한국ITS학회논문지
제8권, 제4호 (2009년 8월)
pp. 28~32

자동차 잡음환경에서의 음성인식에 적용된 두 종류의 일반화된 감마분포 기반의 음성추정 알고리즘 비교

Comparison of Two Speech Estimation Algorithms Based on Generalized-Gamma Distribution Applied to Speech Recognition in Car Noisy Environment

김 형 국* 이 진 호**
(Hyoung-Gook Kim) (Jin-Ho Lee)

요 약

본 논문은 DFT기반의 단일마이크 음성향상 방식에 적용된 두 종류의 generalized-Gamma 분포기반의 음성추정 알고리즘을 비교한다. 음성향상 방식으로서는 최소잡음성분에 의한 회귀적인 평균스펙트럼 값으로부터 유도되는 잡음 추정을 각각 $\kappa=1$인 경우와 $\kappa=2$인 경우의 Gamma 분포를 이용한 음성추정 기법에 결합하여 음질을 향상시켰다. 각 방식에 의해 향상된 음성신호를 자동차 환경에서의 음성인식에 적용하여 그 성능을 비교하였다.

Abstract

This paper compares two speech estimators under a generalized Gamma distribution for DFT-based single-microphone speech enhancement methods. For the speech enhancement, the noise estimation based on recursive averaging spectral values by spectral minimum noise is applied to two speech estimators based on the generalized Gamma distribution using $\kappa=1$ or $\kappa=2$. The performance of two speech enhancement algorithms is measured by recognition accuracy of automatic speech recognition(ASR) in car noisy environment.

Key words: Speech estimators, generalized gamma distribution, noise estimation, speech recognition

Ⅰ. 서 론

잡음의 크기가 크며 잡음의 종류 및 특성이 다양한 자동차 환경에서는 잡음에 의해 음성신호가 왜곡되어 음성인식 성능을 저하시킨다. 이러한 자동차 환경에서 강인한 음성인식 성능을 획득하기 위해서는 전처리 영역에서의 효과적인 음성향상 알고리즘이 필요하다.

단일 마이크로폰를 사용하는 음성인식엔진에 적용되는 DFT(discrete Fourier transform)기반에서의 대부분의 음성향상 알고리즘은 잡음추정과 음성추정의 2가지 요소로 구성된다.

† 이 논문은 2009년도 광운대학교 교내 학술연구비 지원에 의해 연구되었음.
* 주저자 : 광운대학교 전파공학과 부교수(교신저자)
** 공저자 : 광운대학교 전파공학과 학사과정
† 논문접수일 : 2009년 8월 3일
† 논문심사일 : 2009년 8월 25일
† 게재확정일 : 2009년 8월 26일

비정상 잡음 환경에서 효과적인 잡음추정은 DFT를 통해 획득된 파워스펙트럼의 회귀적인 평균값에서 추적된 최소 잡음성분을 기반으로 음성구간을 검출하여 잡음을 추정하는 방식들[1, 2]이 우수한 성능을 나타내고 있다.

음성추정방식으로서는 최근에 DFT coefficients의 generalized Gamma 분포 [3] 기반의 음성추정 방식이 기존에 사용되어 오고 있는 Gaussian 분포[4]보다 음성향상에 있어서 우수한 평가를 받고 있으므로, 본 논문에서는 $\kappa=1$인 경우와 $\kappa=2$인 경우의 Gamma 분포를 적용한 음성추정기법을 자동차 환경에서의 음성인식에 적용하여 두 방식의 성능을 비교한다.

본 논문은 다음과 같이 구성된다. 2장에서는 generalized Gamma분포를 적용한 음성향상 알고리즘을 설명한다. 3장에서는 제시된 음성향상 알고리즘을 이용하여 실험을 수행하고 실험 결과를 논의한다. 마지막으로 4장에서는 결론을 제시한다.

Ⅱ. 음성향상 알고리즘

본 논문에서 사용한 음성향상 알고리즘은 잡음추정과 음성추정의 2가지 요소로 구성된다.

