2024-12-19

【서지사항】

【서류명】 특허출원서

【참조번호】 2024P0150

【출원구분】 특허출원

【출원인】

【명칭】 서울대학교산학협력단

【특허고객번호】 1-2007-050924-2

【대리인】

【성명】 김건우

【대리인번호】 9-2003-000483-2

【발명의 국문명칭】 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법

【발명의 영문명칭】 SEGMENT-LEVEL SPEAKER DIARIZATION SYSTEM AND METHOD BASED ON SPEAKER STATE CHANGE DETECTION

【발명자】

【성명】 김남수

【성명의 영문표기】 Nam Soo Kim

【국적】 KR

【주민등록번호】 651018-1XXXXXX

【우편번호】 08826

【주소】 서울특별시 관악구 관악로 1 서울대학교

【거주국】 KR

【발명자】

2024-12-19

【성명】 문찬영

【성명의 영문표기】 Moon chanyeong

【국적】 KR

【주민등록번호】 970115-1XXXXXX

【우편번호】 06272

【주소】 서울특별시 강남구 논현로 218

【거주국】 KR

【출원언어】 국어

【심사청구】 청구

【이 발명을 지원한 국가연구개발사업】

【과제고유번호】 2710002086

【과제번호】 00235082

【부처명】 과학기술정보통신부

【과제관리(전문)기관명】 과학기술사업화진흥원

【연구사업명】 과학치안공공연구성과실용화촉진시범사업

【연구과제명】 성문분석 프로그램 고도화 기술 개발

【과제수행기관명】 (주)민트기술

【연구기간】 2024.01.01 ~ 2024.12.31

【취지】 위와 같이 특허청장에게 제출합니다.

대리인 김건우 (서명 또는 인)

【수수료】

【출원료】	0 면	46,000	원
【가산출원료】	35 면	0	원
【우선권주장료】	0 건	0	원
【심사청구료】	9 항	625,000	원
【합계】	671,000원		
【감면사유】	전담조직(50%감면)[1]		
【감면후 수수료】	335,500 원		
【첨부서류】	1.기타첨부서류[개별위임장]_1통		

1 : 기타첨부서류
PDF 파일 첨부

【발명의 설명】

【발명의 명칭】

화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법{SEGMENT-LEVEL SPEAKER DIARIZATION SYSTEM AND METHOD BASED ON SPEAKER STATE CHANGE DETECTION}

【기술분야】

【0001】 본 발명은 화자 분할 시스템 및 방법에 관한 것으로서, 보다 구체적으로는 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법에 관한 것이다.

【발명의 배경이 되는 기술】

【0002】 이 부분에 기술된 내용은 단순히 본 발명의 일실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

【0004】 본 발명은 딥 러닝 기반 화자 인식의 한 갈래인 화자 분할 기술 분야에 속하며, 다중 화자가 포함된 음성 데이터에서 시간에 따른 각 화자의 발화 여부를 정확하게 판별하기 위한 기술에 관한 것이다.

【0006】 화자 분할은 다수의 화자가 포함된 음성 신호에서 시간 순서에 따라 각 화자의 발화 여부를 판별하는 기술로, 입력 음성 신호에서 프레임 단위 특징을 추출하고 각 프레임마다 화자 존재 확률을 계산하는 과정을 포함한다. 기존 화자

분할 시스템은 입력 음성을 프레임 단위로 처리하여 특징 벡터를 생성한 후, 각 프레임의 화자 존재 확률을 독립적으로 예측하여 다중 화자를 구분하는 방식을 사용한다. 그러나 이러한 프레임 기반 접근 방식은 화자 상태가 변하지 않는 구간에서도 일관된 결과를 유지하기 어렵고 정확도가 저하되는 문제가 있다.

【0008】 딥러닝 기술의 발전으로 화자 분할 기술의 성능이 크게 향상되었으나, 여전히 기존 프레임 단위 접근 방식은 동일한 화자가 발화하는 구간에서도 결과의 일관성을 유지하기 어려운 한계를 가지고 있다. 이는 프레임 단위로 화자 존재 확률을 계산하기 때문에 발생하는 문제로, 구간 내에서의 화자 변화를 적절히 반영하지 못하기 때문이다. 따라서 이러한 한계를 극복하고 동일한 화자가 발화하는 구간에서 일관성을 유지할 수 있는 기술의 개발이 필요하다.

【0010】 한편, 본 발명과 관련된 선행기술로, 공개특허 제10-2022-0114378호(발명의 명칭: 텍스트 기반의 화자변경검출을 활용한 화자분할 보정 방법 및 시스템, 공개일자: 2022. 08. 17.) 등이 개시된 바 있다.

【0012】 전술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다.

【발명의 내용】

【해결하고자 하는 과제】

【0013】 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 입력 음성의 프레임 단위 특징에서 화자 상태 변화 확률을 계산하고, 화자 상태 변화 시점을 판별하여 화자가 바뀌지 않는 구간을 묶어서 구간 단위로 처리함으로써, 프레임 단위가 아닌 구간 단위에서 보다 정밀하게 화자 존재 확률을 계산하고, 다중 화자 환경에서도 정확하고 일관된 화자 인식 결과를 제공할 수 있는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법을 제공하는 것을 그 목적으로 한다.

【0015】 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수 있는 목적이나 효과도 이에 포함됨은 물론이다.

【과제의 해결 수단】

【0016】 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템은,

【0017】 화자 분할 시스템으로서,

【0018】 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산부;

【0019】상기 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별부;

【0020】상기 화자 상태 변화 시점 판별부의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음부;

【0021】상기 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제부; 및

【0022】상기 구간 단위 특징 벡터와 상기 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산부를 포함하는 것을 그 구성상의 특징으로 한다.

【0024】바람직하게는, 상기 유사도 계산부는,

【0025】상기 구간 단위 특징 벡터와 상기 화자 분할 임베딩 간의 행렬 곱을 계산하고, 계산 결과를 시그모이드 함수에 통과시켜 프레임마다 각 화자의 존재 확률을 계산할 수 있다.

【0027】바람직하게는,

【0028】입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 상기 프레임 단위 특징 벡터를 얻는 프레임 단위 특징 추출부를 더 포함할 수 있다.

