(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2013-0127077
(43) 공개일자 2013년11월22일

(51) 국제특허분류(Int. Cl.)
G10L 15/00 (2006.01)
(21) 출원번호 10-2012-0050706
(22) 출원일자 2012년05월14일
심사청구일자 2012년05월14일

(71) 출원인
고려대학교 산학협력단
서울 성북구 안암동5가 1
(72) 발명자
고한석
서울특별시 용산구 이촌1동 삼성리버스위트아파트 103동 1602호
박진수
서울특별시 강남구 대치1동 롯데캐슬리베아파트 101동 803호
문성규
경기도 안양시 동안구 평안동 초원 LG아파트 502동 301호
(74) 대리인
최관락, 송인호, 민영준

전체 청구항 수 : 총 10 항

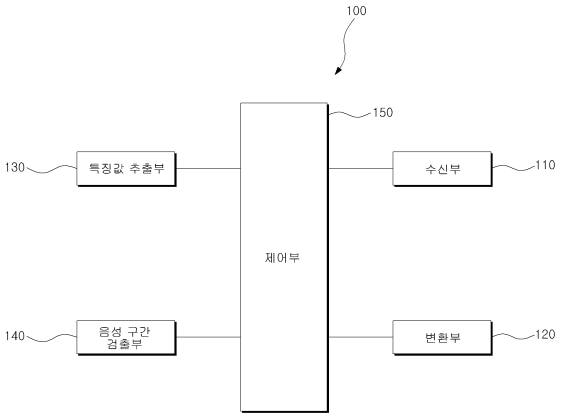
(54) 발명의 명칭 **음성 인식 방법 및 장치**

(57) 요 약

본 발명은 잡음과 음성이 혼합된 환경에서 주파수 변환 및 원점 대칭 필터를 이용하여 음성 구간을 검출하는 음성 인식 방법 및 장치에 관한 것이다.

본 발명의 일실시예에 의한 음성 인식 방법은 잡음과 음성 신호가 혼합된 시간 영역의 오디오 입력 신호를 수신하는 단계; 상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변환하는 단계; 음성 부재 확률에 대한 음성 존재 확률의 비를 이용하여 상기 오디오 입력 신호의 특징값을 추출하는 단계-상기 음성 존재 확률 및 음성 부재 확률은 상기 변환된 주파수 영역의 신호를 이용하여 획득됨-; 상기 추출된 특징값에 원점 대칭 필터를 적용하여 음성 구간을 검출하는 단계를 포함할 수 있다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	000476190111
부처명	중소기업청
연구사업명	산학연공동기술개발사업
연구과제명	텔레매틱스 음성 인터페이스를 위한 멀티 채널 마이크를 이용한 방향성 잡음 제거 기술 개발
기 여 율	1/1
주관기관	고려대학교 산학협력단
연구기간	2011.08.01 ~ 2012.05.31

특허청구의 범위

청구항 1

잡음과 음성 신호가 혼합된 시간 영역의 오디오 입력 신호를 수신하는 단계;

상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변환하는 단계;

음성 부재 확률에 대한 음성 존재 확률의 비를 이용하여 상기 오디오 입력 신호의 특징값을 추출하는 단계-상기 음성 존재 확률 및 음성 부재 확률은 상기 변환된 주파수 영역의 신호를 이용하여 획득됨-;

상기 추출된 특징값에 원점 대칭 필터를 적용하여 음성 구간을 검출하는 단계를 포함하는 것을 특징으로 하는 음성 인식 방법.

청구항 2

제 1 항에 있어서, 상기 음성 구간 검출 단계는

상기 추출된 오디오 신호의 특징값에 원점 대칭 필터를 적용하여 출력 신호를 획득하는 단계;

상기 획득된 출력 신호에서 설정된 상한값 및 하한값을 이용하여 음성 구간의 시작 지점과 음성 구간의 종료 지점을 검출하는 단계를 포함하는 것을 특징으로 하는 음성 인식 방법.

청구항 3

제 2 항에 있어서, 상기 음성 구간의 종료 지점은

상기 출력 신호가 상기 상한값과 하한값 사이에 연속적으로 존재하는 회수에 근거하여 검출되는 것을 특징으로 하는 음성 인식 방법.

청구항 4

제 1 항에 있어서, 상기 주파수 영역의 신호 변환 단계는

상기 시간 영역의 오디오 입력 신호에 대해 2번의 고속 퓨리에 변환을 수행하는 단계를 포함하는 것을 특징으로 하는 음성 인식 방법.

청구항 5

제 1 항에 있어서, 상기 음성 부재 확률은

상기 변환된 주파수 영역의 신호 중 기 설정된 프레임에 대한 특징값을 이용하여 획득되는 것을 특징으로 하는 음성 인식 방법.

청구항 6

잡음과 음성 신호가 혼합된 시간 영역의 오디오 입력 신호를 수신하는 수신부;

상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변환하는 변환부;

음성 부재 확률에 대한 음성 존재 확률의 비를 이용하여 상기 오디오 입력 신호의 특징값을 추출하는 특징값 추출부-상기 음성 존재 확률 및 음성 부재 확률은 상기 변환된 주파수 영역의 신호를 이용하여 획득됨-;

상기 추출된 특징값에 원점 대칭 필터를 적용하여 음성 구간을 검출하는 음성 구간 검출부; 및

상기 수신부, 상기 변환부, 상기 특징값 추출부, 및 상기 음성 구간 검출부를 제어하는 제어부를 포함하는 것을 특징으로 하는 음성 인식 장치.

청구항 7

제 6 항에 있어서, 상기 음성 구간 검출부는

상기 추출된 오디오 신호의 특징값에 원점 대칭 필터를 적용하여 출력 신호를 획득하고, 상기 획득된 출력 신호

에서 설정된 상한값 및 하한값을 이용하여 음성 구간의 시작 지점과 음성 구간의 종료 지점을 검출하는 것을 특징으로 하는 음성 인식 장치.

