(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2016년12월20일
(11) 등록번호 10-1688240
(24) 등록일자 2016년12월14일

(51) 국제특허분류(Int. Cl.)
*G10L 15/26* (2006.01) *G10L 15/16* (2006.01)
(21) 출원번호 10-2011-7013340
(22) 출원일자(국제) 2009년11월12일
심사청구일자 2014년10월27일
(85) 번역문제출일자 2011년06월10일
(65) 공개번호 10-2011-0095314
(43) 공개일자 2011년08월24일
(86) 국제출원번호 PCT/US2009/064214
(87) 국제공개번호 WO 2010/056868
국제공개일자 2010년05월20일
(30) 우선권주장
12/616,723 2009년11월11일 미국(US)
61/113,910 2008년11월12일 미국(US)
(56) 선행기술조사문헌
US05799276 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
**에스씨티아이 홀딩스, 인크.**
미국 매릴랜드 20910 실버 스프링 수트 520 펜톤 스트리트 8630
(72) 발명자
**핀손, 마크**
미국 캘리포니아 91344 그라나다 힐즈 리날디 스트리트 6721
**핀손, 데이비드, 에스알.**
미국 메릴랜드 20720 보위 하이 브리지 로드 7101
*(뒷면에 계속)*
(74) 대리인
**홍순우, 김해중**

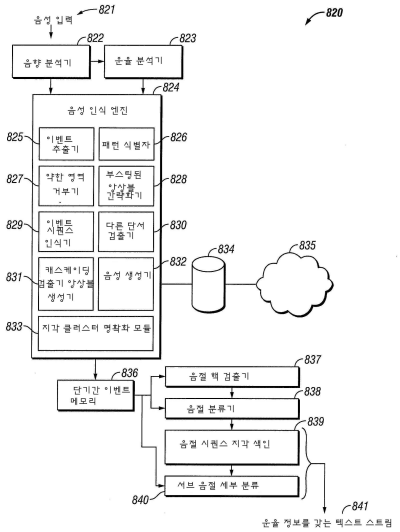
전체 청구항 수 : 총 19 항 심사관 : 정성윤

(54) 발명의 명칭 **자동 음성-텍스트 변환을 위한 시스템 및 방법**

## (57) 요 약

음성 인식은 거의 실시간으로 수행되고 이벤트 및 이벤트 시퀀스를 이용하고, 부스팅된 분류기, 앙상블, 검출기 및 캐스케이드를 포함하는 머신 학습 기법들을 이용하고, 지각 클러스터를 이용함으로써 향상된다. 음성 인식은 또한 탠덤 프로세싱을 이용하여 향상된다. 자동 분리기는 구두점을 인식된 텍스트 스트림에 주입한다.

*대 표 도* - 도8b



(72) 발명자

**플라나간, 메리**

미국 메릴랜드01701 프레이밍햄 빌링스 웨이 4

**마칸반드, 샤로크**

미국 캘리포니아 91390 산타 클라리타 로렐 플레이스 22414

# 명 세 서

## 청구범위

### 청구항 1

디지털 음성 신호에 대응하는 음성을 인식하기 위한 시스템에 있어서,

상기 시스템은, 음성 인식 엔진, 및 음성 인식 엔진에 결합되는 모듈을 포함하고,

상기 음성 인식 엔진은,

음성 인식 엔진에 연결된 적어도 하나의 프로세서; 및

디지털 음성 신호로부터 음성 신호 이벤트와 음성 신호 이벤트의 패턴을 추출하기 위한 이벤트 추출기를 포함하고,

음성 음식 엔진과 연결된 적어도 하나의 프로세서는,

알려진 클래스의 디지털화된 말(known-class digitized speech utterances)의 트레이닝 코퍼스(training corpus)와,

복수의 약 분류기를 포함하는 앙상블 검출기(ensemble detector)로서, 복수의 약 분류기의 각각은 트레이닝 코퍼스 내의 이벤트의 존재를 검출하기 위한 판정 기능을 포함하여 음성 신호 이벤트의 존재 검출시, 약 분류기들이 함께 동작할 때가 혼자 동작할 때보다 더 효과적인 앙상블 검출기에 액세스하도록 구성되고,

음성 음식 엔진과 연결된 적어도 하나의 프로세서는, 복수의 연산(operations)을 수행하도록 구성되고,

상기 복수의 연산은,

디지털 음성 신호 내에서 연관된 음성 신호 이벤트의 위치를 검출하는 연산 - 여기서 음성 신호 이벤트의 각각은 스펙트럼 정보(spectral information)와 시간 정보(temporal information)를 포함하고 있음 - ;

모든 음성 신호 이벤트들 사이의 시간 관계(temporal relation)와 모든 음성 신호 이벤트들의 스펙트럼 피처(features)를 캡처하는 연산;

검출된 음성 신호 이벤트의 검출 위치에 기반하여 디지털 음성 신호를 세그먼트화하는 연산;

세그먼트화된 디지털 음성 신호를 분석하는 연산 - 여기서 분석은 음성 신호 이벤트와 동기화됨 - ;

캡처된 스펙트럼 정보, 시간 관계, 및 분석된 디지털 음성 신호로 디지털 음성 신호 내 패턴을 검출하는 연산 - ;

디지털 음성 신호 내의 검출된 패턴에 대응하는 인식된 음성 데이터에 대한 지각적 대안(perceptual alternative)의 리스트(list)를 제공하는 연산; 및

향상된 인식 음성 데이터를 얻기 위해, 음성 신호 이벤트의 하나 이상의 분석에 기반하여, 인식된 음성 데이터에 대한 지각적 대안 사이를 구별하는(disambiguating) 연산을 포함하고,

적어도 하나의 프로세서는 액세스된 앙상블 검출기를 이용하여 복수의 연산 중 하나 이상을 수행하도록 구성되고

상기 모듈은 향상된 인식 음성 데이터를 출력하도록 구성된 것을 특징으로 하는

시스템.

### 청구항 2

제1항에 있어서,

향상된 인식 음성 데이터 출력의 적어도 일부분에 응답하여 적어도 하나의 동작(action)을 개시하기 위한 메커

니즘을 더 포함하는

시스템.

**청구항 3**

제2항에 있어서,

적어도 하나의 동작은,

향상된 인식 음성 데이터를 적어도 하나의 텍스트 스트림(text stream)으로 변환하는 변환; 및

소정의 단어(word)가 검출될 때 오디오 출력을 억제하는 억제; 중 하나 이상을 포함하는

시스템.

**청구항 4**

제2항에 있어서,

향상된 인식 음성 데이터 내의 적어도 하나의 명령어를 검출하기 위한 메커니즘을 더 포함하고,

적어도 하나의 동작은, 검출된 명령어에 대한 응답을 개시하는 것인

시스템.

**청구항 5**

제1항에 있어서,

복수의 약 분류기를 포함하는 앙상블 검출기는 적어도 하나의 프로세서에 의해 구현(constructed)되는

시스템.

**청구항 6**

제5항에 있어서,

적어도 하나의 프로세서는 부스트된 앙상블 검출기를 구성하기 위해 부스팅 알고리즘(boosting algorithm)으로 앙상블 검출기를 반복적으로 구성하도록 이루어진

시스템.

**청구항 7**

제6항에 있어서,

적어도 하나의 프로세서는 구성된 부스트 앙상블 검출기를 간략화하도록 구성된

시스템.

**청구항 8**

제7항에 있어서,

적어도 하나의 프로세서는 간략화된 구성된 부스트 앙상블 검출기를 캐스케이딩(cascading) 검출기로 변환하도

록 구성된

시스템.

**청구항 9**

제1항에 있어서,

인식 음성 데이터에 대한 지각적 대안의 리스트는 복수의 지각 클러스터(perceptual cluster를 포함하는

시스템.

**청구항 10**

제1항에 있어서,

적어도 하나의 프로세서는 음성 신호 이벤트의 하나 이상을 포함하지 않은 디지털 음성 신호의 하나 이상의 영역(regions)을 거부(reject)하도록 더 구성된

시스템

**청구항 11**

제1항에 있어서,

적어도 하나의 프로세서는 검출된 패턴에 기반하여 음성 신호 이벤트의 시퀀스를 검출하도록 더 구성된

시스템.

**청구항 12**

제1항에 있어서,

적어도 하나의 프로세서는 인식 강화를 위해 대안적인 음성 큐(speech cue)를 인식하도록 더 구성된

시스템.

**청구항 13**

제1항에 있어서,

사전 세그먼트화 필터(pre-segmentation filter); 및 피처 추출기를 더 포함하고,

사전 세그먼트화 필터는 피처 계산(feature computation)을 동기화하는데 이용되는 간격(interval)을 정의하도록 이루어지고,

디지털 음성 신호의 세그먼트화는 정의된 간격의 지각적 차이(perceptual differences)에 기반하며,

피처 추출기는 세그먼트화된 디지털 음성 신호로부터 음성 신호 이벤트에 대한 피처를 추출하도록 구성된,

시스템.

**청구항 14**

제1항에 있어서,

적어도 하나의 프로세서는

향상된 인식 음성 데이터를 적어도 하나의 텍스트 스트림으로 변환하고,

적어도 하나의 텍스트 스트림 내에 구두점을 자동으로 삽입하도록 더 구성된

시스템.

**청구항 15**

음성 인식 방법에 있어서,

음향 분석기(acoustical analyzer)에서,

음성 신호를 수신하는 단계;

수신된 음성 신호를 디지털화하는 단계;

적어도 하나의 프로세서를 포함하는 음성 인식 엔진에서,

복수의 약 분류기를 포함하는 앙상블 검출기(ensemble detector)를 액세스하는 단계 - 여기서, 복수의 약 분류기의 각각은 알려진 클래스의 디지털화된 말의 트레이닝 코퍼스 내의 이벤트의 존재를 검출하기 위한 판정 기능을 포함하여 음성 신호 이벤트의 존재 검출시, 약 분류기들이 함께 동작할 때가 혼자 동작할 때보다 더 효과적임 - ;

디지털 음성 신호 내에서 연관된 음성 신호 이벤트의 위치를 검출하는 단계 - 여기서 음성 신호 이벤트의 각각은 스펙트럼 정보(spectral information)와 시간 정보(temporal information)를 포함하고 있음 - ;

모든 음성 신호 이벤트들 사이의 시간 관계(temporal relation)와 모든 음성 신호 이벤트들의 스펙트럼 피처(features)를 캡처하는 단계;

음성 신호 이벤트의 검출 위치에 기반하여 디지털 음성 신호를 세그먼트화하는 단계;

세그먼트화된 디지털 음성 신호를 분석하는 단계 - 여기서 분석은 음성 신호 이벤트와 동기화됨 - ;

캡처된 스펙트럼 정보, 시간 관계, 및 분석된 디지털 음성 신호로 디지털 음성 신호 내 패턴을 검출하는 단계 - ;

분석된 디지털 음성 신호에 대응하는 음성 데이터를 인식하는 단계로서,

· 디지털 음성 신호 내의 검출된 패턴에 대응하는 인식된 음성 데이터에 대한 지각적 대안(perceptual alternative)의 리스트(list)를 제공하는 단계; 및

· 향상된 인식 음성 데이터를 얻기 위해, 음성 신호 이벤트의 하나 이상의 분석에 기반하여, 인식된 음성 데이터에 대한 지각적 대안 사이를 구별하는(disambiguating) 단계를 포함하는 인식 단계; 및

음성 인식 엔지에 연결된 모듈에서, 향상된 인식 음성 데이터를 출력하는 단계;

를 포함하는 것을 특징으로 하는 음성 인식 방법.

**청구항 16**

제15항에 있어서,

적어도 하나의 프로세서를 포함하는 음성 인식 엔진에서,

복수의 약 분류기를 형성(establishing)하는 단계; 및

앙상블 검출기를 구성(constructing)하는 단계를 더 포함하고,

상기 앙상블 검출기를 구성하는 단계는,

복수의 음성 신호를 저장하는 단계 - 음성 신호는 자동 음성 인식 시스템 내에 저장된 복수의 저장된 트레이닝

실례(training examples)를 포함함 - ;

복수의 저장된 트레이닝 실례로부터 이벤트 패턴을 추출하는 단계 - 이벤트 패턴은 저장된 복수의 음성 신호 내에 구별되는 특징적 위치(distinctive characteristic location)를 포함함 - ;

반복적인 수행 단계를 포함하고, 반복적인 수행 단계는,

· 매칭 이벤트 패턴을 가진 복수의 음성 신호의 샘플에 액세스하는 단계;

· 샘플 중에서 개별 음성 신호로부터의 이벤트를 정렬하는 단계 - 상기 정렬은 상기 매칭 이벤트 패턴에 기초하여 상기 개별 음성 신호로부터의 상기 이벤트를 시간적으로 라이닝 업(lining up)하는 것을 포함함 - ;

· 이벤트 패턴의 검출시 복수의 약 검출기의 유효성을 평가하는 단계;

· 약 검출기의 상대적인 유효성에 기초하여 복수의 약 검출기에 가중 방식을 적용하는 단계 - 가장 유효한 약한 검출기에 가장 높은 가중치가 부여됨 - ;

· 복수의 약 검출기에 적어도 하나의 부가적인 약 검출기를 추가하는 단계를 반복적으로 수행하고,

상기 반복적인 수행 단계는 가중 방식의 유효성이 이벤트 패턴 검출에 대한 효율의 설정 표준까지 작동될 때까지 수행되는

음성 인식 방법.

**청구항 17**

제16항에 있어서,

매칭 이벤트 패턴을 가진 복수의 음성 신호의 샘플에 액세스 하는 단계는, 이벤트 패턴을 포함하는 복수의 음성 신호 내의 영역을 자동으로 식별하는 단계를 더 포함하고,

영역을 자동으로 식별하는 단계는,

복수의 음성 신호를 공통 시간축에 대해 정렬하는 단계;

개별 음성 신호의 하나 이상의 이벤트 위치를 공통 시간축 상에 투영하는 단계; 및

이벤트 패턴을 포함하는 복수의 음성 신호 내의 영역의 형태로, 이벤트 위치가 집중된 시간축 상의 영역을 식별하는 단계를 포함하는

음성 인식 방법.

**청구항 18**

제16항에 있어서,

매칭 이벤트 패턴을 가진 복수의 음성 신호의 샘플에 액세스 하는 단계는, 이벤트 패턴을 포함하는 복수의 음성 신호 내의 영역을 자동으로 식별하는 단계를 더 포함하고,

영역을 자동으로 식별하는 단계는,

트레이닝 세트에 액세스하는 단계;

음성 신호를 포지티브 트레이닝 실례로부터 모든 음성 신호 이벤트를 포함하는 시간-궤도 공간 영역(time-trajectory space regions)으로 변환하는 단계; 및

반복적인 수행 단계를 포함하고, 반복적인 수행 단계는,

모든 시간-궤도 공간 영역에 대한 네거티브 실례를 카운트하는 단계;

네거티브 트레이닝 실례로부터 최소 이벤트를 가진 시간-궤도 공간 영역의 중 하나의 영역을 선택하는 단계; 및

상기 선택된 영역 내에서 음성 신호 이벤트가 없는 네거티브 실례를 추가 고려 대상으로부터 제거하는 단계;를

상기 트레이닝 세트 상에서 완전하게 동작하는 캐스케이드가 생성될 때까지반복하는

음성 인식 방법.

**청구항 19**

디지털 음성 신호에 대응하는 음성을 인식하기 위한 시스템에 있어서,

알려진 클래스의 디지털화된 말(known-class digitized speech utterances)의 트레이닝 코퍼스(training corpus)와, 복수의 약 분류기를 포함하는 앙상블 검출기(ensemble detector)에 액세스하도록 구성된 음성 인식 엔진 - 복수의 약 분류기의 각각은 트레이닝 코퍼스 내의 이벤트의 존재를 검출하기 위한 판정 기능을 포함하여 음성 신호 이벤트의 존재 검출시 약 분류기들이 함께 동작할 때가 혼자 동작할 때보다 더 효과적임 - ;

음성 인식 엔진에 결합되어 인식된 음성 데이터를 출력하도록 구성되는 모듈;

트레이닝 데이터를 포함하는 데이터베이스에 결합되는 자동 구두점 엔진(automatic punctuation engine) - 자동 구두점 엔진은 통계 기반의 구두점이 있는 텍스트(statistical-based punctuated text)의 형태로 트레이닝 데이터를 사용하여 향상된 음성 인식 데이터에 구두점을 추가하기 위한 적어도 하나는 통계 프로세서를 포함함 - ;

사전적 규칙(lexical rule) 데이터베이스와 결합되는 규칙 기반 펀처터(rule-based punctuator) - 규칙 기반 펀처터는 규칙 기반의 구두점이 있는 텍스트(rule-based punctuated text)의 형태로 사전적 규칙 데이터 베이스로부터의 규칙을 사용하여 향상된 인식 음성 데이터에 구두점을 추가함 - ;

구두점이 있는 텍스트 또는 통계 기반의 구두점이 있는 텍스트가 보다 나은 구두점이 있는 결과를 생성하는 지의 여부를 판단하는 판정 모듈; 및

판정에 기반하여 더 나은 구두점이 있는 결과를 출력하도록 구성된 메커니즘;

을 포함하고,

상기 음성 인식 엔진은, 디지털 음성 신호로부터 음성 신호 이벤트와, 음성 신호 이벤트의 패턴을 추출하기 위한 이벤트 추출기를 포함하되 음성 신호 이벤트와 음성 신호 이벤트의 패턴은 음성 인식과 연관되고,

상기 음성 인식 엔진은 복수의 연산(operations)을 수행하도록 구성된 적어도 하나의 프로세서를 포함하고,

복수의 연산은,

· 디지털 음성 신호 내에서 연관된 음성 신호 이벤트의 위치를 검출하는 연산 으로서 음성 신호 이벤트의 각각은 스펙트럼 정보(spectral information)와 시간 정보(temporal information)를 포함하는 음성 신호 이벤트 위치 검출 연산;

· 모든 음성 신호 이벤트들 사이의 시간 관계(temporal relation)와 모든 음성 신호 이벤트들의 스펙트럼 피처(features)를 캡처하는 연산;

· 검출된 음성 신호 이벤트의 검출 위치에 기반하여 디지털 음성 신호를 세그먼트화하는 연산;

· 세그먼트화된 디지털 음성 신호를 분석하는 연산 - 여기서 분석은 음성 신호 이벤트와 동기화됨 - ;

· 캡처된 스펙트럼 정보, 시간 관계, 및 분석된 디지털 음성 신호로 디지털 음성 신호 내 패턴을 검출하는 연산 - ;

· 디지털 음성 신호 내의 검출된 패턴에 대응하는 인식된 음성 데이터에 대한 지각적 대안(perceptual alternative)의 리스트(list)를 제공하는 연산; 및

· 향상된 인식 음성 데이터를 얻기 위해, 음성 신호 이벤트의 하나 이상의 분석에 기반하여, 인식된 음성 데이터에 대한 지각적 대안 사이를 구별하는(disambiguating) 연산을 포함하고,

적어도 하나의 프로세서는 상기 앙상블 검출기를 이용하여 복수의 연산 중 하나 이상을 수행하도록 구성되는 것

을 특징으로 하는 시스템.

**청구항 20**

삭제

## 발명의 설명

### 기 술 분 야

[0001] 본 출원은 2009년 11월 1일에 출원된, 발명의 명칭이 "System and Method for Automatic Speech to Text Conversion"인 미국 특허출원번호 12/616,723 및 2008년 11월 12일 출원된, 발명의 명칭이 "Automated Speech Processors and Automated Punctuator"인 미국 가특허출원번호 61/113,910을 우선권 주장하며, 이들 각각은 전부 참고로서 본 출원에 포함된다.

[0002] 본 발명은 일반적으로 자동 음성 인식에 관한 것이다. 보다 구체적으로, 본 발명은 시간 정보를 포함하는 음성 신호의 가장 강하고(robust) 관련 있는 특징들 및 개념적인 클러스터로부터 유도된 패턴을 이용하고 새로운 머신 학습 기법을 이용하여 정보를 처리함으로써 자동 음성 인식을 향상시키는 기법에 관한 것이다.

### 배 경 기 술

[0003] 음성 인식 정보는 주파수, 진폭 및 시간에 있어 균일하지 않게 분산된다. 모든 측면에서, 음성은 매우 가변적이다. 대부분의 자동 음성 인식 시스템은 단일 스케일로 균일하게 이격된 간격으로 정보를 추출한다. 인간의 음성 지각에서, 일부 음성 클래스(class)는 시간 특성에 대한 어필(appeal)에 의해 구별되는 것으로 알려져 있지만, 통상의 현 기술 수준의 음성 인식 시스템에서는 음성의 시간적 특성이 충분히 이용되지 않는다.

[0004] 대부분의 현 기술 수준의 자동 음성 인식 시스템은 균일한 짧은 기간(통상 20 내지 30 밀리초)의 분석 프레임을 사용하여 균일한 시간 단계들(통상 10 내지 15 밀리초)에서 음성 신호로부터 정보를 추출하는 프로세스를 포함한다. 단일의 짧은 기간의 관측 벡터(observation vector)에 기초한 음성 인식의 분류는, 음성 신호가 매우 동적이고 다양한 언어음(speech sound)이 만들어져 계속적으로 변하기 때문에, 신뢰할 수 없다. 이용가능한 시스템을 생성하기 위해 확실히 보다 긴 기간의 패턴이 이용되어야 한다.

[0005] 보다 긴 기간의 패턴을 이용가능하게 하는 공지된 방법은 음성 분류자(speech classifier)에게 동시에 제공되는 다수의 단기간 관측 벡터의 메모리를 포함한다. 이 방법에 사용되는 분류자는 흔히 인공 신경망 또는 상관 템플릿(correlation template)이다. 단기간 관측 벡터의 메모리를 포함하면 개선된 결과를 가져오지만, 여러 문제점이 존재한다.

[0006] 첫째, 모든 프레임 기반 방법에 공통인 시간 스텝 샘플링이 음성 신호와 동기하지 않는다. 따라서 음성 이벤트와 관측 프레임의 관계가 임의적이다. 그 결과, 추출된 피처의 다양성이 증가하고 시간적 디테일의 양자화가 증가한다.

[0007] 다음으로, 균일한 분석 프레임에 기초한 추출은 최적이 아니다. 언어음을 사람이 감지하는데 사용되는 정보는 많은 상이한 시간 스케일(time scale)로 발생한다. 예컨대, "t" 음(sound)의 파열음은 지속기간이 수 밀리초에 불과하지만, 모음은 1초 이상 지속될 수 있다. 일련의 많은 단기간 관측이 장기간 관측과 동일한 정보를 제공하지 않으며 그 역도 마찬가지다.

[0008] 음성의 일부 측면들은 시간 차원에서 매우 가변적이다. 예컨대, 모음이 유지되는 기간은 스피커, 음성의 레이트, 그 모음이 강세 음절에 있는 지의 여부 및 문장에서 그 음절을 포함하는 단어가 어디에서 발견되는 지에 의존한다. 이 시간적인 가변성은 음성 정보가 다른 관련 관측 프레임으로 이동하게 하여, 동일한 음성 클래스의 여러 실례에 대해 추출된 값들의 다양성을 크게 증가시키며 메모리 내의 의미있는 패턴들의 검출을 어렵게 만든다.

[0009] 또한, 프레임 기반의 시스템은 통상적으로 모든 프레임을 동등하게 중요한 것으로 처리한다. 반면에, 인간의 지각은 최고의 신호대 잡음비를 가지며 필요한 구별을 만드는데 가장 관련이 있고 신뢰할 수 있는 특징을 포함하는 신호의 부분을 사용한다.

[0010] 대부분의 현 기술 수준의 자동 음성 시스템은 은닉 마르코프 모델(Hidden Markov Model)을 포함한다. 은닉 마르코프 모델은 확률 상태 머신이다. 은닉 마르코프 모델은 관측 벡터로부터 추정된 클래스 확률을 은닉(관측되지 않은) 클래스 프로덕션의 유사한 시퀀스로 맵핑한다. 은닉 마르코프 모델을 사용하면, 자신으로의 천이에 대한 각각의 비발생 상태를 허용함으로써 전술한 시간적인 가변성 문제가 해결된다. 자기 천이 상태를 사용함

으로써 시간 가변성이 흡수된다. 불행히도, 이 방법은 기간 정보를 명시적으로 추출하도록 변경되지 않으면, 이 방법은 원치 않는 그리고 바람직한 않은 시간 정보를 제거한다. 음성 이벤트의 시간적인 관계는 특히 파열음, 파찰음 및 마찰음의 구별에서 언어음에 대한 중요한 정보를 전달한다. 또한, 클래스 확률의 로버스트 추정은 많은 양의 트레이닝 데이터를 요구한다. 사용 조건이 트레이닝 조건과 상이한 경우, 확률 평가는 매우 부정확해져서 인식이 불량하게 된다.

