(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl.
G10L 11/02 (2006.01)
G10L 15/14 (2006.01)
G10L 11/06 (2006.01)

(11) 공개번호 10-2006-0134882
(43) 공개일자 2006년12월28일

(21) 출원번호 10-2006-0118920
(22) 출원일자 2006년11월29일
심사청구일자 2006년11월29일

(71) 출원인 인하대학교 산학협력단
인천 남구 용현동 253 인하대학교

(72) 발명자 장준혁
서울 서초구 서초동 1660-31 B동 301호

(74) 대리인 김건우

전체 청구항 수 : 총 1 항

(54) 음성 검출을 위한 통계 모델을 적응적으로 결정하는 방법

(57) 요약

본 발명은 음성 검출을 위한 통계 모델을 적응적으로 결정하는 방법에 관한 것으로서, 보다 구체적으로는 (1) 과거 m개의 이전 프레임에서의 잡음이 섞인 음성에 대한 DFT 계수들을 포함하는 벡터에 대하여, KS 테스트 통계량을 구하는 단계와, (2) 상기 구해진 KS 테스트 통계량을 이용하여 KS 테스트 통계량에 대한 평가치를 구하는 단계와, (3) 각각의 통계 모델에 대해 구해진 상기 평가치들을 비교하여, 평가치가 최소가 되는 통계 모델을 최적 통계 모델로 선택하는 단계와, (4) 상기 단계 (1) 내지 (3)를 일정 구간의 프레임마다 반복하여 해당 일정 구간의 프레임에 대한 최적 통계 모델을 적응적으로 선택하는 단계를 포함하는 방법에 관한 것이다.

본 발명은, 음성 검출을 위한 통계 모델을 잡음의 유형 및 SNR 조건에 따라서 적응적으로 결정함으로써, 기존의 가우시안 통계 모델을 사용하는 경우에 비해 음성 검출 성능을 향상시킬 수 있다.

대표도

도 4

특허청구의 범위

청구항 1.

입력된 음성 신호에 대하여 고속 푸리에 변환(FFT)을 수행하여 주파수 영역의 음성 신호로 변환시키는 단계와, 상기 주파수 영역으로 변환된 음성 신호로부터 잡음 신호의 파워를 추정하는 단계와, 상기 추정된 잡음 신호의 파워에 기초하여 음

성의 존재와 비존재에 대하여, 순수 음성과 잡음의 DFT 계수가 특정 통계 모델을 따른다고 가정함으로써 주파수 영역에 대한 우도비(Likelihood Ratio)를 결정하는 단계와, 상기 결정된 우도비로부터 음성 검출에 대한 결정 규칙(Decision Rule)을 계산하는 단계로 구성되는 음성 검출 방법(VAD; Voice Activity Detection)에 있어서, 상기 특정 통계 모델을 적응적으로 결정하는 방법으로서,

(1) X(t)가 프레임 t에서의 잡음이 섞인 음성에 대한 DFT 계수라고 하면, 과거 m개의 이전 프레임에서의 잡음이 섞인 음성에 대한 DFT 계수들을 포함하는 $\mathbf{X}_m = \{\mathbf{X}(t), \mathbf{X}(t-1), \ldots, \mathbf{X}(t-m)\}$ 에 대하여, 다음 수학식을 이용하여 KS 테스트 통계량 $T_t(X_m)$을 구하는 단계;

$$T(\mathbf{X}) = \max_i |F_X(X_i) - F(X_i)|$$

여기서, $F_x(X_i)$는 아래 수학식과 같이 정의되는 경험적인(empirical) 누적 분포 함수이며, $F(X_i)$는 특정 통계 모델에 대한 누적 분포 함수이다.

$$F_X(z) = \begin{cases} 0, & z < X_{(1)} \\ \frac{n}{N}, & X_{(n)} \le z < X_{(n+1)}, \quad n = 1, 2, \ldots, N-1 \\ 1, & z \ge X_{(N)} \end{cases}$$

여기서, $X_{(n)}, n = 1, \ldots, N$ 은 데이터 X의 순서 통계(order statistics)로서, X의 원소들을 분류하고 정렬하여 X의 가장 작은 원소인 $X_{(1)}$에서부터 X의 가장 큰 원소인 $X_{(N)}$까지 순서대로 배열한 것이다.

(2) 상기 구해진 KS 테스트 통계량 $T_t(X_m)$을 이용하여 다음 수학식에 의해 KS 테스트 통계량 $T_t(X_m)$에 대한 평가치를 구하는 단계;

$$\hat{T}_t = \lambda_T \hat{T}_{t-p} + (1 - \lambda_T) T_t(\mathbf{X}_m)$$

여기서, λ_T 는 평탄화 파라미터이다.

(3) 각각의 통계 모델에 대해 구해진 상기 평가치들을 비교하여, 평가치가 최소가 되는 통계 모델을 최적 통계 모델로 선택하는 단계; 및

(4) 상기 단계 (1) 내지 (3)를 일정 구간의 프레임마다 반복하여 해당 일정 구간의 프레임에 대한 최적 통계 모델을 적응적으로 선택하는 단계

를 포함하는 방법.

명세서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 음성 검출을 위한 통계 모델을 결정하는 방법에 관한 것으로서, 특히 음성 검출을 위한 통계 모델을 일정 프레임 구간마다 적응적으로 결정하는 방법에 관한 것이다.