【0030】 바람직하게는,

【0031】 상기 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력하는 프레임 단위 특징 요약부; 및

【0032】 상기 문맥 벡터를 입력받아 화자 수만큼의 상기 화자 분할 임베딩을 추출하는 화자 분할 임베딩 추출부를 더 포함할 수 있다.

【0034】 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법은,

【0035】 컴퓨터에서 각 단계가 수행되는 화자 분할 방법으로서,

【0036】 (1) 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산 단계;

【0037】 (2) 상기 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별 단계;

【0038】 (3) 상기 화자 상태 변화 시점 판별 단계의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음 단계;

【0039】 (4) 상기 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제 단계; 및

【0040】(5) 상기 구간 단위 특징 벡터와 상기 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산 단계를 포함하는 것을 그 구성상의 특징으로 한다.

【0042】 바람직하게는, 상기 유사도 계산 단계에서는,

【0043】 상기 구간 단위 특징 벡터와 상기 화자 분할 임베딩 간의 행렬 곱을 계산하고, 계산 결과를 시그모이드 함수에 통과시켜 프레임마다 각 화자의 존재 확률을 계산할 수 있다.

【0045】 바람직하게는, 상기 단계 (1) 이전에는,

【0046】(0) 입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 상기 프레임 단위 특징 벡터를 얻는 프레임 단위 특징 추출 단계를 더 포함할 수 있다.

【0048】 바람직하게는, 상기 단계 (5) 이전에는,

【0049】(a) 상기 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력하는 프레임 단위 특징 요약 단계; 및

【0050】(b) 상기 문맥 벡터를 입력받아 화자 수만큼의 상기 화자 분할 임베딩을 추출하는 화자 분할 임베딩 추출 단계를 더 포함할 수 있다.

【0052】상기한 목적을 달성하기 위한 본 발명의 특징에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법을 컴퓨터에서 실행시키기 위해 컴퓨터 판독 가능한 기록 매체에 저장되는 컴퓨터 프로그램을 포함하는 것을 그 구성상의 특징으로 한다.

【발명의 효과】

【0053】본 발명에서 제안하고 있는 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법에 따르면, 입력 음성의 프레임 단위 특징에서 화자 상태 변화 확률을 계산하고, 화자 상태 변화 시점을 판별하여 화자가 바뀌지 않는 구간을 묶어서 구간 단위로 처리함으로써, 프레임 단위가 아닌 구간 단위에서 보다 정밀하게 화자 존재 확률을 계산하고, 다중 화자 환경에서도 정확하고 일관된 화자 인식 결과를 제공할 수 있다.

【0055】더불어, 본 발명의 다양하면서도 유익한 장점과 효과는 상술한 내용에 한정되지 않으며, 본 발명의 구체적인 실시 형태를 설명하는 과정에서 보다 쉽게 이해될 수 있을 것이다.

【도면의 간단한 설명】

【0056】도 1은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 구성을 도시한 도면.

도 2는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템에서, 전처리 모듈의 세부적인 구성을 도시한 도면.

도 3은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템에서, 임베딩 모듈의 세부적인 구성을 도시한 도면.

도 4는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 전체 구조를 도시한 도면.

도 5는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법의 흐름을 도시한 도면.

도 6은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법에서, 전처리 과정의 세부적인 흐름을 도시한 도면.

도 7은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법에서, 임베딩 과정의 세부적인 흐름을 도시한 도면.

【발명을 실시하기 위한 구체적인 내용】

【0057】 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

【0059】 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이

에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한, 이하에서 기재되는 "포함하다", "구비하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것으로 해석되어야 하며, 하나 또는 그 이상의 다른 특징들이나, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다. 또한, 본 발명에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

【0061】 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.

【0063】 또한, 명세서에 기재된 "…부", "…기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

【0065】 또한, 본 발명에 있어서 단말, 장치 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말, 장치 또는 디바이스와 연결된 서버에서 대신 수행될 수 있다. 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말, 장치 또는 디바이스에서 수행될 수도 있다.

【0067】특히, 본 발명의 각 실시예에 따른 시스템을 실행시키기 위한 수단으로는 애플리케이션(Application), 또는 웹 서버일 수 있으며, 이 애플리케이션, 또는 웹 서버를 기록한 기록매체를 읽을 수 있는 수단인 단말로는, 일반적인 데스크톱이나 노트북 등의 일반 PC뿐만 아니라, 스마트 폰, 태블릿 PC, 등의 모바일 단말기를 포함할 수 있다.

【0069】이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

【0071】이하, 첨부된 도면을 참고하여 본 발명의 실시예들을 상세히 설명하도록 한다.

【0073】도 1은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 구성을 도시한 도면이다. 도 1에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템은, 화자 상태 변화 확률 계산부(110), 화자 상태 변화 시점 판별부(120), 동일 화자 구간 묶음부(130), 동일 화자 구간 복제부(140) 및 유사도 계산부(150)를 포함하여 구성될 수 있으며, 이를 화자 분할 모듈로 구성할 수 있다.

【0075】 기존의 화자 분할 기법은 프레임 단위에서 화자 존재 확률을 계산하여 처리하는데, 이에 따라 연속적으로 같은 화자가 발화하는 구간에서 화자 판별의 일관성이 떨어지는 문제가 발생할 수 있다. 본 발명은 이러한 문제를 해결하기 위해 화자 상태 변화를 감지하는 부분을 추가적으로 도입하여, 화자의 상태가 변하는 시점을 감지하고 동일한 화자 상태가 유지되는 구간으로 프레임들을 묶어 처리하는 방식을 채택하였다. 이를 통해 기존의 프레임 단위 접근 방식보다 더 높은 정확성과 안정성을 제공하며, 다중 화자 환경에서도 화자 인식 성능을 크게 향상시킬 수 있다.

【0077】 도 1에 도시된 바와 같이, 화자 분할 모듈은 전처리 모듈에서 전처리된 프레임 단위 특징 벡터를 전달받으며, 임베딩 모듈의 화자 분할 임베딩을 전달받아 유사도 계산부(150)를 통해 화자 존재 확률을 출력할 수 있다.