[0011] 대부분의 현 수준의 자동 음성 인식 시스템에 의해 사용되는 피처는 단기간 스펙트럼 프로파일로부터 주로 유도된다. 이 방법은 많은 언어음이 다소 포먼트(formant)라고 하는 특징적인 주파수 피크를 갖기 때문에 빈번하게 채용된다. 다른 현 시스템에 의해 채용되는 매우 다른 방법은 주파수 대역의 장기 궤도에 초점을 맞춘다. TRAP(Temporal Pattern)라고 하는 방법에서는, 언어음이 소리의 실례의 평균 기간(~1 초) 궤도로서 모델링된다. 분류는 각각의 TRAP 모델과의 음성 신호 엔벨로프의 상관에 기초하여 수행된다. 이 방법의 일부 버전은 단기간 스펙트럼 방법과 비교할 수 있다고 보고된 결과를 갖고 있다. 이들 결과는 언어음의 식별에 유용한 정보가 음소 세그먼트의 경계를 넘어 시간에 걸쳐 퍼져 있음을 보여 준다. 이 방법에 사용된 평균화 및 윈도잉(windowing) 때문에, TRAP의 중심 근방의 정보가 멀리 떨어진 정보에 비해 강조된다. TRAP의 총 캡처는 시간적인 디테일을 캡처하지는 않는다.

[0012] 프레임 기반 피처 추출에 대한 또 다른 방법은 "이벤트"라고 하는 소정의 검출가능한 신호 조건들의 위치에서 음성을 세그먼트화하는 것이다. 각각의 세그먼트화된 부분은 단일 클래스 식별자를 갖는 것으로 고려된다. 일반적으로 모델과의 시간적 정렬은 피처 궤도들이 공통 시간 스케일로 투영될 수 있게 하는 동적 시간 왜곡(dynamic time warping)에 의해 수행된다. 왜곡된 시간 스케일에서 피처 궤도는 리샘플링되고 템플릿과 상관되거나 은닉 마르코프 모델에 대한 관측으로서 사용된다. 동적 시간 왜곡의 프로세스는 음성 세그먼트의 많은 시간 가변성을 제거한다. 그러나, 신뢰할 수 있는 세그먼트화 이벤트는 이벤트 기반 방법에 대한 챌린지를 제공한다. 이벤트 삽입 또는 삭제는 상당한 오정렬을 일으킨다.

[0013] 자동 음성 인식의 효율 및 유효성을 향상시키기 위한 개선된 기술이 당해 기술분야에서 분명히 요구된다.

[0014] 인간의 음성 인식은 상당 부분 음성 신호에서의 이벤트의 상대적인 타이밍에 의존한다. 음성 인식에 대한 큐는 다양한 시간 스케일에 대해 발생하며 인식 자체로부터 시간적으로 오프셋될 수도 있다. 음성 이벤트의 시간적인 관계는 음성의 인식을 변화시킬 수 있다. 이것은 침묵 및 협착적 기식음의 기간이 조작된 인식 실험에 의해, B.Repp 등의 Perceptual Integration of Acoustic Cues for Stop, Fricative, and Performance 1978, Vol. 4, Num.4, 621-637에 설명되어 있다. 하나의 그러한 실험은 "Say" 및 "Shop" 이란 단어들 사이에 짧은 침묵의 간격이 삽입되는데, 이것이 듣는 사람으로 하여금 "Say Chop"으로 들리게 한다. 이벤트들의 상대적인 타이밍이 인식에 어떤 영향을 미치는 지에 대한 다른 실례는 음성 개시 시간(voice onset time)으로 지칭되며, 일반적으로 VOT로 약칭된다. VOT는 폐쇄음이 개방된 후 성대의 떨림이 시작할 때까지 경과되는 시간 간격이다. VOT는 다양한 폐쇄 자음을 구별하는 중요한 단서이다. 타이밍의 중요성은 또한 음성 현상(speech phenomena)의 기간의 가변성으로부터 유도된다. 일부 인식가능한 음성 현상은 매우 짧지만 다른 현상들은 상당히 길다. 예컨대, 음소적으로 발음기호로 표시된 영어 음성의 TIMIT 코퍼스는 5 밀리초보다 짧은 기간의 폐쇄음 버스트를 갖는 한편, 일부 모음 세그먼트는 500 밀리초보다 오래 지속된다.

[0015] 관련 이벤트 타이밍이 인식에 대한 중요한 단서이지만, 가장 일반적인 피처 추출 방법은 음성 이벤트의 타이밍에 민감하지 않다. 거의 모든 현재의 음성 및 화자 인식 애플리케이션은 고정된 스텝 사이즈만큼 시간적으로 진행된 고정된 길이 분석 프레임에 기초하여 신호 세그먼트화 방법을 이용함으로써 피처들을 추출한다. 이들 분석 프레임들은 크기가 고정되기 때문에, 이들은 거의 항상 이들이 캡처를 시도하고 있는 지각 현상의 길이보다 상당히 더 짧거나 또는 상당히 더 길다.

[0016] 구현하기 쉽지만, 일반적인 방법은 신호와 제 1 프레임의 시작점 사이의 임의의 관계 및 분석 프레임의 크기와 다양한 음성 현상의 시간 스케일 사이의 임의의 관계를 갖도록 피처들을 추출한다. S. Basu 등의 Time shift invariant speech recognition, ICSLP98에 개시된 프레임 기반의 음성 인식 시스템은 10 밀리초만큼 진행된 25 밀리초 프레임, 즉 스펙트럼 추정 및 동일한 데이터베이스 상의 워드 에러 레이트의 10%까지의 변화를 일으키는 전단(front-end)에 의해 생성된 멜 주파수 셉스트럼 계수(mel-frequency cepstral coefficients)의 큰 변화를 일으키는 10 밀리초보다 작은 제 1 프레임과 신호의 시작 관계에서의 시프트에 기초한다.

[0017] 음성 신호에서의 가변성에는 화자의 성도(vocal tract) 길이, 액센트, 음성 레이트, 건강 및 감정 상태와, 배경 잡음 등과 같은 많은 소스가 있다. 그러나, Basu 등에 의해 보고된 변화는 전적으로 프레임 사이즈 및 프레임 정렬이 신호와 임의의 관계를 갖는 피처 추출 방법의 사용에 기인한다. Ittycheriah 등의 미국 특허

5,956,671(1997년 6월 4일 출원)에서 개시된 기술 기술은 분석 프레임과 음성 신호 사이의 임의의 관계에 의해 비롯된 피처 가변성을 감소시키는 것을 목표로 한다. 이들의 발명의 일측면은 신호의 복수의 시간 변이된 버전(time-shifted version)을 거침으로써 트레이닝 세트의 가변성을 별도의 트레이닝 실례로서 고정된 프레임 분석 프로세스로 확장한다. 이들은 또한 고정된 프레임 분석의 결과를 복수의 시간 지연된 신호 버전으로 평균화함으로써 피처 값이 계산되는 인식 시간에 사용된 기법을 개시한다.

[0018] 이들 기법은 고정된 프레임 및 고정된 시간 스텝을 이용하여 피처들을 추출함으로써 비롯된 문제점들을 충분히 완화시키지 않는다. 또한, 실례들의 수를 확장하는 것은 트레이닝 시간을 증가시키고 원래의 음성 신호에 존재하지 않는 모델에 추가적인 가변성을 포함시킨다. 시간 변이된 평균화는 계산의 복잡도를 증가시키고 일부 개념적으로 관련있는 음성 특성을 평균화할 수 있다.

[0019] Moncur의 미국 특허 6,470,311(1999년 10월 15일 출원)에는, 대략 피치와 동일한 중심 주파수를 갖는 대역 통과 필터의 출력의 포지티브 영교차율(positive zero crossing)에 기초한 유성음(voiced speech)의 피치 동기 세그먼트화의 방법은 부분적으로 동기화를 다룬다. 무성음(unvoiced speech)은 지정되지 않은 시간 프레임에 걸쳐 계산된 평균 피치 주기를 사용하여 분할된다. 낮은 신호대 잡음 상태 및 작은 DC 신호 오프셋을 갖는 신호는 영교차율 기반 세그먼트화에 대한 문제를 일으키는 것으로 알려져 있음에 유의하라. 고품질 음성 신호에 있어서, Moncur의 방법은 유성음 동안 일반적인 고정 프레임 분석에 대한 개선을 나타낸다. 불행히도 무성음에 대해서는 이 방법은 임의의 고정 프레임 및 시간 스텝으로 되돌아간다. 고정 프레임 및 시간 스텝의 사용은 여전히 폐쇄음 버스트와 같은 이벤트의 정확한 위치를 해결하지 않은 채로 남겨둔다. 또한, 속삭임에 대한 어떠한 해결책도 전혀 제공하지 않는다.

[0020] 음성 현상과의 임의의 변하는 관계를 갖는 고정된 균일한 프레임에 의해서보다는 음성 신호 자체의 이벤트와 동시에 피처들을 추출하는 해결책이 요구된다. 세그먼트화 기법은 유성음과 무성음 모두를 포함하는 전체 신호에 적용되어야 한다. 또한, 음성 분석은 검출되는 특정 타입의 이벤트 각각에 대해 적절한 시간 스케일에 대해 수행되어야 한다.

[0021] 오늘날의 통상의 자동 음성 인식 엔진은 검출된 묵음(silence)을 기다려 출력을 분석하고 생성하는데, 그 이유는 이것이 자연적인 세그먼트화를 허용하고 따라서 증가된 컨텍스트로 인해 정확도가 보다 높아지기 때문이다. 애플리케이션이 거의 실시간으로, 텔레비전 방송에 대한 폐쇄 자막의 자동 생성과 같은 애플리케이션에 요구되는, 출력을 생성해야 하는 경우, 보다 작은 세그먼트화가 분석에 이용가능한 가용 컨텍스트를 감소시키고, 보다 낮은 정확도가 예상되며 생성된다. 이들 유형의 애플리케이션에 있어서는 낮은 지연을 갖는 높은 정확도가 요구된다.

## *발명의 내용*

### *과제의 해결 수단*

[0022] 본 발명의 일부 실시예는 음성 인식을 위한 검출기 및 분류기의 자동 학습과 관련된다. 보다 구체적으로, 본 발명은 특정 검출 또는 분류 작업을 위한 음성 신호의 가장 강인하고 관련있는 특성에 초점을 두는 검출기 및 분류기의 자동 학습에 관한 것이다.

[0023] 본 발명의 일부 실시예는 신호의 주목할만한 특성을 나타내는 음성 신호 스파이크 또는 이벤트의 추출을 포함한다. 이들 실시예는 또한 이벤트들 사이의 시간적인 관계를 캡처하는 것을 포함한다. 현재의 바람직한 실시예에서는, 가중된 분류기의 체계가 이벤트를 추출하는데 사용된다. 본 발명의 일부 실시예는 자동 음성 인식 엔진에 사용하기 위한 가중된 분류기의 체계를 구성하는 것을 포함한다. 본 발명의 일부 실시예는 개별 이벤트를 검출하는 대신에 또는 이에 더하여 이벤트의 시퀀스를 검출하는 것을 포함한다. 본 발명의 일부 실시예에서, 대안적인 단서에 기초한 검출기가 개발된다.

[0024] 본 발명의 일부 실시예에서, 인식 성능을 높이기 위해 적응 부스팅 알고리즘이 사용된다. 본 발명의 일부 실시예는 적응 부스팅 알고리즘에 의해 생성된 앙상블의 복잡도를 감소시키기 위한 프로세스를 포함한다.

[0025] 본 발명의 일부 실시예에서, 이벤트 기반의 검출기 캐스케이드를 자동으로 생성하는 방법은 매우 불안정한 트레이닝 세트로부터의 학습 또는 희귀 개체를 검출하기 위한 학습의 문제를 극복한다. 결과의 검출기 캐스케이드는 조기 스테이지에서 대다수의 네거티브 실례의 제거에 의해 희귀 개체의 효율적인 검출을 제공한다.

[0026] 본 발명의 일부 실시예에서, 음성을 지각 클러스터로 분류하는 프로세스가 수행된다. 프로세스는 그 다음에 다른 지각들 사이를 구별한다.

[0027] 본 발명의 일부 실시예는 지각적으로 중요한 위치에 있는 음성 신호를 분할하는 것을 포함한다. 이것은 지각적으로 관련있는 타이밍들을 추출하는 것뿐만 아니라, 신호의 분석을 음성 이벤트와 동기시키는 수단을 제공하며, 따라서 비동기 고정 프레임 분석의 모든 문제를 회피한다. 이 방법은 먼저 인간의 지각의 소정 특성 및 검출하고자 하는 음성 현상에 기초하는 낮은 복잡도 필터를 사용하여 사전 세그먼트화(pre-segmentation) 필터를 수행한다. 이들 필터는 음성의 시작(speech onset), 폐쇄(closure), 파열(burst), 성문 펄스(glottal pulse) 및 기타 중요한 음성 신호 이벤트를 나타내는 지각할 수 있는 패턴의 위치를 검출한다. 사전 세그먼트화 이벤트 필터링은 소정의 피처 계산을 동기화하는데 사용되는 간격을 정의한다. 동기식으로 추출된 피처의 패턴은 추가로 처리되어 보다 긴 시간 스케일에 걸쳐 피처들을 생성하고 음소 바운다리, 음절 핵 등과 같은 보다 높은 레벨의 지각 이벤트를 검출한다.

[0028] 바람직하게는, 높은 레벨의 음성 인식 시스템은 이들 기법들을 모두 이용한다. 본 발명의 일부 실시예에서, 자동 음성 인식을 위한 시스템에 복수의 방법이 이용된다. 시스템은 음성 입력을 수신하고, 하나 이상의 처리 수단을 음성 입력에 적용하며, 어느 처리 수단이 가장 정확한 지 판정하고 결과의 텍스트 스트림을 출력한다. 본 발명의 현재의 바람직한 실시예에서, 자동 음성 인식 시스템은 실시간 텔레비전 폐쇄 자막 및 워드 스포팅(word spotting) 및 워드 스포팅 환경에 사용된다. [다른 실시예들은 자막 또는 번역 모임 또는 전화 회의, 실시간 받아쓰기 또는 구내 전화 메시지를 기록 형태로 변환하는 것을 포함하는 임의의 음성 변환의 형태를 포함한다.] 본 발명의 일부 실시예는 지연을 감소시키기 위해 시간적으로 중복된 버스트 모드에서 n-탠덤 병렬 자동 음성 인식 엔진을 이용하여 음성 신호를 처리하는 것을 포함한다. 본 발명의 일부 실시예는 구두점 기호를 구두점이 없는 텍스트에 자동으로 삽입하는 것을 포함한다.

## *도면의 간단한 설명*

[0029] 도 1은 본 발명의 일부 실시예에 따른 자동 음성 인식 엔진의 프로세싱 모듈에 사용하기 위한 가중 분류기(weighted classifier)의 구조를 구성하는 워크플로(workflow)의 일례를 도시한 도면.

도 2는 본 발명의 일부 실시예에 따른 복수의 음성 신호 내에서 이벤트를 포함하는 영역(region)을 자동으로 식별하는 워크플로를 도시한 도면.

도 3a는 본 발명의 일부 실시예에 따른 이벤트의 시간 관계를 도시한 도면.

도 3b는 본 발명의 일부 실시예에 따른 시간의 그리드 유닛(grid units of time) 내에서 발생하는 이벤트의 카운팅을 도시한 도면.

도 3c는 본 발명의 일부 실시예에 따른 이벤트에 기초한 합산 맵(summation map)의 구조를 도시한 도면.

도 4는 본 발명의 일부 실시예에 따른 검출기 캐스케이드를 생성하기 위한 워크플로(400)를 도시한 도면.

도 5는 본 발명의 일부 실시예에 따른 모든 포지티브 실례들(positive examples)로부터의 이벤트를 포함하는 영역의 일례를 도시한 도면.

도 6a는 본 발명의 일부 실시예에 따른 모든 포지티브 실례들로부터의 이벤트를 포함하는 시간 특성 공간 내의 영역의 다른 실례를 도시한 도면.

도 6b는 본 발명의 일부 실시예에 따른 모든 포지티브 실례들로부터의 이벤트를 포함하는 비정렬(non-aligned) 영역을 도시한 도면.

도 6c는 본 발명의 일부 실시예에 따른 모든 포지티브 실례들로부터의 이벤트를 포함하는 비직사각형 영역의 일례를 도시한 도면.

도 7은 본 발명의 일부 실시예에 따른 영역의 하나의 프로젝션 내의 가장 타이트한 경계 및 가장 느슨한 경계에 대한 최대 기하학적 경계의 관계를 도시한 도면.

도 8a는 본 발명의 일부 실시예에 따른 자동 음성-텍스트 시스템을 도시한 도면.

도 8b는 본 발명의 일부 실시예에 따른 자동 음성-텍스트 시스템을 도시한 도면.

도 8c는 본 발명의 일부 실시예에 따른 이벤트 인식 및 워드 스포팅(word spotting)을 위한 시스템을 도시한 도면.

도 9는 본 발명의 일부 실시예에 따른 음성 신호의 세그먼트화의 일례를 도시한 도면.

도 10은 본 발명의 일부 실시예에 따른 인식의 변화를 계산하는데 사용되는 지각적 대조 공식을 도시한 도면.

도 11a는 본 발명의 일부 실시예에 따른 순환 큐 메모리를 도시한 도면.

도 11b는 본 발명의 일부 실시예에 따른 업데이트된 순환 큐 메모리를 도시한 도면.

도 11c는 본 발명의 일부 실시예에 따른 업데이트된 순환 큐 메모리를 도시한 도면.

도 12는 본 발명의 일부 실시예에 따른 2개의 실행 합을 유지하기 위한 구획된 순환 큐(sectioned circular queue)를 도시한 도면.

도 13은 본 발명의 일부 실시예에 따른 부분 순환 큐를 도시한 도면.

도 14는 본 발명의 일부 실시예에 따른 유성음의 작은 세그먼트에 대한 성문 펄스 검출기의 출력을 도시한 도면.

도 15는 본 발명의 일부 실시예에 따른 음절의 핵을 도시한 도면.

도 16은 본 발명의 일부 실시예에 따른 포먼트(formant) 추출을 수행하기 위한 워크플로(workflow)를 도시한 도면.

도 17은 본 발명의 일부 실시예에 따른 고조파 추출을 수행하기 위한 워크플로를 도시한 도면.

도 18은 본 발명의 일부 실시예에 따른 발성의 순서로 동작하는, 시간적으로 중복되는 2개의 탠덤 처리 엔진을 도시한 도면.

도 19는 본 발명의 일부 실시예에 따른 자동 분리기를 포함하는 음성-텍스트 시스템을 도시한 도면.

## *발명을 실시하기 위한 구체적인 내용*

[0030] 본 발명은 음성 인식을 위한 검출기(detector) 및 분류기(classifier)의 자동 학습에 관한 것이다. 보다 구체적으로, 본 발명은 가까이에 특정 검출 또는 분류 작업에 대한 시간 정보를 포함하는 음성 신호의 가장 강인하고 관련있는 측면에 중점을 두는 검출기 및 분류기의 자동 학습에 관한 것이다.

[0031] 본 발명의 현재의 바람직한 실시예에서, 자동 음성 인식 시스템은 실시간 텔레비전 폐쇄 자막 및 워드 스포팅(word spotting) 및 워드 스포팅 환경에 사용된다.

[0032] 자동 음성 인식은 수년간 개선되어 왔으며, 여전히 인간의 능력에는 미치지 못하고 있다. 사람이 듣기에는 아무런 어려움이 없는 노이즈의 레벨이 흔히 현 기술 수준의 자동 음성 인식 시스템을 사용할 수 없게 만들 수 있다. 게다가 정확도의 개선은 처리 시간 및 계산 복잡도를 증가시켰다. 많은 부분에서 이들 어려움은 음성 인식을 위해 사람이 사용하는 정보가 주파수, 진폭 및 시간에 있어 균일하지 않게 분포된다는 사실에 기인한다. 대부분의 자동 음성 인식 시스템은 모든 시점을 음성 인식과 동등하게 관련되는 것으로 처리하고 모든 종류를 동일한 피처 세트에 기초하여 결정한다. 반면에, 인간은 음성 신호의 인식에 필요한 구별을 하기 위해 가장 관련되고 강인한 특징들을 선택할 수 있다.

[0033] 귀 속의 신경 수용체는 음향 신호를 그 동적 진폭 및 주파수 분산 특성에 대한 스파이크의 시간 패턴으로 변환한다. 시간적인 스파이크 패턴은 정보를 인코딩하고 추가적인 처리를 위해 뇌의 신경으로 정보를 전달한다. 뇌의 계산 단위를 형성하는 신경 및 시냅스는 스파이크 패턴을 사용하여 정보를 인코딩하여 서로 주고 받는다. 인간의 신경 기관의 패턴 인식의 효율 및 유효성은 주목할 만하다. 스파이크 인코딩은 매우 드문 신호 표현을 생성한다. 인간의 지각의 소정 측면에 의해 영감을 받아, 본 발명은 음성 신호로부터 추출된 정보를 이하에서 "이벤트"로 지칭되는 스파이크로서 인코딩한다.

[0034] 본 발명의 현재의 바람직한 실시예에서, 이벤트 기반의 추출은 신호의 주요 측면들에 초점을 맞추고 이들 측면들의 시간적인 관계를 캡처한다. 이벤트의 유형의 일례는 주파수 통과 대역의 에너지 엔벨로프 내의 피크일 수 있다. 이들 피크는 음성 신호 내에서 각각의 대역 내의 음성 에너지가 배경 노이즈에 비해 가장 강한 위치이다. 피크와 이벤트 사이의 시간적인 거리는 말해지는 것과 강하게 관련된다. 이벤트 추출은 대역 통과

필터로부터 엔벨로프 피크를 발견하는 것에 한정되지 않는다. 다른 이벤트는 서브패턴 검출기의 출력을 포함하는 보다 복잡한 신호 분석에 의해 생성된 이벤트 및 온셋 및 호프셋을 포함한다. 임의의 공지된 방법에 기초한 분류기 및 검출기는 이들이 설계된 상태가 검출될 때 이들이 시동하게(fire) 함으로써 이벤트 패턴에 포함될 수 있다.

[0035] 관련 자동 검출기 및 분류기 구축(BUILDING RELEVANT AUTOMATIC DETECTORS AND CLASSIFIERS)

[0036] 본 명세서에서 사용된 "분류기(classifier)"는 피처 벡터, 이벤트 및/또는 이벤트의 시퀀스에 클레스 라벨(class label)을 할당하는 방법 및 장치를 지칭한다. 검출기(detector)는 "존재" 또는 "부재"의 클레스 라벨을 각각의 피처 벡터, 이벤트 및/또는 이벤트의 시퀀스에 할당하는 분류기이다.

[0037] 약한 분류기(weak classifier)는 가능성 이상(better than chance)을 수행하는 판정 기능이다. 앙상블 분류기는 복수의 약한 분류기의 결과들을 결합함으로써 형성된다. 부스팅(boosting)은 앙상블(ensemble)의 판정이 약한 분류기들 중 어느 하나의 판정보다 낫도록 약한 분류기들을 선택하고 가중함으로써 앙상블 분류기를 자동으로 구성하기 위한 당해 기술분야에서 알려진 방법이다. 선택은 약한 분류기들의 비교적 큰 세트로부터 각각의 약한 분류기를 반복적으로 평가하고 라벨링된 트레이닝 실례들의 가중된 분산에 대해 최선의 성능을 갖는 하나를 선택함으로써 행해진다. 선택된 약한 분류기는 앙상블에 더해지고 그것의 판정은 에러율에 기초하여 가중치(weight)를 할당받는다. 그 다음에 앙상블에 의해 만들어진 에러를 강조하도록 분산 가중치가 조정되고 다음 반복이 시작된다. 정확하게 분류되지 않은 실례들은 분산에서 강조되기 때문에, 앙상블의 에러를 정정하는 경향이 있는 약한 분류기가 후속 단계에서 추가되고 앙상블의 전체적인 판정이 향상된다.