잡음추정은 5단계 과정을 통해 수행된다. 첫 번째 단계인 평균스펙트럼 계산에서는, 입력 음성신호의 DFT를 통해 획득된 스펙트럼 $X(k,l)$에서 주파수 축과 시간 축에 대해 스무딩을 적용한 파워 성분의 평균 $X_T(k,l)$을 일차 회귀 방정식에 의해 구한다.

두 번째 단계에서는 각 시간 축 프레임 지수 l로부터 구해진 평균 스펙트럼의 최소값을 정의된 프레임 수 이내에서 비교함으로써 평균 스펙트럼 $X_T(k,l)$의 스펙트럼 최소잡음성분을 구한다.

세 번째 단계는 입력된 음성신호의 평균 스펙트럼과 최소잡음성분 스펙트럼간의 비율을 이용하여 각 프레임의 시간-주파수 성분에서의 음성존재구간과 비 음성 존재구간을 구별하는 음성구간 검출을 수행한다.

네 번째 단계로, 음성구간을 기반으로 음성존재확률 $p(k,l)$를 추정하고, 추정된 $p(k,l)$를 이용하여 잡음추정을 위한 최적 스무딩함수 $a_d(k.l)$를 계산한다.

음성존재구간과 비 음성 존재구간에서 스무딩 파라미터 $a_d(k.l)$을 이용하여 잡음은 식 (1)과 같이 추정된다.

$$\lambda_d(k,l)=\alpha_d(k,l)\lambda_d(k,l-1) +(1-\alpha_d(k,l))|X(k,l)|^2 \tag{1}$$

즉, 변화하는 환경에 따른 잡음의 파워를 추정하기 위해 현재 프레임이 잡음구간이라고 판단될 경우에만 잡음의 파워가 갱신되게 한다. 위의 잡음추정 방식은 [5]에 구체적으로 설명이 되어 있다.

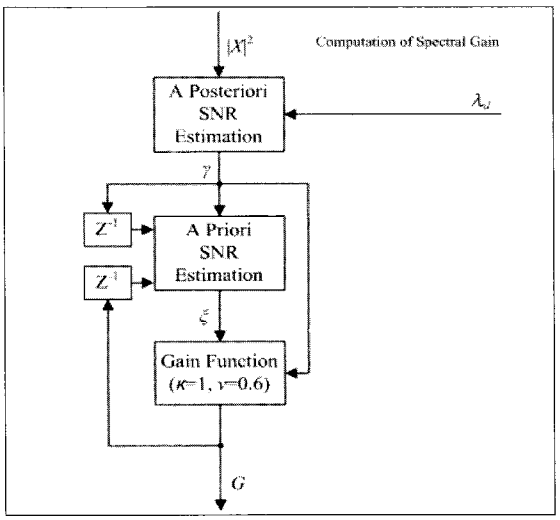
효과적인 음성향상을 위해서는 MMSE (minimum mean-square error) 기반의 음성추정방식이 사용되며, 음성추정은 <그림 1>과 같이 구성된다.

먼저, 추정된 잡음전력을 이용하여 사전 신호대잡음비(a priori SNR) $\xi(k,l)$와 사후 신호대잡음비(a posteriori SNR) $\gamma(k,l)$를 다음과 같이 구한다.

$$\gamma(k,l)=\begin{cases}\dfrac{|X(k,l)|^2}{\lambda_d(k,l)}, & \text{if} \quad |X(k,l)|^2>\lambda_d(k,l)\\ 1 & else\end{cases} \tag{2}$$

$$\xi(k,l)=\beta G^2(k,l-1)\gamma(k,l) +(1-\beta)\xi(k,l-1) \tag{3}$$

$$v(k,l)=\left(\frac{\xi(k,l)}{1+\xi(k,l)}\right)\gamma(k,l) \tag{4}$$



<그림 1> 스펙트럼 이득 계산의 구성도
<Fig. 1> Block diagram of gain function

여기서 $\beta(0<\beta<1)$은 스무딩 파라미터를 나타낸다. 목표하는 음성향상 스펙트럼은 잡음제거이득과 오염된 음성신호의 곱을 통해 다음과 같이 획득된다.

$$\tilde{S}(k,l)=X(k,l)\cdot G(k,l) \tag{5}$$

위에서 기술된 바와 같이 잡음제거이득 $G(k,l)$은 MMSE 음성추정에서 최적의 이득을 계산하기 위해 Rayleigh 분포보다 우수한 성능을 보이는 generalized Gamma분포기반의 음성추정 이득함수 $G(k,l)$로 구성된다.

