(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(51) Int. Cl.[6]
G10L 9/00

(11) 공개번호 특2000-0004972
(43) 공개일자 2000년01월25일

(21) 출원번호 10-1998-0707581
(22) 출원일자 1998년09월21일
번역문제출일자 1998년09월21일
(86) 국제출원번호 PCT/GB1997/00837
(87) 국제공개번호 WO 1997/37346
(86) 국제출원출원일자 1997년03월25일
(87) 국제공개일자 1997년10월09일
(81) 지정국 AP ARIPO특허 : 케냐 레소토 말라위 수단 스와질랜드

EA 유라시아특허 : 아르메니아 아제르바이잔 벨라루스

EP 유럽특허 : 오스트리아 벨기에 스위스 리히텐슈타인 독일 덴마크 스페인 프랑스 영국 그리스 이탈리아 룩셈부르크 모나코 네덜란드 포르투칼 오스트리아 스위스 리히텐슈타인 독일 덴마크 스페인 핀랜드 영국

국내특허 : 아일랜드 알바니아 오스트레일리아 보스니아-헤르체고비나 바베이도스 불가리아 브라질 캐나다 중국 쿠바 체크 에스토니아 그루지야 헝가리 이스라엘 아이슬란드 일본

(30) 우선권주장 96302236.3 1996년03월29일 EPO(EP)
(71) 출원인 브리티쉬 텔리커뮤니케이션즈 파블릭 리미티드 캄퍼니 내쉬 로저 윌리엄
영국 런던(우편번호 이시1에이 7에이제이) 뉴게이트 스트리트 81
(72) 발명자 밀너 벤자민 피터
영국 노퍽 (우편번호 쉬 엔알7 9엘제이) 노리치소어프 에스티.앤드류 암스트롱 로드 16
(74) 대리인 김명신, 김원오

심사청구 : 없음

(54) 음성 반응 장치에서 사용하기 위한 특징 발생 방법과 장치 및음성 인식 방법과 장치

요약

본 발명은 음성 인식에서 사용하기 위한 특징 발생 방법 및 장치에 관한 것으로서, 상기 방법은 입력 음성 신호의 프레임의 소정수 n 각각의 로그 프레임 에너지값을 계산하는 단계; 및 입력 음성 신호를 나타내는 시간 행렬을 형성하기 위해 n 로그 프레임 에너지값에 행렬 변환을 적용하는 단계를 포함하며, 행렬 변환은 이산 코사인 변환이 될 수 있는 것을 특징으로 한다.

대표도

도3

명세서

본 발명은 음성 인식에 관한 것으로서, 특히 음성 인식에서 사용되는 특징의 발생에 관한 것이다.

자동화 음성 인식 시스템은 일반적으로 특별한 목적을 위해 설계된다. 예를 들어, 공용으로 액세스되는 서비스에서는 임의의 사용자로부터의 음성을 인식하기 위해 설계된 일반적 음성 인식 시스템을 요구한다. 사용자 고유 데이터와 관련된 자동화 음성 인식기는 사용자를 인식하거나 또는 사용자 요구 아이덴티티를 검증(소위 화자(話者) 인식)하기 위해 사용된다.

자동화 음성 인식 시스템은 마이크로부터의 입력 신호를 직접 또는 간접적으로 (예를 들어 전기통신 링크를 통해) 수신한다. 그리고, 입력 신호는 일반적으로 시변화 입력신호의 적절한(스펙트럼) 특성 표현을 생성하므로써 입력신호를 연속적인 시간 세그먼트 또는 프레임으로 나누는 음성 처리 수단에 의해 처리된다. 스펙트럼 분석의 공통적인 기법은 LPC(linear predictive coding) 및 푸리에(Fourier) 변환이다. 다음, 스펙트럼 측정은 입력신호의 폭넓은 음향 특성을 설명하는 특징의 벡터 또는 한 세트의 특징으로 변환된다. 음성 인식에서 사용되는 대부분의 공통 특징은 MFCCs(mel-frequency cepstral coefficients)이다.

그리고, 특징 벡터는 인식될 어구 또는 단어(또는 그 일부)와 어떠한 방법으로 관련되거나 표현하는 다수의 패턴과 비교된다. 비교 결과는 인식된 것으로 생각되는 단어/어구를 나타낸다.

음성 인식에 관한 패턴 정합 접근은 두 가지 기법-템플릿 정합 또는 통계적 모형화중 하나와 관련된다. 전자의 경우, 템플릿은 워드를 나타내는 일반적인 음성 신호의 스펙트럼 특성을 나타내도록 형성된다. 각각의 템플릿은 음성 기간을 통한 스펙트럼 프레임의 연결이 된다. 따라서 패턴에서의 일반적인 음성 프레임 시퀀스는 평균화 프로시져를 통해 생산되고, 입력 신호는 이들 템플릿과 비교된다. 패턴 프레임

의 스펙트럼 특성을 특징짓는 잘 알려지고 널리 사용되는 한가지 통계 방법이 HMM(hidden Markov model) 접근이다. HMM(또는 어떤 다른 타입의 통계 모형)의 기초가 되는 가정은 음성 신호가 매개변수화 임의 프로세스로서 특징지어질 수 있고, 확률적 프로세스의 매개변수가 정확하게 잘 정의된 방법으로 결정될 수 있다는 것이다.

현재의 패턴 정합 기법, 특히 HMMs의 공지된 결함은 특징 추출의 상관 관계의 사용을 위한 유효 매카니즘의 부족이다. 좌우(left-right) HMM은 하나의 상태에서 다른 상태로의 음성 스펙트럼 특징의 시간 전개를 모형화하는 시간 구조를 제공하지만, 각각의 상태에서 관측 벡터는 IID(independent and identically distributed)인 것으로 가정된다. IID 가정이란 연속적인 음성 벡터간에 상관 관계가 전혀 없는 것을 나타낸다. 이것은 각각의 상태에서 음성 벡터가 동일한 평균 접근 시간 및 공분산을 갖는 PDFs(probability density functions)와 관련이 있다는 것을 의미한다. 이것은 또한 각각의 상태에서 스펙트럼적 시간 곡선이 고정된 평균 접근 시간에 임의로 변동하는 커브라는 것을 의미한다. 그러나, 실제로 스펙트럼적 시간 곡선은 명확하게 그것이 하나의 음성 이벤트에서 다른 음성 이벤트로 움직이는 것에 따른 정확한 방향을 갖는다.