청구항 8

제 7 항에 있어서, 상기 음성 구간의 종료 지점은

상기 출력 신호가 상기 상한값과 하한값 사이에 연속적으로 존재하는 회수에 근거하여 검출되는 것을 특징으로 하는 음성 인식 장치.

청구항 9

제 6 항에 있어서, 상기 변환부는

상기 시간 영역의 오디오 입력 신호에 대해 2번의 고속 퓨리에 변환을 수행하는 것을 특징으로 하는 음성 인식 장치.

청구항 10

제 6 항에 있어서, 상기 음성 부재 확률은

상기 변환된 주파수 영역의 신호 중 기 설정된 프레임에 대한 특징값을 이용하여 획득되는 것을 특징으로 하는 음성 인식 장치.

명 세 서

기 술 분 야

[0001] 본 발명은 음성 인식 방법 및 장치에 관한 것으로, 보다 상세하게는 잡음과 음성이 혼합된 환경에서 주파수 변환 및 원점 대칭 필터를 이용하여 음성 구간을 검출하는 음성 인식 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 일반적으로 음성 인식(Speech Recognition)이란 사람이 말하는 음성 언어를 컴퓨터가 해석해 처리하는 기술을 의미한다. 이러한 음성 인식 기술은 주변의 잡음이 적은 실험실, 가정, 또는 사무실 등에서 사용되면 성능이 우수하게 느껴지나, 주변의 잡음이 심한 도로, 복도, 전시장, 회의장 등에서 사용되는 경우에는 인식률이 현저히 떨어지게 된다. 이는 주변 잡음과 사람의 음성을 효과적으로 분리하지 못하기 때문이다.

[0003] 한편, 음성 인식에서의 음성구간 검출은 음성인식 성능에 큰 영향을 미치는 요소이다. 음성구간 검출을 통해 음성 구간만의 신호를 취함으로써, 음성인식에 소요되는 시간을 단축시킬 수 있으며 비음성 구간에 존재하는 잡음이 음성인식 성능을 하락시킬 수 있는 가능성을 줄일 수 있다. 그동안 다양한 음성구간 검출기들이 개발되었지만, 신호 대 잡음 비(Signal to Noise Ratio, SNR)가 -5dB 이하인 경우 대부분 좋은 성능을 보이지 않는다. 신호 대 잡음 비가 -5dB 이하의 잡음 환경에서는 사용자의 음성과 잡음이 함께 시스템의 마이크로 입력이 되어 정확한 음성 구간을 검출하기 어렵다.

[0004] 따라서 잡음이 많은 환경에서도 정밀하게 음성 구간을 검출할 수 있는 기술에 대한 연구가 필요한 실정이다.

발명의 내용

해결하려는 과제

[0005] 본 발명의 목적은 잡음이 많은 환경에서도 음성 구간을 정확하게 검출할 ㅅ수 있는 음성 인식 방법 및 장치를 제공하는 데 있다.

[0006] 본 발명의 목적은 주파수 영역의 신호를 이용하여 획득될 수 있는 음성 존재 확률과 음성 부재 확률의 비를 이용하여 음성 구간을 검출할 수 있는 음성 인식 방법 및 장치를 제공하는 데 있다.

과제의 해결 수단

[0007] 상기 목적을 달성하기 위해 본 발명의 일실시예에 의하면, 잡음과 음성 신호가 혼합된 시간 영역의 오디오 입력

신호를 수신하는 단계; 상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변환하는 단계; 음성 부재 확률에 대한 음성 존재 확률의 비를 이용하여 상기 오디오 입력 신호의 특징값을 추출하는 단계-상기 음성 존재 확률 및 음성 부재 확률은 상기 변환된 주파수 영역의 신호를 이용하여 획득됨-; 상기 특징값에 원점 대칭 필터를 적용하여 음성 구간을 검출하는 단계를 포함하는 음성 인식 방법이 제공된다.

[0008] 상기 목적을 달성하기 위해 본 발명의 일실시예에 의하면, 잡음과 음성 신호가 혼합된 시간 영역의 오디오 입력 신호를 수신하는 수신부; 상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변환하는 변환부; 음성 부재 확률에 대한 음성 존재 확률의 비를 이용하여 상기 오디오 입력 신호의 특징값을 추출하는 특징값 추출부-상기 음성 존재 확률 및 음성 부재 확률은 상기 변환된 주파수 영역의 신호를 이용하여 획득됨-; 상기 추출된 특징값에 원점 대칭 필터를 적용하여 음성 구간을 검출하는 음성 구간 검출부; 및 상기 수신부, 상기 변환부, 상기 특징값 추출부, 및 상기 음성 구간 검출부를 제어하는 제어부를 포함하는 음성 인식 장치가 제공된다.

발명의 효과

[0009] 본 발명의 일실시예에 의한 음성 인식 방법 및 장치는 주파수 영역의 신호를 이용하여 획득될 수 있는 음성 존재 확률과 음성 부재 확률의 비를 이용하여 오디오 입력 신호의 특징값을 추출함으로써, 음성 구간의 특징값을 잡은 구간의 특징값과 구별되게 검출할 수 있다.

[0010] 또한, 본 발명의 일실시예에 의하면, 음성 존재 확률과 음성 부재 확률의 비 및 원점 대칭 필터를 이용함으로써, 잡은 환경에서도 음성 구간을 정확하게 검출 할 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 발명의 일실시예와 관련된 음성 인식 장치의 블록도이다.

도 2는 본 발명의 일실시예와 관련된 음성 인식 방법의 흐름도이다.

도 3은 본 발명의 일실시예와 관련된 주파수 변환을 통해 얻어진 오디오 입력 신호의 패턴을 나타내는 그래프이다.

도 4는 본 발명의 일실시예와 관련된 원점 대칭 필터의 일예를 나타내는 그리프이다.

도 5는 본 발명의 일실시예와 관련된 음성 구간 검출을 위한 모델을 설명하기 위한 도면이다.

