[36] 도 19는 일 실시 예에 따른 서버의 블록도를 나타내는 도면이다.

[37] 도 20은 일 실시 예에 따라, 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.

[38] 도 21은 일 실시 예에 따라, 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.

[39] 도 22은 일 실시 예에 따라 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.

[40] 도 23은 일 실시 예에 따라 전자 장치가 목표 음성 신호 및 잡음 신호를 모두 이용하여 음성 인식 모델을 학습시키는 과정을 설명하기 위한 도면이다.

발명의 실시를 위한 최선의 형태

[41] 일 실시 예에 의하면, 전자 장치가 사용자의 음성을 인식하는 방법에 있어서 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 결정하는 단계; 상기 결정된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하는 단계; 상기 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하는 단계; 및 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 단계; 를 포함하는 방법이 제공될 수 있다.

[42] 또한, 일 실시 예에 의하면, 사용자 음성을 인식하는 전자 장치에 있어서, 하나 이상의 인스트럭션을 저장하는 메모리; 및 상기 하나 이상의 인스트럭션을 실행하는 프로세서; 를 포함하고, 상기 프로세서는 상기 하나 이상의 인스트럭션을 실행함으로써, 복수의 프레임 단위로 구분되는 오디오 신호를 획득하고, 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 결정하고, 상기 결정된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하고, 상기 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하고, 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식하는 전자 장치가 제공될 수 있다.

[43] 또한, 일 실시 예에 의하면, 전자 장치가 음성 인식 모델을 학습하기 위한 오디오 신호 처리 방법에 있어서, 상기 음성 인식 모델을 학습하기 위해 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 제1 오디오 신호의 주파수 스펙트럼을 획득하는 단계; 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 상기 제1

오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 단계; 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성하는 단계; 및 상기 생성된 제2 오디오 신호의 특징 벡터를 추출하는 단계; 를 포함하는, 방법이 제공될 수 있다.

- [44] 또한, 일 실시 예에 의하면, 전자 장치가 음성 인식 모델을 학습하는 방법에 있어서, 상기 음성 인식 모델을 학습하기 위해 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득하는 단계; 상기 프레임 단위로 구분되는 상기 제1 오디오 신호의 주파수 스펙트럼을 획득하는 단계; 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 단계; 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성하는 단계; 상기 생성된 제2 오디오 신호의 특징 벡터를 추출하는 단계; 및 상기 추출된 특징 벡터를 이용하여 상기 음성 인식 모델을 학습하는 단계; 를 포함하는, 방법이 제공될 수 있다.

발명의 실시를 위한 형태

- [45] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 개시에 대해 구체적으로 설명하기로 한다.
- [46] 본 개시에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.
- [47] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.
- [48] 아래에서는 첨부한 도면을 참고하여 본 개시의 실시 예에 대하여 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 개시는 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 개시를 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [49] 도 1은 일 실시 예에 따른, 전자 장치가 사용자의 음성을 인식하는 방법 및 음성

인식을 위한 오디오 신호 처리 방법을 개략적으로 나타내는 도면이다.

- [50] 일 실시 예에 의하면, 전자 장치(1000)는 음성 인식 모델(3000)을 포함할 수 있다. 예를 들어, 전자 장치(1000)는 오디오 신호(122)를 획득하고, 획득된 오디오 신호에서 추출된 특징 벡터(132)를 음성 인식 모델(3000)에 입력함으로써 오디오 신호 내 사용자의 음성을 인식할 수 있다. 전자 장치(1000)가 사용자의 음성을 인식하기 위해 이용하는 음성 인식 모델은 자동 음성 인식(Auto Speech Recognition)을 위한 신경망(Neural Network) 기반의 모델일 수 있다.
- [51] 일 실시 예에 따른 전자 장치(1000)는 AI 프로그램이 탑재되고 음성 인식 기능을 포함하는 스마트폰, 태블릿 PC, PC, 스마트 TV, 휴대폰, PDA(personal digital assistant), 랩톱, 미디어 플레이어, 서버, 마이크로 서버, GPS(global positioning system) 장치, 전자책 단말기, 디지털방송용 단말기, 네비게이션, 키오스크, MP3 플레이어, 디지털 카메라, 스피커 기타 모바일 또는 비모바일 컴퓨팅 장치일 수 있으나, 이에 제한되지 않는다.
- [52] 일 실시 예에 의하면, 전자 장치(1000)는 전자 장치와 네트워크를 통하여 연결되는 서버(2000)와 연동함으로써 오디오 신호 내 사용자의 음성을 인식할 수 있다. 일 실시 예에 의하면, 네트워크는 근거리 통신망(Local Area Network; LAN), 광역 통신망(Wide Area Network; WAN), 부가가치 통신망(Value Added Network; VAN), 이동 통신망(mobile radio communication network), 위성 통신망 및 이들의 상호 조합을 포함할 수 있다. 전자 장치(1000)와 네트워크를 통하여 연결되는 서버(2000)는 본원 전자 장치(1000)와 통신 가능한 적어도 하나의 다른 전자 장치를 포함할 수 있다.
- [53] 일 실시 예에 의하면, 전자 장치(1000)는 음성 인식 모델(3000)을 미리 학습시킬 수 있다. 예를 들어, 전자 장치(1000)는 음성 인식 모델을 학습하기 위한 오디오 신호를 적어도 하나의 데이터 증강 기법에 따라 증강(augmentation)시키고, 증강된(augmented) 오디오 신호에 기초하여 음성 인식 모델을 학습할 수 있다. 또한, 전자 장치(1000)는 사용자의 음성을 포함하는 오디오 신호의 주파수 스펙트럼의 필터 बैं크 별 에너지 성분을 평탄화(smoothing)하고, 평탄화된 필터 बैं크 별 에너지 성분을 포함하는 오디오 신호의 주파수 스펙트럼으로부터 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 사용자의 음성을 정확하게 인식할 수 있다.
- [54] 이하에서는 전자 장치(1000)가 증강된 오디오 신호에 기초하여 음성 인식 모델(3000)을 학습하는 방법 및 상기 학습된 음성 인식 모델에, 평탄화된(smoothed) 필터 बैं크 별 에너지 성분을 포함하는 오디오 신호의 주파수 스펙트럼으로부터 추출된 특징 벡터를 입력함으로써 사용자의 음성을 인식하는 방법을 간략하게 설명하기로 한다.
- [55] 예를 들어, 전자 장치(1000)는 사용자로부터 제1 오디오 신호(102)를 획득하고, 제1 오디오 신호를 분할(segment)하며, 분할된 제1 오디오 신호를 소정의 간격으로 중첩(overlap)함으로써, 인접한 프레임과 소정의 간격으로 중첩되는

- 프레임 단위의 제1 오디오 신호를 획득할 수 있다. 제1 오디오 신호는 연속된 신호가 미리 설정된 샘플링 주파수로 샘플링된 이산(discrete) 신호일 수 있다.
- [56] 전자 장치(1000)는 제1 Data Augmentation(104) 기법을 이용하여 제1 오디오 신호(102)를 증강시킬 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 오디오 신호의 특징 벡터가 아닌 오디오 신호의 주파수 스펙트럼(spectral) 자체에 VTLP(Vocal Tract length perturbation)를 적용함으로써 오디오 신호를 증강(augmentation)시킬 수 있다. 즉, 전자 장치(1000)는 제1 오디오 신호의 주파수 스펙트럼에 VTLP를 적용함으로써, 적은 양의 오디오 데이터가 획득되는 경우에도, 적은 양의 오디오 데이터를 기초로 증강된(augmented) 다양한 성도 길이를 가지는 화자들의 오디오 데이터를 획득할 수 있다.
- [57] 보다 상세하게는, 전자 장치(1000)는 프레임 단위로 구분되는 제1 오디오 신호를 고속 푸리에 변환(Fast Fourier Transform, FFT)함으로써 주파수 스펙트럼을 획득하고, 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영(representing)하기 위해 워핑 계수(warping coefficient)에 기초하여 정의되는 워핑 함수(warping function)를 이용하여 주파수 스펙트럼의 주파수 축을 변환(frequency warping)한다.
- [58] 일 실시 예에 의하면, 전자 장치(1000)는 제1 오디오 신호를 고속 푸리에 변환 후 FFT 계수들 사이에 다양한 보간(interpolation) 알고리즘을 적용함으로써 주파수 스펙트럼을 획득할 수 있다. 그러나, 또 다른 실시 예에 의하면, 전자 장치(1000)는 업 샘플링 하고자 하는 샘플의 수에 따라 미리 설정되는 FFT 사이즈계수에 기초하여 제1 오디오 신호를 오버사이즈 고속 푸리에 변환(Fast Fourier Transform, FFT, 106)함으로써 주파수 스펙트럼을 획득할 수도 있다. 주파수 스펙트럼은 고속 푸리에 변환된 제1 오디오 신호의 주파수 성분 별 복소수의 크기를 나타낼 수 있다.
- [59] 전자 장치(1000)는 주파수 축이 변환된 제1 오디오 신호로부터 획득된 주파수 스펙트럼을 역 고속 푸리에 변환함으로써 시간 영역의 신호를 획득하고, 획득된 시간 영역의 신호들을 연결함으로써 재-합성된(re-synthesized) 신호(112)를 획득할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환(Inverse Fast Fourier Transform)하고, 역 고속 푸리에 변환된 제1 오디오 신호의 주파수 스펙트럼을 시간 축에서 중첩하여 더함(Overlap Addition, 112)으로써, 재-합성된(re-synthesized) 신호(112)를 획득할 수도 있다. 그러나, 전자 장치(1000)가 역 고속 푸리에 변환된 시간 영역의 신호들을 연결하기 위한 방법은 전술한 바에 한정되는 것은 아니고, 시간 영역에서 신호들을 연결하기 위한 기타 다양한 방법을 사용할 수 있다.
- [60] 또 다른 실시 예에 의하면, 전자 장치(1000)는 제2 Data Augmentation(114) 기법을 이용하여 재합성(re-synthesized)된 신호(112)를 증강시킬 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 제2 오디오 신호(116)에 상기 제2 오디오 신호가

전달되는 룸(Room)내에서 상기 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room impulse) 필터(예컨대, 룸 임펄스 응답으로써 전달 경로 별 음향 특징을 나타내는 전달함수)를 적용하고, 룸 임펄스 필터가 적용된 제2 오디오 신호의 주파수 스펙트럼에서 추출된 특징 벡터(118)에 기초하여 음성 인식 모델을 미리 학습시킬 수도 있다. 즉, 전자 장치(1000)는 제1 오디오 신호의 주파수 스펙트럼에 VTLP를 적용함으로써 복수 화자들의 성도 길이 변화(variation)가 반영된 증강된 데이터를 획득하고, 증강된 데이터에 RIR 필터를 적용함으로써, 다양한 환경의 룸(room) 내 오디오 신호의 전달 경로 별 특징이 반영된 더 증강된 데이터들을 획득할 수 있다.

[61] 따라서, 전자 장치(1000)는 적은 양의 오디오 데이터를 획득하더라도, 적은 양의 데이터들이 증강된 데이터에 기초하여 음성 인식 모델을 학습시킴으로써 사용자 음성의 인식 정확도를 향상시킬 수 있다.

[62] 종래, 일반적인 전자 장치들은 주파수 스펙트럼에 VTLP를 적용하는 것이 아니고, 위평 함수에 기초하여 멜 필터 बैं크의 위치를 조절하고, 위치가 조절된 멜 필터 बैं크 별 에너지 성분에 기초하여 Mel 특징(feature)을 추출하였다. 따라서, 일반적인 전자 장치들은 멜 필터 बैं크 별 에너지 변환의 단 방향성으로 인하여, Mel 특징 추출 후, RIR 필터를 적용하는데 어려움이 있었다. 그러나, 일 실시 예에 의하면, 전자 장치(1000)는 오디오 신호의 특징 추출 단계 이전의 주파수 스펙트럼에 VTLP를 적용하기 때문에, VTLP 적용 후 재-합성된 오디오 신호에 RIR 필터를 더 적용할 수 있고, 그 결과 오디오 데이터를 효과적으로 증강시킬 수 있다.

[63] 또한, 본 개시에 따른 전자 장치(1000)가 오디오 신호를 증강(augmentation)함에 있어, 서로 다른 화자의 성도 길이 변화를 반영하기 위해, 오디오 신호의 주파수 스펙트럼상에 VTLP를 적용하고, 재-합성된 오디오 신호에 RIR 필터를 적용하는 순서는 실제 룸(room) 내에서 서로 다른 성도 길이를 가지는 화자의 음성 신호가 마이크까지 수신되는 일련의 오디오 신호의 전달 순서에 더 부합한다. 따라서, 본 개시에 따른 전자 장치(1000)는 제1 오디오 신호(102)에, 실제 음성 인식에 필요한 오디오 신호의 전달 순서에 따라 VTLP 및 RIR 기법을 모두 적용함으로써 특징 벡터(118)를 획득하고, 획득된 특징 벡터에 기초하여 음성 인식 모델을 학습시킴으로써 음성 인식 모델의 음성 인식 정확도를 더 향상시킬 수 있다.

[64] 또한, 도 1에는 도시되지 않았지만, 전자 장치(1000)는 제1 오디오 신호를 증강시킴으로써 획득된 제2 오디오 신호(116)의 주파수 스펙트럼으로부터 필터 बैं크 별 에너지 성분을 획득하고, 획득된 필터 बैं크 별 에너지 성분을 평탄화(smoothing)한 후, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 추출된 특징 벡터에 기초하여 음성 인식 모델(3000)을 학습시킬 수 있다. 전자 장치(1000)는 일반적으로 특정 주파수 영역에 집중 분포되는 멜 필터 बैं크 에너지 성분을 평탄화(smoothing) 하고, 균일하게 분포된 에너지 성분을

포함하는 주파수 스펙트럼으로부터 추출된 특징 벡터에 기초하여 음성 인식 모델(3000)을 학습 시킴으로써 음성 인식 모델의 음성 인식 정확도를 향상시킬 수 있다.

- [65] 전자 장치(1000)가 제2 오디오 신호(116)의 필터 뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 음성 인식 모델(3000)을 학습하는 경우, 후술하는 바와 같이, 새로 획득된 오디오 신호(122)의 멜 필터 뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터(132)를 음성 인식 모델에 입력시킴으로써 오디오 신호(122) 내 사용자의 음성을 인식한다.
- [66] 전자 장치(1000)는 오디오 신호를 획득(122)하고, 미리 학습된 음성 인식 모델(3000)을 이용하여 획득된 오디오 신호 내 사용자의 음성을 인식할 수 있다. 예를 들어, 전자 장치는 사용자의 음성을 포함하는 오디오 신호(122)를 획득하고, 획득된 오디오 신호를 프레임 단위로 분할함으로써, 프레임 단위로 구분되는 오디오 신호(122)를 획득할 수 있다.
- [67] 또한, 전자 장치(1000)는 프레임 단위로 구분되는 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된(distributed) 멜(mel) 필터 뱅크(126)를 적용함으로써 멜 특징(mel feature, 126)을 추출하고, 추출된 멜 특징을 포함하는 오디오 특징 벡터를 음성 인식 모델에 입력함으로써 오디오 신호 내 사용자의 음성을 인식할 수 있다.
- [68] 또한 도 1에는 도시되지 않았지만, 전자 장치(1000)가 제2 오디오 신호(116)의 필터 뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 학습된 음성 인식 모델(3000)을 이용하는 경우, 전자 장치(1000)는 새로 획득된 오디오 신호(122)의 멜 필터 뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터(132)를 음성 인식 모델에 입력시킴으로써 오디오 신호(122) 내 사용자의 음성을 인식할 수도 있다.
- [69] 도 2는 일 실시 예에 따라 전자 장치가 사용자의 음성을 인식하는 방법을 나타내는 흐름도이다.
- [70] S210에서, 전자 장치(1000)는 복수의 프레임 단위로 구분되는 오디오 신호를 획득할 수 있다. 예를 들어, 전자 장치(1000)는 소정의 윈도우의 윈도우 길이(window length)를 결정하고, 결정된 윈도우 길이의 윈도우들을 이용하여 오디오 신호를 복수의 프레임 단위로 구분할 수 있다. 또한, 전자 장치(1000)는 상기 윈도우들을 소정의 윈도우 간격으로 중첩(overlap)하고, 소정의 간격으로 중첩된 윈도우들을 이용하여 오디오 신호를 복수의 프레임 단위로 구분할 수도 있다.
- [71] 일 실시 예에 의하면 전자 장치가 이용하는 윈도우는 대역 통과 필터(Band Pass filter)로써, 프레임 내 양 끝단(side edge)의 주파수 성분에 포함된 노이즈(noise)를 제거하는데 사용될 수 있다. 일 실시 예에 의하면 윈도우는 Hanning Window 또는

Hamming Window 를 포함할 수 있으나 이에 한정되는 것은 아니다. 또한, 윈도우 길이(window length)는 30ms~50ms일 수 있으나 이에 한정되는 것은 아니고, 목표 주파수 스펙트럼의 해상도(resolution)에 따라 달라질 수 있다.

- [72] 또한, 일 실시 예에 의하면, 전자 장치(1000)가 제1 오디오 신호의 주파수 스펙트럼에 VTLP를 적용 시 이용하는 윈도우의 길이, RIR 필터를 적용할 때 이용하는 윈도우의 길이, 음성 인식 모델의 학습 후, 실제 획득된 오디오 신호 내 사용자의 음성을 인식할 때 이용하는 윈도우의 길이는 서로 다를 수 있다. 또 다른 실시 예에 의하면, 전자 장치(1000)가 제1 오디오 신호의 주파수 스펙트럼에 VTLP를 적용 시 이용하는 윈도우의 길이는 RIR 필터를 적용할 때 이용하는 윈도우의 길이 및 음성 인식 모델의 학습 후, 실제 획득된 오디오 신호 내 사용자의 음성을 인식할 때 이용하는 윈도우의 길이와 다를 수도 있다.
- [73] S220에서, 전자 장치(1000)는 기 설정된 스케일에 따라 분포된 필터 뱅크를 오디오 신호의 주파수 스펙트럼에 적용함으로써 필터 뱅크 별 에너지 성분을 결정할 수 있다. 전자 장치(1000)가 필터 뱅크를 배치하기 위해 이용하는 기 설정된 스케일은 로그 스케일(log-scale) 또는 멜 스케일(mel-scale)을 포함할 수 있으나, 이에 한정되는 것은 아니고, 저주파를 더 잘 구분하는 사람의 청각 특성에 기초하여 상기 필터 뱅크를 비선형적으로 분포(distribute)하기 위한 기타 비선형적(non-linearity) 스케일을 더 포함할 수 있다.
- [74] 일 실시 예에 의하면, 기 설정된 스케일에 따라 분포되는 필터 뱅크는 오디오 신호의 주파수 스펙트럼을 복수의 부-대역(sub-band) 스펙트럼으로 나누기 위한 복수의 대역 통과필터 열(array)을 의미할 수 있다. 또한, 필터 뱅크 별 에너지 성분은 상기 오디오 신호의 주파수 스펙트럼에 필터 뱅크를 적용함으로써 획득된 에너지 또는 에너지의 확률 밀도 함수(Probability Density Function, PDF)를 의미할 수 있다.
- [75] S230에서, 전자 장치(1000)는 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 생성된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 할 수 있다. 전자 장치(1000)는 균일하게 분포된(uniformly distributed) 에너지 성분을 포함하는 특징 벡터를 음성 인식 모델에 입력함으로써 사용자의 음성을 더 정확하게 인식할 수 있다. 전자 장치(1000)가 필터 뱅크 별 에너지 성분을 평탄화 하는 구체적인 방법은 후술하는 도 4 및 도 5를 참조하여 구체적으로 설명하기로 한다.
- [76] S240에서, 전자 장치(1000)는 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 오디오 신호의 특징 벡터를 추출할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 평탄화된(smoothed) 필터 뱅크 별 에너지 성분을 이산 코사인 변환(Discrete Cosine Transform)함으로써 DCT 계수를 결정하고, 결정된 DCT 계수들 중 적어도 하나의 계수를 원소(element)로 하는 특징 벡터를 추출할 수 있다.
- [77] 전자 장치(1000)가 이산 코사인 변환에 따라 결정된 DCT 계수를 모두 이용할

경우, 필터 뱅크 에너지의 빠른 변화로 인하여 음성 인식 모델의 성능은 떨어질 수 있다. 따라서, 전자 장치(1000)는 결정된 DCT 계수들 중 일부의 DCT 계수들을 제거하고 남은 DCT 계수들을 포함하는 특징 벡터를 음성 인식 모델에 입력함으로써 음성 인식의 정확도를 향상시킬 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 2-13번째 DCT 계수들에 기초하여 특징 벡터를 결정할 수 있으나, 이에 한정되는 것은 아니다. 그러나, 전자 장치(1000)가 평탄화된 필터 뱅크 별 에너지 성분으로부터 오디오 특징 벡터를 추출하는 과정은 DCT 계수를 이용한 방법에 한정되지 않으며, 기타 오디오 특징(feature)을 추출하는 방법에 기초하여 수행될 수도 있다.

- [78] 예를 들어, 전자 장치(1000)가 오디오 신호의 주파수 스펙트럼에 적용하는 필터 뱅크들은 주파수 영역에서 소정의 간격으로 중첩(Overlap)된 상태이기 때문에, 필터 뱅크 별 에너지들은 서로 상관되어 있다. 따라서, 전자 장치(1000)는 필터 뱅크 별 에너지 성분을 이산 코사인 변환함으로써 필터 뱅크 별 에너지 사이의 상관 관계를 분리하고, 분리된 필터 뱅크 별 독립적인 에너지(필터 뱅크의 주파수 대역에 해당하는 대역 에너지)에 기초하여 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 음성 인식의 정확도를 향상시킬 수 있다.
- [79] S250에서, 전자 장치(1000)는 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 오디오 신호 내 사용자의 음성을 인식할 수 있다. 전자 장치(1000)가 추출된 특징 벡터를 음성 인식 모델에 입력할 경우, 소정의 텍스트 정보를 포함하는 레이블에 대한 확률 값에 기초하여 사용자의 음성을 인식하는 구체적인 방법은 후술하는 도 18을 참조하여 구체적으로 설명하기로 한다.
- [80] 도 3은 일 실시 예에 따라 전자 장치가 필터 뱅크 별 에너지 성분을 결정하는 방법을 구체적으로 나타내는 흐름도이다.
- [81] S320에서, 전자 장치(1000)는 분포된 필터 뱅크를 오디오 신호의 주파수 스펙트럼에 적용할 수 있다. 예를 들어, 전자 장치(1000)는 주파수 도메인에서, 기 설정된 스케일에 따라 분포된 필터 뱅크의 채널 별 함수 및 상기 필터 뱅크의 주파수 대역에 대응하는 주파수 스펙트럼의 함수를 곱하는 동작은 전자 장치(1000)가 분포된 필터 뱅크를 오디오 신호의 주파수 스펙트럼에 적용하는 동작에 대응될 수 있다.
- [82] S340에서, 전자 장치(1000)는 필터 뱅크가 적용된 주파수 스펙트럼의 값을 로그 스케일로 변환할 수 있다. 전자 장치(1000)는 선형 스케일(linear-scale)이 아닌 사람의 청각적 특성을 반영하기 위해 주파수 스펙트럼 각각에 log 함수를 적용함으로써 필터 뱅크가 적용된 주파수 스펙트럼의 값을 로그 스케일로 변환한다.
- [83] S360에서, 전자 장치(1000)는 로그 스케일로 변환된 주파수 스펙트럼의 값을 이용하여 필터 뱅크 별 에너지 성분을 결정할 수 있다. 예를 들어, 전자 장치(1000)는 S320과 같이, 기 설정된 스케일에 따라 분포된 필터 뱅크의 채널 별 함수 및 상기 필터 뱅크의 주파수 대역에 대응하는 주파수 스펙트럼의 함수의

급을 제공하고, 제공한 결과에 로그 함수를 적용함으로써, 필터 뱅크 별 에너지 성분을 결정할 수 있다.

[84] 도 4는 일 실시 예에 따라 전자 장치가 미리 학습되는 평탄화 계수를 이용하여 필터 뱅크 별 에너지 성분을 평탄화(smoothing)하는 방법을 나타내는 흐름도이다.

[85] S420에서, 전자 장치(1000)는 필터 뱅크 별 에너지 성분을 평탄화 하기 위한 평탄화 계수를 필터 뱅크 별로 학습할 수 있다. 일 실시 예에 의하면, 전자 장치(1000)는 사람이 인지하는 필터 뱅크의 에너지들 사이의 세기(intensity)의 비선형(non-linearity) 특징을 반영하기 위해, 로그 스케일(log-scale) 또는 멜 스케일(mel-scale)에 따라 분포된 필터 뱅크를 오디오 신호의 주파수 스펙트럼에 적용함으로써 생성된 필터 뱅크 별 에너지 성분에 기초하여 특징 벡터를 추출(예컨대 log-mel 또는 MFCC에 따른 특징 추출)할 수 있다.

[86] 또 다른 실시 예에 의하면, 전자 장치(1000)는 사람이 인지하는 필터 뱅크의 에너지들 사이의 세기 관계를 반영하기 위해, 1/15의 값을 가지는 평탄화 계수를 거듭제곱의 지수로 하는 파워 함수(power-function)를 이용하여, 모든 필터 뱅크 별 에너지의 값을 평탄화(smoothing)하고, 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 특징 벡터를 추출(예컨대 Power-Normalized Cepstral Coefficient, PNCC에 따른 특징 추출)할 수도 있다.

[87] 그러나, 본 개시에 따른 전자 장치(1000)는 사람이 인지하는 필터 뱅크의 에너지들 사이의 세기(intensity)의 비선형(non-linearity) 특징을 반영하기 위해, 필터 뱅크 별로 평탄화 계수를 학습하고, 학습된 평탄화 계수를 거듭 제곱의 지수로 하는 파워 함수를 이용하여, 필터 뱅크 별 에너지의 값을 평탄화 할 수도 있다. 본 개시에 따른 평탄화 계수는 파워 함수(power-function)의 거듭 제곱의 지수인 파워 계수(power-coefficient)를 포함한다. 이하에서는 전자 장치(1000)가 필터 뱅크 별 평탄화 계수를 학습하는 과정을 구체적으로 설명하기로 한다. 먼저 균일하게 분포된 변수 Y를 하기 수학식 1에 기초하여 정의한다.