IID 가정의 스펙트럼 벡터에 의한 이러한 방해는 HMMs의 성능을 제한한다. 음성 특징으로의 일부 시간인 정보를 포함하는 것은 음성이 고정된 독립적 프로세스이고, 인식 성능을 개선하기 위해 사용될 수 있다는 이러한 가정의 효과를 줄일 수 있다.

특징 벡터로의 시간 정보의 포함을 허용하는 종래 방법에서는 셉트럼(cepstrum)의 제 1 및 제 2 차수 미분계수와 함께 특징 벡터를 증가시킨다. 수학적으로 좀더 묵시적인 음성 다이내믹의 표현은 B.P. Milner 및 S.V. Vaseghi의 "An analysis of cepstral-time feature matrices for noise and channel robust speech recognition"(Proc. Eurospeech, pp 519-522, 1995)에서 설명된 바와 같이 시간 정보를 엔코드하기 위해 코사인 변환을 사용하는 셉트럼적 시간 행렬이다.

셉트럼적 시간 행렬, $c_t(m,n)$은 2-D DCT(Discrete Cosine Transform)을 스펙트럼 시간 행렬에 적용하거나, 또는 1-D DCT를 MFCC(mel-frequency cepstral coefficients) 음성 벡터의 스택화에 적용하므로써 얻어진다. M N차원 로그 필터 뱅크 벡터들은 스펙트럼 시간 행렬, $X_t(f,k)$을 형성하기 위해 함께 스택되고, 여기서 t는 시간 프레임, f는 필터 뱅크 채널, 그리고 k는 행렬내 시간 벡터이다. 그러면 스펙트럼 시간 행렬은 2차원 DCT를 사용하여 셉트럼적 시간 행렬로 변환된다. 2차원 DCT는 두 개의 1차원 DCTs로 나눠질 수 있기 때문에, 셉트럼적 시간 행렬의 대안적인 구현에서는 M개의 종래 MFCC 벡터로 구성된 행렬의 시간축에 따라 1-D DCT를 사용한다.

본 발명의 제 1 측면에 따르면 입력 음성 신호 프레임의 소정수 n 각각의 대수적 프레임 에너지값을 계산하는 단계; 및 입력 음성 신호를 나타내는 시간인 행렬을 형성하기 위해 n 대수적 프레임 에너지값에 변환 행렬을 적용하는 단계로 이루어지는, 음성 인식에서 사용하기 위한 특징 발생 방법이 제공된다.

증가된 미분계수를 갖는 셉트럼 벡터로 이루어진 명시적 표현과 비교하면, 음성 전이적 다이내믹은 시간 벡터내에서 묵시적으로 생산된다. 따라서, 그러한 행렬에서 열이 이뤄진 모형들은 소음 강도를 개선하기 위한, PMC(parallel model combination)와 같은 기법에서 선형 필터 뱅크 도메인으로의 변환을 허용하여 역변환이 적용될 수 있다는 이점을 갖는다.

변환은 이산 코사인 변환이 될 수 있다. 대개, 시간 행렬은 n 이하의 요소를 포함하도록 끊긴다. 이것은 관련된 계산량을 감소시켜 우수한 성능 결과를 낳는 것으로 밝혀졌다. 행렬의 고정상태(m=0) 행은 생략될 수 있고, 그래서 행렬에 채널 강도 특징을 만드는 선형 회선상태의 채널 왜곡에 의해 어떠한 음성 신호 왜곡을 제거한다.

본 발명은 또한 본 발명의 방법을 발생시키는 특징을 포함하는 음성 인식 방법과 관련된다.

본 발명의 또다른 측면에서,

입력 음성 신호 프레임의 소정수 n 각각의 에너지의 대수를 계산하는 프로세서; 및

입력 음성을 나타내는 시간 행렬을 형성하기 위해 계산되는 n 대수적 에너지값에 변환을 적용하는 프로세서로 이루어지는, 음성 반응 장치에서 사용하기 위한 특징 발생 장치가 제공된다.

본 발명의 특징 발생 수단은 음성 인식 장치에서 사용하기에 적당하고, 또한 그러한 장치에서 사용하기 위한 인식 데이터를 발생시키기에 적당하다.

지금부터 첨부한 도면을 참조하여 예의 방법으로 본 발명을 설명한다.

도 1은 전기통신 환경에서의 음성 인식기의 사용을 개략적으로 나타낸 도면,

도 2는 음성 인식기를 개략적으로 표현한 도면,

도 3은 본 발명에 따른 특징 추출기의 하나의 실시예의 구성요소를 개략적으로 나타낸 도면,

도 4는 KL(Karhunen-Loeve) 변환을 결정하는 단계를 나타내는 도면,

도 5는 도 2의 음성 인식기의 일부를 형성하는 종래 음성 분류기의 구성요소를 개략적으로 나타낸 도면,

도 6은 도 5의 분류기의 작동을 개략적으로 나타내는 흐름도,

도 7은 도 2의 음성 인식기의 일부를 형성하는 종래 시퀀서의 구성요소를 개략적으로 나타낸 블럭도,

도 8은 도 7의 시퀀서의 일부를 형성하는 기억장치내 필드 내용을 개략적으로 나타낸 도면, 및

도 9는 도 7의 시퀀서의 작동을 개략적으로 나타내는 흐름도이다.