도 6은 본 발명의 일실시예와 관련된 음성 인식 방법을 통해 얻어진 음성 구간 검출 결과를 나타내는 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하, 본 발명의 일실시예와 관련된 음성 인식 방법 및 장치에에 대해 도면을 참조하여 설명하기로 하겠다.

[0013] 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0014] 도 1은 본 발명의 일실시예와 관련된 음성 인식 장치의 블록도이다.

[0015] 도시된 바와 같이, 음성 인식 장치(100)는 수신부(100), 변환부(120), 특징값 추출부(130), 음성 구간 검출부(140), 및 제어부(150)를 포함할 수 있다. 그러나, 도시된 구성요소 모두가 필수구성요소인 것은 아니다. 도시된 구성요소보다 많은 구성요소로 상기 음성 인식 장치(100)가 구성될 수도 있고, 그보다 적은 구성요소에 의해서도 음성 인식 장치(100)가 구성될 수 있다. 상기 음성 인식 장치(100)는 개인용 휴대 단말기, 음성 인식 TV, 대화형 로봇, 음성 기반의 자동차 인터페이스(예: 자동차 내비게이션), 컴퓨터 등 다양한 전자 제품에 적용될 수 있다.

[0016] 수신부(110)는 오디오 입력 신호를 수신할 수 있다. 상기 음성 검출을 위해 상기 음성 인식 장치로 입력되는 신호로, 잡음과 음성이 혼합된 신호일 수 있다. 예를 들어, 잡음 환경에서 사람의 음성이 상기 수신부(110)로 입력될 수 있다. 그리고 상기 오디오 입력 신호는 시간 영역의 신호일 수 있다. 시간 영역의 신호라 함은 시간 요

소가 변수로 표현되는 신호를 의미한다. 예를 들어, 오디오 입력 신호는 시간축의 함수($x(t)$)로 표현될 수 있다.

[0017] 변환부(120)는 상기 시간 영역의 오디오 입력 신호를 주파수 영역의 신호를 변환할 수 있다. 주파수 영역의 신호로의 변환을 위해 고속 퓨리에 변환(FFT, Fast Fourier Transform), 이산 코사인 변환(DCT, Discrete, Cosine Transform) 등의 변환 기법이 사용될 수 있다.

[0018] 특징값 추출부(130)는 주파수 영역으로 변환된 오디오 입력 신호를 이용하여 오디오 입력 신호의 특징값을 추출할 수 있다. 상기 특징값은 오디오 입력 신호의 정보를 잘 표현할 수 있도록 수치화한 값을 말한다.

[0019] 음성 구간 검출부(140)는 상기 추출된 오디오 입력 신호의 특징값에 소정의 필터를 적용하여 음성 구간을 검출할 수 있다. 상기 소정의 필터는 값이 거의 일정한 경우 상쇄시켜 0에 가까운 결과가 나오게 할 수 있는 원점 대칭 필터를 포함할 수 있다. 상기 원점 대칭 필터가 에지 필터의 역할을 수행할 수 있다. 에지 필터는 오디오 신호 내의 에지를 검출할 수 있다. 에지는 오디오 신호 내에서 특징값의 윤곽선에 해당하는 것으로 계조가 급격하게 변화하는 경계이다.

[0020] 또한, 음성 구간 검출(140)는 상기 특징값에 원점 대칭 필터를 적용한 후, 설정된 상한값과 하한값을 이용하여 음성 구간의 시작 지점과 종료 지점을 검출할 수 있다.

[0021] 제어부(150)는 상기 수신부(110), 상기 변환부(120), 상기 특징값 추출부(130), 및 음성 구간 검출부(140)를 전반적으로 제어할 수 있다.

[0022] 도 2는 본 발명의 일실시예와 관련된 음성 인식 방법의 흐름도이다.

[0023] 수신부(110)는 오디오 입력 신호를 수신할 수 있다(S210). 상기 오디오 입력 신호는 잡음과 음성이 혼합된 신호일 수 있다. 그리고 상기 오디오 입력 신호는 시간 영역의 신호이다. 따라서 오디오 입력 신호는 다음과 같이 시간의 함수 $x(t)$ 로 표현될 수 있다.

[0024] 주파수 변환부(120)는 시간의 함수로 표현된 오디오 입력 신호를 주파수 영역의 신호로 변환할 수 있다(S220). 시간 영역의 오디오 입력 신호를 주파수 영역의 신호로 변화는 것은 시간축의 신호는 과거 시간과 미래 시간 모두 무한개의 데이터가 존재하여, 그 신호를 분석하려면 어려움이 있기 때문이다.

[0025] 주파수 영역의 신호로의 변환을 위해 고속 퓨리에 변환(FFT, Fast Fourier Transform), 이산 코사인 변환(DCT, Discrete, Cosine Transform) 등의 변환 기법이 사용될 수 있다.

[0026] 이하, 실시예에서는 고속 퓨리에 변환(FFT, Fast Fourier Transform)을 이용하여 주파수 영역으로 변환하는 방법에 대해 설명하도록 하겠다.

[0027] 시간 축의 오디오 입력 신호 $x(t)$ 는 수학식 1과 같이 고속 퓨리에 변환(FFT, Fast Fourier Transform)을 이용하여 주파수 축의 신호로 변환될 수 있다.

수학식 1

[0028]
$$X(k') = \sum_{m=1}^{M} x(m)\, e^{-j\frac{2\pi}{N}k'm}$$

[0029] 여기서 k' 는 고속 퓨리에 변환 빈 인덱스를 나타내며, M은 고속 퓨리에 변환 사이즈를 나타낸다. k' 번째 인덱스에 대하여 수학식 1을 $X(k')^{1st}$ 라 하고, $|X(k')^{1st}|$ 는 k' 번째 주파수 인덱스에서의 절대값의 크기를 나타낸다. 그 다음 첫 번째 고속 퓨리에 변환을 통한 절대값 크기 (magnitude)를 시간축이라고 가정

하여 두 번째 고속 퓨리에 변환 $X(k)^{2nd}$ 을 수학식 2와 같이 구할 수 있다. 또한, j 는 복소수 및 N은 신호의 주기를 나타내다.