[0038] 부스팅은 양호한 일반화 특성을 갖는 분류기를 생성하는 것으로 보였다. 약한 분류기는 그들의 성능이 가능성 이상인 한 어떠한 형태를 취할 수도 있다.

[0039] 시간적인 패턴 분류를 수행하는 한 방법은 복수의 고정된 간격으로 피처 궤도를 샘플링하고 모든 시간 피처 포인트를 개별 피처로서 제공하는 것이다. 통상적으로, 고정된 수의 시간 피처 포인트는 분류에 사용된다. 고정된 수의 시간 피처 포인트를 사용하는 경우, 하나의 실례의 정보와 다른 실례의 정보 간의 대응이 피처 벡터의 정의에 의해 확립된다.

[0040] 본 발명의 현재의 바람직한 실시예에 따르면, 상이한 접근법이 이용된다. 피처 궤도의 균일한 샘플링이 샘플들 사이에서 발생하는 디테일(details)을 분실하고 균일한 샘플링이 거의 구별되지 않는 정보를 포함하는 많은 샘플을 생성하기 때문에, 본 발명은 대신에 이벤트에 대한 피처 궤도를 샘플링한다. 이벤트는 중요한 정보가 집중받는 궤도 내의 포인트이다. 이벤트 기반 추출은 신호의 희소 표현을 생성한다. 이 방법은 이미지 처리와 같은 다른 상황에서 통상 사용되는 약한 분류기를 정의하는 방법의 변형을 요구하는데, 그 이유는 주어진 클레스의 실례들이 주어진 타입의 0개, 1개 또는 하나보다 많은 이벤트를 가질 수 있기 때문에, 따라서 하나의 실례에서의 정보와 다른 실례에서의 정보 간의 대응을 확립하는 방법이 요구되기 때문이다.

[0041] 피처 값, 이벤트 및 이벤트의 패턴은 검출기의 타겟 클래스와 일치하는 증거를 제공할 수 있거나 또는 반대 증거를 제공할 수 있다. 이벤트의 유형 및 이벤트들 간의 시간적인 관계는 타겟 클래스의 검출 또는 그 반대에 대한 증거의 중요부를 나타낸다. 불행히도, 동일 발음의 다른 실례들에서의 이벤트 패턴들 간의 정확한 대응은 발생하지 않는다. 또한, 노이즈는 이벤트를 가짜로 만들거나 분실시킬 수 있으며, 음성의 레이트는 이벤트 시퀀스에서 시간적인 변화를 일으킬 수도 있다. 일반적으로 머신 학습 기법들은 고정 길이 피처 벡터를 이용하도록 설계된다. 고정 길이 피처 벡터를 이용하면, 각각의 포지티브 및 네거티브 트레이닝 실례는 모든 피처마다의 값을 가지며, 각 실례에 대한 피처 값의 대응이 피처 벡터 내의 동일한 색인된 위치에서 발견된다. 고정 길이 피처 벡터 내의 값들과 달리, 이벤트 및 이벤트의 패턴은 존재할 수도 있고 존재하지 않을 수도 있으며, 서로간의 다소 상이한 시간 관계를 가질 수도 있어 하나의 실례로부터의 어느 이벤트가 다른 실례에서의 이벤트와 일치하는 지를 판정하기 어렵게 할 수 있다.

[0042] 본 발명은 시간 정보가 부스팅된 앙상블 학습자를 위한 약한 검출기를 생하는데 활용될 수 있도록, 이벤트와 실례들 사이의 이벤트 및 이벤트의 패턴의 대응이 결정될 수 있는 방법을 정의한다.

[0043] 본 발명의 현재의 바람직한 실시예에서 시간 원점(temporal origin)은 소정 종류의 이벤트와 연관되고, 모든 실례의 시간 원점은 정렬된다. 소정 음성의 특성을 나타내는 이벤트의 시간적 변화는 시간 원점에 대해 정의된

간격에 의해 경계지워진다. 주어진 간격에 있어서, (소정 종류의) 이벤트가 일정하게 포지티브 클레스 및 네거티브 클래스에 대한 간격 내에 차이를 두고 있으면, 이 차이가 약한 검출기를 생성하는데 이용될 수 있다. 본 발명의 일부 실시예에서, 실례들은 그들의 음절 핵 이벤트(syllable nucleus event)의 위치에 기초하여 정렬된다. 본 발명의 일부 실시예에서, 둘 이상의 이벤트의 세트는 각 세트 내의 이벤트들 중 하나에 대해 정렬된다.

[0044] 이벤트와 연관된 긍정 정보(affirmative information)에 기초하여 이용가능한 약한 검출기를 만들기 위해, 약한 검출기를 정의하는 간격은 포지티브 실례들의 이벤트를 과반수 포함해야 하며 네거티브 실례의 이벤트를 과반수 포함해서는 안 된다. 이러한 간격은 대부분의 포지티브 실례들로부터의 이벤트를 포함하는 모든 간격을 평가함으로써 체계적으로 결정될 수 있다. 첫째, 실례들은 특정한 공통 이벤트에 기초한 정렬에 의해 일반적인 시간적 대응을 하게 된다. 선택적으로, 상이한 전체적인 기간의 실례들이 공통 길이를 갖도록 스케일링될 수 있다. 먼저 모든 실례들에 대해 상이한 센서(예컨대, 주파수 대역 센서)로부터의 이벤트를 2차원 공간에 정렬시키고 가중된 이벤트 수의 누적된 합을 각 이벤트의 위 및 좌측에 기록함으로써, 일관된 간격들이 효율적으로 발견될 수 있다. 그 다음에 누적된 가중 카운트의 단순 차에 의해 임의의 직사각형 간격 내의 이벤트의 수가 결정될 수 있다. 대다수 실례들에 대한 이벤트를 포함하는 각각의 간격에 기초한 약한 검출기가 평가되고 현재 가중된 분산에 대한 최선의 검출기가 유지된다. 합성 검출기가 전체 트레이닝 세트에 대해 평가되고 형성된 에러에 대해 분산 가중치가 조정된다.

[0045] 검출기 성능이 트레이닝 샘플에서 완벽하거나 또는 최대 횟수의 반복에 도달할 때까지 약한 분류기가 위 처리에 따라 추가된다.

[0046] 도 1은 자동 음성 인식 엔진의 처리 모듈에 사용하기 위한 가중된 분류기의 체계를 구성하기 위한 워크플로(100)의 실례를 도시한 것이다. 본 발명의 현재의 바람직한 실시예에서, 가중된 분류 체계는 도 9와 관련하여 후술하는 바와 같이 자동 음성 인식 엔진의 분류 모듈에 사용된다. 도 1의 워크플로(100)는 복수의 음성 신호를 트레이닝 세트로서 저장(101)함으로써 시작하며, 그 후 트레이닝 세트로부터 이벤트 패턴을 추출(102)하는데, 여기서 이벤트 패턴은 음성 신호의 특징적인 측면을 포함한다. 그 다음에, 매칭 이벤트 패턴을 갖는 음성 신호의 샘플이 액세스되고(103) 음성 신호 내에서 이벤트가 발생하는 시간적인 위치에 기초하여 정렬된다(104). 그 후 각각의 신호는 공통 기간에 대해 선택적으로 스케일링된다(105).

[0047] 추출된 신호가 매칭 이벤트 위치를 갖는 공통 기간으로 스케일링되면, 복수의 약한 검출기가 이들 신호에 적용되고 이벤트를 검출하는 능력에 대해 각각의 약한 분류기의 유효성이 테스트된다(106). 측정된 유효성에 기초하여, 약한 분류기는 높은 계수의 수신을 잘 수행하는 것과 낮은 계수를 불완전하게 수행하는 것으로 가중된다(107).

[0048] 그 다음에, 가중이 사전결정된 유효성의 임계치에 기초하여 트레이닝 세트 내의 이벤트를 적절히 인식하는 지의 여부를 판정하기 위한 가중 방식의 유효성이 테스트된다(108). 워크플로는 가중이 이벤트를 적절히 인식하는지 질의한다(109). 만약 가중 방식이 적절히 수행되면, 워크플로(100)는 가중 방식을 저장하고 종료한다(110). 반면에, 가중 방식이 적절히 수행되지 않으면, 부가적인 약한 분류기가 이전에 적용된 약한 분류기 그룹에 추가되고(111), 워크플로는 유효성의 임계 레벨에 도달할 때까지 반복된다.

[0049] 주어진 발음의 상이한 실례의 이벤트 패턴은 일부 유사성을 갖지만, 음성의 임의의 두 실례들 사이에서 이벤트의 정확한 대응이 발생하지는 않는다. 상이한 실례들로부터의 이벤트에 음절의 중심에 대해 만들어지는 것과 같이 공통 시간 기준이 주어지면, 주어진 발음의 상이한 실례들로부터의 대응하는 이벤트는 타임 센서 면(time-sensor plane) 내의 영역 내에서 발생할 것이다. 음성은 매우 가변적이고 감지에 가장 유용한 정보는 주파수, 진폭, 시간 및 시간 스케일에서 균일하지 않게 분산된다. 따라서, 소정의 지각 정보를 제공하는 이벤트를 포함하는 타임 센서 면 내의 영역을 지정하는 것은 단일의 일정한 스케일 또는 형상을 이용하여 효과적으로 행해질 수 없다. 그러나, 대응하는 관련 이벤트의 집합을 포함할 수 있는 영역의 모든 가능한 위치, 형상 및 스케일을 완전히 평가하는 것은 계산적으로 실행불가능하다. 따라서, 음성 인식에 유용한 대응 이벤트의 영역을 자동으로 식별하는 프로세스가 정의된다.

[0050] 복수의 포지티브 트레이닝 실례들로부터의 제 1 이벤트들은 음절 중심(syllable center)과 같은 공통 시간 기준에 대해 만들어지며, 이들 이벤트는 시간 궤도 면(time-trajectory plane) 상으로 투영된다. 선택적으로는, 투영 전에 패턴들이 이들의 존속기간이 1이 되도록 스케일링될 수 있다. 대다수의 포지티브 실례들로부터의 이벤트를 포함하는 시간 궤도 면 내의 영역들은 대응 이벤트의 잠재적인 클러스터로서 유지된다. 이들 영역들의 리스트는 형성되어 약한 검출기를 생성하는 모든 후속 단계에 대해서 사용된다.

[0051] 도 2는 본 발명의 일부 실시예에 따른 이벤트 패턴을 포함하는 복수의 음성 신호에서 영역들을 자동으로 식별하는 워크플로(200)의 일례를 도시한 것이다. 워크플로(200)는 공통 시간축에 대해 속도 신호의 트레이닝 세트로부터 음성 신호의 그룹을 정렬함으로써 시작된다(201). 그 다음에, 워크플로(200)는 선택적으로 그룹 내의 각각의 개별 음성 신호의 기간(duration)을 공통 시간 단위 기간으로 스케일링하고(202) 음성 신호의 음절 중심 및 음성 신호의 이벤트 중심을 공통 시간축 상으로 투영한다(203). 마지막으로, 음절 중심 및 이벤트 중심의 높은 집중을 갖는 시간축 상의 영역들은 이벤트 패턴을 포함하는 영역으로 식별된다(204).

[0052] 이벤트의 높은 집중을 갖는 영역들을 식별하는 개시된 기법들 외에, 본 발명은 또한 이벤트 통합 맵핑(event integration mapping), 실례 밀도 제약의 애플리케이션, 중복 영역의 폐기 및 이들의 조합을 포함하는 강인한 약한 검출기가 될 것 같지 않은 영역들을 거부하는데 이용되는 여러 기법들도 포함한다.

[0053] 이벤트 통합 맵핑(Event integration Mapping)

[0054] 본 발명의 일부 실시예에서, 이벤트 통합 맵핑의 프로세스는 유용한 약한 검출기가 될 것 같지 않은 영역들을 거부하는데 이용된다.

[0055] 직사각형 영역에 걸친 픽셀의 세기 값의 합을 빠르게 계산할 수 있게 하는 이미지 처리 분야에 알려진 기법이 영역 내의 이벤트 카운트에 기초하여 실행 불가능 영역을 신속하게 거부할 수 있도록 수정된다. 원래의 이미지 처리 기법에서는 제 1 단계가 "합산 맵(summation map)"을 계산하는 것으로, 여기서 맵의 각각의 셀은 그 셀에서의 코너 및 대각으로 반대편의 원점에서의 코너에 의해 정의되는 직사각형 영역 내의 픽셀 값의 합에 대응한다. 그러한 합산 맵이 계산된 후에, 이미지의 임의의 직사각형 서브 영역의 픽셀의 합이 2개의 감산 및 하나의 가산 연산에 의해 결정된다. "합산 맵" 기법은 픽셀의 세기 값을 시간 궤도 면 상에 중첩된 그리드의 각각의 그리드 셀 내에서의 이벤트의 카운트로 교체함으로써 지정된 수의 실례보다 많은 수의 실례로부터의 증거를 포함할 수 없는 영역들의 빠른 제거에 적합하다. 그리드 셀 이벤트 카운트의 합산 맵이 계산되면, 2개의 감산 및 하나의 가산 연산만을 사용하여 임의의 직사각형 영역 내의 이벤트의 수의 결정이 이루어질 수 있다. 영역 내의 이벤트의 수를 아는 것은 영역 내의 실례의 수를 아는 것과 동등하지 않지만, 이것은 상부 경계를 확립한다. 따라서, 실례들의 요구된 수 이상의 이벤트의 카운트를 갖지 않는 임의의 영역은 요구된 수의 실례들을 포함할 수 없다.

[0056] 도 3a 내지 3c는 본 발명의 일부 실시예에 따른 이벤트에 기초한 합산 맵의 구성을 도시한 것이다. 도 3a에는 시간 궤도 면 내의 이벤트의 패턴이 도시되어 있다. 도 3b에서 중첩된 그리드의 경계 내에서 발생하는 이벤트의 카운트가 결정된다. 도 3c에는 합산 맵이 도시되어 있는데, 여기서 각 셀은 원점을 코너로 갖고 셀을 대각으로 반대편 코너로 갖는 직사각형 영역 내의 카운트의 합을 포함한다. 도 3c의 중앙의 4개의 셀 내의 이벤트의 수를 판정하기 위해, 문제의 영역의 우상향 셀 내의 값(이 경우에는 '7')으로부터 좌측의 비 포함 영역의 값(이경우에는 '3')이 감산되고 아래 비포함 영역 내의 값(이 경우에는 '4')이 감산되며, 2개의 감산된 영역의 교점의 과 감산된 영역(이 경우에는 '2')이 다시 가산된다. 이 결과는 영역 내의 이벤트의 수이며, 이 경우에는 '2'(7-3-4+2=2)이다. 임의의 크기 또는 형상의 영역의 이벤트 카운트를 결정하는 계산 비용(computational cost)은 동일하다.

[0057] 이벤트 밀도 제약(Event Density Constraint)

[0058] 본 발명의 일부 다른 실시예에서, 유용한 약한 검출기일 것 같지 않은 영역을 거부하는데 이벤트 밀도 제약의 애플리케이션이 이용된다. 예컨대, 최소 밀도 제약이 지정된 양 아래의 이벤트 밀도를 갖는 영역을 거부하는데 선택적으로 적용될 수 있다.

[0059] 중복 영역 거부(Redundant Region Rejection)

[0060] 본 발명의 일부 다른 실시예에서, 유용한 약한 검출기가 될 것 같지 않은 중복 영역은 거부된다. 다른 영역들을 포함하지만 포함된 영역 내에 포함된 이벤트를 넘는 부가적인 포지티브 이벤트를 추가하지 않는 영역들은 영역들의 리스트에 추가되지 않는다.

[0061] 도 2를 참조하면, 영역들이 식별되면, 이들은 약한 검출기를 생성하는데 사용되는 제약을 형성한다. 약한 검출

기는 주어진 실례들이 영역 내에 임의의 이벤트를 갖는 지의 여부를 판정하기 위한 간단한 테스트로 이루어지거나 또는 영역 내에 이벤트를 갖는 포지티브 예들의 피처 값들의 영역에 기초하여 부가적인 제약을 포함하도록 확장될 수 있다.

[0062] 이벤트 시퀀스 기반의 음성 인식(Event Sequence Based Speech Recognition)

[0063] 이벤트의 시퀀스는 일반적으로 이들을 이루는 개별 이벤트보다 자동 음성 인식에서 더 강력한 판별기이다. 본 발명의 일부 실시예는 개별 이벤트를 검출하는 것 대신에 또는 이에 더하여 이벤트들의 이벤트 시퀀스를 검출하는 것을 포함한다.

[0064] 본 발명의 일부 실시예에서, 이벤트들의 시퀀스는 시간 센서 공간 내의 (가능하게는 스케일링된) 간격을 좌표로서 이용함으로써 하이퍼 스페이스(hyper-space) 내의 포인트로서 위치한다. 이 개념을 이해하기 위해, 제 2 이벤트가 제 1 이벤트에 2 단위 시간만큼 후속하고, 제 3 이벤트가 제 2 이벤트의 4 단위 시간만큼 후속하는, 단일 센서에 의해 생성된 3개의 이벤트의 시퀀스를 고려해 보자. 서로에 대한 이들 3개의 이벤트의 시간 시퀀스는 좌표(2, 4)로 표현된다. 시간 시퀀스의 유사도는 투영된 포인트들 사이의 거리 함수를 계산함으로써 판정될 수 있다. 예컨대, 유클리드 거리가 이 목적으로 사용될 수 있다. 어느 시퀀스가 꾸준히 실례들에 나타날 수 있는 지(또는 없는 지)를 평가하기 위해, 포지티브 실례로부터의 이벤트의 시퀀스가 전술한 바와 같이 투영되어 포지티브 실례들과 관련될 수 있는 가능한 시퀀스를 나타내는 표준 포인트 세트를 형성한다. 표준 포인트는 제 1 실례로부터의 각각의 포인트들의 좌표에 기초하여 정의되고, 카운트와 관련된 각각의 표준 포인트는 1로 설정된다. 포지티브 이벤트의 나머지로부터의 이벤트 시퀀스는 제 1 실례와 유사한 방식으로 그들의 간격을 좌표로서 사용하여 하이퍼 스페이스 포인트로 투영된다. 각 시퀀스 포인트가 생성되면 그것은 가장 가까운 표준 포인트와 연관된다. 시퀀스 포인트는 표준 포인트와 연관된 리스트에 추가되고 표준 포인트 카운트는 1씩 증분된다. 그 다음에 표준 포인트의 좌표가 조정되어 관련 실례의 포인트의 좌표의 중간 값이 된다. 모든 실례들이 처리된 후, 높은 카운트를 갖는 표준 포인트는 클래스와 고도로 연관되는 이벤트 시퀀스를 나타낸다. 표준 포인트의 좌표는 시퀀스 내의 제 1 이벤트에 대해 영역의 관련 중심을 나타낸다. 영역의 크기 및 형상은 연관된 실례 시퀀스의 변화에 의해 결정될 수 있다. 본 발명의 일부 실시예에서, 유사한 시퀀스를 병합하는 것이 바람직할 수도 있다. 병합에 대한 후보는 투영된 하이퍼 스페이스 내의 그들의 거리에 의해 쉽게 결정된다.

[0065] 본 발명의 일부 실시예에서, 프로세스는 빈번하게 타겟 클래스와 동시에 발생하는 것으로 보이는 이벤트의 시퀀스를 검출하는 영역들의 조합을 찾는다. 약한 검출기로서의 이들의 활용은 타겟 클래스가 존재하지 않을 때 덜 빈번한 동시 발생에 의존한다.

[0066] 본 명세서에서 기술하는 프로세스는 포지티브 클래스의 긍정적인 증거를 제공하는 이벤트 시퀀스를 발견하는 프로세스를 포함한다. 위반 증거(contravening evidence) 또한 유익하다. 위반 증거를 발견하기 위해, 전술한 프로세스는 반복되지만 이 경우에는 네거티브 실례를 가지고 한다. 억제의 약한 검출기는, 약간의 빈도로 네거티브 실례에서 다시 발생하지만 포지티브 실례에서는 절대 발생하지 않거나 거의 발생하지 않는 시퀀스에 기초하여 형성된다.

[0067] 본 발명의 일부 실시예에서, 약한 검출기의 앙상블은 불균형한 트레이닝 세트를 처리하거나 또는 검출기의 복잡도가 낮아지는 적응 부스팅 알고리즘을 이용하여 형성될 수 있다.

[0068] 부스팅된 앙상블을 단순화함으로써 성능 개선(Performance Improvement by Simplifying Boosted Ensembles)

[0069] 본 발명의 일부 실시예에서, 인식 성능을 향상시키기 위해 적응 부스팅 알고리즘(adaptive boosting algorithm)이 이용된다. 적응 부스팅 알고리즘은 순차적으로 약한 분류기를 호출하고, 이들 분류기를 테스트하고, 그에 따라 가중 계수를 조정하는 반복적인 프로세스를 포함한다. 적응 부스팅 알고리즘은 룩어헤드 및 이전 가중의 정정 없이 반복마다 하나의 약한 검출기를 추가함으로써 앙상블을 생성한다. 그 결과, 최종 앙상블은 필요 이상으로 복잡해질 수 있다.

[0070] 본 발명의 일부 실시예는 적응 부스팅 알고리즘에 의해 생성된 앙상블의 복잡도를 감소시키는 프로세스를 포함한다. 이들 실시예에 따르면, 검출기가 트레이닝 세트에 대해 완벽을 달성하거나 최대 수의 라운드에 도달한 후, 간략화 프로세스가 수행된다. 합성 검출기의 성능은 각각이 제거된 약한 검출기들 중 상이한 하나를 갖는 자신의 버전과 반복적으로 비교된다. 약한 검출기 중 어느 하나를 제거함으로써 에러 레이트가 향상되면, 최고

의 향상을 보이는 제거가 수행되고, 그렇지 않고 약한 검출기 중 어느 하나를 제거함으로써 에러 레이트에 아무런 향상이 없으면, 그러한 약한 검출기는 제거된다. 이 프로세스는 더 이상의 약한 검출기가 제거되지 않을 때까지 계속한다.

[0071] 본 발명의 일부 다른 실시예에서는, 앙상블 구조를 위해 사용되는 새로운 검출기가 추가될 때 선형 프로그래밍 부스팅 알고리즘은 앙상블의 모든 가중을 업데이트한다.

[0072] 대안적인 단서 검출(Alternative Cue Detection)

[0073] 인간의 음성 인식은 음성 신호의 일부 측면이 손상될 경우 대안적인 단서(cue)에 의존할 수 있다. 대안적인 단서는 음성 샘플에서 발견될 수 있고 자동 음성 인식 시스템에서 검출될 수 있다.

[0074] 본 발명의 일부 실시예에서, 전술한 단계를 따라 앙상블 검출기를 생성한 다음 이전에 생성된 검출기에 의해 사용된 약한 검출기가 후속 검출기를 구성하는데 사용될 수 없는 제약으로 후속 검출기를 형성하는 프로세스를 반복함으로써 대안적인 단서에 기초한 검출기가 개발되었다. 이것은 검출기의 독립성을 최대화할 것이다. 또한 복수의 대안적인 단서 검출기는 그러한 변화에 견디는 검출기를 형성하기 위해 앙상블로서 결합될 수 있다.

[0075] 캐스케이딩 검출기에 대한 앙상블의 자동 변환(Automatic Conversion of Ensembles to Cascading Detectors)

[0076] 전체적인 앙상블의 결정은 개별 검출기의 가중된 합이다. 앙상블의 표준 형태에서, 모든 약한 분류기는 음성을 판정하도록 평가되어야 한다. 본 발명의 일부 실시예에서, 검출기 앙상블은 평균적으로 평가되어야 하는 약한 검출기의 수를 감소시키는 캐스케이딩 검출기로 변환된다. 약한 검출기를 가장 강한 것으로부터 가장 약한 것으로 배열하고 각 스테이지에서의 합과 최종 결과 사이의 관계를 분석함으로써, 앙상블을 검출기 캐스케이드로 변환하는 "조기(early out)" 임계치가 확립될 수 있다.

[0077] 다양한 이벤트들의 상대적인 타이밍은 음성 인식에 중요한 정보를 포함한다. 이 유형의 정보는 주어진 단어, 음절, 음소 등의 복수의 실례들로부터 대응하는 이벤트의 지속 패턴을 검사함으로써 활용될 수 있다. 이 분석은 모든 음성 측면에서의 가변성 및 상이한 인식 단서가 상이한 시간 스케일에 걸쳐서 발생한다는 사실로 인해 도전적이다.