【0079】 도 2는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템에서, 전처리 모듈의 세부적인 구성을 도시한 도면이다. 도 2에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템은, 전처리부(11) 및 프레임 단위 특징 추출부(12)를 포함하여 전처리 모듈을 구성할 수 있다.

【0081】도 3은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템에서, 임베딩 모듈의 세부적인 구성을 도시한 도면이다. 도 3에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템은, 프레임 단위 특징 요약부(210) 및 화자 분할 임베딩 추출부(220)를 포함하여 임베딩 모듈을 구성할 수 있다.

【0083】도 4는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 전체 구조를 도시한 도면이다. 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 핵심적인 부분은 화자 상태 변화 확률 계산부(110)로, 이 구성은 각 프레임마다 화자가 변화하는 시점인지에 대한 확률을 계산하여 결과값으로 출력할 수 있다. 그 다음 화자 상태 변화 시점 판별부(120)가, 화자 상태 변화 확률 계산부(110)에서 출력된 확률값을 분석하여, 일정한 역치를 초과하는 경우 화자 상태가 변했다고 판단할 수 있다. 이러한 판단 과정을 통해 화자 상태 변화 시점을 파악할 수 있다. 이후, 동일 화자가 발화하는 구간에 대해서는 평균 풀링(mean pooling) 방법을 활용하여, 각 구간에 대한 프레임 단위 특징 벡터들의 평균을 계산할 수 있다. 이를 통해 동일 화자 구간 묶음부(130)에서는 각 구간마다 구간을 대표하는 하나의 특징 벡터를 생성할 수 있다. 생성된 벡터는 동일 화자 구간 복제부(140)에서 복제되어 원래의 프레임 길이로 복원될 수 있다. 이렇게 복원된 구간 단위 특징 벡터는 이후 유사도 계산부(150)에서 활용될 수 있다.

【0085】 한편, 이러한 과정과 동시에, 프레임 단위 특징 벡터가 프레임 단위 특징 요약부(210)를 통해 전체 음성에 대한 집약적인 정보를 가지고 있는 문맥 벡터를 생성해 화자 분할 임베딩 추출부(220)에 전달할 수 있다. 화자 분할 임베딩 추출부(220)는 전달받은 문맥 벡터를 통해 입력 음성에 대한 화자 분할 임베딩 벡터를 추출할 수 있다. 이러한 화자 분할 임베딩 벡터는 유사도 계산부(150)에서 구간 단위 특징 벡터와의 곱셈 연산과 시그모이드(sigmoid) 함수를 통해 유사도를 계산하며, 계산된 유사도를 바탕으로 각 프레임마다 화자 존재 확률을 산출할 수 있다. 이러한 과정을 통해 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템은 다중 화자가 발화하는 복잡한 환경에서도 더 정확하고 안정적인 화자 분할을 할 수 있다.

【0087】 본 발명은 기존의 프레임 단위 기반 화자 분할 방식에서 발생하는 여러 문제점을 해결하여 다양한 장점을 제공할 수 있다. 먼저, 화자 상태 변화 감지기를 통해 화자의 상태 변화 시점을 정확히 판별하고, 동일한 화자 상태가 유지되는 구간을 그룹화하여 처리함으로써 결과의 일관성을 개선할 수 있다. 또한, 구간 단위 접근 방식과 평균 풀링 기법을 활용하여 각 구간에 대해 평균적인 화자 존재 확률을 계산함으로써, 다중 화자 환경에서도 높은 화자 인식 정확도를 유지할 수 있으며, 다중 화자가 포함된 복잡한 환경에서도 안정적이고 일관된 화자 분할 결과를 제공할 수 있다.

【0089】이하에서는, 도 1 내지 도 4를 참고하여, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템의 각 구성에 대해 상세히 설명하도록 한다.

【0091】전처리부(11)는, 입력 신호로부터 음성 특징 벡터를 생성할 수 있다. 즉, 전처리부(11)에서는 입력된 음성 특징 벡터(예를 들어, MFCC, STFT, Mel Filter Bank 등)로부터 프레임 단위 특징 정보 출력을 계산한다. 음성 특징 벡터들은 입력 wave 신호에 미리 설정된 길이(예를 들어, 32ms)의 윈도우를 씌워 일정 길이로 이동(shift)시켜 처리되며, 화자 $s_1, s_2, \cdots, s_N$의 음성이 혼합된 형태의 입력 음성(입력 신호)으로부터 추출된 음성 특징 벡터들을 $x_t (0 \leq t \leq T)$라 정의할 수 있다. 전처리부(11)는, 음성 특징 벡터들을 만들기 위해 각 프레임마다 추출한 MFCC에 대해 앞과 뒤에 존재하는 n개의 프레임만큼 이어 붙여 총 $2n+1$개의 프레임에 대한 정보를 음성 특징 벡터 $x_t (0 \leq t \leq T)$가 가질 수 있다.

【0093】프레임 단위 특징 추출부(12)는, 입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 얻을 수 있다. 즉, 전처리부(11)에서 생성된 음성 특징 벡터를, 프레임 단위 특징 추출부(12)에서 한 번의 가공 과정을 더 거쳐 프레임 단위 특징을 나타내는 벡터인 $e_t^{frame} (0 \leq t \leq T)$를 얻을 수 있다. 도 2 및 도 4에 도시된 바와 같이, 프레임 단위 특징 벡터는 화자 분할

모듈의 화자 상태 변화 확률 계산부(110)과 임베딩 모듈의 프레임 단위 특징 요약부(210)에 입력되어, 2가지 갈래의 입력으로 사용될 수 있다. 이하에서는, 먼저 화자 분할 모듈을 설명한다.

【0095】 화자 상태 변화 확률 계산부(110)는, 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산할 수 있다. 즉, 화자 상태 변화 확률 계산부(110)는, 프레임 단위 특징 추출부(12)로부터 프레임 단위 특징 벡터 $e_t^{frame}(0 \leq t \leq T)$를 전달받아, 각 프레임마다 화자 상태 변화의 확률이 얼마인지에 대한 결과값인 $p_t^{change}(0 \leq t \leq T)$를 얻을 수 있다.