[88] [수식1]

$$Y = \sigma_p(X) = (X - x_{\min})^{\alpha}$$

[89] 여기에서, Y는 목표로 하는 균일하게 분포된 변수이고, X는 $x_{\max}$ 에서 $x_{\min}$ 사이의 랜덤 변수이며,

$\sigma_p(X)$

는 변수 X를 입력으로 하는 비선형 함수(nonlinearity function)이고,

α

는 평탄화 계수이다. 균일하게 분포된 변수 Y의 간격(Interval)은 변수 X의 최소 및 최대 값에 기초하여

$$I_Y = [0, (x_{\max} - x_{\min})^\alpha]$$

와 같이 정의될 수 있다. 여기에서 I_Y 는 균일하게 분포된 변수 Y 의 Interval이고, $x_{\max}$ 및 $x_{\min}$ 는 변수 X 의 Interval의 최대 및 최소 값이고,

α

는 평탄화 계수이다. 균일하게 분포된 Y 의 Interval에 따라, 변수 Y 를 uniform 함수를 이용하여 하기 수학식 2와 같이 다시 정의한다.

[90] [수식2]

$$Y \sim \mu(0, (X - x_{\min})^\alpha)$$

[91] 상기 수학식 2와 같이, 균일하게 분포된 변수 Y 를 0에서,

$$(X - x_{\min})^\alpha$$

입력을 가지는 유니폼 함수(uniform function)

$\mu()$

를 이용하여 정의할 수 있다. Y 가 균일 분포(uniform distribution)을 따른다고 가정하면, Y 의 확률 밀도 함수(PDF)는 하기 수학식 4와 같다.

[92] [수식3]

$$p_Y(y) = \begin{cases} \frac{1}{(x_{\max} - x_{\min})^\alpha}, & 0 \leq y \leq (x_{\max} - x_{\min})^\alpha \\ 0, & \text{otherwise} \end{cases}$$

[93] 여기서,

$p_Y(y)$

는 Y 의 확률 밀도 함수(PDF)이고, Y 의 확률 밀도 함수는

$$0 \leq y \leq (x_{\max} - x_{\min})^\alpha$$

범위에서, 균일하게 분포된

$$\frac{1}{(x_{\max} - x_{\min})^\alpha}$$

값이고, 기타 다른 범위에서는 0의 확률 밀도를 가질 수 있다. 상기 수학식 3과 같이 정의된 Y 의 확률 밀도 함수에 대하여, $x_{\max}$ 에서 $x_{\min}$ 사이의 선택된 랜덤 변수 X 의 확률 밀도 함수를 다음과 같이 정의한다.

[94]

[수식4]

$$\begin{aligned}
 pX(x) &= pY(y) \frac{dy}{dx} \\
 &= pY(y) [\alpha(x - x_{\min})^{\alpha-1}] \\
 &= \begin{cases} \frac{\alpha(x-x_{\min})^{\alpha-1}}{x_{\max}-x_{\min}^{\alpha}}, & x_{\min} \leq x \leq x_{\max} \\ 0, & \text{otherwise} \end{cases}
 \end{aligned}$$

[95] Y의 확률 밀도 함수에 대한, X의 확률 밀도 함수

 $pX(x)$

를 구하기 위해,

 $pY(y)$ 를 x 에 대하여 미분한다. $pY(y)$ 를 x 에 대하여 미분함으로써, 변수 X 에 대한 확률 밀도 함수는 $x_{\min} \leq x \leq x_{\max}$

범위에서,

$$\frac{\alpha(x - x_{\min})^{\alpha-1}}{x_{\max} - x_{\min}^{\alpha}}$$

값을 가지고, 다른 범위에서는 0의 값을 나타낼 수 있다. 상기 수학식 4에서 변수 X 는 하기의 수학식 5와 같이 정의될 수 있다.

[96] [수식5]

$$X = \{x_0, x_1, \dots, x_{N-1}\}$$

[97] 즉, 변수 X 는 0에서 $N-1$ 까지의 총 N 개의 랜덤 변수를 포함하는 벡터로 표현될 수 있다. 이제, 변수 X 의 확률 밀도 함수의 로그 우도(log likelihood)를 최대화하는 평탄화 계수

 α

를 결정(Maximum Likelihood Estimation, MLE)하기 위해 다음과 같이 변수 X 에 대한 우도 함수를 정의한다.

[98]

[수식6]

$$\begin{aligned}
 L(\alpha|X) &= \sum_{i=0}^{N-1} \ln p_X(x_i) \\
 &= \sum_{i=0}^{N-1} \ln \left[\frac{\alpha(x_i - x_{\min})^{\alpha-1}}{(x_{\max} - x_{\min})^\alpha} \right] \\
 &= N \ln(\alpha) + (\alpha - 1) \sum_{i=0}^{N-1} \ln(x_i - x_{\min}) - N \alpha \ln(x_{\max} - x_{\min}) .
 \end{aligned}$$

[99] 여기서,

 $L(\alpha|X)$ 는 변수 X 의 평탄화 계수 α

에 대한 로그 우도(Log Likelihood)이고,

 $p_X(x_i)$ 는 i 번째 변수 x 의 확률 밀도 함수를 나타낸다. $L(\alpha|X)$

는 상기 수식 6에서와 같이 3개의 로그 항으로 표현될 수 있는데, 마지막 로그항(term)

 $\ln(x_i - x_{\min})$

은

 $x_i = x_{\min}$

에서 정의될 수 없다. 따라서, 올림 계수(flooring coefficient)

 δ

를 이용하여

 $x_i = x_{\min}$

에서

 $L(\alpha|X)$

를 표현하면 다음과 같다.

[100] [수식7]

$$L(\alpha|X) = N \ln(\alpha) + (\alpha - 1) \sum_{i=0}^{N-1} \ln(\max\{x_i - x_{\min}, \delta\}) - N \alpha \ln(x_{\max} - x_{\min})$$

[101] 여기서,

$L(\alpha|X)$

는 변수 X 의 평탄화 계수

α

에 대한 로그 우도(Log Likelihood)이고, 변수 x 가

$x_i = x_{\min}$

일 때 로그 항은 올림 계수(flooring coefficient)

δ

를 이용하여

$$(\alpha - 1) \sum_{i=0}^{N-1} \ln(\max\{x_i - x_{\min}, \delta\}) - N\alpha \ln(x_{\max} - x_{\min})$$

와 같이 표현될 수 있다. 일 실시 예에 의하면 올림 계수(flooring coefficient)는 10^{-1} - 100 일 수 있으나, 이에 한정되는 것은 아니다. 이제 평탄화 계수

α

에 대한 최대 우도(Maximum Likelihood)를 결정하기 위해

$L(\alpha|X)$

를 평탄화 계수

α

에 대하여 미분한다. 따라서 변수 X 의 평탄화 계수

α

에 대한 로그 우도(Log Likelihood)를 최대화하는 평탄화 계수

$\hat{\alpha}$

는 다음과 같다.

[102] [수식8]

$$\hat{\alpha} = \frac{1}{\ln(x_{\max} - x_{\min}) - \frac{1}{N} \sum_{i=0}^{N-1} \ln(\max\{x_i - x_{\min}, \delta\})}$$

[103] 여기서,

$\hat{\alpha}$

는 변수 X 의 평탄화 계수

α

에 대한 로그 우도(Log Likelihood)를 최대화하는 평탄화 계수로써, 변수 X 의 Interval의 최대값 $x_{\max}$, 최소값 $x_{\min}$ 및 올림 계수

δ

를 이용하여 상기 수학식8과 같이 표현될 수 있다.

- [104] 즉, 전자 장치는 수학식 1 내지 8에서와 같이, 균일하게 분포된 타겟 변수 Y 를 가정하고, 타겟 변수 Y 에 대한 현재 변수 X 의 로그 우도(log likelihood)를 결정하며, 결정된 로그 우도가 최대화되는지 여부에 따라, 평탄화 계수를 결정할 수 있다. 상기 수학식 1 내지 8에서의 타겟 변수 Y 는 균일하게 분포된 필터 뱅크 별 에너지 성분에 대응되고, 변수 X 는 현재 수신된 오디오 신호의 필터 뱅크 별 에너지 성분에 대응될 수 있다.
- [105] 예를 들어, 전자 장치(1000)는 균일하게 분포된 필터 뱅크 별 에너지 성분(예컨대 필터 뱅크 별 에너지의 PDF)에 대한, 현재 수신된 오디오 신호의 필터 뱅크 별 에너지 성분(예컨대 필터 뱅크 별 에너지의 PDF)의 로그 우도를 결정하고, 로그 우도가 최대화되는지 여부에 따라 평탄화 계수를 결정할 수 있다. 즉, 전자 장치(1000)는 오디오 학습 신호를 수신하고, 균일하게 분포된 필터 뱅크 별 에너지 성분에 대한, 오디오 학습 신호의 필터 뱅크 별 에너지 성분의 로그 우도가 최대화되는지 여부에 기초하여, 필터 뱅크 별로 평탄화 계수를 미리 학습할 수 있다. 즉, 전자 장치(1000)가 필터 뱅크 별로 학습한 평탄화 계수는, 사람이 인지하는 필터 뱅크의 에너지들 사이의 세기(intensity)의 비선형(non-linearity) 특징을 반영하기 위해 모든 필터 뱅크 별로 동일한 평탄화 계수인 1/15를 적용한 PNCC 특징 추출 방법과는 달리, 필터 뱅크 별로 서로 다른 값을 가질 수 있다. 이하 도 5를 참조하여 필터 뱅크 별로 서로 다르게 학습되는 평탄화 계수에 대해 설명하기로 한다.
- [106] 도 5는 필터 뱅크 채널 별로 학습되는 평탄화 계수를 설명하기 위한 도면이다.
- [107] 모든 필터 뱅크 별로 동일한 값(예컨대 1/15)의 평탄화 계수(606)가 적용되는 PNCC와 달리, 본원 전자 장치(1000)는 서로 다른 채널을 가지는 필터 뱅크 별로 서로 다른 평탄화 계수를 학습할 수 있다.
- [108] 예를 들어, 전자 장치(1000)는 노이즈 신호를 포함하지 않는 오디오 학습 신호를 수신하고, 균일하게 분포된(uniformly distributed) 필터 뱅크 별 에너지 성분에 대한, 오디오 학습 신호의 필터 뱅크 별 에너지 성분의 로그 우도가 최대화되도록 평탄화 계수(602)를 학습할 수 있다.
- [109] 또한, 전자 장치(1000)는 노이즈 신호를 포함하는 오디오 학습 신호를 수신하고, 균일하게 분포된(uniformly distributed) 필터 뱅크 별 에너지 성분에 대한, 오디오 학습 신호의 필터 뱅크 별 에너지 성분의 로그 우도가 최대화되도록 평탄화 계수(604)를 학습할 수 있다.
- [110] 즉, 도 5에 도시된 바와 같이, 전자 장치(1000)는 사람이 인지하는 필터 뱅크의 에너지들 사이의 세기(intensity)의 비선형(non-linearity) 특징을 반영하기 위해,

1/15와 같이 필터 뱅크 채널 별로 동일한 평탄화 계수를 결정할 수도 있지만, 노이즈 신호를 포함하지 않는 오디오 학습 신호 또는 노이즈 신호를 포함하는 오디오 학습 신호에 기초하여, 필터 뱅크 별 서로 다른 평탄화 계수를 학습할 수도 있다. 본 개시에 따른 전자 장치(1000)는 필터 뱅크 별로 서로 다르게 학습된 평탄화 계수를 이용하여, 현재 수신된 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화(smoothing)함으로써 사용자의 음성을 더 정확하게 인식할 수 있다.

[111] 도 6은 일 실시 예에 따라 전자 장치가 타겟 히스토그램을 매핑하기 위한 매핑 함수를 이용하여 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하는 방법을 나타내는 흐름도이다.

[112] 전자 장치(1000)는 도 4 내지 도 5와 달리, 분포된 필터 뱅크를 오디오 신호의 주파수 스펙트럼에 적용함으로써 필터 뱅크 별 에너지 성분을 획득하고, 획득된 필터 뱅크 별 에너지 성분 별 크기에 관한 히스토그램을 생성하며, 생성된 히스토그램을 타겟 히스토그램으로 매핑(mapping)하기 위한 매핑 함수(mapping function)에 기초하여, 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화 할 수도 있다.

[113] 구체적으로, S520에서, 전자 장치(1000)는 오디오 신호의 주파수 스펙트럼에 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 획득하고, 획득된 필터 뱅크 별 에너지 성분의 크기에 관한 히스토그램을 생성할 수 있다. 예를 들어, 전자 장치(1000)는 오디오 신호의 필터 뱅크 별 에너지 성분의 크기를 소정의 주파수 영역(bin)으로 분할하고, 분할된 주파수 영역 별 필터 뱅크 에너지 성분의 크기에 기초하여 히스토그램을 생성할 수 있다. 주파수 성분은 단위 주파수 영역 또는 주파수 bin을 포함할 수 있다.

[114] 또 다른 실시 예에 의하면, 전자 장치(1000)는 필터 뱅크 별 에너지 성분의 주파수 성분 별 빈도수(frequency)에 기초하여 히스토그램을 생성할 수도 있다. 예를 들어, 전자 장치(1000)는 오디오 신호에 필터 뱅크를 적용함으로써 획득된 필터 뱅크 별 에너지 성분(예컨대 필터 뱅크 별 에너지의 PDF)의 주파수 스펙트럼을 소정의 주파수 영역(bin)으로 분할하고, 분할된 주파수 영역 별로 나타나는 스펙트럼 값의 빈도수에 기초하여 히스토그램을 생성할 수도 있다.

[115] S540에서, 전자 장치(1000)는 상기 생성된 히스토그램을, 상기 주파수 성분 별 크기가 균일하게 분포된 타겟 히스토그램으로 매핑하기 위한 매핑 함수를 결정할 수 있다. 타겟 히스토그램은 필터 뱅크 별 에너지 성분이 균일하게 분포된(uniformly distributed) 오디오 신호의 필터 뱅크 별 에너지 값을 획득하기 위해 가정되는, 오디오 신호의 필터 뱅크 별 에너지 성분 내 주파수 성분 별 크기를 나타낼 수 있다. 매핑 함수는 하기 수학식 9로부터 결정될 수 있다.

[116]

[수식9]

$$Y = \sigma_{np}(X) = F_u^{-1}(\widehat{F}_x(X))$$

[117] 여기에서,

$\sigma_{np}(X)$

는 변수 X (예컨대 필터뱅크별 에너지 성분, PDF)의 히스토그램을 변수 Y 의 히스토그램으로 매핑하기 위한, 변수 X 의 히스토그램을 입력으로 하는 비선형 함수(nonlinearity function)이고,

$F_u(\cdot)$

는 균일하게 분포된 타겟 히스토그램의 확률 밀도 함수이며,

$\widehat{F}_x^{-1}(\cdot)$

의 누적 분포 함수(Cumulative distribution function, CDF)이고,

F_u^{-1}

는 변수 X 의 히스토그램을 타겟 히스토그램으로 매핑하기 위한 매핑 함수이다.

[118] S560에서, 전자 장치(1000)는 결정된 매핑 함수를 이용하여 상기 오디오 신호의 필터뱅크별 에너지 성분을 변환할 수 있다. 예를 들어, 전자 장치(1000)는 현재 수신된 오디오 신호의 필터뱅크별 에너지 성분 내 주파수 성분별 필터뱅크 에너지 값을 매핑 함수에 입력하고, 상기 매핑 함수에서 출력된 결과에 기초하여, 주파수 성분별 크기 또는 빈도수가 타겟 히스토그램을 나타내도록 하는 필터뱅크별 에너지 값을 획득함으로써, 현재 수신된 오디오 신호의 필터뱅크별 에너지 성분을 변환할 수 있다. 전자 장치(1000)는 매핑 함수를 이용하여 오디오 신호의 필터뱅크별 에너지 성분을 변환함으로써, 오디오 신호의 필터뱅크별 에너지 성분을 평탄화(smoothing)할 수 있다. 전자 장치(1000)는 평탄화된 필터뱅크별 에너지 성분에 기초하여, 오디오 신호의 특징 벡터를 추출하고, 추출된 오디오 특징 벡터를 이용하여 오디오 신호 내 사용자의 음성을 인식함으로써, 정확하게 사용자의 음성을 인식할 수 있다. S560은 도 2의 S220 내지 S230에 대응될 수 있으므로 구체적인 설명은 생략하기로 한다.

[119] 도 7은 일 실시예에 따른 특정 채널 인덱스에 해당하는 필터뱅크의 에너지 성분을 나타내는 도면이다.

[120] 도 7을 참조하면, 전자 장치가 멜 스케일에 따라 분포된 필터뱅크들 중, 채널 인덱스가 3인 필터뱅크를 주파수 스펙트럼에 적용함으로써 획득한 멜 필터뱅크 에너지의 PDF가 도시된다.

[121] 예를 들어, 전자 장치(1000)가 전술한 도 4 및 도 6에서 설명한 바에 따라 멜 필터뱅크 에너지의 PDF를 평탄화하지 않는 경우(702), 전자 장치는 0 Hz의 주파수 영역 부근에 집중적으로 분포된 멜 필터뱅크 에너지의 PDF를 획득할 수

있다. 전자 장치(1000)가 균일하게 분포되지 않은 에너지 성분을 포함하는 주파수 스펙트럼으로부터 추출된 특징 벡터를 이용할 경우, 전자 장치의 음성 인식의 정확도는 떨어질 수 있다.

- [122] 그러나, 일 실시 예에 따라, 전자 장치(1000)가 미리 학습된 평탄화 계수를 이용하여 채널 인덱스 $l=3$ 인 필터 뱅크 에너지의 PDF를 평탄화(smoothing)할 경우(704), -1에서 1까지 정규 분포 모양에 유사한 필터 뱅크 에너지의 PDF를 획득할 수 있다. 또 다른 실시 예에 의하면, 전자 장치(1000)는 히스토그램 기반 매핑 함수를 이용하여 채널 인덱스 $l=3$ 인 필터 뱅크 에너지의 PDF를 평탄화(smoothing)할 경우(706), 균일하게 분포된(uniformly distributed)된 필터 뱅크 에너지의 PDF를 획득할 수 있다. 전자 장치(1000)는 오디오 신호의 필터 뱅크 별 에너지 PDF를 평탄화(smoothing)함으로써, 오디오 신호 내 사용자 음성을 정확하게 인식할 수 있다.
- [123] 도 8은 일 실시 예에 따른 필터 뱅크 에너지에 대한 비선형 출력(nonlinearity output)을 설명하기 위한 도면이다.
- [124] 전자 장치(1000)는 비선형(non-linearity) 특징을 나타내는 사람의 청각적 특성을 반영하기 위해, 오디오 신호의 필터 뱅크 별 에너지 성분에 동일한 평탄화 계수 $1/15$ 를 적용(708)할 수도 있지만, 균일하게 분포된 필터 뱅크 별 에너지 성분에 대한, 현재 수신된 오디오 신호의 필터 뱅크 별 에너지 성분의 로그 우도를 최대화하는 평탄화 계수를 필터 뱅크 별로 학습함으로써, 필터 뱅크 별로 서로 다르게 학습된 평탄화 계수를, 오디오 신호의 필터 뱅크 별 에너지 성분에 적용(704)할 수도 있다. 또한, 전자 장치(1000)는 오디오 신호의 주파수 스펙트럼의 주파수 성분 별 크기에 관한 히스토그램을 생성하고, 생성된 히스토그램을 타겟 히스토그램으로 매핑(mapping)하기 위한 매핑 함수(mapping function)에 기초하여, 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화(706)할 수도 있다.
- [125] 전자 장치(1000)가 필터 뱅크 별로 학습된 평탄화 계수를 이용하여, 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화(smoothing)한 경우(704), 비선형 출력(non-linearity output)은, 필터 뱅크의 에너지 성분이 감소함에 따라, 동일한 평탄화 계수 $1/15$ 를 이용하여 필터 뱅크 별 에너지를 평탄화 한 경우(708)보다, 더 가파르게(sharply) 감소할 수 있다. 그러나, 전자 장치(1000)가 히스토그램 기반 매핑 함수를 이용하여 필터 뱅크 별 에너지를 평탄화 한 경우(706) 비선형 출력은 필터 뱅크의 에너지 성분의 변화에 상관없이 균일하게 유지될 수 있다. 비선형 출력(non-linearity output)은 비선형 특징을 나타내는 사람의 청각적 특징을 반영하기 위해, 필터 뱅크 별 에너지의 비선형(non-linearity) 특징의 정도를 나타낼 수 있다.
- [126] 도 9는 일 실시 예에 따른 전자 장치가 음성 인식 모델 학습을 위한 오디오 신호 처리 방법을 나타내는 흐름도이다.
- [127] 전자 장치(1000)는 오디오 신호를 획득하고, 오디오 신호로부터 추출된 특징

벡터를 미리 학습된 음성 인식 모델에 입력함으로써 오디오 신호 내 사용자의 음성을 인식할 수 있다. 이하에서는 전자 장치(1000)가 음성 인식 모델을 미리 학습하기 위한 오디오 신호를 처리하는 방법을 설명하기로 한다.

- [128] S910에서, 전자 장치(1000)는 음성 인식 모델을 학습하기 위한 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득할 수 있다. 예를 들어, 전자 장치(1000)는 기 설정된 윈도우 길이를 가지는 hanning 윈도우를 소정의 간격으로 중첩(overlap)하여 배치하고, 중첩하여 배치된 hanning 윈도우를 이용하여 프레임 단위로 구분되는 제1 오디오 신호를 획득할 수 있다. 일 실시 예에 의하면 hanning 윈도우의 길이는 50ms일 수 있으나, 이에 한정되는 것은 아니고, 목표로 하는 주파수 해상도(resolution)에 따라 달라질 수 있다. S910은 도 2의 전자 장치(1000)가 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 동작에 대응될 수 있으므로 구체적인 설명은 생략하기로 한다.

- [129] S920에서, 전자 장치(1000)는 프레임 단위로 구분되는 제1 오디오 신호의 주파수 스펙트럼을 획득할 수 있다. 예를 들어, 전자 장치(1000)는 16 KHz의 주파수에서, $K=1024$ 개의 샘플수로 샘플링된 프레임 단위의 제1 오디오 신호를 획득할 수 있다. 전자 장치(1000)는 주파수축을 변환하는 동안(frequency warping) 더 나은 주파수 해상도(resolution)을 획득하기 위해, 오버 사이즈 FFT 계수 U 를 16으로 설정하고, 설정된 오버 사이즈 FFT 계수에 따라 제1 오디오 신호를 업 샘플링하며, 업 샘플링된 제1 오디오 신호를 고속 푸리에 변환(Fast Fourier Transform, FFT, 106)함으로써 주파수 스펙트럼을 획득할 수 있다. FFT 사이즈 계수가 K 인 스펙트럼의 주파수 축이 변환된(frequency warped)된 오디오 신호를

$$Y_K[m, e^{jw_k}]$$

라 할 때, FFT 사이즈 계수가 UK 인 주파수 축 변환 전 원(original) 신호의 스펙트럼

$$X_{UK}[m, e^{jv_k}]$$

의 관계는 하기 수식 10과 같다.

- [130] [수식 10]

$$\begin{aligned} Y_K[m, e^{jw_k}] &= X_{UK}[m, e^{jv_{k_0}}] \\ &= X_{UK}[m, e^{j\phi_\alpha(w_k)}], 0 \leq k \leq \frac{K}{2} \end{aligned}$$

- [131] 여기서, m 은 전자 장치가 획득한 오디오 신호의 프레임 인덱스이고,

$$\phi_\alpha(w_k)$$

은 워핑 함수이며, 이산 시간 주파수

$$w_k = \frac{2\pi k}{K}$$

이고, 이산 시간 주파수

$$v_{k_0} = \frac{2\pi k_0}{UK}$$

이며, U는 오버사이즈 FFT 사이즈 계수, K는 FFT 사이즈 계수이다.

- [132] 주파수 축이 변환된 오디오 신호의 주파수 스펙트럼의 대칭 성을 고려하여 주파수 축이 변환된 오디오 신호의 스펙트럼을

$$0 \leq k \leq \frac{K}{2}$$

범위에서

$$Y(e^{jw_k})$$

로 정의할 수 있다. 상기 수학식 10의

$$Y_K[m, e^{jw_k}]$$

를 획득함에 있어서, 주파수 축 변환 전 원(original) 신호의 스펙트럼

$$X_{UK}[m, e^{jv_k}]$$

는 이산 값(discrete value) k에 대해서만 정의될 수 있으므로,

$$\emptyset_\alpha(w_k)$$

는 소정의 정수 k_0 을 이용하여 v_{k_0} 에 대해서 표기되어야 한다. v_{k_0} 는

$$v_{k_0} = \frac{2\pi k_0}{UK}$$

와 같이 정의되었으므로, 위평 함수

$$\emptyset_\alpha(w_k)$$

는 하기 수학식 11과 같이 정의될 수 있다.

- [133] [수식11]

$$\frac{2\pi k_0}{UK} \approx \emptyset_\alpha(w_k) \cdot$$

- [134] 여기에서 U는 오버사이즈 FFT 계수이고, K는 FFT 사이즈의 계수이며,

$$\emptyset_\alpha(w_k)$$

은 위평 함수이다. 따라서, 상기 수학식 11에서 k_0 은 하기 수학식 12와 같이 정의된다.

- [135]

[수식12]

$$K_0 \approx \frac{UK\emptyset_{\alpha}(w_k)}{2\pi}$$

[136] 상기 수학식 12에서 k_0 는 하기 수학식 13에 의해 반올림될 수 있다.