도 1을 참조하면, 음성 인식을 포함하는 전기통신 시스템은 일반적으로 (대개 전화 송수화기의 일부를

형성하는) 마이크(1), 전기통신 네트워크(2)(대개 PSTN(public switched telecommunications network)), 네트워크(2)로부터 음성 신호를 수신하기 위해 연결된 음성 인식기(3), 및 음성 인식기(3)와 연결되고 그로부터 음성 인식 신호를 수신하도록 배치되어, 특정한 단어 또는 어구의 인식 또는 인식하지 않음을 나타내며 그에 따른 응답 행위를 취하는 이용 장치(4)로 이루어진다. 예를 들어, 이용 장치(4)는 은행 업무, 정보 서비스 등을 하기 위해 원격 작동 터미널이 될 수 있다.

여러 경우에서, 이용 장치(4)는 일반적으로 사용자의 송수화기 일부를 형성하는 확성기(5)로 네트워크(2)를 통해 전송되는, 사용자에게 들리는 응답을 발생시킬 것이다.

작동중에, 사용자는 마이크(1)로 말하고, 신호는 네트워크(2)를 통해 마이크(1)로부터 음성 인식기(3)로 전송된다.

음성 인식기는 음성 신호를 분석하고, 특정한 단어 또는 어구의 인식 또는 인식하지 않음을 나타내는 신호가 생성되어 이용 장치(4)로 전송되며, 음성 인식 상황에서의 적절한 동작을 취한다.

일반적으로, 음성 인식기(3)는 신호에 의해 취해진 마이크(1)로부터 네트워크(2)를 통한 네트워크(2)로의 루트를 알지 못한다. 매우 다양한 타입 또는 특성의 송수화기중의 하나가 사용될 수 있다. 유사하게, 네트워크(2)내에서, 다양한 전송 경로중 임의의 한 경로가 결정될 수 있고, 무선 링크, 아나로그 및 디지털 경로 등을 포함한다. 따라서, 음성 인식기(3)에 이르는 음성 신호(Y)는 마이크(1)로 수신된 음성 신호(S)에 대응하고, 마이크(1)의 변환 특성, 네트워크(2)로의 링크, 네트워크(2)를 통한 채널, 및 음성 인식기(3)로의 링크와 관련되며, 하나의 전송 특성(H)으로 일괄되고 설계될 수 있다.

일반적으로, 음성 인식기(3)는 음성 신호를 확인하기 위해 음성과 관련된 데이터를 요구할 필요가 있고, 이러한 데이터 확인은 음성 인식기(3)가 그 단어 또는 어구를 위한 인식 데이터를 형성하기 위해 마이크(1)로부터 음성 신호를 수신하는 작동의 트레이닝 모드에서 음성 인식기에 의해 수행된다. 그러나, 음성 인식 데이터를 요구하는 다른 방법 또한 가능하다.

도 2를 참조하면, 음성 인식기는 (디지털 네트워크 또는 아나로그 디지털 변환기로부터) 디지털 형태로 음성을 수신하는 입력(31); 연속적인 디지털 샘플을 연속적인 샘플의 연속적인 프레임으로 나누는 프레임 발생기(32); 샘플의 프레임으로부터 대응하는 특징 벡터를 발생시키는 특징 추출기(33); 연속적인 특징 벡터를 수신하고, 인식 결과를 생성하는 분류기(34); 입력 신호가 가장 큰 유사성을 나타내는 소정의 발음을 결정하는 시퀀서(35); 및 인식 신호가 인식된 음성 발성을 나타내도록 제공되는 출력 포트(35)로 이루어진다.

상기한 바와 같이, 음성 인식기는 일반적으로 트레이닝 단계동안 인식 데이터를 습득한다. 트레이닝동안, 음성 신호는 음성 인식기(3)로 입력되고, 특징은 본 발명에 따른 특징 추출기(33)에 의해 추출된다. 이러한 특징은 잇따른 인식을 위해 음성 인식기(3)에 의해 저장된다. 후술되는 바와 같이, 특징은 음성 처리에서 잘 알려진 기법, 예를 들어 HMMs에 의해 모형화된 임의의 편리한 형태로 저장될 수 있다. 인식동안, 특징 추출기는 공지되지 않은 입력 신호로부터 유사한 특징을 추출하고, 공지되지 않은 신호 특징을 인식될 각각의 단어/어구를 위해 저장된 특징(들)과 비교한다.

간단하게 하기 위해, 인식 단계에서의 음성 인식기의 작동을 후술한다. 트레이닝 단계에서, 추출된 특징은 상기 기술분야에서 잘 알려진 바와 같이 적당한 분류기(34)를 트레인하기 위해 사용된다.

프레임 발생기(32)

프레임 발생기(32)는 예를 들어 초당 8000 샘플의 속도로 음성 샘플을 수신하고, 매 16ms당 1 프레임의 프레임 속도로 256개 연속적인 샘플로 이루어지는 프레임을 형성하도록 배치된다. 대개, 각각의 프레임은 프레임 가장자리에서 발생된 가짜 인공물을 감소시키기 위해 예를 들어 해밍(Hamming) 윈도우를 이용하여 창이 내어진다(즉, 프레임 가장자리로 향하는 샘플들은 소정의 부가 정수가 곱해진다). 적절한 실시예에서, 프레임은 창을 내는 효과를 개선하도록 (예를 들어 50%씩) 중첩된다.

특징 추출기(33)

특징 추출기(33)는 프레임 발생기(32)로부터 프레임을 수신하고, 각각의 프레임으로부터 특징 또는 특징 벡터를 생성한다. 도 3은 본 발명에 따른 특징 추출기의 실시예를 나타낸다. 다른 특징, 예를 들어 LPC 셉트럼 계수 또는 MFCCs를 생성하기 위해 수단이 추가적으로 제공될 수 있다.

입력 음성 신호의 각각의 프레임 j은 데이터 프레임의 평균 에너지를 계산하는 프로세서(331)로 입력된다, 즉 에너지 계산기 프로세서(331)는:

$$E_{av_j} = \frac{1}{256} \sum_{i=1}^{256} x_i^2$$

여기서 x_i는 프레임 j내 샘플 i의 에너지값이다.