【0097】 화자 상태 변화 시점 판별부(120)는, 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별할 수 있다. 즉, 화자 상태 변화 시점 판별부(120)에서는 각 시점에 대한 화자 상태 변화의 확률이 일정 역치값 τ를 넘기는지 확인하여, 그 값을 넘기면 그 프레임 $i_l(0 \leq l \leq L)$에서 화자가 변화한다고 판별하게 되고, 그 판별 값으로 1을 할당해 총 L개의 구간으로 음성을 나눌 수 있다. 이때, 화자 상태 변화 확률이 역치를 넘기지 않는 경우에는 반대로 변화하지 않는다고 판별하며 그 판별 값을 0으로 만들 수 있다.

【0099】 동일 화자 구간 묶음부(130)는, 화자 상태 변화 시점 판별부(120)의

판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출할 수 있다. 즉, 화자 상태 변화 시점 판별부(120)의 판별 값이 1이 나온 시점을 기준으로 다음 1이 나오기 전까지의 구간을 화자가 바뀌지 않는 구간(또는 '동일 화자 구간')이라고 명명할 수 있으며, 그 구간마다 다음 수학식 1과 같은 풀링의 계산을 통해 구간의 평균적인 특징을 나타내는 구간 특징 벡터를 추출할 수 있다.

【0101】 **【수학식 1】**

$$e_l^{pooling} = \frac{1}{i_{l+1} - i_l} \sum_{t=i_l}^{i_{l+1}} e_t^{frame}$$

【0103】 동일 화자 구간 복제부(140)는, 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성할 수 있다. 즉, 동일 화자 구간 묶음부(130)의 풀링을 통해 프레임 단위였던 특징 벡터의 길이가 구간 단위로 길이가 변화하는 일이 발생하게 되는데, 다시 길이를 복원시켜 주고자 동일 화자 구간 복제부(140)에서 구간 특징 벡터를 원래 길이만큼 복제해주는, $e_t^{segment} = e_l^{pooling} (i_l \leq t \leq i_{l+1})$ 형식으로 길이를 복원할 수 있다. 이를 통해 구간 단위 특징 벡터인 $e_t^{segment} (0 \leq t \leq T)$를 추출해낼 수 있다.

【0105】 한편, 도 2 및 도 4에 도시된 바와 같이, 프레임 단위 특징 벡터는 전술한 바와 같은 화자 분할 모듈 외에, 임베딩 모듈에도 입력될 수 있다. 다른 갈래의 경우 $e_t^{frame}(0 \leq t \leq T)$는 화자 분할 임베딩을 만들기 위해 프레임 단위 특징 요약부(210)와 화자 분할 임베딩 추출부(220)를 통과할 수 있다. 프레임 단위 특징 요약부(210)와 화자 분할 임베딩 추출부(220)는 RNN(recurrent neural network)의 동일한 구조를 가질 수 있다. 이하에서는 도 3 및 도 4를 참고하여 임베딩 모듈에 대해 상세히 설명하도록 한다.

【0107】 프레임 단위 특징 요약부(210)는, 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력할 수 있다. 즉, 프레임 단위 특징 요약부(210)는 $e_t^{frame}(0 \leq t \leq T)$를 입력으로 받아 혼합 화자 입력 음성에 대한 요약된 정보를 담고 있는 문맥 벡터(context vector) h_T, c_T를 출력할 수 있다.

【0109】 화자 분할 임베딩 추출부(220)는, 문맥 벡터를 입력받아 화자 수만큼의 화자 분할 임베딩을 추출할 수 있다. 즉, 프레임 단위 특징 요약부(210)에서 요약된 h_T, c_T는 화자 분할 임베딩 추출부(220)의 입력으로 활용되어, 화자 수만큼의 화자 분할 임베딩인 $a_s(1 \leq s \leq N)$을 추출할 수 있다.

【0111】유사도 계산부(150)는, 구간 단위 특징 벡터와 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득할 수 있다. 보다 구체적으로, 유사도 계산부(150)는, 구간 단위 특징 벡터와 화자 분할 임베딩 간의 행렬 곱을 계산하고, 계산 결과를 시그모이드 함수에 통과시켜 프레임마다 각 화자의 존재 확률을 계산할 수 있다.

【0113】즉, 유사도 계산부(150)는 화자 분할 임베딩 $a_s(1 \le s \le N)$와 구간 단위 특징 벡터 $e_t^{segment}(0 \le t \le T)$ 간의 행렬 곱과 시그모이드 함수를 씌우는 2가지의 과정을 통해 각 프레임마다 화자 존재 확률을 계산할 수 있다. 보다 구체적으로, 특정한 시간 t에 대해 특정한 화자 s가 발화하는 확률을 구하고자 유사도를 얻는 과정을 살펴보면, $e_t^{segment}(0 \le t \le T)$는 $1 \times F$ (단, F는 각 프레임 단위 특징 벡터의 크기) 형태를 갖는 벡터이고, $a_s(1 \le s \le N)$는 $1 \times F$ 형태의 벡터이기 때문에, 두 벡터를 행렬 곱을 하게 되면 $p_{t,s} = e_t \cdot a_s^T$로 표현될 수 있다. 이때, $p_{t,s}$는 $1 \times F$과 $F \times 1$의 행렬 곱이기 때문에 1×1 크기의 스칼라(scalar)값이며, 이를 시그모이드 함수에 통과 시킴으로써 $P_{t,s} = \sigma(p_{t,s})$의 형태의 0과 1 사이의 값을 얻을 수 있다. 따라서 $P_{t,s}$는 특정 t 시점에 화자 s가 존재할 확률을 의미하고 이를 일반화하여 전체 확률값인 P에 대해 표현하면 다음 수학식 2와 같이 표현할 수 있다.

【0115】【수학식 2】

$$P = \sigma([e_1 e_2 e_3 \cdots e_T]\,[a_1 a_2 a_3 \cdots a_N]^T)$$

【0117】즉, 최종적인 결과로 $T \times S$ 크기의 $0 \sim T$ 시간에서 $s_1 \sim s_N$ 화자에 대한 존재 확률을 획득할 수 있다.

【0119】도 5는 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법의 흐름을 도시한 도면이다. 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법은, 컴퓨터에서 각 단계가 수행되는 화자 분할 방법으로서, 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산 단계(S110); 계산한 화자 상태 변화 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별 단계(S120); 화자 상태 변화 시점 판별 단계(S120)의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음 단계(S130); 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제 단계(S140); 및 구간 단위 특징 벡터와 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산 단계(S150)를 포함하여 구현

될 수 있다.