[137] [수식13]

$$K_0 = \left\lfloor \frac{UK\emptyset_{\alpha}(w_k)}{2\pi} + 0.5 \right\rfloor$$

[138] 여기에서, U는 오버사이즈 FFT 계수이고, K는 FFT 사이즈의 계수이며,

$$\emptyset_{\alpha}(w_k)$$

은 워핑 함수이고,

$$\lfloor \cdot \rfloor$$

는 바닥 연산자(floor operator)로써 함수 내 입력 값 이하의 최대 정수를 구하기 위한 연산자이다. 0.5는 소수점 이하의 반올림 연산을 위하여 추가된 상수이다. 상기 수학식 10 및 수학식 13을 이용하여 하기 수학식 14와 같이, 주파수 축이 변환된(frequency warped)된 오디오 신호를

$$Y_K[m, e^{jw_k}]$$

를 정의할 수 있다.

[139] [수식14]

$$\begin{aligned} Y_K[m, e^{jw_k}] &= X_{UK}[m, e^{jv_{k_0}}] \\ &= X_{UK} \left[m, e^{jv \left\lfloor \frac{UK\emptyset_{\alpha}(w_k)}{2\pi} + 0.5 \right\rfloor} \right], \\ &0 \leq k \leq \frac{K}{2} \end{aligned}$$

[140] 여기에서,

$$Y_K[m, e^{jw_k}]$$

는 주파수 축이 변환된(frequency warped)된 오디오 신호의 스펙트럼이고,

$$X_{UK}[m, e^{jv_k}]$$

는 주파수 축 변환 전 원 신호의 스펙트럼이며, 이산 시간 주파수

$$w_k = \frac{2\pi k}{K}$$

이고, 이산 시간 주파수

$$v_{k_0} = \frac{2\pi k_0}{UK}$$

이다.

- [141] S930에서, 전자 장치(1000)는 복수 화자들의 서로 다른 성도 길이(Vocal Tract Length) 변화를 반영하기 위해 상기 수학식 10 및 14에서 정의된 워핑 함수

$$\phi_\alpha(w_k)$$

를 이용하여, 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환할 수 있다. 예를 들어, 전자 장치(1000)는 워핑 계수(warping coefficient) 및 상기 워핑 계수에 기초하여 정의되는 워핑 함수(warping function)를 이용하여, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환(frequency warping, 108)할 수 있다. 전자 장치가 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 구체적인 방법은 후술하는 도 11 및 도 12를 참조하여 구체적으로 설명하기로 한다.

- [142] S940에서, 전자 장치(1000)는 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성할 수 있다. 예를 들어, 전자 장치(1000)는 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환(Inverse Fast Fourier Transform)하고, 역 고속 푸리에 변환된 제1 오디오 신호의 주파수 스펙트럼을 시간 축에서 중첩하여 더함(Overlap Addition, 112)으로써, 재-합성된(re-synthesized) 제2 오디오 신호(116)를 획득할 수 있다.

- [143] S950에서, 전자 장치(1000)는 룸(Room) 내 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스 필터를 제2 오디오 신호에 적용할 수 있다. 예를 들어, 룸 임펄스 필터는, 전달함수로서 룸 임펄스 응답(Room impulse Response)일 수 있다. 예를 들어, 룸 임펄스 필터는 룸 내 오디오 신호의 전달 경로 별 서로 다른 음향 특징을 나타낼 수 있다.

- [144] 룸 임펄스 필터가 나타내는 음향 특징은 룸 내 전달 경로 별 반향 시간(wide distribution of reverberation), 오디오 신호의 전달 경로 별 룸 차원(dimension) 및 룸 내 전달 경로 별 신호 대 잡음 비(Signal To Noise Ratio)를 포함할 수 있으나, 이에 한정되는 것은 아니다. 또한 룸 내 오디오 신호의 전달 경로는 오디오 신호가 생성되는 음원 및 상기 오디오 신호가 수신되는 마이크의 위치에 기초하여 정의될 수 있다.

- [145] 전자 장치(1000)가 룸 임펄스 필터를 제2 오디오 신호에 적용하는 동작은 룸 임펄스 응답을 시간 영역에서 제2 오디오 신호와 컨벌루션하는 동작 또는 주파수 영역에서 제2 오디오 신호를 룸 임펄스 응답과 곱하는 동작에 대응될 수 있다. 전자 장치(1000)는 재-합성된 제2 오디오 신호에 룸 임펄스 필터를 적용함으로써, 다양한 환경의 룸 내에서 오디오 신호가 전달되는 상황을

시뮬레이션 할 수 있으며, 실제로는 오디오 신호가 전달되지 않았던 전달 경로의 음향 특징이 반영된 오디오 신호를 획득할 수 있다. 즉, 전자 장치(1000)는 제2 오디오 신호에 룸 임펄스(RIR) 필터를 적용함으로써 실제로는 획득하지 않았던 증강된(Augmented) 오디오 신호들을 획득할 수 있다.

- [146] 도 10은 일 실시 예에 따른 전자 장치가 음성 인식 모델 학습을 위한 오디오 신호를 처리 과정을 개략적으로 설명하기 위한 도면이다.
- [147] 전자 장치(1000)는 오디오 신호를 획득하고, 획득된 오디오 신호들을 소정의 데이터 증강 기법을 이용하여 증강(augment)시킬 수 있다. 예를 들어, 전자 장치(1000)는 사용자로부터 사용자의 음성을 포함하는 제1 오디오 신호(1002) 및 1에서 I-1번째 잡음 원들(noise sources)로부터 잡음 신호들(1004, 1006, 1008)을 획득할 수 있다.
- [148] 전자 장치(1000)는 제1 오디오 신호(1002)의 주파수 스펙트럼을 획득하고, 위핑 계수 및 위핑 계수에 기초하여 정의되는 위핑 함수(warping function)를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환(frequency warping, 1010)한다. 전자 장치(1000)는 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 룸 임펄스 필터(1020)과 주파수 영역에서 곱할 수 있다. 룸 임펄스 필터(1020)는 룸(room) 내 0번째 음원으로부터 j번째 마이크까지 오디오 신호가 전달되는 전달 경로의 음향 특징을 나타내는 룸 임펄스 응답(room impulse response)이다.
- [149] 또한, 전자 장치(1000)는 1에서 I-1번째 잡음 원들(noise sources)로부터 획득된 잡음 신호들(1004, 1006, 1008)에 잡음 신호들 사이의 상대적인 세기를 조절하기 위한 이득 계수(gain factor) g를 곱하고, 이득 계수가 곱해진 잡음 신호들 각각에 룸 임펄스 필터를 주파수 영역에서 곱할 수 있다. 전자 장치는 잡음 신호들(1004, 1006, 1008) 및 이득 계수 g의 곱한 결과에, 룸 임펄스 필터를 적용한 결과와 주파수 축이 변환된 제1 오디오 신호에 룸 임펄스 필터(1020)를 적용한 결과를 모두 더함으로써 증강된(augmented) 오디오 신호 $y_j[n]$ 을 획득할 수 있다. 증강된 오디오 신호 $y_j[n]$ 은 하기의 수학식으로 정의될 수 있다.
- [150] [수식15]

$$y_j[n] = F(x_0[n]|\alpha) * h_{0j}[n] + g \sum_{i=1}^{I-1} x_i[n] * h_{ij}[n]$$

- [151] 여기에서, $y_j[n]$ 은 j번째 마이크에서 획득된 오디오 신호이고, $x_0[n]$ 은 인식 대상이 되는 사용자의 음성 신호이며, $h_{0j}[n]$ 은 0번째 잡음 원으로부터 j번째 잡음 원까지의 룸 임펄스 응답이고, $x_i[n]$ 은

$$1 \leq i \leq I - 1$$

범위 내에서 i번째 잡음 원으로부터 획득된 잡음 신호이며, $h_{ij}[n]$ 은 i번째 잡음

원에서 j 번째 마이크까지 잡음 신호 $x_i[n]$ 이 전달되는 경로의 음향 특징을 나타내는 룸 임펄스 응답이다. 또한,

$$F(\cdot | \alpha)$$

는 워핑 함수(frequency warping)이고,

$$\alpha$$

는 워핑 계수를 나타낸다.

- [152] 일반적으로 서로 다른 화자의 성도 길이(Vocal tract length) 변화를 모델링하기 위한 VTLP의 경우, 시간 영역이 아닌, 주파수 영역에서, 오디오 신호의 주파수 스펙트럼 내 필터 뱅크의 스케일을 조정함으로써 서로 다른 화자의 성도 길이 변화를 모델링 하였으나, VTLP 신호 처리 과정 중, 필터 뱅크가 적용된 이후에는 또 다른 데이터 증강 기술로써 RIR 필터를 적용하기 어려운 문제가 있었다.
- [153] 본 개시에 따른 전자 장치(1000)는 제1 오디오 신호를 획득하고, 획득된 제1 오디오 신호로부터 특징 벡터를 추출하기에 앞서, 먼저 서로 다른 화자의 성도 길이 변화를 모델링 하기 위해, 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 재-합성(re-synthesize)하기 때문에, VTLP 및 음향 시뮬레이션(acoustic simulation)으로써 RIR을 동시에 적용할 수 있는 효과가 있다.
- [154] 또한, 본 개시에 따른 전자 장치(1000)가 오디오 신호를 증강(augmentation)함에 있어서, 서로 다른 화자의 성도 길이 변화를 반영하기 위한 워핑 계수 및 워핑 함수를 이용하여 오디오 신호를 재-합성하고, 재-합성된 오디오 신호에 RIR 필터를 적용하는 순서는, 실제 룸(room) 내에서 서로 다른 성도 길이를 가지는 화자의 음성 신호가 마이크까지 수신되는 일련의 오디오 신호의 전달 순서를 반영할 수 있다. 따라서, 본 개시에 따른 전자 장치(1000)는 서로 다른 화자의 성도 길이 변화를 반영하기 위한 워핑 계수 및 워핑 함수를 이용하여 오디오 신호를 재-합성한 후, 재합성된 오디오 신호에 RIR 필터를 적용함으로써, 음성 인식의 정확도가 향상된 음성 인식 모델을 학습시킬 수 있다.
- [155] 도 11은 일 실시 예에 따른 전자 장치가 주파수 축을 변환하는 방법을 구체적으로 나타내는 흐름도이다.
- [156] 도 11을 참조하여 전자 장치(1000)가 서로 다른 화자의 성도 길이 변화를 모델링하기 위해, 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 과정을 구체적으로 설명하기로 한다. S1112에서, 전자 장치(1000)는 워핑 계수(warping coefficient)를 결정한다. 예를 들어, 전자 장치(1000)는 제1 오디오 신호의 프레임 별로 워핑 계수를 랜덤(random)하게 결정할 수 있다.
- [157] S1114에서, 전자 장치(1000)는 결정된 워핑 계수에 기초하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하기 위한 워핑 함수(warping function)을

결정할 수 있다. 전자 장치(1000)는 주파수 스펙트럼의 주파수 축을 변환하기 위해, piecewise linear rule 또는 bilinear rule에 기초하여 워핑 함수를 결정할 수 있다. 먼저 piecewise linear rule에 따른 워핑 함수는 하기 수학과 같이 정의될 수 있다.

[158] [수식16]

$$w'_k = \begin{cases} w\alpha, & w \leq w_{hi} \frac{\min(\alpha, 1)}{\alpha} \\ \pi - \frac{\pi - w_{hi} \frac{\min(\alpha, 1)}{\alpha}}{\pi - w_{hi} \frac{\min(\alpha, 1)}{\alpha}} (\pi - w), & \text{otherwise} \end{cases}$$

[159] 여기서,

α

는 워핑 계수이고, 워핑 함수는 입력 주파수 W 와 출력 주파수로써, 주파수 축이 변환된 W'_k 의 관계를 정의한다. 예를 들어, 입력 주파수 W 가

$$w \leq w_{hi} \frac{\min(\alpha, 1)}{\alpha}$$

를 만족할 때, 주파수 축이 변환된(frequency warped)된 출력 주파수 W'_k 는 W

이고, 기타 범위에서 주파수 축이 변환된(frequency warped)된 출력 주파수 W'_k 는

$$\pi - \frac{\pi - w_{hi} \frac{\min(\alpha, 1)}{\alpha}}{\pi - w_{hi} \frac{\min(\alpha, 1)}{\alpha}} (\pi - w)$$

와 같이 표기될 수 있다. 전자 장치(1000)는 상기 수학적 식 16에 기초하여, piecewise linear rule를 따르는, 주파수 W 를 입력으로 하고, 출력 주파수 W'_k 를 출력하는 워핑 함수를 결정할 수 있다.

[160] 또 다른 실시 예에 의하면, 전자 장치(1000)는 bilinear rule에 기초하여 워핑 함수를 결정하고, 결정된 워핑 함수에 따라 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환할 수 있다.

[161] [수식17]

$$w'_k = w_k + 2 \tan^{-1} \left(\frac{(1 - \alpha) \sin(w_k)}{1 - (1 - \alpha) \cos(w_k)} \right)$$

[162] 여기서, W_k 는 입력 주파수이고, W'_k 는 주파수 축이 변환된(frequency warped)된 출력 주파수이다. 전자 장치(1000)는 상기 수학적 식 17에 기초하여, bilinear rule를 따르는, 주파수 W_k 를 입력으로 하고, 출력 주파수 W'_k 를 출력하는

워핑 함수를 결정할 수 있다. S1116에서, 전자 장치(1000)는 상기 수학식 16 또는 17에 따라 워핑 함수를 결정하고, 결정된 워핑 함수를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환할 수 있다.

- [163] 도 12는 워핑 계수에 따라 입력 주파수가 워핑(warped)되는 과정을 설명하기 위한 도면이다.
- [164] 전자 장치(1000)는 워핑 계수를 결정하고, 결정된 워핑 계수에 기초하여 정의되는 워핑 함수를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환한다. 예를 들어, 도 12를 참조하면, 전자 장치(1000)는 워핑 계수를 0.8 내지 1.2 사이로 결정하고, 결정된 워핑 계수에 기초하여 bilinear rule를 따르는 워핑 함수를 이용하여 입력 주파수를 주파수 워핑(frequency warping) 한다.
- [165] 예를 들어, 전자 장치(1000)가 워핑 계수를 0.8로 설정하는 경우, bilinear rule를 따르는 워핑 함수에서 출력되는 출력 주파수(예컨대, warped frequency)의 분포는 입력 주파수에 대하여, 아래 쪽으로 볼록한 모양의 주파수 분포를 나타낼 수 있다. 또 다른 실시 예에 의하면, 전자 장치(1000)가 워핑 계수를 1로 설정하는 경우, 워핑 함수에서 출력되는 출력 주파수의 분포는 선형 분포를 나타낼 수 있다.
- [166] 또한, 전자 장치(1000)가 워핑 계수를 1.2로 설정하는 경우, 워핑 함수에서 출력되는 출력 주파수의 분포는 위쪽으로 볼록한 모양의 주파수 분포를 나타낼 수 있다. 즉, 전자 장치(1000)는 복수 화자의 서로 다른 성도 길이(vocal tract length)를 반영하기 위해, bilinear rule 또는 piecewise linear rule을 따르는 워핑 함수를 워핑 계수에 기초하여 정의하고, 정의된 워핑 함수를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수축을 변환함으로써, 화자의 성도 길이 변이를 모델링할 수 있다.
- [167] 도 13은 일 실시 예에 따라 전자 장치가 주파수 축이 변환된 제1 오디오 신호를 이용하여 제2 오디오 신호를 재-합성(re-synthesize)하는 방법을 나타내는 도면이다.
- [168] S1302에서, 전자 장치(1000)는 주파수 축이 변환된(frequency warped)된 제1 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환 (Inverse Fast Fourier Transform, IFFT)할 수 있다. S1304에서, 전자 장치(1000)는 역 고속 푸리에 변환된 제1 오디오 신호의 주파수 스펙트럼을 시간 축 상에서 중첩(Overlap)하고 더함(Add)으로써 제1 오디오 신호가 재-합성된 제2 오디오 신호를 생성할 수 있다.
- [169] 예를 들어, 전자 장치(1000)는 소정의 윈도우 길이(window length)를 가지는 윈도우를 이용하여 오디오 신호를 분할함으로써, 프레임 단위로 구분되는 오디오 신호를 획득하고, 오디오 신호를 프레임 별로 주파수 영역으로 변환한다. 즉, 전자 장치는 오디오 신호를 프레임 별로 주파수 영역으로 변환하였기 때문에, 주파수 영역으로 변환된 오디오 신호의 주파수 스펙트럼을 시간 영역으로 역 변환할 경우, 시간 영역으로 변환된 오디오 신호의 주파수

스펙트럼을 중첩하고, 중첩된 시간 영역의 오디오 신호들을 더함으로써 재-합성된 오디오 신호를 획득할 수 있다.

- [170] 도 14는 일 실시 예에 따른 전자 장치의 블록도이다. 도 15는 또 다른 실시 예에 따른 전자 장치의 블록도이다.
- [171] 도 14에 도시된 바와 같이, 일 실시 예에 따른 전자 장치(1000)는 프로세서(1400) 및 메모리(1402)를 포함할 수 있다. 그러나 도시된 구성 요소가 모두 필수구성요소인 것은 아니다. 도시된 구성요소보다 많은 구성요소에 의해 전자 장치(1000)가 구현될 수도 있고, 그보다 적은 구성요소에 의해서도 전자 장치(1000)는 구현될 수 있다. 예를 들어, 도 15에 도시된 바와 같이, 일 실시 예에 따른 전자 장치(1000)는 사용자 입력부(1502), 통신부(1504), 마이크(1506), 디스플레이(1507), 스피커(1508)를 더 포함할 수도 있다.
- [172] 프로세서(1400)는 메모리(1402)내 하나 이상의 인스트럭션을 실행함으로써 전자 장치(1000)의 전반적인 동작을 제어한다. 예를 들어, 프로세서(1400)는 메모리(1402)에 저장된 하나 이상의 인스트럭션을 실행함으로써 사용자 입력부(1502), 통신부(1504), 마이크(1506), 디스플레이(1507), 스피커(1508) 등을 전반적으로 제어할 수 있다. 또한, 프로세서(1400)는 메모리(1402)에 저장된 하나 이상의 인스트럭션들을 실행함으로써 도 1 내지 도 12에 기재된 전자 장치(1000)의 기능을 수행할 수 있다.
- [173] 또한, 프로세서(1400)는 하나 또는 복수의 프로세서로 구성될 수 있고, 하나 또는 복수의 프로세서는 CPU, AP, DSP(Digital Signal Processor) 등과 같은 범용 프로세서, GPU, VPU(Vision Processing Unit)와 같은 그래픽 전용 프로세서 또는 NPU와 같은 인공지능(AI) 전용 프로세서일 수 있다. 일 실시 예에 의하면, 프로세서(1400)가 복수의 프로세서 또는 그래픽 전용 프로세서 또는 NPU와 같은 인공지능 전용 프로세서로 구현될 때, 복수의 프로세서 또는 그래픽 전용 프로세서 또는 NPU와 같은 인공지능 전용 프로세서 중 적어도 일부는 전자 장치(1000) 및 전자 장치(1000)와 연결된 다른 전자 장치 또는 복수의 서버에 탑재될 수도 있다.
- [174] 예를 들어, 프로세서(1400)는 하나 이상의 인스트럭션을 실행함으로써, 복수의 프레임 단위로 구분되는 오디오 신호를 획득하고, 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하며, 상기 결정된 필터 बैं크 별 에너지 성분을 평탄화(smoothing) 하고, 상기 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출하며, 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오 신호 내 상기 사용자의 음성을 인식할 수 있다.
- [175] 예를 들어, 프로세서(1400)는 하나 이상의 인스트럭션을 실행함으로써, 상기 프레임 단위로 구분하기 위한 윈도우의 윈도우 길이(window length)를 결정하고, 상기 결정된 윈도우 길이의 윈도우들을 소정의 윈도우 간격으로

중첩(overlap)하며, 상기 중첩된 윈도우들을 이용하여 상기 오디오 신호를 복수의 프레임 단위로 구분할 수 있다.

- [176] 또한, 프로세서(1400)는 분포된 필터 뱅크를 상기 오디오 신호의 주파수 스펙트럼에 적용하고, 상기 필터 뱅크가 적용된 상기 주파수 스펙트럼의 값을 로그 스케일로 변환하며, 상기 로그 스케일로 변환된 상기 주파수 스펙트럼의 값을 이용하여 상기 필터 뱅크 별 에너지 성분을 결정할 수 있다.
- [177] 또한, 프로세서(1400)는 균일하게 분포된 타겟 에너지 성분에 기초하여 상기 필터 뱅크 별 에너지 성분을 평탄화 하기 위한 평탄화 계수를 필터 뱅크 별로 학습하고, 상기 필터 뱅크 별로 학습된 평탄화 계수를 이용하여 상기 필터 뱅크 별 에너지 성분을 평탄화(smoothing)할 수 있다.
- [178] 또한, 프로세서(1400)는 오디오 신호의 주파수 스펙트럼의 주파수 성분 별 크기에 관한 히스토그램을 생성하고, 상기 생성된 히스토그램을, 상기 주파수 성분 별 크기가 균일하게 분포된 타겟 히스토그램으로 매핑하기 위한 매핑 함수를 결정하고, 상기 결정된 매핑 함수를 이용하여 상기 오디오 신호의 주파수 스펙트럼을 변환하고, 상기 변환된 오디오 신호의 주파수 스펙트럼에 상기 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 결정된 필터 뱅크 별 에너지 성분을 평탄화할 수도 있다.
- [179] 또 다른 실시 예에 의하면, 프로세서(1400)는 음성 인식 모델을 학습하기 위한 오디오 신호를 처리할 수도 있다. 예를 들어, 프로세서(1400)는 음성 인식 모델을 학습하기 위해 복수의 프레임 단위로 구분되는 제1 오디오 신호를 획득하고, 상기 프레임 단위로 구분되는 상기 제1 오디오 신호의 주파수 스펙트럼을 획득하며, 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 상기 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 이용하여 제2 오디오 신호를 생성하며, 상기 생성된 제2 오디오 신호로부터 특징 벡터를 추출할 수도 있다.
- [180] 또 다른 실시 예에 의하면, 프로세서(1400)는 재합성된 제2 오디오 신호로부터 추출된 특징 벡터를 이용하여 사용자의 음성을 인식하기 위한 음성 인식 모델을 학습시킬 수도 있다.
- [181] 일 실시 예에 의하면, 프로세서(1400)는 소정의 룸(Room) 내에서 상기 제2 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room Impulse) 필터를 상기 제2 오디오 신호에 적용하고, 룸 임펄스 필터가 적용된 상기 제2 오디오 신호로부터 특징 벡터를 추출할 수 있다.
- [182] 또한, 프로세서(1400)는 프레임 별로 랜덤하게 생성되는 워핑(warping) 계수를 결정하고, 결정된 워핑 계수에 기초하여, 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하기 위한 워핑 함수를 결정하며, 결정된 워핑 함수를 이용하여 상기 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환할 수도 있다.

- [183] 또한, 프로세서(1400)는 주파수 축이 변환된 제1 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환(Inverse Fast Fourier Transform)하고, 상기 역 고속 푸리에 변환된 제1 오디오 신호의 주파수 스펙트럼을 시간 축 상에서 중첩(overlap)함으로써 상기 제2 오디오 신호를 생성할 수도 있다.
- [184] 또한, 프로세서(1400)는 프레임 단위로 구분되는 상기 제2 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하고, 상기 결정된 필터 बैं크 별 에너지 성분을 평탄화(smoothing) 하며, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 상기 오디오 신호의 특징 벡터를 추출할 수 있다.
- [185] 일 실시 예에 의하면, 프로세서(1400)가 미리 학습 시키는 음성 인식 모델은 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 프레임 단위로 획득된 오디오 학습 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 상기 주파수 축이 변환된 오디오 학습 신호의 주파수 스펙트럼을 이용하여 재-합성(re-synthesize)된 오디오 학습 신호의 특징 벡터를 기초로 미리 학습될 수 있다.
- [186] 또한, 음성 인식 모델은 오디오 신호가 전달되는 룸(room)내에서 상기 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room impulses) 필터를 상기 재-합성된 오디오 학습 신호에 적용하고, 상기 룸 임펄스 필터가 적용된 상기 재-합성된 오디오 학습 신호에 기초하여 미리 학습될 수도 있다. 본 개시에 따른 전자 장치(1000)가 이용하는 음성 인식 모델의 구체적인 구조 및 학습 방법은 후술하는 도 17 및 도 18을 참조하여 보다 상세하게 설명하기로 한다.
- [187] 메모리(1402)는 전자 장치(1000)의 동작을 제어하기 위한 하나 이상의 인스트럭션(instruction)을 포함할 수 있다. 또한, 메모리(1402)는 오디오 신호 내 사용자의 음성을 인식하기 위한 하나 이상의 음성 인식 모델을 포함할 수도 있다. 또한, 메모리(1402)에 저장된 프로그램들은 그 기능에 따라 UI 모듈, 터치 스크린 모듈, 알림 모듈 등으로 분류될 수 있다.
- [188] 일 실시 예에 의하면, 메모리(1402)는 예를 들어, 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(RAM, Random Access Memory) SRAM(Static Random Access Memory), 롬(ROM, Read-Only Memory), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있으나, 이에 제한되지 않는다.
- [189] 사용자 입력부(1502)는 전자 장치의 동작을 제어하기 위한 사용자 입력을 수신할 수 있다. 예를 들어, 사용자 입력부(1502)는, 키 패드(key pad), 돔 스위치(dome switch), 터치 패드(접촉식 정전 용량 방식, 압력식 저항막 방식, 적외선 감지 방식, 표면 초음파 전도 방식, 적분식 장력 측정 방식, 피에조 효과

방식 등), 조그 휠, 조그 스위치 등을 포함할 수 있으나, 이에 한정되는 것은 아니다.