음성 검출(VAD; Voice Activity Detection)은, 활성 음성(active speech)을 비음성으로부터 구분하는 고전적인 문제를 지칭하는 것으로, 음성 코딩, 음성 인식, 잡음이 섞인 음성의 향상, 핸즈프리 회의, 및 에코 제거와 같은 다양한 음성 통신

시스템에 대한 애플리케이션을 가지고 있다. 특히, VAD는, 통신 대역폭의 용량 및 커버리지를 향상시키는 가장 정밀한 방식 중 하나인 가변 음성 코딩의 주요한 부분이 되어 왔다. 이러한 이유로, 다양한 유형의 VAD 알고리즘이 제안되어 왔다. VAD에 대한 초기 알고리즘들은 대부분 선형 예측 코딩(LPC; linear prediction coding) 파라미터들, 에너지 레벨들, 포먼트 형상, 영 교차 비율(ZCR; zero crossing rate), 켑스트럴 특성(cepstral features), 음성 신호의 적응 모델링, 및 주기성 평가에 기초하고 있다. 보다 최근에는, 패턴 인식 및 LPC 잔여의 고차 큐뮬런트(higher order cumulants)에 기초한 VAD 접근법이 새로운 전략으로서 제안되어 왔다. 예를 들면, 특히 에너지 차이, ZCR, 및 스펙트럼 차이를 ITU-T G.729 Annex B에 적용시켜 왔다. 유사한 접근법을, 3세대 파트너십 프로젝트2(3GPP2; 3rd Generation Partnership Project2), ITU-T G.723.1 및 유럽 전화통신 표준 협회(European Telecommunications Standards Institutes) 적응형 다-비율(AMR; Adaptive Multi-Rate) VAD 옵션 2에서 선택 가능한 모드 보코더에 적용하는 시도도 있었다.

최근에, 일련의 가설이 적용된 우도비 테스트(LRT; Likelihood Ratio Test)로부터 구해지는 결정 규칙과 함께, 통계 모델을 이용함으로써 VAD의 성능을 개선하려는 노력이 이루어져 왔다. VAD 알고리즘에 기초한 통계 모델은 Ephraim과 Malah의 연구로부터 비롯되었다. 일부 연구자들은 가우시안 통계 모델을 VAD에 적용하는 알고리즘을 시도하였는데, 여기서 파라미터들의 평가가 DD(decision-directed) 접근법에 의해 수행되었다. 이 알고리즘은, 비록 관련 파라미터들의 최적화가 요구되기는 하였으나, 기존의 알고리즘들에 비해 높은 검출 정확도를 나타내었다. 통계 모델에 기초한 방법은 소프트 결정 스킴(soft decision scheme)과 결합됨으로써 더욱 개선되었다.

잘 알려진 바와 같이, DFT 도메인에서 주로 동작하는 기존의 VAD 알고리즘의 대부분은 순수 음성(clean speech)과 잡음의 스펙트럼이 가우시안 분포에 의해 특성지어진다고 가정한다. 그러나 최근의 연구 결과에 따르면, 순수 음성과 잡음의 DFT 계수들은 가우시안 분포가 아닌 감마 및 라플라시안 분포와 같은 다른 확률 밀도 함수(probability density function; 이하 'PDF'라 함)에 의해 보다 효과적으로 설명된다고 보고되고 있다. 따라서 음성 검출 성능을 향상시키기 위하여, 순수 음성과 잡음의 DFT 계수를 모델링할 새로운 분포 모델이 요구되고 있다.

발명이 이루고자 하는 기술적 과제

본 발명은 상기와 같은 문제 인식에서 비롯된 것으로서, 순수 음성과 잡음의 DFT 계수를 기존의 가우시안 통계 모델 대신에 잡음의 유형 및 SNR 조건에 따라서 적합한 통계 모델을 적응적으로 결정하는 방법을 제공하는 것을 그 목적으로 한다.

발명의 구성

상기한 목적을 달성하기 위한 본 발명의 특징에 따른, 음성 검출을 위한 통계 모델을 적응적으로 결정하는 방법은,

입력된 음성 신호에 대하여 고속 푸리에 변환(FFT)을 수행하여 주파수 영역의 음성 신호로 변환시키는 단계와, 상기 주파수 영역으로 변환된 음성 신호로부터 잡음 신호의 파워를 추정하는 단계와, 상기 추정된 잡음 신호의 파워에 기초하여 음성의 존재와 비존재에 대하여, 순수 음성과 잡음의 DFT 계수가 특정 통계 모델을 따른다고 가정함으로써 주파수 영역에 대한 우도비(Likelihood Ratio)를 결정하는 단계와, 상기 결정된 우도비로부터 음성 검출에 대한 결정 규칙(Decision Rule)을 계산하는 단계로 구성되는 음성 검출 방법(VAD; Voice Activity Detection)에 있어서, 상기 특정 통계 모델을 적응적으로 결정하는 방법으로서,

(1) X(t)가 프레임 t에서의 잡음이 섞인 음성에 대한 DFT 계수라고 하면, 과거 m개의 이전 프레임에서의 잡음이 섞인 음성에 대한 DFT 계수들을 포함하는 $\mathbf{X}_m = \{\mathbf{X}(t), \mathbf{X}(t-1), \ldots, \mathbf{X}(t-m)\}$ 에 대하여, 다음 수학식 1을 이용하여 KS 테스트 통계량 $T_t(X_m)$을 구하는 단계;

$$T(\mathbf{X}) = \max_i |F_X(X_i) - F(X_i)|$$

여기서, $F_x(X_i)$는 다음 수학식 2와 같이 정의되는 경험적인(empirical) 누적 분포 함수이며, $F(X_i)$는 특정 통계 모델에 대한 누적 분포 함수이다.

$$F_X(z) = \begin{cases} 0, & z < X_{(1)} \\ \frac{n}{N}, & X_{(n)} \le z < X_{(n+1)}, \quad n = 1, 2, \ldots, N-1 \\ 1, & z \ge X_{(N)} \end{cases}$$

여기서, $X_{(n)}, n = 1, \ldots, N$ 은 데이터 X의 순서 통계(order statistics)로서, X의 원소들을 분류하고 정렬하여 X의 가장 작은 원소인 $X_{(1)}$에서부터 X의 가장 큰 원소인 $X_{(N)}$까지 순서대로 배열한 것이다.