【0121】 도 6은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법에서, 전처리 과정의 세부적인 흐름을 도시한 도면이다. 도 6에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법은, 단계 S110 및 단계 S210 이전에, 입력 신호로부터 음성 특징 벡터를 생성하는 전처리 단계(S11), 및 입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 얻는 프레임 단위 특징 추출 단계(S12)를 더 포함하여 구현될 수 있다.

【0123】 도 7은 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법에서, 임베딩 과정의 세부적인 흐름을 도시한 도면이다. 도 7에 도시된 바와 같이, 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법은, 단계 S150 이전에, 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력하는 프레임 단위 특징 요약 단계(S210); 및 문맥 벡터를 입력받아 화자 수만큼의 화자 분할 임베딩을 추출하는 화자 분할 임베딩 추출 단계(S220)를 더 포함하여 구현될 수 있다.

【0125】 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템과 관련하여 충분히

설명되었으므로, 상세한 설명은 생략하기로 한다.

【0127】 한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위해 컴퓨터 판독 가능한 기록매체에 저장되는 컴퓨터 프로그램을 제공하는 것을 그 구성상의 특징으로 한다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 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 장치를 포함할 수 있다.

【0129】 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

【0131】 전술한 바와 같이, 본 발명에서 제안하고 있는 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법에 따르면, 입력 음성의 프레임 단위 특징에서 화자 상태 변화 확률을 계산하고, 화자 상태 변화 시점을 판별하여 화자가 바뀌지 않는 구간을 묶어서 구간 단위로 처리함으로써, 프레임 단위가 아닌 구간 단위에서 보다 정밀하게 화자 존재 확률을 계산하고, 다중 화자 환경에서도 정확하고 일관된 화자 인식 결과를 제공할 수 있다.

【0133】 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명된 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명된 구성 요소들도 결합된 형태로 실시될 수 있다.

【0135】 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

【부호의 설명】

【0136】11: 전처리부

12: 프레임 단위 특징 추출부

110: 화자 상태 변화 확률 계산부

120: 화자 상태 변화 시점 판별부

130: 동일 화자 구간 묶음부

140: 동일 화자 구간 복제부

150: 유사도 계산부

220: 화자 분할 임베딩 추출부

S11: 전처리 단계

S12: 프레임 단위 특징 추출 단계

S110: 화자 상태 변화 확률 계산 단계

S120: 화자 상태 변화 시점 판별 단계

S130: 동일 화자 구간 묶음 단계

S140: 동일 화자 구간 복제 단계

S150: 유사도 계산 단계

S210: 프레임 단위 특징 요약 단계

S220: 화자 분할 임베딩 추출 단계

【청구범위】

【청구항 1】

화자 분할 시스템으로서,

입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산부(110);

상기 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별부(120);

상기 화자 상태 변화 시점 판별부(120)의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음부(130);

상기 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제부(140); 및

상기 구간 단위 특징 벡터와 상기 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산부(150)를 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템.

【청구항 2】

제1항에 있어서, 상기 유사도 계산부(150)는,

상기 구간 단위 특징 벡터와 상기 화자 분할 임베딩 간의 행렬 곱을 계산하고, 계산 결과를 시그모이드 함수에 통과시켜 프레임마다 각 화자의 존재 확률을 계산하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템.

【청구항 3】

제1항에 있어서,

입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 상기 프레임 단위 특징 벡터를 얻는 프레임 단위 특징 추출부(12)를 더 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템.

【청구항 4】

제1항에 있어서,

상기 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력하는 프레임 단위 특징 요약부(210); 및

상기 문맥 벡터를 입력받아 화자 수만큼의 상기 화자 분할 임베딩을 추출하는 화자 분할 임베딩 추출부(220)를 더 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템.

【청구항 5】

컴퓨터에서 각 단계가 수행되는 화자 분할 방법으로서,

(1) 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산 단계(S110);

(2) 상기 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별 단계(S120);

(3) 상기 화자 상태 변화 시점 판별 단계(S120)의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음 단계(S130);

(4) 상기 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제 단계(S140); 및

(5) 상기 구간 단위 특징 벡터와 상기 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산 단계(S150)를 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법.

【청구항 6】

제5항에 있어서, 상기 유사도 계산 단계(S150)에서는,

상기 구간 단위 특징 벡터와 상기 화자 분할 임베딩 간의 행렬 곱을 계산하고, 계산 결과를 시그모이드 함수에 통과시켜 프레임마다 각 화자의 존재 확률을

계산하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법.

【청구항 7】

제5항에 있어서, 상기 단계 (1) 이전에는,

(0) 입력되는 음성 특징 벡터를 이용해 프레임 단위의 특징을 나타내는 상기 프레임 단위 특징 벡터를 얻는 프레임 단위 특징 추출 단계(S12)를 더 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법.

【청구항 8】

제5항에 있어서, 상기 단계 (5) 이전에는,

(a) 상기 프레임 단위 특징 벡터를 입력받아 전체 음성에 대한 요약된 정보를 담고 있는 문맥 벡터를 출력하는 프레임 단위 특징 요약 단계(S210); 및

(b) 상기 문맥 벡터를 입력받아 화자 수만큼의 상기 화자 분할 임베딩을 추출하는 화자 분할 임베딩 추출 단계(S220)를 더 포함하는 것을 특징으로 하는, 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법.

【청구항 9】

제5항 내지 제8항 중 어느 한 항의 화자 상태 변화 감지 기반 구간 단위 화자 분할 방법을 컴퓨터에서 실행시키기 위해 컴퓨터 판독 가능한 기록 매체에 저장

되는 컴퓨터 프로그램.