[0078] 그러나, 본원 명세서에서 설명하는 바와 같이, 대부분의 머신 학습 분류 기법들은 동질의 정보의 고정된 길이 벡터에 기초하여 결정을 배우도록 설계된다. 이벤트 기반 추출의 경우, 신호 상태에 따라서 이벤트가 발생하기도 하고 발생하지 않기도 한다. 이것은 주어진 실례가 동일한 음절, 단어, 음소 등의 다른 실례보다 많거나 또는 더 적은 이벤트를 가질 수도 있음을 의미한다. 이벤트 기반 추출을 이용하여 검출기를 효과적으로 훈련시키기 위해서는, 음절, 단어, 음소 등의 하나의 실례로부터 어느 이벤트가 다른 실례들 내의 동일한 인식 지원에 대응하는 지를 검출할 필요가 있다. 본원 명세서의 후반부에 이들 대응 이벤트의 경계를 자동으로 찾아내는 방법들이 설명되어 있다.

[0079] 관련 지원 및 위반 정보 및 결정 가중을 발견하여 검출 결정을 하기 위해 트레이닝 실례들을 자동으로 사용하는 방법 및 기법들(METHODS AND TECHNIQUES TO AUTOMATICALLY USE TRAINING EXAMPLES TO DISCOVER RELEVANT SUPPORTING AND CONTRAVENING INFORMATION AND DETERMINING WEIGHTINGS TO MAKE A DETECTION DECISION)

[0080] 고도로 불안정한 트레이팅 세트에 대한 이벤트 기반 캐스케이드(Event-Based Cascades for Highly Unbalanced Training Sets)

[0081] 본 발명의 일부 실시예에서, 이벤트 기반 검출기 캐스케이드를 자동으로 생성하는 방법은 고도로 불안정한 트레이닝 센트로부터의 학습 또는 희귀 개체(rare object)를 검출하기 위한 학습의 문제점들을 해결한다. 결과의 검출기 캐스케이드는 이른 스테이지에서 대다수의 네거티브 실례들의 제거에 의해 희귀 개체들의 효율적인 검출을 제공한다.

[0082] 본 발명의 일부 실시예에서, 이벤트 기반 검출기 캐스케이드를 생성하는 것은 드물게 발생하는 특정 단어에 대한 검출기를 생성하는 것을 포함한다. 희귀 단어를 검출하는 것은 단순히 본 발명을 설명하기 위해 사용되며

당업자라면 다른 검출을 적용가능함을 쉽게 알 수 있을 것이다. 예컨대, 다른 기법들은 특정 음절, 특정 음소, 넓은 음절 클래스 및 넓은 음소 클래스와 같은 서브워드(sub-word) 음성 클래스의 검출을 포함한다. 또한, 본 발명은 산업상의 프로세스 모니터링, 자동 시스템 장애 검출 및 의료 장비 모니터링과 같은 음성 인식과 관련이 없는 많은 애플리케이션에 적용될 수도 있다.

[0083] 적은 수의 포지티브 실례들 및 다수의 네거티브 실례들을 갖는 고도로 불안정한 트레이닝 세트는 총 에러 수를 최소화하려고 시도하는 머신 학습 기법들에 의해 양호하게 처리되지 않는다. 포지티브 실례들이 드물게, 예컨대 1/100,000,000의 비율로 발생하면, 이 발생을 검출하는데 실패하는 검출기는 매우 낮은 에러율(에러율=0.00000001)을 가질 것이다. 그러나, 그것이 낮은 에러율을 갖는다 해도 절대로 잘못된 검출을 만들기 때문에, 그것은 본질적으로 무용하다.

[0084] 클래스의 멤버인 개체들은 그들의 값이 소정 범위 내에 있는 특징들을 공유한다. 따라서, 값이 이들 범위 밖에 있는 특징을 갖는 개체들은 클래스에 속하지 않기 때문에 완전히 거부될 수도 있다. 그러나, 값이 완전히 범위 내에 있는 특징을 갖는 개체들은 값이 클래스와 연관되는 범위 내에 있는 일부 특징을 가질 수도 있다. 따라서, 개체가 단일의 범위 밖의 피처 값을 가지면 개체에 대한 클래스 멤버쉽을 배제하는 것도 가능할 수 있다. 본 발명의 일부 실시예에서, 클래스 멤버쉽의 확인은 모든 관련 피처 값이 클래스와 일관된 범위 내에 있을 것을 요구한다.

[0085] 음성 인식에 적용될 경우, 이벤트 기반 피처 추출은 시간 정보를 포함하는 음성 클래스의 인식과 가장 관련있는 정보를 보존하는 희소 표현(sparse representation)을 생성한다. 추출되는 이벤트의 유형의 일례로는 소정의 피처 궤도의 엔벨로프 내의 피크의 발생을 들 수 있다. 피처 궤도 엔벨로프는, 예컨대 음성 신호가 소정의 대역 통과 필터를 통과할 때 생성된 출력에 대해 계산될 수 있다. 많은 이러한 궤도들이 계산되면, 이벤트들이 시간 궤도 공간에 분포된다. 단어 클래스 식별에 유용한 모든 증거는 시간 궤도 공간 내의 이벤트와 연관된다. 이벤트 시간이 음절의 중심과 같은 공통 시간 기준에 대해 만들어지고, 동일한 클래스의 복수의 실례들로부터의 이벤트가 시간 궤도 공간에 배치되면(plotted), 관련 이벤트의 클러스터를 포함하는 영역들이 형성된다.

[0086] 이들 클러스터를 포함하는 영역들의 위치, 형상 및 스케일은 클래스에 대해 특유하다. 이들 영역들 중 일부는 클래스와 강하게 연관되어 클래스의 모든 포지티브 실례가 영역 내에 있는 이벤트를 가질 것이다. 전술한 바와 같이, 그러한 영역 내에 이벤트를 갖지 않는 개체는 클래스의 멤버로서 거부될 수도 있다. 다수의 피처 값이 각각의 이벤트와 연관될 수도 있다. 영역 내의 포지티브 클래스 실례들로부터의 이벤트와 연관된 피처들 각각에 대한 값들의 범위는 부가적인 차원의 공간에 간격을 형성한다. 개체는 클래스 멤버로서 받아들여지는 모든 관련 피처 차원의 범위 내에 관련 값들을 갖는 이벤트를 가져야 한다. 클래스의 모든 개체들로부터 하나의 넌클래스(non-class) 개체를 구별하는 특징은 클래스의 모든 개체들로부터 다른 하나의 넌클래스 개체를 구별하는 특징과 상이하다.

[0087] 본 발명의 일부 실시예에 따르면, 이들 관련 고려사항들은 검출기를 생성하기 위해 자동으로 발견될 수 있다. 도 4는 본 발명의 일부 실시예에 따른 검출기 캐스케이드를 생성하는 워크플로(400)를 도시한 것이다.

[0088] 워크플로(400)는 제로 검출기 스테이지를 포함하도록 검출기 캐스케이드를 초기화함으로써 시작한다(401). 그 다음에, 모든 포지티브 트레이닝 실례들로부터의 이벤트를 포함하는 시간 궤도 공간 내의 모든 영역들이 식별되고 각각의 식별된 영역 내에 이벤트를 갖는 네거티브 실례들의 수가 계산된다(402).

[0089] 그 다음에, 모든 포지티브 트레이닝 실례들로부터의 이벤트를 포함하는 각각의 영역에 있어서, 영역의 정의가 부가적인 피처 차원을 포함하도록 선택적으로 확장될 수 있다(403). 임의의 부가적인 차원의 영역의 경계는 이들이 포지티브 실례의 값의 전 범위를 포함하도록 선택된다. 그 다음에, 그렇게 확립된 모든 경계 내의 피처 값을 포함하지 않는 네거티브 실례가 거부되고, 영역 내에 포함된 네거티브 실례의 카운트가 그에 따라 감소된다(404). 부가적인 차원은, 만약 존재한다면, 주어진 수의 차원에서 포함된 네거티브 실례의 카운트를 최소화하도록 선택된다. 이것은 상이한 영역에 사용된 피처 차원이 가장 잘 판별하는 차원이고 영역으로부터 영역으로 변할 수도 있다는 것을 의미한다.

[0090] 그 다음에, 가장 적은 네거티브 트레이닝 실례로부터의 이벤트를 포함하는 리스트 내의 영역이 검출기 캐스케이드 스테이지로서 선택된다(405). 본 발명의 일부 실시예에서, 검출기 스테이지의 최대 수는 사전 결정된다. 또한, 선택된 영역 내에 이벤트가 없는 네거티브 실례들은 추가적인 고려대상에서 제외된다(406).

[0091] 그 다음에, 워크플로는 얼마나 많은 네거티브 실례들이 남아 있는 지에 대한 질의를 한다(407). 네거티브 실례들이 남아있지 않으면, 트레이닝 실례들에 대해 완벽하게 수행하는 검출기 캐스케이드가 생성되었으므로, 워크

플로(400)는 검출기를 출력하고(408) 종료한다.

[0092] 이전의 반복에서보다 더 적은 네거티브 실례가 존재하지 않으면 더 이상의 향상은 이루어질 수 없다. 이 경우, 워크플로(400)는 방금 더해진 스테이지를 제거하며, 불완전한 검출기(409)를 기록하고 종료한다.

[0093] 역으로, 이전 반복에서보다 더 적은 네거티브 실례들이 존재하면, 워크플로는 최대 수의 검출기 스테이지가 더해졌는 지 질의한다(410). 최대 수의 검출기 스테이지가 캐스케이드에 더해졌다면, 워크플로(400)는 불완전한 검출기(411)를 출력하고 종료한다.

[0094] 나머지 네거티브 실례들이 존재하고 최대 수의 검출기 스테이지에 도달하지 않았다면, 워크플로(400)는 다시 반복하고 단계(402)로 되돌아가서 추가적인 스테이지를 더함으로써 검출기 캐스케이드를 계속해서 구축한다.

[0095] 검출기 캐스케이드들이 생성된 후, 이들은 다음 방법에 따라 사용된다. 먼저, 이벤트가 검출되고 트레이닝 프로세스 동안 공통 기준이 주어진다. 그 다음에, 캐스케이드의 제 1 스테이지에서 시작하여, 영역 내에 이벤트가 있는 지를 판정하기 위해 리스트 내의 이벤트들이 평가된다. 임의의 이벤트가 영역 내에 있는 것으로 발견되면, 적어도 하나의 이벤트가 그 스테이지에 의해 사용된 영역 내에 있는 것으로 발견되는 한, 리스트 내의 이벤트는 후속 스테이지에 의해 평가된다.

[0096] 그 다음에, 개체가 캐스케이드의 모든 스테이지의 영역 내부에 이벤트를 가지면, 개체는 클래스의 멤버로서 검출된다. 마지막으로, 개체가 스테이지들 중 어느 한 스테이지 내에 이벤트를 갖지 않으면, 그 스테이지에 의해 클래스의 멤버로서 거부되고, 더 이상의 처리는 수행되지 않는다.

[0097] 이들 실례에서, 축 정렬된 (하이퍼) 직사각형 영역이 이용된다. 본 발명의 일부 다른 실시예에서는, (하이퍼) 형상 또는 (하이퍼) 타원, 또는 다른 영역 또는 다른 차원 내의 바운다리 형상들의 조합과 같은 다른 바운다리 구성이 사용된다. 또한, 축 정렬되지 않은 (하이퍼) 직사각형 영역이 사용될 수도 있다. 이 관측은 모든 약한 검출기 기준에도 적용된다.

[0098] 도 5 내지 6c는 본 발명의 일부 실시예에 따른 시간-피처 값 면 상에 트레이닝 실례 이벤트를 투영한 다양한 실례들을 도시한 것이다. 도 5는 모든 포지티브 실례로부터의 이벤트를 포함하는 영역의 실례를 도시한 것이다. 도 6a는 모든 포지티브 실례로부터의 이벤트를 포함하는 영역의 다른 실례를 도시한 것이다. 도 6b는 모든 포지티브 실례들로부터의 이벤트를 포함하는 비정렬 영역을 도시한 것이다. 도 6c는 모든 포지티브 실례로부터의 이벤트를 포함하는 비직사각형 영역의 실례를 도시한 것이다.

[0099] 기하학적 마진을 최대화함으로써 일반화를 개선시킴(Improving Generalization by Maximizing Geometric Margin)

[0100] 시간-궤도 면 내의 영역들을 식별하는데 사용된 방법은 영역에 포함된 포지티브 트레이닝 실례 이벤트 주위에 꼭 맞춰지는 바운다리를 생성한다. 이러한 꼭 맞춰진 바운다리는 검출기로서 사용되면 영역의 외부 경계에 있는 트레이닝 실례 이벤트들과 약간만 상이한 값을 갖는 경우도 거부한다. 바운다리가 추가적인 네거티브 실례 이벤트를 감싸지 않고 가능한 많이 팽창되면, 검출기는 영역 내의 포지티브 트레이닝 실례들 중 어느 하나의 값의 범위와 유사하지만 이를 넘는 값을 갖는 경우를 검출할 수 있을 것이다. 그러나, 이들 최대로 느슨한 바운다리는 바운다리를 한정하는 네거티브 실례 이벤트의 값과 약간만 다른 값을 갖는 경우의 잘못된 검출을 야기할 수 있다.

[0101] 일반화는 검출된 포지티브 실례 이벤트와 거절된 네거티브 실례 이벤트 사이의 기하학적 마진을 최대화하기 위해 영역 내의 각각의 경계를 조정함으로써 향상될 수 있다. 최대 기하학적 마진의 바운다리는 타이트한 최소 바운다리와 느슨한 최대 바운다리 사이의 중간이다. 기하학적 마진을 최대화하면, 트레이닝 실례에서 보이지 않는 경우에 일반화할 최고의 기회를 허용한다. 도 7은 영역의 한 투영 내의 최대로 타이트한 바운다리와 최대로 느슨한 바운다리에 대한 최대 기하학적 바운다리의 관계를 도시한 것이다.

[0102] 제약 지각에 대한 신뢰할 수 있는 일반적인 카테고리의 시퀀스의 사용(THE USE OF SEQUENCES OF RELIABLE GENERAL CATEGORIES TO CONSTRAIN PERCEPTIONS)

[0103] 통상의 자동 음성 인식 시스템은 음소(phoneme) 또는 서브 음소(sub-phoneme) 클래스와 같은 디테일(detail)을 인식하고, 단어와 같은 보다 높은 레벨의 패턴을 판정하기 위해 그러한 디테일을 사용함으로써 동작한다. 이들

낮은 레벨의 디테일은 확실성(certainty)과 구별되지 않고, 대신에 피처 값의 관측 벡터가 주어진 각각의 클래스에 대해 확률 평가가 이루어진다. HMM(Hidden Markov Model)은 변화 확률과 함께 클래스 확률 평가를 이용하여 가장 유망한 의도된 언어음의 시퀀스를 계산한다. "디테일로부터의 구축"의 방법이 일반적이며 상당히 효과적이지만, 인간의 능력에 필적하는 자동 음성 인식 시스템에 이르지는 못한다. 이 방법의 단점들 중 하나는 세부적인 분류가 신뢰성이 높지 않고 보다 높은 레벨의 컨텍스트를 적용함으로써 고정되어야 한다는 사실이다. 또한, 세부적인 분류는 매우 컨텍스트 의존적이지만, 컨텍스트는 음성 클래스의 식별을 결정할 때 알려져 있지 않다. 또한, 컨텍스트는 부정확하게 또는 낮은 신뢰도로 표현될 수 있다. 또한, 드물게 발생하는 컨텍스트 내의 디테일에 대해 평가하는데 정확한 통계가 어렵다. 모델의 통계적 분산에서 표현되지 않는 음성의 방식 또는 음향 상태에서의 변화는 통계적 평가가 매우 부정확하게 되게 한다. 마지막으로, 대안적인 해결책의 큰 탐색 공간이 계산적으로 어려울 수 있다. 통상적으로 탐색은 가장 유망한 'n'만 간직하는 것과 같은 임의의 수단에 의해 감소된다. 본 발명의 목적은 일반적인 방법으로 상기 문제들 및 고유한 제한들을 극복하는 것이다.

[0104] 일반적으로, 넓은 카테고리로의 분류는 세분된 카테고리로의 분류보다 더 신뢰할 수 있게 수행될 수 있다. 예컨대, 어류와 조류 간을 구별하는 것은 특정 유형의 조류 또는 어류를 결정하는 것보다 더 신뢰성이 높게 행해질 수 있다. 마찬가지로, 음성 인식의 경우에, 넓은 분류는 세분된 분류보다 더 정확하게 수행될 수 있다.

[0105] 또한, 인간의 지각은 주로 넓은 분류에 대해 작동하고 이들에 초점을 맞추는 이유가 있는 경우에만 디테일을 고려하는 것으로 볼 수 있다. 유창한 연속 음성에서, 사전과 같은 단어들은 드물게 생성되지만, 이것은 지각을 지원하는 충분한 증거가 존재하는 한 듣는 사람들에게 문제를 거의 일으키지 않는다. 듣는 사람은 음성의 측면들이 일반적으로 음성의 예상 타이밍을 따르는 예상된 신뢰할 수 있는 넓은 카테고리 내로 들어가는 한 대체 또는 생략을 허용할 수 있다.

[0106] 예를 들어, 다음과 같은 질의 및 응답을 고려해보자. "Why you cryin?", "See hit me!". 이 질의는 "are"라는 단어를 생략하였고 음절 'in'이 'ing'를 대체하였다. 이들 변화들 중 어느 것도 인간의 지각에 많은 영향을 미치지 않는다. 마찬가지로, 응답은 필요한 'sh' 음이 유사 음인 's'로 되어도 "She hit me!"로 인식될 것이다. 이들 실례에서 디테일의 대체 및 생략이 지각에 거의 영향을 미치지 않고 아마도 인간에 의해 무시될 것이다. 넓은 음절 카테고리의 시퀀스의 패턴은 많은 경우에 세분된 클래스의 특정 식별을 요구하지 않고 모호하지 않은 지각을 일으키는 지각적인 유닛을 색인하기에 충분한 것으로 보인다.

[0107] 본 발명은 다음 관측에 기초한다.

[0108] - 상당한 범위까지, 넓은 음성 카테고리의 시퀀스 패턴이 가능한 지각적 대안을 제한할 수 있다. 지각적 대안은 지각적 클러스터를 형성한다.

[0109] - 넓은 음성 카테고리의 시퀀스 자체는 지각적 대안의 리스트에 직접 액세스하는데 사용될 수 있다.

[0110] - 추가적인 계산 노력이 지각적 클러스터 내의 나머지 대안들 사이에서 모호하지 않게 하는데 필요한 것으로 적용된다.

[0111] - 클러스터 내의 대안들은 트레이닝 시에 알려지기 때문에, 각각의 지각적 클러스터에 있어서 명확화의 프로세스가 최대 신뢰도 또는 최소 계산 노력에 대해 최적화될 수 있다. 결과적으로, 어떠한 상황에서도 가장 신뢰할 수 있는 구별이 적용될 수 있다. 이것은 단어 통계학, 운율학, 문법 등을 포함하는 다양한 소스로부터의 정보가 적용될 수 있다는 것을 의미한다.

[0112] - 대안적인 지각들 사이의 명확화 시에, 대안들의 발음 및 단어 컨텍스트가 알려지고, 이것에 의해 관련 있고 가장 신뢰할 수 있는 것들에 대해 피처를 구별하는 것의 계산을 제한한다. 또한, 컨텍스트 특정 검출기 및 분류기가 보다 높은 신뢰도를 위해 사용될 수 있다.

[0113] 이들 실시예에 따르면, 디테일에 대한 어필이 요구되는 것은 넓은 카테고리의 시퀀스 패턴이 지각을 완전히 명확화하지 않을 때뿐이다. 그 때에도 다른 세분된 구별보다 더 신뢰할 수 있게 만들어지도록 알려져 있는 세분된 구별을 우선적으로 사용하는 것은 가능하다. 예컨대, 두 위치에서 상이한 음소에 의해 서로 구별될 수 있는 2개의 가능한 지각에 대해 색인된 넓은 음절 카테고리의 시퀀스 패턴을 고려해 보자. 음소 쌍들 중 하나가 다른 것들보다 더 신뢰할 수 있게 구별되는 것으로 알려지면, 이 차이는 보다 신뢰할 수 있는 분류로 만들어질 것이다.

[0114] 마찬가지로, 컨텍스트가 지각에 매우 중요하다. 이전에 주어진 실례의 응답이 "cuzsee hit me!" 였다면, 이것은 "cause, he hit me!"로서 인식될 것이다. 세그먼트 'see'의 디테일들은 변경되지 않았지만, 지각은 그 세그

먼트의 세부 사항에 의존하지 않는다.

[0115] 본 발명의 일부 실시예에서, 고유 알고리즘은 음성을 지각적인 클러스터로 분류하는데 사용되며 이용가능한 정보를 최적으로 액세스함으로써 대안적인 지각들 사이를 명확화한다. 이들 실시예에 따르면, 각각의 시간 단계(즉, 소정 기간 내에 아무런 음성도 발생하지 않으면 다른 음절 패턴 또는 널 음절의 도달)에서, 이 알고리즘은 넓은 음절 카테고리와 같은 넓지만 신뢰할 수 있는 패턴의 시퀀스로 음성을 분류한다. 그 다음에, 각각의 넓은 카테고리는 카테고리 번호와 연관된다. 우선적으로, 유사한 카테고리가 유사한 번호를 할당받는다.

[0116] 그 다음에, 이 알고리즘은 상태 공간 내의 좌표로서 카테고리 번호를 사용함으로써 넓은 카테고리의 시퀀스를 지각 패턴으로 맵핑한다. 상태 공간 내의 각각의 포인트는 지각 클러스터 및 명확화 전략과 연관된다. 트레이닝 동안에 확립된 명확화 전략은 지각 클러스터가 액세스될 때 수행되는 단계들의 시퀀스이다. 명확화 전략의 목표는 이용가능한 정보에 최적으로 액세스하도록 대안 지각들 사이를 구별하는 것이다. 명확화 전략은 계산 요건 및 상이한 순서 및 상이한 조합에 적용된 다양한 명확화 기법들의 성공을 평가함으로써 결정된다. 전략을 적용한 최종 결과는 대안적인 지각을 작은 수, 바람직하게는 1로 감소시키는 것이다.

[0117] 대안들이 단일 지각으로 감소되면, 그 지각은 행해진다. 음성-텍스트 시스템에서, 이것은 지각에 대응하는 단어들을 출력하는 것을 포함할 것이다. 음성 제어 시스템에서, 지각과 연관된 동작들이 실행될 것이다.

[0118] 대안들이 단일 지각으로 감소되지 않고 최대 지연 임계치에 도달하면, 가장 유력한 지각이 지각으로 받아들여지고 그에 따라 동작들이 생성된다. 최대 지연 임계치에 도달하지 않았으면, 가능한 나머지 대안적인 지각들은 유지되고 시간 스텝에서 지각의 명확화를 돕고 이들 시간 스텝에서 이용가능한 정보에 의해 명확화되도록 후속 시간 스텝과 상호작용한다.

[0119] 자동 음성 인식 인식 엔진(AUTOMATIC SPEECH RECOGNITION ENGINE)

[0120] 본 발명의 현재의 바람직한 실시예에서, 본 발명의 신규한 특징의 모두를 수행하는 장치가 제공된다. 본 발명의 현재의 바람직한 실시예에서, 자동 음성 인식 시스템은 실시간 텔레비전 폐쇄 자막 및 워드 스포팅(word spotting) 환경에 사용된다.

[0121] 도 8a는 넓은 음절 분류의 음절 스케일에서의 이벤트 기반 추출 및 인식을 포함하는 음성-텍스트 시스템(800)을 도시한 것이다. 자동 음성-텍스트 시스템(800)은 명확화에 필요한 음소 레벨 디테일에 대한 지각 유닛의 리스트로 색인하기 위해 넓은 음절 분류의 시퀀스의 패턴을 사용한다. 본 발명의 현재 바람직한 실시예에서, 자동 음성-텍스트 시스템(800)은 어느 음소 분류를 만들 지를 선택하거나 또는 이들 분류 또는 방법의 신뢰도에 기초하여 이용할 다른 명확화 방법을 선택한다.

