Maximum a Posterior Probability and Cumulative Distribution Function Equalization Methods for Speech Spectral Estimation with Application in Noise Suppression Filtering

Tran Huy Dat[1], Kazuya Takeda[1], and Fumitada Itakura[2]

[1] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
{dat, takeda}@sp.is.nagoya-u.ac.jp
[2] Graduate school of Information Engineering, Meijo University, Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan
itakuraf@ccmfs.meijo-u.ac.jp

Abstract. In this work, we develop and compare noise suppression filtering systems based on maximum a posterior probability (MAP) and cumulative distribution function equalization (CDFE) estimation of speech spectrum. In these systems, we use a double-gamma modeling for both the speech and noise spectral components, in which the distributions are adapted to the actual parameters in each frequency bin. The performances of the proposed systems are tested using the Aurora database they are shown to be better than conventional systems derived from the MMSE method. Whereas the MAP-based method performed best in the SNR improvement, the CDFE-based system provides a lower musical noise level and shows a higher recognition rate.

1 Introduction

Noise reduction is an important problem in speech and audio processing. Among single channel approaches, the statistical methods for speech spectrum estimation have been frequently used [1]. The MMSE and MAP estimations for the Gaussian model of the speech spectrum were proposed by Ephraim and Malah [2] and Wolfe and Godsill [3], respectively. Later, a MAP based on the super-gaussian modeling of speech was derived by Lotter[4] and the MMSE based on gamma modeling was investigated by Martin [5],[6]. However, in both cases, the prior distribution parameters were chosen blindly without any adaptation. In previous work, we proposed an improved version of MAP estimation for the speech spectral magnitude by using generalized gamma modeling of the speech spectral magnitude [6]. However, that work was limited by the Gaussian assumption of the noise spectrum and therefore, was not effective under certain noise conditions. In this work, we extend gamma modeling for both speech and noise spectra and derive the MAP and cumulative distribution equalization (CDFE) estimation for the spectral components. As in our previous work [7], the prior

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 328–337, 2005.

distribution is adapted from observed signals and the estimations are derived for an arbitrary set of distribution parameters. The reason for applying MAP or CDFE instead of MMSE is that, the last generally provides non closed form solution, which is complicated even for numerical methods. Cumulative distribution equalization has frequently been used in data-driven approaches, where the empirical histogram is used. In this work, we show that, this method can also be usefully applied in the model-based manner, where the cumulative distribution function (cdf) is used. To overcome the difficulties of applying cdf, we develop an cdf estimation method via the characteristic function, which implies a multiplication for the additive model. The organization of this paper is as follows. In section 2, we describe the double-gamma modeling of the speech and noise spectral components. Section 3 contains a review of the MMSE estimation of speech spectral estimation. In sections 4 and 5, we develop the MAP and CDFE estimation of speech spectral components using the proposed modeling of speech and noise. In section 6, we reports an experimental evaluation of implemented noise suppression filtering systems, and section 7 is a summary of the present work.

2 Statistical Modeling of Speech and Noise Spectral Components

2.1 Double-Gamma Modeling of Speech and Noise Spectra

Consider the additive model of the noisy speech as below:

$$\mathbf{X}[k,m] = \mathbf{S}[k,m] + \mathbf{N}[k,m], \tag{1}$$

where $\mathbf{X}[k,m]$, $\mathbf{S}[k,m]$, and $\mathbf{N}[k,m]$ are noisy, clean speech and noise complex spectrum. The pair $[k,m]$ indicates the frequency-frame index. Each complex spectrum is presented in terms of the spectral components (real and imaginary parts) as follows:

$$\mathbf{C}[k,m] = C_R[k,m] + jC_I[k,m]. \tag{2}$$

The following assumptions are assumed for speech and noise spectral components: (1) spectral components are independent and zero-mean, (2) spectral component pdf is symmetrical, (3) The variances of spectral components are power density and determined at each frequency-frame index $[k,m]$. In this work, we investigated double-gamma modeling for both speech and noise.

$$p_{double-gamma}\left(C\left[k,m\right]\right) = \frac{ba^{b-1}}{2\sigma_C^b\left[k,m\right]\Gamma\left(a\right)} C^{b-1}\left[k,m\right] \exp\left(-b\frac{C\left[k,m\right]}{\sigma_C\left[k,m\right]}\right) \tag{3}$$

As an alternative, the conventional Gaussian model is also investigated and noted as follows:

$$p_{gauss}\left(C\left[k,m\right]\right) = \frac{1}{\sqrt{2\pi}\sigma_C\left[k,m\right]} \exp\left(-\frac{C^2\left[k,m\right]}{2\sigma_C^2\left[k,m\right]}\right), \tag{4}$$

where $C\left[k,m\right]$ denotes the spectral component (real or imaginary part) and $\sigma_C^2\left[k,m\right]$ denotes the local power density at each frequency-frame index $[k,m]$.