그리고, 대수 프로세서(332)는 프레임 j을 위한 이 평균값의 로그를 형성한다. 로그 에너지값은 예를 들어 n=7인 n 프레임을 위한 로그 에너지 값을 저장하기에 충분한 길이를 갖는 버퍼(333)로 입력된다. 일단 데이터의 7 프레임 가치가 계산되면, 스택된 데이터는 변환 프로세서(334)로 출력된다.

프레임 에너지 벡터 또는 시간 행렬의 형성에서, 변환 프로세서(334)로 입력된 스택된 로그 에너지값의 스펙트럼 시간 벡터는 변환 행렬로 곱해진다. 즉,

MH=T

여기서 M은 스택된 로그 에너지값의 벡터, H는 시간 정보를 엔코드할 수 있는 변환, T는 프레임 에너지 벡터이다.

변환 H의 행은 시간 정보를 엔코딩하기 위한 기본적인 기능이다. 시간 정보를 엔코딩하는 이러한 방법을 이용하여, 폭넓은 범위의 변환이 시간 변환 행렬 H와 같이 사용될 수 있다.

변환 H은 시간 정보를 엔코드한다, 즉 변환 H는 로그 에너지값 스택의 공분산 행렬이 대각선화되도록 한다. 즉, H에 의해 변환된 로그 에너지값의 공분산 행렬의 대각선에서 벗어난 요소(즉, 주도적이지 않은 대각선)는 제로로 간주된다. 공분산 행렬의 대각선에서 벗어난 요소는 각각의 샘플들간의 상관 관계의 등급을 나타낸다. 이것을 성취하는 최적의 변환이 N.S. Jayant 및 P. Noll의 "Digital coding of waveforms"(Prentice-Hall, 1984)에 설명된 바와 같은 KL(Karhunen-Loeve) 변환이다.

특징 벡터에 의해 전달된 시간 정보를 엔코딩하기 위한 최적의 KL 변환을 찾기 위해서, 연속적인 벡터의 상관관계를 고려한 통계가 필요하다. 그리고, 이러한 상관관계 정보를 사용하여, KL 변환이 계산될 수 있다. 도 4는 음성 데이터로부터 KL 변환을 결정하는 것과 관련된 프로시져를 나타내고 있다.

KL 변환을 정확하게 결정하기 위해, 전체 트레이닝 데이터 세트가 먼저 로그 에너지값으로 매개변수화된다. 시간내에 n개 연속적인 로그 에너지값을 포함하는 벡터 x_t가 생성된다:

$$x_t=[c_t,c_{t-1},\ldots,c_{t+n-1}]$$

트레이닝 세트를 통한 이들 벡터들의 전체 세트로부터, 공분산 행렬 Σ_{xx}는 $Qxx=E\{xx^T\}-\mu_x\,\mu_x^T$ 로 계산되고, 여기서 μ_x는 로그 에너지값의 평균 벡터이다.

상기한 바와 같이, 이것은 상관관계 행렬 $E\{xx^T\}$과 매우 밀접하게 관련되고, 그와 마찬가지로 시간인 음성 다이내믹을 고려한 정보를 포함한다. KL 변환은 공분산 행렬의 고유벡터로부터 결정되고, 예를 들어 단수값 분해를 이용하여 계산될 수 있다.

$$H^T QxxH=dia(\lambda_0,\lambda_1,\ldots,\lambda_M)=\Lambda$$

결과적인 행렬 H는 공분산 행렬의 고유벡터로부터 만들어진다. 이들은 그 각각의 고유값 λ_i의 크기에 따라 랭크된다. 이 행렬은 KL 유도된 시간 변환 행렬이다.

Legendre, Laguerre 등과 같은 시간 변환 행렬을 생성하기 위해 다른 다항식이 사용될 수 있다. KL 변환은 각각의 트레이닝 데이터 세트를 위한 그자신의 변환을 계산하기 위한 필요성에 의해 복잡해진다. 대신, DCT(Discrete Cosine Transform)도 사용될 수 있다. 이러한 경우, 변환 프로세서(334)는 n 프레임을 위한 로그 에너지값과 관련된 스택된 데이터의 DCT를 계산한다.

1차원 DCT는 다음과 같다:

$$F(u)=\sqrt{\frac{2}{n}}\,C(u)\sum_{i=0}^{n-1}f(i)\cos\left[\frac{(2i+1)upi}{2n}\right]$$

여기서, f(i) = 프레임 i에서의 로그 에너지값

C(u) = u=0일 경우 $1/\sqrt{2}$

= 그렇지 않은 경우 1

u는 0 내지 n-1까지의 정수이다.

변환 프로세서(334)는 데이터의 n 프레임으로부터 발생된 n DCT 계수를 출력한다. 이들 계수들은 입력 신호의 에너지 레벨과 관련된 프레임 에너지 벡터를 형성한다.

프레임 에너지 벡터는 n=7인 경우 예를 들어 프레임 0 내지 6, 1 내지 7, 2 내지 8 등에서 입력 신호의 연속적인 n 프레임 각각을 위해 형성된다. 프레임 에너지 벡터는 음성 프레임을 위한 특징 벡터의 일부를 형성한다. 이러한 특징은 다른 특징, 예를 들어 MFCCs 또는 미분 MFCC를 증대시키기 위해 사용될 수 있다.

분류기(34)

도 5를 참조하면, 분류기(34)는 종래 디자인중의 하나이고, 상기 실시예에서는 HMM 분류 프로세서(341), HMM 상태 메모리(342), 및 모드 메모리(343)로 이루어진다.

상태 메모리(342)는 인식될 다수의 음성 부분 각각에서 상태 필드(3421,3422,...)로 이루어진다. 예를 들어, 상태 필드는 인식될 각각의 워드 음소를 위해 상태 메모리(342)에 제공될 수 있다. 소음/침묵을 위한 상태 필드도 제공될 수 있다.