[0122] 자동 음성-텍스트 시스템(800)은 음향 분석기(802)를 포함한다. 음향 분석기는 입력 음성 신호(801)를 수신하고 상기 입력 신호(801)를 디지털화한다. 음향 분석기(802)는 운율 분석기(803) 및 이벤트 추출기(804)와 선택적으로 결합된다. 본 발명의 일부 실시예에서, 디지털화된 신호는 운율 분석기(803)에 의해 처리되고, 이에 따라 리듬, 강세, 억양 또는, 화자의 감정 상태, 발음이 평서문인 지, 의문문인 지 또는 명령문인 지의 여부, 반어법, 풍자, 강조, 초점 등을 반영하는 기타 운율 정보 등을 포함하는 화자의 다양한 언어적 특성이 추출된다. 이들 실시예에 따르면, 운율적 정보 및 디지털화된 신호는 이벤트 추출기(804)로 보내진다.

[0123] 이벤트 추출기(804)는 이벤트 패턴을 포함하는 복수의 음성 신호에서 영역들을 자동으로 식별하고, 음성 인식을 위해 상기 이벤트를 추출하는 처리 엔진을 포함한다. 본 발명의 현재의 바람직한 실시예에서, 이벤트 인식 및 추출에 대한 전술한 프로세스 및 방법은 이벤트 추출기(804)에 의해 이용된다. 이벤트 추출기(804)는 추출된 음성 이벤트를 저장하는 단기간 이벤트 메모리(805)와 결합된다. 단기간 이벤트 메모리(805)는 결과의 텍스트 스트림을 출력하기 위해 추출된 이벤트를 사용하는 복수의 이벤트-텍스트 스트림 처리 모듈과 결합된다. 본 발명의 현재의 바람직한 실시예에서, 이벤트-텍스트 스트림 처리 모듈은 음절 핵 검출기(806), 음절 분류기(807), 음절 시퀀스 지각 색인 모듈(808) 및 서브 음절 세부 분류 모듈(809)를 포함한다. 이벤트-텍스트 스트림 처리 모듈은 내부에 삽입된 추가된 운율 정보(811)를 갖는 텍스트 스트림을 출력한다.

[0124] 도 8a에 도시된 자동 음성-텍스트 시스템(800)은 자동 음성 인식 및 이를 개선시키는 장치의 일례를 포함한다. 당업자라면, 임의의 수의 시스템, 구성, 하드웨어 부품 등이 자동 음성 인식 및 이를 향상시키는 방법 및 프로세스를 실행하는데 사용될 수 있음을 쉽게 알 수 있을 것이다.

[0125] 도 8b는 본 발명의 일부 실시예에 따른 입력 음성 신호(821)를 처리하는 음성 인식 엔진(824)을 포함하는 음성-

텍스트 시스템(820)을 도시한 것이다. 본 발명의 현재의 바람직한 실시예에서, 음향 분석기(822)는 입력 음성 신호(821)를 수신하고 상기 입력 음성 신호(821)를 디지털화한다. 음향 분석기(822)는 운율 분석기(823) 및 음성 인식 엔진(824)와 결합된다. 본 발명의 일부 실시예에서, 디지털화된 신호는 운율 분석기(823)에 의해 처리되고, 이것에 의해 전술한 바와 같이 운율 정보가 추출된다.

[0126] 본 발명의 현재의 바람직한 실시예에서, 음성 인식 엔진(824)은 다양한 음성 인식 처리 단계를 수행하는 복수의 처리 모듈을 포함한다. 도시된 바와 같이, 음성 인식 처리 엔진(824)은 이벤트 추출기(825), 패턴 식별기(826), 약한 영역 거부기(827), 부스트 앙상블 간략화기(boosted ensemble simplifier)(828), 이벤트 시퀀스 인식기(829), 대안 단서 검출기(830), 캐스케이딩 검출기 앙상블 생성기(831), 음성 일반화기(832), 지각 클러스터 명확화 모듈(perceptual cluster disambiguating module)(833)을 포함한다. 특정 처리 모듈을 열거하였지만, 당업자라면 현재 알려져 있거나 차후에 개발될 임의의 음성 인식 툴이 음성 인식 엔진(824) 내의 처리 모듈로서 이용될 수 있음을 쉽게 알 수 있을 것이다.

[0127] 본 발명의 일부 실시예에서, 이벤트 추출기(825)는 음성 인식 엔진(824)에서 사용하기 위한 가중 분류기의 체계를 구성하기 위한 이벤트 기반 음성 인식 모듈을 포함한다. 본 발명의 일부 실시예에서, 패턴 인식기(826)는 이벤트 패턴을 포함하는 복수의 음성 신호 내의 영역들을 자동으로 식별한다. 본 발명의 일부 실시예에서, 약한 영역 거부기(827)는 강인한 약한 검출기가 될 것 같지 않은 영역들을 거부하기 위해 여러 기법들을 이용했다. 본 발명의 일부 실시예에서, 부스트 앙상블 간략화기(828)는 적응 부스팅 알고리즘에 의해 생성된 검출기 앙상블의 복잡도를 감소시킨다. 본 발명의 일부 실시예에서, 이벤트 시퀀스 인식기(829)는 개별 이벤트를 검출하는 것 대신에 또는 이에 더하여 이벤트의 시퀀스를 검출한다. 본 발명의 일부 실시예에서, 대안 단서 검출기(830)는 음성 신호의 특징이 손상되는 경우에 대안 음성 단서를 인식한다. 본 발명의 일부 실시예에서, 캐스케이딩 검출기 앙상블 생성기(831)은 검출기의 앙상블을 자동으로 생성한다. 본 발명의 일부 실시예에서, 음성 일반화기(832)는 전술한 바와 같이, 기하학적 마진을 최대화함으로써 일반화를 향상시킨다. 본 발명의 일부 실시예에서, 감지 클러스터 명확화 모듈(833)은 전술한 바와 같이 감지 클러스터링을 이용하여 음성을 명확화한다. 본 발명의 이들 실시예에 따르면, 음성 인식 엔진(824)은 음성 데이터를 출력한다.

[0128] 본 발명의 일부 실시예에서, 인식된 음성 데이터는 하나 이상의 데이터베이스(834)에 저장되고, 하나 이상의 데이터베이스(834)는 바람직하게는 네트워크(835)와 결합된다. 본 발명의 일부 다른 실시예에서, 인식된 음성 데이터는 음성-텍스트 프로세싱을 위해 단기간 이벤트 메모리(836)로 자동으로 보내진다.

[0129] 본 발명의 일부 실시예에서, 단기간 이벤트 메모리(836)는 결과의 텍스트 스트림을 출력하기 위해 추출된 이벤트를 사용하는 복수의 이벤트-텍스트 스트림 처리 모듈과 결합된다. 본 발명의 현재의 바람직한 실시예에서, 이벤트-텍스트 스트림 처리 모듈은 음절 핵 검출기(837), 음절 분류기(838), 음절 시퀀스 지각 색인 모듈(839) 및 서브 음절 세부 분류 모듈(840)을 포함한다. 이벤트-텍스트 스트림 처리 모듈은 그 내부에 삽입된 처가된 운율 정보(841)를 갖는 텍스트 스트림을 출력한다.

[0130] 본 발명의 일부 다른 실시예에서, 장치는 음성 신호 및 스포팅 워드로부터 이벤트 데이터를 추출하는 장치가 제공된다. 도 8c는 특정 단어의 이벤트 기반 추출 및 인식을 포함하는 이벤트 인식 및 워드 스포팅을 위한 시스템(850)을 도시한 것이다. 자동 음성-텍스트 시스템(850)은 입력 음성 신호(851)를 수신하는 음향 분석기(852)를 포함한다. 음향 분석기(852)는 운율 분석기(853) 및 이벤트 추출기(854)와 선택적으로 결합된다. 이벤트 추출기(854)는 이벤트 패턴을 포함하는 복수의 음성 신호 내의 영역들을 자동으로 식별하고 워드 스포팅을 위해 상기 이벤트를 추출하기 위한 처리 엔진을 포함한다. 이벤트 추출기(854)는 추출된 음성 이벤트를 저장하기 위한 단기간 이벤트 메모리(855)와 결합된다. 단기간 이벤트 메모리(855)는 복수의 워드 스포팅 처리 모듈과 결합된다. 본 발명의 일부 실시예에서, 워드 스포팅 처리 모듈은 음절 핵 검출기(856) 및 워드 검출기(857)를 포함한다. 워드 스포팅 처리 모듈은 단어가 스포팅될 때 하나 이상의 동작을 개시한다.

[0131] 제 2 처리 모듈(862)은 스파이킹 신경망 분류기(spiking neural net classifier)를 포함한다. 음성 지각에 사용된 정보는 주파수, 진폭 및 시간 내에서 균일하지 않게 분산된다. 시간 패턴은 음성 인식에 매우 중요하다. 스파이킹 신경망은 스파이크의 시간 패턴으로 음성 정보의 코딩을 허용하고 퍼지 메모리 구조는 시간 가변성의 공차를 허용한다. 제 3 처리 모듈(863)은 후술하는 바와 같이 하나 이상의 직렬 음성 인식 엔진을 포함한다.

[0132] 대안적 음성-텍스트 시스템(860)은 또한 입력 음성 신호(867)를 분석하고 디지털화하기 위한 음향 분석기(866)를 포함한다. 디지털화된 음성 신호는 3개의 처리 모듈(861, 862 또는 863) 중 하나 이상에 의해 처리되고, 그 결과는 판정 모듈(868)로 공급되며, 판정 모듈은 가장 잘 인식된 결과를 선택하여 텍스트 출력(869)을 전달한다.

[0133] 본 발명의 일부 실시예는 지각적으로 중요한 위치에서 음성 신호를 분할하는 것을 포함한다. 이것은 지각적으로 관련있는 타이밍들만을 추출하는 것뿐만 아니라, 신호의 분석을 음성 이벤트와 동기시키는 수단을 제공하며, 따라서 전술한 바와 같이 비동기 고정 프레임 분석의 모든 문제를 회피한다.

[0134] 이 방법은 먼저 인간의 지각의 소정 특성 및 검출하고자 하는 음성 현상에 기초하는 낮은 복잡도 필터를 사용하여 사전 세그먼트화(pre-segmentation) 필터를 수행한다. 이들 필터는 음성의 시작(speech onset), 폐쇄(closure), 파열(burst), 성문 펄스(glottal pulse) 및 기타 중요한 음성 신호 이벤트를 나타내는 지각할 수 있는 패턴의 위치를 검출한다.

[0135] 사전 세그먼트화 이벤트 필터링은 소정의 피처 계산을 동기화하는데 사용되는 간격을 정의한다. 동기식으로 추출된 피처의 패턴은 추가로 처리되어 보다 긴 시간 스케일에 걸쳐 피처들을 생성하고 음소 바운다리, 음절 핵 등과 같은 보다 높은 레벨의 지각 이벤트를 검출한다.

[0136] 도 9는 본 발명의 일부 실시예에 따른 음성 신호의 세그먼트화의 일례를 도시한 것이다. 도 9의 음성 신호는 "Once" 발음을 포함한다. 신호는 파형을 볼 때 시각적으로 명확한 방식으로 말의 진행에 걸쳐 특성을 수차례 변화시킨다. 그래프의 바닥에서 짧은 수직 표시로 표시된 세그먼트화는 단어의 "말해진(voiced)" 부분 동안의 성문 펄스에 해당한다.

[0137] 긴 수직 라인은 다양한 유형의 언어음 바운다리 이벤트에 대응한다. 참고로, 세그먼트 라벨이 세그먼트의 음성 식별을 나타내는 그래프 상에 위치하고 있다. 음소들 사이의 천이에서의 신호 상태는 천이 유형에 의해 변한다. 일부 바운다리에서 총 에너지는 급격하게 변하며, 다른 부분에서는 스펙트럼 변화가 이벤트와 연관된다. 이 모든 것을 고려할 때, 이들 다양한 이벤트는 피처 추출이 음성 이벤트와 동기식으로 수행될 수 있게 하며 지각적으로 관련있는 세그먼트화를 제공한다.

[0138] 본 발명의 일부 실시예에서, 신호 세그먼트화는 음성 신호 내에 존재하는 지각 차에 기초한다. 흔히, 음성 지각에 사용된 정보는 시간에 대해 균일하게 분포되지 않는다. 인간의 지각은 자극의 변화에 민감하다. 음성과 같은 시간적 신호에서, 중요한 변화(즉, 이벤트)의 시간적 위치는 신호의 지각적 구성을 제공한다. 이벤트의 상대적 타이밍 및 이들의 이웃에서의 자극의 특성이 많은 지각 정보를 인코딩한다. 일반적으로, 크기 지각은 비선형이다. 예컨대, 소리의 세기(sound intensity)의 지각은 대수적이며 일반적으로 데시벨로 측정된다. 넓은 범위의 감지에 있어서, 자극에서의 최소한의 감지가능한 차이(just-noticeable-difference)는 자극의 원래의 레벨과 관련된다. 그러나, 이것은 극단에서는 유효하지 않으며 자극의 레벨이 신경 활동에 대한 최소 레벨에 도달할 때까지 로우엔드(low end)에서의 감지는 없다. 하이 엔드에서는, 뉴런이 포화되기 시작하면, 추가적인 자극의 증가는 감지되지 않는다. 동작 범위 내에서, 많은 유형의 자극에 있어서, 지각의 반응에 필요한 변화는 베버의 법칙 $K=\Delta I/I_0$으로 요약될 수 있다. 여기서 $I_0$는 원래의 자극 레벨이고, $\Delta I$는 자극 레벨의 변화이며, K는 최소한의 감지가능한 차이의 임계치를 정의하는 경험적으로 결정된 상수이다.

[0139] 베버의 법칙 공식의 우측은 상수로서 인식될 수 있다. 본 발명에서, 이벤트는 관련 특성에서의 변화가 지각 임계치를 초과할 때 선언된다(즉, 검출기가 시동한다). 본 발명에서, 지각 변화는 베버의 법칙과 관련된 지각 대비(contrast) 계산을 이용하여 계산된다.

[0140] 도 10은 본 발명의 일부 실시예에 따른 지각 변화를 계산하는데 사용된 지각 대비 공식을 도시한 것이다. 이 공식에서, 우측의 비의 분모는, 그것이 대비되는 값의 합을 포함하고 부가적인 요소 ε을 포함한다고 하는 두 방식에서 표준 베버의 법칙과 상이하다. 요소 ε는 매우 낮은 레벨의 자극에 대한 지각 반응을 보다 잘 흉내내기 위해 매우 낮은 레벨에서의 활동을 억제한다. 요소 ε는 또한 자극이 존재하지 않을 때 제로에 의한 분할을 회피함으로써 수치상 안정된 공식을 만든다.

[0141] 대비 값의 합을 포함하면, 매우 낮은 레벨 및 매우 높은 레벨에서의 지각 대비 응답이 더욱 평평해진다. 각각의 측정된 지각 특성(예컨대, 에너지 또는 주파수)에 대해, ε 및 지각 임계치의 적절한 값이 경험적으로 확립된다. 본 발명의 일부 실시예에서, 복수의 이종 지각 이벤트 검출기가 생성되는데, 이들 각각은 일부 특정한 신호 특성에 기초하고, 일부 특정 시간 스케일로 측정되며, 자신의 특정 ε 및 지각 임계치를 갖는다.

[0142] 본 발명의 이벤트 검출기는 다양한 스케일에서 다양한 신호의 특성에 대해 동작한다. 먼저, 파열, 폐쇄 및 성문 펄스의 시간적 위치를 검출하는 낮은 복잡도 필터를 통해 에너지 값을 처리함으로써 사전 세그먼트화가 수행된다. 그 다음에 사전 세그먼트화 이벤트에 대해 피처 추출이 수행된다. 보다 높은 레벨의 피처 및 이벤트를 추출하기 위해 부가적인 필터 및 검출기가 동기적으로 추출된 피처에 적용된다.

[0143] 부가적인 피처 추출 및 처리 기법들(ADDITIONAL FEATURE EXTRACTION AND PROCESSING TECHNIQUES)

[0144] 구획된 순환 큐 메모리(Sectioned Circular Queue Memory)

[0145] 이벤트 검출기의 여러 요소들은 서로에 대해 특정 시간 관계로 정렬된 다양한 길이의 분석 윈도우를 사용하여 계산된 피처값의 합의 비교를 포함한다. 이벤트 검출기의 계산 부담을 최소화하기 위해, 이들 합은 구획된 순환 큐 메모리를 사용하여 유지된다. 순환 큐는 새로운 정보가 메모리 내의 가장 오래된 정보의 인덱스인 $I_0$으로 메모리에 기록되는 선입선출(FIFO) 메모리 구조이다. 새로운 정보를 메모리에 기록한 후, 인덱스 $I_0$ 모듈로는 메모리의 길이만큼 진행한다(즉, 인덱스 $I_0$이 메모리의 끝에 도달하면 0으로 돌아간다). 메모리 내의 값의 실행 합은 후술하는 프로세스에 따라서 유지될 수 있다.

[0146] 먼저, 순환 큐 메모리 위치, 실행 합 및 인덱스 $I_0$을 초기화한다. 그 다음에, 각각의 시간 스텝에서, 실행 합으로부터 색인된 값을 감산하고, 실행 합에 새로운 값을 더하며, 새로운 값을 순환 큐에 기록하고, 인덱스 $I_0$ 모듈로를 메모리의 길이만큼 진행시킨다.

[0147] 실행 합의 효율적인 계산을 위한 순환 큐의 동작 및 그 유용성이 도 11a 내지 11c에 도시되어 있다. 도 11a는 본 발명의 일부 실시예에 따른 순환 큐 메모리를 도시한 것이다. 도 11a에서, 새로운 값 "7"이 기억되는 시간 "t"에서의 5 원소 순환 큐 메모리가 도시되어 있다. 새로운 값은, 도시된 예에서 값 "9"를 갖는 메모리에서 가장 오래된 값에 겹쳐쓰기된다. 새로운 값을 기억하기 전, 이 예에서의 값들의 합은 25이다. 새로운 값이 가장 오래된 값에 겹쳐쓰기되기 때문에, 실행 합은 가장 오래된 값을 빼고 새로운 값을 더함으로써 유지될 수 있다. 쉽게 알 수 있듯이, 이 방식에서 나머지 실행 합을 유지하는 계산 복잡도는 메모리의 길이와 무관하다. 메모리 길이와 관계없이 감산 및 가산만이 요구된다.

[0148] 도 11b 및 도 11c는 본 발명의 일부 실시예에 따른 업데이트된 순환 큐 메모리를 도시한 것이다. 보다 구체적으로, 도 11b 및도 11c는 다음 2개의 시간 스텝을 통해 지속되는 업데이트 프로세스를 도시한 것이다. 메모리의 다양한 서브섹션에 걸쳐 값들의 복수의 실행 합을 유기하기 위해, 각각이 인덱스 $I_0$으로부터 고정된 오프셋을 갖는 부가적인 인덱스를 사용하여 순환 큐가 구획된다. 각각의 서브섹션의 실행 합은 단순히 서브섹션으로부터 막 제거되는 값을 감산하고 막 서브섹션의 부분이되는 값을 가산함으로써 유지된다.

[0149] 도 12는 본 발명의 일부 실시예에 따른 2개의 실행 값을 유지하기 위한 구획된 순환 큐를 도시한 것이다. 구획된 순환 큐는 2개의 실행 합, 즉 순환 큐 내의 값들의 가장 오래된 절반(즉, 서브섹션 A)에 대해 계산된 실행 합 및 순환 큐 내의 값들의 가장 최근의 절반(즉, 서브섹션 B)에 대해 계산된 실행 합의 유지를 용이하게 하도록 구성된다. 이들 합은 각각 $\Sigma_A$ 및 $\Sigma_B$로 지칭된다. 이제, 인덱스 $I_0$으로부터 메모리의 길이의 1/2의 오프셋에서 유지된 제 2 인덱스 $I_1$이 존재한다. 각각의 시간 스텝에서 $I_0$으로 색인된 값(즉, 전체 메모리에서 가장 오래된 값)이 $\Sigma_A$로부터 감산되고, $I_1$로 색인된 값이 $\Sigma_A$에 가산되며, I1로 색인된 값이 $\Sigma_B$로부터 감산되고 메모리에 기록될 새로운 값이 $\Sigma_B$에 가산된다. 새로운 값은 인덱스 $I_0$의 위치에 기록되고, 그 다음에 인덱스 $I_0$ 및 $I_1$이 모두 메모리의 길이만큼 모듈로 증분된다. 이 예에서, 메모리의 서브섹션은 크기가 동일하고, 공통 원소를 갖지 않는 집합을 형성하며, 함께 전체 메모리를 커버한다. 이들 조건들 중 어느 것도 방법에 의해 요구되지 않는다.

[0150] 도 13은 본 발명의 일부 실시예에 따른 구획된 순환 큐를 도시한 것이다. 도 13에서, 서브섹션 "A"는 서브섹션 B 내에 완전히 포함되도록 배치된다. 메모리의 전체 크기 및 각각의 서브섹션의 크기와 서브섹션의 시간적 배치는 합이 유지되는 목적에 따라서 결정된다.

[0151] 본 발명의 일부 실시예에서, 순환 큐는 갑작스런 변화의 위치를 검출하는데 사용된다. 시작, 폐쇄, 정지음 개방(stop burst) 등과 같은 몇몇 중요한 음성 이벤트는 신호의 일부 특징의 레벨에서 급격한 유사 단조 변화와 연관된다. 도 13에서와 같이 일반적으로 구성된 구획된 순환 큐는 급격한 유사 단조 변화를 검출하는데 이용될 수 있다. 서브섹션 "A" 및 "B"의 길이를 적절히 설정하면, 서브섹션 "A" 및 "B"의 실행 합들 사이의 지각 차이가 각각의 시간 스텝에서 계산된다. 지각 차이가 최대에 도달하고 그 크기가 지각 임계치를 초과하는 시간은 후보 세그먼트화 포인트가 된다. 검출된 이벤트들 사이의 최소 시간 분리를 위해 인간의 지각 특성을 가장 유사하게 흉내내도록 추가적인 자격이 적용된다. 이미 이 스테이지에서, 이벤트는 이벤트의 변화 방향에 기초하

여 전체적으로 분류되기 시작할 수 있다. 예컨대, 폐쇄로 인한 이벤트는 천이를 가로지른 에너지 변화의 방향에 의해 시작 및 파열로부터 구별된다.

[0152] 본 발명의 일부 다른 실시예에서, 순환 큐는 음성 신호에서의 임펄스 및 갭을 검출하는데 사용된다. 일부 중요한 음성 이벤트는 일부 신호의 특성이 매우 짧은 기간 동안 급격하게 변한 다음 변화 이전의 레벨과 유사한 레벨로 복귀하는 시간 내의 위치와 연관된다. 짧은 변화가 보다 높은 값이 될 경우, 이 변화를 "임펄스"라고 한다. 짧은 변화가 보다 낮은 값으로 될 경우, 이 변화를 "갭"이라고 한다. 일반적으로 도 5에서와 같이 구성된 구획된 순환 큐가 임펄스 및/또는 갭을 검출하는데 이용될 수 있다. 서브섹션 "A" 및 "B" 의 길이가 적절히 설정되면, 임펄스(갭)는 지각가능한 조정 임계치에 의해 서브섹션 "A"에서의 평균값이 서브섹션 "B"에서의 평균값을 상회(하회)할 때 발견된다. 전술한 바와 같이, 임계치 함수는 경험적으로 결정된다. 서브섹션 "A" 및 "B"의 길이는 인간의 지각의 성질 및 검출되는 신호 특성의 시간적인 특성에 따라서 결정된다.

[0153] 성문 펄스 검출(Glottal Pulse Detection)

[0154] 이 방법의 사용을 예시하는 중요한 특별한 경우는 성문 펄스 이벤트의 검출이다. 성문 펄스 이벤트는 다음 절차를 통해 발견된다. 먼저, 신호가 제 1 포먼트(formant)의 범위 내에서 대역 통과 필터링된다. 그 다음에, 대역 통과 필터의 출력에 대해 티저 에너지(teager energy)가 계산된다. 티저 에너지는 Teager(t)=x(t)*x(t)-x(t-1)*x(t+1)로서 계산된다. 여기서 x(t)는 시간 t에서의 입력 값이다.