- [190] 통신부(1504)는 서버(2000)와의 통신을 위한 하나 이상의 통신 모듈을 포함할 수 있다. 예를 들어, 통신부(1504)는, 근거리 통신부 또는 이동 통신부 중 적어도 하나를 포함할 수 있다.
- [191] 근거리 통신부(short-range wireless communication unit)는, 블루투스 통신부, BLE(Bluetooth Low Energy) 통신부, 근거리 무선 통신부(Near Field Communication unit), WLAN(와이파이) 통신부, 지그비(Zigbee) 통신부, 적외선(IrDA, infrared Data Association) 통신부, WFD(Wi-Fi Direct) 통신부, UWB(ultra wideband) 통신부, Ant+ 통신부 등을 포함할 수 있으나, 이에 제한되는 것은 아니다.
- [192] 이동 통신부는, 이동 통신망 상에서 기지국, 외부의 단말, 서버 중 적어도 하나와 무선 신호를 송수신한다. 여기에서, 무선 신호는, 음성 호 신호, 화상 통화 호 신호 또는 문자/멀티미디어 메시지 송수신에 따른 다양한 형태의 데이터를 포함할 수 있다.
- [193] 마이크(1506)는 사용자의 음성을 포함하는 오디오 신호를 수신할 수 있다. 또한, 마이크는 사용자의 음성 외에 복수의 잡음 원들로부터 발생된 잡음 신호를 포함하는 오디오 신호를 수신할 수도 있다. 마이크(1506)는 획득한 오디오 신호를 프로세서(1400)로 전달함으로써, 음성 인식 모델에 의한 음성 인식이 수행되도록 할 수 있다.
- [194] 스피커(1508)는 오디오 신호 내 사용자의 음성을 인식한 결과를 오디오 신호로 출력할 수 있다. 예를 들어, 스피커(1508)는 전자 장치(1000)에서 수행되는 기능(예를 들어, 호신호 수신음, 메시지 수신음, 알람음)과 관련된 신호를 소리로 출력할 수 있다.
- [195] 디스플레이(1507)는 전자 장치(1000)에서 처리되는 정보를 표시 출력할 수 있다. 예를 들어, 디스플레이(1507)는, 오디오 신호 내 사용자의 음성을 인식한 결과를 텍스트 형식으로 표시할 수 있다. 디스플레이(1507)는, 전자 장치(1000)의 제어를 위한 인터페이스, 전자 장치(1000)의 상태 표시를 위한 인터페이스 등을 디스플레이할 수 있다.
- [196] 도 16은 일 실시 예에 따른 음성 인식 모델을 학습시키는 전자 장치의 구조를 설명하기 위한 도면이다.
- [197] 일 실시 예에 의하면, 전자 장치(1000)는 복수의 중앙 처리 장치(CPU)(1642, 1644)를 포함하는 CPU 클러스터(1640) 및 복수의 그래픽 처리 장치(GPU)(1622, 1624, 1626)를 포함하는 GPU 클러스터(1620)를 이용하여 음성 인식 모델을 학습시킬 수 있다. 일반적으로 중앙 처리 장치(CPU)가 오디오 신호에 대하여 VTLP 및 Acoustic Simulation을 적용하고, VTLP 및 Acoustic Simulation을 통하여 재-합성된 오디오 신호로부터 특징을 추출하는 데에는 많은 컴퓨팅 자원이 소모되기 때문에, 전자 장치(1000)는 복수의 CPU를 포함하는 클러스터를

이용하여 효과적으로 VTLP 및 Acoustic Simulation을 수행하고, VTLP 및 Acoustic Simulation을 통하여 재-합성된 오디오 신호로부터 특징 벡터를 추출할 수 있다.

- [198] 또한, 전자 장치(1000)는 복수의 중앙 처리 장치(CPU)에서 출력된 특징 벡터 및 상기 특징 벡터에 대응되는 텍스트 정보를 복수의 그래픽 처리 장치(GPU) 및 복수의 그래픽 처리 장치를 제어하기 위한 복수의 중앙 처리 장치(CPU)를 포함하는 GPU노드(1612)로 전송함으로써, GPU 클러스터 각각에 의해 음성 인식 모델을 병렬적으로 학습시킬 수 있다. 이하에서는, 복수의 중앙 처리 장치 및 그래픽 처리 장치를 이용하여 음성 인식 모델을 학습시키기 위한 과정을 구체적으로 설명하기로 한다.
- [199] 먼저, 전자 장치(1000)는 CPU 클러스터(1640)를 이용하여 복수의 오디오 신호들을 수신할 수 있다. 전자 장치(1000)가 수신하는 복수의 오디오 신호들은 랜덤하게 나누어진 시간 영역의 오디오 신호들일 수 있다. 예를 들어, 전자장치가 수신하는 ShardedTFRecords(1652)는 시간적 순서가 아닌, 랜덤하게 선택된 오디오 신호들을 포함할 수 있다. 또한, ShardedTFRecords(1652)는 복수의 화자들로부터 수신된, 소정의 프레임 단위로 나누어진(Sharded)된 오디오 신호들을 포함할 수 있다. 예를 들어, ShardedTFRecords(1652)는 A 화자로부터 수신된 "안녕하세요"와 같은 오디오 신호 및 B 화자로부터 수신된 "하이 빅스비"와 같은 오디오 신호들을 포함할 수 있다.
- [200] 전자 장치(1000)는 A 화자로부터 수신된 "안녕하세요"와 같은 오디오 신호 및 B 화자로부터 수신된 "하이 빅스비"와 같은 오디오 신호들을 랜덤하게 스플릿(split)하여, Example Queue(1654)에 임시로 저장할 수 있다. 전자 장치(1000)는 Example Queue(1654)에 저장된 스플릿(split)된 오디오 신호들을 CPU 클러스터(1640)내의 복수의 CPU들에 전달할 수 있다. 전자 장치(1000)는 Example Queue(1654)에 저장된 스플릿(split)된 오디오 신호들을 순차적으로 복수의 CPU들로 전달할 수도 있지만, QueueQuere에 쌓인 스플릿된 오디오 신호들을 무작위 순서로 CPU로 전달할 수도 있다.
- [201] CPU 클러스터(1640)내 각각의 CPU들은 오디오 신호들을 수신하고, 위평함수를 이용하여, 수신된 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하며, 변환된 주파수 축을 포함하는 주파수 스펙트럼을 역변환함으로써 재-합성된 오디오 신호를 획득(VTLP)할 수 있다. 또한, CPU 클러스터(1640)내 각각의 CPU들은 재-합성된 오디오 신호에 대하여 룸(Room)내 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room impulse) 필터(예컨대, 룸 임펄스 응답으로써 전달 경로 별 음향 특징을 나타내는 전달함수)를 적용함으로써 Acoustic Simulation을 적용할 수 있다.
- [202] 또한, CPU 클러스터(1640)내 각각의 CPU들은 Acoustic Simulation을 통하여 획득된 오디오 신호들의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하고, 결정된 필터

뱅크 별 에너지 성분이 평탄화된 주파수 스펙트럼으로부터 특징 벡터를 추출(Power-Function based MUD)할 수 있다. CPU 클러스터(1640)내 각각의 CPU들은 추출된 특징 벡터에 대응하는 텍스트 정보를 포함하는 Label(1662, 1664)로부터, 상기 특징 벡터에 대응하는 텍스트 정보를 획득하고, 획득된 텍스트 정보 및 추출된 특징 벡터를 함께 GPU 노드(1612) 내의 중앙 처리 장치(CPU, 1628)로 전송한다.

- [203] GPU 노드(1612) 내의 중앙 처리 장치(CPU, 1628)는 CPU 클러스터(1640)로부터 수신한 특징 벡터 및 상기 특징 벡터에 대응되는 텍스트 정보를 복수의 GPU로 전달함으로써, 복수의 GPU 각각에 의해 음성 인식 모델이 학습되도록 할 수 있다. 일 실시 예에 의하면, GPU 노드(1612) 내의 중앙 처리 장치(CPU, 1628)는 CPU 클러스터(1640)로부터 수신한 특징 벡터 및 상기 특징 벡터에 대응되는 텍스트 정보를 복수의 GPU로 전달하기 위한 우선 순위(Priority) 정보를 결정하고, 결정된 우선 순위 정보에 기초하여 특징 벡터 및 상기 특징 벡터에 대응되는 텍스트 정보를 복수의 GPU에 전달할 수 있다.
- [204] GPU 노드(1612) 내의 복수의 GPU들은 수신된 특징 벡터 및 텍스트 정보에 기초하여 음성 인식 모델을 학습시킬 수 있다. 예를 들어, 복수의 GPU들(1622, 1624, 1626)들은 특징 벡터 및 텍스트 정보에 기초하여 음성 인식 모델에 대한 inference 출력을 결정할 수 있고, 이를 통해 출력되는 출력 값과 정답 값을 비교하여 loss를 결정함으로써 음성 인식 모델을 학습시킬 수 있다. 특징 벡터 및 텍스트 정보에 기초하여 inference를 결정하는 과정은 후술하는 도 17 및 도 18의 음성 인식 모델 내 디코딩(decoding)과정에 대응될 수 있다. 전자 장치(1000)가 음성 인식 모델을 학습시키는 구체적인 방법은 후술하는 도 17 및 도 18을 참조하여 더 상세하게 설명하기로 한다.
- [205] 일 실시 예에 의하면, GPU 노드(1612) 내의 중앙 처리 장치들(CPU, 1628)는 GPU 노드 내 복수의 GPU들의 현재 상태를 결정하고, 복수의 GPU들의 상태에 관한 정보를 피드백 정보로써 CPU 클러스터(1640)로 전달할 수 있다. CPU 클러스터(1640)는 GPU 클러스터(1620)로부터 수신된 피드백 정보를 이용하여 복수의 CPU들이 출력한 특징 벡터 및 텍스트 정보의 전달 속도를 조절할 수 있다.
- [206] 즉, 본 개시에 따른 전자 장치(1000)는 중앙 처리 장치(CPU)보다 상대적으로 속도가 빠른 복수의 그래픽 처리 장치(GPU)들을 포함하고, 복수의 중앙 처리 장치 및 그래픽 처리 장치의 장치 비(device ratio)를 결정함으로써 효과적으로 음성 인식 모델을 학습시킬 수 있다. 전자 장치(1000)는 VTLP 및 Acoustic Simulation을 수행하지 않은 오디오 신호에서 특징 벡터를 추출하고, 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습하는 속도 값에 기초하여, 복수의 중앙 처리 장치 및 그래픽 처리 장치의 장치 비(device ratio)를 결정할 수 있다.
- [207] 도 17은 일 실시 예에 따른 전자 장치가 음성 인식 모델을 학습하는 과정을 나타내는 도면이다.

- [208] 일 실시 예에 의하면, 전자 장치(1000)는 오디오 신호 내 사용자의 음성을 인식하기 위해 음성 인식 모델을 미리 학습시킬 수 있다. 전자 장치(1000)가 이용하는 음성 인식 모델은 자동 음성 인식(Auto Speech Recognition, ASR) 모델로써, End-to-End 방식의 Attention based 신경망 모델일 수 있다. 그러나, 전자 장치(1000)가 이용하는 음성 인식 모델이 End-to-End 방식의 Attention based 신경망 모델에 한정되는 것은 아니며, 음성 인식 모델을 훈련하기 위해 필요한 기타 신경망 기반의 모델을 모두 포함할 수 있다. 전자 장치(1000)가 이용하는 음성 인식 모델은 Attention based 신경망 모델로써, 오디오 신호의 특징 벡터를 부호화 위한 복수의 LSTM(Long Short-Term Memory) 레이어 및 맥스 풀링(max-pooling) 레이어를 포함하는 인코더, 상기 부호화된 특징 벡터를 디코딩함으로써 소정의 텍스트를 출력하기 위한 LSTM 레이어를 포함하는 디코더, 상기 인코더 및 상기 디코더의 상관 관계에 기초하여 상기 인코더에서 출력되는 특징 벡터들에 가중치(attention)를 설정하기 위한 어텐션 레이어 및 상기 가중치에 기초하여 가중화된 특징 벡터가 특정 키워드에 관한 레이블(label)을 나타낼 확률을 출력하는 소프트 맥스 레이어를 포함할 수 있다. 그러나, 전술한 바와 같이, 전자 장치(1000)가 이용하는 음성 인식 모델은 도 17에 도시된 음성 인식 모델의 구조에 한정되는 것은 아니며, 음성 인식을 위한 기타 신경망 기반의 모델을 모두 포함할 수 있다.
- [209] 예를 들어, 전자 장치(1000)는 프레임 단위로 구분되는 오디오 신호 및 워핑 계수를 획득할 수 있다. 전자 장치(1000)는 사용자 또는 전자 장치와 연결되는 외부 디바이스로부터 워핑 계수(warping coefficient)를 획득할 수도 있지만, 전자 장치 스스로 오디오 신호의 프레임 별로 워핑 계수를 생성할 수도 있다. 전자 장치(1000)는 프레임 별로 생성된 워핑 계수를 이용하여 워핑 함수를 결정하고, 결정된 워핑 함수를 이용하여 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 주파수 축이 변환된 주파수 스펙트럼을 다시 시간 영역에서 중첩(Overlap) 및 더함(Add)으로써 재-합성된 오디오 신호를 생성할 수 있다.
- [210] 즉 전자 장치는 오디오 신호의 주파수 스펙트럼에 대하여, 복수 화자들의 성도 길이를 반영하기 위한 VTLP Processing(1702)을 수행한다. 또 다른 실시 예에 의하면, 전자 장치(1000)는 VTLP 처리를 통하여 재-합성된 오디오 신호에 대하여 Acoustic Simulation을 더 수행함으로써 음성 인식 모델 학습을 위한 다양한 데이터를 확보할 수도 있다.
- [211] 전자 장치(1000)는 VTLP를 통하여 재-합성된 오디오 신호로부터 특징 벡터를 추출한다. 예를 들어, 전자 장치(1000)는 상기 재-합성된 오디오 신호의 주파수 스펙트럼에 멜 스케일에 따라 분포된 필터 बैं크를 적용함으로써 필터 बैं크 별 에너지 성분을 결정하고, 결정된 필터 बैं크 별 에너지 성분에 평탄화 계수 1/15를 거듭 제곱의 지수로 하는 파워 함수를 적용함으로써 필터 बैं크 별 에너지 성분을 평탄화(smooth)할 수 있다. 또한, 전자 장치는 평탄화 계수 1/15를 거듭 제곱의 지수로 하는 파워 함수를 이용하여 평탄화된 필터 बैं크 별 에너지 성분에

- 기초하여 특징 벡터를 추출(1704)할 수 있다.
- [212] 전자 장치(1000)가 추출한 특징 벡터는 m 차원(dimension)의 주파수 특징을 포함하는 특징 벡터로써, 프레임 단위로 구성될 수 있다. 즉 하나의 프레임(1706)은 m 차원의 주파수 특징들을 포함하는 특징 벡터로 표현될 수 있다. 전자 장치(1000)는 프레임 단위로 추출된 특징 벡터를 순차로(sequential) 음성 인식 모델에 입력할 수 있다.
- [213] 음성 인식 모델의 엔코더(1720)는 복수의 LSTM 레이어 및 맥스 풀링 레이어를 포함할 수 있다. 엔코더(1720)는 순차로 입력된 복수의 프레임들에 non-linearity를 적용함으로써 프레임의 차원(dimension)을 변환하고, 변환된 차원(dimension)을 가지는 프레임들 중 소정의 비율로 프레임을 선택함으로써 입력된 프레임들 중 일부의 프레임들만을 출력할 수 있다.
- [214] 예를 들어, 엔코더(1720)가 4쌍의 LSTM 레이어(1721) 및 맥스 풀링 레이어(1722)를 포함하는 경우를 가정한다. LSTM 레이어(1721) 및 맥스 풀링 레이어(1722)는 엔코더 내 단일 레이어(Single layer)를 구성하고, 따라서, 엔코더(1720)는 총 4개의 단일 레이어를 포함할 수 있다. 예를 들어, 엔코더(1720)에 m 차원(dimension)을 가지는 프레임이 8개가 순차로 입력되는 경우, 각 프레임은 LSTM 레이어(172)를 통해 non-linearity가 적용됨으로써 m 차원과 다른 차원(dimension)으로 변환되고, 차원이 변환된 프레임들을 수신한 맥스 풀링 레이어(1722)는 2개의 프레임들의 오디오 특징들 중 더 큰 오디오 특징을 나타내는 프레임을 1/2비율로 선택한다.
- [215] 따라서, 총 4개의 레이어를 포함하는 엔코더(1720)는 m 차원의 8개 프레임을 수신하고, 수신된 m 차원의 8개 프레임들 중, m 차원과 다른 차원을 가지는 1개의 프레임을 출력할 수 있다. m 차원의 8개 프레임들을 입력 받는 엔코더(1720)에서 출력되는 프레임의 차원(1732)은 LSTM 레이어(1721)의 non-linearity 특성으로 인하여 알 수 없기 때문에 hidden dimension 상태일 수 있다. 또 다른 실시 예에 의하면, 40차원(dimension)의 1000프레임으로 구분되는 오디오 신호가 엔코더(1720)로 입력될 경우, 4개의 맥스 풀링 레이어를 통하여 8배 풀링이 이루어지므로, 엔코더(1720)는 1000/8개의 미지 차원의 프레임을 출력할 수도 있다.
- [216] 어텐션 레이어(1740)는 엔코더(1720) 및 상기 디코더(1760)의 상관 관계에 기초하여 상기 엔코더에서 출력되는 특징 벡터들에 가중치(attention)를 설정할 수 있다. 예를 들어, 어텐션 레이어(1740)는 엔코더(1720)에서 출력된 hidden dimension 상태의 특징 벡터(1732) 및 이전 시간의 디코더(1760)에서 출력된 hidden dimension 상태의 특징 벡터(1734)와 어텐션 레이어(1740)에서 결정된 어텐션 가중치(attention weight) 피드백을 이용하여 엔코더(1720)에서 출력된 특징 벡터들 사이의 가중치를 설정할 수 있다.
- [217] 예를 들어, 어텐션 레이어(1740)는 엔코더(1740)에서 출력된 이전 시간의 특징 벡터 및 디코더(1760)에서 출력된 이전 시간의 특징 벡터(1734)에 기초하여

어텐션 가중치 피드백을 결정할 수 있다. 어텐션 레이어(1740)는 어텐션 가중치 피드백을 이용하여 엔코더에서 출력된 특징 벡터들에 가중치를 적용하고, 가중치가 적용된 특징 벡터들을 가중합함으로써 컨텍스트 특징 벡터(1744)를 결정할 수 있다. 즉, 어텐션 레이어(1740)가 출력하는 컨텍스트 특징 벡터(1744)는 어텐션 가중치가 적용된 엔코더에서 출력된 프레임들의 가중합(weighted sum)일 수 있다. 또는 컨텍스트 특징 벡터(1744)는 어텐션 레이어가 결정한 어텐션 가중치에 기초하여 엔코더에서 출력된 프레임들 중 선택된 일부 프레임일 수 있다.

- [218] 디코더(1760)는 어텐션 레이어(1740)에서 출력된 가중합된 특징 벡터들을 디코딩한다. 예를 들어, 디코더(1760)는 하나의 LSTM 레이어를 포함할 수 있다. 디코더(1760)는 현재 시간의 컨텍스트 특징 벡터(1744), 어텐션 레이어(1740)에서 출력된 가중합된 특징 벡터, 디코더(1760)내 LSTM 자체의 Recurrent state(1762)에 기초하여, 이전 시간의 컨텍스트 특징 벡터(1742)에 non-linearity를 적용함으로써 이전 시간의 컨텍스트 특징 벡터의 차원(dimension)을 변환할 수 있다. 디코더(1760)에서 출력하는 특징 벡터의 차원은, LSTM의 non-linearity 특성으로 인하여 알 수 없기 때문에 hidden dimension 상태일 수 있다.
- [219] 소프트 맥스 레이어(1780)는 디코더(1760)에서 출력된 hidden dimension 상태의 특징 벡터가 소정의 텍스트 레이블에 대응될 확률 값을 결정할 수 있다. 예를 들어, 소프트 맥스 레이어(1780)는 디코더(1760)에서 이전 시간에 출력된 hidden dimension 상태의 특징 벡터, 어텐션 레이어(1740)에서 출력된 현재 시간의 컨텍스트 특징 벡터(1744) 및 소프트 맥스 레이어(1780)에서 이전 시간에 출력된 텍스트 정보를 입력 받고, 디코더(1760)에서 출력된 특징 벡터가 '하이 빅스비' 또는 '안녕하세요'와 같은 소정의 텍스트 레이블에 대응될 확률 값을 결정할 수 있다.
- [220] 일 실시 예에 의하면, 소프트 맥스 레이어(1780)는 어텐션 레이어(1740)에서 출력된 현재 시간의 컨텍스트 특징 벡터(1744) 및 소프트 맥스 레이어(1780)에서 이전 시간에 출력된 텍스트 정보를 입력 받지 않고도, 디코더(1760)에서 출력된 hidden dimension 상태의 특징 벡터가 소정의 텍스트 레이블에 대응될 확률 값을 결정할 수도 있다. 소프트 맥스 레이어(1780)는 텍스트 레이블(label)에 대한 확률 값들 중, 가장 높은 확률 값에 대응되는 텍스트 레이블의 텍스트 정보(1782)를 출력할 수 있다. 전술한 구조의 음성 인식 모델은 VTLP Processing을 통하여 재합성된 오디오 신호로부터 추출된 특징 벡터에 기초하여, 엔코더, 디코더, 어텐션 레이어 및 소프트 맥스 레이어들 간의 연결 강도에 관한 가중치를 수정(modify) 및 갱신(refine)함으로써 미리 학습될 수 있다.
- [221] 도 18은 일 실시 예에 따른 전자 장치가 필터뱅크 별 에너지 성분이 평탄화된 특징 벡터를 음성 인식 모델에 입력함으로써 음성 인식을 수행하는 과정을 나타내는 도면이다.

- [222] 전자 장치(1000)는 오디오 신호의 주파수 스펙트럼의 필터 뱅크 별 에너지 성분을 평탄화 하고, 필터 뱅크 별 에너지 성분이 평탄화된 주파수 스펙트럼으로부터 추출된 특징 벡터를 미리 학습된 음성 인식 모델에 입력함으로써, 사용자의 음성을 정확하게 인식할 수 있다.
- [223] 예를 들어, 전자 장치(1000)는 오디오 신호의 주파수 스펙트럼의 필터 뱅크 별 에너지 성분에 power-non linearity(1804)를 적용하기 위해, 필터 뱅크 별로 평탄화 계수를 결정하고, 평탄화 계수를 거듭 제곱의 지수로 하는 파워 함수를 이용하여 필터 뱅크 별 에너지 성분을 평탄화할 수 있다. 또 다른 실시 예에 의하면, 전자 장치(1000)는 오디오 신호의 주파수 스펙트럼의 주파수 성분 별 크기 또는 빈도수에 관한 히스토그램에 기초하여 결정되는 매핑 함수를 이용하여 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화 할 수도 있다. 전자 장치가 주파수 스펙트럼의 필터 뱅크 별 에너지 성분에 power-non linearity(1804)를 적용하는 과정은 도 4 내지 도 7의 과정에 대응될 수 있다.
- [224] 전자 장치(1000)는 필터 뱅크 별 에너지 성분이 평탄화된 주파수 스펙트럼으로부터 특징 벡터(1806)를 추출하고, 추출된 특징 벡터(1806)를 미리 학습된 음성 인식 모델에 입력할 수 있다. 전자 장치(1000)가 미리 학습하는 음성 인식 모델의 구조 및 동작방법은 도 17에서 전술한 음성 인식 모델의 구조 및 동작 방법에 대응될 수 있으므로 생략하기로 한다.
- [225] 도 19는 일 실시 예에 따른 서버의 블록도를 나타내는 도면이다.
- [226] 일 실시 예에 따르면, 서버(2000)는 통신부(2100), 데이터 베이스(Data Base, 2200) 및 프로세서(2300)를 포함할 수 있다.
- [227] 통신부(2100)는 도 15에 도시된 전자 장치(1000)의 통신부(1504)에 대응될 수 있다. 예를 들어, 통신부(2100)는 전자 장치(1000)로부터 오디오 신호를 수신하거나, 전자 장치로부터 수신된 오디오 신호에 대한 음성 인식 결과에 대한 정보를 전자 장치로 전송할 수 있다. 또한, 통신부(2100)는 다른 서버 내 통신부와 통신함으로써 전자 장치로부터 수신된 오디오 신호를 다른 서버와 송수신하거나, 공유된 오디오 신호에 대한 음성 인식 결과 정보를 송수신할 수도 있다.
- [228] 데이터 베이스(2200)는 도 15의 메모리(1402)에 대응될 수 있다. 예를 들어, 데이터 베이스(2200)는 음성 인식 모델을 저장할 수 있다. 또한, 데이터 베이스(2200)는 음성 인식 모델을 이용한 음성 인식 서비스를 제공하건, 음성 인식 모델을 학습하는데 필요한 데이터 Augmentation 알고리즘 및 특징(feature) 추출 알고리즘을 더 저장할 수도 있다.
- [229] 프로세서(2300)는 통상적으로 서버(2000)의 전반적인 동작을 제어한다. 예를 들어, 프로세서(2300)는, 서버(2000)의 DB(2200)에 저장된 프로그램들을 실행함으로써, DB(2200) 및 통신부(2100) 등을 전반적으로 제어할 수 있다. 또한, 프로세서(2300)는 DB(2100)에 저장된 프로그램들을 실행함으로써, 도 1 내지 도 18에서의 전자 장치(1000)의 동작의 일부를 수행할 수 있다.