수학식 2

[0030] $$X(k)^{2nd} = \sum_{k=1}^{M} | X(k')^{1st} | e^{-j\frac{2\pi}{N}kk'}$$

[0031] 여기서 k 는 고속 퓨리에 변환 빈 인덱스를 나타내다. 고속 퓨리에 변환 사이즈는 256 샘플인 것을 예를 들어 설명하기로 한다. 이러한 고속 퓨리에 변환 과정은 신호의 기본주파수 (fundamental frequency)를 계산하여, 스펙트럼 절대값의 연속적인 정점을 가지는 하모닉 성분을 분석하는데 적합한 방법이다.

[0032] 도 3은 본 발명의 일실시예와 관련된 주파수 변환을 통해 얻어진 오디오 입력 신호의 패턴을 나타내는 그래프이다.

[0033] 도 3(a)는 음성에 잡음이 부가된 신호와 잡음 신호를 첫 번째 고속 퓨리에 변환을 한 절대값을 나타낸다. 도 3(a)에서 알 수 있듯이 음성에 잡음이 부가된 신호와 잡음 신호의 구별이 쉽지 않을 수 있다.

[0034] 그러나 도 3(b)에서 알 수 있듯이 도 3(a)의 결과에 또 한번의 고속 퓨리에 변환을 수행하면, 음성이 존재하는 구간에서 주기적으로 변동을 거듭하는 하모닉 성분 때문에 음성에 잡음이 부가된 신호와 잡음 신호의 구별이 잘 됨을 확인할 수 있다.

[0035] 즉, 본 발명의 일실시예에 의하면, 오디오 입력 신호를 주파수 영역의 신호로 변환할 때, 주파수 변환 시, 변환부(120)는 2번의 고속 퓨리에 변환을 수행을 통해 시간 영역의 신호를 주파수 영역의 신호로 변환할 수 있다.

[0036] 도시된 그래프에서 Noise signal은 잡음 신호를 의미하고, Noisy signal은 잡음과 음성이 혼합된 신호를 의미한다. 상기 Noise signal은 오디오 입력 신호에서 기 설정된 프레임(예: 처음 10 프레임)에 대한 평균으로 추정될 수 있다. 왜냐하면 오디오 입력 신호가 처음부터 음성이 존재할 확률이 낮기 때문이다.

[0037] 특징값 추출부(130)는 변환부(120)에서 획득한 주파수 영역의 신호를 이용하여 특징값을 추출할 수 있다(S230). 상기 특징값 추출부(130)는 음성 부재 확률에 음성 부재 확률의 비를 이용하여 오디오 입력 신호의 특징값을 추출할 수 있다. 본 실시예에서는 음성 신호와 잡음 신호가 독립 가우시안 확률 과정 (Independent Gaussian Random Process)을 따른다고 가정하기로 한다.

[0038] 이하에서 오디오 입력 신호의 특징값을 추출하는 방법에 대해 상세히 설명하기로 한다.

[0039] 두 번의 고속 퓨리에 변환을 통한 L 차원의 벡터는 수학식 3으로 나타낼 수 있다.

수학식 3

[0040] $$\lambda_Y(k) = \lambda_S(k) + \lambda_N(k)$$

[0041] $\lambda_Y(k)$ 는 $|X(k)^{2nd}|$ 를 의미하며, 음성에 잡음이 부가된 신호의 번째 주파수 인덱스의 절대값 크기를 나타낸다. $\lambda_S(k)$ 는 깨끗한 음성 신호의 절대값의 크기를 나타내고, $\lambda_N(k)$ 는 잡음 신호의 절대값 크기를 의미한다.

[0042] 현재의 입력 프레임이 음성 구간이라는 가설(Hypothesis)을 H_1, 비음성 구간이라는 가설을 H_0라고 하면, H_0

와 H_1는 수학식 4와 같이 표현될 수 있다.

수학식 4

$$H_0 : speech\,absent : Y = N$$

$$H_1 : speech\,present : Y = S + N$$

[0043]

[0044] H_0와 H_1 일 때의 관측값 Y가 생성될 조건부 확률 밀도 함수(Conditional Probability Density Function)는 수학식 5와 수학식 6으로 표현될 수 있다.

수학식 5

$$P(Y|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} exp(-\frac{|\lambda_Y(k)|}{\lambda_N(k)})$$

[0045]

수학식 6

$$P(Y|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi [\lambda_N(k) + \lambda_S(k)]} exp(-\frac{|\lambda_Y(k)|}{\lambda_N(k) + \lambda_S(k)})$$

[0046]

[0047] 즉, H_0는 음성이 부재할 가설이고, 그 때의 관측값 Y는 N(노이즈)이 된다. 따라서 수학식 5는 음성이 부재할 가설을 세워둔 상태에서 음성이 부재할 확률(편의상, '음성 부재 확률'이라 함)을 나타낸다. $P(Y|H_0)$ 값이 클수록 음성이 존재하지 않을 확률이 높게 된다.

[0048] 또한, H_1은 음성이 존재할 가설이고, 그 때의 관측값 Y는 S(음성)+N(노이즈)이 된다. 수학식 6은 음성이 존재할 가설을 세워둔 상태에서 음성이 존재할 확률(편의상, '음성 존재 확률'이라 함)을 나타낸다. $P(Y|H_1)$ 값이 크다는 것은 음성이 존재할 확률이 높다는 의미로 해석될 수 있다.

[0049] 본 발명의 일실시예에 의하면, 음성 존재 확률과 음성 부재 확률의 비를 이용하여 오디오 입력 신호의 특징값을 추출하여 음성 존재와 음성 부재를 판별할 수 있다.

[0050] 예를 들어, 특징값 추출부(130)는 수학식 7로 표현되는 음성 부재 확률에 대한 음성 존재 확률의 로그 우도 비를 이용하여 오디오 입력 신호의 특징값을 추출할 수 있다.