식 (7)에 나타난 generalized Gamma분포기반의 음성추정기법을 통해 음성향상 스펙트럼은 식 (6)과 같이 표현되며, $f_{S|X}(S_r,S_i|X)$는 오염된 음성 신호 X가 존재할 때 실수부S_r와 허수부 S_i를 가지는 음성신호 S의 조건부 확률분포를 나타낸다.

$$S(k,l)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(S_r+jS_i)f_{S|X}(S_r,S_i|X)dS_rdS_i \tag{6}$$

$$f_A(a)=\frac{\kappa\beta^{\nu}}{\Gamma(\nu)}a^{\kappa\nu-1}\exp(-\beta a^{\kappa}) \tag{7}$$

여기서 κ는 scale 파라미터, β, ν는 shape 파라미터를 나타내며, 랜덤변수 a는 음성 신호의 스펙트럼 크기를 나타낸다.

추정된 음성 $\tilde{S}(k,l)$는 오염된 음성신호와 음성추정 이득의 곱으로 표현되고 베이즈 법칙을 통해 식 (8)과 같이 유도된다.

$$G(k,l)=\frac{1}{r}\frac{\int_0^{\infty}\int_{-\pi}^{+\pi}ae^{j(\phi-\theta)}f_{x|s}(x|a,\phi)f_A(a)d\phi da}{\int_0^{\infty}\int_{-\pi}^{+\pi}f_{x|s}(x|a,\phi)f_A(a)d\phi da} \tag{8}$$

1. 이득 추정방법, $\kappa=1$

식 (8)로부터 특정한 값 κ, ν의 closed-form solution와 Bessel 함수를 추정함으로써 실질적으로 음성향상에 generalized Gamma 분포를 적용시킨 음성추정 이득 $G(k,l)$을 식 (9)와 같이 구할 수 있다.

$$G(k,l)=\frac{1}{r}\frac{\int_0^{\infty}a^{\nu}\exp\left[-\frac{a^2}{\lambda_d}-\beta a\right]I_1\left(\frac{2ar}{\lambda_d^2}\right)da}{\int_0^{\infty}a^{\nu-1}\exp\left[-\frac{a^2}{\lambda_d}-\beta a\right]I_0\left(\frac{2ar}{\lambda_d^2}\right)da} \tag{9}$$

$\nu(k,l)$에 의해 각 프레임을 low SNR과 high SNR으로 구분하고, 구분된 환경에 적합한 음성향상 이득을 구한다.

SNR환경에 따른 최적의 음성향상 이득을 구하기 위해 $\mu=\sqrt{\nu(\nu+1)}$, $x=2ar/\lambda_d$를 사용하여 식 (9)는 식 (10)과 같이 간소화된다.

$$G(k,l)=\frac{1}{2\gamma}\frac{\int_0^{\infty}x^{\nu}\exp\left[-\frac{x^2}{4\gamma}-\frac{\mu x}{2\sqrt{\xi\gamma}}\right]I_1(x)dx}{\int_0^{\infty}x^{\nu-1}\exp\left[-\frac{x^2}{4\gamma}-\frac{\mu x}{2\sqrt{\xi\gamma}}\right]I_0(x)dx} \tag{10}$$

Low SNR에서 잡음환경을 고려한 음성향상 이득은 강한잡음으로부터 음성신호를 향상시키기 위해 Bessel함수를 식 (11)로 근사화하여 식 (12)로 나타내진다.