- [230] 예를 들어, 프로세서(2300)는 전자 장치(1000)로부터 오디오 신호를 획득하고, 복수 화자들의 서로 다른 성도 길이 변화를 반영하기 위해, 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 주파수 축이 변환된 오디오 신호의 주파수 스펙트럼을 이용하여 재합성된 오디오 신호를 생성할 수 있다. 또한, 프로세서(2300)는 재합성된 오디오 신호에 대하여 림 임펄스 응답(RIR)을 적용함으로써 림 내 오디오 신호의 전달 경로 별 서로 다른 음향 특징을 반영할 수 있다.
- [231] 또한, 프로세서(2300)는 음성 인식 모델의 음성 인식 정확도를 향상하기 위해, 재합성된 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화하고, 평탄화된 필터 뱅크 별 에너지 성분을 포함하는 주파수 스펙트럼으로부터 특징 벡터를 추출할 수도 있다. 예를 들어, 프로세서(2300)는 필터 뱅크 별로, 평탄화 계수를 미리 학습하고, 학습된 평탄화 계수를 거듭제곱의 지수로 하는 파워 함수를 이용하여 오디오 신호의 필터 뱅크 별 에너지를 평탄화 할 수 있다. 또 다른 실시 예에 의하면, 프로세서(2300)는 오디오 신호의 주파수 스펙트럼의 주파수 성분 별 크기 또는 빈도수에 관한 히스토그램 및 상기 히스토그램의 매핑 함수를 이용하여 오디오 신호의 필터 뱅크 별 에너지 성분을 평탄화 할 수도 있다.
- [232] 도 20은 일 실시 예에 따라, 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.
- [233] S2002에서, 전자 장치(1000)는 제1 오디오 신호를 획득한다. S2004에서, 전자 장치(1000)는 획득한 제1 오디오 신호를 서버(2000)로 전송한다. 일 실시 예에 의하면, 전자 장치(1000)는 기 설정된 윈도우 길이를 가지는 윈도우를 이용하여 제1 오디오 신호를 프레임 단위로 구분하고, 프레임 단위로 구분되는 제1 오디오 신호를 서버(2000)로 전송할 수도 있다.
- [234] S2006에서, 서버(2000)는 수신된 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환(frequency warped)한다. 예를 들어 서버(2000)는 전자 장치로부터 미리 결정된 워핑 계수(warping coefficient)를 수신하고, 워핑 계수에 기초하여 상기 수학적 식 15 또는 16을 따르는 워핑 함수를 결정한다. 서버(2000)는 워핑 계수 및 워핑 함수를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환함으로써 복수 화자들의 성도 길이(vocal tract length) 변이를 반영할 수 있다.
- [235] S2008에서, 서버(2000)는 주파수 축이 변환된 오디오 신호의 주파수 스펙트럼을 이용하여 오디오 신호를 재-합성한다. 예를 들어, 서버(2000)는 주파수 축이 변환된 오디오 신호의 주파수 스펙트럼을 역 고속 푸리에 변환하고, 역 고속 푸리에 변환된 오디오 신호의 주파수 스펙트럼을 시간 축 상에서 중첩하여 더함으로써 재-합성된 오디오 신호를 생성(VTLP)할 수 있다. 일 실시 예에 의하면 서버(2000)는 재-합성된 오디오 신호에 대하여 RIR 필터를 적용함으로써 림 내 오디오 신호의 전달 경로 별 서로 다른 음향 특징을 나타내는 오디오 신호들을 더 생성할 수도 있다.
- [236] S2010에서, 서버(2000)는 재-합성된 오디오 신호로부터 특징 벡터를 추출한다.

일 실시 예에 의하면, 서버(2000)는 RIR 필터가 적용된 재-합성된 오디오 신호로부터 특징 벡터를 추출할 수도 있다. 또한, 일 실시 예에 의하면, S2010에서, 서버(2000)는 재-합성된 오디오 신호의 필터뱅크별 에너지 성분을 평탄화하고, 평탄화된 필터뱅크별 에너지 성분에 기초하여 특징 벡터를 추출할 수도 있다. 서버(2000)가 재-합성된 특징 벡터를 추출하는 구체적인 방법은 도 2 내지 도 7에서 전자 장치가 재-합성된 오디오 신호로부터 특징 벡터를 추출하는 과정에 대응될 수 있으므로 구체적인 설명은 생략하기로 한다.

[237] S2012에서, 서버(2000)는 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습시킨다. 일 실시 예에 의하면, 서버(2000)는, 서버(2000)가 재-합성된 오디오 신호의 필터뱅크별 에너지 성분을 평탄화하고, 평탄화된 필터뱅크별 에너지 성분에 기초하여 특징 벡터를 추출하는 경우, 필터뱅크별 에너지 성분이 평탄화된 오디오 신호로부터 특징 벡터를 추출하고, 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습시킬 수도 있다.

[238] 서버(2000)가 필터뱅크별 에너지 성분을 평탄화하고, 평탄화된 필터뱅크별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 음성 인식 모델(3000)을 학습하는 경우, 서버(2000)는 새로 획득된 오디오 신호 내 사용자 음성을 인식함에 있어, 새로 획득된 오디오 신호(122)의 멜 필터뱅크별 에너지 성분을 평탄화하고, 평탄화된 필터뱅크별 에너지 성분에 기초하여 추출된 특징 벡터(132)를 음성 인식 모델에 입력시킴으로써 사용자의 음성을 인식할 수도 있다. 서버(2000)가 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습하는 과정은 도 17 내지 도 18에서 전자 장치(1000)가 음성 인식 모델을 학습시키는 과정에 대응될 수 있다.

[239] S2014에서, 전자 장치(1000)는 사용자로부터 제2 오디오 신호를 획득할 수 있다. S2016에서, 전자 장치(1000)는 획득한 제2 오디오 신호를 서버(2000)로 전송한다. S2020에서, 서버(2000)는 필터뱅크별 에너지 성분이 평탄화된 제2 오디오 신호의 주파수 스펙트럼으로부터 특징 벡터를 추출한다. 서버(2000)가 특징 벡터를 추출하는 동작은 도 2의 S240에서 전자 장치가 특징 벡터를 추출하는 동작에 대응될 수 있다.

[240] 또한, 일 실시 예에 의하면, 서버(2000)가 제1 오디오 신호의 필터뱅크별 에너지 성분을 평탄화하고, 평탄화된 필터뱅크별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 학습된 음성 인식 모델을 이용하는 경우, S2020에서, 서버(2000)는 제2 오디오 신호의 주파수 스펙트럼의 필터뱅크별 에너지 성분을 평탄화(smoothing)하고, 평탄화된 필터뱅크별 에너지 성분을 포함하는 제2 오디오 신호로부터 특징 벡터를 추출할 수도 있다. 서버(2000)가 제2 오디오 신호의 주파수 스펙트럼의 필터뱅크별 에너지 성분을 평탄화 하는 동작은 도 4 내지 도 7에서 전자 장치(1000)가 오디오 신호의 필터뱅크별 에너지 성분을 평탄화 하는 과정에 대응될 수 있다.

[241] S2022에서, 서버(2000)는 추출된 특징 벡터를 미리 학습된 음성 인식 모델에

입력함으로써, 제2 오디오 신호 내의 사용자의 음성을 인식한다. S2024에서, 서버(2000)는 음성 인식 모델이 출력한 음성 인식 결과를 전자 장치(1000)로 전송한다. S2026에서, 전자 장치(1000)는 서버(2000)로부터 수신된 음성 인식 결과를 출력할 수 있다.

- [242] 도 21은 일 실시 예에 따라, 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.
- [243] 일 실시 예에 의하면, 본 개시에 따른 사용자의 음성을 인식하는 방법, 음성 인식 모델을 학습하기 위한 오디오 신호 처리 방법은 전자 장치 및 상기 전자 장치와 연동되는 복수의 서버에 의해 수행될 수도 있다.
- [244] S2012에서, 전자 장치(1000)는 제1 오디오 신호를 획득한다. S2104에서, 전자 장치(1000)는 제1 오디오 신호를 제1 서버(2500)로 전송한다. 또 다른 실시 예에 의하면, 전자 장치(1000)는 제1 오디오 신호를 제2 서버(2600)로 전송함으로써, 제2 서버(2600)를 경유하여 제1 오디오 신호를 제1 서버(2500)로 전송할 수도 있다.
- [245] S2106에서, 제1 서버(2500)는 미리 결정된 위핑 계수 및 수학식 15 또는 16을 따르는 위핑 함수를 이용하여 제1 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환(frequency warped)한다. S2108에서, 제1 서버(2500)는 주파수 축이 변환된 오디오 신호의 주파수 스펙트럼을 이용하여 오디오 신호를 재-합성함으로써 제1 오디오 신호에 대해 VTLP Processing을 수행할 수 있다. S2110에서, 제1 서버(2500)는 재-합성된 제1 오디오 신호를 제2 서버(2600)로 전송한다. 즉, 일 실시 예에 의하면, 도 21에 도시된 제1 서버(2500)는 네트워크 상 오디오 데이터를 증강시키기 위한 전용 서버일 수도 있다.
- [246] S2112에서, 제2 서버(2600)는 제1 서버(2500)로부터 수신된 재-합성된 제1 오디오 신호로부터 특징 벡터를 추출한다. 일 실시 예에 의하면, 제2 서버(2600)는 재-합성된 제1 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 특징 벡터를 추출할 수도 있다. 제2 서버(2600)가 재-합성된 제1 오디오 신호로부터 특징 벡터를 추출하는 동작은 도 20의 S2010에서 제1 서버가 재-합성된 오디오 신호로부터 특징 벡터를 추출하는 동작에 대응될 수 있다.
- [247] S2114에서, 제2 서버(2600)는 특징 벡터를 이용하여 음성 인식 모델을 학습시킨다. 일 실시 예에 의하면, 제2 서버(2600)가 재-합성된 제1 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 특징 벡터를 추출하는 경우, 제2 서버(2600)는 필터 बैं크 별 에너지 성분이 평탄화된 제1 오디오 신호로부터 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습시킬 수도 있다. 제2 서버(2600)가 특징 벡터를 이용하여 음성 인식 모델을 학습시키는 동작은 도 2의 서버(2000)가 S2012에서 음성 인식 모델을 학습시키는 동작에 대응될 수 있다.
- [248] S2116에서, 전자 장치(1000)는 제2 오디오 신호를 획득한다. S2118에서, 전자

장치(1000)는 제2 오디오 신호를 제2 서버(2600)로 전송한다. 일 실시 예에 의하면, 전자 장치(1000)는 제2 오디오 신호를 제1 서버(2500)를 통하여 제2 서버(2600)로 전송할 수도 있다. S2120에서, 제2 서버(2600)는 제2 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화 한다. S2120 은 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화하는 과정으로써, 도 4 내지 도 7의 전자 장치(1000)가 오디오 신호의 필터 बैंक 별 에너지 성분의 로그 우도가 최대화되는지 여부에 기초하여 평탄화 계수를 결정하고, 결정된 평탄화 계수를 거듭제곱의 지수로 하는 파워함수를 이용하여 필터 बैंक 별 에너지 성분을 평탄화 하는 과정에 대응될 수 있다.

- [249] S2122에서, 제2 서버(2600)는 필터 बैंक 별 에너지 성분이 평탄화된 제2 오디오 신호로부터 특징 벡터를 추출한다. 일 실시 예에 의하면, 제2 서버(2600)가 제1 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैंक 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 학습된 음성 인식 모델을 이용하는 경우, 제2 서버(2000)는 새로 획득된 제2 오디오 신호의 멜 필터 बैंक 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैंक 별 에너지 성분에 기초하여 추출된 특징 벡터를 음성 인식 모델에 입력시킴으로써 제2 오디오 신호 사용자의 음성을 인식할 수도 있다. S2124에서, 제2 서버(2600)는 추출된 특징 벡터를 음성 인식 모델에 입력함으로써, 제2 오디오 신호 내의 사용자의 음성을 인식한다. S2126에서, 제2 서버(2600)는 음성 인식 모델이 출력한 음성 인식 결과를 전자 장치(1000)로 전송한다. S2128에서, 전자 장치(1000)는 제2 서버(2600)로부터 수신된 음성 인식 결과를 출력할 수 있다.
- [250] 전술한 바와 같이, 사용자의 음성을 인식하는 방법은 획득된 오디오 신호를 증강(augmentation)시키는 제1 서버(2500) 및 상기 제1 서버로부터 증강(augmented)된 오디오 신호를 수신하고, 증강된 오디오 신호에 기초하여 음성 인식 모델을 학습하고, 학습된 음성 인식 모델을 이용하여 제2 오디오 신호 내 사용자의 음성을 인식하는 제2 서버(2600)를 포함하는 복수의 서버들을 이용하여 수행될 수도 있다.
- [251] 도 22은 일 실시 예에 따라 전자 장치 및 서버가 연동됨으로써 사용자의 음성을 인식하는 방법을 나타내는 도면이다.
- [252] 일 실시 예에 의하면, 본 개시에 따른 전자 장치는 복수의 서버와 연동함으로써 복수 화자의 음성을 정확하게 인식할 수 있다.
- [253] 예를 들어, S2202에서, 전자 장치(1000)는 어린이 음성을 포함하는 제1 오디오 신호를 획득한다. S2204에서, 전자 장치(1000)는 어린이 음성을 포함하는 제1 오디오 신호를 제1 서버(2500)로 전송한다. S2206에서, 제1 서버(2500)는 어린이 음성을 포함하는 제1 오디오 신호에 대하여 VTLP Processing을 수행한다. 제1 서버(2500)는 어린이 음성을 포함하는 제1 오디오 신호에 대하여 VTLP Processing을 수행하는 동작은 도 11 내지 도 12에서 전자 장치가 워핑 계수 및 상기 워핑 계수를 포함하는 워핑 함수를 이용하여 오디오 신호의 주파수

스펙트럼의 주파수 축을 변환하는 동작에 대응될 수 있다.

- [254] 일 실시 예에 의하면, 제1 서버(2500)는 VTLP Processing을 통해 재-합성된 제1 오디오 신호에 대하여 RIR 필터를 더 적용할 수도 있다. S2208에서, 제1 서버(2500)는 VTLP Processing을 통해 재-합성된 오디오 신호로부터 특징 벡터를 추출한다. 일 실시 예에 의하면, 제1 서버(2500)는 VTLP Processing을 통해 재-합성된 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 특징 벡터를 추출할 수도 있다.
- [255] S2210에서, 제1 서버(2500)는 추출된 특징 벡터를 이용하여 제1 음성 인식 모델을 학습시킬 수 있다. 일 실시 예에 의하면, 제1 서버(2500)는 RIR 필터가 적용된 VTLP Processing을 통해 재-합성된 오디오 신호에서 특징 벡터를 추출하고, 추출된 특징 벡터를 이용하여 제1 음성 인식 모델을 학습시킬 수도 있다. 또 다른 실시 예에 의하면, 제1 서버(2500)가 재-합성된 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 특징 벡터를 추출하는 경우, 제1 서버(2500)는 필터 बैं크 별 에너지 성분이 평탄화된 재-합성된 오디오 신호로부터 추출된 특징 벡터를 이용하여 음성 인식 모델을 학습시킬 수도 있다.
- [256] S2212에서, 제1 서버(2500)는 어린이 음성을 포함하는 제2 오디오 신호를 획득한다. S2214에서, 제1 서버(2500)는 어린이 음성을 포함하는 제2 오디오 신호를 제1 서버(2500)로 전송한다. S2216에서, 제1 서버(2500)는 제2 오디오 신호의 주파수 스펙트럼의 필터 बैं크 별 에너지 성분들이 Maximum Uniformity에 따라 분포되도록 하기 위하여, 필터 बैं크 별로 미리 학습되는 평탄화 계수를 거듭 제곱의 지수로 하는 파워 함수를 이용하여, 제2 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 한다.
- [257] S2218에서, 제1 서버(2500)는 추출된 특징 벡터를 제1 음성 인식 모델에 입력함으로써 어린이 음성을 포함하는 제2 오디오 신호 내 어린이의 음성을 인식한다. 일 실시 예에 의하면, 제1 서버(2500)가 어린이 음성을 포함하는 제1 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 학습된 음성 인식 모델을 이용할 경우, 제1 서버(2500)는 새로 획득된 어린이 음성을 포함하는 제2 오디오 신호의 필터 बैं크 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैं크 별 에너지 성분에 기초하여 추출된 특징 벡터를 음성 인식 모델에 입력시킴으로써 제2 오디오 신호 내 어린이 음성을 인식할 수도 있다. S2220에서, 제1 서버(2500)는 음성 인식 모델에서 출력되는 음성 인식 결과를 전자 장치(1000)로 전송한다. S2221에서, 전자 장치(1000)는 제1 서버(2500)로부터 수신된 음성 인식 결과를 출력한다.
- [258] S2222에서, 전자 장치(1000)는 어른 음성을 포함하는 제1 오디오 신호를 획득한다. S2224에서, 전자 장치(1000)는 어른 음성을 포함하는 제1 오디오 신호를 제2 서버(2600)로 전송한다. 일 실시 예에 의하면, 전자 장치(1000)는 어른

음성을 포함하는 제1 오디오 신호를 제1 서버(2500)를 통하여 제2 서버(2600)로 전송할 수도 있다. S2226에서, 제2 서버(2600)는 어른 음성을 포함하는 제1 오디오 신호에 대하여 VTLP Processing을 수행한다. 제2 서버(2600)가 어른 음성을 포함하는 제1 오디오 신호에 대하여 VTLP Processing을 수행하는 동작은 도 11 내지 도 12에서 전자 장치가 위평 계수 및 상기 위평 계수를 포함하는 위평 함수를 이용하여 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 동작에 대응될 수 있다.

- [259] 일 실시 예에 의하면, 제2 서버(2600)는 VTLP Processing을 통해 재-합성된 어른 음성을 포함하는 제1 오디오 신호에 대하여 RIR 필터를 더 적용할 수도 있다. S2228에서, 제2 서버(2600)는 VTLP Processing을 통해 재-합성된 어른 음성을 포함하는 제1 오디오 신호로부터 특징 벡터를 추출한다. 일 실시 예에 의하면, 제2 서버(2600)는 RIR 필터가 적용된 어른 음성을 포함하는 재-합성된 제1 오디오 신호에서 특징 벡터를 추출할 수도 있다. 또한, 도 22에는 도시되지 않았지만, S2228에서, 제2 서버(2600)는 VTLP Processing을 통해 재-합성된 어른 음성을 포함하는 제2 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैंक 별 에너지 성분에 기초하여 특징 벡터를 추출할 수도 있다. 즉, 제2 서버(2600)는 음성을 포함하는 제2 오디오 신호의 주파수 스펙트럼의 필터 बैंक 별 에너지 성분들이 Maximum Uniformity 에 따라 분포되도록 하기 위하여, 필터 बैंक 별로 미리 학습되는 평탄화 계수를 거듭 제곱의 지수로 하는 파워 함수를 이용하여, 제2 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화 할 수 있다.
- [260] S2230에서, 제2 서버(2500)는 추출된 특징 벡터를 이용하여 제2 음성 인식 모델을 학습시킬 수 있다. 일 실시 예에 의하면, 제2 서버(2500)는 필터 बैंक 별로 에너지 성분이 평탄화된, 어른 음성을 포함하는 제2 오디오 신호로부터 특징 벡터를 추출하고, 추출된 특징 벡터를 이용하여 제2 음성 인식 모델을 학습시킬 수도 있다.
- [261] S2232에서, 전자 장치(1000)는 어른 음성을 포함하는 제2 오디오 신호를 획득한다. S2234에서, 전자 장치(1000)는 어른 음성을 포함하는 제2 오디오 신호를 제2 서버(2600)로 전송한다. S2236에서, 제2 서버(2600)는 어른 음성을 포함하는 제2 오디오 신호로부터 특징 벡터를 추출할 수 있다. 일 실시 예에 의하면, 제2 서버(2600)는 제2 오디오 신호의 주파수 스펙트럼의 필터 बैंक 별 에너지 성분들이 Maximum Uniformity 에 따라 분포되도록 하기 위하여, 필터 बैंक 별로 미리 학습되는 평탄화 계수를 거듭 제곱의 지수로 하는 파워 함수를 이용하여, 어른 음성을 포함하는 제2 오디오 신호의 필터 बैंक 별 에너지 성분을 평탄화 하고, 필터 बैंक 별 에너지 성분이 평탄화된, 어른 음성을 포함하는 제2 오디오 신호로부터 특징 벡터를 추출할 수 있다.
- [262] S2238에서, 제2 서버(2600)는 특징 벡터를 제2 음성 인식 모델에 입력함으로써, 어른 음성을 포함하는 제2 오디오 신호 내 어른 음성을 인식할 수 있다. 일 실시

예에 의하면, 제2 서버(2600)가 어른 음성을 포함하는 제2 오디오 신호의 필터뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 학습된 음성 인식 모델을 이용할 경우, 제2 서버(2600)는 새로 획득된 어른 음성을 포함하는 제2 오디오 신호의 멜 필터뱅크 별 에너지 성분을 평탄화 하고, 평탄화된 필터뱅크 별 에너지 성분에 기초하여 추출된 특징 벡터를 음성 인식 모델에 입력시킴으로써 제2 오디오 신호 내 어른 음성을 인식할 수도 있다.

- [263] S2240에서, 제2 서버(2600)는 제2 음성 인식 모델에서 출력되는 음성 인식 결과를 전자 장치(1000)로 전송한다. S2242에서, 전자 장치(1000)는 제2 서버(2600)로부터 수신된 음성 인식 결과를 출력할 수 있다. 즉, 본 개시에 따른 전자 장치는, 서로 다른 화자들의 음성을 인식하기 위한 음성 인식 모델을 각각 포함하는 복수의 서버와 연동함으로써, 복수 화자의 음성을 정확하게 인식할 수도 있다.
- [264] 도 23은 일 실시 예에 따라 전자 장치가 목표 음성 신호 및 잡음 신호를 모두 이용하여 음성 인식 모델을 학습시키는 과정을 설명하기 위한 도면이다.
- [265] 일 실시 예에 의하면, 전자 장치(1000)는 음성 인식 모델이 목표 음성 신호 및 잡음 신호를 구별하여 인식하도록, 상기 음성 인식 모델을 학습시킬 수 있다. S2302에서, 전자 장치(1000)는 목표 음성 신호에 관한 소량의 근거리 음성 데이터 베이스(DB)를 획득할 수 있다. 근거리 음성 데이터 베이스(DB)는 다양한 음성 특징(발음 시간, 발성 주파수, 피치, 톤 등)으로 표현되는 목표 음성 신호에 관한 소량의 음성 신호들을 포함할 수 있다.
- [266] S2304에서, 전자 장치(1000)는 소량의 근거리 음성 데이터 베이스 내 소량의 목표 음성 신호들에 VTLP Processing을 수행할 수 있다. 전자 장치(1000)가 소량의 근거리 음성 데이터 베이스 내 목표 음성 신호에 관한 음성 신호들에 VTLP Processing을 수행하는 동작은 도 11 내지 도 12에서 전자 장치가 워핑 계수 및 상기 워핑 계수를 포함하는 워핑 함수를 이용하여 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 동작에 대응될 수 있다.
- [267] S2306에서, 전자 장치(1000)는 VTLP Processing을 통해 재-합성된 목표 음성 신호에 관한 음성 신호들에 대하여 RIR 필터를 더 적용할 수도 있다. 즉, 전자 장치(1000)는 원거리 환경을 포함하는 가상의 룸(Room) 내 오디오 신호의 전달 경로 별 음향 특징을 VTLP Processing을 통해 재-합성된 목표 음성 신호에 관한 음성 신호들에 적용함으로써, 가상의 room에서 전달되는 목표 음성 신호들을 모델링할 수 있다. 전자 장치(1000)는 전자 장치(1000)는 VTLP Processing을 통해 재-합성된 목표 음성 신호에 관한 음성 신호들에 대하여 RIR 필터를 더 적용함으로써 훈련용 목표 음성(2308)을 획득할 수 있다.
- [268] 전자 장치(1000)는 획득된 훈련용 목표 음성(2308)을 이용하여 전자 장치(1000)의 음성 인식 모델(3000)을 학습시킬 수 있다. 도 23에는 도시되지는 않았지만, 전자 장치(1000)는 훈련용 목표 음성(2308) 신호의 필터뱅크 별

에너지 성분을 평탄화 하고, 평탄화된 필터 बैंक 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 음성 인식 모델(3000)을 학습시킬 수도 있다.

- [269] S2314에서, 전자 장치(1000)는 소량의 근거리 음성 데이터 베이스 내, 잡음 신호들에 VTLP Processing을 수행할 수 있다. 전자 장치(1000)가 소량의 근거리 음성 데이터 베이스 내 잡음 신호에 관한 음성 신호들에 VTLP Processing을 수행하는 동작은 도 11 내지 도 12에서 전자 장치가 위평 계수 및 상기 위평 계수를 포함하는 위평 함수를 이용하여 오디오 신호의 주파수 스펙트럼의 주파수 축을 변환하는 동작에 대응될 수 있다.
- [270] S2314에서, 전자 장치(1000)는 VTLP Processing을 통해 재-합성된 잡음 신호에 관한 음성 신호들에 대하여 RIR 필터를 더 적용할 수도 있다. 즉, 전자 장치(1000)는 원거리 환경을 포함하는 가상의 룸(Room) 내 잡음 신호의 전달 경로 별 음향 특징을 VTLP Processing을 통해 재-합성된 잡음 신호에 관한 음성 신호들에 적용함으로써, 가상의 room에서 전달되는 잡음 신호들을 모델링할 수 있다. 전자 장치(1000)는 VTLP Processing을 통해 재-합성된 잡음 신호에 관한 음성 신호들에 대하여 RIR 필터를 더 적용함으로써 훈련용 잡음 음성(2318)을 획득할 수 있다.
- [271] 전자 장치(1000)는 획득된 훈련용 잡음 음성(2318)을 이용하여 전자 장치(1000)의 음성 인식 모델(3000)을 학습시킬 수 있다. 도 23에는 도시되지는 않았지만, 전자 장치(1000)는 훈련용 잡음 음성(2318) 신호의 필터 बैंक 별 에너지 성분을 평탄화 하고, 평탄화된 필터 बैंक 별 에너지 성분에 기초하여 추출된 특징 벡터를 이용하여 음성 인식 모델(3000)을 학습시킬 수도 있다.
- [272] 전자 장치(1000)는 소량의 목표 음성 신호들 및 소량의 잡음 신호들을 증강(augmentation)시킴으로써 획득된 훈련용 목표 음성(2308) 및 훈련용 잡음 음성(2318)에 기초하여 학습된 음성 인식 모델(3000)을 이용하여, 사용자의 음성을 정확하게 인식할 수 있다. 예를 들어, 전자 장치(1000)는 사용자(2332)로부터 목표 음성 신호(2322) 및 잡음원(예컨대, TV, 스피커 등, 2334)으로부터 잡음 신호가 함께 획득되는 경우에도, 훈련용 목표 음성(2308) 및 훈련용 잡음 음성(2318)을 기초로 학습된 음성 인식 모델(3000)을 이용하여 잡음 신호와 사용자의 목표 음성 신호를 정확하게 구분함으로써, 사용자의 음성을 정확하게 인식할 수 있다.
- [273] 일 실시예에 따른 전자 장치가 사용자의 음성을 인식하는 방법, 음성 인식 모델을 학습하기 위한 오디오 신호 처리 방법 및 전자 장치가 음성 인식 모델을 학습시키는 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 또한, 상기 전자 장치가 상기 전자 장치에

표시되는 콘텐츠의 적어도 일부를 통하여, 상기 전자 장치에 관한 가상 장치를 제공하는 방법을 수행하도록 하는 프로그램이 저장된 기록 매체를 포함하는 컴퓨터 프로그램 제품이 제공될 수 있다.