(2) 상기 구해진 KS 테스트 통계량 $T_t(X_m)$을 이용하여 다음 수학식 3에 의해 KS 테스트 통계량 $T_t(X_m)$에 대한 평가치를 구하는 단계;

$$\hat{T}_t = \lambda_T \hat{T}_{t-p} + (1 - \lambda_T) T_t(\mathbf{X}_m)$$

여기서, λ_T 는 평탄화 파라미터이다.

(3) 각각의 통계 모델에 대해 구해진 상기 평가치들을 비교하여, 평가치가 최소가 되는 통계 모델을 최적 통계 모델로 선택하는 단계; 및

(4) 상기 단계 (1) 내지 (3)를 일정 구간의 프레임마다 반복하여 해당 일정 구간의 프레임에 대한 최적 통계 모델을 적응적으로 선택하는 단계

를 포함하는 것을 그 특징으로 한다.

이하에서는 본 발명의 실시예를 도면을 참조하여 상세히 설명한다.

본 발명의 실시예는 다양한 잡음 조건에서 잡음에 오염된 음성신호의 DFT 계수를 VAD에 적용하기 위해 주어진 조건에 적합한 통계 모델을 적응적으로 결정하는 방법을 제안한다.

본 발명의 실시예에서는 먼저 다양한 잡음에 오염된 음성 조건하에서 적합도(Goodness Of Fit; 이하 'GOF'라 함) 테스트를 적용함으로써 가우시안 모델, 복소수 라플라시안 모델 및 복소수 감마 모델을 서로 비교한 후, 우도비 테스트 (Likelihood Ration Test; 이하 'LRT'라 함)에 기반한 결정 규칙(Decision Rule)을 고려한다.

1. 통계적 모델

먼저, 잡음 신호 N(t)와 음성 신호 S(t)의 합이 X(t)라고 가정하고, 이때 가설 H_0과 H_1이 각각 음성 신호가 없는 경우와 음성 신호가 있는 경우를 나타낸다고 하면, 가설 H_0과 H_1에서, X(t)는 각각 다음의 수학식 4 및 5를 만족한다.

$$H_0 : \text{speech absent} : \mathbf{X}(t) = \mathbf{N}(t)$$

$$H_1 : \text{speech present} : \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t)$$

여기서, $\mathbf{X}(t) = [X_0(t), X_1(t), \ldots, X_{M-1}(t)]^T$ 는 잡음이 섞인 음성의 DFT 계수, $\mathbf{N}(t) = [N_0(t), N_1(t), \ldots, N_{M-1}(t)]^T$ 는 잡음의 DFT 계수, $\mathbf{S}(t) = [S_0(t), S_1(t), \ldots, S_{M-1}(t)]^T$ 는 순수 음성의 DFT 계수이다.

상기한 통계적 모델은 적절한 DFT 계수 분포를 선택함으로써 완료된다. 본 발명의 실시예에서는 DFT 계수의 적절한 분포로서 기존에 흔히 사용되고 있는 가우시안 확률 밀도 함수를 채택하는 대신 다중 확률 밀도 함수 중 하나를 잡음의 유형 및 SNR 조건에 따라서 적응적으로 선택한다.

(1) 가우시안 통계 모델

가우시안 PDF에서, 양 가설(H0, H1)에 의해 결정되는 잡음 스펙트럼 성분(noisy spectral component)의 분포는 다음의 수학식 6 및 7과 같이 주어진다.

$$p_G(X_k|H_0) = \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}}\right\}$$

$$p_G(X_k|H_1) = \frac{1}{\pi[\lambda_{n,k}+\lambda_{s,k}]} \exp\left\{-\frac{|X_k|^2}{\lambda_{n,k}+\lambda_{s,k}}\right\}$$

여기서, $\lambda_{n,k}$와 $\lambda_{s,k}$는 각각 잡음 N_k 및 순수 음성 S_k의 분산(variance)을 나타낸다.

(2) 복소수 라플라시안 통계 모델

다음으로, 복소수 라플라시안 PDF에서, DFT 계수 X_k의 실수부(real part)와 허수부(imaginary part)를 각각 $X_{k(R)}$와 $X_{k(I)}$로 가정하면, $X_{k(R)}$와 $X_{k(I)}$는 다음의 수학식 8 및 9와 같이 분포된다.

$$p_L(X_{k(R)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(R)}|}{\sigma_x}\right\}$$

$$p_L(X_{k(I)}) = \frac{1}{\sigma_x} \exp\left\{-\frac{2|X_{k(I)}|}{\sigma_x}\right\}$$

여기서, σ_x^2는 X_k의 분산을 나타낸다. 만약 X_k의 실수부 및 허수부가 서로 독립적이라고 더 가정하면, X_k의 분포는 다음 수학식 10과 같이 될 수 있다.

$$\begin{aligned} p_L(X_k) &= p_L(X_{k(R)}) \cdot p_L(X_{k(I)}) \\ &= \frac{1}{\sigma_x^2} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sigma_x}\right\} \end{aligned}$$

상기 수학식 10으로부터, 양 가설(H0, H1)에 의해 결정되는 DFT 계수의 분포는 다음 수학식 11 및 12와 같이 주어질 수 있다.

$$p_L(X_k|H_0) = \frac{1}{\lambda_{n,k}} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k}}}\right\}$$

$$p_L(X_k|H_1) = \frac{1}{\lambda_{n,k}+\lambda_{s,k}} \exp\left\{-\frac{2(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{\lambda_{n,k}+\lambda_{s,k}}}\right\}$$

(3) 복소수 감마 통계 모델

마지막으로 복소수 감마 PDF에 대해 살펴본다. 각각의 DFT 계수의 실수부와 허수부 $X_{k(R)}$와 $X_{k(I)}$는 각각 다음의 수학식 13 및 14에 따라 분포된다.