상태 메모리(342)내 각각의 상태 필드는 모드 메모리(343)내 모드 필드 세트(361,362,...)로의 포인터 어드레스를 저장하는 포인터 필드(3421b,3422b,...)를 포함한다. 각각의 모드 필드 세트는 다수의 모드 필드(3611,3612,...)로 이루어지고, 각각은 질의내 상태를 특징짓는 특징 계수값의 다차원 가우스 분산을 정의한다. 예를 들어, 만일 각각의 특징(예를 들어 본 발명의 에너지 행렬의 7개 계수 및 제 1의 8 MFCC 계수)내에 d 계수가 있는 경우, 각각의 모드를 특징짓는 각각의 모드 필드(3611,3612,...)내에 저장된 데이터는: 상수 C, 한 세트의 d 특징 평균값 μ_i, 및 한 세트의 d 특징 편차 σ_i이다; 다시 말해서, 전체 2d+1개이다.

각각의 모드 필드 세트(361,362,...)내 모드 필드(3611,3612,...)의 개수 N_i는 가변적이다. 모드 필드는 트레이닝 단계동안 발생되고, 특징 추출기에 의해 유도된 특징(들)을 나타낸다.

인식하는 동안, 분류 프로세서(34)는 메모리(342)내 각각의 상태 필드를 차례로 읽도록 배치되고, 본 발명의 특징 추출기(33)에 의해 출력된 현재 입력 특징 계수 세트를 사용하여 각각에서 입력 특징 세트 또는 벡터가 대응하는 상태와 대응할 확률을 계산한다. 그렇게 하기 위해, 도 6에 도시된 바와 같이, 프로세서(341)는 그것이 가리키는 모드 메모리(343)내 모드 필드 세트를 액세스하고, 모드 필드 세트내 각각의 모드 필드 j에서의 지정 확률 P_j를 계산하기 위해, 상태 필드내 포인터를 읽도록 배치된다.

다음, 프로세서(341)는 지정 확률 P_j을 합하므로써 상태 확률을 계산한다. 따라서, 분류 프로세서(341)의 출력은 다수의 상태 확률 P이 되고, 상태 메모리(342)내 각각의 상태에서의 상태 확률은 입력 특징 벡터가 각각의 상태와 대응하는 유사성을 나타낸다.

도 6은 분류 프로세서(341)의 작동을 간단하게 설명하고 있다는 것을 알 수 있다. 실제로, 모드 확률 각각은 모드가 대응하는 음소와 관련된 모든 상태 확률의 계산에 사용되도록, 일단 계산되고, 시간으로 저장될 수 있다.

분류 프로세서(341)는 적당하게 프로그램된 DSP(digital signal processing) 장치가 될 수 있고, 특히 특징 추출기(33)와 동일한 DSP 장치가 될 수 있다.

시퀀서(35)

도 7을 참조하면, 시퀀서(35)는 종래 디자인이고, 상기 실시예에서는 각각의 프로세스된 프레임에서 분류 프로세서(341)에 의해 출력된 상태 확률을 저장하는 상태 확률 메모리(353); 상태 시퀀스 메모리(352); 분석 프로세서(351); 및 시퀀서 출력 버퍼(354)로 이루어진다.

상태 시퀀스 메모리(352)는 다수의 상태 시퀀스 필드(3521,3522,...)로 이루어지고, 각각은 상기 실시예에서 하나의 음소 문자열로 이루어지는 인식될 단어 또는 어구 시퀀스에 대응한다. 상태 시퀀스 메모리(352)내 각각의 상태 시퀀스는 도 8에 도시된 바와 같이 다수의 상태 $P_1,P_2,...P_N$ 및 각각의 상태에서의 두가지 확률; 반복 확률(P_{i1}) 및 다음 상태로의 전이 확률(P_{i2})로 이루어진다. 따라서, 일련의 프레임과 관련된 관측된 상태 시퀀스는 각각의 상태 시퀀스 모형(3521등)내 각각의 상태 P_i의 일부 반복으로 이루어진다. 예를 들면 다음과 같다.

프레임 번호	1	2	3	4	5	6	7	8	9	...Z	Z+1
상태	P1	P1	P1	P2	P2	P2	P2	P2	P2	...Pn	Pn

도 9에 도시된 바와 같이, 시퀀스 프로세서(351)는 각각의 프레임에서 분류 프로세서(341)에 의해 출력된 상태 확률 및 상태 확률 메모리(353)내 먼저 저장된 상태 확률을 읽고, 시간내내 일자를 기입하기 위해 가장 유사한 상태 경로를 계산하며, 이것을 상태 시퀀스 메모리(352)내에 저장된 상태 시퀀스 각각과 비교하도록 배치된다.

계산을 위해 S.J. Cox의 "Hidden Markov Models for Automatic Speech Recognition: theory and applications"(British Telecom Technology Journal, 1998, 4, p105)에 일반적으로 설명된 잘 알려진 HMM 기법을 채택한다. 편리하게, 시퀀스 프로세서(351)에 의해 수행된 HMM 프로세싱에서는 공지된 비터비(Viterbi) 알고리즘을 사용한다. 시퀀스 프로세서(351)는 예를 들어 Intel(TM) i-486(TM) 마이크로프로세서 또는 Motorola(TM) 68000 마이크로프로세서와 같은 마이크로프로세서가 될 수 있거나, 또는 대안적으로 DSP 장치(예를 들어, 상기 프로세서중의 임의의 하나에서 사용된 것과 동일한 DSP 장치)가 될 수 있다.

따라서, (인식될 단어, 어구, 또는 다른 음성 시퀀스에 대응하는) 각각의 상태 시퀀스에서 확률 스코어는 입력 음성의 각각의 프레임에서 시퀀스 프로세서(351)에 의해 출력된다. 예를 들어, 상태 시퀀스는 전화번호부내 이름으로 이루어질 수 있다. 발성의 끝이 검출되는 경우, 가장 확률있는 상태 시퀀스를 나타내는 라벨 신호는 대응하는 이름, 단어 또는 어구가 인식되었는지를 나타내기 위해 시퀀스 프로세서(351)에서 출력 포트(38)로 출력된다.