$$\begin{aligned}I_0(x;K)&=\sum_{j=0}^{K-1}\left(\frac{x}{2}\right)^{2j}\frac{1}{(j!)^2}\\ I_1(x;K)&=\sum_{j=0}^{K-1}\left(\frac{x}{2}\right)^{2j+1}\frac{1}{j!(j+1)!}\end{aligned} \tag{11}$$

$$G_{\ll,K}(k,l)=\frac{1}{2}\frac{\sum_{k=0}^{K-1}\frac{1}{k!(k+1)!}\left(\frac{\gamma}{2}\right)^k\Gamma(n+2)D_{-(n+2)}(\chi)}{\sum_{k=0}^{K-1}\frac{1}{k!}\left(\frac{\gamma}{2}\right)^k\Gamma(n)D_{-(n)}(\chi)} \tag{12}$$

여기서 $D_v(\cdot)$은 ν차 parabolic cylinder function, $\Gamma(\cdot)$는 gamma function이며, $\chi=\frac{\mu}{\sqrt{2\xi}}$, $n=2\nu+2k$를 각각 나타낸다.

High SNR에서 잡음환경을 고려한 음성향상 이득은 음성구간에서 최소한의 잡음을 제거하여 음성왜곡을 방지하기 위해 식 (13)의 Bessel함수 근사화를 통해 식 (14)와 같이 표현된다.

$$I_0(x)=\frac{1}{\sqrt{2\pi x}}exp(x) \tag{13}$$

$$G_{\gg}(k,l)=\left[\frac{\left(\nu-\frac{1}{2}\right)D_{-(\nu+0.5)}(\chi')+\left(\nu+\frac{1}{2}\right)D_{-(\nu+1.5)}(\chi')}{2\zeta D_{-(\nu-0.5)}(\chi')}\right] \quad (14)$$

여기서 $\chi'=\frac{\mu}{\sqrt{2\xi}}-\sqrt{2\gamma}$를 나타낸다.

2. 이득 추정방법, $\kappa=2$

식 (8)로부터 $\kappa=2$인 경우의 음성향상 이득 $G(k,l)$는 식 (15)와 같이 구할 수 있다.

$$G(k,l)=\frac{\nu\xi(k,l)}{\nu+\xi(k,l)}\frac{M\left(\nu+1;2;\frac{\gamma(k,l)\xi(k,l)}{\nu+\xi(k,l)}\right)}{M\left(\nu;1;\frac{\gamma(k,l)\xi(k,l)}{\nu+\xi(k,l)}\right)} \quad (15)$$

여기서 $M(a;b;z)$은 CHF(confluent hypergeometric function)을 나타내며, a와 b는 CHF의 coefficient, z는 랜덤변수이다. $M(a;b;z)$는 아래와 같이 표현된다.

$$M(a;b;z)=1+\frac{az}{b}+\frac{(a)_2z^2}{(b)_2 2!}+\ldots+\frac{(a)_n z^n}{(b)_n n!}\ldots \quad (16)$$

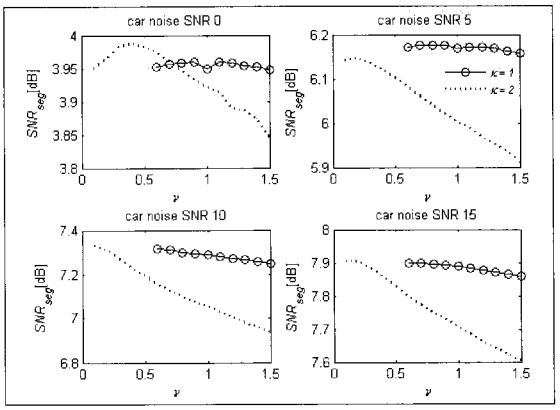
Ⅲ. 실험 및 결과고찰

본 논문에서는 두 종류의 음성추정 방식을 segmental SNR 측정과 자동차 환경에서의 음성인식에 적용하여 성능을 평가하였다.

