

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2024년 6월 27일 (27.06.2024)



(10) 국제공개번호
WO 2024/136409 A1

- (51) 국제특허분류:
G10L 17/02 (2013.01) *G10L 15/02* (2006.01)
G10L 25/93 (2013.01)
- (21) 국제출원번호: PCT/KR2023/021003
- (22) 국제출원일: 2023년 12월 19일 (19.12.2023)
- (25) 출원언어: 한국어
- (26) 공개언어: 한국어
- (30) 우선권정보:
10-2022-0178297 2022년 12월 19일 (19.12.2022) KR
- (71) 출원인: 네이버 주식회사 (NAVER CORPORATION)
[KR/KR]: 13561 경기도 성남시 분당구 정자일로 95,
Gyeonggi-do (KR).
- (72) 발명자: 최민석 (CHOI, Min-Seok); 13529 경기도
성남시 분당구 정자일로 95, NAVER 1784, Gyeonggi-do
(KR). 이봉진 (LEE, Bong-Jin); 13529 경기도 성남시
분당구 정자일로 95, NAVER 1784, Gyeonggi-do (KR).

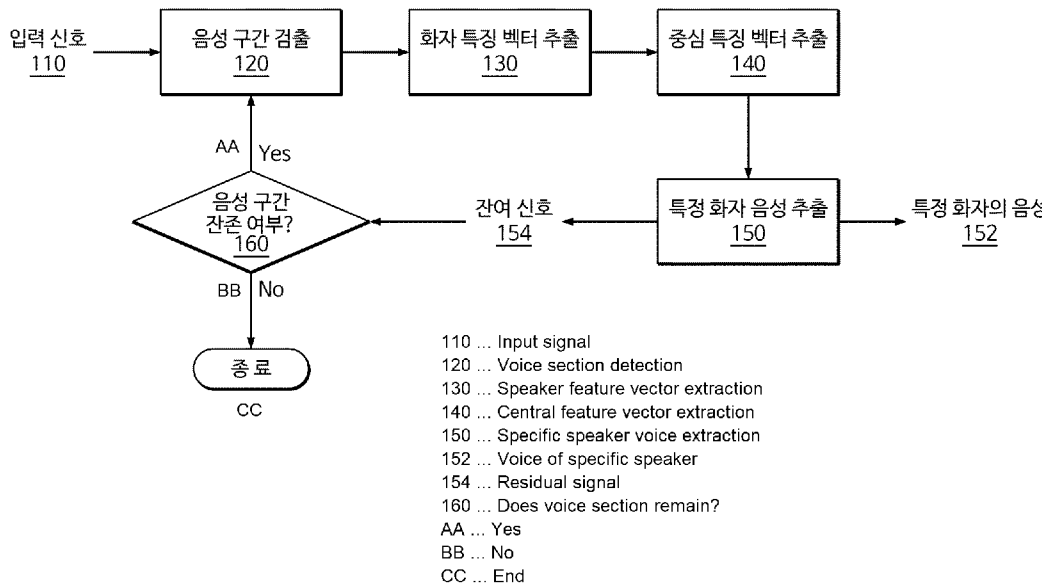
최소연 (CHOE, Soyeon); 13529 경기도 성남시 분당구
정자일로 95, NAVER 1784, Gyeonggi-do (KR). 허희수
(HEO, Hee Soo); 13529 경기도 성남시 분당구 정자일로
95, NAVER 1784, Gyeonggi-do (KR).

(74) 대리인: 안제성 등 (AHN, Je Sung et al.); 06253
서울특별시 강남구 도곡로 117, 옥신타워 14층, AnK
특허법률사무소, 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, CV, CZ, DE, DJ, DK, DM, DO, DZ,
EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR,
HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH,
KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD,
MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: SPEAKER DIARIZATION METHOD AND SYSTEM

(54) 발명의 명칭: 화자 분할 방법 및 시스템



(57) Abstract: The present disclosure relates to a speaker diarization method performed by at least one processor. The speaker diarization method comprises the steps of: extracting a first set of speaker feature vectors on the basis of a voice section included in an input signal; extracting a first central feature vector having the highest reliability, on the basis of the extracted first set of speaker feature vectors; and extracting the voice of a first speaker associated with the first central feature vector from the input signal.

(57) 요약서: 본 개시는 적어도 하나의 프로세서에 의해 수행되는 화자 분할 방법에 관한 것이다. 화자 분할 방법은, 입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하는 단계, 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하는 단계 및 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하는 단계를 포함한다.

[다음 쪽 계속]

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의
역내 권리의 보호를 위하여): ARIPO (BW, CV, GH, GM,
KE, LR, LS, MW, MZ, NA, RW, SC, 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,
ME, 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: 화자 분할 방법 및 시스템

기술분야

- [1] 본 개시는 화자 분할 방법 및 시스템에 관한 것으로, 구체적으로, 화자 특징 벡터 중 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출하고, 추출된 중심 특징 벡터와 연관된 특정 화자의 음성을 추출하는 화자 분할 방법 및 시스템에 관한 것이다.

[2]

배경기술

- [3] IT 기술의 발달 및 음성 인식 기술의 발달로 인해, 다양한 음성 인식 관련 서비스들이 제공되고 있다. 예를 들어, 회의 녹음 파일에 기초하여 자동으로 회의록을 생성해주는 회의록 생성 서비스, 드라마, 영화, 동영상 등의 자막을 자동으로 생성해주는 자동 자막 생성 서비스 등이 제공되고 있다.
- [4] 이와 같은 음성 인식 기능을 포함하는 서비스에서, 인식 대상 음성 데이터에 여러 화자의 음성이 포함된 경우, 여러 화자의 발화 구간을 구분하여 인식하기 위해 화자 분할이 수행될 수 있다. 다만, 인식 대상 음성에 잡음이 섞이거나 여러 화자의 음성이 중첩된 구간이 있는 경우, 화자 특징 벡터가 오염되어 화자 분할 결과에 오차가 발생할 수 있다는 문제점이 있다.

[5]

발명의 상세한 설명

기술적 과제

- [6] 본 개시는 상기와 같은 문제점을 해결하기 위한 화자 분할 방법, 명령어들을 기록한 컴퓨터 판독가능한 비일시적 기록 매체 및 장치(시스템)를 제공한다.

[7]

과제 해결 수단

- [8] 본 개시는 방법, 장치(시스템) 또는 명령어들을 기록한 컴퓨터 판독가능한 비일시적 기록 매체를 포함한 다양한 방식으로 구현될 수 있다.
- [9] 본 개시의 일 실시예에 따르면, 적어도 하나의 프로세서에 의해 수행되는, 화자 분할 방법은, 입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하는 단계, 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하는 단계 및 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하는 단계를 포함한다.
- [10] 본 개시의 일 실시예에 따른 방법을 컴퓨터에서 실행하기 위한 명령어들을 기록한 컴퓨터 판독가능한 비일시적 기록 매체가 제공된다.
- [11] 본 개시의 일 실시예에 따른 정보 처리 시스템은, 메모리 및 메모리와 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구

성된 적어도 하나의 프로세서를 포함하고, 적어도 하나의 프로그램은, 입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하고, 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하고, 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하기 위한 명령어들을 포함한다.

[12]

발명의 효과

[13] 본 개시의 일부 실시예에 따르면, 음성 구간으로부터 추출된 화자 특징 벡터들 중에서 특정 화자를 나타내는데 있어서 신뢰도가 높은 중심 특징 벡터를 추출함으로써, 화자 특징 벡터들 중 일부가 오염된 경우에도, 특정 화자의 음성 특징을 잘 반영하는 대표성 있는 화자 특징 벡터를 추출할 수 있다.

[14] 본 개시의 일부 실시예에 따르면, 입력 신호로부터 중심 특징 벡터와 연관된 특정 화자의 음성이 제거된 잔여 신호에 대해서 중심 특징 벡터의 추출과 이와 연관된 다른 특정 화자의 음성을 추출하는 일련의 과정을 반복하여 실시함으로써, 화자 인식 및 음성 추출 과정에서 여러 화자의 음성이 중첩되는 구간이나 여러 화자의 음성이 밀접한 구간이 줄어들 수 있어, 결과적으로 더 신뢰도를 갖고 하나 이상의 화자를 나타낼 수 있는 화자 특징 벡터를 추출할 수 있다.

[15] 본 개시의 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급되지 않은 다른 효과들은 청구범위의 기재로부터 본 개시가 속하는 기술분야에서 통상의 지식을 가진 자(“통상의 기술자”라 함)에게 명확하게 이해될 수 있을 것이다.

[16]

도면의 간단한 설명

[17] 본 개시의 실시예들은, 이하 설명하는 첨부 도면들을 참조하여 설명될 것이며, 여기서 유사한 참조 번호는 유사한 요소들을 나타내지만, 이에 한정되지는 않는다.

[18] 도 1은 본 개시의 일 실시예에 따른 화자 분할 방법의 예시를 나타내는 도면이다.

[19] 도 2는 본 개시의 일 실시예에 따른 정보 처리 시스템이 복수의 사용자 단말과 통신 가능하도록 연결된 구성을 나타내는 개요도이다.

[20] 도 3은 본 개시의 일 실시예에 따른 사용자 단말 및 정보 처리 시스템의 내부 구성을 나타내는 블록도이다.

[21] 도 4는 본 개시의 일 실시예에 따라 입력 신호에 포함된 음성 구간에 기초하여 복수의 화자 특징 벡터를 추출하는 예시를 나타내는 도면이다.

[22] 도 5는 본 개시의 일 실시예에 따라 화자 특징 벡터에 대해 클러스터링을 수행하는 예시를 나타내는 도면이다.

[23] 도 6은 본 개시의 일 실시예에 따라 추출된 복수의 화자 특징 벡터에 기초하여 중심 특징 벡터를 추출하는 예시를 나타내는 도면이다.

- [24] 도 7은 본 개시의 일 실시예에 따라 입력 신호로부터 특정 화자의 음성을 추출하는 예시를 나타내는 도면이다.
- [25] 도 8은 본 개시의 일 실시예에 따른 목적 음성 추출 모델의 예시를 나타내는 도면이다.
- [26] 도 9는 본 개시의 일 실시예에 따른 목적 음성 추출 모델을 학습시키는 방법의 예시를 나타내는 도면이다.
- [27] 도 10은 본 개시의 일 실시예에 따른 화자 분할 방법의 예시를 나타내는 흐름도이다.

[28]

발명의 실시를 위한 형태

- [29] 이하, 본 개시의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 개시의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기로 한다.
- [30] 첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지 않는다는 점에 유의한다.
- [31] 개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.
- [32] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 개시에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 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.
- [33] 본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 명세서 전체에서 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

- [34] 또한, 명세서에서 사용되는 '모듈' 또는 '부'라는 용어는 소프트웨어 또는 하드웨어 구성요소를 의미하며, '모듈' 또는 '부'는 어떤 역할들을 수행한다. 그렇지만, '모듈' 또는 '부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '모듈' 또는 '부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서, '모듈' 또는 '부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 또는 변수들 중 적어도 하나를 포함할 수 있다. 구성요소들과 '모듈' 또는 '부'들은 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '모듈' 또는 '부'들로 결합되거나 추가적인 구성요소들과 '모듈' 또는 '부'들로 더 분리될 수 있다.
- [35] 본 개시의 일 실시예에 따르면, '모듈' 또는 '부'는 프로세서 및 메모리로 구현될 수 있다. '프로세서'는 범용 프로세서, 중앙 처리 장치(CPU), 마이크로프로세서, 디지털 신호 프로세서(DSP), 제어기, 마이크로제어기, 상태 머신 등을 포함하도록 넓게 해석되어야 한다. 몇몇 환경에서, '프로세서'는 주문형 반도체(ASIC), 프로그램가능 로직 디바이스(PLD), 필드 프로그램가능 게이트 어레이(FPGA) 등을 지칭할 수도 있다. '프로세서'는, 예를 들어, DSP와 마이크로프로세서의 조합, 복수의 마이크로프로세서들의 조합, DSP 코어와 결합한 하나 이상의 마이크로프로세서들의 조합, 또는 임의의 다른 그러한 구성들의 조합과 같은 처리 디바이스들의 조합을 지칭할 수도 있다. 또한, '메모리'는 전자 정보를 저장 가능한 임의의 전자 컴포넌트를 포함하도록 넓게 해석되어야 한다. '메모리'는 임의 액세스 메모리(RAM), 판독-전용 메모리(ROM), 비-휘발성 임의 액세스 메모리(NVRAM), 프로그램가능 판독-전용 메모리(PROM), 소거-프로그램가능 판독 전용 메모리(EPROM), 전기적으로 소거가능 PROM(EEPROM), 플래쉬 메모리, 자기 또는 광학 데이터 저장장치, 레지스터들 등과 같은 프로세서-판독가능 매체의 다양한 유형들을 지칭할 수도 있다. 프로세서가 메모리로부터 정보를 판독하고/하거나 메모리에 정보를 기록할 수 있다면 메모리는 프로세서와 전자 통신 상태에 있다고 불린다. 프로세서에 집적된 메모리는 프로세서와 전자 통신 상태에 있다.
- [36] 본 개시에서, '시스템'은 서버 장치와 클라우드 장치 중 적어도 하나의 장치를 포함할 수 있으나, 이에 한정되는 것은 아니다. 예를 들어, 시스템은 하나 이상의 서버 장치로 구성될 수 있다. 다른 예로서, 시스템은 하나 이상의 클라우드 장치로 구성될 수 있다. 또 다른 예로서, 시스템은 서버 장치와 클라우드 장치가 함께 구성되어 동작될 수 있다.
- [37] 본 개시에서, '기계학습 모델'은 주어진 입력에 대한 해답(answer)을 추론하는데 사용하는 임의의 모델을 포함할 수 있다. 일 실시예에 따르면, 기계학습 모델은 입력 레이어(층), 복수 개의 은닉 레이어 및 출력 레이어를 포함한 인공신경망 모델을 포함할 수 있다. 여기서, 각 레이어는 복수의 노드를 포함할 수 있다. 본 개

사에서, 기계학습 모델은 인공신경망 모델을 지칭할 수 있으며, 인공신경망 모델은 기계학습 모델을 지칭할 수 있다.