$$p_M(X_{k(R)}) = \left(\frac{\sqrt{6}}{8\pi\sigma_x|X_{k(R)}|}\right)^{0.5} \exp\left\{-\frac{\sqrt{3}|X_{k(R)}|}{\sqrt{2}\sigma_x}\right\}$$

$$p_M(X_{k(I)}) = \left(\frac{\sqrt{6}}{8\pi\sigma_x |X_{k(I)}|}\right)^{0.5} \exp\left\{-\frac{\sqrt{3}|X_{k(I)}|}{\sqrt{2}\sigma_x}\right\}$$

만약 실수부와 허수부가 라플라시안 경우에서처럼 서로 독립적이라고 더 가정하면, DFT 계수 X_k의 분포는 다음 수학식 15와 같이 주어진다.

$$p_M(X_k) = \frac{\sqrt{6}}{8\pi\sigma_x |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \exp\left\{-\frac{\sqrt{3}(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sigma_x}\right\}$$

수학식 15를 2개의 가설 H_0과 H_1에 적용하면, DFT 계수의 분포는 다음 수학식 16 및 17과 같이 얻어진다.

$$p_x M(X_k|H_0) = \frac{\sqrt{6}}{8\pi\sqrt{\lambda_{n,k}} |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \exp\left\{-\frac{\sqrt{3}(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sqrt{\lambda_{n,k}}}\right\}$$

$$p_M(X_k|H_1) = \frac{\sqrt{6}}{8\pi\sqrt{\lambda_{n,k}+\lambda_{s,k}} |X_{k(R)}|^{0.5} |X_{k(I)}|^{0.5}} \exp\left\{-\frac{\sqrt{3}(|X_{k(R)}| + |X_{k(I)}|)}{\sqrt{2}\sqrt{\lambda_{n,k}+\lambda_{s,k}}}\right\}$$

성공적인 VAD 동작을 위해서, 본 발명의 실시예에서는 H_0과 H_1에 의해 결정되는 잡음 스펙트럼 요소를 위한 통계적 적합성 테스트(statistical fitting test)를 수행한다. 확률 밀도 함수(PDF)의 선택을 위해, 본 발명의 실시예에서는 적합도 테스트(GOF test; Goodness-of-Fit test)로서 잘 알려진 Kolomogorov-Sriminov 테스트(이하, 'KS 테스트'라 함)를 적용한다. KS 테스트를 적용함으로써, 각 통계적인 가정에 대한 신뢰성 있는 관측을 보장할 수 있다. KS 테스트는 경험적인(empirical) 누적 분포 함수(cumulative distribution function; 이하 'CDF'라 함) F_x를 소정의 분포 함수 F와 비교한다. $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ 가 잡음이 섞인 음성의 DFT 계수를 나타내는 벡터라고 가정할 때, 앞서 언급한 수학식 2로서 나타낼 수 있는 함수를 경험적 CDF로서 사용하였다(참조를 위하여 수학식 2를 다시 기재함).

<수학식 2>

$$F_X(z) = \begin{cases} 0, & z < X_{(1)} \\ \frac{n}{N}, & X_{(n)} \le z < X_{(n+1)}, \quad n = 1, 2, \ldots, N-1 \\ 1, & z \ge X_{(N)} \end{cases}$$

여기서, $X_{(n)}, n = 1, \ldots, N$ 은 데이터 X의 순서 통계(order statistics)이다. 순서 통계를 계산하기 위해, X의 원소들을 분류하고 정렬하여, X의 가장 작은 원소인 $X_{(1)}$에서부터 X의 가장 큰 원소인 $X_{(N)}$까지 순서대로 배열한다.

잡음 환경을 모의 실험하기 위해, 서로 다른 4명의 남성과 여성 화자로부터, 64초 구간의 음성 자료(speech material)를 수집하였으며, NOISEX-92 데이터베이스로부터 추출한 백색 잡음(white noise), 차량 잡음(vehicular noise), 및 배블 잡음(babble noise)을 신호 대 잡음비(SNR)가 10dB인 순수 음성 신호에 첨가하였다. 수집한 상기 데이터를 사용하여, 표본 평균과 분산을 계산하였으며, 주어진 가우시안 분포, 라플라시안 분포 및 감마 분포에 적용하였다.

경험적인 CDF와 주어진 분포 사이의 거리 측정을 규정하기 위해, 본 발명의 실시예에서는 KS 테스트 통계량(statistic)을 사용한다. KS 테스트 통계량(T)은 앞서 언급한 수학식 1과 같이 정의된다(참조를 위하여 수학식 1을 다시 기재함).

<수학식 1>

$$T(\mathbf{X}) = \max_i |F_X(X_i) - F(X_i)|$$

여기서, 표본 포인트Xi에서 구한 $F_x(X_i)$와 $F(X_i)$ 사이의 최대 차이가 거리에 해당한다. 몇 가지 분포에 대하여 데이터를 테스트하는 경우, 최소의 KS 통계량을 나타내는 분포가 주어진 데이터에 가장 적합한 것으로 고려될 수 있다.

도 1 내지 도 3은 각각 다양한 잡음 조건 하에서 평가된 KS 통계량의 결과를 도시한다. 결과를 요약하면, 다음과 같은 사항들을 발견할 수 있다.

(1) 도 1은 백색 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 1(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 1(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면이다. 도 1에 따르면, 라플라시안 모델이 모든 SNR 레벨에서 H_0에 가장 적합하며, 가우시안 모델이 다음으로 적합하다. 백색 잡음이 가우시안 분포되어지는지 여부가 의문일 수 있다. 이에 대해, 프레임 길이가 10ms로 제한되기 때문에, 백색 잡음에 대한 DFT 계수가 나머지 모델들보다 라플라시안에 의해 보다 잘 근사화되어진다는 것을 확인할 수 있다. 큰 프레임 길이(>200ms)에 대해서는, DFT 계수의 PDF는 가우시안으로 접근할 수 있으나, 이러한 프레임 길이는 (VAD를 포함하는) 실시간 신호 처리 애플리케이션에 적합하지 않다.