[0155] 티저 에너지가 진폭 및 주파수의 함수이면, 티저 에너지는 에너지 및 고주파수 성분의 로컬 최대치와 연관되는 성문 펄스의 위치를 강조한다. 마지막으로, 신호는 도 13에서와 같이 일반적으로 구성된 임펄스 검출기를 사용하여 구획된다. 검출기는 티저 에너지의 절대치의 실행 합에 기초한다. 바람직한 실시예에서, 서브섹션 "A" 및 "B"의 길이는 각각 2ms 내지 10ms로 설정된다. 검출기는 서브섹션 "A" 내의 평균 티저 에너지가 서브섹션 "B" 내의 평균 티저 에너지에 지각 임계치 K를 곱한 값보다 더 클 때마다 하이 상태가 된다. K의 값은 1.3으로 선택되었다. 서브섹션 "A" 및 "B"의 길이 및 승수(multiplier) K는 성문 펄스 위치를 검출하는데 유용한 것으로 확인되었다. 본 발명의 범위 내에서 본 명세서에 개시하지 않은 다른 값들이 사용될 수도 있다.

[0156] 전술한 성문 펄스 검출기는 각 성문 펄스에 대해 2개의 이벤트 위치를 생성하는데, 하나는 펄스의 상승 에지에 다른 하나는 펄스의 하강 에지에 생성한다. 피치 주기는 2개의 연속적인 상승 에지 이벤트 사이의 주기로서 정의한다. 펄스의 지속기간은 상승 에지 및 후속 하강 에지 사이의 시간에 의해 추정된다. 총 피치 주기에 대한 펄스 지속기간의 비는, 일부 음성 처리 애플리케이션에서 유용할 수 있는 언어음의 피처인 "성문 개대율(open quotient)"과 관련이 있다. 또한, 피치 주기의 개방 부분 동안에 서브 성문 공동은 폐쇄 부분의 패턴에 비해 이 부분 동안에 다소 상이한 포먼트 패턴을 생성하는 구강과 음향적으로 결합된다. 이 사실은 이들 이벤트에 대해 피처 추출을 정렬함으로써 유익하게 이용될 수 있다.

[0157] 도 14는 본 발명의 일부 실시예에 따른 음성의 작은 세그먼트에 대한 성문 펄스 검출기의 출력을 도시한 것이다. 도 14에서, 성문 펄스 검출기 출력은 "하이" 및 "로우" 세그먼트로 신호를 분할한다. 하이 세그먼트는 관련 피처(이 경우에는 티저 에너지)가 지각적으로 표준 이상인 시간을 나타낸다. 이 구성은 펄스 또는 갭의 기간동안 세그먼트를 생성한다. 일부 애플리케이션에 있어서, 세그먼트보다 펄스 또는 갭에 마킹하는 것이 바람직할 수도 있다. 그러한 경우에는 특정 이벤트 시간의 선택이 다음을 포함하는 여러 대안적인 방법들 중 하나에 의해 결정될 수 있지만 이들로 한정되는 것은 아니다.

[0158] - 상승(하강) 및 하강(상승) 에지 사이의 중간지점 선택

[0159] - 세그먼트의 상승 에지 선택

[0160] - 세그먼트의 하강 에지 선택

[0161] - 세그먼트 내의 최대(최소) 피처 값 선택

[0162] - 세그먼트 내의 최대 지각 대비의 포인트 선택

[0163] 위에서 약술한 성문 펄스는, 소정의 신호 특성(예컨대, 티저 에너지)의 평균 값이 보다 긴 기간에 걸쳐 평균한 동일 특성으로부터 상당히 이탈할 때를 검출하는 것에 기초한다. 도 13에서와 같이 일반적으로 구성된 구획된 순환 큐는 선택된 음성 특성(예컨대, 에너지 또는 포먼트 주파수)이 그것의 장기간 표준으로부터 지각할 수 있게 이탈하는 영역들을 식별함으로써 임의의 복조 신호를 분할하는데 사용될 수 있다. 검출기에 의해 사용된 실

행 합을 유지하기 위한 계산 비용이 서브섹션의 길이에 독립적이기 때문에, 이들은 짧은 임펄스와 마찬가지로 큰 스케일의 변조를 분할하는데 사용될 수 있다.

[0164] 음절 핵 검출(Syllable nucleus Detection)

[0165] 이 점을 예시하기 위해, 음절 핵 검출기는, 서브섹션 "A"의 길이가 60ms로 설정되었고 서브섹션 "B"의 길이는 100ms로 설정되었다는 점을 제외하고 성문 펄스 검출기에서와 같이 정확하게 계산된 티저 에너지의 실행 합을 유지하도록, 도 13에서와 같이 일반적으로 구성된 구획된 순환 큐를 사용하여 구성되었다.

[0166] 도 15는 본 발명의 일부 실시예에 따른 파형 출력을 도시한 것이다. 도 15는 첫째는 정상적으로 두번째는 속삭임으로 2회 말해진 단어 "Once"에 대한 파형 및 검출기 출력을 보여준다. 도면에서 볼 수 있듯이, 이 검출기는 일반적으로 음절의 중심부를 일괄 처리한다.

[0167] 본 발명의 일부 실시예는 포먼트 추출을 이용하여 음성 패턴을 인식하는 방법을 포함한다. 음성이 생성될 때, 발음 기관(즉, 혀, 턱, 입술)의 구성이 포먼트라고 하는 주파수 스펙트럼 내에 공진 및 반공진(anti-resonance)의 동적 패턴을 생성한다. 유성음 동안에, 확산되어 있는 "공기 소음(air noise)" 및 강하게 조직된 고조파 구조(harmonic structure) 모두에 의해 음(sound)이 생성된다. 확산(diffuse) 및 고조파(harmonic)는 음성의 이해에 기여하며, 상이한 잡음 상태 하에서 다양하게 의존한다. 확산되어 있는 "공기 소음"은 포먼트와 상호작용하며 이들에 의해 공유되고, 이들을 비교적 부드러워지도록 노출한다. 강하게 분해된 고조파는 스펙트럼에서 상당히 선명한 피크를 생성하며, 적절히 처리되지 않으면, 근방의 포머트를 정확히 찾아내기 어렵게 많든다. 고조파 시리즈는 피치 주기 주파수 자체가 신호로부터 빠져 있는 경우에도, 피치를 결정하기 위한 유수한 수단을 제공한다. 실험에 의하면, 진폭 변조된 고조파는 잡음을 "무시하는" 이해불가능한 음성을 재생하는데 사용될 수 있다. 무성음 동안에 지각할 수 있는 변화는 신호를 유사 동질의 세그먼트로 시간적으로 분할한다.

[0168] 포먼트 추출

[0169] 본 발명의 일부 실시예에서, 포먼트 추출의 프로세스는 도 16에 도시된 바와 같이 수행된다. 도 16은 본 발명의 일부 실시예에 따른 포먼트 추출을 수행하는 워크플로(1600)를 도시한 것이다.

[0170] 워크 플로(1600)는 세그먼트의 샘플이 세그먼트 길이와 동일한 윈도우 길이로 해밍 윈도우잉(Hamming-windowing)될 때(1601) 시작하며, 여기서 세그먼트는 유성음 동안의 한 피치에 대응한다. 그 후 광대역 통과 필터의 필터 뱅크를 통해 윈도우잉된 샘플이 처리된다(1602). 일부 실시예에서, 대역 통과 필터는 400Hz 대역폭을 가지며, 50Hz에 중심을 두고 450Hz 내지4000Hz 범위를 커버한다. 그 다음에, 워크플로는 순간 진폭을 계산하고, DESA-1 기법을 이용하여 각 필터의 주파수가 계산된다(1603). 이들의 수치 품질(numeric quality)에 기초하여, 단계 1604에서 계산된 값이 "유효(valid)" 또는 "무효(not valid)"로서 판정된다. 그 다음에 "유효" 평가치가 카운팅되어 임시 버퍼에 저장된다.

[0171] 그 다음에, 주파수 범위를 나타내는 빈(bin)을 갖는 히스토그램을 초기화하고(1606), 각각의 유효 평가치에 대해 평가된 순간 주파수를 나타내는 히스토그램 빈이 대응하는 로그 압축된 평가된 순간 진폭만큼 증분된다. 그 다음에, 평활화된 히스토그램의 피크가 포먼트 후보로서 선택되고(1607), 포먼트 주파수, 대역폭(시그마) 및 진폭이 피처로서 유지되며(1608), 델타 피처가 라인 피팅에 의해 포먼트 트랙에 대해 계산된다(1609). 마지막으로, 포먼트 패턴에서의 지각할 수 있는 변화의 위치에서 이벤트가 생성된다(1610).

[0172] 12번째 옥타브 필터 뱅크 처리($12^{th}$ Octave Filter Bank Processing)

[0173] 본 발명의 일부 다른 실시예에서, 12번째 옥타브 필터 뱅크 처리의 프로세스는 보다 낮은 주파수에서의 좁은 통과 대역 및 인간의 청취에서 발견된 주파수 분해능 트렌드를 흉내내는 보다 높은 주파수에서의 보다 넓은 통과 대역을 이용하여 분할된 신호에 대해 수행된다. 도 17은 본 발명의 일부 실시예에 따른 포먼트 추출을 수행하는 워크플로(1700)를 도시한 것이다.

[0174] 워크플로(1700)는 세그먼트의 샘플이 세그먼트 길이와 동일한 윈도우 길이를 갖는 해밍 윈도우잉된 신호와 동기하는 것으로 시작하는데(1701), 여기서 세그먼트는 하나의 피치 주기에 해당한다. 그 다음에, 윈도우잉된 샘플

들이 12 옥타브 이격된 필터 뱅크를 통해 처리되고(1702), 각 필터의 순간 진폭 및 주파수가 DESA-1 기법을 이용하여 계산된다(1703). 수치 품질에 기초하여, 계산된 값은 "유효" 또는 "무효"로 판단되며(1704), 여기서 "유효" 평가치가 카운트되고 간격 동안 임시 버퍼에 저장된다(1705).

[0175] 그 다음에, 빈이 12 옥타브 필터 뱅크 내의 각 필터의 중심 주파수에 대응하는 히스토그램이 구성되며(1706), 여기서 각각의 유효 평가치에 대해, 평가된 순간 주파수를 포함하는 범위를 갖는 히스토그램 빈이 대응하는 로그 압축된 평가된 순간 진폭만큼 증분된다. 그 다음에, 히스토그램 가중치에 상이한 주파수에서 귀의 민감도에 기초하는 가중 함수가 곱해진다(1707). 히스토그램 계산 후에, 가장 강한 에너지를 갖는 가장 강한 고조파 시퀀스를 검출하기 위해 히스토그램 빈 에너지 패턴이 고조파 조합에 더해지며(1708), 여기서 가장 강한 고조파 시퀀스의 기본이 피치의 평가로서 이용된다. 애플리케이션이 훨씬 더 정교한 평가를 요구하면, 보다 높은 대역통과 필터를 평가된 고조파 주파수의 중심에 두고 재계산한다(1709). 이 프로세스는 매우 정확한 평가치에 신속하게 수렴한다. 마지막으로, 고조파 에너지대 총 에너지의 비가 음성의 측정치로서 계산되며(1710), 여기서 고조파의 진폭비 패턴이 피처로서 유지되고, 이 비는 자동 음성 인식에 사용된다.

[0176] 피치 주기의 사용(Use of Pitch Periods)

[0177] 본 발명의 일부 실시예에서, 고조파 트랙의 온셋(onset) 및 오프셋(offset)은 피치 주기로부터 피치 주기로의 상대적인 진폭에 의해 결정될 수 있다. 고조파 트랙의 진폭에서의 급격한 변화는 고조파의 포먼트와의 상호작용과 연관되고, 이 급격한 변화는 반복 내에서의 변화를 나타내며, 이는 피치에서의 변화 또는 포먼트에서의 변화를 나타낸다. 이러한 변화는 천이 위치를 나타낸다. 이벤트는 위에서 약술한 필터 방법을 이용하여 이들 변화에 응답하여 생성될 수 있다. 이들 이벤트는 발생시에 성문 펄스 타이밍과 동기될 수 있다는 점에 유의하라.

[0178] 성도 정규화 및 연음 인식(Vocal Tract Normalization and Soft Phoneme Segment Recognition)

[0179] 본 발명의 일부 실시예에서, 포먼트 패턴을 피처로서 사용하는데 있어서의 고유한 복잡성을 치유하기 위해 성도 정규화 및 연음 인식의 프로세스가 이용된다. 화자에 의해 생성된 포먼트 패턴은 생성되는 언어음 및 화자의 성도 길이에 대한 정보를 동시에 인코딩한다. 이것은 포먼트 패턴을 피처로서 사용하는 것을 복잡하게 한다.

[0180] 이것은 Wantanabe, et al., Reliable methods for estimating relative vocal tract lengths from formant trajectories of common words, IEEE transactions on audio, speech, and language processing, 2006, vol. 14, pp. 1193-1204에 동일한 언어음을 생성하는 두 화자에 대한 포먼트가 그들의 성도 길이의 비에 역비례하는 관계, 즉 $L_A/L_B=F_{nB}/F_{nA}$를 갖는 것으로 개시되어 있다.

[0181] 다른 언어음이 생성될 때, 화자의 성도 길이는 발음 기관의 동적 인식에 의해 연속적으로 수정된다. 소정의 화자에 대해, 각각의 음이 생성됨에 따라 포먼트들은 성도 길이를 변형시키므로 아래 또는 위로 이동할 것이다. Watanabe의 공식을 소정의 언어음을 발음하는 화자 "A"의 포먼트 패턴 및 동일한 음을 발음하는 화자 B의 포먼트 패턴에 적용하면, 각각의 측정된 포먼트에 대한 관련 성도 길이의 하나의 평가가 제공된다. 본 발명의 일부 특징은 다음 관측에 기초한다. 먼저, 화자 "A" 및 화자 "B"가 동일 음을 생성하고 있다면, 다양한 측정 포먼트의 각각에 기초하는 관련 성도 평가는 대략 참 값에 근접할 것이며, 따라서 서로 유사할 것이다. 다음으로, 화자 "A" 및 화자 "B"가 상이한 음을 생성하고 있다면, 다양한 측정 포먼트의 각각에 기초한 관련 성도 길이의 평가가 달라질 것이다. 또한, 소정의 언어음으로부터의 천이가 화자 "A"에 의해 말해질 때의 성도 길이를 길어지게(짧게) 하는 것을 포함하면, 이는 또한 화자 "B"의 성도 길이를 그들의 생리학에 기초하여 상이한 양만큼 길어지게(짧게)하는 것을 포함할 것이다.

[0182] 일부 실시예에서, 기준 화자에 의해 말해진 각각의 언어음에 대한 포먼트 값이 기록된다. 기준 화자의 포먼트 측정은 단일 화자 또는 다수에 기초할 수 있으며, 바람직하게는 다수 화자의 측정으로부터 평균으로 취해질 수 있다. 인식 시간에, 각각의 세그먼트는 전술한 바와 같이 포먼트 값을 생성하도록 처리된다. 각각의 언어음(즉, 음속 또는 부분 음소)은 차례로 말해지는 언어음이고, 현 세그먼트의 포먼트 값은 기준 화자의 성도 길이에 대한 현재 화자의 상대적인 성도 길이의 평가치를 계산하도록 사용된다. 평가치의 일치(consistency)가 각각의 음에 대해 기록된다. 일치의 리스트에 기초하여, 각 언어음의 상대적인 확률이 확립될 수도 있다. 음성의 궤도가 각각의 정규 포먼트 패턴의 타겟 구성에 접근함에 따라, 평가치의 일치는 증가할 것이며 그러한 타겟 시간은 지각된 언어음에 대해 최대로 되는 경향이 있을 것이다. 이러한 지각에 적용될 수 있는 신뢰도는 언어

음 및 잡음 상태에 의존한다. 언어음이 높은 신뢰도를 갖는 것으로 판정되면, 이들은 낮은 신뢰도를 갖는 영역 내의 가능한 패턴을 제약하는데 유용한 신호에서의 기준 포인트가 된다.

[0183] 탠덤 병렬 자동 음성 인식 엔진(TANDEM PARALLEL AUTOMATIC SPEECH RECOGNITION ENGINES)

[0184] 본 발명의 일부 실시예는 지연을 감소시키고 정확도를 향상시키기 위해 시간적으로 중첩되는 버스트 모드에서 복수의 탠덤 병력 자동 음성 인식(ASR) 엔진을 사용하는 것을 포함한다. 각각의 ASR 엔진은 유사하거나 상이한 설계 및 기원을 가질 수 있지만, 모두가 최소 세그먼트화 시간 프레임 내에서 세그먼트의 중심부에 타겟 랭귀지에서 받아들일 수 있는 결과를 생성할 수 있어야 한다. 탠덤 프로세서의 결과는 시작 및 끝에서의 단어보다 더 높은 각 세그먼트의 중심부 동안에 생성된 단어를 가중하고 가장 잘 맞는 것에 의해 세그먼트들을 싱킹함으로써 분석되며, 보다 높은 가중치를 갖는 단어들이 출력용으로 선택된다.

[0185] 이들 실시예는 지연을 감소시키고 정확도를 향상시키기 위해 음성 세그먼트들을 중첩할 시에 복수의 ASR 엔진을 사용하는 것을 포함한다. 탠덤 병렬 방법은 지연을 감소시키면서 정확도를 증가시킨다.

[0186] 예컨대, 하나의 ASR이 임의로 x초에서 인입하는 음성 신호를 분할하면, 그 출력은 x/2 위치에서 가장 정확하고, 세그먼트의 시작 및 끝에서 가장 덜 정확한 경향이 있는데, 이는 전방 및 후방으로의 최고 컨택스트가 중심 위치에서 발견되기 때문이다. 이 관측 동작을 고려하면, 단순히 배치 모드에서 ASR 엔진의 n개의 인스턴스를 실행하고, 인입 신호를 x/n 초만큼 중복되는 x초 버스트로 분할하며, 각 엔진 사이에 이들 세그먼트의 라우팅을 교체시킴으로써 이 정보를 레버리지(leverage)로서 사용할 수 있다. n=2이면, 엔진 B가 그 세그먼트를 인식하는 동작을 하고 있는 동안, 엔진 A로부터의 단어들을 통계적으로 부스팅하고, 수정하며 출력하기 위해 이전에 출력된 단어 스트림과 함께 엔진 A로부터의 출력이 분석된다. 그 다음에, n 초 입력 바운다리에서, 출력 분석기 및 프로세싱 작업들이 엔진들 사이에서 작업을 교환한다.

[0187] 탠덤 구성에서 유용한 통상적인 ASR 엔진을 관측하는 중에, 3000 개의 단어 WSJ 영어 언어 모델을 사용할 때 대략 3초를 설정할 때 x가 가장 잘 동작하는 것을 본다. 이것은 낮은 지연이 필요한 환경에 사용하기에 적합해지도록, 긴 발음에서 작동하도록 설계되고 최적화되는 엔진을 사용할 가능성을 허용한다.

[0188] 즉, x=3이면, 0.0 내지 3.0 초에서의 제 1 음성 세그먼트가 엔진 A로 렌더링렌더링을 위해 제공될 것이다. 그 다음에 1.5 내지 4.5로부터의 세그먼트가 엔진 B로 제공되는 등으로 될 것이다.

[0189] 도 18은 본 발명의 일부 실시예에 따른 발음의 시퀀스로 동작하는 시간적으로 중복되는 2개의 탠덤 처리 엔진을 도시한 것이다. 도 18에 도시된 바와 같이, 단어들 "is falling from the sky"는 엔진 A로부터의 출력이고, "done the sky today at"는 엔진 B로부터 온다. 통계적 방법을 이용함으로써 이들 단어에 대한 신뢰성 요소를 고려하는 각 세그먼트의 단부에서의 각 단어에 대한 가중치를 할인하면, 3초의 고정된 지연을 갖는 "is falling from the sky today at"와 유사한 명백한 연속적인 단어 스트림으로 끝날 수 있다.

[0190] 가중 분석 및 출력 엔진은 후속 카테고리 내의 하나 이상의 알고리즘 및 어느 단어가 최종 출력 스트림에 가산될 수 있는 지를 결정하기 위한 다른 알고리즘을 포함할 수 있다. 예컨대, 세그먼트의 에지에 있는 단어보다 더 높은 값을 갖는 세그먼트 내의 중심 단어의 단순한 가중, 원래의 음성 신호로부터 획득된 음향 및 운율적 힌트, 보다 유망한 출력의 가중치를 부시팅하도록 출력되는 단어들의 통계적 분석, 보다 유망한 출력을 선택하기 위한 문법 규칙 또는 기타 머신 학습 및 통계 방법을 포함할 수도 있다.

[0191] 자동 분리기(AUTOMATIC PUNCTUATOR)

[0192] 본 발명의 일부 실시예는 구두점 표시를 구두점이 없는 텍스트에 자동으로 삽입하는 것을 포함한다. 자동 분리기는 구두점 표시(마침표, 쉼표, 의문 부호, 느낌표, 아포스트로피, 인용 부호, 괄호, 타원, 세미콜론 및 콜론)를 구두점이 없는 텍스트에 삽입하는 시스템이다.

[0193] 도 19는 본 발명의 일부 실시예에 따른 자동 분리기를 포함하는 음성-텍스트 시스템(1900)을 도시한 것이다. 본 발명의 일부 실시예에서, 구두점이 없는 텍스트는 텍스트(1901) 또는 음성 언어(1902)로서 시작할 수 있으며, 음성 언어는 그 후 자동 음성 인식 시스템(1903)에 의해 텍스트로 변환된다.

[0194] 변환된 텍스트 또는 1901로부터의 텍스트는 자동 분리기(1905)로 보내진다. 자동 분리기(1905)는 구두점의 적절한 배치로 인해 보다 쉽게 읽을 수 있고 덜 모호한 텍스트를 생성한다.

[0195] 본 발명의 일부 실시예에서, 자동 분리기(1905)는 트레이닝 데이터를 포함하는 데이터베이스(1904)와 결합된다. 자동 분리기는 정확하게 구두점이 찍힌 많은 양의 트레이닝 텍스트로 트레이닝되는 하나 이상의 베이지안(Bayesian) 알고리즘을 사용한다. 트레이닝 데이터 내의 구두점 패턴은 텍스트 내의 구두점 패턴을 기술하는 규칙 세트를 생성하도록 분석된다.

[0196] 충분한 양의 텍스트에 대해 구두점이 트레이닝되면, 이 규칙은 구두점 기호가 어디에 삽입되어야 하는 지를 예측하기 위해 새로운 텍스트에 적용될 수 있다.

[0197] 본 발명의 일부 실시예에서, 자동 분리기(1905)는 복수의 처리 모듈을 포함한다. 도시된 바와 같이, 자동 분리기는 제 1 통계 프로세서(1906), 제 2 통계 프로세서(1907) 및 제 3 통계 프로세서(1908)를 포함한다.

[0198] 일부 실시예에서, 제 1 통계 프로세서(1906)는 통계적 규칙에 기초하여 구두점이 삽입되어야 할 위치를 식별한다. 트레이닝 프로세스는 이들 규칙을 발전시키도록 동작한다. 트레이닝 프로세스는 특정 단어와 다량의 적절하게 구두점이 찍힌 텍스트 내의 구두점 기호 사이의 상관의 분석을 포함한다. 규칙 세트는 이 분석으로부터 도출된다. 그러면 규칙 세트는 구두점 기호에 대해 유망한 위치를 예측하기 위해 새로운 구두점이 찍히지 않은 텍스트에 적용될 수 있다. 이 프로세스의 출력은 구두점 기호가 어디에 삽입되어야 하는 지에 대한 일련의 견해이다.

[0199] 일부 실시예에서, 제 2 통계 프로세서(1907)는 음성부와 구두점 기호와의 상관을 트레이닝한다. 이 프로세스는 트레이닝 데이터 내의 문장의 구조를 분석하고 각 단어에 음성부 태그를 할당하는 음성부 태거(part-of-speech tagger)에 의존한다. 음성부 태그의 예들은 명사, 동사, 형용사 전치사 등이 있다.

[0200] 그 다음에 이 프로세스는 소정의 음성부를 구두점 기호와 상관시키는 방법의 관측에 기초하여 규칙 세트를 구축한다. 그 다음에 규칙 세트는 새로운 텍스트에 적용될 수 있다. 이 프로세스의 출력은 구두점이 텍스트 내의 어디에 삽입되어야 하는 지에 대한 일련의 견해이다.