Note that the normalization condition $\langle C^2 \rangle = \sigma_C^2$ implies the following relationship between a and b:

$$\frac{a\,(a+1)}{b^2} = 1 \rightarrow b = \sqrt{a\,(a+1)}, \tag{5}$$

Since the spectral components are assumed to be identical independent variables, the additive model of the complex noisy speech spectral (3) can be simply denoted in terms of the spectral component as

$$X = S + N, \tag{6}$$

where each symbol in (5) corresponds to the real and imaginary parts of complex spectrum. The following three models of the speech and noise distributions are consequently investigated in this work: Gaussian/Gaussian (Model 1), gamma/Gaussian (Model 2) and gamma/gamma (Model 3).

2.2 Actual Adaptation of the Modeled Distribution Parameters

Since the prior distributions of speech and noise are scaled by their local power densities (3), which are estimated separately, the prior parameter should be adapted from each observed noisy speech. In this work, we develop a parameter estimation method, in which the prior pdf is adapted in each frequency bin. As done in our previous work [7], the high-order moments of observed noisy speech spectrum are used to derive estimation equation. In this case, it is done for both the speech and noise prior pdf. For the gamma-speech, Gaussian-noise model (Model 1), the four moments of the noisy speech spectral component are expressed as

$$\langle X^4 \rangle = \bar{\sigma}_S^4 M_4\,(a_S) + \bar{\sigma}_N^4 M_4\,(a_N) + 6\bar{\sigma}_S^2 \bar{\sigma}_N^2, \tag{7}$$

where the fourth moments of the noise and speech spectral components are given below following the Gaussian distribution,

$$M_4\,(a_N) = 3, \tag{8}$$

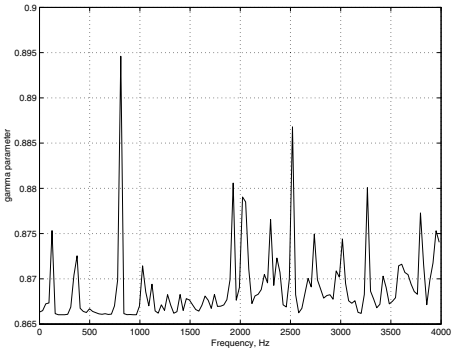

Fig. 1. Example of double-gamma estimation of speech spectral components

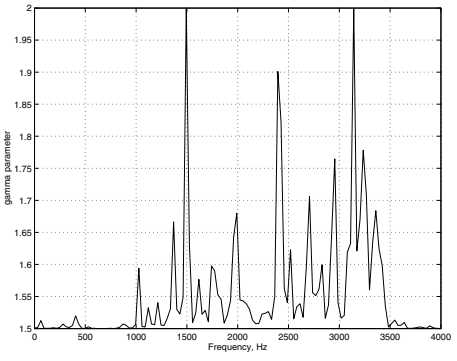

Fig. 2. Example of double-gamma estimation of noise spectral components

and through gamma distribution as

$$M_4\left(a_S\right) = \frac{(a_S+2)(a_S+3)}{b_S^2}. \tag{9}$$

Substituting (8) and (9) into (7) and taking (5) into account, the speech prior distribution parameter is given in a closed-form solution. Analogously, for the speech-gamma and noise-gamma model (model 3), the distribution parameter is estimated using the pair of fourth-order and sixth-order moments of the observed noisy spectrum. Figures 1 and 2 show examples of double-gamma parameter estimation for a noisy speech signal under the 5dB street noise condition.