가설 H_1에 대해서도, 라플라시안 모델이 SNR = 20 dB를 제외하고는 가장 적합했다. 특히, 감마 PDF로부터 계산된 KS 통계치는 SNR이 증가할수록 작아지는 반면에, 다른 모델들은 보다 큰 값들을 생성했다. 마지막으로, 감마 모델의 KS 통계치는 SNR = 20 dB에서 가장 최소였다. 이러한 이유로, H_1 하에서 백색 잡음이 섞인 음성이 순수 음성 스펙트럼에 접근할수록 감마 모델이 점점 적합해진다는 결론을 얻을 수 있다.

2) 도 2는 차량 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 2(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 2(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면이다. 도 2로부터, H_0이 참일 경우 SNR = 20 dB를 제외하고는 모든 경우에서 감마 모델의 KS 통계량이 라플라시안 및 가우시안 모델들의 KS 통계량보다 더 작다는 것을 확인할 수 있다. 차량 잡음에 대한 거의 모든 경우에서, 가우시안 분포 가정은 잡음이 섞인 음성의 DFT 스펙트럼을 특징짓는데 좋지 못한 선택인 것으로 관찰되었다.

3) 도 3은 배블 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 3(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 3(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면이다. 도 3에서는, 가설 H_1 하에서 얻어진 결과가 가설 H_0 하에서 생성된 결과와 상당히 다른 경향을 나타내고 있다. H_0에 대해서는, 라플라시안 모델의 KS 통계량이 거의 모든 SNR 값들에서 최소였으며, 감마 모델의 KS 통계량이 항상 최대였다. 반면에, H_1이 참일 경우에는 감마 모델이 잡음이 섞인 음성의 DFT 계수의 분포에 가장 적합한 것으로 관찰되었다. 또한, 가우시안 모델의 성능은 거의 모든 조건에서 가장 나쁜 것으로 관찰되었다.

KS 테스트 결과를 요약하면, 잡음이 섞인 음성의 DFT 스펙트럼에 대하여 잡음의 유형 및 SNR 조건에 따라 서로 다른 통계 모델을 적용하는 것이 유리할 수 있다는 것을 알 수 있다.

2. LRT에 기초한 결정 규칙

본 발명의 실시예는 가정된 통계적 모델에 기초하여, k번째 주파수 빈(bin)에 대한 우도비(likelihood ratio; 이하 'LR'이라 함)를 다음 수학식 18과 같이 정의하였다.

$$\Lambda_k \equiv \frac{p(X_k|H_1)}{p(X_k|H_0)}$$

VAD에 대한 결정 규칙은 개별 주파수 빈에 대하여 계산된 LR들의 기하 평균으로서 다음 수학식 19와 같이 구해질 수 있다.

$$\log \Lambda = \frac{1}{M}\sum_{k=0}^{M-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

여기서, M은 주파수 빈의 전체 수를 나타내며, η은 검출 임계치를 나타낸다.

H_0과 H_1 모두에 대하여 가우시안 PDF가 음성 분포 모델링에 대하여 사용되는 경우에, LR은 다음 수학식 20에 의해 주어진다.

$$\Lambda_k^G = \frac{p_G(X_k|H_1)}{p_G(X_k|H_0)} = \frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\}$$

여기서, $\xi_k = \lambda_{s,k}/\lambda_{n,k}$ 이고, $\gamma_k = |X_k|^2/\lambda_{n,k}$ 이며, 이들은 각각 선행 SNR(a priori SNR) 및 후행 SNR(a posteriori SNR)로 지칭된다.

한편, 라플라시안 PDF의 경우에는, LR이 다음 수학식 21과 같이 구해진다.

$$\Lambda_k^L = \frac{p_L(X_k|H_1)}{p_L(X_k|H_0)} = \frac{1}{1+\xi_k}\exp\left\{2\left(|X_{k(R)}|+|X_{k(I)}|\right)\left(\frac{\sqrt{\lambda_{n,k}+\lambda_{s,k}}-\sqrt{\lambda_{n,k}}}{\sqrt{\lambda_{n,k}+\lambda_{s,k}}\sqrt{\lambda_{n,k}}}\right)\right\}$$

실질적으로, 수학식 21을 직접 적용하는 대신에, $\sqrt{\lambda_{n,k}+\lambda_{s,k}}$ 를 순간 진폭 스펙트럼 $|X_k|$로 대체하는 것이 보다 유리하다고 알려져 있다. 이 접근법은 순간 형태(instantaneous form)로서 지칭된다.

LRT의 성능을 보다 향상시키기 위하여, LR을 평탄화된 형태(smoothed form)로 지칭될 수 있는 다음 수학식 22와 같은 형태로 수정하는 대안적인 방법을 제안한다.

$$\hat{\Lambda}_k^L = \frac{1}{1+\xi_k}\exp\left\{2\left(|X_{k(R)}|+|X_{k(I)}|\right)\left(\frac{|\hat{X}_k|-\sqrt{\lambda_{n,k}}}{|\hat{X}_k|\sqrt{\lambda_{n,k}}}\right)\right\}$$

여기서, $|\hat{X}_k| = \kappa_L|X_k| + (1-\kappa_L)\sqrt{\lambda_{n,k}+\lambda_{s,k}}$ 이며, κ_L 은 평탄화 파라미터(smoothing parameter)이다.

평탄화된 형태에서는, 진폭 스펙트럼에 대하여 보다 평탄한 평가치가 순간 값 대신에 적용된다. 다수의 실험을 통하여, 평탄화된 형태가 순간 형태보다 우수하며, 특히 LRT의 성능은 평탄화 파라미터 κ_L 의 선택에 크게 의존하는데, κ_L 의 값으로는 0.9를 선택하는 것이 바람직한 결과를 준다는 것을 관찰하였다.