- [38] 본 개시에서, '디스플레이'는 컴퓨팅 장치와 연관된 임의의 디스플레이 장치를 지칭할 수 있는데, 예를 들어, 컴퓨팅 장치에 의해 제어되거나 컴퓨팅 장치로부터 제공된 임의의 정보/데이터를 표시할 수 있는 임의의 디스플레이 장치를 지칭할 수 있다.
- [39] 본 개시에서, '복수의 A의 각각' 또는 '복수의 A 각각'은 복수의 A에 포함된 모든 구성 요소의 각각을 지칭하거나, 복수의 A에 포함된 일부 구성 요소의 각각을 지칭할 수 있다.
- [40] 본 개시에서, '신뢰도'란, 특정 화자 특징 벡터가 특정 화자의 음성 특징을 대표할 것으로 추정되는 확률을 지칭할 수 있다. 따라서, 본 개시에서, 가장 높은 신뢰도를 갖는 화자 특징 벡터란, 특정 화자의 음성 특징을 가장 잘 대표할 것으로 추정되는 화자 특징 벡터, 또는 특정 화자의 음성 특징을 가장 잘 반영할 것으로 추정되는 화자 특징 벡터를 지칭할 수 있다. 화자 특징 벡터의 신뢰도는 다양한 척도에 의해 측정될 수 있다. 일 실시예에 따르면, 특정 입력 신호로부터 추출된 복수의 화자 특징 벡터가 위치한 벡터 공간 내에서, 특정 화자 특징 벡터의 주위 영역에 해당 화자 특징 벡터들이 많이 밀집되어 있을수록 높은 신뢰도를 가질 수 있다.
- [41] 본 개시에서, '중심 특징 벡터'는 특정 입력 신호로부터 추출된 복수의 화자 특징 벡터 중 가장 높은 신뢰도를 갖는 화자 특징 벡터를 지칭할 수 있다.
- [42] 도 1은 본 개시의 일 실시예에 따른 화자 분할 방법의 예시를 나타내는 도면이다. 화자 분할 방법은 복수의 화자의 음성이 포함된 입력 신호(110)를 분석하여 각 화자의 발화 구간을 구분하여 인식하고, 해당 화자의 음성을 추출할 수 있다. 먼저, 정보 처리 시스템은 입력 신호(110)를 수신할 수 있다. 여기서, 입력 신호(110)는 복수의 화자의 발화 음성 및 배경 소음을 포함할 수 있다.
- [43] 그런 다음, 정보 처리 시스템은 입력 신호(110)로부터 음성 구간을 검출할 수 있다(120). 여기서, 음성 구간은 입력 신호(110) 중 화자의 발화 음성이 포함된 구간일 수 있다. 음성 구간이 검출된 경우, 정보 처리 시스템은 검출된 음성 구간에 대응하는 입력 신호(110)의 적어도 일부에 기초하여 복수의 화자 특징 벡터를 추출할 수 있다(130). 정보 처리 시스템이 입력 신호(110)로부터 음성 구간을 검출하고, 복수의 화자 특징 벡터를 추출하는 과정은 도 4를 참조하여 상세히 후술된다.
- [44] 정보 처리 시스템은 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출할 수 있다(140). 여기서, 특정 화자란 입력 신호(110)에 발화 음성이 포함된 복수의 화자들 중 사전 결정된 한 명의 화자를 지칭하는 것 대신에, 발화가 포함된 여러 명의 화자 중에서 중심 특징 벡터와 연관된 음성을 발화한 임의의 한 화자(any one speaker)를 지칭할 수 있다. 따라서, 복수의 화자 특징 벡터 중 가장 높은 신뢰도를 갖는 중심 특징 벡터는, 입력 신호(110)에 발화 음성이 포함된 복수의 화자들 중 구체적으로 어느 화자인지를 특정하거나

사전에 인지할 수 없으나, 임의의 한 화자의 음성 특징을 가장 잘 대표할 것으로 추정되는 화자 특징 벡터를 지칭할 수 있다.

- [45] 일 실시예에 따르면, 화자와 연관된 정보를 획득할 수 있는 경우, 화자 식별 (Speaker Identification; SID) 과정이 추가적으로 수행될 수도 있다. 예를 들어, 일부 화자의 음성 샘플 또는 일부 화자의 음성 샘플로부터 추출된 화자 특징 정보를 사전에 획득한 경우, 정보 처리 시스템은 화자 식별 과정을 추가적으로 수행함으로써, 입력 신호(110)로부터 추출된 화자 특징 벡터가 어느 화자와 연관되는지 추정할 수 있다. 이 경우, 중심 특징 벡터는, 입력 신호(110)에 포함된 복수의 화자들 중 사전 결정되거나 사전 식별된 화자의 음성 특징을 가장 잘 대표할 것으로 추정되는 화자 특징 벡터를 지칭하는 의미로 사용될 수도 있다. 다만, 이하의 설명에서는 '특정 화자'를 입력 신호에 발화 음성이 포함된 복수의 화자 중 임의의 한 화자(any one speaker)를 지칭하는 의미로 사용하고자 한다. 정보 처리 시스템이 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출하는 과정은 도 6을 참조하여 상세히 후술된다. 상술한 바와 같이, 복수의 화자 특징 벡터에 기초하여 가장 신뢰도가 높은 중심 특징 벡터를 추출함으로써, 화자 특징 벡터들 중 일부가 오염된 경우에도, 특정 화자의 음성 특징을 잘 반영하는 대표성 있는 화자 특징 벡터를 추출할 수 있다.
- [46] 중심 특징 벡터를 추출한 정보 처리 시스템은, 입력 신호(110)로부터 중심 특징 벡터와 연관된 특정 화자의 음성(152)을 추출할 수 있다(150). 예를 들어, 정보 처리 시스템은 목적 음성 추출 모델을 이용하여 입력 신호(110)로부터 중심 특징 벡터와 연관된 특정 화자의 음성(152)을 추출할 수 있다. 정보 처리 시스템이 입력 신호(110)로부터 중심 특징 벡터와 연관된 특정 화자의 음성(152)을 추출하는 과정은 도 7 및 도 8을 참조하여 상세히 후술된다.
- [47] 추가적으로, 정보 처리 시스템은 입력 신호(110)로부터 추출된 특정 화자의 음성(152)을 제외한 잔여 신호(154)에 음성 구간이 남아 있는지 여부를 판정할 수 있다(160). 잔여 신호(154)에 음성 구간이 남아 있다고 판정되는 경우, 상술한 일련의 과정을 반복하여 수행함으로써 입력 신호(110)에 포함된 다른 화자(들)의 음성을 추출할 수 있다.
- [48] 예를 들어, 정보 처리 시스템은 입력 신호(110)에 포함된 음성 구간 중, 추출된 특정 화자의 음성(152)이 포함된 구간을 제외한 잔여 음성 구간에 대해, 화자 특징 벡터 추출(130), 중심 특징 벡터 추출(140), 특정 화자 음성 추출(150) 등의 과정을 반복 수행함으로써, 다른 화자(들)의 음성을 추출할 수 있다.
- [49] 다른 예로, 정보 처리 시스템은 입력 신호(110)에 기초하여 추출된 복수의 화자 특징 벡터 중 특정 화자의 음성(152)이 포함된 구간을 제외한 잔여 음성 구간에 기초하여 추출된 일부의 화자 특징 벡터에 대해, 중심 특징 벡터 추출(140), 특정 화자 음성 추출(150) 등의 과정을 반복 수행함으로써, 다른 화자(들)의 음성을 추출할 수 있다.

- [50] 또 다른 예로, 정보 처리 시스템은 잔여 신호(154)를 입력 신호(110)로 하여, 상술한 일련의 과정(120, 130, 140, 150 등)을 반복하여 수행할 수 있다.
- [51] 상술한 과정을 잔여 신호(154)에 음성 구간이 더 이상 남아 있지 않을 때까지 반복하는 경우, 입력 신호(110)에 발화 음성이 포함된 모든 화자의 음성을 분리하여 인식할 수 있다.
- [52] 상술한 바와 같이, 먼저, 가장 신뢰도가 높은 화자 특징 벡터를 이용하여, 입력 신호(110)로부터 특정 화자의 음성(152)을 추출함으로써, 특정 화자의 음성(152)을 높은 정확도를 갖고 분리할 수 있다. 그런 다음, 입력 신호(110)로부터 특정 화자의 음성(152)이 제거된 잔여 신호(154)에 대해 일련의 과정을 반복하여 실시함으로써, 이러한 과정 중에 여러 화자의 음성이 중첩되는 구간이나 여러 화자의 음성이 밀접한 구간이 점차적으로 감소할 수 있어, 잔여 신호(154)에 발화 음성이 포함된 나머지 화자에 대해서도 신뢰도가 높은 화자 특징 벡터가 추출될 수 있어, 고품질의 화자 분할을 수행할 수 있다.
- [53] 이하에서는 도 2 내지 도 10을 참조하여, 본 개시의 화자 분할 방법에 대해 보다 상세히 후술된다. 이하에서는, 설명의 편의를 위해, 입력 신호(410)로부터 가장 먼저 중심 특징 벡터가 추출되는 음성과 연관된 화자를 제1 화자로 지칭하기로 한다. 이와 유사하게, n번째로 중심 특징 벡터가 추출되는 음성과 연관된 화자를 제n 화자로 지칭하기로 한다.
- [54] 상술한 설명에서, 화자 분할 방법은 정보 처리 시스템에 의해 수행되는 것으로 기재되었으나, 이에 한정되지 않으며, 본 개시의 화자 분할 방법에 포함된 적어도 일부 단계가 사용자 단말에 의해 수행될 수 있다. 예를 들어, 본 개시의 화자 분할 방법과 연관된 일련의 과정(120, 130, 140, 150 등)이 사용자 단말에 의해 수행될 수 있다. 다만, 이하에서는 설명의 편의를 위해, 본 개시의 화자 분할 방법이 정보 처리 시스템에 의해 수행되는 것으로 가정하여 서술하고자 한다.
- [55] 도 2는 본 개시의 일 실시예에 따른 정보 처리 시스템(230)이 복수의 사용자 단말(210_1, 210_2, 210_3)과 통신 가능하도록 연결된 구성을 나타내는 개요도이다. 도시된 바와 같이, 복수의 사용자 단말(210_1, 210_2, 210_3)은 네트워크(220)를 통해 화자 분할 서비스, 음성 인식 서비스 또는 화자 분할 및/또는 음성 인식 기능을 이용한 다양한 애플리케이션(예를 들어, 자동 회의록 생성 애플리케이션, 자동 자막 재생 애플리케이션 등)을 제공할 수 있는 정보 처리 시스템(230)과 연결될 수 있다. 여기서, 복수의 사용자 단말(210_1, 210_2, 210_3)은 화자 분할 서비스 및/또는 음성 인식 서비스를 제공받을 사용자의 단말을 포함할 수 있다. 일 실시예에서, 정보 처리 시스템(230)은 화자 분할 서비스 및/또는 음성 인식 서비스 등과 관련된 컴퓨터 실행 가능한 프로그램(예를 들어, 다운로드 가능한 애플리케이션) 및 데이터를 저장, 제공 및 실행할 수 있는 하나 이상의 서버 장치 및/또는 데이터베이스, 또는 클라우드 컴퓨팅 서비스 기반의 하나 이상의 분산 컴퓨팅 장치 및/또는 분산 데이터베이스를 포함할 수 있다.

- [56] 대안적으로, 사용자 단말(210_1, 210_2, 210_3)은 네트워크(220)를 통하지 않고, 화자 분할 서비스, 음성 인식 서비스 등과 관련된 컴퓨터 실행 가능한 프로그램 (예를 들어, 다운로드 가능한 애플리케이션)을 통해 사용자에게 화자 분할 서비스 및/또는 음성 인식 서비스를 제공할 수도 있다.
- [57] 정보 처리 시스템(230)에 의해 제공되는 화자 분할 서비스 및/또는 음성 인식 서비스는, 복수의 사용자 단말(210_1, 210_2, 210_3)의 각각에 설치된 화자 분할 애플리케이션, 음성 인식 애플리케이션, 자동 회의록 생성 애플리케이션, 자동 자막 생성 애플리케이션, 음성 편집 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등을 통해 사용자에게 제공될 수 있다. 예를 들어, 정보 처리 시스템(230)은 화자 분할 애플리케이션 등을 통해 사용자 단말(210_1, 210_2, 210_3)로부터 수신되는 화자 분할 요청, 음성 인식 요청, 회의록 생성 요청, 자막 생성 요청 등에 대응하는 정보를 제공하거나 대응하는 처리를 수행할 수 있다.
- [58] 복수의 사용자 단말(210_1, 210_2, 210_3)은 네트워크(220)를 통해 정보 처리 시스템(230)과 통신할 수 있다. 네트워크(220)는 복수의 사용자 단말(210_1, 210_2, 210_3)과 정보 처리 시스템(230) 사이의 통신이 가능하도록 구성될 수 있다. 네트워크(220)는 설치 환경에 따라, 예를 들어, 이더넷(Ethernet), 유선 홈 네트워크(Power Line Communication), 전화선 통신 장치 및 RS-serial 통신 등의 유선 네트워크, 이동통신망, WLAN(Wireless LAN), Wi-Fi, Bluetooth 및 ZigBee 등과 같은 무선 네트워크 또는 그 조합으로 구성될 수 있다. 통신 방식은 제한되지 않으며, 네트워크(220)가 포함할 수 있는 통신망(일례로, 이동통신망, 유선 인터넷, 무선 인터넷, 방송망, 위성망 등)을 활용하는 통신 방식뿐만 아니라 사용자 단말(210_1, 210_2, 210_3) 사이의 근거리 무선 통신 역시 포함될 수 있다.
- [59] 도 2에서 휴대폰 단말(210_1), 태블릿 단말(210_2) 및 PC 단말(210_3)이 사용자 단말의 예로서 도시되었으나, 이에 한정되지 않으며, 사용자 단말(210_1, 210_2, 210_3)은 유선 및/또는 무선 통신이 가능하고 화자 분할 애플리케이션, 음성 인식 애플리케이션, 자동 회의록 생성 애플리케이션, 자동 자막 생성 애플리케이션, 음성 편집 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등이 설치되어 실행될 수 있는 임의의 컴퓨팅 장치일 수 있다. 예를 들어, 사용자 단말은, AI 스피커, 스마트폰, 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스, 셋톱 박스 등을 포함할 수 있다. 또한, 도 2에는 3개의 사용자 단말(210_1, 210_2, 210_3)이 네트워크(220)를 통해 정보 처리 시스템(230)과 통신하는 것으로 도시되어 있으나, 이에 한정되지 않으며, 상이한 수의 사용자 단말이 네트워크(220)를 통해 정보 처리 시스템(230)과 통신하도록 구성될 수도 있다.
- [60] 일 실시예에 따르면, 정보 처리 시스템(230)은 복수의 사용자 단말(210_1, 210_2, 210_3)로부터 복수의 화자의 발화를 포함하는 입력 신호를 수신할 수 있