(57) 청구의 범위

청구항 1

음성 반응 장치에서 사용하기 위한 특징 발생 방법에 있어서,

입력 음성 신호 프레임의 소정수 n 각각의 대수적 프레임 에너지값을 계산하는 단계; 및 입력 음성 신호를 나타내는 시간인 행렬을 형성하기 위해 n 대수적 프레임 에너지값에 변환 행렬을 적용하는 단계로 이루어지는 것을 특징으로 하는 방법.

청구항 2

제 1 항에 있어서,

연속적인 특징은 입력 신호의 n 프레임 그룹을 중첩시키는 것을 나타내는 것을 특징으로 하는 방법.

청구항 3

제 1 항 또는 제 2 항에 있어서,

변환 행렬은 이산 코사인 변환인 것을 특징으로 하는 방법.

청구항 4

제 1 항 내지 제 3 항 중 어느 한 항에 있어서,

시간 행렬은 n 이하의 요소를 포함하도록 끊기는 것을 특징으로 하는 방법.

청구항 5

음성을 나타내는 입력 신호를 수신하는 단계;

입력 음성 신호의 소정수 n 프레임 각각의 대수적 프레임 에너지값을 계산하으로써 특징을 발생시키고, 입력 음성 신호를 나타내는 시간 행렬을 형성하기 위해 n 대수적 프레임 에너지값에 매트릭스 변환을 적용하는 단계;

발생된 특징과 허용된 발성을 나타내는 인식 데이터를 비교하는 단계; 및

비교 단계를 기초로 인식 또는 인식하지 않음을 나타내는 단계로 이루어지고,

상기 입력 신호는 프레임으로 나누어지며, 상기 인식 데이터는 특징과 관련되는 것을 특징으로 하는 음성 인식 방법.

청구항 6

제 5 항에 있어서,

변환 프로세서는 n 대수적 에너지값에 이산 코사인 변환을 적용하도록 배치되는 것을 특징으로 하는 음성 인식 방법.

청구항 7

음성 반응 장치에서 사용하기 위한 특징 발생 장치에 있어서,

입력 음성 신호 프레임의 소정수 n 각각의 에너지의 대수를 계산하는 프로세서; 및

입력 음성을 나타내는 시간 행렬을 형성하기 위해 계산되는 n 대수적 에너지값에 변환을 적용하는 프로세서로 이루어지는 것을 특징으로 하는 장치.

청구항 8

제 7 항에 있어서,

변환 프로세서는 n 대수적 에너지값에 이산 코사인 변환을 적용하도록 배치되는 것을 특징으로 하는 장치.

청구항 9

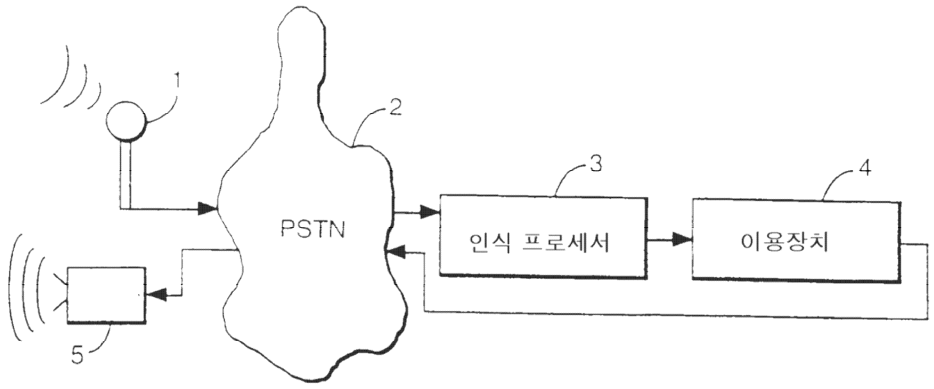
제 7 항 또는 제 8 항에 따른 특징 발생 장치를 포함하는 음성 인식 장치.

청구항 10

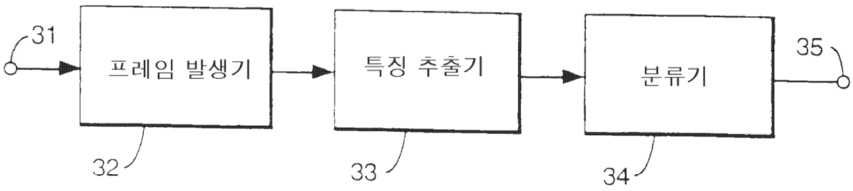
제 1 항에 따라 발생된 특징과 관련된 인식 데이터를 수신하기 위한 입력을 포함하는 음성 인식 장치.