수학식 7

$$\Lambda = \frac{1}{L} log \frac{P(Y|H_1)}{P(Y|H_0)} = \frac{1}{L} \sum_{k=0}^{L-1} \left\{ \frac{\lambda_Y(k)}{\lambda_N(k)} - \log \frac{\lambda_Y(k)}{\lambda_N(k)} - 1 \right\}$$

[0051]

[0052] 잡음 구간(음성 부재 구간)에서는 오디오 입력 신호의 특징과 추정된 잡음(예: 오디오 입력 신호 중 처음 10프레임의 평균)의 특징의 패턴이 유사하여 로그 우도 비 갑이 적게 나올 것이고, 음성 구간에서는 음성 특징의 패턴의 변화로 인하여 로그 우도 비 값이 크게 나올 수 있다. 이와 같은 특징을 이용하여, 음성 부재 구간과 음성 구간을 구분할 수 있다.

[0053] 음성 구간 검출부(140)는 상기 특징 추출부(130)에서 추출한 특징값에 소정의 필터를 적용하여 음성 구간을 검출할 수 있다(S240). 상기 소정 필터의 예로 원점 대칭 필터가 포함될 수 있다. 상기 원점 대칭 필터는 에지 필터의 역할을 수행할 수 있다. 예를 들어, 상기 원점 대칭 필터는 수학식 8과 같이 표현될 수 있다.

수학식 8

$$f(x) = e^{Ax}[K_1 \sin(Ax) + K_2 \cos(x)] + e^{-Ax}[K_3 \sin(Ax) + K_4 \cos(x)] + K_5 + K_6 e^{sx}$$

$$h(i) = \begin{cases} -f(i), & -W \le i \le 0 \\ f(i), & 1 \le i \le W \end{cases}$$

[0054]

[0055] 도 4는 수학식 8과 표현되는 원점 대칭 필터를 나타낸다. 여기서 W는 필터 길이와 관계되는 변수이며, i는 -W부터 W까지의 정수, A, K는 필터 파라미터를 나타낸다. 도시된 필터는 W=7 일 때의 필터 응답 그래프를 나타낸다. 도시된 그래프에서 A=0.41, K_1=1.538, K_2=1.468, K_3=-0.078, K_4=-0.036, K_5=-0.872, K_6=-0.56이다.

[0056] 음성 구간 검출부(140)는 추출된 특징값에 원점 대칭 필터를 적용하여 출력 $F(n)$을 구할 수 있다. 상기 출력 $F(n)$은 수학식 9와 같이 나타낼 수 있다.

수학식 9

$$F(n) = \sum_{i=-W}^{W} h(i) \Lambda(n+i)$$

[0057]

[0058] 여기서 n은 프레임 인덱스를 나타낸다. 상기 수학식 9에 따라 출력 $F(n)$을 계산하면, 비음성 구간(음성 부재 구간)에서의 특징값이 그 크기에 상관없이 일정하면 상기 출력값이 0에 가까운 결과가 나오다가 음성이 존재하여 특징값이 커지면 출력 역시 커지게 된다. 상기 특징값이 작아지면 필터 출력 역시 작아지게 된다.

[0059] 본 발명의 일실시예에 의하면, 음성 구간 검출부(140)는 출력 $F(n)$의 상태 천이 동작을 통하여 음성 구간의 시작 지점과 종료 지점을 검출할 수 있다. 즉, 음성 구간 검출부(140)는 상태 천이 모델을 이용하여 음성 구간의 시작 지점과 종료 지점을 검출할 수 있다.

[0060] 도 5는 음성 구간의 시작 지점과 종료 지점 검출을 위한 모델의 일례를 나타낸다.

[0061] 도 5에서 Silence는 비음성 구간(음성 부재 구간)을 나타내고, In speech는 음성 구간을 나타낸다. Leaving

speech는 음성 구간이지만 비음성 구간으로 변할 수 있는 구간을 의미한다. 하한값 T_L(lower threshold), 상한값 T_U (upper threshold), Gap은 끝나는 점을 결정하기 위한 허용치로써 실험적으로 정하는 상수이다. 단, 상한값은 하한값 보다 크다고 가정한다.

[0062] 도 5에 도시된 상태 천이 모델을 이용하면, 음성 구간 검출부(140)는 $F(n)$이 T_U 보다 작으면 음성이 없는 비음성 구간(Silence)으로 판단하고, $F(n)$ 이 T_U 보다 커지면 음성 구간이 시작된 것으로 판단할 수 있다. 그리고 음성 구간 검출부(140)는 그 부분을 음성 구간(In speech)의 시작점으로 검출할 수 있다.

[0063] 한편, 음성 구간 검출부(140)는 $F(n)$ 이 T_L 보다 작아지면 아직 음성 구간이긴 하지만 비음성 구간으로 바뀔 가능성이 있는 구간(Leaving speech)으로 판단하고, Count를 0으로 간주한다. Count는 $F(n)$이 T_L과 T_U 사이에 연속적으로 존재하는 회수를 의미한다. 음성 구간 검출부(140)는 Count가 Gap 보다 작으면 Leaving speech로 판단하고, Count가 Gap 보다 크면 Silence로 판단한다. Silence로 판단되는 해당 프레임이 음성 구간의 종료 지점이 된다. 또한, 음성 구간 검출부(140)는 $F(n)$ 이 T_L 보다 작아지면 Count를 0으로 잡고 해당 프레임을 Leaving speech 단계로 유지할 수 있다. 그리고 $F(n)$ 이 T_U 보다 커지면 다시 In speech 구간으로 간주할 수 있다.

[0064] 도 5에 도시된 상태 천이 모델을 이용하면, 잡음의 크기에 따라 상한과 하한을 조절함으로써, 정밀한 음성 구간 검출이 가능하다.

[0065] 도 6은 본 발명의 일실시예와 관련된 음성 인식 방법을 통해 얻어진 음성 구간 검출 결과를 나타내는 그래프이다.

[0066] 도 6은 자동차 잡음 환경(자동차 안에서 잡음이 존재하는 환경)에서 오디오 입력 신호를 대상으로 로그 우도 비 및 원점 대칭 필터를 이용하여 음성 구간을 검출한 실제 결과를 나타낸다.