3 MMSE Estimation

In general, the MMSE estimation is given by the conditional expectation,

$$\hat{S} = E\left[S|X\right] = \frac{\int\limits_{-\infty}^{\infty} Sp\left(X,S\right)dS}{p\left(X\right)} = \frac{\int\limits_{-\infty}^{\infty} Sp\left(X|S\right)p\left(S\right)dS}{\int\limits_{-\infty}^{\infty} p\left(X|S\right)p\left(S\right)dS}, \tag{10}$$

where the conditional pdf $p\left(X|S\right)$ is given by the noise pdf and the prior distribution $p\left(S\right)$ is the Gaussian (3) or double-gamma distribution (4). The MMSE estimation of the speech spectral components for the Gaussian modeling of noise and speech spectra yields the conventional Wiener filtering:

$$\hat{S} = \frac{\sigma_S^2}{\sigma_S^2+\sigma_N^2}X. \tag{11}$$

The MMSE estimation using gamma prior was investigated by R.Martin [5],[6], for two special cases of double-gamma distribution, including the Laplacian distribution, in which a closed-form solutions are given. However, the MMSE

in other cases of gamma modeling does not yield a closed-form solution of the estimation. For our proposed system, where the prior distribution parameters are adapted from each observed signal, the numerical calculation of integral (10) should be implemented. However the main drawback of this method is highly expensive computational cost and therefore we don't use this method for our system.

4 MAP Estimation

MAP is a general estimation method and is used in this work to estimate the speech spectral components. In contrast to the estimations presented in [3] and [7], where the spectral magnitude domain was uses, in this work, we use the spectral components domain to derive the estimation. The advantage of using this domain is that exactly matches the additive model of noisy speech and the estimation is given not only for the Gaussian model of noise spectrum. The general form of MAP estimation

$$\hat{S} = \arg\max_{S} \log\left(p\left(S|X\right)\right) = \arg\max_{S} \log\left(p\left(X|S\right)p\left(S\right)\right), \tag{12}$$

yields an equation of the derivatives

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right) + \log\left(p\left(S\right)\right)\right] = 0. \tag{13}$$

Since the MAP estimation for the model 1 implies the classical Wiener filter, we begin this section with model 2.

4.1 Model 2: Gamma Speech and Gaussian Noise

For this model, the conditional and prior distributions are derived as follows:

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right)\right] = \frac{X-S}{\sigma_N^2}, \tag{14}$$

$$\frac{\partial}{\partial S}\left[\log\left(p\left(S\right)\right)\right] = \frac{\left(a_S-1\right)}{S} - sign\left(S\right)\frac{b_S}{\sigma_S}. \tag{15}$$

Equations (14) and (15) imply the following second-order equation for the gain function $G = \frac{\hat{S}}{X}$:

$$G^2 - G\left(1 - \frac{sign\left(X\right)b_S}{\sqrt{\gamma\xi}}\right) + \frac{\left(a_S-1\right)}{\gamma} = 0, \tag{16}$$

where: $\gamma = \frac{X^2}{\sigma_N^2}$, and $\xi = \frac{\sigma_S^2}{\sigma_N^2}$ are posterior and prior SNRs, respectively which are estimated separately [2]. Obtaining a closed form solution for the MAP estimation is important because then the global maximum of posterior probability in (12) can be found strictly:

$$G = \max\left\{u \pm \sqrt{u^2+v}, 0\right\}, \tag{17}$$

where

$$u = \left(0.5 - \frac{sign\,(X)\,b_S}{\sqrt{4\gamma\xi}}\right), \tag{18}$$

$$v = \frac{(a_S - 1)}{4\gamma}. \tag{19}$$

4.2 Model 3: Gamma Speech and Gamma Noise

For this model, the conditional distribution in (13) can be expressed as

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right)\right] = -\frac{(a_N - 1)}{X - S} - sign\,(X - S)\,\frac{b_N}{\sigma_N}. \tag{20}$$

Analogously, a second order equation for the gain function is derived.

$$G^2 - G\left(1 - \frac{(a_S + a_N - 2)}{\sqrt{\gamma}sign\,(X)\left(b_N - \frac{b_S}{\sqrt{\xi}}\right)}\right) + \frac{(a_S - 1)}{\sqrt{\gamma}sign\,(X)\left(b_N - \frac{b_S}{\sqrt{\xi}}\right)} = 0. \tag{21}$$

The solution of Eq.(21) is given in the same manner as for (17).

5 Cumulative Distribution Function Equalization

One remaining problem of the above MAP estimation is the relative sensitivity to the "poor fit" prior estimation, or other words it requires a sufficiently "good" prior. Therefore, in addition to the MAP estimation, we investigate an alternative estimation based on cumulative distribution function equalization (CDFE).