다. 그리고 나서, 정보 처리 시스템(230)은 입력 신호에 포함된 음성 구간에 기초하여, 복수의 화자 특징 벡터를 추출하고, 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출할 수 있다. 그런 다음, 정보 처리 시스템(230)은 입력 신호로부터 중심 특징 벡터와 연관된 특정 화자의 음성을 추출할 수 있다. 정보 처리 시스템(230)은 상술한 과정을 여러 번 반복하여 수행함으로써 입력 신호에 포함된 각 화자의 음성을 분리할 수 있다. 추가적으로, 정보 처리 시스템(230)은 분리된 각 화자의 음성 및/또는 각 화자의 음성에 추가적인 처리를 수행한 결과를 화자 분할의 결과로서 사용자 단말(210)로 전송할 수 있다. 예를 들어, 정보 처리 시스템은 분리된 각 화자의 음성에 기초하여, 자동 음성 인식(Automatic Speech Recognition; ASR) 또는 STT(Speech to Text) 등을 수행함으로써, 회의록 또는 자막 등을 생성하여 사용자 단말(210)로 전송할 수 있다.

[61] 도 3은 본 개시의 일 실시예에 따른 사용자 단말(210) 및 정보 처리 시스템(230)의 내부 구성을 나타내는 블록도이다. 사용자 단말(210)은 화자 분할 애플리케이션, 음성 인식 애플리케이션, 자동 회의록 생성 애플리케이션, 자동 자막 생성 애플리케이션, 음성 편집 애플리케이션, 모바일 브라우저 애플리케이션 또는 웹 브라우저 등을 실행 가능하고 유/무선 통신이 가능한 임의의 컴퓨팅 장치를 지칭할 수 있으며, 예를 들어, 도 2의 휴대폰 단말(210_1), 태블릿 단말(210_2), PC 단말(210_3) 등을 포함할 수 있다. 도시된 바와 같이, 사용자 단말(210)은 메모리(312), 프로세서(314), 통신 모듈(316) 및 입출력 인터페이스(318)를 포함할 수 있다. 이와 유사하게, 정보 처리 시스템(230)은 메모리(332), 프로세서(334), 통신 모듈(336) 및 입출력 인터페이스(338)를 포함할 수 있다. 도 3에 도시된 바와 같이, 사용자 단말(210) 및 정보 처리 시스템(230)은 각각의 통신 모듈(316, 336)을 이용하여 네트워크(220)를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다. 또한, 입출력 장치(320)는 입출력 인터페이스(318)를 통해 사용자 단말(210)에 정보 및/또는 데이터를 입력하거나 사용자 단말(210)로부터 생성된 정보 및/또는 데이터를 출력하도록 구성될 수 있다.

[62] 메모리(312, 332)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(312, 332)는 RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저장 장치로서 사용자 단말(210) 또는 정보 처리 시스템(230)에 포함될 수 있다. 또한, 메모리(312, 332)에는 운영체제와 적어도 하나의 프로그램 코드(예를 들어, 사용자 단말(210)에 설치되어 화자 분할 애플리케이션 등을 위한 코드)가 저장될 수 있다.

- [63] 이러한 소프트웨어 구성요소들은 메모리(312, 332)와는 별도의 컴퓨터에서 판독가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독가능한 기록매체는 이러한 사용자 단말(210) 및 정보 처리 시스템(230)에 직접 연결가능한 기록 매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독가능한 기록매체가 아닌 통신 모듈을 통해 메모리(312, 332)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 애플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 네트워크(220)를 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(312, 332)에 로딩될 수 있다.
- [64] 프로세서(314, 334)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(312, 332) 또는 통신 모듈(316, 336)에 의해 프로세서(314, 334)로 제공될 수 있다. 예를 들어, 프로세서(314, 334)는 메모리(312, 332)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.
- [65] 통신 모듈(316, 336)은 네트워크(220)를 통해 사용자 단말(210)과 정보 처리 시스템(230)이 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 사용자 단말(210) 및/또는 정보 처리 시스템(230)이 다른 사용자 단말 또는 다른 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성 또는 기능을 제공할 수 있다. 일례로, 사용자 단말(210)의 프로세서(314)가 메모리(312) 등과 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청 또는 데이터(예를 들어, 화자 분할 요청, 음성 인식 요청, 회의록 생성 요청, 자막 생성 요청 등)는 통신 모듈(316)의 제어에 따라 네트워크(220)를 통해 정보 처리 시스템(230)으로 전달될 수 있다. 역으로, 정보 처리 시스템(230)의 프로세서(334)의 제어에 따라 제공되는 제어 신호나 명령이 통신 모듈(336)과 네트워크(220)를 거쳐 사용자 단말(210)의 통신 모듈(316)을 통해 사용자 단말(210)에 수신될 수 있다. 예를 들어, 사용자 단말(210)은 정보 처리 시스템(230)으로부터 통신 모듈(316)을 통해 입력 신호로부터 분리된 각 화자의 음성 및/또는 각 화자의 음성에 추가적인 처리를 수행한 결과(예를 들어, 회의록, 자막 등)를 화자 분할의 결과로서 수신할 수 있다.
- [66] 입출력 인터페이스(318)는 입출력 장치(320)와의 인터페이스를 위한 수단일 수 있다. 일 예로서, 입력 장치는 오디오 센서 및/또는 이미지 센서를 포함한 카메라, 키보드, 마이크로폰, 마우스 등의 장치를, 그리고 출력 장치는 디스플레이, 스피커, 햅틱 피드백 디바이스(haptic feedback device) 등과 같은 장치를 포함할 수 있다. 다른 예로, 입출력 인터페이스(318)는 터치스크린 등과 같이 입력과 출력을 수행하기 위한 구성 또는 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 사용자 단말(210)의 프로세서(314)가 메모리(312)에 로딩된 컴퓨터 프로그램의 명령을 처리함에 있어서 정보 처리 시스템(230)이나 다른 사용자 단말이 제공하는 정보 및/또는 데이터를 이용하여 구성되는 서비스 화

면 등이 입출력 인터페이스(318)를 통해 디스플레이에 표시될 수 있다. 도 3에서는 입출력 장치(320)가 사용자 단말(210)에 포함되지 않도록 도시되어 있으나, 이에 한정되지 않으며, 사용자 단말(210)과 하나의 장치로 구성될 수 있다. 또한, 정보 처리 시스템(230)의 입출력 인터페이스(338)는 정보 처리 시스템(230)과 연결되거나 정보 처리 시스템(230)이 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 3에서는 입출력 인터페이스(318, 338)가 프로세서(314, 334)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(318, 338)가 프로세서(314, 334)에 포함되도록 구성될 수 있다.

[67] 사용자 단말(210) 및 정보 처리 시스템(230)은 도 3의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 일 실시예에 따르면, 사용자 단말(210)은 상술된 입출력 장치(320) 중 적어도 일부를 포함하도록 구현될 수 있다. 또한, 사용자 단말(210)은 트랜시버(transceiver), GPS(Global Positioning system) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수 있다. 예를 들어, 사용자 단말(210)이 스마트폰인 경우, 일반적으로 스마트폰이 포함하고 있는 구성요소를 포함할 수 있으며, 예를 들어, 가속도 센서, 자이로 센서, 카메라 모듈, 각종 물리적인 버튼, 터치패널을 이용한 버튼, 입출력 포트, 진동을 위한 진동기 등의 다양한 구성요소들이 사용자 단말(210)에 더 포함되도록 구현될 수 있다. 일 실시예에 따르면, 사용자 단말(210)의 프로세서(314)는 화자 분할 서비스를 제공하는 애플리케이션 등이 동작하도록 구성될 수 있다. 이 때, 해당 애플리케이션 및/또는 프로그램과 연관된 코드가 사용자 단말(210)의 메모리(312)에 로딩될 수 있다.

[68] 화자 분할 애플리케이션 등을 위한 프로그램이 동작되는 동안에, 프로세서(314)는 입출력 인터페이스(318)와 연결된 터치 스크린, 키보드, 오디오 센서 및/또는 이미지 센서를 포함한 카메라, 마이크로폰 등의 입력 장치를 통해 입력되거나 선택된 텍스트, 이미지, 영상, 음성 및/또는 동작 등을 수신할 수 있으며, 수신된 텍스트, 이미지, 영상, 음성 및/또는 동작 등을 메모리(312)에 저장하거나 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다. 예를 들어, 프로세서(314)는 화자 분할을 요청하는 사용자 입력을 수신하여, 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다. 다른 예로서, 프로세서(314)는 여러 화자의 발화가 포함된 입력 신호 등에 대한 사용자의 선택을 나타내는 입력을 수신하여, 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시스템(230)에 제공할 수 있다.

[69] 사용자 단말(210)의 프로세서(314)는 입력 장치(320), 다른 사용자 단말, 정보 처리 시스템(230) 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다. 프로세서(314)에 의해 처리된 정보 및/또는 데이터는 통신 모듈(316) 및 네트워크(220)를 통해 정보 처리 시

시스템(230)에 제공될 수 있다. 사용자 단말(210)의 프로세서(314)는 입출력 인터페이스(318)를 통해 입출력 장치(320)로 정보 및/또는 데이터를 전송하여, 출력할 수 있다. 예를 들면, 프로세서(314)는 수신한 정보 및/또는 데이터를 사용자 단말의 화면에 디스플레이할 수 있다.

- [70] 정보 처리 시스템(230)의 프로세서(334)는 복수의 사용자 단말(210) 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다. 프로세서(334)에 의해 처리된 정보 및/또는 데이터는 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)에 제공할 수 있다. 일 실시예에 따르면, 정보 처리 시스템(230)은 사용자 단말(210)로부터 수신된 입력 신호에 포함된 음성 구간에 기초하여, 복수의 화자 특징 벡터를 추출하고, 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출할 수 있다. 그런 다음, 정보 처리 시스템(230)은 입력 신호로부터 중심 특징 벡터와 연관된 특정 화자의 음성을 추출할 수 있다. 정보 처리 시스템(230)은 상술한 과정을 여러 번 반복하여 수행함으로써 입력 신호에 포함된 각 화자의 음성을 분리할 수 있다. 추가적으로, 정보 처리 시스템(230)은 분리된 각 화자의 음성 및/또는 각 화자의 음성에 추가적인 처리를 수행한 결과를 화자 분할의 결과로서 사용자 단말(210)로 제공할 수 있다. 예를 들어, 정보 처리 시스템은 분리된 각 화자의 음성에 기초하여, 자동 음성 인식(ASR) 또는 STT 등을 수행함으로써, 회의록 또는 자막 등을 생성하여 사용자 단말(210)로 전송할 수 있다.

- [71] 정보 처리 시스템(230)의 프로세서(334)는 사용자 단말(210)의 디스플레이 출력 가능 장치(예: 터치 스크린, 디스플레이 등), 음성 출력 가능 장치(예: 스피커) 등의 출력 장치(320)를 통해 처리된 정보 및/또는 데이터를 출력하도록 구성될 수 있다. 예를 들어, 정보 처리 시스템(230)의 프로세서(334)는 분리된 각 화자의 음성을 화자 분할의 결과로서 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)로 제공하고, 각 화자의 음성을 사용자 단말(210)의 사운드 출력 가능 장치 등을 통해 출력하도록 구성될 수 있다. 다른 예로서, 정보 처리 시스템(230)의 프로세서(334)는 통신 모듈(336) 및 네트워크(220)를 통해 사용자 단말(210)로 분리된 각 화자의 음성에 기초하여, 자동 음성 인식(ASR) 또는 STT 등을 수행함으로써 생성된 회의록 또는 자막 등을 제공하고, 사용자 단말(210)의 디스플레이 출력 가능 장치 등을 통해 출력하도록 구성될 수 있다.

- [72] 도 4는 본 개시의 일 실시예에 따라 입력 신호에 포함된 음성 구간에 기초하여 복수의 화자 특징 벡터를 추출하는 예시를 나타내는 도면이다. 일 실시예에 따르면, 정보 처리 시스템 또는 정보 처리 시스템의 프로세서는 입력 신호(410)를 수신할 수 있다. 여기서, 입력 신호(410)는 여러 명의 화자의 발화 음성(및 주변 소음)을 포함할 수 있다. 또한, 여기서, 화자는 인간 뿐만 아니라, 합성된 음성 발화가 가능한 가상의 인물 또는 캐릭터, 음성을 포함하는 사운드 생성 및 출력이 가능한 소프트웨어 또는 하드웨어 모듈 등을 포함하는 개념일 수 있다. 예를 들어, 입력 신호(410)는 다수가 참가한 회의 도중 녹음된 음성을 포함할 수 있다. 다른

예로, 입력 신호(410)는 드라마 또는 영화 등의 영상으로부터 추출된 음성을 포함할 수 있다.