【요약서】

【요약】

본 발명은 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템에 관한 것으로서, 보다 구체적으로는 화자 분할 시스템으로서, 입력 음성의 프레임 단위의 특징을 나타내는 프레임 단위 특징 벡터를 이용해 프레임마다 화자 상태 변화의 확률을 계산하는 화자 상태 변화 확률 계산부; 상기 계산한 화자 상태 변화의 확률을 역치값과 비교해 화자 상태가 변화하는 프레임을 판별하는 화자 상태 변화 시점 판별부; 상기 화자 상태 변화 시점 판별부의 판별 결과에 따라, 화자가 바뀌지 않는 구간마다 평균 풀링을 적용해 구간 특징 벡터를 추출하는 동일 화자 구간 묶음부; 상기 구간 특징 벡터를 복제해 원래 프레임 길이로 복원하여 구간 단위 특징 벡터를 생성하는 동일 화자 구간 복제부; 및 상기 구간 단위 특징 벡터와 상기 프레임 단위 특징 벡터에서 추출된 화자 분할 임베딩을 결합해 각 프레임 화자에 대한 존재 확률을 획득하는 유사도 계산부를 포함하는 것을 그 구성상의 특징으로 한다.

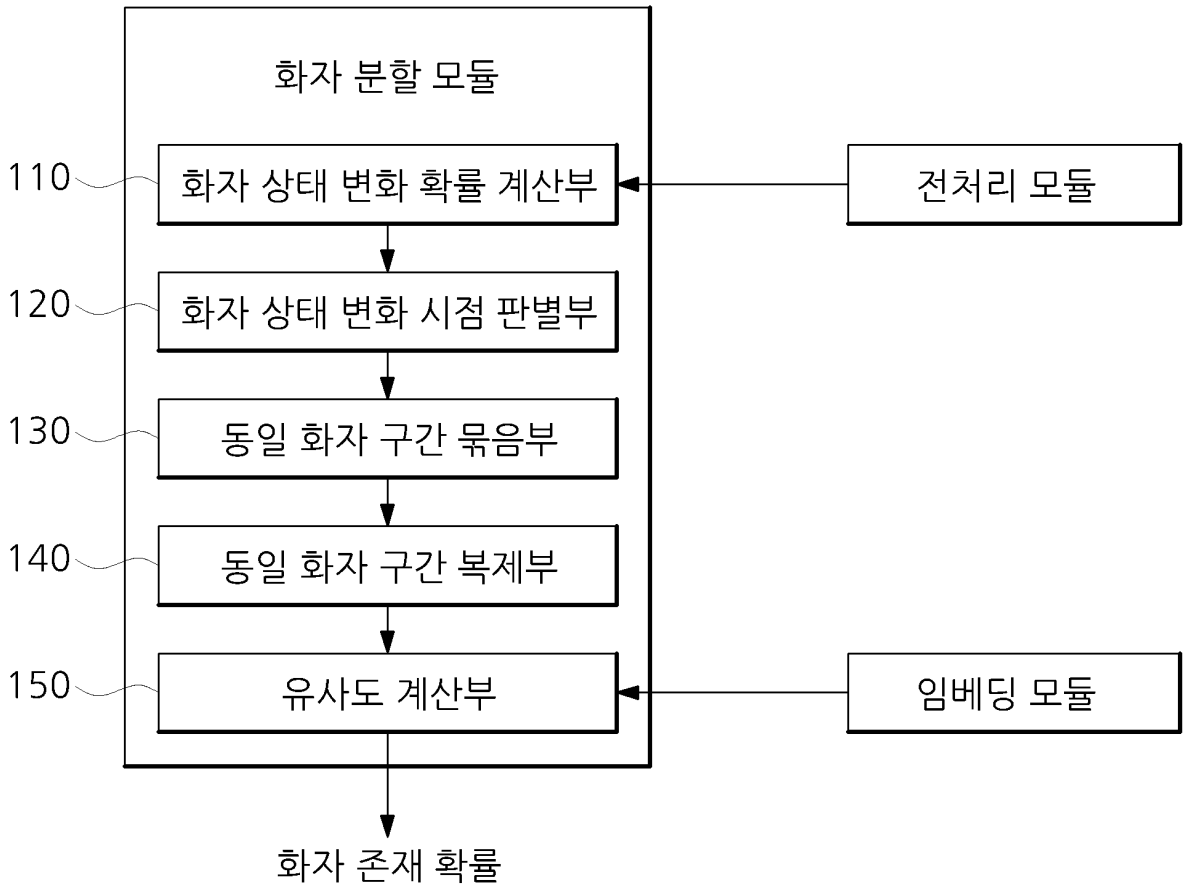
본 발명에서 제안하고 있는 화자 상태 변화 감지 기반 구간 단위 화자 분할 시스템 및 방법에 따르면, 입력 음성의 프레임 단위 특징에서 화자 상태 변화 확률을 계산하고, 화자 상태 변화 시점을 판별하여 화자가 바뀌지 않는 구간을 묶어서 구간 단위로 처리함으로써, 프레임 단위가 아닌 구간 단위에서 보다 정밀하게 화자 존재 확률을 계산하고, 다중 화자 환경에서도 정확하고 일관된 화자 인식 결과를 제공할 수 있다.