- [274] 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [275] 일부 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다.
- [276] 예를 들어, 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적 저장매체'는 실재(tangible)하는 장치이고, 신호(signal)(예: 전자기파)를 포함하지 않는다는 것을 의미할 뿐이며, 이 용어는 데이터가 저장매체에 반영구적으로 저장되는 경우와 임시적으로 저장되는 경우를 구분하지 않는다. 예로, '비일시적 저장매체'는 데이터가 임시적으로 저장되는 버퍼를 포함할 수 있다.
- [277] 일 실시예에 따르면, 본 문서에 개시된 다양한 실시예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로 배포되거나, 또는 어플리케이션 스토어(예: 플레이 스토어™)를 통해 또는 두개의 사용자 장치들(예: 스마트폰들) 간에 직접, 온라인으로 배포(예: 다운로드 또는 업로드)될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품(예: 다운로드를 앱(downloadable app))의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 기기로 읽을 수 있는 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.
- [278] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속한다.

청구범위

- [청구항 1] 사용자의 음성을 인식하는 방법에 있어서,
 복수의 프레임 단위로 구분되는 오디오 신호를 획득하는 단계;
 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기
 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별
 에너지 성분을 결정하는 단계;
 상기 결정된 필터 뱅크 별 에너지 성분을 평탄화(smoothing) 하는 단계;
 상기 평탄화된 필터 뱅크 별 에너지 성분에 기초하여 상기 오디오 신호의
 특징 벡터를 추출하는 단계; 및
 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오
 신호 내 상기 사용자의 음성을 인식하는 단계; 를 포함하는, 방법.
- [청구항 2] 제1항에 있어서, 상기 오디오 신호를 획득하는 단계는
 상기 프레임 단위로 구분하기 위한 윈도우의 윈도우 길이(window
 length)를 결정하는 단계;
 상기 결정된 윈도우 길이의 윈도우들을 소정의 윈도우 간격으로
 중첩(overlap)하는 단계; 및
 상기 중첩된 윈도우들을 이용하여 상기 오디오 신호를 복수의 프레임
 단위로 구분하는 단계; 를 포함하는, 방법.
- [청구항 3] 제1항에 있어서, 상기 필터 뱅크 별 에너지 성분을 결정하는 단계는
 상기 분포된 필터 뱅크를 상기 오디오 신호의 주파수 스펙트럼에
 적용하는 단계;
 상기 필터 뱅크가 적용된 상기 주파수 스펙트럼의 값을 로그 스케일로
 변환하는 단계; 및
 상기 로그 스케일로 변환된 상기 주파수 스펙트럼의 값을 이용하여 상기
 필터 뱅크 별 에너지 성분을 결정하는 단계; 를 포함하는, 방법.
- [청구항 4] 제1항에 있어서, 상기 평탄화 하는 단계는
 균일하게 분포된 타겟 에너지 성분에 기초하여 상기 필터 뱅크 별 에너지
 성분을 평탄화 하기 위한 평탄화 계수를 필터 뱅크 별로 학습하는 단계;
 및
 상기 필터 뱅크 별로 학습된 평탄화 계수를 이용하여 상기 필터 뱅크 별
 에너지 성분을 평탄화(smoothing) 하는 단계; 를 포함하는, 방법.
- [청구항 5] 제1항에 있어서, 상기 평탄화 하는 단계는
 상기 오디오 신호의 필터 뱅크 별 에너지 성분의 크기에 관한
 히스토그램을 생성하는 단계;
 상기 생성된 히스토그램을, 상기 필터 뱅크 별 에너지 성분의 크기가
 균일하게 분포된 타겟 히스토그램으로 매핑하기 위한 매핑 함수를
 결정하는 단계; 및

상기 결정된 매핑 함수를 이용하여 상기 오디오 신호의 필터 뱅크 별 에너지 성분을 변환함으로써 상기 필터 뱅크 별 에너지 성분을 평탄화하는 단계; 를 포함하는, 방법.

[청구항 6] 제1항에 있어서, 상기 오디오 신호의 특징 벡터를 추출하는 단계는 상기 평탄화된 필터 뱅크 별 에너지 성분을 이산 코사인 변환(Discrete Cosine Transform)함으로써 DCT 계수를 결정하는 단계; 및 상기 결정된 DCT 계수들 중 적어도 하나의 계수를 원소로 포함하는 상기 특징 벡터를 추출하는 단계; 를 포함하는, 방법.

[청구항 7] 제1항에 있어서, 상기 음성 인식 모델은 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기 위해, 프레임 단위로 획득된 오디오 학습 신호의 주파수 스펙트럼의 주파수 축을 변환하고, 상기 주파수 축이 변환된 오디오 학습 신호의 주파수 스펙트럼을 이용하여 재-합성(re-synthesize)된 오디오 학습 신호의 특징 벡터를 기초로 미리 학습되는 것인, 방법.

[청구항 8] 제7항에 있어서, 상기 오디오 학습 신호의 주파수 스펙트럼의 주파수 축은, 프레임 별로 랜덤하게 생성된 위핑 계수 및 상기 위핑 계수에 기초하여 상기 오디오 학습 신호의 주파수 스펙트럼의 주파수 축을 변환하기 위한 위핑 함수에 기초하여 변환되는 것인, 방법.

[청구항 9] 제7항에 있어서, 상기 음성 인식 모델은 상기 오디오 신호가 전달되는 룸(room)내에서 상기 오디오 신호의 전달 경로 별 음향 특징을 나타내는 룸 임펄스(Room impulses) 필터를 상기 재-합성된 오디오 학습 신호에 적용하고, 상기 룸 임펄스 필터가 적용된 상기 재-합성된 오디오 학습 신호에 기초하여 미리 학습되는 것을 특징으로 하는, 방법.

[청구항 10] 제7항에 있어서, 상기 재-합성된 오디오 학습 신호는, 상기 주파수 축이 변환된 오디오 학습 신호의 주파수 스펙트럼을 역 고속 푸리에 변환(Inverse Fast Fourier Transform)하고, 상기 역 고속 푸리에 변환된 오디오 학습 신호의 주파수 스펙트럼을 시간 축 상에서 중첩(overlap)함으로써 생성되는 것을 특징으로 하는 방법.

[청구항 11] 사용자 음성을 인식하는 전자 장치에 있어서, 하나 이상의 인스트럭션을 저장하는 메모리; 및 상기 하나 이상의 인스트럭션을 실행하는 프로세서; 를 포함하고, 상기 프로세서는 상기 하나 이상의 인스트럭션을 실행함으로써, 복수의 프레임 단위로 구분되는 오디오 신호를 획득하고, 상기 프레임 단위로 구분되는 상기 오디오 신호의 주파수 스펙트럼에 기 설정된 스케일에 따라 분포된 필터 뱅크를 적용함으로써 필터 뱅크 별 에너지 성분을 결정하고,

상기 결정된 필터뱅크별 에너지 성분을 평탄화(smoothing) 하고,
 상기 평탄화된 필터뱅크별 에너지 성분에 기초하여 상기 오디오 신호의
 특징 벡터를 추출하고,
 상기 추출된 특징 벡터를 음성 인식 모델에 입력함으로써 상기 오디오
 신호 내 상기 사용자의 음성을 인식하는 전자 장치.

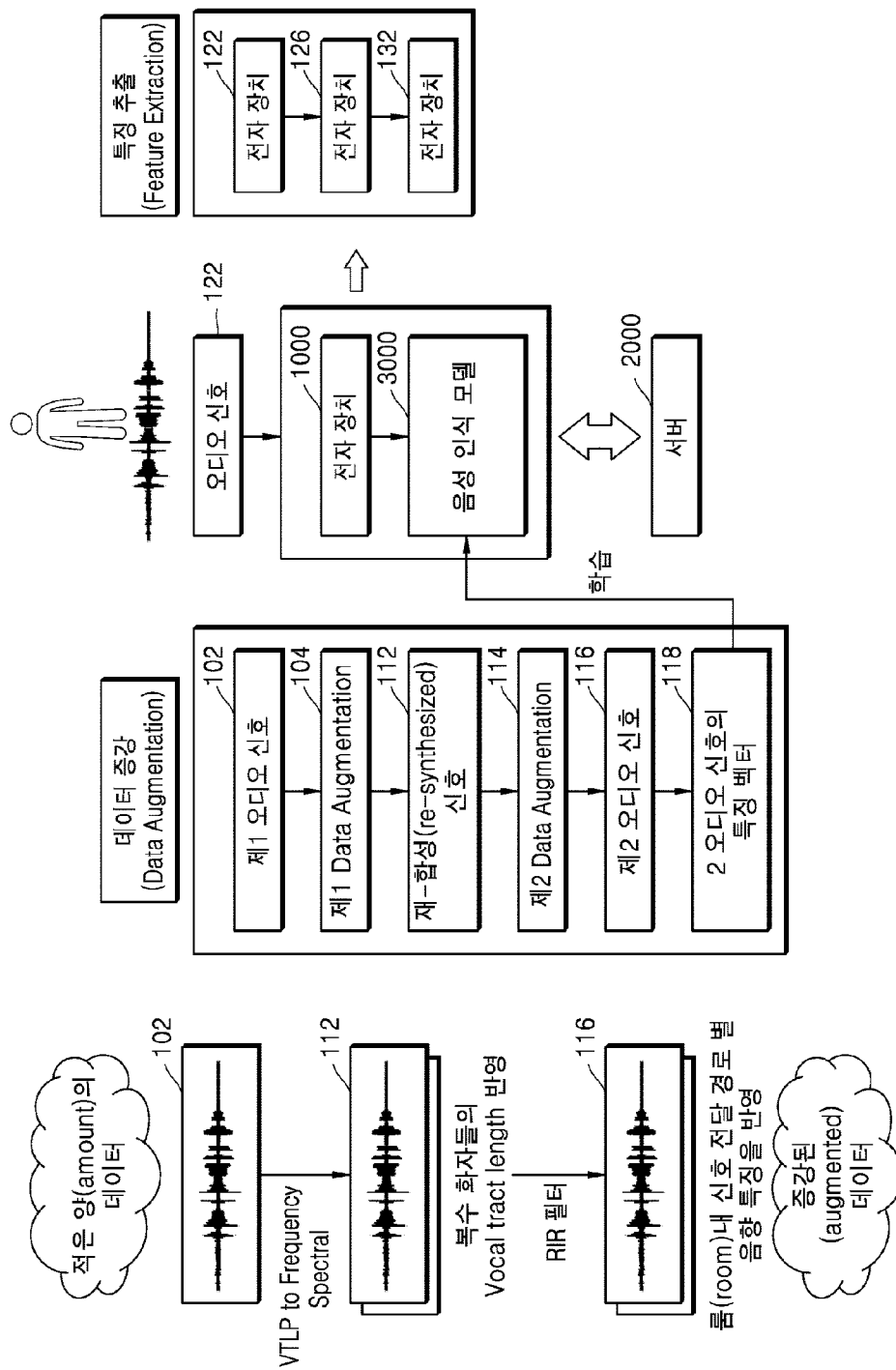
[청구항 12] 제11항에 있어서, 상기 프로세서는 상기 하나 이상의 인스트럭션을
 실행함으로써,
 균일하게 분포된 타겟 에너지 성분에 기초하여 상기 필터뱅크별 에너지
 성분을 평탄화 하기 위한 평탄화 계수를 필터뱅크별로 학습하고,
 상기 필터뱅크별로 학습된 평탄화 계수를 이용하여 상기 필터뱅크별
 에너지 성분을 평탄화(smoothing) 하는 전자 장치.

[청구항 13] 제11항에 있어서, 상기 프로세서는 상기 하나 이상의 인스트럭션을
 실행함으로써,
 상기 오디오 신호의 필터뱅크별 에너지 성분의 크기에 관한
 히스토그램을 생성하고,
 상기 생성된 히스토그램을, 상기 필터뱅크별 에너지 성분의 크기가
 균일하게 분포된 타겟 히스토그램으로 매핑하기 위한 매핑 함수를
 결정하고,
 상기 결정된 매핑 함수를 이용하여 상기 오디오 신호의 필터뱅크별
 에너지 성분을 변환함으로써 상기 필터뱅크별 에너지 성분을 평탄화
 하는 전자 장치.

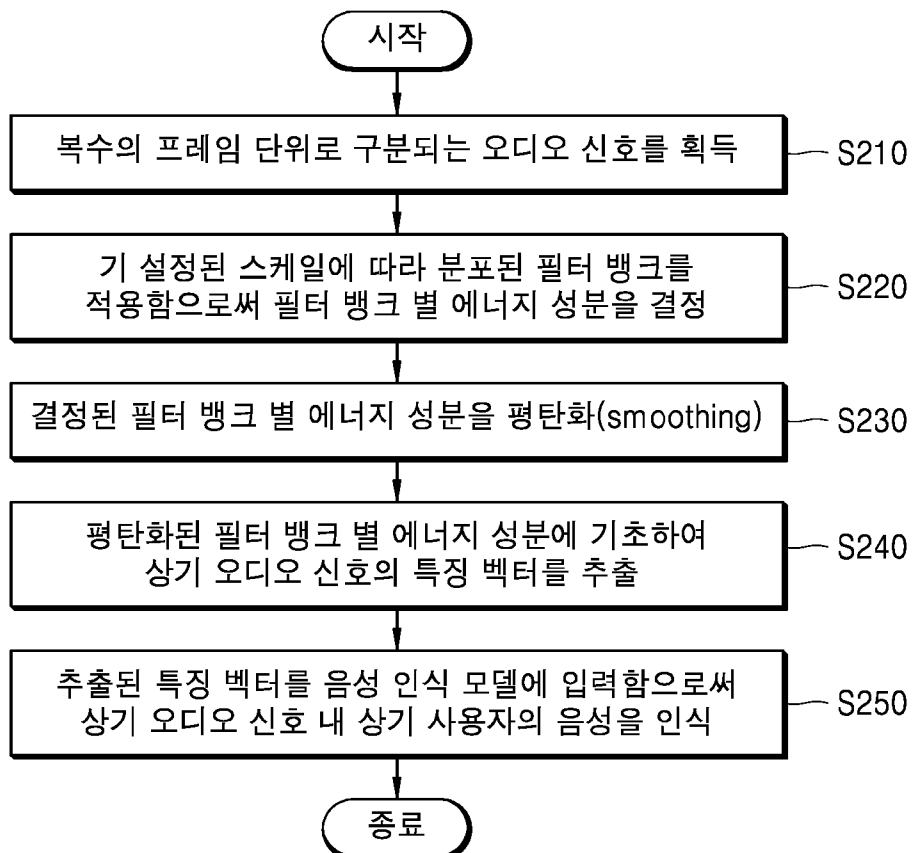
[청구항 14] 제11항에 있어서, 상기 음성 인식 모델은
 복수 화자들의 서로 다른 성도 길이(Vocal tract length) 변화를 반영하기
 위해, 프레임 단위로 획득된 오디오 학습 신호의 주파수 스펙트럼의
 주파수 축을 변환하고, 상기 주파수 축이 변환된 오디오 학습 신호의
 주파수 스펙트럼을 이용하여 재-합성(re-synthesize)된 오디오 학습 신호의
 특징 벡터를 기초로 미리 학습되는 것을 특징으로 하는 전자 장치.

[청구항 15] 제1항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한
 컴퓨터로 읽을 수 있는 기록 매체.

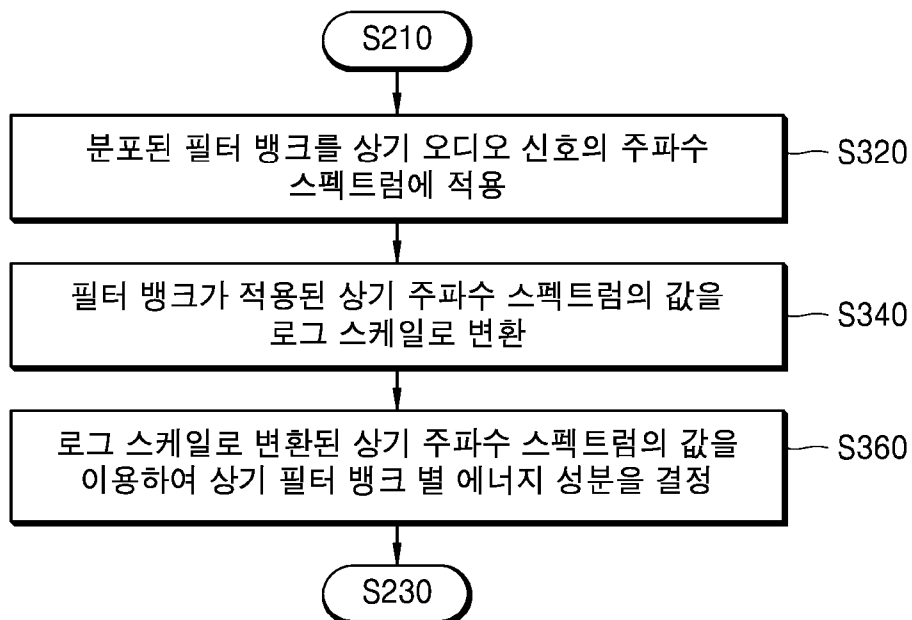
[도 1]



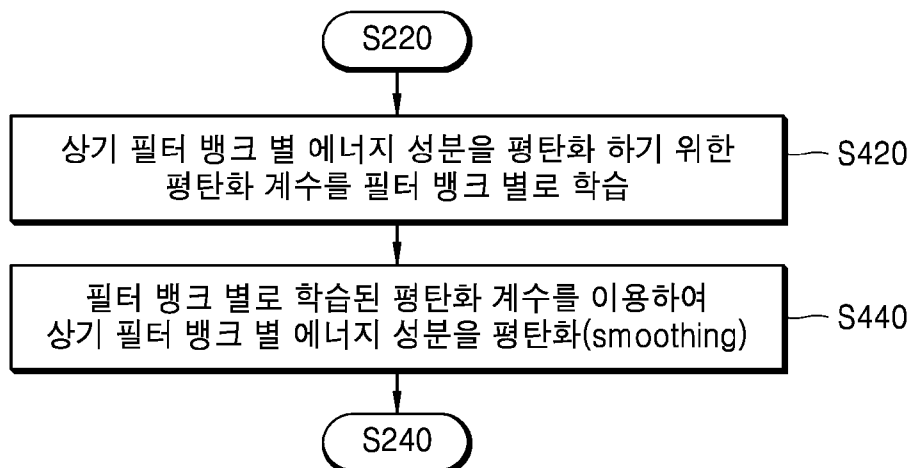
[도2]



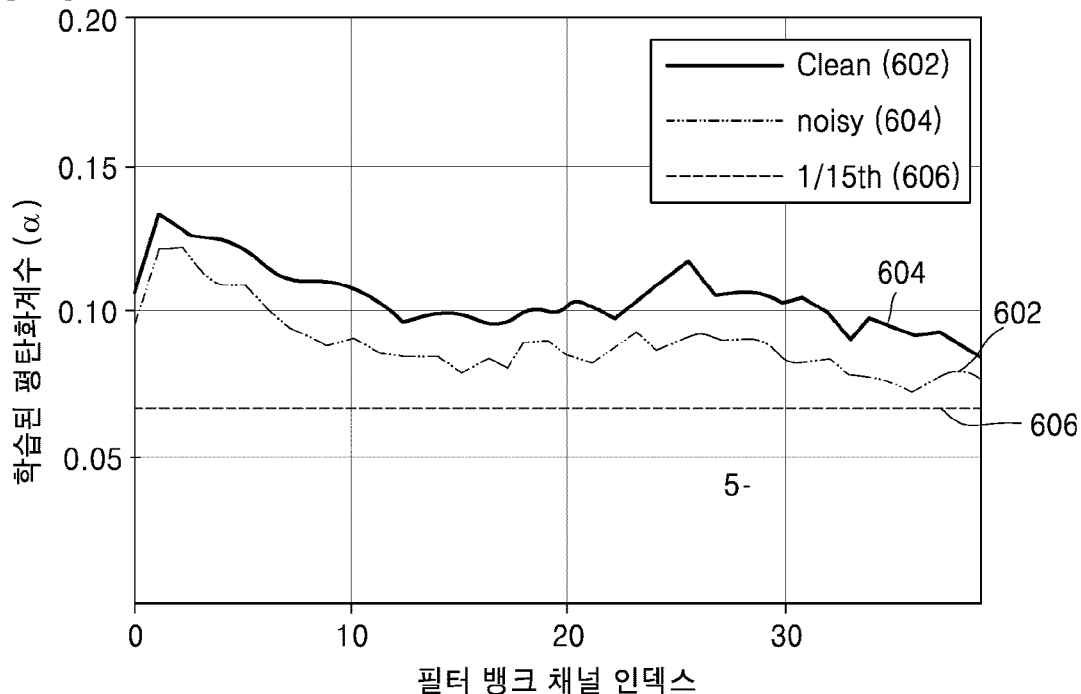
[도3]



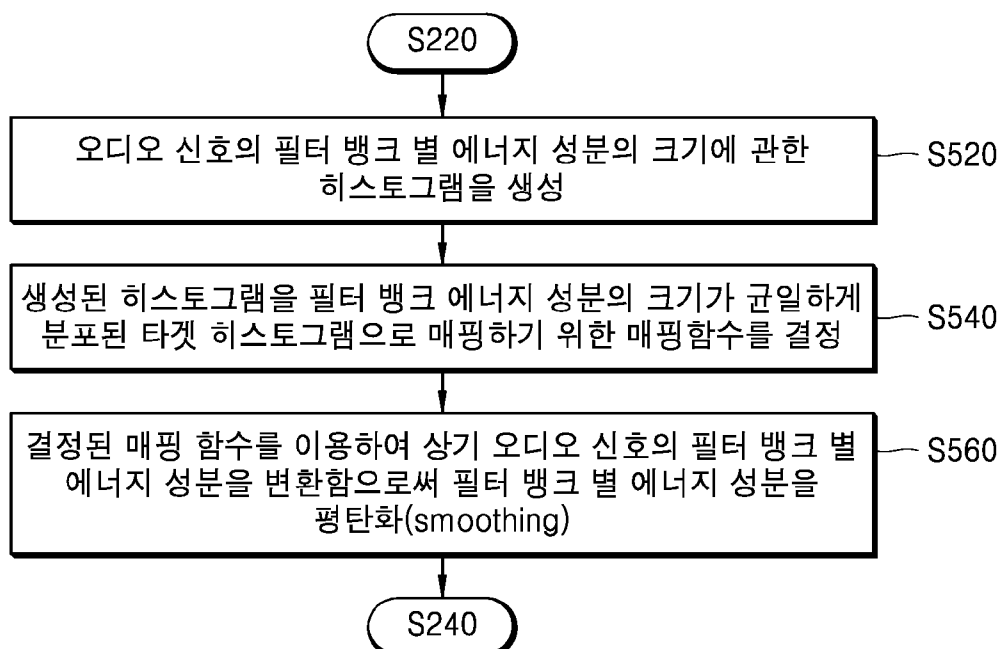
[도4]



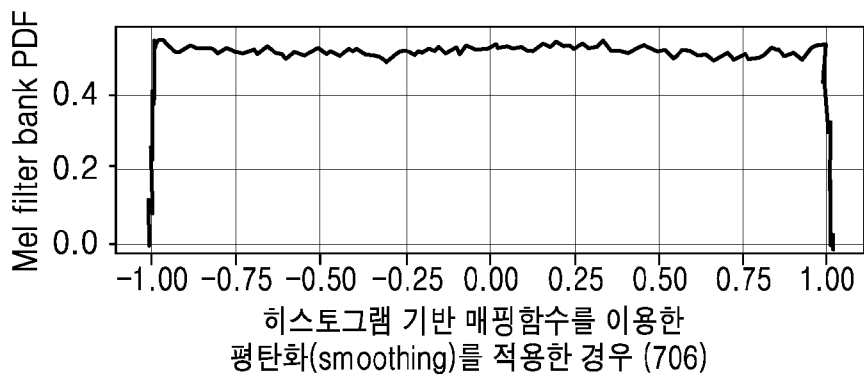
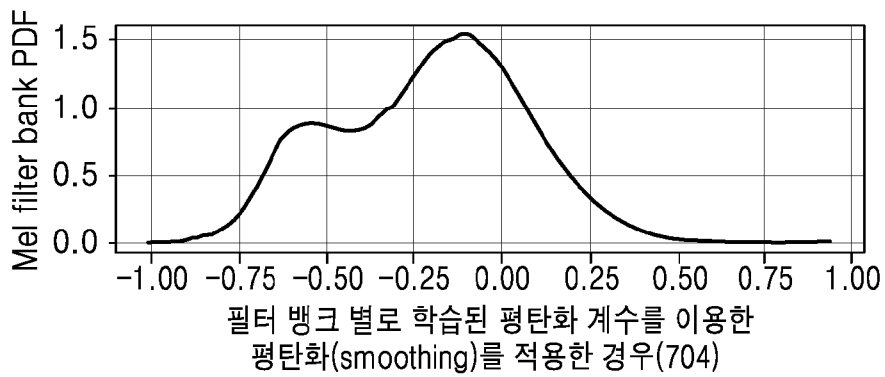
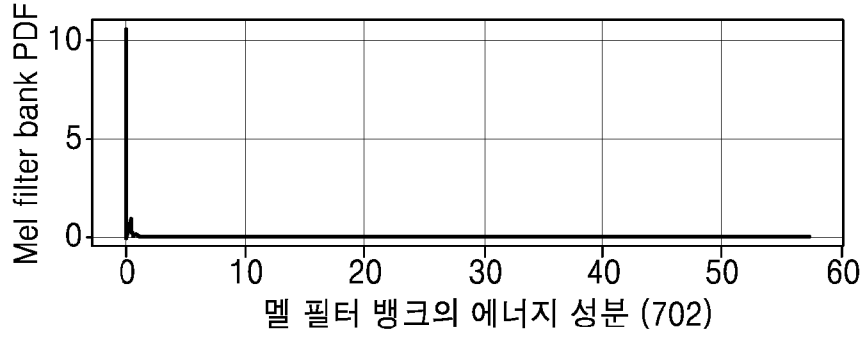
[도5]