- [73] 입력 신호(410)는 다양한 형태의 사운드일 수 있다. 예를 들어, 입력 신호(410)는 웨이브 폼(waveform) 형태의 사운드이거나, 스펙트로그램(spectrogram) 형태의 사운드일 수 있다. 일 실시예에 따르면, 정보 처리 시스템은 웨이브 폼 형태의 사운드를 수신하고, 수신된 웨이브 폼 형태의 사운드를 기초로 주파수 변환을 수행하여 스펙트로그램 형태로 변환된 사운드를 입력 신호(410)로서 사용할 수 있다.
- [74] 그런 다음, 정보 처리 시스템은 입력 신호(410)로부터 음성 구간을 검출할 수 있다(420). 여기서, 음성 구간은 입력 신호(110) 중 화자의 발화 음성이 포함된 구간일 수 있다. 예를 들어, 정보 처리 시스템은 음성 끝점 검출(End Point Detection; EPD)을 수행하여 발화의 시작점과 끝점을 검출함으로써, 입력 신호(410)에 포함된 음성 구간을 검출할 수 있다.
- [75] 음성 구간이 검출된 경우, 정보 처리 시스템은 검출된 음성 구간에 기초하여 복수의 화자 특징 벡터(432)를 추출할 수 있다(430). 화자 특징 벡터는 음성 구간에 포함된 발화 음성의 특징에 대한 정보를 포함할 수 있다. 즉, 화자 특징 벡터는 음성을 발화한 화자에 대한 정보를 포함할 수 있다.
- [76] 일 실시예에 따르면, 정보 처리 시스템은 각 단위 음성 구간 당 하나의 화자 특징 벡터를 추출함으로써, 검출된 음성 구간에 기초하여 복수의 화자 특징 벡터(432)를 추출할 수 있다. 예를 들어, 단위 음성 구간의 길이가 0.5초인 실시예를 가정하면, 검출된 제1 음성 구간(422)의 길이가 1초인 경우, 정보 처리 시스템은 제1 음성 구간(422)에 기초하여 2개의 화자 특징 벡터(도시된 예에서, E_1 및 E_2)를 추출할 수 있다. 또한, 검출된 제2 음성 구간(424)의 길이가 1.5초인 경우, 정보 처리 시스템은 제2 음성 구간(424)에 기초하여 3개의 화자 특징 벡터(도시된 예에서, E_3 , E_4 및 E_5)를 추출할 수 있다.
- [77] 정보 처리 시스템은 다양한 길이의 단위 음성 구간에 따른 음성 구간 검출 및 화자 특징 벡터 추출 방법을 이용하여 검출된 음성 구간으로부터 복수의 화자 특징 벡터(432)를 추출할 수 있다.
- [78] 일 실시예에 따르면, 정보 처리 시스템은 화자 특징 추출 모델(Speaker Embedding Extractor)을 이용하여 음성 구간에 기초하여 복수의 화자 특징 벡터(432)를 추출할 수 있다. 여기서, 화자 특징 추출 모델은 음성 구간에 기초하여 화자 특징 벡터를 추출하도록 학습된 기계학습 모델일 수 있다. 예를 들어, 화자 특징 추출 모델은 동일한 화자의 발화 음성을 포함하는 음성 구간들로부터는 동일/유사한 화자 특징 벡터를 추출하도록 학습되고, 상이한 화자의 발화 음성을 포함하는 음성 구간들로부터는 각 화자의 특징 차이가 잘 구분될 수 있는 화자 특징 벡터를 추출하도록 학습된 기계학습 모델일 수 있다.
- [79] 다른 실시예에 따르면, 정보 처리 시스템은 입력 신호(410)로부터 검출된 각 음성 구간(예를 들어, 단위 음성 구간)을 가공함으로써 복수의 화자 특징 벡터(432)

를 추출할 수 있다. 예를 들어, 정보 처리 시스템은 스펙트로그램 형태의 각 음성 구간의 평균 값을 산출함으로써 복수의 화자 특징 벡터(432)를 추출할 수 있다.

- [80] 도 5는 본 개시의 일 실시예에 따라 화자 특징 벡터에 대해 클러스터링을 수행하는 예시를 나타내는 도면이다. 일 실시예에 따르면, 정보 처리 시스템은 입력 신호에 포함된 음성 구간으로부터 추출된 복수의 화자 특징 벡터에 대해 클러스터링을 수행할 수 있다. 복수의 화자 특징 벡터에 대해 클러스터링이 수행되는 경우, 복수의 화자 특징 벡터 중 유사한 특징 벡터들이 그룹핑될 수 있다. 예를 들어, 정보 처리 시스템은 K-means Clustering, Spectral Clustering 등 임의의 클러스터링 방법을 이용하여 클러스터링을 수행할 수 있다. 일 실시예에 따르면, 정보 처리 시스템은 추출된 복수의 화자 특징 벡터로부터 클러스터의 개수를 추정하고, 복수의 화자 특징 벡터를 추정된 개수의 클러스터로 그룹핑할 수 있다.
- [81] 종래의 화자 분할 방법에 따르면, 입력 신호에 포함된 각 화자의 음성을 구분하기 위해 클러스터링 과정을 필수적으로 포함하지만, 본 개시의 화자 분할 방법에 따르면, 클러스터링 과정은 선택적으로 수행될 수 있다.
- [82] 도 5에는 세 명의 화자의 발화 음성이 포함된 두 입력 신호에 기초하여 추출된 두 세트의 화자 특징 벡터에 대해 클러스터링을 수행한 결과가 도시되어 있다. 제1 클러스터링 결과(520)는 세 명의 화자의 발화 음성이 포함된 제1 입력 신호에 포함된 음성 구간으로부터 추출된 제1 세트의 화자 특징 벡터(510)에 대해 클러스터링을 수행한 결과이다. 제1 클러스터링 결과(520)를 살펴보면, 제1 세트의 화자 특징 벡터(510)가 세 개의 클러스터(522, 524, 526)로 그룹핑된 것을 확인할 수 있다. 동일한 클러스터에 속한 화자 특징 벡터들은 서로 동일한 화자의 발화 음성을 포함하는 음성 구간에서 추출된 것으로 추정될 수 있다.
- [83] 제2 클러스터링 결과(540)는 세 명의 화자의 발화 음성이 포함된 제2 입력 신호에 포함된 음성 구간으로부터 추출된 제2 세트의 화자 특징 벡터(530)에 대해 클러스터링을 수행한 결과이다. 제2 입력 신호는 제1 화자 및 제2 화자의 발화 음성이 중첩된 중첩 음성 구간(532)을 포함할 수 있다. 이 경우, 중첩 음성 구간(532)에 기초하여 추출된 화자 특징 벡터(534)는 오염되어, 세 명의 화자 중 어느 한 화자의 특징을 잘 반영하지 않을 수 있다. 예를 들어, 중첩 음성 구간(532)에 기초하여 추출된 화자 특징 벡터(534)는 제1 화자의 발화 음성만 포함하는 음성 구간으로부터 추출된 화자 특징 벡터들이 위치한 영역과 제2 화자의 발화 음성만 포함하는 음성 구간으로부터 추출된 화자 특징 벡터들이 위치한 영역 사이에 위치할 수 있다. 결국, 중첩 음성 구간(532)에 기초하여 추출된 화자 특징 벡터(534)에 의해, 제1 화자에 대응되는 화자 특징 벡터의 영역과 제2 화자에 대응되는 화자 특징 벡터의 영역의 구분이 불분명해질 수 있다.
- [84] 제2 입력 신호에 기초하여 추출된 제2 세트의 화자 특징 벡터(530)에 대해 클러스터링을 수행한 결과인 제2 클러스터링 결과(540)를 살펴보면, 제2 입력 신호는 실제로 세 명의 화자의 발화 음성을 포함하고 있음에도 불구하고, 화자의 수가

잘못 추정되어, 제2 세트의 화자 특징 벡터(530)가 두 개의 클러스터(542, 544)로 그룹핑된 것을 확인할 수 있다.

- [85] 상술한 바와 같이, 입력 신호에 중첩 음성 구간(532)이 존재하는 경우, 중첩 음성 구간(532)에 의해 화자 특징 벡터에 오염이 발생하여, 화자 분할의 결과에 오류가 발생할 수 있다. 또한, 입력 신호에 중첩 음성 구간(532)이 존재하는 경우 뿐만 아니라, 입력 신호에 주변 소음과 같은 잡음이 포함된 경우에도 화자 특징 벡터에 오염이 발생될 수 있다.
- [86] 본 개시의 일 실시예에 따르면, 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출함으로써 특정 화자의 대표성 있는 화자 특징 벡터를 추출할 수 있다. 따라서, 클러스터링 결과에 오류가 있더라도, 혹은 클러스터링을 수행하지 않고도, 오류 없이 화자 분할을 수행할 수 있다. 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출하는 과정에 관하여서는 도 6을 참조하여 후술된다.
- [87] 도 6은 본 개시의 일 실시예에 따라 추출된 복수의 화자 특징 벡터에 기초하여 중심 특징 벡터를 추출하는 예시를 나타내는 도면이다. 도 6에는 세 명의 화자(화자 A, 화자 B 및 화자 C)의 발화 음성을 포함하는 입력 신호에 기초하여 추출된 제1 세트의 화자 특징 벡터(600)의 예시가 도시되어 있다. 제1 세트의 화자 특징 벡터(600)는 화자 A의 발화 음성만 포함된 음성 구간에 기초하여 추출된 복수의 화자 특징 벡터를 포함하는 제1 그룹(610), 화자 B의 발화 음성만 포함된 음성 구간에 기초하여 추출된 복수의 화자 특징 벡터를 포함하는 제2 그룹(620), 화자 A 및 화자 B의 발화가 중첩된 음성 구간에 기초하여 추출된 복수의 화자 특징 벡터를 포함하는 제3 그룹(630), 및 화자 C의 발화 음성만 포함된 음성 구간에 기초하여 추출된 제4 그룹(640)을 포함할 수 있다. 이와 같이, 입력 신호에 두 명 이상의 화자의 발화 음성이 중첩된 구간이 포함된 경우, 화자 특징 벡터들에 대해 클러스터링을 수행하면, 제1 그룹(610), 제2 그룹(620) 및 제3 그룹(630)이 모두 동일한 클러스터로 분류될 수 있다. 따라서, 종래의 방법에 따라, 클러스터링 결과만을 기초로 화자 분할을 수행하는 경우, 화자 분할의 결과에 오류가 발생할 수 있다.
- [88] 본 개시의 일 실시예에 따르면, 정보 처리 시스템은 추출된 제1 세트의 화자 특징 벡터(600)에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출할 수 있다. 일 실시예에 따르면, 정보 처리 시스템은 화자 특징 벡터의 밀집도를 신뢰도의 척도로서 사용할 수 있다.
- [89] 예를 들어, 정보 처리 시스템은 먼저 벡터 공간 내에서 화자 특징 벡터가 가장 밀집된 영역을 중심 영역으로 결정할 수 있다. 구체적 예로, 정보 처리 시스템은 벡터 공간 내에서 추출된 제1 세트의 화자 특징 벡터(600)의 각각을 중심으로 한 동일한 크기의 복수의 영역 중 가장 많은 수의 화자 특징 벡터를 포함하는 영역을 중심 영역으로 결정할 수 있다. 도시된 예에서, 제1 특징 벡터(650)를 중심으로 한 제1 영역(652)에는 제1 특징 벡터(650)를 제외한 6개의 화자 특징 벡터가 존재

한다. 또한, 제2 특징 벡터(660)를 중심으로 한 제2 영역(662)(제1 영역(652)과 동일한 모양 및 크기의 영역)에는 제2 특징 벡터(660)를 제외한 5개의 화자 특징 벡터가 존재한다. 따라서, 이 경우 제1 영역(652)이 중심 영역으로 결정될 수 있다.

[90] 그런 다음, 정보 처리 시스템은 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 중심 특징 벡터를 결정할 수 있다. 예를 들어, 정보 처리 시스템은 중심 영역에 포함된 하나 이상의 화자 특징 벡터의 평균을 중심 특징 벡터로 결정할 수 있다. 이 실시예에 따르면, 정보 처리 시스템은 중심 영역으로 결정된 제1 영역(652)에 포함된 7개의 화자 특징 벡터의

[91] 평균을 중심 특징 벡터로 결정할 수 있다. 다른 예로, 정보 처리 시스템은 중심 영역의 중심 또는 중심에 가장 근접하여 위치한 화자 특징 벡터를 중심 특징 벡터로 추출할 수 있다. 이 실시예에 따르면, 정보 처리 시스템은 중심 영역으로 결정된 제1 영역(652)의 중심에 위치하는 제1 특징 벡터(650)를 중심 특징 벡터로 추출할 수 있다. 중심 특징 벡터로 추출된 제1 특징 벡터(650)는 화자 A의 발화 음성만 포함된 음성 구간에 기초하여 추출된 복수의 화자 특징 벡터를 포함하는 제1 그룹(610) 중에서도 가장 밀집된 영역에 기초하여 추출되어, 화자 A의 음성을 가장 잘 대표할 것으로 추정될 수 있다.

[92] 일 실시예에서, 중심 특징 벡터를 추출하기 전에 제1 세트의 화자 특징 벡터(600)에 대해 클러스터링이 수행된 경우, 정보 처리 시스템은 클러스터링 결과, 가장 많은 수의 화자 특징 벡터를 포함하는 클러스터로부터 중심 특징 벡터를 추출할 수 있다. 예를 들어, 제1 세트의 화자 특징 벡터(600)에 대해 클러스터링을 수행한 결과, 제1 그룹(610), 제2 그룹(620) 및 제3 그룹(630)을 포함하는 제1 클러스터, 제4 그룹(640)을 포함하는 제2 클러스터로 분류될 수 있다. 이 경우, 가장 많은 수의 화자 특징 벡터를 포함하는 제1 클러스터에 대응되는 공간 내에서 중심 영역을 결정하고, 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 중심 특징 벡터를 추출할 수 있다.