[0201] 일부 실시예에서, 제 3 통계 프로세서(1908)는 평균 문장 길이에 기초한 가중치를 이용한다. 통계적 구두점의 제 3 성분은 통상적으로 특성 텍스트 내에 문장을 구성하는 단어의 수에 기초한다. 다른 프로세스와 마찬가지로, 이 프로세스는 다량의 정확하게 구두점이 찍힌 텍스트에 대해 트레이닝한다. 이들 규칙은 구두점에 의해 구분되는 텍스트의 유닛들 내에서 발생하는 n-그램의 수에 기초하여 개발된다.

[0202] 본 발명의 일부 실시예에서, 제 1 통계 프로세서(1906) 및 제 2 통계 프로세서(1907)는 구두점이 텍스트 내의 어디에 삽입되어야 하는 지에 대한 2 세트의 견해이다. 제 3 통계 프로세서(1908)로부터의 결과는 판정이 충돌할 때의 상황을 해결하기 위해 일종의 타이 브레이커(tie-breaker)로서 사용된다. 예컨대, 제 1 통계 프로세서(1906)가 스트링 내의 제 5 단어 다음에 마침표가 필요하다고 예측하고, 제 2 통계 프로세서(1907)가 제 3 단어 후에 마침표가 필요하다고 예측하면, 2 단어 문장이 형성될 것이기 때문에 둘 다 올바를 것 같지는 않으므로 결정을 내리기 위해 제 3 통계 프로세서(1908)로부터의 결과가 호출될 것이다.

[0203] 일부 실시예에서, 제 3 통계 프로세서(1908)는 이 유형의 문서 내에서 통상의 문장 길이의 정보에 기초하여 보다 높은 가중치를 제 1 통계 프로세서(1906) 또는 제 2 통계 프로세서(1907)로부터의 결과에 할당한다. 이 문서 유형의 문장이 통상적으로 매우 짧다면, 제 3 통계 프로세서(1908)는 제 2 통계 프로세서(1907)의 출력에 보다 높은 가중치를 할당할 수 있다. 반면에, 문서 유형의 문장이 일반적으로 5 단어 또는 그 이상이면, 제 3 통계 프로세서는 제 1 통계 프로세서(1906)에 의해 생성된 견해에 보다 높은 가중치를 할당할 것이다.

[0204] 결정 단계가 완료되면, 그 결과는 규칙 기반의 구두점 모듈(1910) 및 피치/정지 모듈(1911)로부터의 정보와 협력하여 구두점을 어디에 삽입할 것인지에 대한 최종 결정을 내리는 결정 모듈(1909)로 진행한다.

[0205] 일부 실시예에서, 규칙 기반의 구두점 모듈(1910)은 구두점 기호가 텍스트 내의 어디에 삽입되어야 하는 지를 결정하기 위해 언어 구조에 대한 규칙 세트를 사용한다. 규칙 기반 분리기 모듈(1910)은 사전적 데이터베이스(1916)와 결합된다.

[0206] 규칙 기반 분리기 모듈(1910)은 주격 대명사, 목적격 대명사, 관계 대명사, 법(modal), 접속사, 정관사, 날짜 및 동사의 소정 카테고리를 포함하는 단어의 몇몇 기능적 클래스를 식별할 수 있다. 일부 실시예에서, 사전 데이터베이스(1916)는 음성 부분 정보를 포함한다.

[0207] 프로그램이 기능적 카테고리들 중 하나의 멤버를 식별하면, 근방의 문맥을 탐색하여 식별된 항목 및 선행 및 후행하는 2 단어로 이루어지는 텍스트의 윈도우를 조사한다. 음성부의 단어들의 특정 카테고리는 스트링 내의 일부 지점에서 쉼표의 필요를 나타낼 것이다. 언어 규칙은 쉼표가 어디에 삽입되어야 하는 지에 대한 인스트럭션

리스트로서 작용한다. 일례로서, 프로그램이 주격 대명사(I, he, she, we, they)를 식별하면, 프로그램은 다른 카테고리의 발생에 대한 컨텍스트 윈도우를 검사한다. 예컨대, 주격 대명사가 형용사 또는 분사(소정의 동사 분사가 예측됨)에 선행하면, 프로그램은 식별된 단어에 선행하는 단어 다음에 쉼표가 있을 것으로 예측할 것이다. 규칙 기반의 구두점은 텍스트의 스트림 또는 사전 존재하는 텍스트 파일을 처리할 수 있다. 규칙 기반의 구두점은 쉼표가 어디에 삽입되어야 하는 지에 대한 일련의 견해이다.

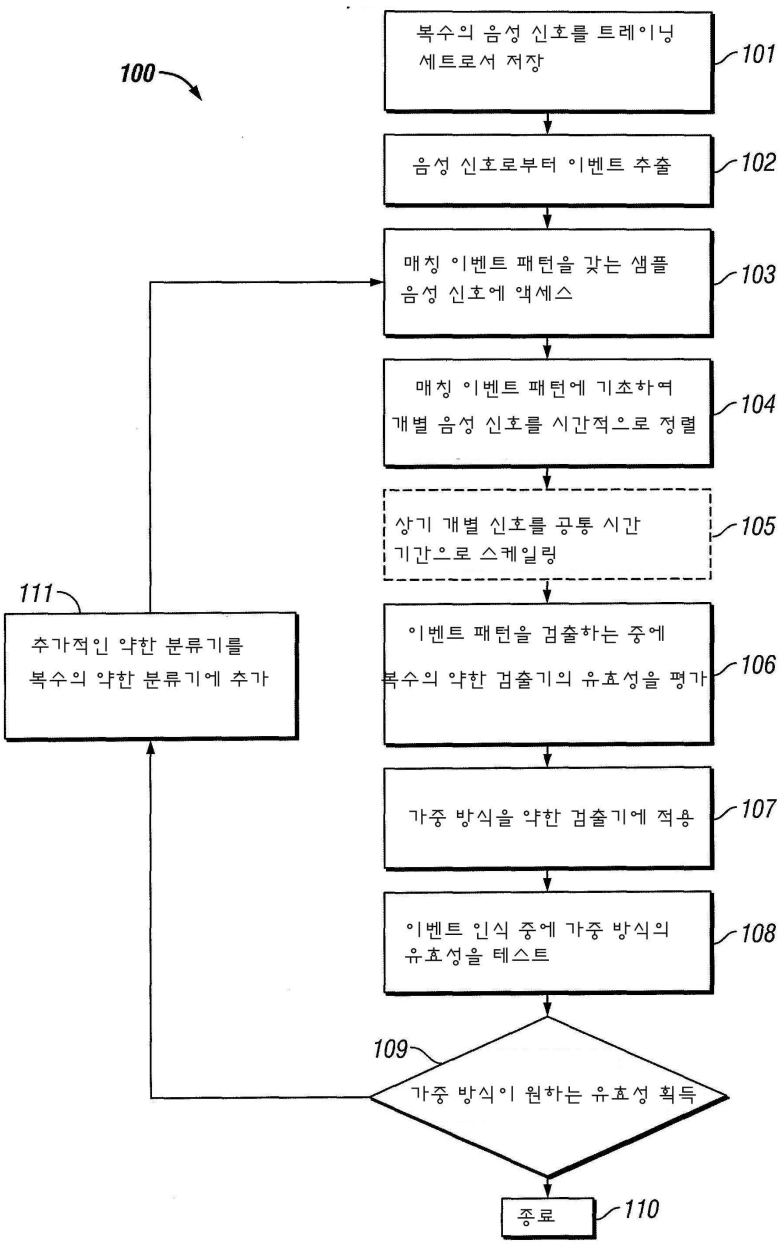
[0208] 일부 실시예에서, 피치/중지(pitch/pause) 모듈(1911)은 그 입력이 사람의 음성을 포함하는 오디오 파일이라는 점에서 다른 요소들과 상이하다. 다른 요소들은 비록 차후에 변환된 오디오 데이터로서 시작한 텍스트일 수도 있지만, 텍스트에 대해 동작한다. 피치/중지 모듈(1911)은 사람의 음성에서 짧은 기간에 걸쳐 발생하여 침묵 기간과 상관되는 중요한 피치 변화가 일반적으로 구두점에 대한 필요성을 나타낸다고 하는 관측에 기초하여 동작한다. 예를 들면, 오디오 파일 내의 소정 지점은 짧은 기간(275 ㎳) 내에 발생하는 피치의 가파른 하락(30% 이상)을 보여주는데, 즉 화자가 문장의 끝에 도달했다는 것을 나타내는 지표이다.

[0209] 이 패턴 다음의 중지는 구두점 기호에 대한 위치가 식별되었다는 것을 확인하는 경향이 있다. 피치/중지 구두점은 정확한 상태가 구두점을 나타내도록 만났을 때 오디오 파일 및 신호의 피치를 추적한다. 피치/중지 구두점은 구두점 기호가 어디에 삽입되어야 하는 지에 대한 견해를 출력한다.

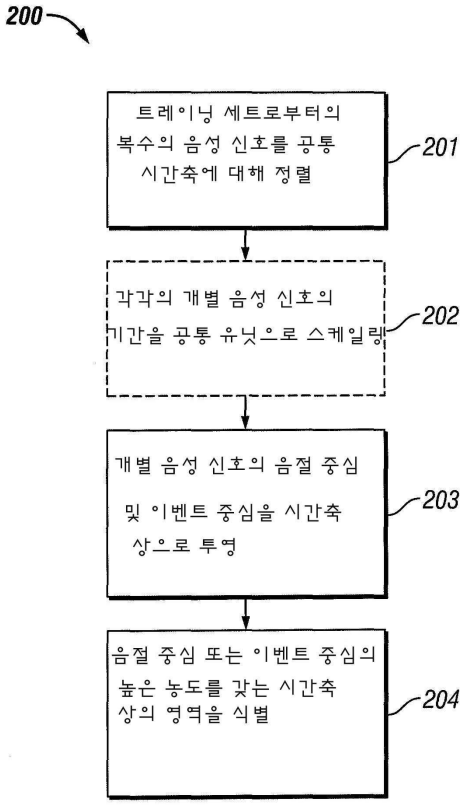
[0210] 일부 실시예에서, 결정 모듈(1909)은 자동 분리기(1905), 규칙 기반 분리기(1910) 및 피치/중지 모듈(1911)로부터 입력을 취한다. 텍스트의 유형의 특징에 기초하여, 결정 모듈(1909)은 이들 결정의 각각에 보다 높거나 낮은 가중치를 할당하여 구두점이 텍스트 내 주어진 지점에 삽입되어야 할 지의 여부에 대한 최종 결정을 내린다.

도면

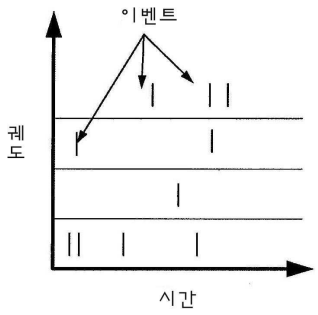
도면1


100
복수의 음성 신호를 트레이닝 세트로서 저장
101
음성 신호로부터 이벤트 추출
102
매칭 이벤트 패턴을 갖는 샘플 음성 신호에 액세스
103
매칭 이벤트 패턴에 기초하여 개별 음성 신호를 시간적으로 정렬
104
상기 개별 신호를 공통 시간 기간으로 스케일링
105
이벤트 패턴을 검출하는 중에 복수의 약한 검출기의 유효성을 평가
106
가중 방식을 약한 검출기에 적용
107
이벤트 인식 중에 가중 방식의 유효성을 테스트
108
109
가중 방식이 원하는 유효성 획득
종료
110
111
추가적인 약한 분류기를 복수의 약한 분류기에 추가

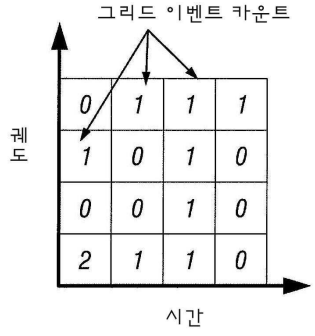
도면2



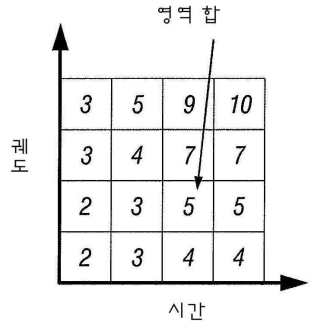

도면3a



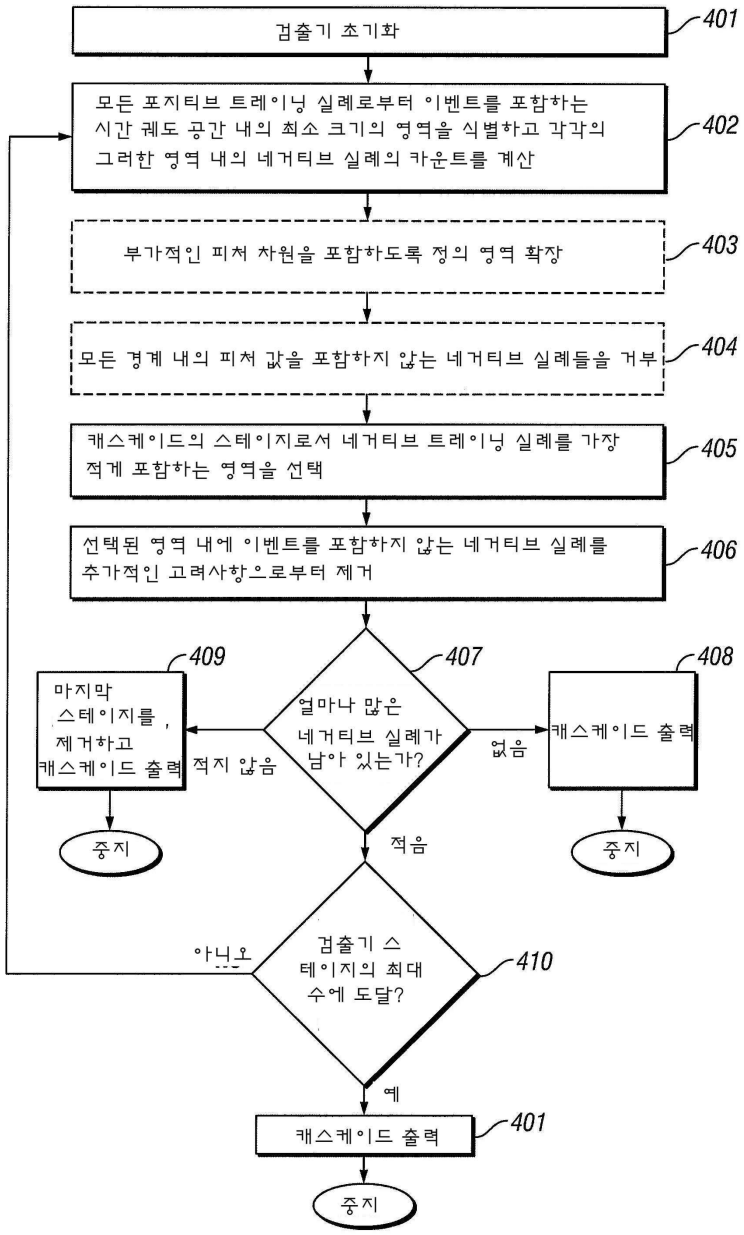

도면3b



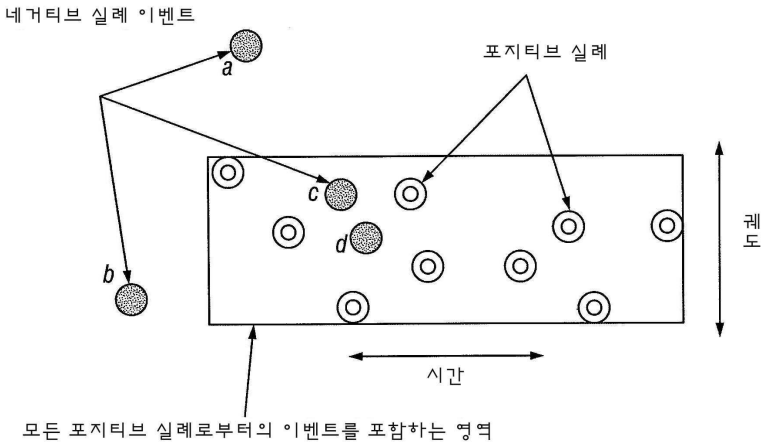
도면3c


영역 합
3 5 9 10
3 4 7 7
2 3 5 5
2 3 4 4
궤도
시간

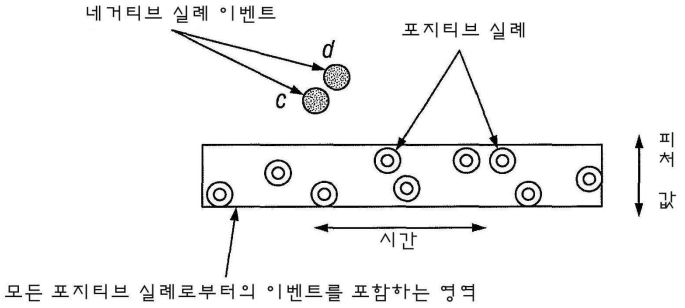
도면4


검출기 초기화
401
모든 포지티브 트레이닝 실례로부터 이벤트를 포함하는 시간 궤도 공간 내의 최소 크기의 영역을 식별하고 각각의 그러한 영역 내의 네거티브 실례의 카운트를 계산
402
부가적인 피처 차원을 포함하도록 정의 영역 확장
403
모든 경계 내의 피처 값을 포함하지 않는 네거티브 실례들을 거부
404
캐스케이드의 스테이지로서 네거티브 트레이닝 실례를 가장 적게 포함하는 영역을 선택
405
선택된 영역 내에 이벤트를 포함하지 않는 네거티브 실례를 추가적인 고려사항으로부터 제거
406
얼마나 많은 네거티브 실례가 남아 있는가?
407
적지 않음
마지막 스테이지를 제거하고 캐스케이드 출력
409
중지
없음
캐스케이드 출력
408
중지
적음
검출기 스테이지의 최대 수에 도달?
410
아니오
예
캐스케이드 출력
401
중지

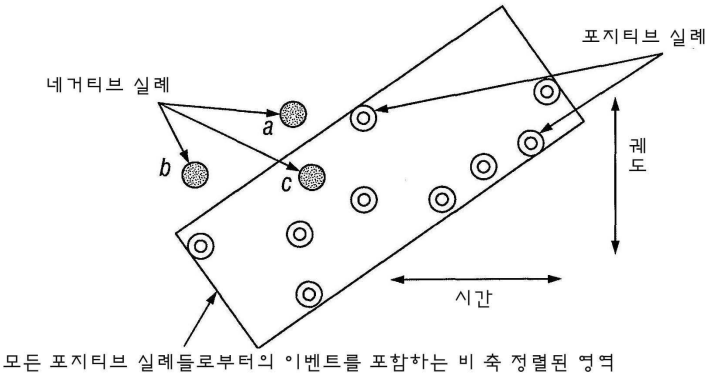
도면5



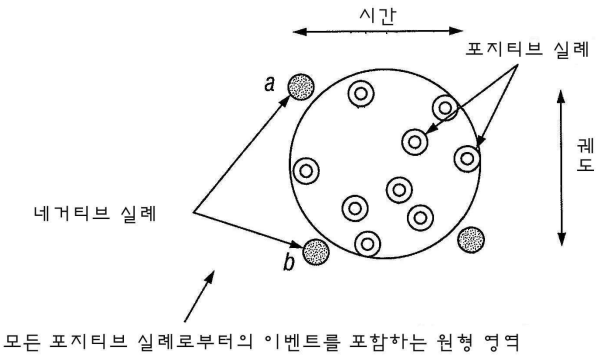

도면6a



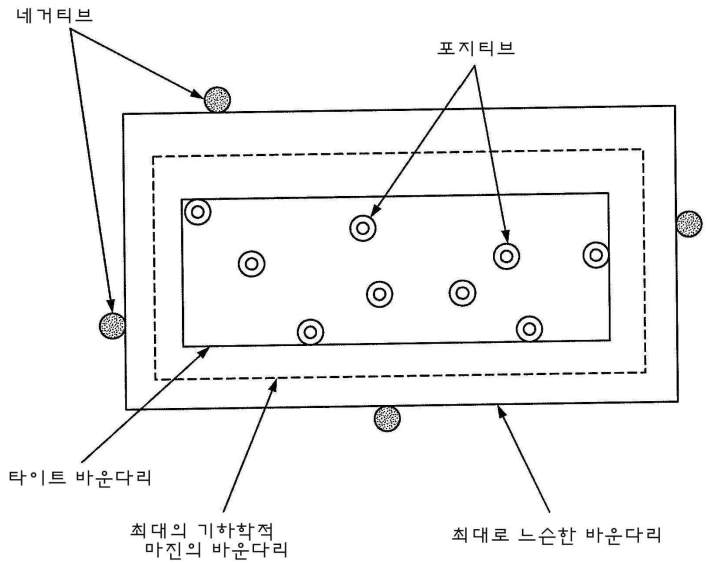

도면6b



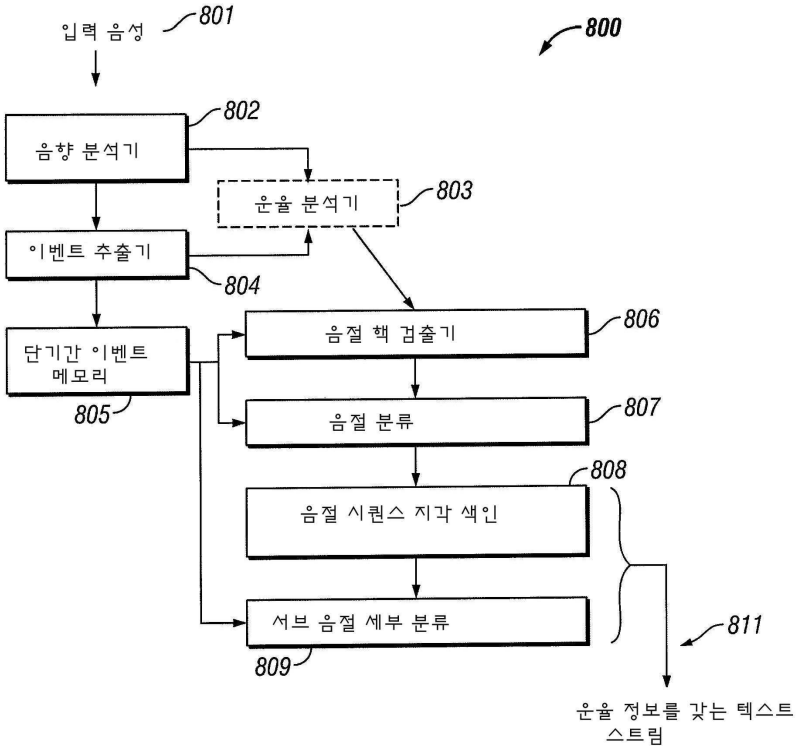
도면6c


시간
포지티브 실례
a
네거티브 실례
b
궤도
모든 포지티브 실례로부터의 이벤트를 포함하는 원형 영역

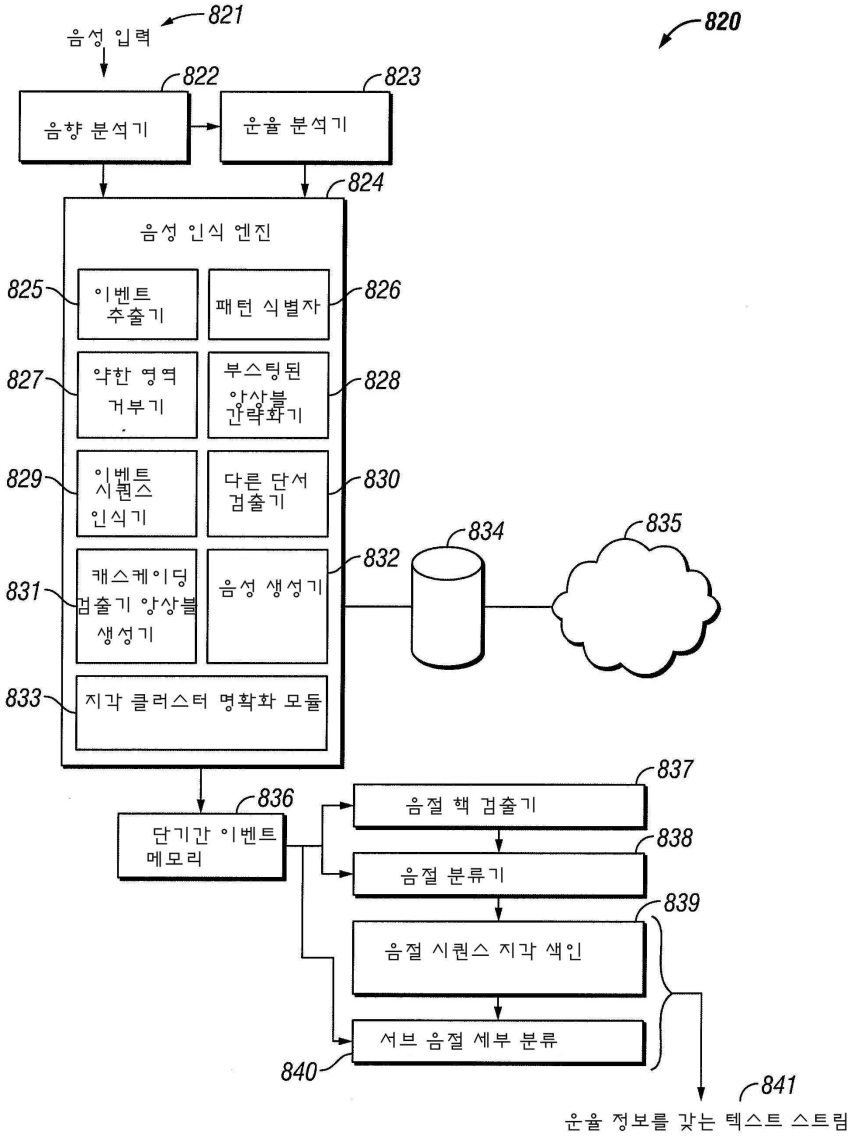
도면7


네거티브
포지티브
타이트 바운다리
최대의 기하학적
마진의 바운다리
최대로 느슨한 바운다리

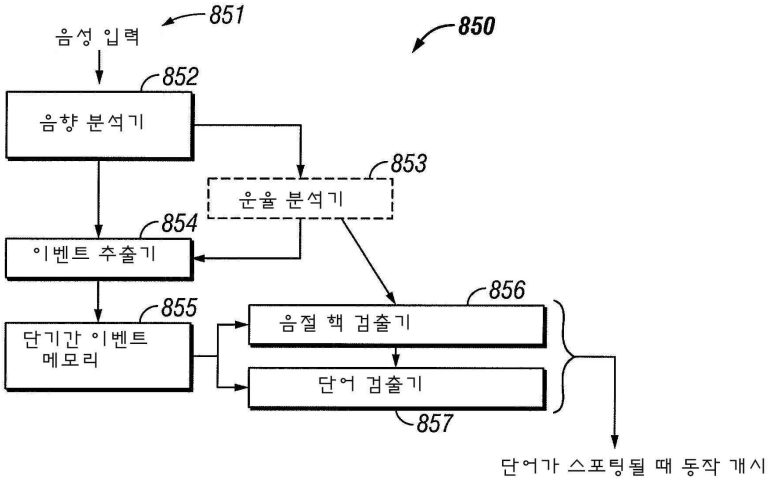
도면8a


입력 음성
801
800
802
음향 분석기
803
운율 분석기
이벤트 추출기
804
단기간 이벤트 메모리
805
806
음절 핵 검출기
807
음절 분류
808
음절 시퀀스 지각 색인
서브 음절 세부 분류
809
811
운율 정보를 갖는 텍스트 스트림