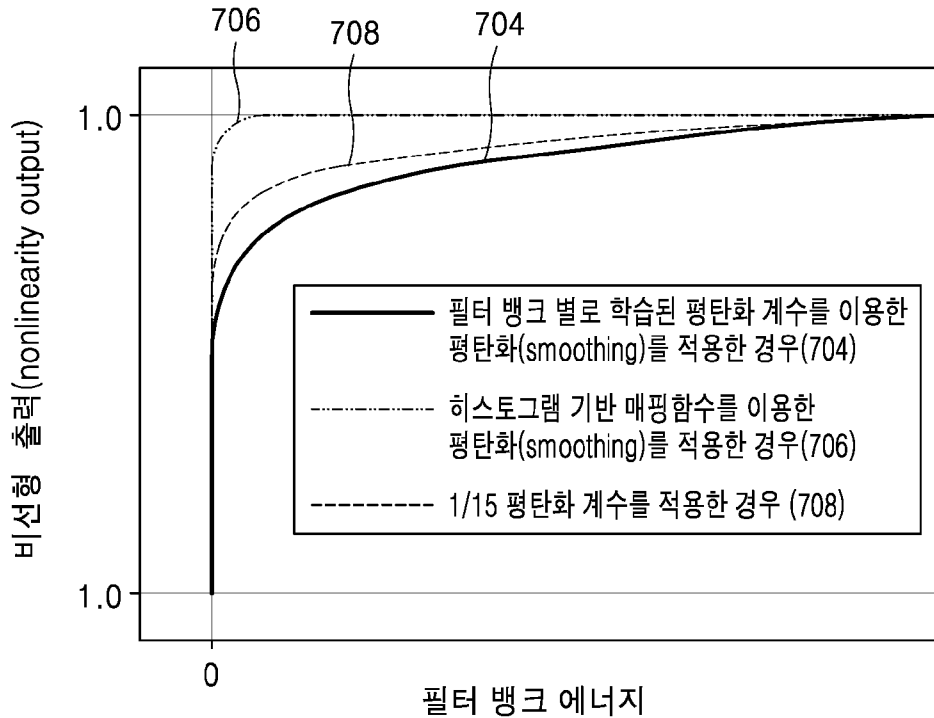
[도6]



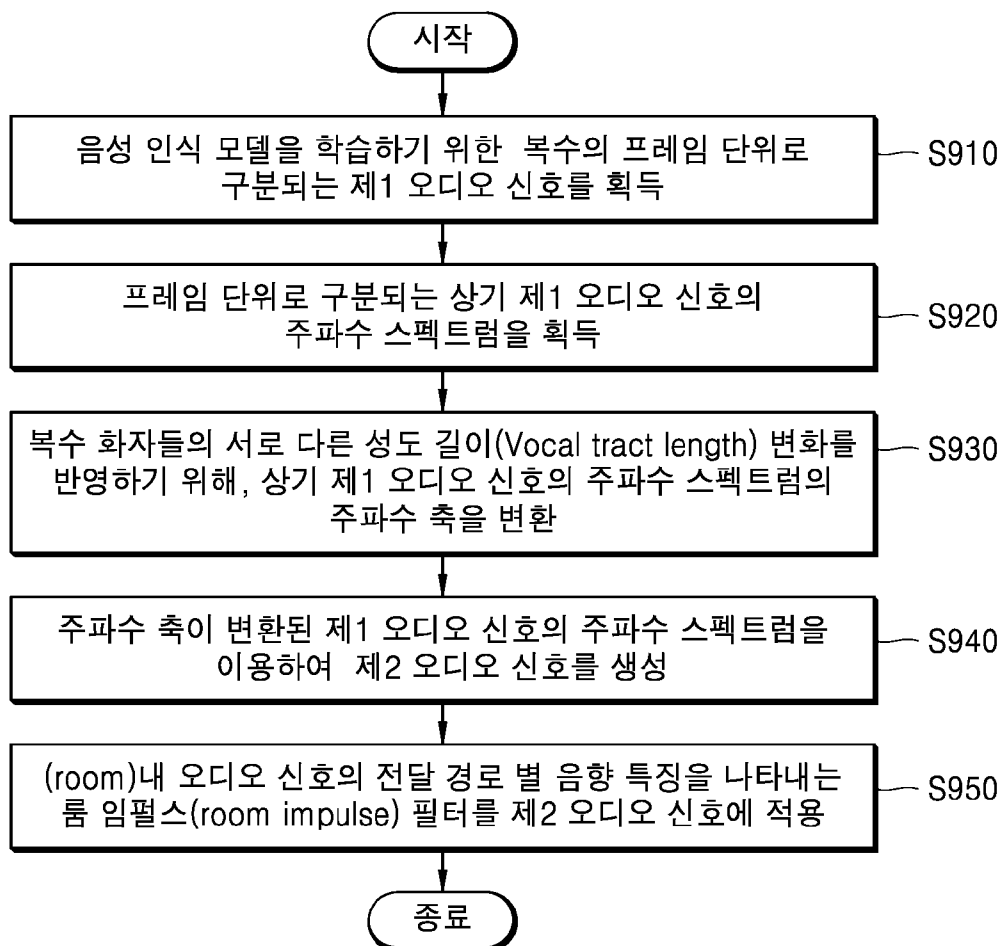
[도7]



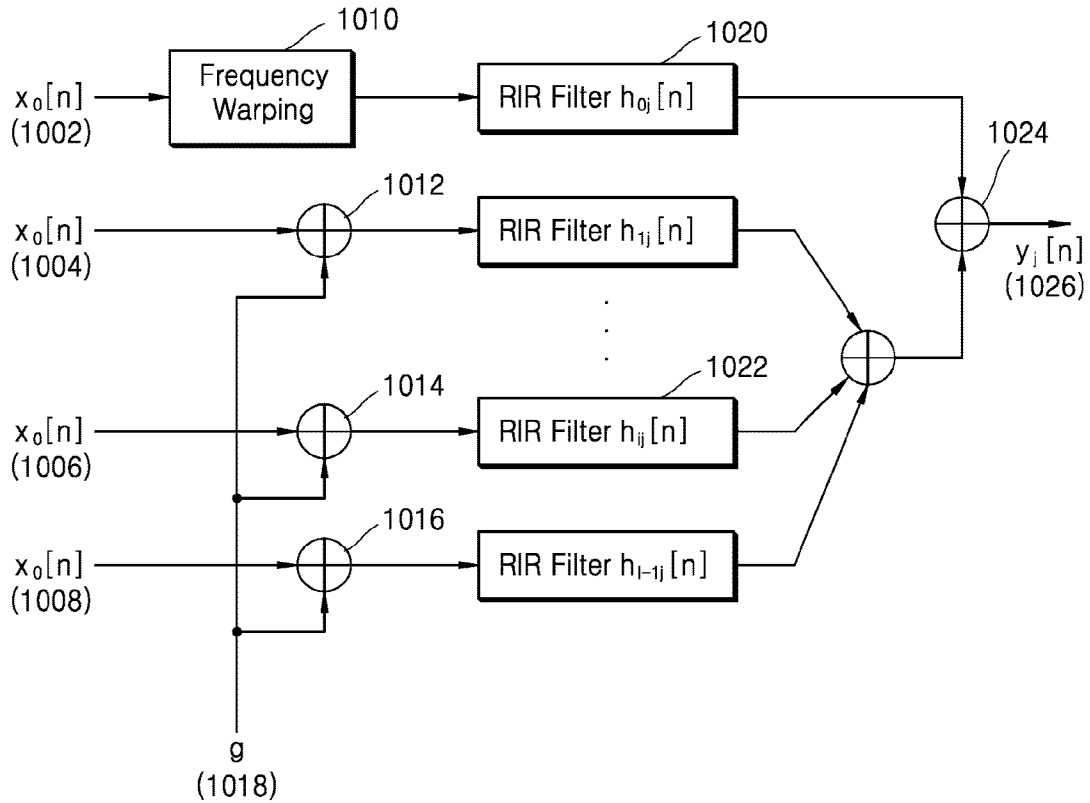
[도8]



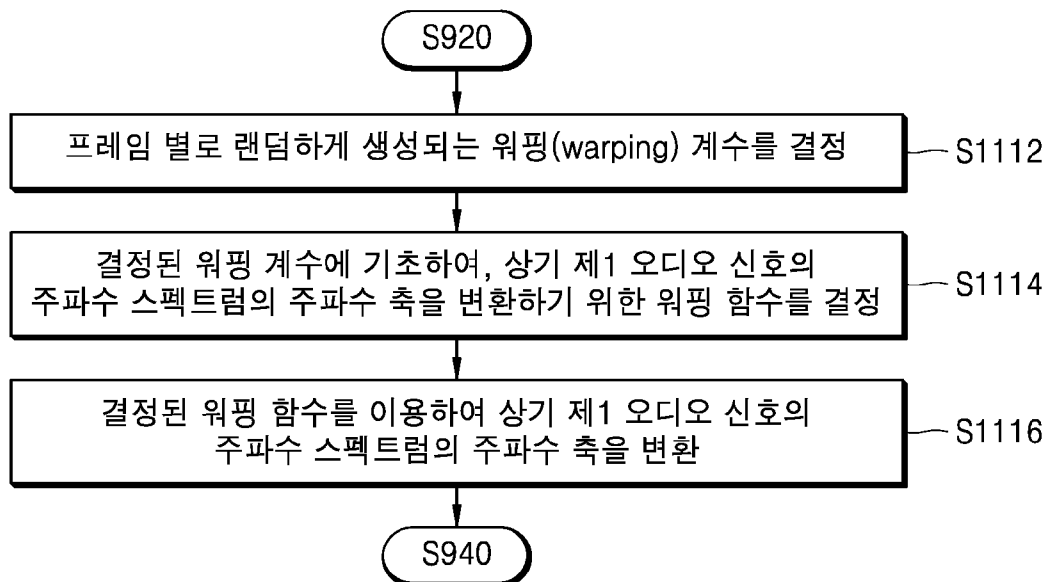
[도9]



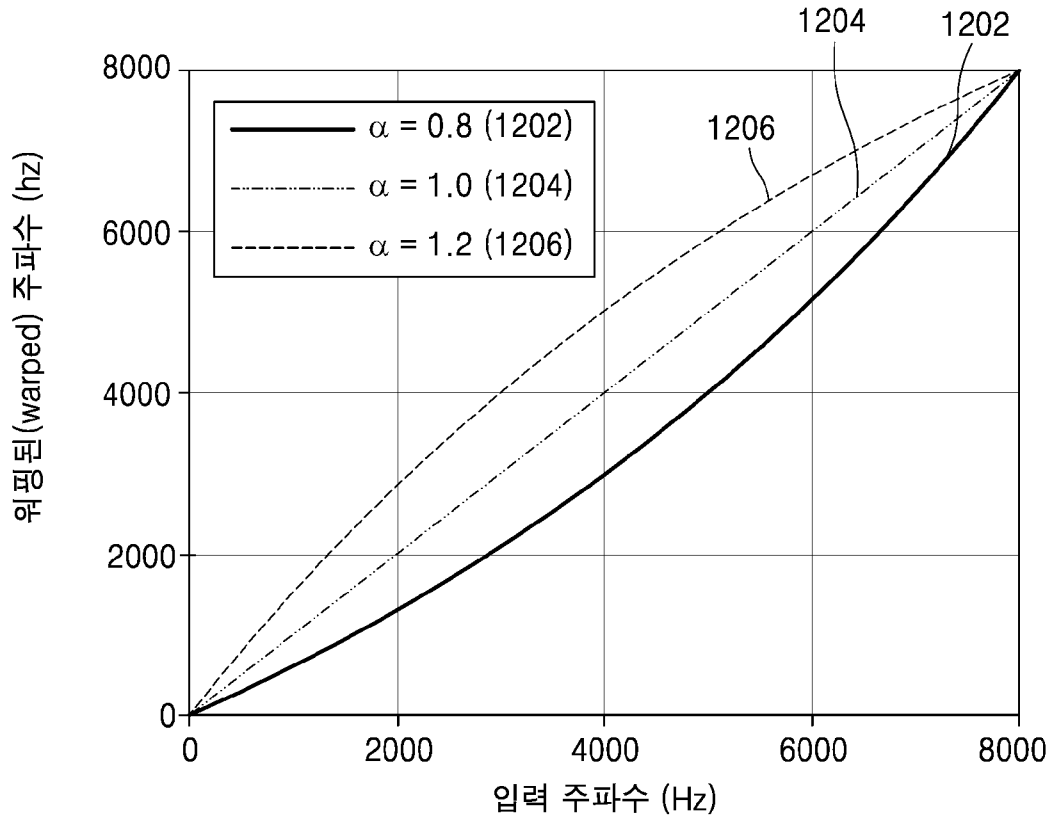
[도10]



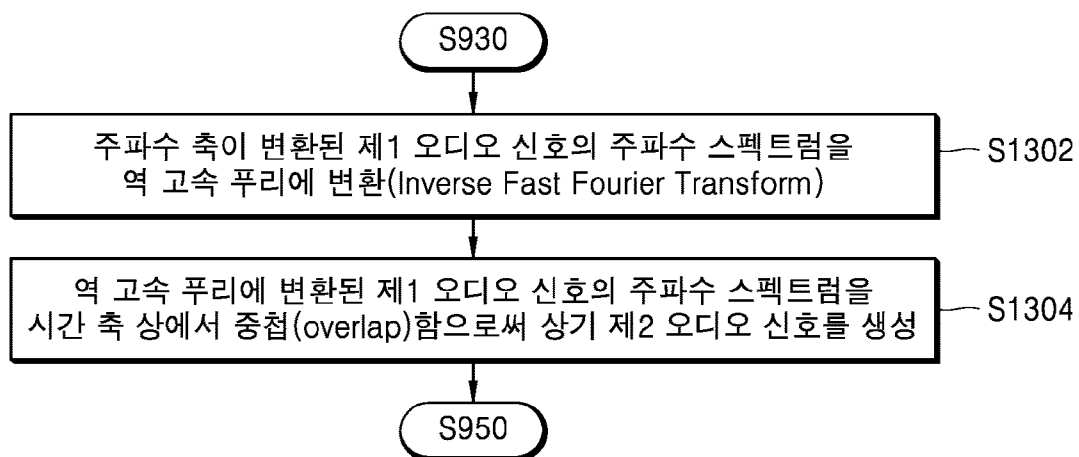
[도11]



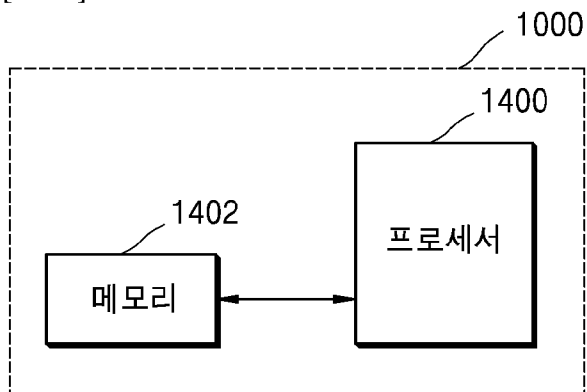
[도12]



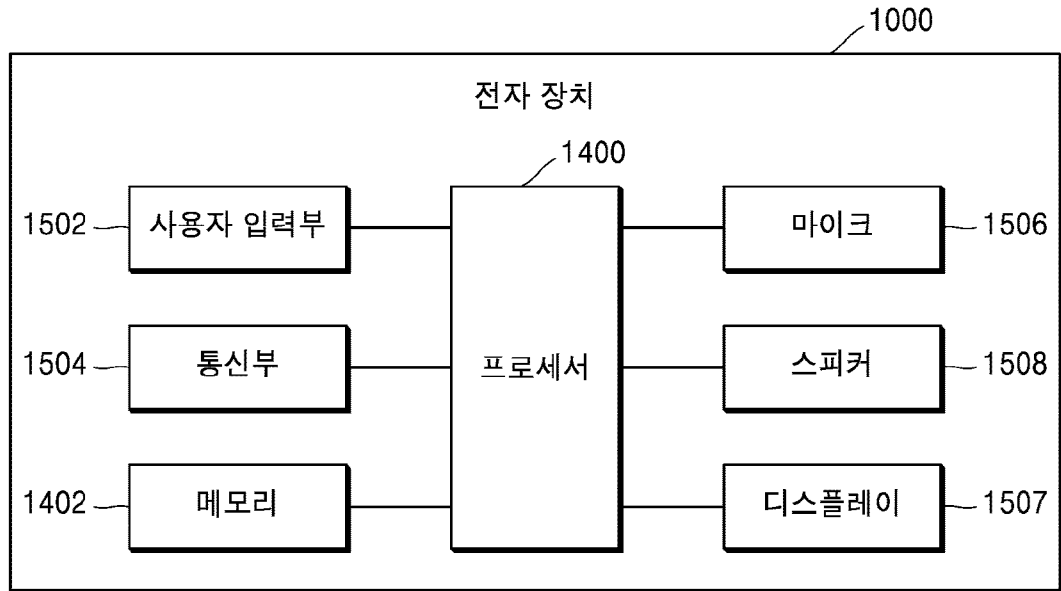
[도13]



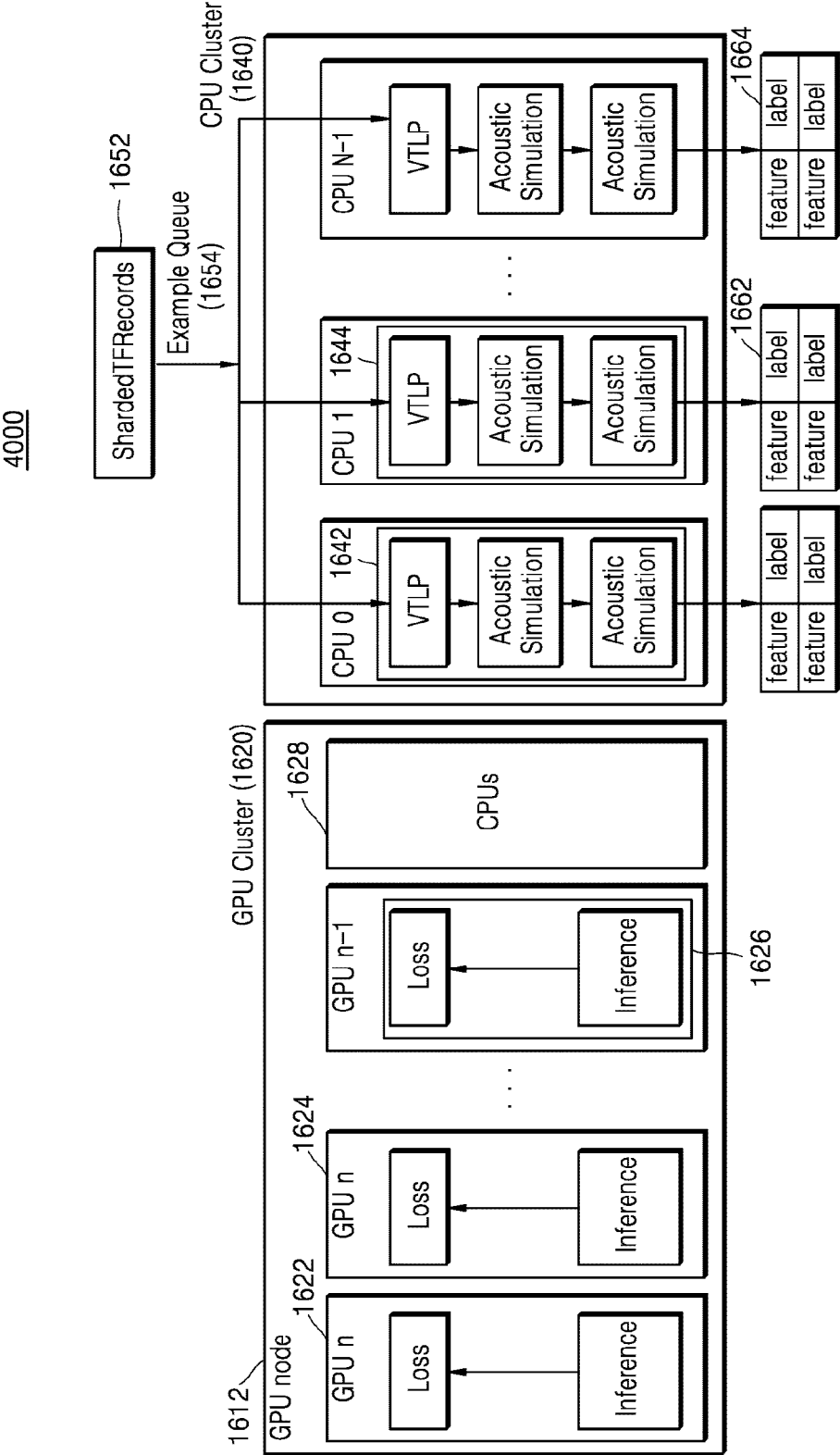
[도14]



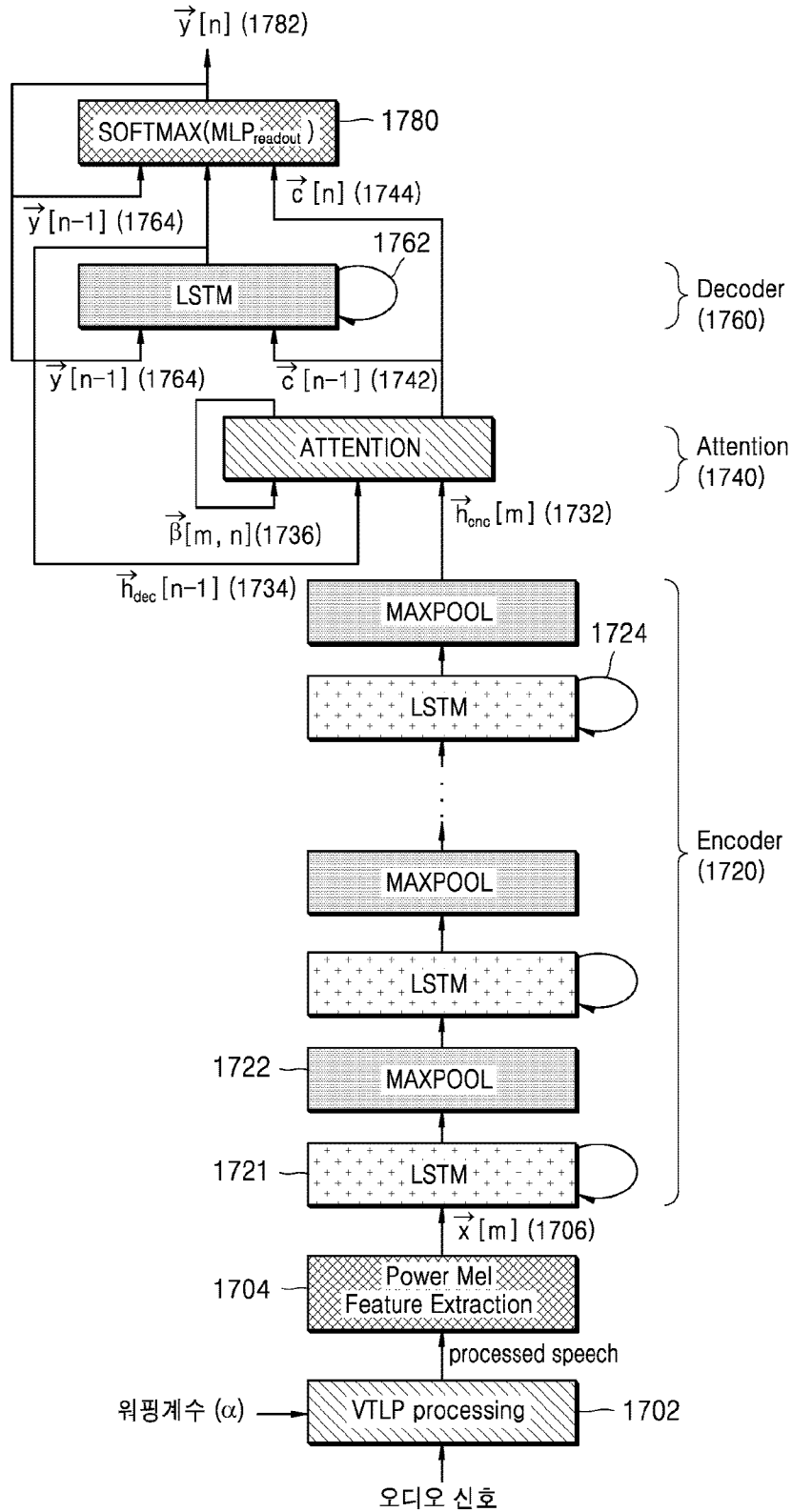
[도15]