[93] 상술한 바와 같이, 입력 신호로부터 추출된 복수의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 중심 특징 벡터를 추출함으로써 특정 화자의 대표성 있는 화자 특징 벡터를 추출할 수 있어, 클러스터링 결과에 오류가 있더라도, 혹은 클러스터링을 수행하지 않고도, 높은 정확도를 갖고 화자 분할을 수행할 수 있다.

[94] 도 7은 본 개시의 일 실시예에 따라 입력 신호(710)로부터 특정 화자의 음성(720)을 추출하는 예시를 나타내는 도면이다. 정보 처리 시스템은 입력 신호(710)로부터 중심 특징 벡터와 연관된 특정 화자의 음성(720)을 추출할 수 있다. 예를 들어, 정보 처리 시스템은 목적 음성 추출 모델(700)을 이용하여, 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출할 수 있다. 특정 화자의 음성(720)을 추출하기 위해 이용되는 목적 음성 추출 모델(700)의 구체적인 예시 및 학습 방법에 관하여서는 도 8 및 도 9를 참조하여 상세히 후술된다.

- [95] 제1 화자의 음성을 추출한 뒤, 정보 처리 시스템은 입력 신호(710)로부터 추출된 제1 화자의 음성을 제외한 잔여 신호에 음성 구간이 남아 있는지 여부를 판정할 수 있다. 잔여 신호에 음성 구간이 남아 있다고 판정되는 경우, 도 4 내지 도 7을 참조하여 상술한 일련의 과정 중 적어도 일부 과정을 반복하여 수행할 수 있다.
- [96] 예를 들어, 정보 처리 시스템은 입력 신호(710)에 포함된 음성 구간 중, 추출된 특정 화자의 음성(720)이 포함된 음성 구간을 제외한 잔여 음성 구간에 기초하여, 복수의 화자 특징 벡터를 추출하고, 추출된 복수의 화자 특징 벡터에 기초하여 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호(710) 또는 잔여 신호로부터 추출된 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.
- [97] 다른 예로, 정보 처리 시스템은 입력 신호(710)에 기초하여 추출된 제1 세트의 화자 특징 벡터 중 특정 화자의 음성(720)이 포함된 음성 구간을 제외한 잔여 음성 구간에 기초하여 추출된 제1 부분 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호(710) 또는 잔여 신호로부터 추출된 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.
- [98] 또 다른 예로, 정보 처리 시스템은 잔여 신호에 포함된 음성 구간에 기초하여, 복수의 화자 특징 벡터를 추출하고, 추출된 복수의 화자 특징 벡터에 기초하여 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호(710) 또는 잔여 신호로부터 추출된 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.
- [99] 상술한 일련의 과정을 잔여 신호에 음성 구간이 남아 있지 않을 때까지 반복하여 수행하는 경우, 입력 신호(710)에 발화 음성이 포함된 모든 화자의 음성을 분리할 수 있다.
- [100] 추가적으로, 정보 처리 시스템은 분리된 각 화자의 음성 및/또는 각 화자의 음성에 추가적인 처리를 수행한 결과를 화자 분할의 결과로서 사용자 단말로 전송할 수 있다. 예를 들어, 정보 처리 시스템은 분리된 각 화자의 음성에 기초하여, 자동 음성 인식(Automatic Speech Recognition; ASR) 또는 STT(Speech to Text) 등을 수행함으로써, 회의록 또는 자막 등을 생성하여 사용자 단말로 전송할 수 있다.
- [101] 도 8은 본 개시의 일 실시예에 따른 목적 음성 추출 모델(800)의 구체적인 예시를 나타내는 도면이다. 일 실시예에 따르면, 목적 음성 추출 모델(800)은 혼합 음성(810) 및 목적 화자 특징 벡터(820)를 입력으로 수신하여, 목적 화자 특징 벡터(820)와 연관된 마스크(830)를 출력하는 모델일 수 있다.
- [102] 여기서, 혼합 음성(810)은 목적 화자의 음성과 다른 화자의 음성 및/또는 잡음이 혼합된 음성일 수 있다. 예를 들어, 혼합 음성(810)은 웨이브 폼(waveform) 형태의 음성이거나, 스펙트로그램(spectrogram) 형태의 음성일 수 있다. 일 실시예에 따르면, 웨이브 폼 형태의 음성에 대해 주파수 변환을 수행하여 스펙트로그램 형태

로 변환함으로써, 스펙트로그램 형태의 음성을 혼합 음성(810)으로서 사용할 수 있다.

- [103] 목적 화자 특징 벡터(820)는 목적 화자에 대한 정보를 포함하는 벡터일 수 있다. 또한, 목적 화자 특징 벡터(820)와 연관된 마스크(830)는 입력 신호로부터 목적 화자를 제외한 다른 화자의 음성 및 잡음을 제거하기 위한 마스크일 수 있다. 출력된 마스크(830) 및 혼합 음성(810)을 이용하면, 목적 화자의 음성(840)을 추출할 수 있다. 예를 들어, 혼합 음성(810)에 마스크(830)를 곱함으로써, 목적 화자의 음성(840)을 추출할 수 있다.
- [104] 일 실시예에서, 혼합 음성(810)으로서 스펙트로그램 형태의 음성을 사용한 경우, 추출된 목적 화자의 음성(840) 역시 스펙트로그램 형태일 수 있다. 이 경우, 스펙트로그램 형태의 목적 화자의 음성에 주파수 변환의 역변환을 수행함으로써 웨이브폼 형태의 음성으로 변환할 수 있다.
- [105] 일 실시예에 따르면, 정보 처리 시스템은 입력 신호를 혼합 음성(810)으로서, 추출된 중심 특징 벡터를 목적 화자 특징 벡터(820)로서 목적 음성 추출 모델(800)에 입력하여, 중심 특징 벡터와 연관된 마스크(830)를 추정할 수 있으며, 입력 신호에 마스크(830)를 곱함으로써 특정 화자의 음성을 추출할 수 있다. 추가적으로 또는 대안적으로, 정보 처리 시스템은 목적 화자 특징 벡터(820)로서 상술한 중심 특징 벡터 추출 방법과 유사하거나 상이한 방법에 의해 추출된 다른 화자 특징 벡터를 이용할 수도 있다.
- [106] 도 8에 도시된 목적 음성 추출 모델(800)은 일 예시일 뿐이며, 다른 실시예에서 목적 음성 추출 모델(800)은 도시된 모델과 다르게 구현될 수 있다. 예를 들어, 목적 음성 추출 모델(800)은 목적 화자 특징 벡터(820)와 연관된 마스크(830)를 출력하는 대신, 곧바로 목적 화자의 음성(840)을 출력하도록 구현될 수도 있다.
- [107] 도 9는 본 개시의 일 실시예에 따른 목적 음성 추출 모델(800)을 학습시키는 방법의 예시를 나타내는 도면이다. 일 실시예에 따르면, 목적 음성 추출 모델(800)은 혼합 음성(810) 및 목적 화자 특징 벡터(820)를 입력으로 수신하여, 목적 화자 특징 벡터(820)와 연관된 마스크(830)를 출력하도록 학습된 기계학습 모델일 수 있다. 예를 들어, 목적 음성 추출 모델(800)은 CNN(convolutional neural network), LSTM(long short term memory), FC(fully connected) 레이어 중 적어도 하나를 포함하는 인공신경망 모델일 수 있다.
- [108] 일 실시예에 따르면, 정보 처리 시스템은 목적 음성 추출 모델(800)을 학습시키기 위한 학습 데이터로서 제1 목적 화자 참조 음성(822), 제2 목적 화자 참조 음성(812), 타 화자 참조 음성(814)을 획득할 수 있다.
- [109] 제1 목적 화자 참조 음성(822) 및 제2 목적 화자 참조 음성(812)은 목적 화자의 발화 음성만을 포함하는 음성일 수 있다. 제1 목적 화자 참조 음성(822) 및 제2 목적 화자 참조 음성(812)은 서로 동일하거나 상이할 수 있다. 또한, 타 화자 참조 음성(814)은 목적 화자가 아닌 다른 화자의 발화 및/또는 잡음을 포함하는 음성일 수 있다. 정보 처리 시스템은 제2 목적 화자 참조 음성(812)과 타 화자 참조 음

성(814)을 혼합함으로써, 혼합 음성(810)을 생성할 수 있다. 상술한 각 음성(810, 812, 814, 822)은 웨이브 폼 형태의 음성이거나, 스펙트로그램 형태의 음성일 수 있다. 일 실시예에 따르면, 웨이브 폼 형태의 음성에 대해 주파수 변환을 수행하여 스펙트로그램 형태의 음성으로 변환하여 사용할 수 있다.

[110] 먼저, 정보 처리 시스템은 제1 목적 화자 참조 음성(822)에 기초하여, 목적 화자 특징 벡터(820)를 추출할 수 있다. 제1 목적 화자 참조 음성(822)에 기초하여, 목적 화자 특징 벡터(820)를 추출하는 과정은 도 4를 참조하여 상술한 방법과 동일하거나 유사하게 수행될 수 있다.

[111] 그런 다음, 정보 처리 시스템은 혼합 음성(810) 및 목적 화자 특징 벡터(820)를 목적 음성 추출 모델(800)에 입력하여, 목적 화자 특징 벡터(820)와 연관된 마스크(830)를 추정할 수 있다. 그런 다음, 혼합 음성(810)에 추정된 마스크(830)를 곱함으로써 목적 화자의 음성(840)을 추출할 수 있다.

[112] 정보 처리 시스템은 추출된 목적 화자의 음성(840)과 제2 목적 화자 참조 음성(812)의 비교에 기초하여, 예측 손실(Loss)(850)을 산출할 수 있다. 그런 다음, 산출된 예측 손실(850)을 이용하여 목적 음성 추출 모델(800)에 포함된 복수의 레이어들 또는 노드들 사이의 연결에 대응하는 가중치를 업데이트할 수 있다.

[113] 도 9 및 상술한 학습 방법은 일 예시일 뿐이며, 다른 실시예에서는 목적 음성 추출 모델(800)이 도시된 모델과 다르게 구현되거나 다른 방법에 의해 학습될 수 있다.

[114] 도 10은 본 개시의 일 실시예에 따른 화자 분할 방법(1000)의 예시를 나타내는 흐름도이다. 일 실시예에 따르면, 화자 분할 방법(1000)은 프로세서(예를 들어, 정보 처리 시스템 또는 사용자 단말의 적어도 하나의 프로세서)가 입력 신호에 포함된 음성 구간에 기초하여 제1 세트의 화자 특징 벡터를 추출함으로써 개시될 수 있다(S1010). 예를 들어, 프로세서는 입력 신호로부터 음성 구간을 검출하고, 검출된 음성 구간에 기초하여 제1 세트의 화자 특징 벡터를 추출할 수 있다. 여기서, 제1 세트의 화자 특징 벡터는 복수의 화자 특징 벡터를 포함할 수 있다. 일 실시예에 따르면, 프로세서는 입력 신호에 포함된 복수의 단위 음성 구간의 각각에 기초하여, 복수의 화자 특징 벡터를 추출할 수 있다.

[115] 그런 다음, 프로세서는 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출할 수 있다(S1020).

[116] 예를 들어, 프로세서는 먼저, 추출된 제1 세트의 화자 특징 벡터에 기초하여, 벡터 공간 중 제1 중심 영역을 결정할 수 있다. 일 실시예에서, 프로세서는 벡터 공간 내에서 화자 특징 벡터가 가장 밀집된 영역을 제1 중심 영역으로 결정할 수 있다. 예를 들어, 프로세서는 벡터 공간 내에서 추출된 제1 세트의 화자 특징 벡터의 각각을 중심으로 한 동일한 크기의 복수의 영역 중 가장 많은 수의 화자 특징 벡터를 포함하는 영역을 제1 중심 영역으로 결정할 수 있다.

[117] 제1 중심 영역이 결정된 다음, 프로세서는 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 제1 중심 특징 벡터를 추출할 수 있다. 예를 들어, 프

로세서는 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터의 평균을 제1 중심 특징 벡터로 결정할 수 있다. 다른 예로, 프로세서는 제1 중심 영역의 중심 또는 중심에 가장 근접하여 위치한 화자 특징 벡터를 제1 중심 특징 벡터로서 추출할 수 있다.

[118] 일 실시예에 따르면, 프로세서는 제1 중심 특징 벡터를 추출하기 전에, 추출된 제1 세트의 화자 특징 벡터에 대해 클러스터링을 수행할 수도 있다. 클러스터링이 수행된 경우, 프로세서는 클러스터링 결과 가장 많은 수의 화자 특징 벡터를 포함하는 클러스터로부터 제1 중심 특징 벡터를 추출할 수 있다. 예를 들어, 프로세서는 벡터 공간 중 클러스터링 결과 가장 많은 수의 화자 특징 벡터를 포함하는 클러스터에 대응되는 공간 내에서 제1 중심 영역을 결정하고, 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 제1 중심 특징 벡터를 추출할 수 있다.

[119] 그런 다음, 프로세서는 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출할 수 있다(S1030). 일 실시예에 따르면, 프로세서는 목적 음성 추출 모델을 이용하여, 입력 신호로부터 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출할 수 있다. 예를 들어, 프로세서는 목적 음원 추출 모델을 이용하여, 입력 신호 및 제1 중심 특징 벡터를 기초로 제1 중심 특징 벡터와 연관된 마스크를 추정할 수 있으며, 입력 신호 및 추정된 마스크에 기초하여, 제1 화자의 음성을 추출할 수 있다.

[120] 추가적으로, 프로세서는 입력 신호로부터 추출된 제1 화자의 음성을 제외한 잔여 신호에 음성 구간이 남아 있는지 여부를 판정할 수 있다. 잔여 신호에 음성 구간이 남아 있다고 판정되는 경우, 상술한 일련의 과정을 반복하여 수행함으로써 입력 신호 또는 잔여 신호에 포함된 다른 화자의 음성을 추출할 수 있다.