5.1 Cumulative Distribution Function Equalization

This method (CDFE) was originally called as histogram equalization and has been used in data-driven approaches. In this work, we investigate the use of cdf for the model-based approaches, in which modeled distributions are used. The principle of this method is to identify a non-linear transform from noisy to clean features, which matches the cumulative distribution function. Denoting the general equalization

$$\widehat{s} = g\,(x)\,, \tag{22}$$

the criterion for our estimation here is expressed as

$$F_{g(x)}\left(g\,(x)\right) = F_s\,(s)\,. \tag{23}$$

The key point of the method is that, the cumulative distribution function (cdf) is invariant though arbitrary nonlinear functional, that is,

$$F_{g(x)}\left(g\,(x)\right) = F_x\,(x)\,. \tag{24}$$

From (23) and (24), the "best" nonlinear transform is obtained by equalizing cdf of noisy to clean signals.

$$g\,(x) = F_s^{-1}\left(F_x\,(x)\right) \tag{25}$$

5.2 Model 1: Gaussian Speech and Gaussian Noise

For Gaussian modeling of both noise and speech spectral components, the noisy speech spectral components are also Gaussian

$$X \sim \mathrm{N}\left(0, \sigma_S^2 + \sigma_N^2\right). \tag{26}$$

Since both cdf $F_X(.)$, and $F_S(.)$ are Gaussian, the CDFE operation is carried out without any difficulties.

5.3 CDF Estimation Via the Characteristic Function

CDFE has the following main problem. Excepting the Gaussian model considered above, the cdf of noisy speech, presented as an addition of speech and noise, has no analytical form. To overcome this problem, we develop a cdf estimation method by using the characteristic function as follows:

$$F(x) = \begin{cases} 1 & x \geq m + 4\sigma \\ \frac{1}{2} - sign(x)\frac{2}{\pi}\int\limits_{\varepsilon}^{2\pi} \frac{f(u)}{u}\sin(ux)\,du & m + 4\sigma > x > m - 4\sigma, \\ 0 & x \leq m - 4\sigma \end{cases} \tag{27}$$

where $f(u)$ denotes the characteristic function [8] of noisy speech spectral components. The main point here is that the characteristic function of the additive model (6) is multiple and therefore convenient for implementation. Note that, according to the symmetrical assumption of the pdf of speech and noise spectral components, the characteristic function of noisy speech is always a real function.

5.4 Model 2: Gamma Speech and Gaussian Noise

Denoting characteristic function of the Gaussian distribution of noise and double-gamma distribution of speech spectral components, respectively, as follows:

$$f_N(u) = e^{-\frac{u^2\sigma_N^2}{2}}, \tag{28}$$

$$f_S(u) = Re\left[\left(\frac{a_S}{a_S - iu}\right)^{b_S}\right] = \cos\left(b_S a \cos\left(\frac{a_S^2}{a_S^2 + u^2}\right)\right), \tag{29}$$

the characteristic function of the noisy speech spectral component is obtained by multiplying (28) and (29). The CDFE is then derived using (27) and (25).

5.5 Model 3: Gamma Speech and Gamma Noise

For gamma modeling of both speech and noise spectral components, the cdf of noisy speech spectral components is estimated from the characteristic functions of speech and noise and is denoted by

$$f_X(u) = \cos\left(b_N a \cos\left(\frac{a_N^2}{a_N^2 + u^2}\right)\right)\cos\left(b_S a \cos\left(\frac{a_S^2}{a_S^2 + u^2}\right)\right). \tag{30}$$

6 Experiment

The proposed noise suppression filtering systems are tested using AURORA2 to determine the ASR performance [10], where the speech enhancement is applied for both testing and training databases. The noise and signal powers are estimated using the minimum statistic [9] and decision directed [2]. The three models of speech and noise modeling described above are investigated. Each system is identified according to the estimation method (MMSE,MAP ,CDFE) and the assumed models (1, 2, 3). For reference, the Ephraim-Malah LSA version based on Gaussian modeling [1] is also implemented. The reference MMSE versions