마지막으로, 감마 PDF에 대한 LR의 공식은 다음의 수학식 23과 같이 주어진다.

$$\Lambda_k^M = \frac{p_M(X_k|H_1)}{p_M(X_k|H_0)} = \frac{1}{\sqrt{1+\xi_k}}\exp\left\{\frac{\sqrt{2}}{\sqrt{3}}\left(|X_{k(R)}|+|X_{k(I)}|\right)\cdot\left(\frac{\sqrt{\lambda_{n,k}+\lambda_{s,k}}-\sqrt{\lambda_{n,k}}}{\sqrt{\lambda_{n,k}+\lambda_{s,k}}\sqrt{\lambda_{n,k}}}\right)\right\}$$

라플라시안 모델의 경우와 유사하게, 상기 수학식 23을 대응하는 평탄화된 형태인 다음의 수학식 24와 같이 변환할 수 있다.

$$\hat{\Lambda}_k^M = \frac{1}{\sqrt{1+\xi_k}} \exp\left\{ \frac{\sqrt{2}}{\sqrt{3}} \left(|X_{k(R)}| + |X_{k(I)}| \right) \left(\frac{|\hat{X}_k| - \sqrt{\lambda_{n,k}}}{|\hat{X}_k| \sqrt{\lambda_{n,k}}} \right) \right\}$$

여기서, $|\hat{X}_k| = \kappa_M |X_k| + (1-\kappa_M)\sqrt{\lambda_{n,k} + \lambda_{s,k}}$ 이며, κ_M 은 감마 모델에 대한 평탄화 파라미터로서, 다수의 실험을 통하여 그 값을 0.9로 선택하는 것이 바람직한 결과를 주는 것으로 관찰되었다.

3. 온라인 KS 테스트를 사용하는 다중 통계 모델에 기초한 VAD

앞서 살펴본 KS 테스트 통계량 결과에 따르면, 잡음의 유형, SNR 조건, 및 서로 다른 가설 조건에 따라서 적합한 통계 모델이 변하는 것을 확인할 수 있었다. 그러나 가우시안 통계 모델에 기초한 기존의 VAD 알고리즘에서는 VAD의 처리 전에 미리 통계 모델이 선택되었기 때문에, 보다 적극적으로 VAD의 성능 향상을 도모할 수 없었다. 이러한 문제점을 극복하기 위하여, 본 발명에서는 잡음의 유형, SNR 조건, 및 가설 조건에 따라서 적합한 통계 모델을 VAD 처리 시 적응적으로 결정하는 방법을 제안한다. 제안하는 방법은, 비정적 방식으로 시간이 흘러감에 따라 더해진 잡음의 통계적 특성 또는 SNR이 변하는 입력 신호를 다룰 때 특히 유리하다. 이를 위해서, 추가적인 계산상의 부하를 허용 가능한 레벨 아래로 유지하면서 주어진 조건에 가장 적합한 모델을 추적하기 위한 강인한 방법이 필요하다.

본 발명의 일 실시예에서는, 매 p 프레임마다 최근 입력 데이터 $\mathbf{X}_m = \{\mathbf{X}(l), \mathbf{X}(l-1), \ldots, \mathbf{X}(l-m)\}$ (과거 m개의 이전 프레임을 포함)을 이용하여, 앞서 언급한 수학식 3과 같이 주어지는 온라인 KS 평가치 T를 평가함으로써 주어진 조건에 가장 적합한 모델을 적응적으로 선택한다(참조를 위하여 수학식 3을 다시 기재함).

<수학식 3>

$$\hat{T}_l = \lambda_T \hat{T}_{l-p} + (1-\lambda_T) T_l(\mathbf{X}_m)$$

여기서, λ_T 는 평탄화 파라미터이다.

도 4는 잡음이 섞인 음성 샘플과 관련하여 주어진 통계 모델들에 대하여 온라인 KS 통계량 $\hat{T}_l$ 을 도시한다. 보다 구체적으로는, 도 4(a)는 도 4(b)와 같이 주어지는 음성 샘플에 대하여 각각의 통계 모델에 기초한 온라인 KS 통계량 $\hat{T}_l$ 을 도시하고 있다. 도 4로부터, 제안된 온라인 KS 테스트 통계량이 실제의 통계 특성을 효과적으로 따라갈 수 있다는 것을 확인할 수 있다. 특히, 도 4에서 도시된 배블 잡음의 경우에, 음성 세그먼트 동안 온라인 KS 통계량에 의해 라플라시안 모델이 감마 모델 및 가우시안 모델보다 우수한 성능을 보이는 것을 확인할 수 있다.

4. 실험 결과

우선, 다양한 잡음 조건 하에서 검출 성능을 조사함으로써, LRT를 적용하기 위한 각각의 파라미터 모델에 대한 임계치의 집합을 선택하였다. LRT의 임계치를 선택된 값으로 고정시킨 후, 본 발명에서 제안된 VAD 알고리즘의 성능을 평가하였으며, 그 결과를 ITU-T G.729B 알고리즘의 성능 및 가우시안 모델, 라플라시안 모델, 감마 모델 각각에 기초한 알고리즘의 성능과 비교하였다.

성능의 비교 대상으로서는, 거짓 경고 확률(순수 음성이 아닌 데 순수 음성으로 잘못 검출하는 확률) 및 순수 음성을 놓칠 확률(순수 음성인데 순수 음성으로 검출하지 못하는 확률)의 합인 음성 검출 오차 확률(P_E)을 평가하였다. 도 5 내지 도 7은 각각 백색, 자동차 및 배블 잡음 하에서의 검출 결과를 도시한다. 도 5 내지 도 7로부터, 다음과 같은 사항들을 관찰할 수 있다.

1) 가우시안, 라플라시안, 감마 중 어느 하나의 통계 모델에 기초한 VAD 알고리즘의 경우에, 거의 모든 평가된 조건 하에서 G.729B보다 더 나은 성능을 보였다.

2) 다양한 조건에서의 결과에 따르면, 나머지 통계 모델에 기초한 VAD 알고리즘에 비하여 라플라시안에 기초한 VAD 알고리즘이 우수한 성능을 나타내었다.