[121] 예를 들어, 프로세서는 단계 S1010으로 돌아가 일련의 과정을 반복하여 수행할 수 있다. 구체적으로, 프로세서는 잔여 신호에 포함된 음성 구간에 기초하여, 제2 세트의 화자 특징 벡터를 추출할 수 있다. 그런 다음, 추출된 제2 세트의 화자 특징 벡터에 기초하여 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호 또는 잔여 신호로부터 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.

[122] 다른 예로, 프로세서는 단계 S1020로 돌아가 일련의 과정을 반복하여 수행할 수 있다. 구체적으로, 프로세서는 제1 세트의 화자 특징 벡터 중 잔여 신호에 포함된 음성 구간을 기초로 추출된 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호 또는 잔여 신호로부터 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.

[123] 또 다른 예로, 프로세서는 입력 신호에 포함된 음성 구간 중 제1 화자의 음성이 포함된 음성 구간을 제외한 잔여 음성 구간에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하고, 입력 신호 또는 잔여 신호로부터 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출할 수 있다.

- [124] 도면에 포함된 흐름도 및 상술한 설명은 일 예시일 뿐이며 본 개시의 범위가 이에 한정되는 것은 아니다. 예를 들어, 다른 실시예에 따르면, 일부 단계가 추가/변경/삭제될 수 있으며, 각 단계의 순서가 변경될 수 있다.
- [125] 상술한 방법은 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램으로 제공될 수 있다. 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록 수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.
- [126] 본 개시의 방법, 동작 또는 기법들은 다양한 수단에 의해 구현될 수도 있다. 예를 들어, 이러한 기법들은 하드웨어, 펌웨어, 소프트웨어, 또는 이들의 조합으로 구현될 수도 있다. 본원의 개시와 연계하여 설명된 다양한 예시적인 논리적 블록들, 모듈들, 회로들, 및 알고리즘 단계들은 전자 하드웨어, 컴퓨터 소프트웨어, 또는 양자의 조합들로 구현될 수도 있음을 통상의 기술자들은 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 대체를 명확하게 설명하기 위해, 다양한 예시적인 구성요소들, 블록들, 모듈들, 회로들, 및 단계들이 그들의 기능적 관점에서 일반적으로 위에서 설명되었다. 그러한 기능이 하드웨어로서 구현되는지 또는 소프트웨어로서 구현되는지의 여부는, 특정 애플리케이션 및 전체 시스템에 부과되는 설계 요구사항들에 따라 달라진다. 통상의 기술자들은 각각의 특정 애플리케이션을 위해 다양한 방식으로 설명된 기능을 구현할 수도 있으나, 그러한 구현들은 본 개시의 범위로부터 벗어나게 하는 것으로 해석되어서는 안된다.
- [127] 하드웨어 구현에서, 기법들을 수행하는 데 이용되는 프로세싱 유닛들은, 하나 이상의 ASIC들, DSP들, 디지털 신호 프로세싱 디바이스들(digital signal processing devices; DSPD들), 프로그램가능 논리 디바이스들(programmable logic devices; PLD들), 필드 프로그램가능 게이트 어레이들(field programmable gate arrays; FPGA들), 프로세서들, 제어기들, 마이크로제어기들, 마이크로프로세서들, 전자 디바이스들, 본 개시에 설명된 기능들을 수행하도록 설계된 다른 전자 유닛들, 컴퓨터, 또는 이들의 조합 내에서 구현될 수도 있다.
- [128] 따라서, 본 개시와 연계하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 및 회로들은 범용 프로세서, DSP, ASIC, FPGA나 다른 프로그램 가능 논리 디바이스, 이산 게이트나 트랜지스터 로직, 이산 하드웨어 컴포넌트들, 또는 본원에 설명된 기능들을 수행하도록 설계된 것들의 임의의 조합으로 구현되거나 수행될

수도 있다. 범용 프로세서는 마이크로프로세서일 수도 있지만, 대안으로, 프로세서는 임의의 종래의 프로세서, 제어기, 마이크로제어기, 또는 상태 머신일 수도 있다. 프로세서는 또한, 컴퓨팅 디바이스들의 조합, 예를 들면, DSP와 마이크로프로세서, 복수의 마이크로프로세서들, DSP 코어와 연계한 하나 이상의 마이크로프로세서들, 또는 임의의 다른 구성의 조합으로서 구현될 수도 있다.

[129] 펌웨어 및/또는 소프트웨어 구현에 있어서, 기법들은 랜덤 액세스 메모리(random access memory; RAM), 판독 전용 메모리(read-only memory; ROM), 비휘발성 RAM(non-volatile random access memory; NVRAM), PROM(programmable read-only memory), EPROM(erasable programmable read-only memory), EEPROM(electrically erasable PROM), 플래시 메모리, 콤팩트 디스크(compact disc; CD), 자기 또는 광학 데이터 스토리지 디바이스 등과 같은 컴퓨터 판독가능 매체 상에 저장된 명령들로서 구현될 수도 있다. 명령들은 하나 이상의 프로세서들에 의해 실행 가능할 수도 있고, 프로세서(들)로 하여금 본 개시에 설명된 기능의 특정 양태들을 수행하게 할 수도 있다.

[130] 소프트웨어로 구현되는 경우, 상기 기법들은 하나 이상의 명령들 또는 코드로서 컴퓨터 판독 가능한 매체 상에 저장되거나 또는 컴퓨터 판독 가능한 매체를 통해 전송될 수도 있다. 컴퓨터 판독가능 매체들은 한 장소에서 다른 장소로 컴퓨터 프로그램의 전송을 용이하게 하는 임의의 매체를 포함하여 컴퓨터 저장 매체들 및 통신 매체들 양자를 포함한다. 저장 매체들은 컴퓨터에 의해 액세스될 수 있는 임의의 이용 가능한 매체들일 수도 있다. 비제한적인 예로서, 이러한 컴퓨터 판독가능 매체는 RAM, ROM, EEPROM, CD-ROM 또는 다른 광학 디스크 스토리지, 자기 디스크 스토리지 또는 다른 자기 스토리지 디바이스들, 또는 소망의 프로그램 코드를 명령들 또는 데이터 구조들의 형태로 이송 또는 저장하기 위해 사용될 수 있으며 컴퓨터에 의해 액세스될 수 있는 임의의 다른 매체를 포함할 수 있다. 또한, 임의의 접속이 컴퓨터 판독가능 매체로 적절히 칭해진다.

[131] 예를 들어, 소프트웨어가 동축 케이블, 광섬유 케이블, 연선, 디지털 가입자 회선(DSL), 또는 적외선, 무선, 및 마이크로파와 같은 무선 기술들을 사용하여 웹사이트, 서버, 또는 다른 원격 소스로부터 전송되면, 동축 케이블, 광섬유 케이블, 연선, 디지털 가입자 회선, 또는 적외선, 무선, 및 마이크로파와 같은 무선 기술들은 매체의 정의 내에 포함된다. 본원에서 사용된 디스크(disk)와 디스크(disc)는, CD, 레이저 디스크, 광 디스크, DVD(digital versatile disc), 플로피디스크, 및 블루레이 디스크를 포함하며, 여기서 디스크들(disks)은 보통 자기적으로 데이터를 재생하고, 반면 디스크들(disks)은 레이저를 이용하여 광학적으로 데이터를 재생한다. 위의 조합들도 컴퓨터 판독가능 매체들의 범위 내에 포함되어야 한다.

[132] 소프트웨어 모듈은, RAM 메모리, 플래시 메모리, ROM 메모리, EPROM 메모리, EEPROM 메모리, 레지스터들, 하드 디스크, 이동식 디스크, CD-ROM, 또는 공지된 임의의 다른 형태의 저장 매체 내에 상주할 수도 있다. 예시적인 저장 매체는, 프로세서가 저장 매체로부터 정보를 판독하거나 저장 매체에 정보를 기록할

수 있도록, 프로세서에 연결될 수 있다. 대안으로, 저장 매체는 프로세서에 통합될 수도 있다. 프로세서와 저장 매체는 ASIC 내에 존재할 수도 있다. ASIC은 유저 단말 내에 존재할 수도 있다. 대안으로, 프로세서와 저장 매체는 유저 단말에서 개별 구성요소들로서 존재할 수도 있다.

- [133] 이상 설명된 실시예들이 하나 이상의 독립형 컴퓨터 시스템에서 현재 개시된 주제의 양태들을 활용하는 것으로 기술되었으나, 본 개시는 이에 한정되지 않고, 네트워크나 분산 컴퓨팅 환경과 같은 임의의 컴퓨팅 환경과 연계하여 구현될 수도 있다. 또 나아가, 본 개시에서 주제의 양상들은 복수의 프로세싱 칩들이나 장치들에서 구현될 수도 있고, 스토리지는 복수의 장치들에 걸쳐 유사하게 영향을 받게 될 수도 있다. 이러한 장치들은 PC들, 네트워크 서버들, 및 휴대용 장치들을 포함할 수도 있다.
- [134] 본 명세서에서는 본 개시가 일부 실시예들과 관련하여 설명되었지만, 본 개시의 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 개시의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다. 또한, 그러한 변형 및 변경은 본 명세서에 첨부된 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

청구범위

- [청구항 1] 적어도 하나의 프로세서에 의해 수행되는, 화자 분할 방법에 있어서, 입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하는 단계;
상기 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하는 단계; 및
상기 입력 신호로부터 상기 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하는 단계를 포함하는, 화자 분할 방법.
- [청구항 2] 제1항에 있어서, 상기 제1 중심 특징 벡터를 추출하는 단계는, 상기 추출된 제1 세트의 화자 특징 벡터에 기초하여, 벡터 공간 중 제1 중심 영역을 결정하는 단계; 및
상기 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 상기 제1 중심 특징 벡터를 추출하는 단계를 포함하는, 화자 분할 방법.
- [청구항 3] 제2항에 있어서, 상기 제1 중심 영역을 결정하는 단계는, 상기 벡터 공간 내에서 상기 화자 특징 벡터가 가장 밀집된 영역을 상기 제1 중심 영역으로 결정하는 단계를 포함하는, 화자 분할 방법.
- [청구항 4] 제2항에 있어서, 상기 제1 중심 영역을 결정하는 단계는, 상기 벡터 공간 내에서 상기 추출된 제1 세트의 화자 특징 벡터의 각각을 중심으로 한 동일한 크기의 복수의 영역 중 가장 많은 수의 화자 특징 벡터를 포함하는 영역을 상기 제1 중심 영역으로 결정하는 단계를 포함하는, 화자 분할 방법.
- [청구항 5] 제2항에 있어서, 상기 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 제1 중심 특징 벡터를 추출하는 단계는, 상기 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터의 평균을 상기 제1 중심 특징 벡터로 결정하는 단계를 포함하는, 화자 분할 방법.
- [청구항 6] 제2항에 있어서, 상기 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 제1 중심 특징 벡터를 추출하는 단계는,

상기 제1 중심 영역의 중심 또는 중심에 가장 근접하여 위치한 화자 특징 벡터를 상기 제1 중심 특징 벡터로서 추출하는 단계
를 포함하는, 화자 분할 방법.

[청구항 7] 제1항에 있어서,
상기 추출된 제1 세트의 화자 특징 벡터에 대해 클러스터링을 수행하는 단계
를 더 포함하는, 화자 분할 방법.

[청구항 8] 제7항에 있어서,
상기 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하는 단계는,
상기 클러스터링 결과 가장 많은 수의 화자 특징 벡터를 포함하는 클러스터로부터 상기 제1 중심 특징 벡터를 추출하는 단계
를 포함하는, 화자 분할 방법.

[청구항 9] 제7항에 있어서,
상기 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하는 단계는,
벡터 공간 중 상기 클러스터링 결과 가장 많은 수의 화자 특징 벡터를 포함하는 클러스터에 대응되는 공간 내에서 제1 중심 영역을 결정하는 단계; 및
상기 제1 중심 영역에 포함된 하나 이상의 화자 특징 벡터에 기초하여, 상기 제1 중심 특징 벡터를 추출하는 단계
를 포함하는, 화자 분할 방법.

[청구항 10] 제1항에 있어서,
상기 입력 신호로부터 상기 추출된 제1 화자의 음성을 제외한 잔여 신호에 음성 구간이 남아 있는지 여부를 판정하는 단계
를 더 포함하는, 화자 분할 방법.

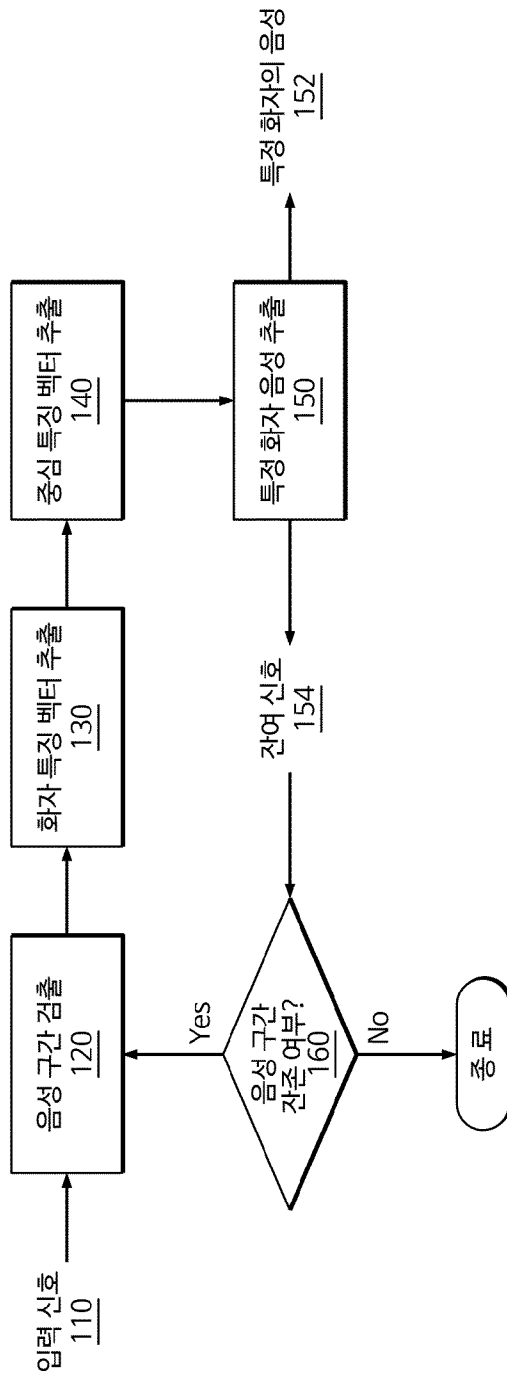
[청구항 11] 제10항에 있어서,
상기 잔여 신호에 음성 구간이 남아 있다고 판정되는 경우, 상기 제1 세트의 화자 특징 벡터 중 상기 잔여 신호에 포함된 음성 구간을 기초로 추출된 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하는 단계; 및
상기 입력 신호 또는 상기 잔여 신호로부터 상기 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출하는 단계
를 더 포함하는, 화자 분할 방법.