【대표도】

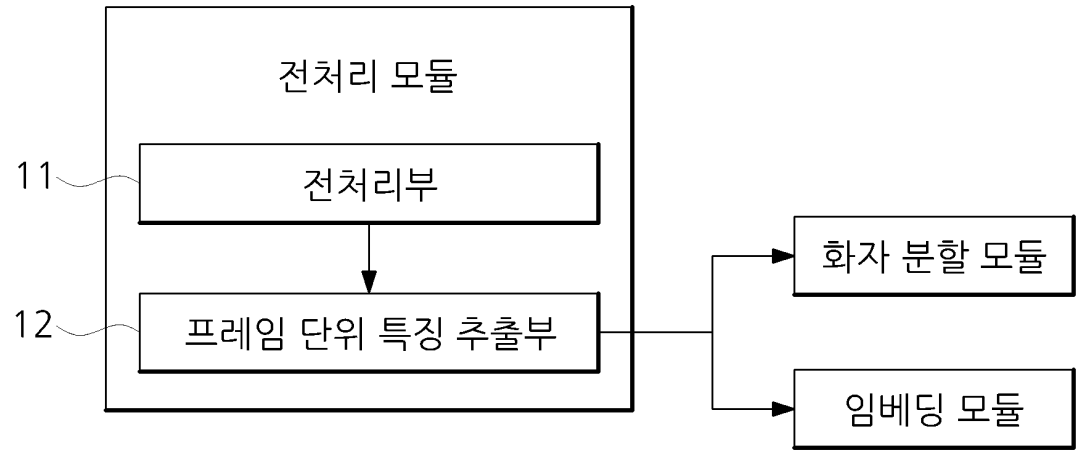
도 1

【도면】

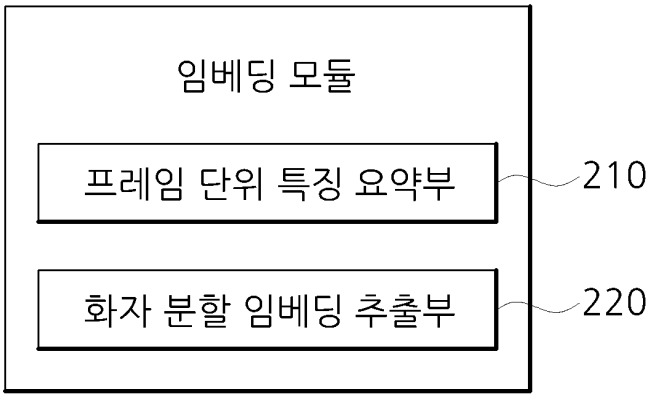
【도 1】


화자 분할 모듈
110
화자 상태 변화 확률 계산부
120
화자 상태 변화 시점 판별부
130
동일 화자 구간 묶음부
140
동일 화자 구간 복제부
150
유사도 계산부
전처리 모듈
임베딩 모듈
화자 존재 확률