3) 본 발명에 따른 MSM에 기초한 VAD 알고리즘은 모든 테스트 조건에서 라플라시안에 기초한 VAD 알고리즘보다 우수하거나 대등한 성능을 나타내었다.

결론적으로, MSM이 VAD에 대하여 잡음이 섞인 음성의 DFT 계수들에 대한 통계 분포를 특징짓기 위한 보다 바람직한 방식이라는 점을 결과로부터 명백하게 확인할 수 있다.

이상 설명한 본 발명은 본 발명이 속한 기술분야에서 통상의 지식을 가진 자에 의하여 다양한 변형이나 응용이 가능하며, 본 발명에 따른 기술적 사상의 범위는 아래의 특허청구범위에 의하여 정해져야 할 것이다.

발명의 효과

본 발명은, 기존의 가우시안 통계 모델 대신에 잡음의 유형, SNR 조건, 가설 조건에 따라서 다중 통계 모델 중 어느 하나를 VAD 처리 중 적응적으로 선택하는 음성 검출 방법을 제안함으로써, 음성 검출의 성능을 향상시킬 수 있다.

도면의 간단한 설명

도 1은 백색 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 1(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 1(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면.

도 2는 차량 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 2(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 2(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면.

도 3은 배블 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 KS 통계량을 나타내는 도면으로서, 도 3(a) 및 (b)는 각각 가설 H_0인 경우의 DFT 계수의 실수부와 허수부를, 도 3(c) 및 (d)는 각각 가설 H_1인 경우의 DFT 계수의 실수부와 허수부를 나타내는 도면.

도 4는 잡음이 섞인 음성 샘플과 관련하여 주어진 통계 모델들에 대하여 온라인 KS 통계량  을 도시하는 도면.

도 5는 백색 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 음성 검출 오차 확률을 나타내는 도면.

도 6은 차량 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 음성 검출 오차 확률을 나타내는 도면.

도 7은 배블 잡음 조건 하에서, SNR = 5, 10, 15, 20 dB에서 평가된 각각의 모델에 대한 음성 검출 오차 확률을 나타내는 도면.

<도면 중 주요 부분에 대한 부호의 설명>

Gaussian : 가우시안 통계 모델

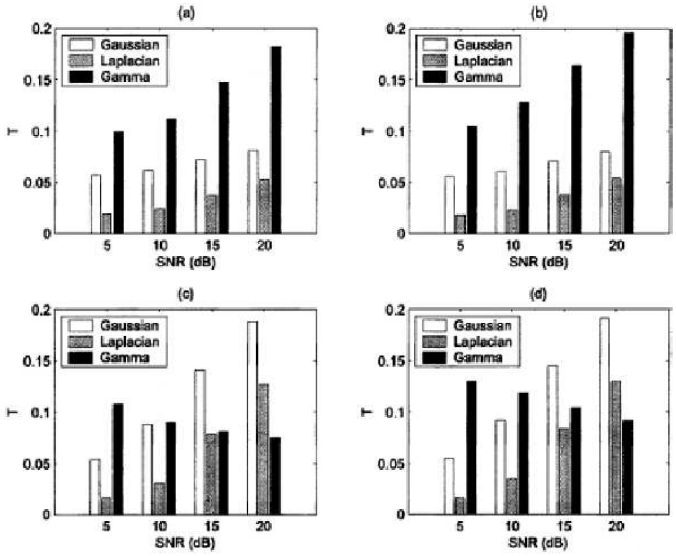
Laplacian : 복소수 라플라시안 통계 모델

Gamma : 복소수 감마 통계 모델

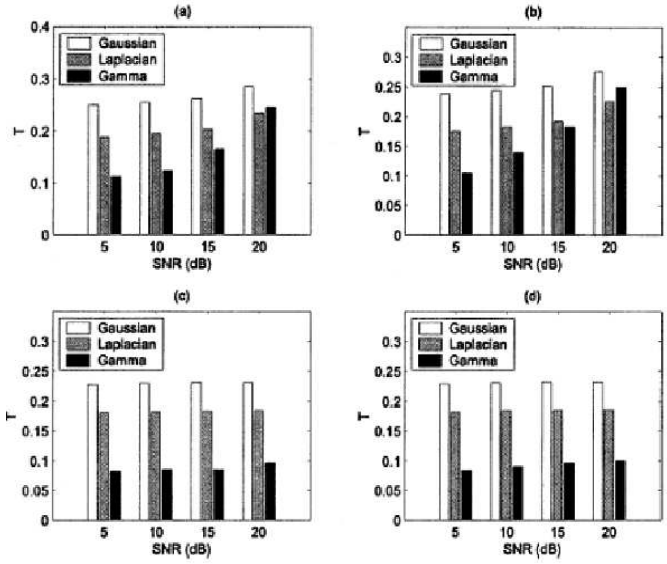
MSM : 다중 통계 모델

도면

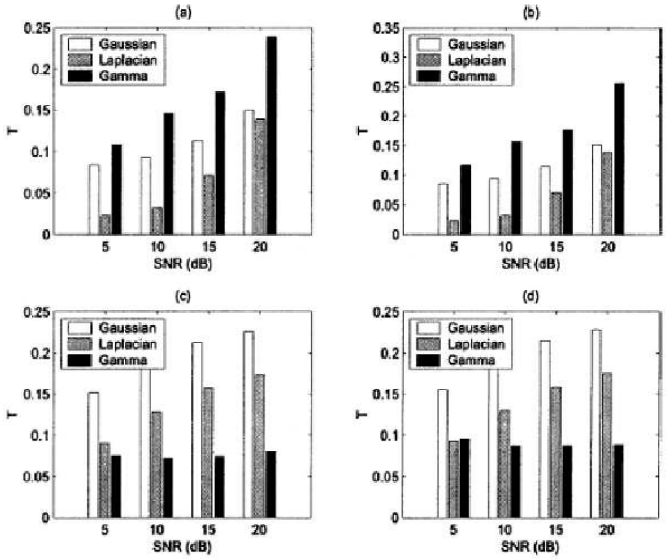
도면1


(a)
(b)
(c)
(d)
Gaussian
Laplacian
Gamma
T
SNR (dB)