도면

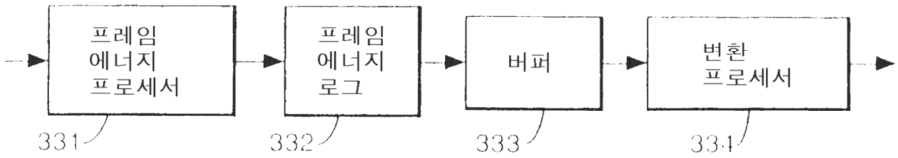
도면1



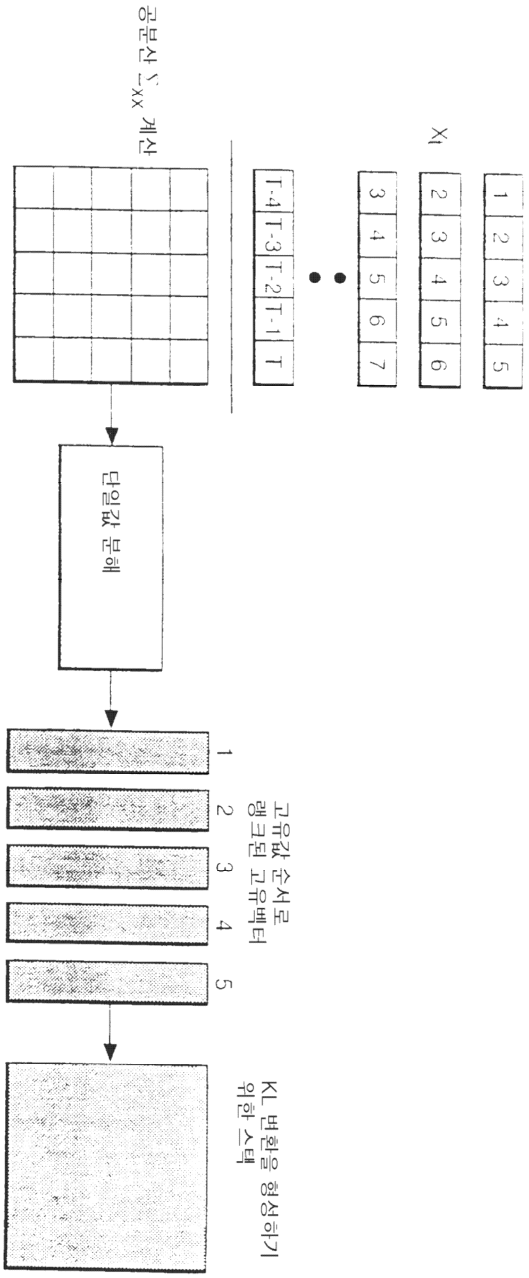
도면2


31
프레임 발생기
32
특징 추출기
33
분류기
34
35

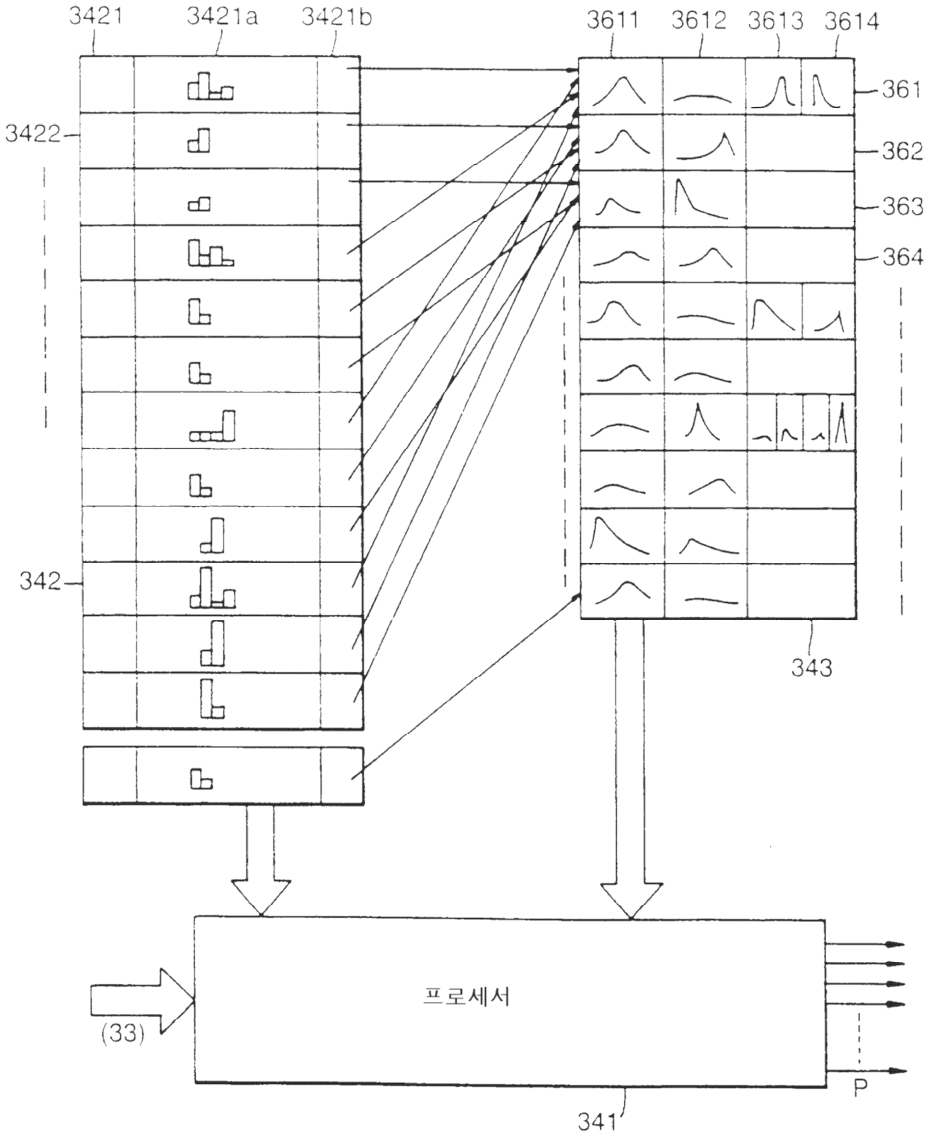
도면3


프레임
에너지
프로세서
331
프레임
에너지
로그
332
버퍼
333
변환
프로세서
334

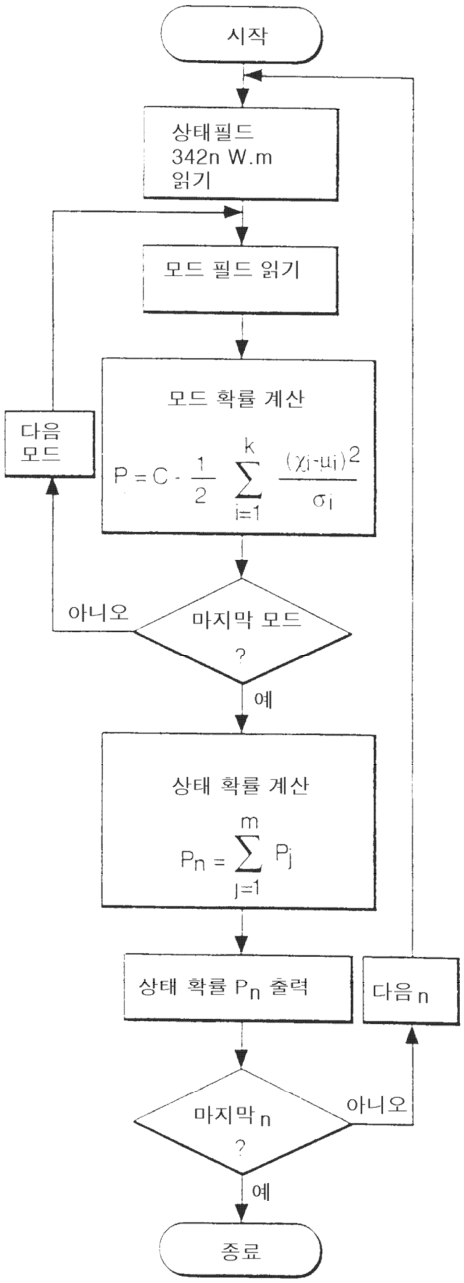
도면4


공분산 Σ_{xx} 계산
X_t
1 2 3 4 5
2 3 4 5 6
3 4 5 6 7
T-4 T-3 T-2 T-1 T
단일값 분해
1
2
3
4
5
고유값 순서로 랭크된 고유벡터
KL 변환을 형성하기 위한 스택

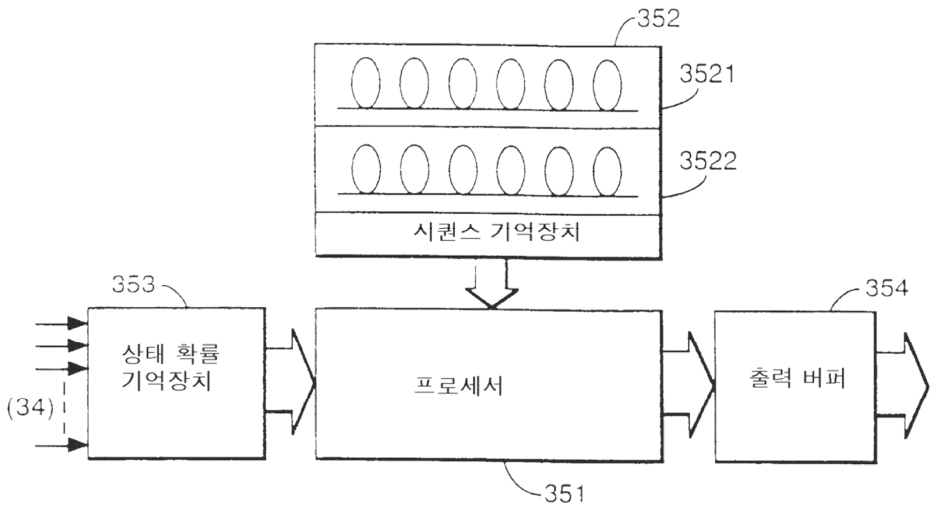
도면5


3421
3421a
3421b
3611
3612
3613
3614
3422
342
361
362
363
364
343
프로세서
(33)