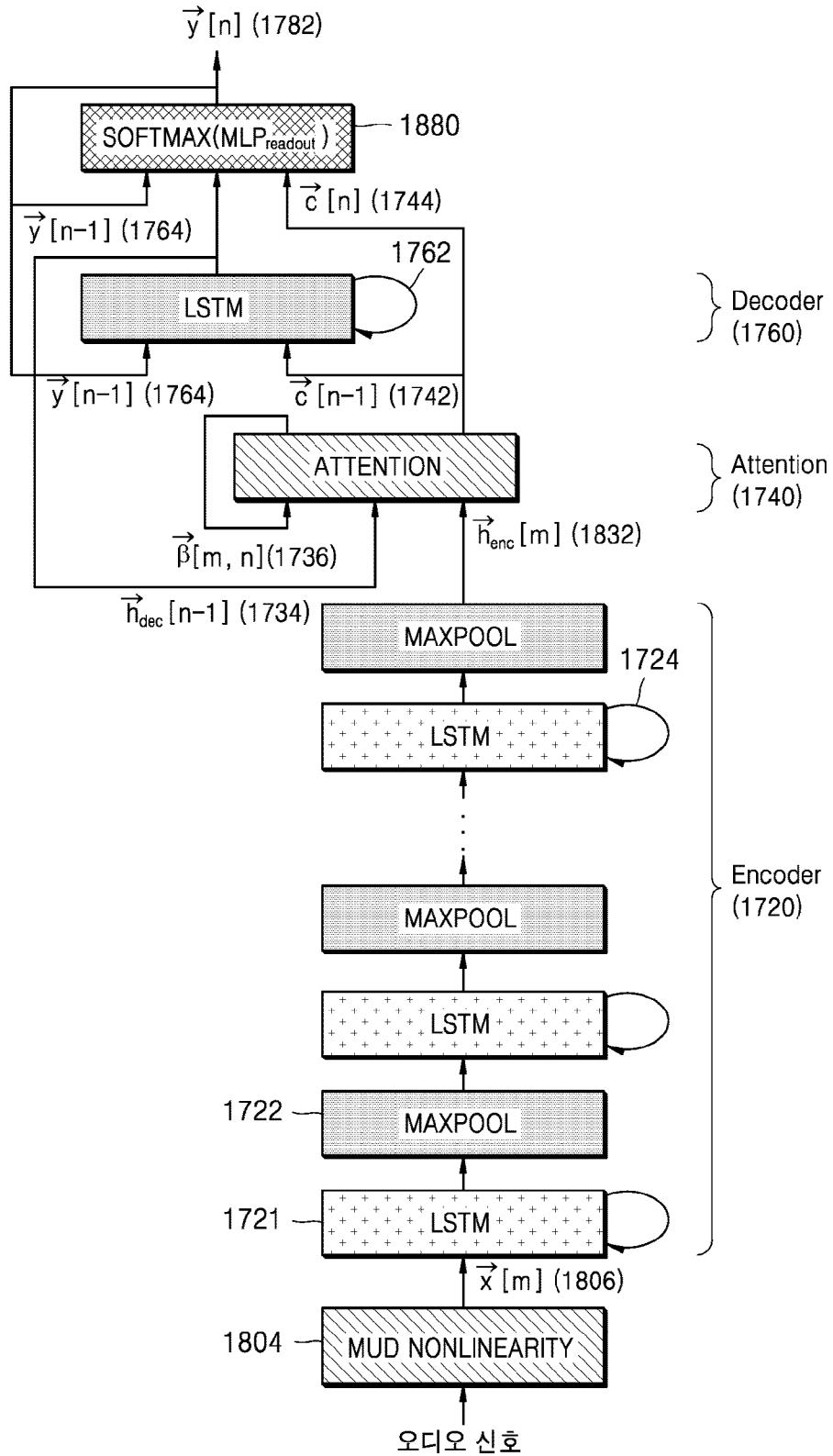
[도16]



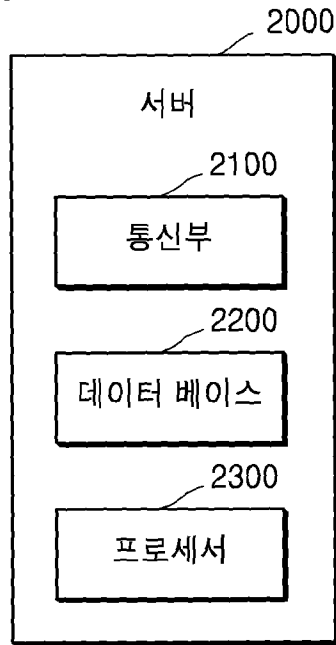
[도17]



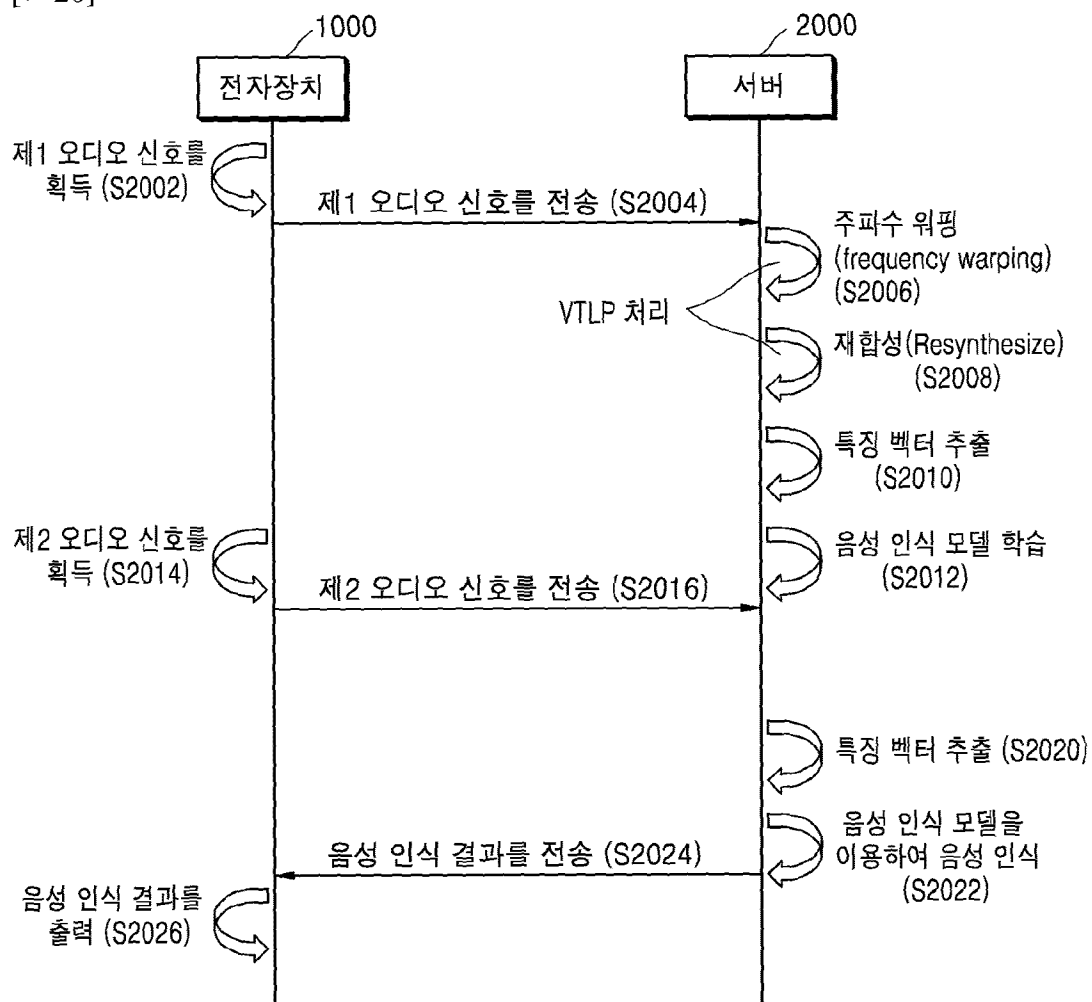
[도18]



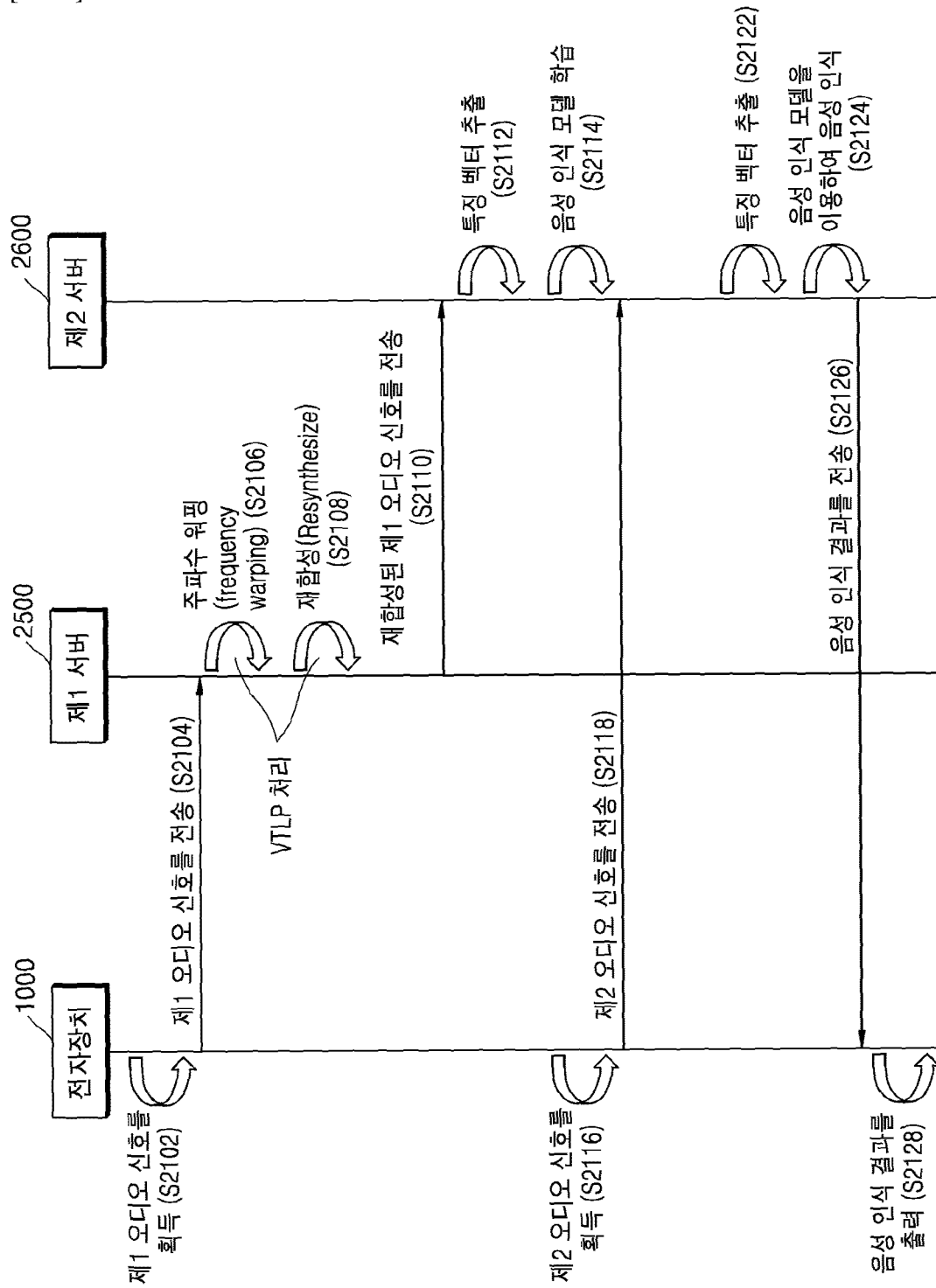
[도19]



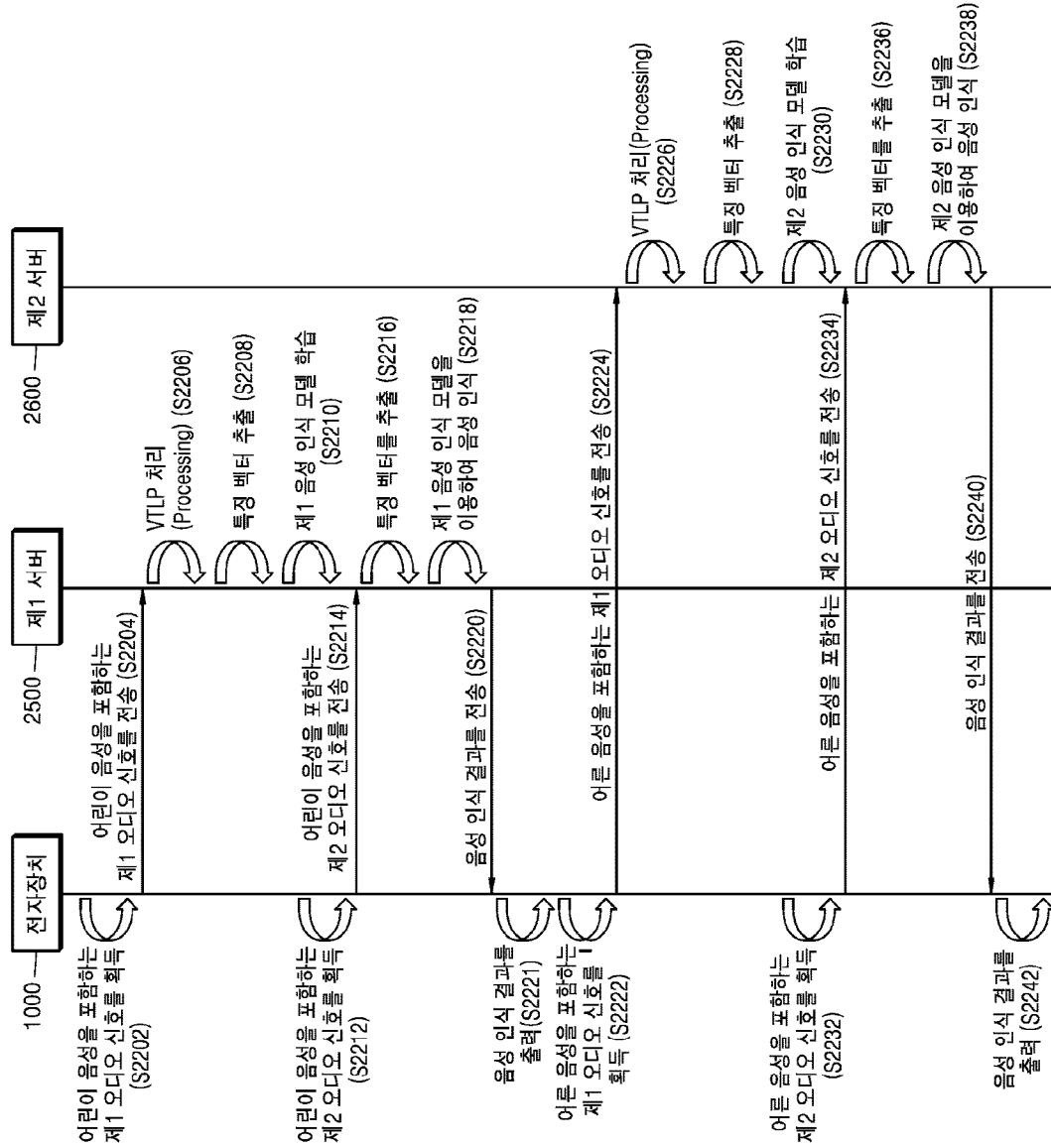
[도20]



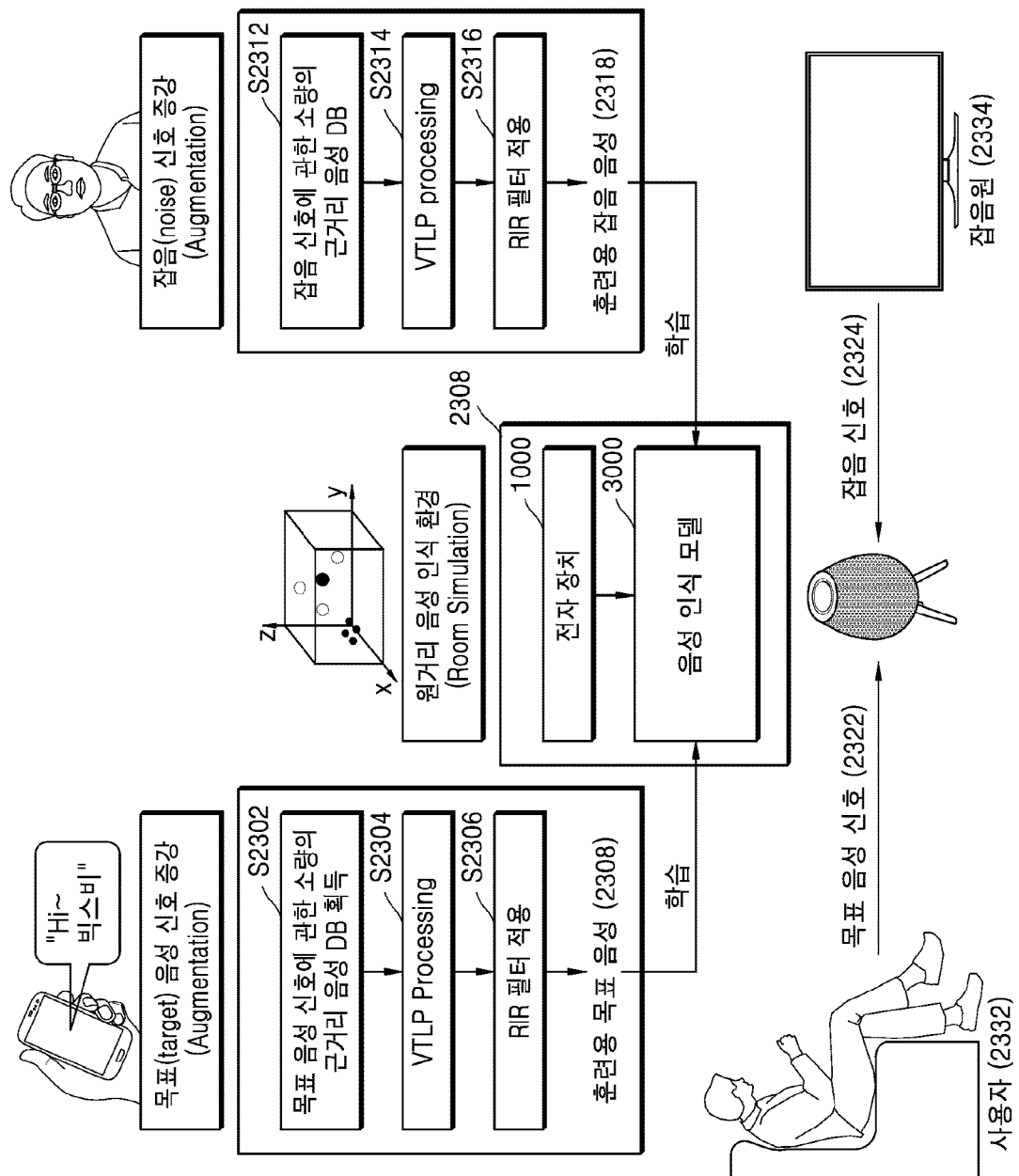
[도21]



[도22]



[도23]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2019/016181

A. CLASSIFICATION OF SUBJECT MATTER

G10L 25/75(2013.01)i, G10L 15/06(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L 25/75; G10L 15/02; G10L 15/20; G10L 19/00; G10L 21/02; H03G 3/00; H03G 3/30; H03G 5/00; H04B 15/00; H04R 5/02; H04S 7/00; G10L 15/06

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above

Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: voice recognition, frequency spectrum, filter bank, smoothing, feature vector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2008-0077874 A (SAMSUNG ELECTRONICS CO., LTD.) 26 August 2008 See paragraphs [0022]-[0026], [0038]-[0039]; and claim 1.	1-15
A	US 9954506 B2 (DOLBY LABORATORIES LICENSING CORPORATION) 24 April 2018 See column 20, line 3-column 21, line 27.	1-15
A	US 8103020 B2 (AVENDANO, Carlos et al.) 24 January 2012 See column 3, lines 26-42; claim 1; and figure 1.	1-15
A	KR 10-1295727 B1 (TRANSONO INC.) 16 August 2013 See paragraph [0018]; and claim 1.	1-15
A	US 9031268 B2 (FEJZO, Zoran et al.) 12 May 2015 See claims 1-2.	1-15
PX	KIM, Chanwoo et al. Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system. INTERSPEECH 2019, 19 September 2019 See pages 739-741.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

04 MARCH 2020 (04.03.2020)

Date of mailing of the international search report

04 MARCH 2020 (04.03.2020)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex Daejeon Building 4, 189, Cheongsu-ro, Seo-gu,
Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2019/016181

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-2008-0077874 A	26/08/2008	None	
US 9954506 B2	24/04/2018	CN 101048935 A	03/10/2007
		CN 101048935 B	23/03/2011
		CN 101421781 A	29/04/2009
		EP 2262108 A1	15/12/2010
		EP 2262108 B1	01/03/2017
		JP 2009-532739 A	10/09/2009
		JP 4981123 B2	18/07/2012
		JP 5101292 B2	19/12/2012
		KR 10-1261212 B1	07/05/2013
		KR 10-2007-0068379 A	29/06/2007
		US 10454439 B2	22/10/2019
		US 10476459 B2	12/11/2019
		US 2019-0115887 A1	18/04/2019
		US 2019-0115888 A1	18/04/2019
		US 2019-0115889 A1	18/04/2019
		US 2019-0115890 A1	18/04/2019
		US 2019-0115891 A1	18/04/2019
		US 2019-0115892 A1	18/04/2019
		US 2019-0115893 A1	18/04/2019
		US 2019-0115894 A1	18/04/2019
		US 2019-0115895 A1	18/04/2019
		US 2019-0115896 A1	18/04/2019
		US 9960743 B2	01/05/2018
		US 9966916 B2	08/05/2018
		US 9979366 B2	22/05/2018
US 8103020 B2	24/01/2012	US 2008-0049951 A1	28/02/2008
		US 2008-0212795 A1	04/09/2008
		US 7277550 B1	02/10/2007
		US 7353169 B1	01/04/2008
		US 8086448 B1	27/12/2011
		US 8321206 B2	27/11/2012
KR 10-1295727 B1	16/08/2013	None	
US 9031268 B2	12/05/2015	CN 103621110 A	05/03/2014
		CN 103621110 B	23/03/2016
		EP 2708039 A1	19/03/2014
		EP 2708039 B1	10/08/2016
		JP 2014-517596 A	17/07/2014
		JP 6023796 B2	09/11/2016
		KR 10-2014-0034817 A	20/03/2014
		KR 10-2036359 B1	24/10/2019
		US 2012-0288124 A1	15/11/2012
		US 2015-0230041 A1	13/08/2015
		US 9641952 B2	02/05/2017
		WO 2012-154823 A1	15/11/2012

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G10L 25/75(2013.01)i, G10L 15/06(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G10L 25/75; G10L 15/02; G10L 15/20; G10L 19/00; G10L 21/02; H03G 3/00; H03G 3/30; H03G 5/00; H04B 15/00; H04R 5/02; H04S 7/00; G10L 15/06

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 음성 인식(voice recognition), 주파수 스펙트럼(frequency spectrum), 필터 뱅크(filter bank), 평탄화(smoothing), 특징 벡터(feature vector)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	KR 10-2008-0077874 A (삼성전자주식회사) 2008.08.26 단락 [0022]-[0026], [0038]-[0039]; 및 청구항 1	1-15
A	US 9954506 B2 (DOLBY LABORATORIES LICENSING CORPORATION) 2018.04.24 컬럼 20, 라인 3 - 컬럼 21, 라인 27	1-15
A	US 8103020 B2 (CARLOS AVENDANO 등) 2012.01.24 컬럼 3, 라인 26-42; 청구항 1; 및 도면 1	1-15
A	KR 10-1295727 B1 ((주)트란소노) 2013.08.16 단락 [0018]; 및 청구항 1	1-15
A	US 9031268 B2 (ZORAN FEJZO 등) 2015.05.12 청구항 1-2	1-15
PX	CHANWOO KIM 등, 'Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system', INTERSPEECH 2019, 2019.09.19 페이지 739-741	1-15

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

“D” 본 국제출원에서 출원인이 인용한 문헌

“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후 “X”에 공개된 선출원 또는 특허 문헌

“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

“&” 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일

2020년 03월 04일 (04.03.2020)

국제조사보고서 발송일

2020년 03월 04일 (04.03.2020)

ISA/KR의 명칭 및 우편주소



대한민국 특허청
(35208) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)

팩스 번호 +82-42-481-8578

심사관

권성호

전화번호 +82-42-481-3547



국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2008-0077874 A	2008/08/26	없음	
US 9954506 B2	2018/04/24	CN 101048935 A CN 101048935 B CN 101421781 A EP 2262108 A1 EP 2262108 B1 JP 2009-532739 A JP 4981123 B2 JP 5101292 B2 KR 10-1261212 B1 KR 10-2007-0068379 A US 10454439 B2 US 10476459 B2 US 2019-0115887 A1 US 2019-0115888 A1 US 2019-0115889 A1 US 2019-0115890 A1 US 2019-0115891 A1 US 2019-0115892 A1 US 2019-0115893 A1 US 2019-0115894 A1 US 2019-0115895 A1 US 2019-0115896 A1 US 9960743 B2 US 9966916 B2 US 9979366 B2	2007/10/03 2011/03/23 2009/04/29 2010/12/15 2017/03/01 2009/09/10 2012/07/18 2012/12/19 2013/05/07 2007/06/29 2019/10/22 2019/11/12 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2019/04/18 2018/05/01 2018/05/08 2018/05/22
US 8103020 B2	2012/01/24	US 2008-0049951 A1 US 2008-0212795 A1 US 7277550 B1 US 7353169 B1 US 8086448 B1 US 8321206 B2	2008/02/28 2008/09/04 2007/10/02 2008/04/01 2011/12/27 2012/11/27
KR 10-1295727 B1	2013/08/16	없음	
US 9031268 B2	2015/05/12	CN 103621110 A CN 103621110 B EP 2708039 A1 EP 2708039 B1 JP 2014-517596 A JP 6023796 B2 KR 10-2014-0034817 A KR 10-2036359 B1 US 2012-0288124 A1 US 2015-0230041 A1 US 9641952 B2 WO 2012-154823 A1	2014/03/05 2016/03/23 2014/03/19 2016/08/10 2014/07/17 2016/11/09 2014/03/20 2019/10/24 2012/11/15 2015/08/13 2017/05/02 2012/11/15