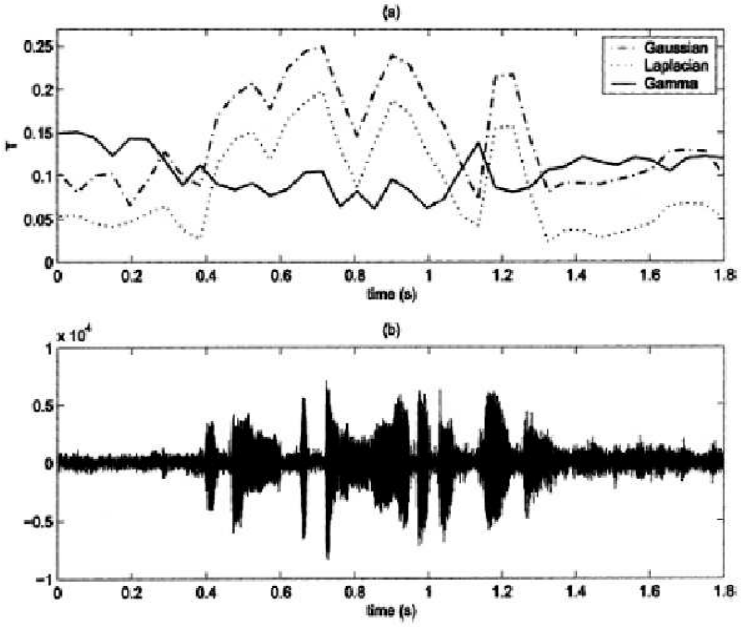
도면2


(a)
(b)
(c)
(d)
Gaussian
Laplacian
Gamma
T
SNR (dB)

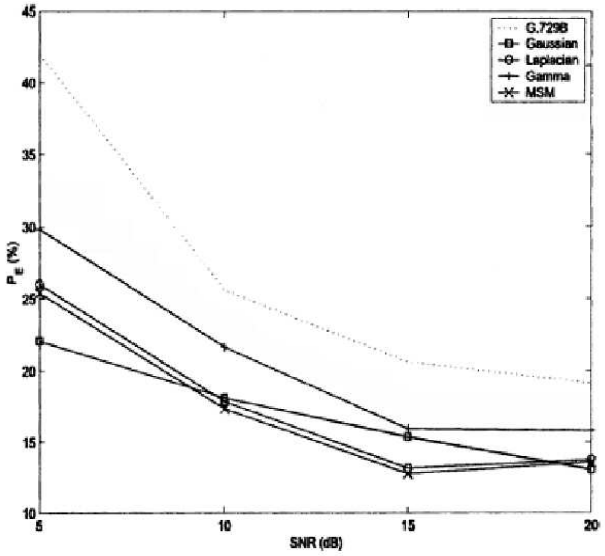
도면3


(a)
(b)
(c)
(d)
Gaussian
Laplacian
Gamma
T
SNR (dB)

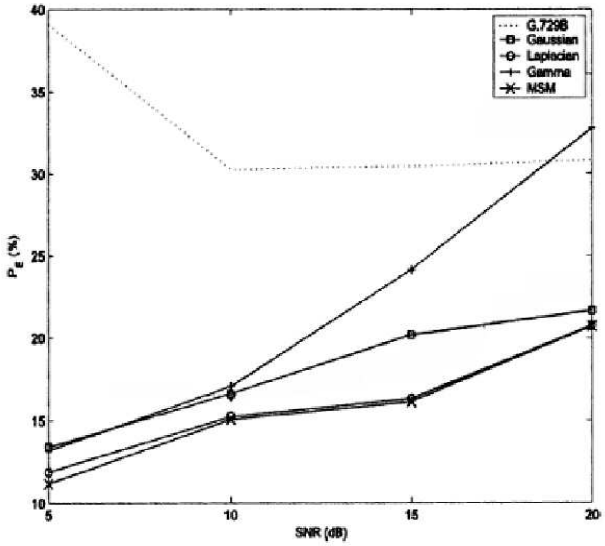
도면4


(a)
Gaussian
Laplacian
Gamma
T
time (s)
(b)
x 10^4

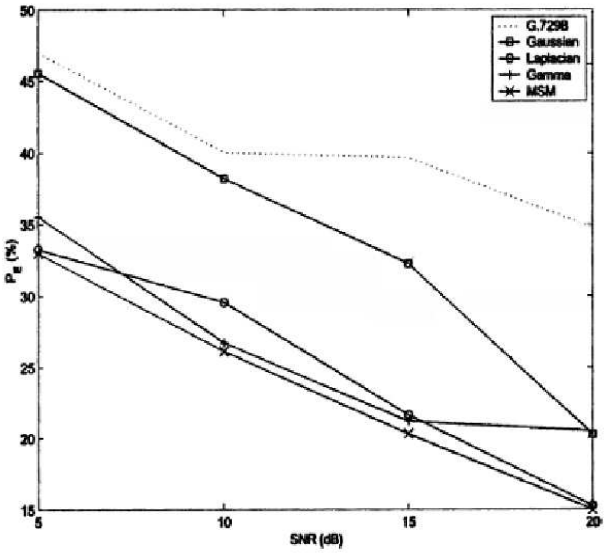
도면5


G.729B
Gaussian
Laplacian
Gamma
MSM
P_E (%)
SNR (dB)

도면6


G.729B
Gaussian
Laplacian
Gamma
MSM
P_E (%)
SNR (dB)

도면7


G.729B
Gaussian
Laplacian
Gamma
MSM
P_e (%)
SNR (dB)
50
45
40
35
30
25
20
15
5
10
15
20