341
P

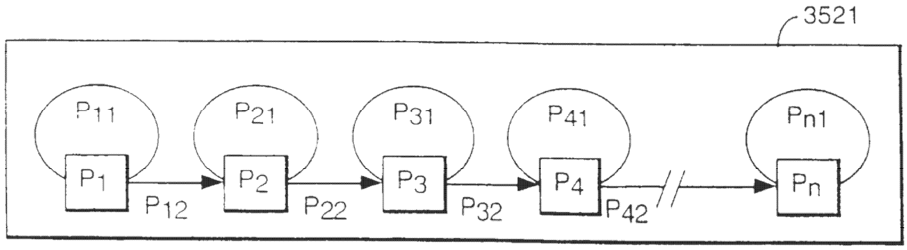
도면6


시작
상태필드
342n W.m
읽기
모드 필드 읽기
모드 확률 계산
$P = C \cdot \frac{1}{2} \sum_{i=1}^{k} \frac{(\chi_i - u_i)^2}{\sigma_i}$
다음
모드
아니오
마지막 모드
?
예
상태 확률 계산
$P_n = \sum_{j=1}^{m} P_j$
상태 확률 P_n 출력
다음 n
마지막 n
?
아니오
예
종료

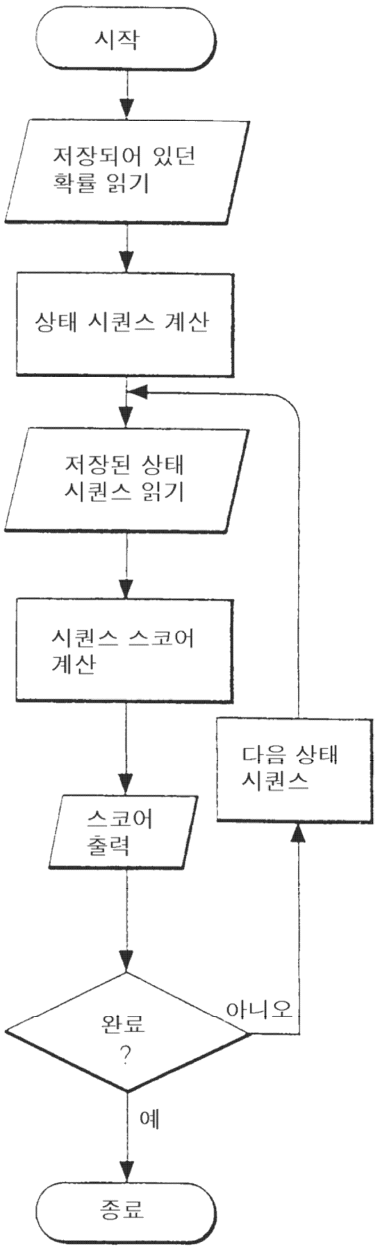
도면7


352
3521
3522
시퀀스 기억장치
353
상태 확률
기억장치
(34)
프로세서
354
출력 버퍼
351

도면8


3521
P11
P21
P31
P41
Pn1
P1
P2
P3
P4
Pn
P12
P22
P32
P42

도면9


시작
저장되어 있던 확률 읽기
상태 시퀀스 계산
저장된 상태 시퀀스 읽기
시퀀스 스코어 계산
다음 상태 시퀀스
스코어 출력
완료 ?
아니오
예
종료