[0067] 도 6(a)는 상기 자동차 잡음 환경(잡음과 음성이 혼합된 환경)에서의 샘플 단위의 특징값을 나타낸다. 그리고 도 6(b)는 잡음이 없는 깨끗한 환경에의 음성 신호의 특징값을 나타낸다. 좀 더 구체적으로 도 6(a)는 도 6(b)와 같이 깨끗한 환경에의 음성 신호에 잡음 혼합한 오디오 입력 신호의 특징값을 나타낸다.

[0068] 그리고 도 6(c)는 수학식 7을 이용하여 오디오 입력 신호의 특징값을 추출한 그래프이다. 도시된 바와 같이, 도 6(c)는 로그 우도 비를 이용하여 오디오 입력 신호의 특징값을 추출하기 때문에, 비음성 구간에서는 0에 가까운 특징값을 가지다가 음성 구간에서 특징값이 크게 나오는 것을 확인할 수 있다.

[0069] 도 6(c)는 수작업을 통해 획득된 도 6(b)와 유사하게 음성 구간의 특징값이 나타나기에, 에지 필터를 적용하지 않더라고 음성 구간이 예측 가능하다.

[0070] 또한, 도 6(d)는 도 6(c)와 같은 오디오 입력 신호의 특징값에 수학식 8에 해당하는 원점 대칭 필터를 적용한 결과를 나타낸 그래프이다. 본 발명의 일실시예에 의한 음성 구간 검출은 도 6(d)를 통해 확인할 수 있듯이(도 6(d)에서 나타내는 음성 구간과 도 6(b)에서 나타내는 음성 구간이 유사함), 잡음 환경에서도 정확한 음성 구간 검출이 가능하다.

[0071] 상술한 음성 인식 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터로 판독 가능한 기록 매체에 기록될 수 있다. 이때, 컴퓨터로 판독 가능한 기록매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 한편, 기록매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0072] 컴퓨터로 판독 가능한 기록매체에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행

하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0073] 한편, 이러한 기록매체는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다.

[0074] 또한, 프로그램 명령에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0075] 상기와 같이 설명된 음성 인식 방법 및 장치는 상기 설명된 실시예들의 구성과 방법이 한정되게 적용될 수 있는 것이 아니라, 상기 실시예들은 다양한 변형이 이루어질 수 있도록 각 실시예들의 전부 또는 일부가 선택적으로 조합되어 구성될 수도 있다.