<그림 2>는 자동차 잡음이 섞인 음성신호의 SNR 0dB, 5dB, 10dB, 15dB 환경에서 $\kappa=1$와 $\kappa=2$의 음성추정방식을 segmental SNR을 통해 측정한 결과를 나타낸다.

<그림 2>에서 나타난 바와 같이, 모든 자동차 잡음환경에서 $\kappa=1$을 사용한 음성추정 방식이 $\kappa=2$를 사용한 방식보다 segmental SNR 결과가 우수했으며, ν의 값이 상승할수록 $\kappa=1$과 $\kappa=2$의 segmental SNR 결과의 차이가 커짐을 알 수 있다.

음성 인식기는 HTK를 사용하여 단어당 16개의 상태를 가지고 단어 단위 HMM으로 모델링하였다. 또한 특징벡터로서 12개의 MFCC와 로그 프레임 에너지, 그것의 증분, 증분의 증분으로 정의되는 총 39차를 NOISEX-92 음성 데이터베이스에 적용하여 자동차 잡음에 대한 SNR -6dB부터 18dB사이에서의 음성인식률 평균으로 <표 1>에 나타내었다.



<그림 2> 다양한 자동차 잡음 환경에 대한 segmental SNR 측정 결과
<Fig. 2> Segmental SNR results in various car noisy environment

<표 1> 인식률 성능
<Table 1> Accuracy of the recognition rates

음성추정 방식	평균 음성인식 정확도
$\kappa=1,\ \nu=0.7$	84.5%
$\kappa=2,\ \nu=0.7$	84.3%

<표 1>에 나타난 바와 같이 $\kappa=1$을 사용한 음성추정 방식과 $\kappa=2$를 사용한 음성추정 방식을 사용한 음성인식 정확도는 거의 동일한 결과를 나타내었다. 즉, $\kappa=1$과 $\kappa=2$의 음성추정 방식에 대한 segmental SNR 측정결과의 미소한 차이는 음성인식성능에 영향이 없음을 알 수 있었다.

Ⅳ. 결 론

본 논문에서는 $\kappa=1$과 $\kappa=2$를 갖는 일반화된 Gamma 분포기반의 음성추정 방식을 자동차 환경에서의 음성인식에 적용하여 성능을 측정하였다. 실험을 통해 연산량이 $\kappa=1$ 보다 작은 $\kappa=2$의 방식이 $\kappa=1$의 방식과 동일한 음성인식 결과를 나타냄을 알

수 있었다.

차후 계획으로는 새로운 잡음추정 방식을 일반화된 Gamma 분포기반의 음성추정 방식에 적용하여 다양한 잡음환경에서의 음성인식성능을 측정하고자 한다.

참 고 문 헌

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no 5, pp. 504-512, July 2001.

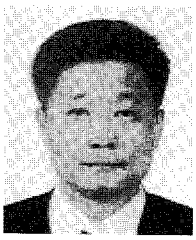
[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Processing*, vol. 81, no. 11, pp. 2403-2418, Nov. 2001.

[3] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 6, pp. 1741-1752, Aug. 2007.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. 33, no. 2, pp. 443-445, Dec. 1985.

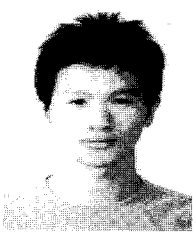
[5] 김형국, 신동, 이진호, "잡음에 강인한 음성인식을 위한 generalized Gamma 분포기반과 spectral gain floor를 결합한 음성향상기법," *한국ITS학회 논문지*, 제3권, 제2호, pp. 64-70, 2009. 6.

저자소개



김 형 국 (Kim, Hyoung-Gook)

2007년 3월 ~ 현재 : 광운대학교 전파공학과 부교수
2005년 4월 ~ 2007년 2월 : 삼성종합기술원 수석연구원
2002년 8월 ~ 2005년 3월 : 독일 베를린 공과대학교 Adjunct Professor
1999년 1월 ~ 2002년 7월 : 독일 Cortologic AG 책임연구원



이 진 호 (Lee, Jin-Ho)

2003년 3월 ~ 현재 : 광운대학교 전파공학과 학사과정