[청구항 12] 제10항에 있어서,
상기 잔여 신호에 포함된 음성 구간에 기초하여, 제2 세트의 화자 특징 벡터를 추출하는 단계;
상기 추출된 제2 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하는 단계; 및

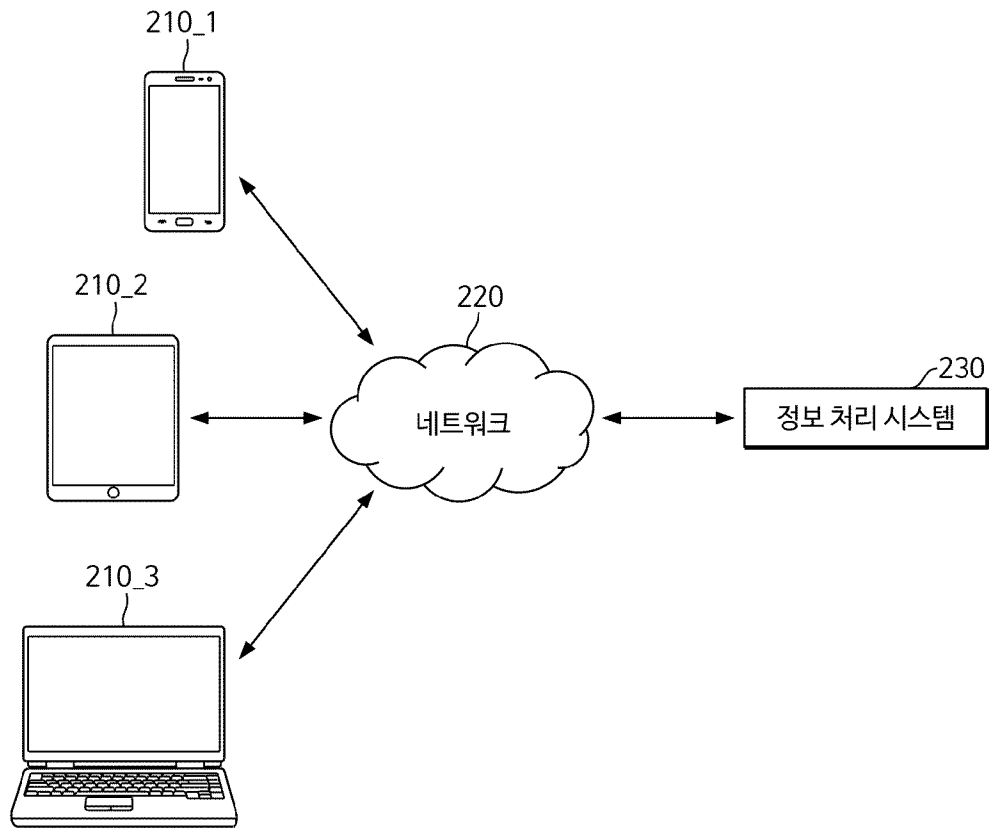
상기 입력 신호 또는 상기 잔여 신호로부터 상기 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출하는 단계를 더 포함하는, 화자 분할 방법.

- [청구항 13] 제10항에 있어서,
상기 잔여 신호에 음성 구간이 남아 있다고 판정되는 경우, 상기 입력 신호에 포함된 음성 구간 중 상기 제1 화자의 음성이 포함된 음성 구간을 제외한 잔여 음성 구간에 기초하여, 가장 높은 신뢰도를 갖는 제2 중심 특징 벡터를 추출하는 단계; 및
상기 입력 신호 또는 상기 잔여 신호로부터 상기 제2 중심 특징 벡터와 연관된 제2 화자의 음성을 추출하는 단계를 더 포함하는, 화자 분할 방법.
- [청구항 14] 제1항에 있어서,
입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하는 단계는,
상기 입력 신호에 포함된 복수의 단위 음성 구간의 각각에 기초하여, 상기 제1 세트의 화자 특징 벡터를 추출하는 단계를 포함하는, 화자 분할 방법.
- [청구항 15] 제1항에 있어서,
상기 입력 신호로부터 상기 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하는 단계는,
목적 음성 추출 모델을 이용하여, 상기 입력 신호로부터 상기 제1 중심 특징 벡터와 연관된 상기 제1 화자의 음성을 추출하는 단계를 포함하는, 화자 분할 방법.
- [청구항 16] 제1항에 따른 방법을 컴퓨터에서 실행하기 위한 명령어들을 기록한 컴퓨터 판독 가능한 비일시적 기록 매체.
- [청구항 17] 정보 처리 시스템으로서,
메모리; 및
상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서
를 포함하고,
상기 적어도 하나의 프로그램은,
입력 신호에 포함된 음성 구간에 기초하여, 제1 세트의 화자 특징 벡터를 추출하고,
상기 추출된 제1 세트의 화자 특징 벡터에 기초하여, 가장 높은 신뢰도를 갖는 제1 중심 특징 벡터를 추출하고,
상기 입력 신호로부터 상기 제1 중심 특징 벡터와 연관된 제1 화자의 음성을 추출하기 위한 명령어들을 포함하는, 정보 처리 시스템.

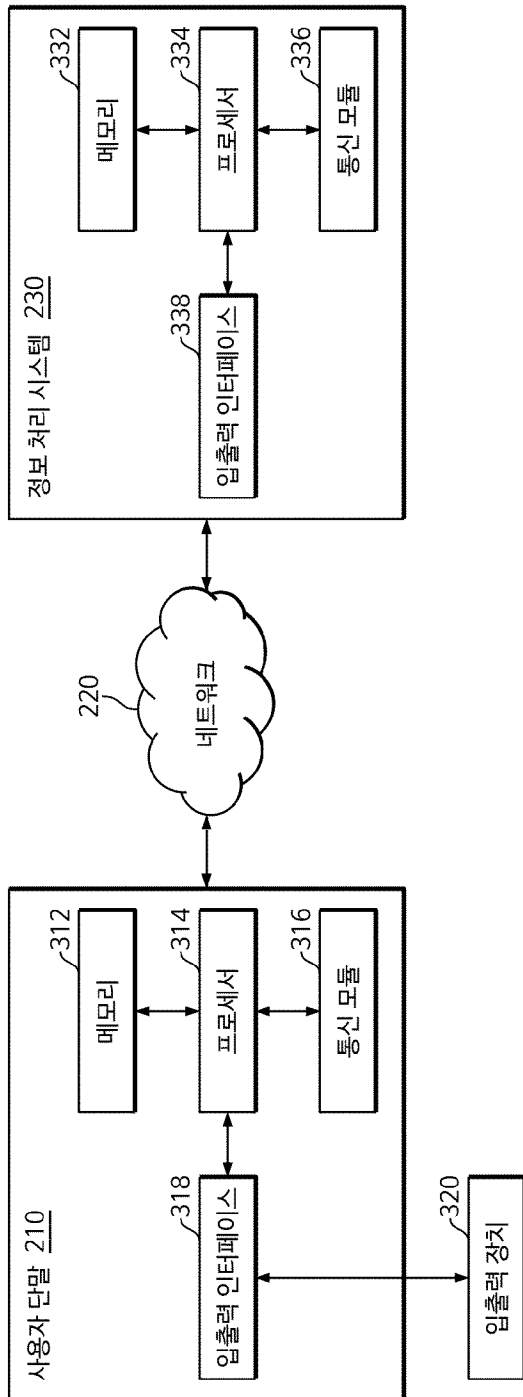
[도 1]



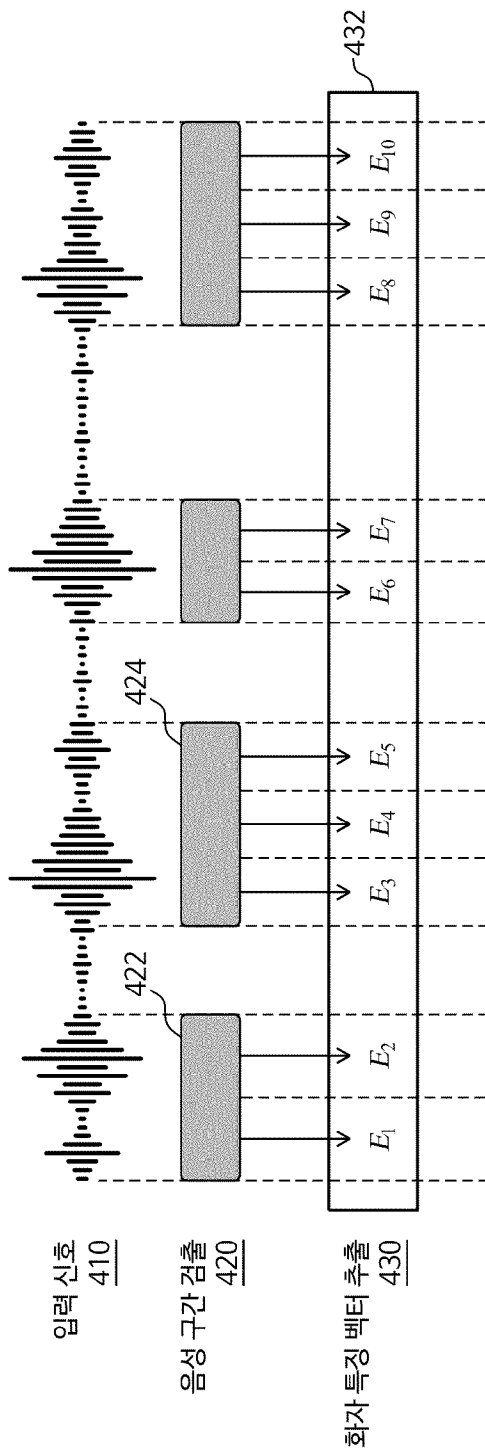
[도2]



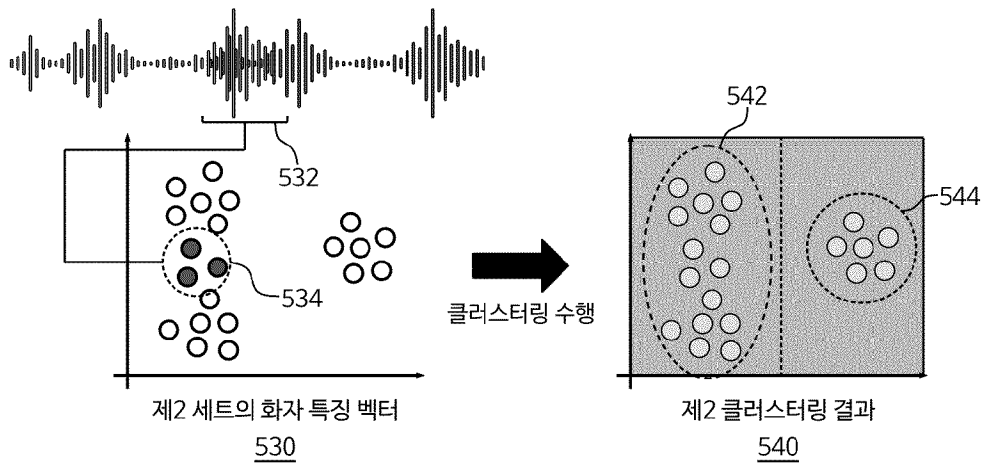
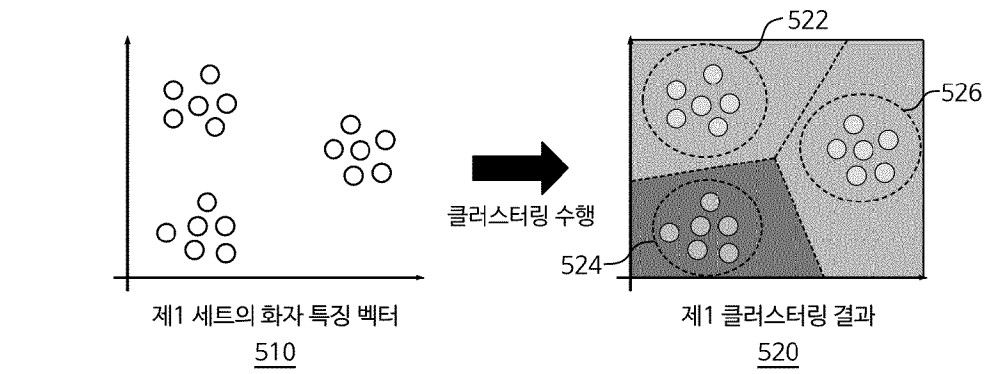
[도3]