부호의 설명

[0076] 100: 음성 인식 장치

110: 수신부

120: 변환부

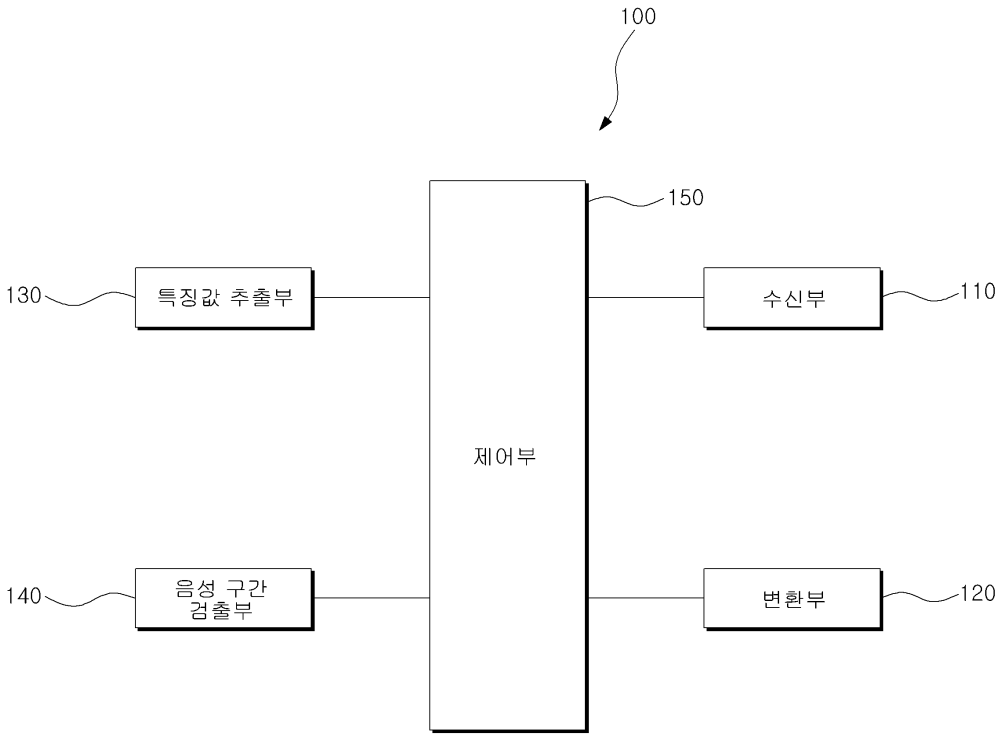
130: 특징값 추출부

140: 음성 구간 검출부

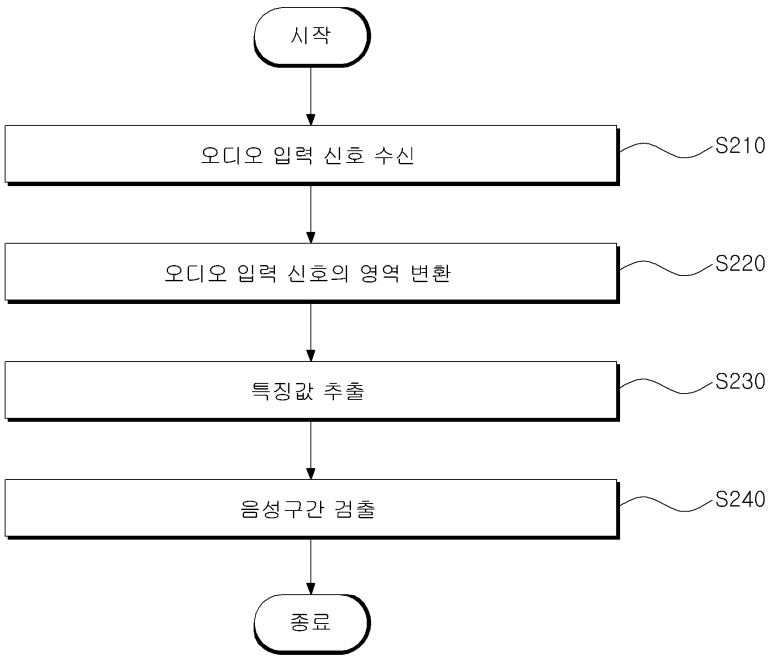
150: 제어부

도면

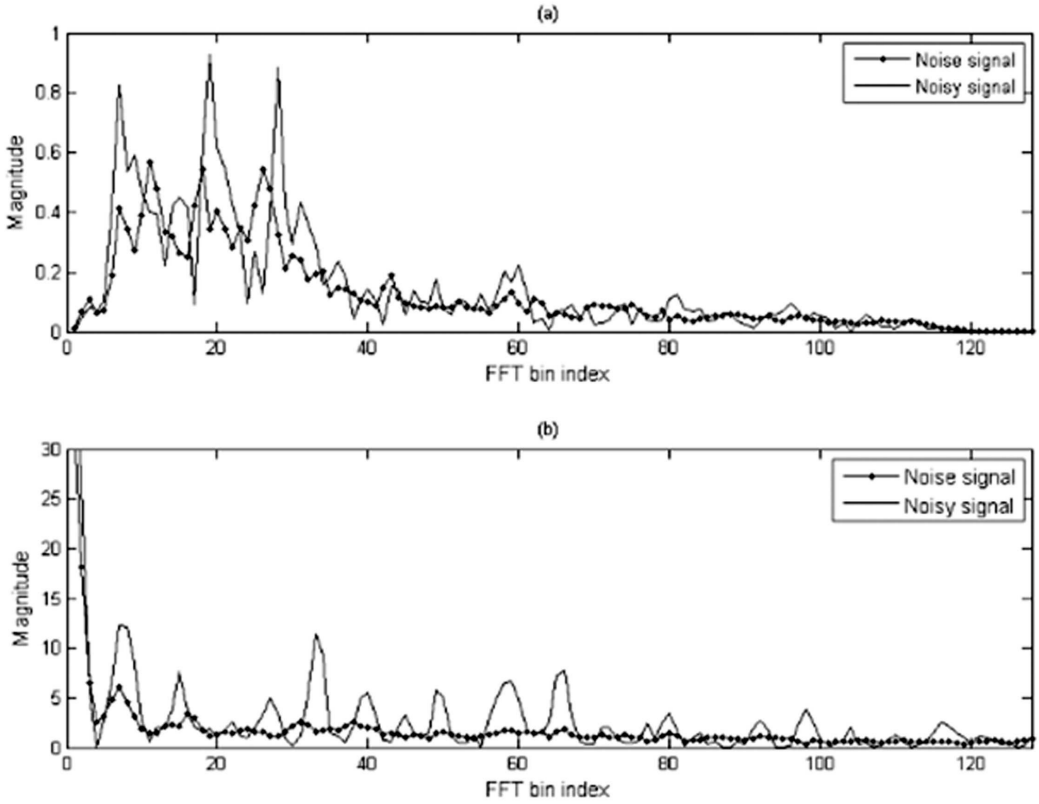
도면1



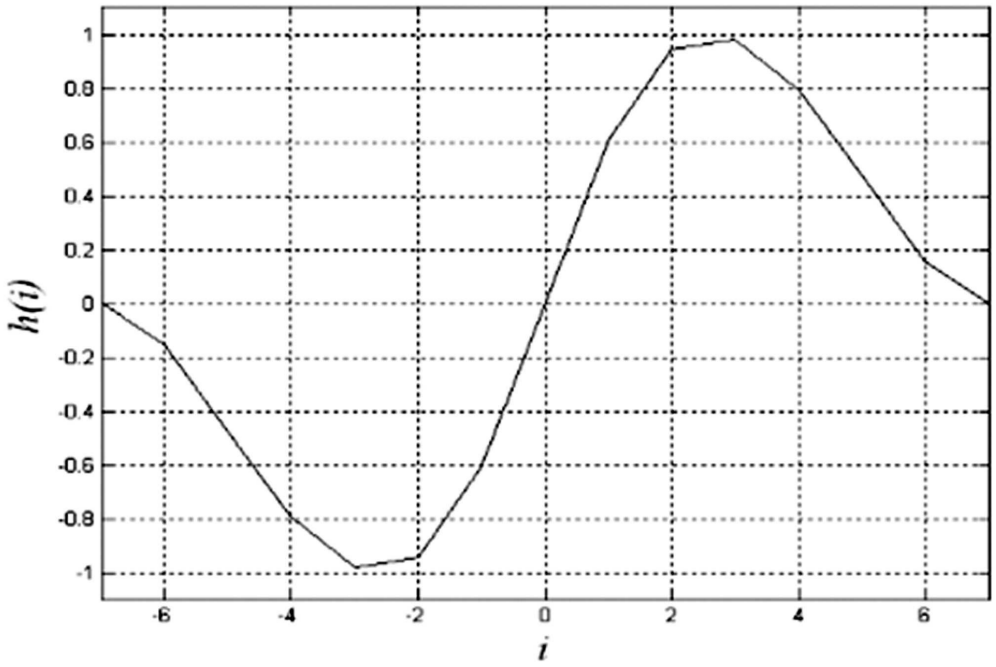
도면2


시작
오디오 입력 신호 수신
S210
오디오 입력 신호의 영역 변환
S220
특징값 추출
S230
음성구간 검출
S240
종료

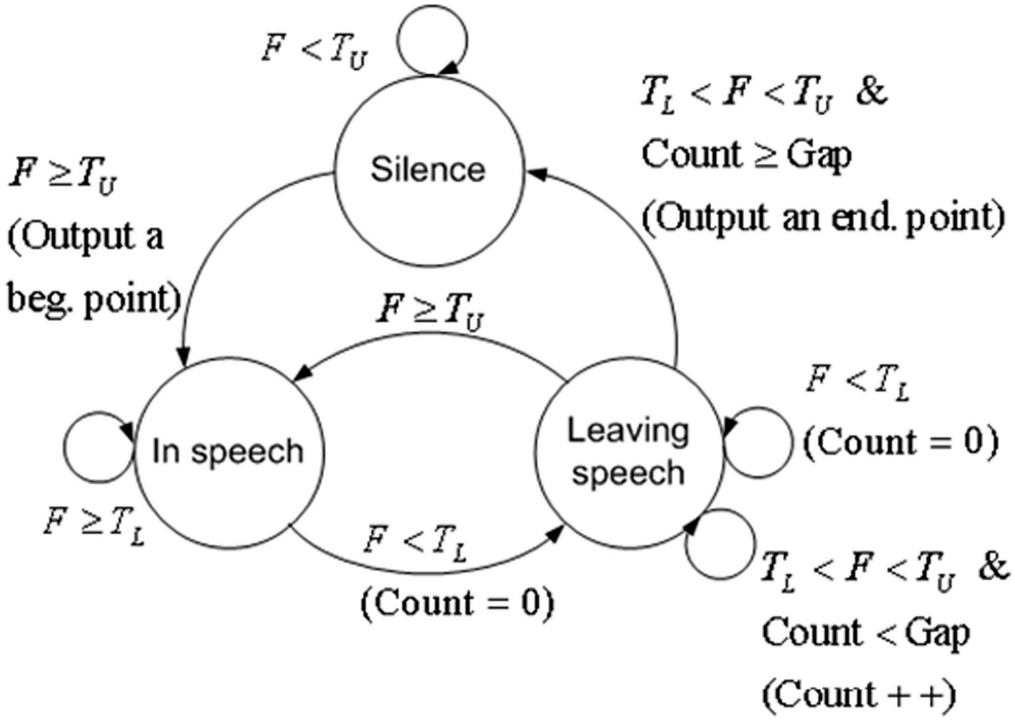
도면3


(a)
Noise signal
Noisy signal
Magnitude
FFT bin index
(b)
Noise signal
Noisy signal
Magnitude
FFT bin index

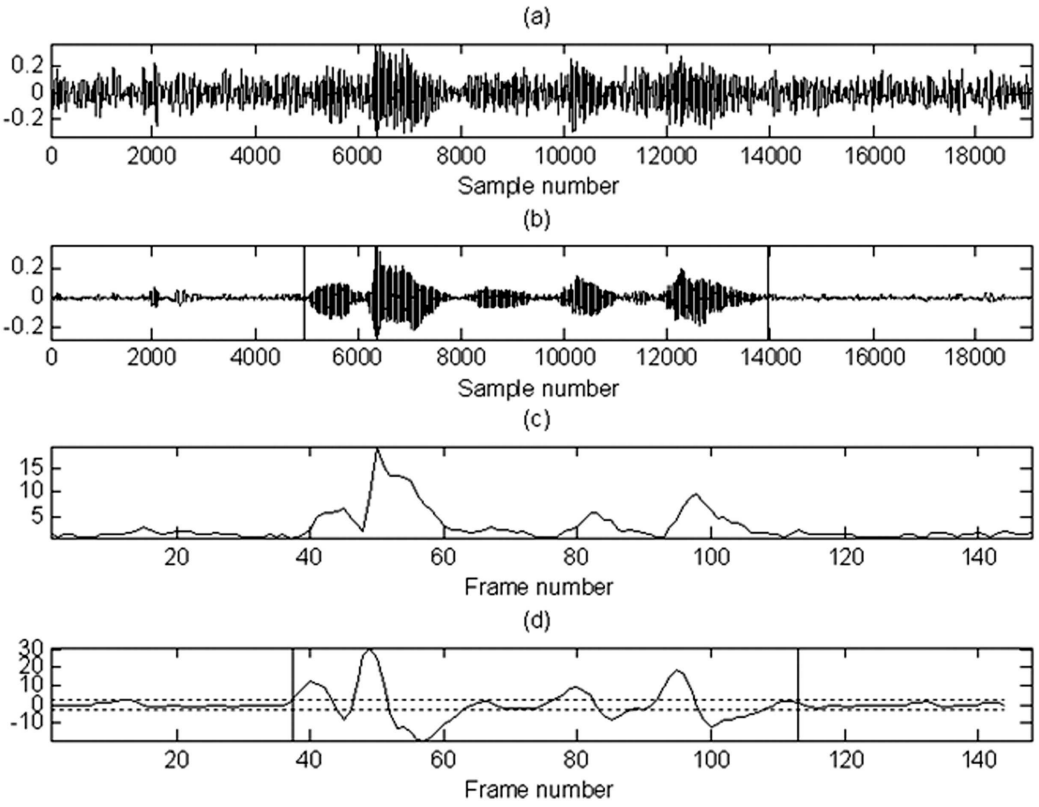
도면4


h(i)
i
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-6
-4
-2
0
2
4
6

도면5


$F < T_U$
Silence
$T_L < F < T_U$ &
Count ≥ Gap
(Output an end. point)
$F \geq T_U$
(Output a
beg. point)
$F \geq T_U$
In speech
Leaving
speech
$F < T_L$
(Count = 0)
$F \geq T_L$
$F < T_L$
(Count = 0)
$T_L < F < T_U$ &
Count < Gap
(Count + +)

도면6


(a)
0.2
0
-0.2
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
Sample number
(b)
0.2
0
-0.2
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
Sample number
(c)
15
10
5
20
40
60
80
100
120
140
Frame number
(d)
30
20
10
0
-10
20
40
60
80
100
120
140
Frame number