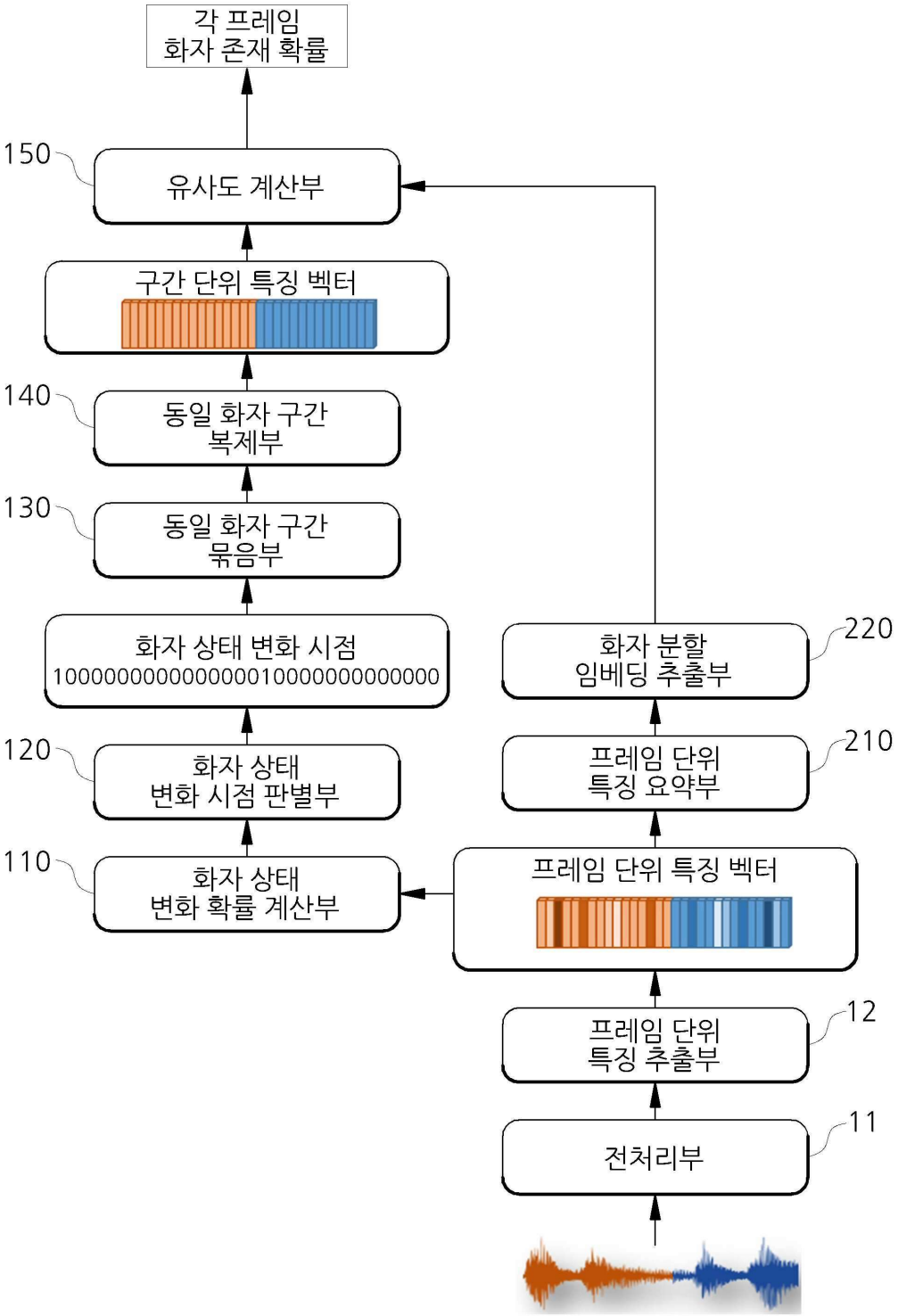
【도 2】


전처리 모듈
11
전처리부
12
프레임 단위 특징 추출부
화자 분할 모듈
임베딩 모듈

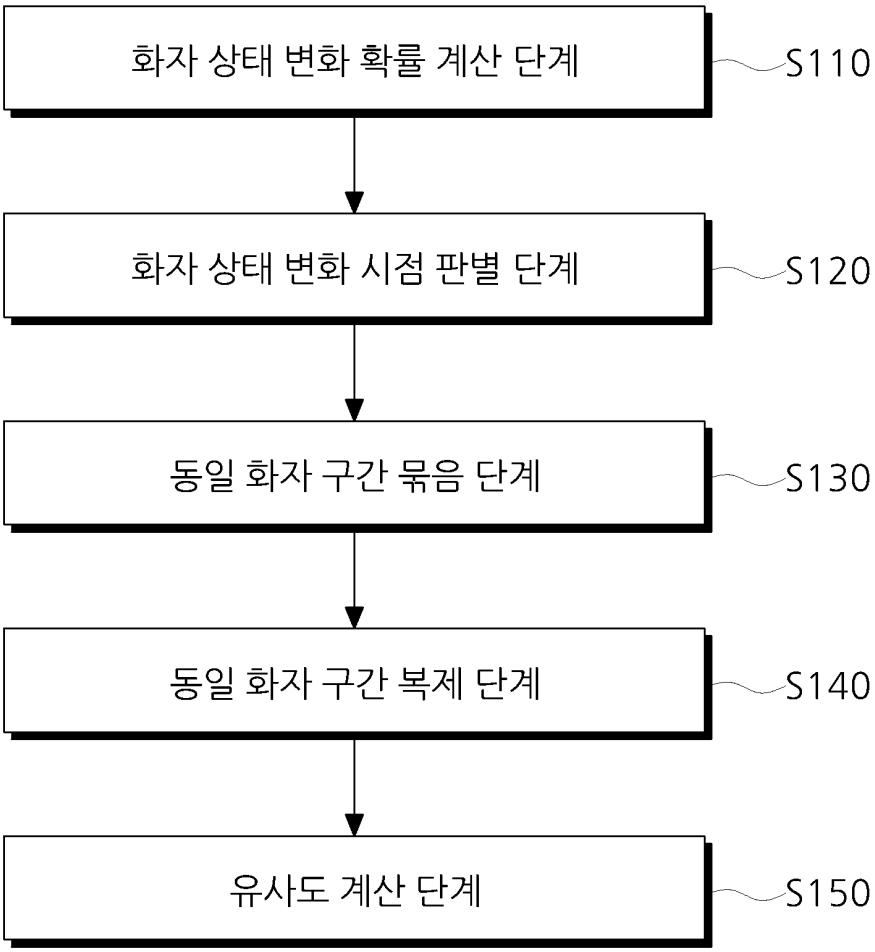
【도 3】



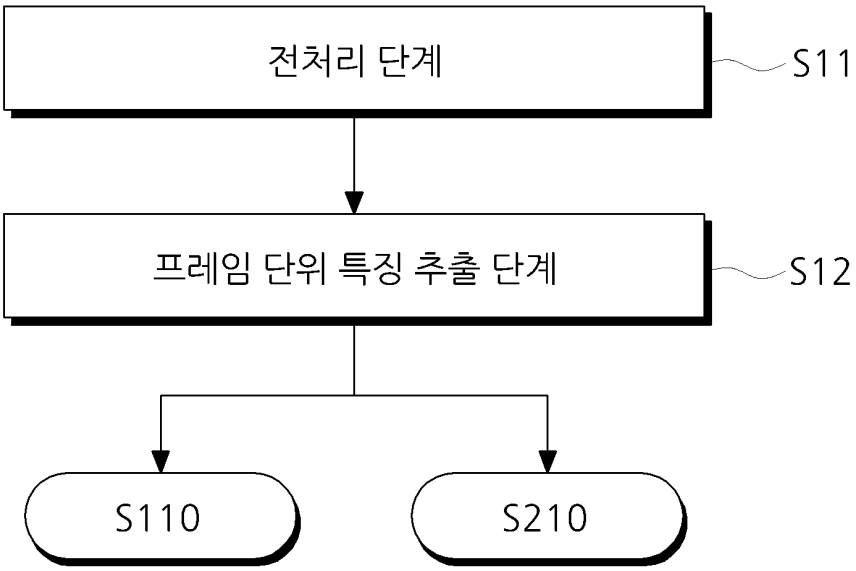
【도 4】


각 프레임
화자 존재 확률
150
유사도 계산부
구간 단위 특징 벡터
140
동일 화자 구간
복제부
130
동일 화자 구간
묶음부
화자 상태 변화 시점
10000000000000001000000000000
120
화자 상태
변화 시점 판별부
110
화자 상태
변화 확률 계산부
화자 분할
임베딩 추출부
220
프레임 단위
특징 요약부
210
프레임 단위 특징 벡터
프레임 단위
특징 추출부
12
전처리부
11

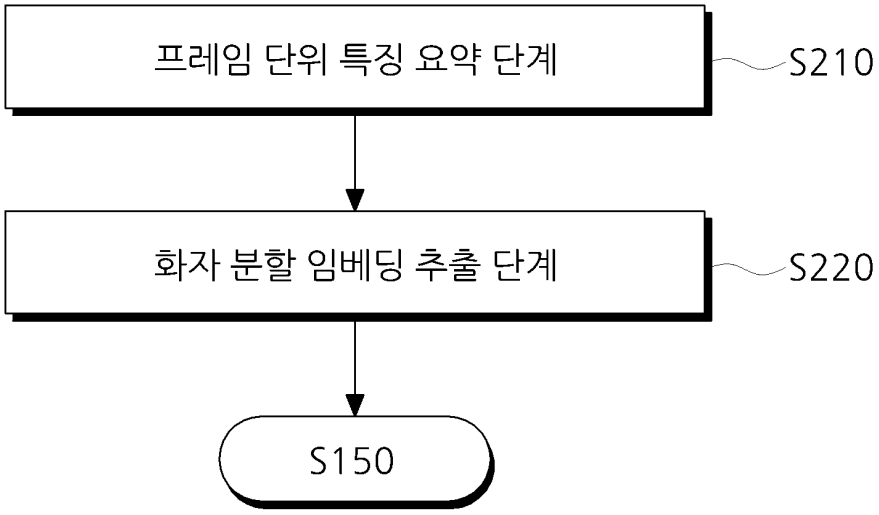
【도 5】


화자 상태 변화 확률 계산 단계
S110
화자 상태 변화 시점 판별 단계
S120
동일 화자 구간 묶음 단계
S130
동일 화자 구간 복제 단계
S140
유사도 계산 단계
S150

【도 6】


전처리 단계
S11
프레임 단위 특징 추출 단계
S12
S110
S210

【도 7】


프레임 단위 특징 요약 단계
S210
화자 분할 임베딩 추출 단계
S220
S150