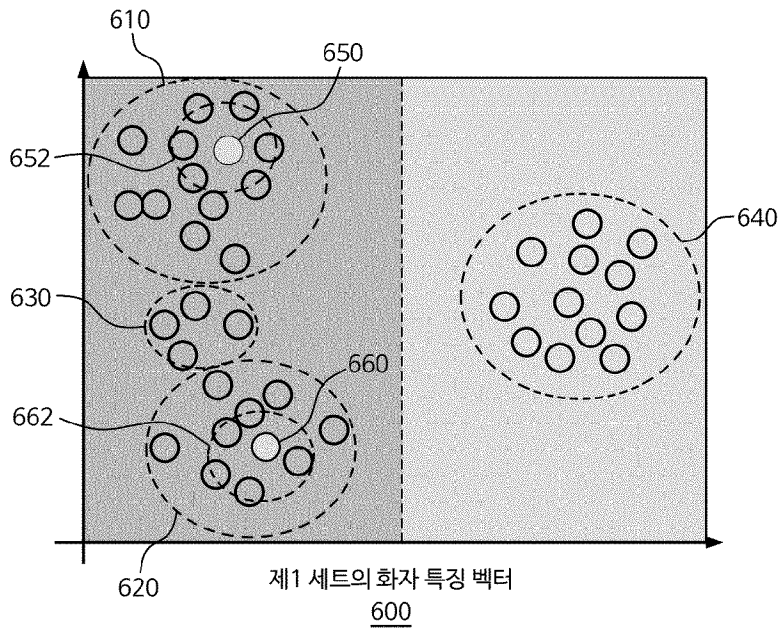
[도4]



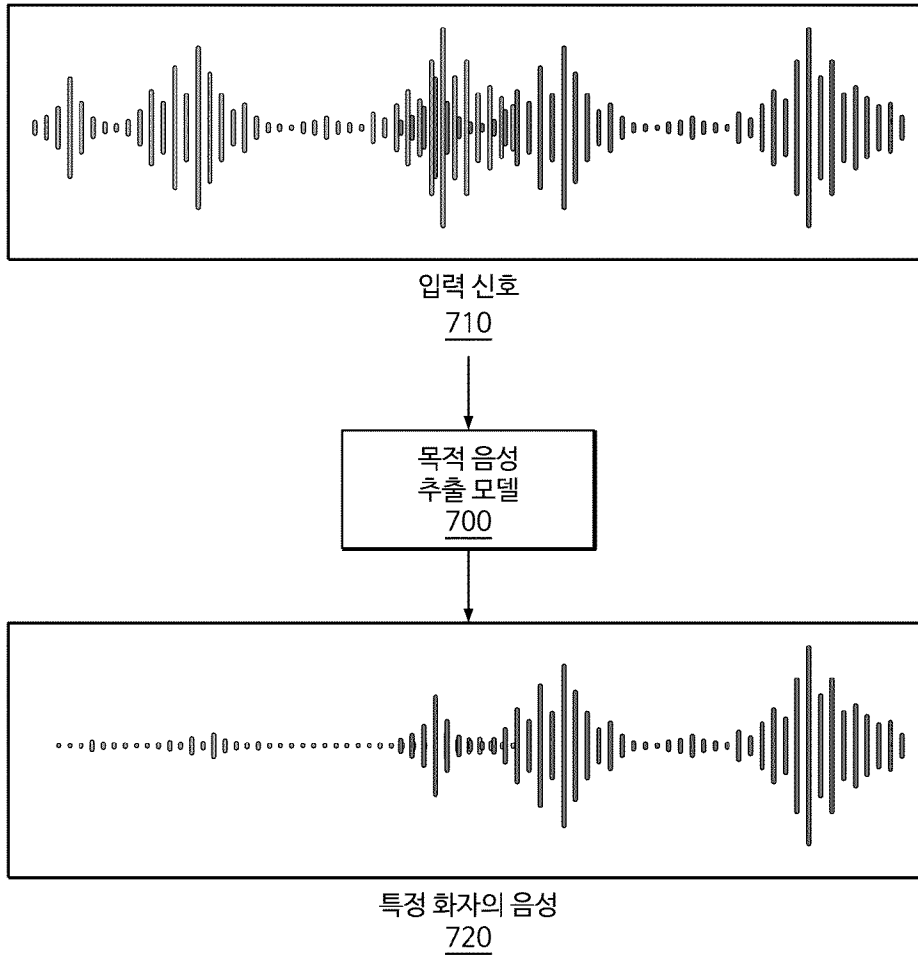
[도5]



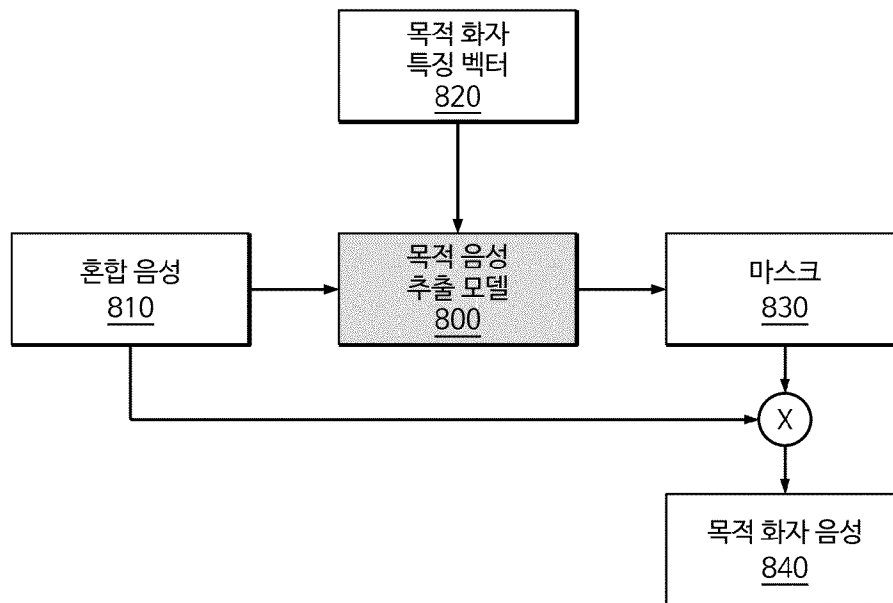
[도6]



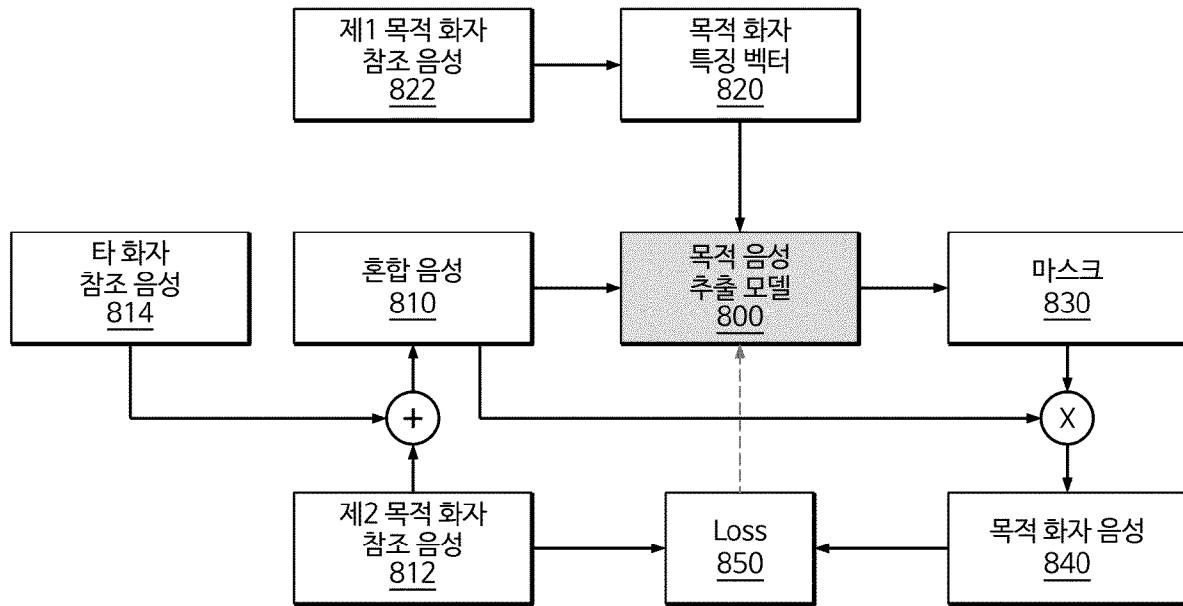
[도7]



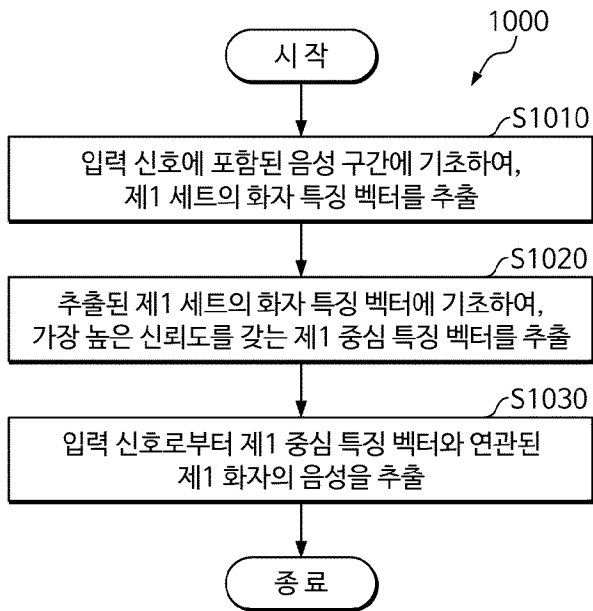
[도8]



[도9]



[도10]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2023/021003

A. CLASSIFICATION OF SUBJECT MATTER

G10L 17/02(2013.01)i; G10L 25/93(2013.01)i; G10L 15/02(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 17/02(2013.01); G10L 15/06(2006.01); G10L 17/00(2006.01); G10L 17/04(2013.01); G10L 17/14(2013.01)

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: 발화자(speaker), 식별(recognition), 중심(center), 벡터(vector), 클러스터(cluster), 신뢰도(credibility)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 6350148 B2 (FUJITSU LTD.) 04 July 2018 (2018-07-04) See claim 5.	1-17
Y	US 2013-0006633 A1 (GROKOP, Leonard Henry et al.) 03 January 2013 (2013-01-03) See claims 1 and 5.	1-17
Y	KR 10-2022-0103507 A (NAVER CORPORATION et al.) 22 July 2022 (2022-07-22) See claim 1.	10-13
A	JP 09-244683 A (SEIKO EPSON CORP.) 19 September 1997 (1997-09-19) See paragraphs [0063]-[0075]; and figures 8-11.	1-17
A	JP 5486565 B2 (NIPPON TELEGRAPH AND TELEPHONE CORP.) 07 May 2014 (2014-05-07) See claims 1-3.	1-17



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

“D” document cited by the applicant in the international application

“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

15 March 2024

Date of mailing of the international search report

15 March 2024

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex-Daejeon Building 4, 189 Cheongsaro, Seo-gu, Daejeon 35208

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/KR2023/021003

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
JP	6350148	B2	04 July 2018	JP	2016-057461	A	21 April 2016
				US	2016-0071520	A1	10 March 2016
				US	9536525	B2	03 January 2017
US	2013-0006633	A1	03 January 2013	WO	2013-006489	A1	10 January 2013
KR	10-2022-0103507	A	22 July 2022	JP	2022-109867	A	28 July 2022
				JP	7348445	B2	21 September 2023
				KR	10-2560019	B1	27 July 2023
				TW	202230342	A	01 August 2022
				US	2022-0230648	A1	21 July 2022
JP	09-244683	A	19 September 1997	JP	4027443	B2	26 December 2007
JP	5486565	B2	07 May 2014	JP	2013-037108	A	21 February 2013

A. 발명이 속하는 기술분류(국제특허분류(IPC)) G10L 17/02(2013.01)i; G10L 25/93(2013.01)i; G10L 15/02(2006.01)i		
B. 조사된 분야 조사된 최소문헌(국제특허분류를 기재) G10L 17/02(2013.01); G10L 15/06(2006.01); G10L 17/00(2006.01); G10L 17/04(2013.01); G10L 17/14(2013.01) 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: 발화자(speaker), 식별(recognition), 중심(center), 벡터(vector), 클러스터(cluster), 신뢰도(credibility)		
C. 관련 문헌		
카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	JP 6350148 B2 (FUJITSU LTD.) 2018.07.04 청구항 5	1-17
Y	US 2013-0006633 A1 (LEONARD HENRY GROKOP 등) 2013.01.03 청구항 1, 5	1-17
Y	KR 10-2022-0103507 A (네이버 주식회사 등) 2022.07.22 청구항 1	10-13
A	JP 09-244683 A (SEIKO EPSON CORP.) 1997.09.19 단락 [0063]-[0075]; 및 도면 8-11	1-17
A	JP 5486565 B2 (NIPPON TELEGRAPH AND TELEPHONE CORP.) 2014.05.07 청구항 1-3	1-17
☐ 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 이후에 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 2024년03월15일(15.03.2024)	국제조사보고서 발송일 2024년03월15일(15.03.2024)	
ISA/KR의 명칭 및 우편주소 대한민국 특허청 (35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사) 팩스 번호 +82-42-481-8578	심사관 양정록 전화번호 +82-42-481-5709	

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
JP 6350148 B2	2018/07/04	JP 2016-057461 A	2016/04/21
		US 2016-0071520 A1	2016/03/10
		US 9536525 B2	2017/01/03
US 2013-0006633 A1	2013/01/03	WO 2013-006489 A1	2013/01/10
KR 10-2022-0103507 A	2022/07/22	JP 2022-109867 A	2022/07/28
		JP 7348445 B2	2023/09/21
		KR 10-2560019 B1	2023/07/27
		TW 202230342 A	2022/08/01
		US 2022-0230648 A1	2022/07/21
JP 09-244683 A	1997/09/19	JP 4027443 B2	2007/12/26
JP 5486565 B2	2014/05/07	JP 2013-037108 A	2013/02/21