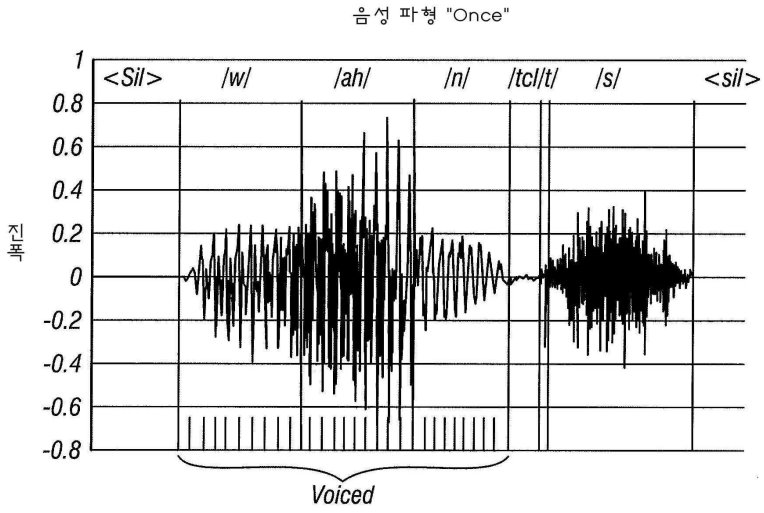
도면8b


820
821
음성 입력
822
음향 분석기
823
운율 분석기
824
음성 인식 엔진
825
이벤트 추출기
826
패턴 식별자
827
약한 영역 거부기
828
부스팅된 앙상블 간략화기
829
이벤트 시퀀스 인식기
830
다른 단서 검출기
831
캐스케이딩 검출기 앙상블 생성기
832
음성 생성기
833
지각 클러스터 명확화 모듈
834
835
836
단기간 이벤트 메모리
837
음절 핵 검출기
838
음절 분류기
839
음절 시퀀스 지각 색인
840
서브 음절 세부 분류
841
운율 정보를 갖는 텍스트 스트림

도면8c




도면9




도면10

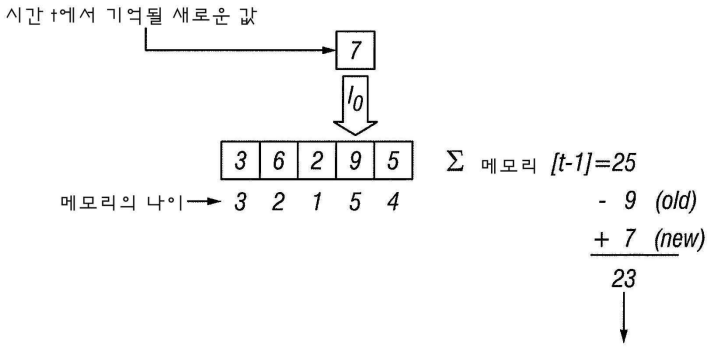
지각 대비 공식

$$C_{AB} = \frac{(A_{공식} - B_{공식})}{(A_{공식} + B_{공식} + \varepsilon)};$$

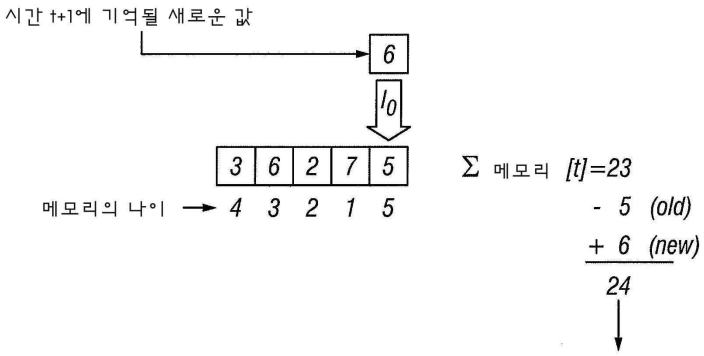
AAverage 및 BAverage는 평균 간격 값이다.

파라미터 ε은 최소 지각 활성화 레벨 값이다.

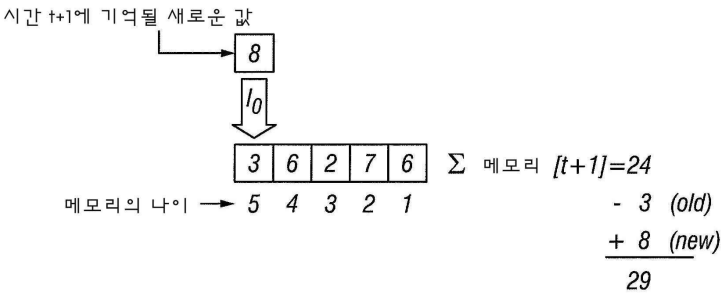
도면11a



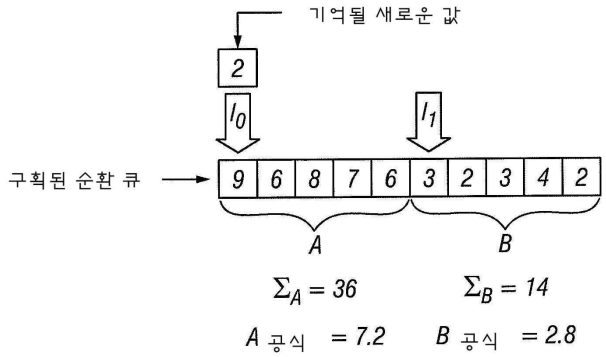

도면11b



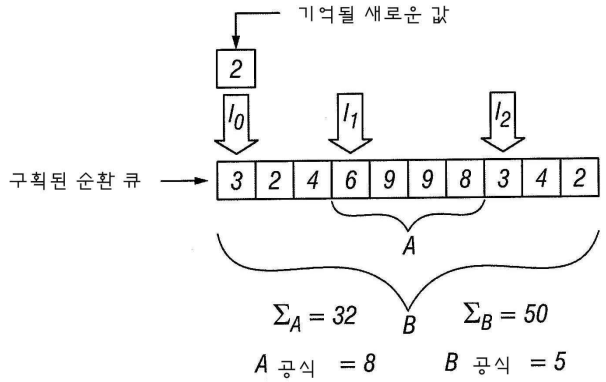

도면11c



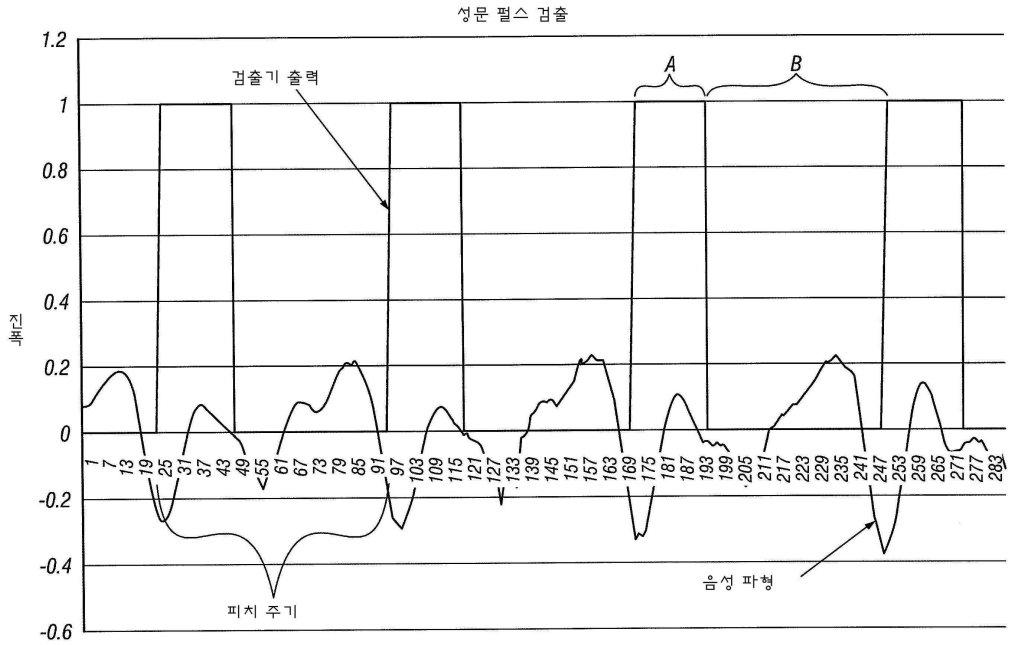
도면12



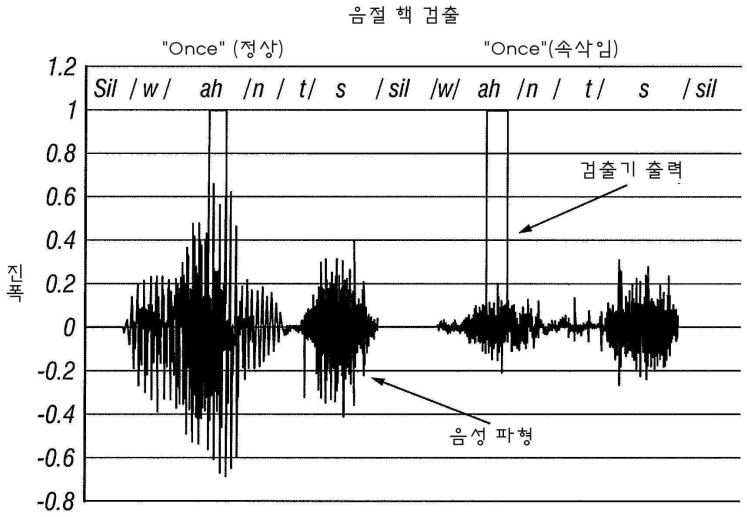

도면13



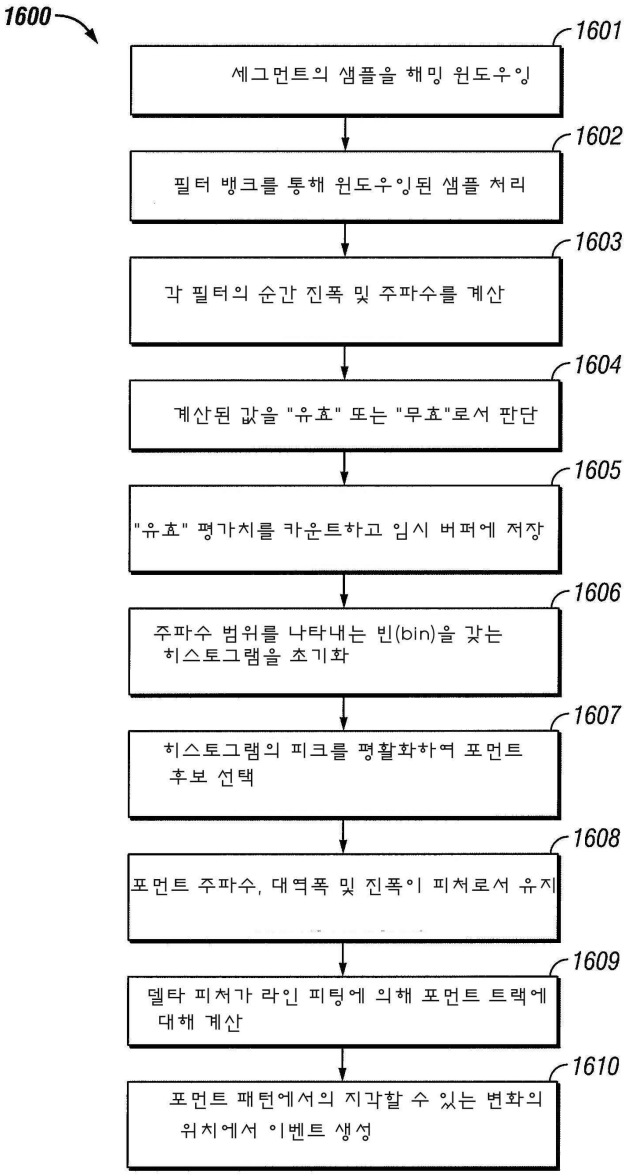

도면14



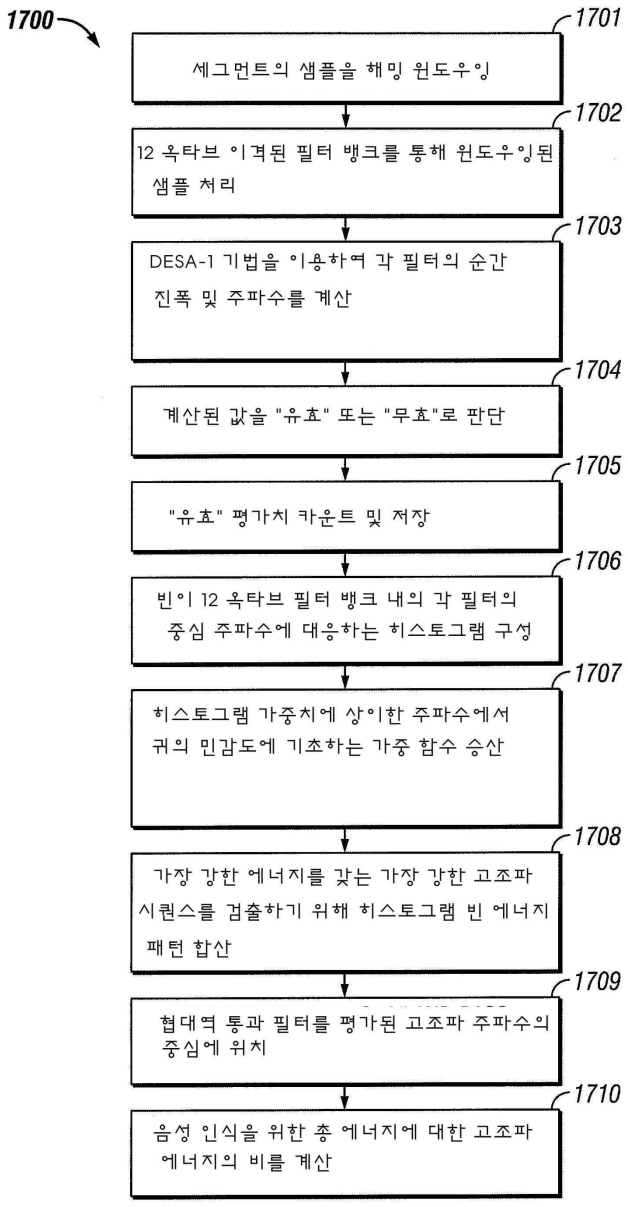
도면15


음절 핵 검출
"Once" (정상)
"Once"(속삭임)
Sil /w/ ah /n/ t/ s /sil /w/ ah /n/ t/ s /sil
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
진폭
검출기 출력
음성 파형

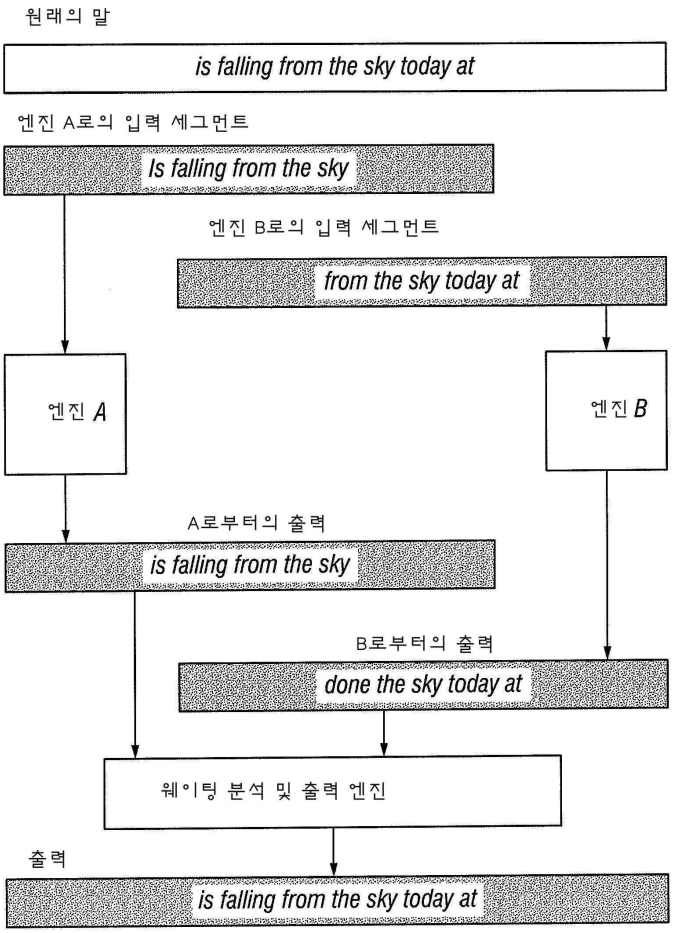
도면16


1600
1601
세그먼트의 샘플을 해밍 윈도우잉
1602
필터 뱅크를 통해 윈도우잉된 샘플 처리
1603
각 필터의 순간 진폭 및 주파수를 계산
1604
계산된 값을 "유효" 또는 "무효"로서 판단
1605
"유효" 평가치를 카운트하고 임시 버퍼에 저장
1606
주파수 범위를 나타내는 빈(bin)을 갖는 히스토그램을 초기화
1607
히스토그램의 피크를 평활화하여 포먼트 후보 선택
1608
포먼트 주파수, 대역폭 및 진폭이 피처로서 유지
1609
델타 피처가 라인 피팅에 의해 포먼트 트랙에 대해 계산
1610
포먼트 패턴에서의 지각할 수 있는 변화의 위치에서 이벤트 생성

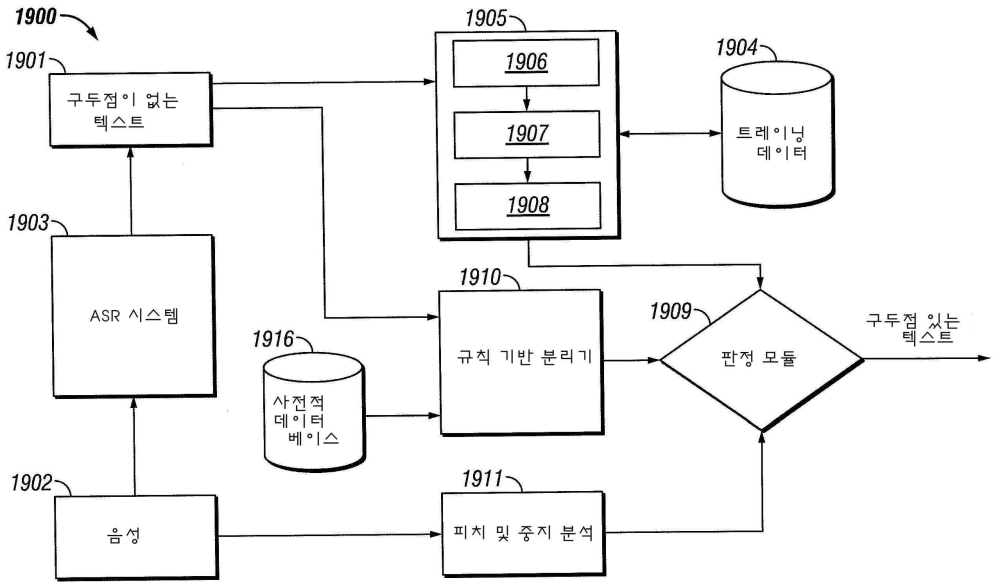
도면17


1700
1701
세그먼트의 샘플을 해밍 윈도우잉
1702
12 옥타브 이격된 필터 뱅크를 통해 윈도우잉된 샘플 처리
1703
DESA-1 기법을 이용하여 각 필터의 순간 진폭 및 주파수를 계산
1704
계산된 값을 "유효" 또는 "무효"로 판단
1705
"유효" 평가치 카운트 및 저장
1706
빈이 12 옥타브 필터 뱅크 내의 각 필터의 중심 주파수에 대응하는 히스토그램 구성
1707
히스토그램 가중치에 상이한 주파수에서 귀의 민감도에 기초하는 가중 함수 승산
1708
가장 강한 에너지를 갖는 가장 강한 고조파 시퀀스를 검출하기 위해 히스토그램 빈 에너지 패턴 합산
1709
협대역 통과 필터를 평가된 고조파 주파수의 중심에 위치
1710
음성 인식을 위한 총 에너지에 대한 고조파 에너지의 비를 계산

도면18


원래의 말
is falling from the sky today at
엔진 A로의 입력 세그먼트
Is falling from the sky
엔진 B로의 입력 세그먼트
from the sky today at
엔진 A
엔진 B
A로부터의 출력
is falling from the sky
B로부터의 출력
done the sky today at
웨이팅 분석 및 출력 엔진
출력
is falling from the sky today at

도면19


1900
1901
구두점이 없는 텍스트
1903
ASR 시스템
1902
음성
1905
1906
1907
1908
1904
트레이닝 데이터
1910
규칙 기반 분류기
1916
사전적 데이터 베이스
1911
피치 및 중지 분석
1909
판정 모듈
구두점 있는 텍스트