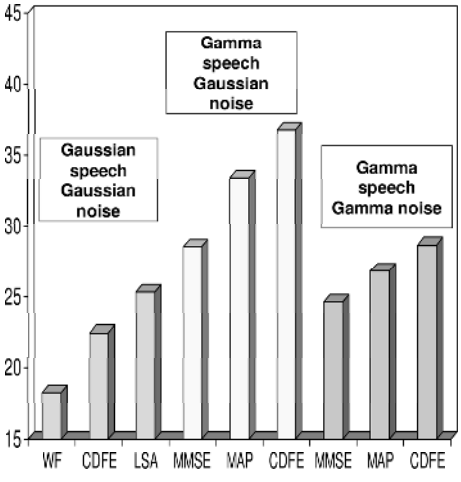

Fig. 3. ASR relative improvement of clean training: overall results

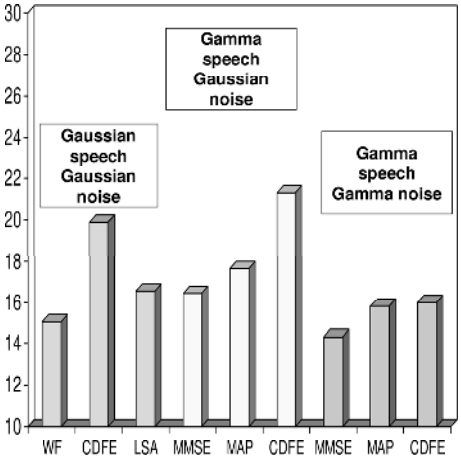

Fig. 4. ASR relative improvement of multi-conditions training: overall results

Table 1. Listening test: Q1-Which one is less distorted? Q-2 Which one is less noisy? Q-3 Which one is best?

Q	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station
1	CDFE-1	CDFE-2	CDFE-3	CDFE-3	CDFE-3	CDFE-3	CDFE-2	CDFE-2
2	MAP-2	MAP-2	MAP-2	MAP-2	CDFE-3	CDFE-3	MAP-2	MAP2
3	CDFE-1	CDFE-2	MAP-2	CDFE-3	CDFE-3	CDFE-1	MAP-2	CDFE-2

Table 2. Best ASR performance in each noise condition

Cond.	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station
CL	CDFE-1	CDFE-2	MAP-2	MAP-2	CDFE-3	CDFE-3	MAP-2	CDFE-2
MT	CDFE-1	MAP-2	MAP-2	CDFE-1	CDFE-3	CDFE-3	MAP-2	CDFE-2

Table 3. SNR improvement-Overall results [dB]

Meth	WF	CDFE-1	LSA	MMSE-2	MAP-2	CDFE-2	MMSE-3	MAP-3	CDFE-3
dB	3.25	3.52	4.25	5.65	6.32	6.07	5.45	6.11	6.12

using Laplacian/Gaussian and Laplacian/Laplacian modeling of speech and noise [5] are implemented. A simple listening test is performed with four subjects listening to 25 random chosen utterances of each noise conditions. Table 1 shows the results of the listening test, table 2 shows the best method in terms of ASR for each noise condition, and table 3 shows the noise reduction comparison in terms of SNR improvement. From the tables, we can conclude that, CDFE-3 is superior to other methods only for the restaurant and street noise conditions. Meanwhile, the MAP-2 is dominated in SNR improvements. The overall results of ASR performance using clean HMM and multi-conditions training are shown in Figure 3-4. The results in Figure 3 indicates that, CDFE-2 performs the best, as CDFE-3 is even worse than CDFE-1. For multi conditions training, the best performances are shown by the CDFE-1 and CDFE-2. This means that, double-gamma model for speech always performs better than Gaussian model but the Gaussian modeling is better for noise modeling under most of noise conditions.

7 Conclusion

We develop a maximum posterior probability and cumulative distribution equalization method for the speech spectral estimation using the double-gamma modeling of speech and noise spectral components. The main point of the

proposed method is that a solution is given for an arbitrary set of prior distributions and therefore it is possible to combine the estimation method to a prior adaptation to improve the performances of system. Double-gamma modeling of speech and noise spectral component was shown to be adaptable to the actual distribution, without any use of a training data. The results of the experimental evaluation shows the advantage of the proposed MAP and CDFE comparing to the conventional MMSE method. Gamma modeling is superior to the Gaussian for the speech spectral modeling in all cases, but is better for noise modeling only for some particular noise conditions (restaurant and street). The CDFE shows the best ASR performance, while the MAP is better in noise reduction.

References

1. Y. Ephraim, "Statistical model based speech enhancement systems," *IEEE Proc.*, vol. 80, pp. 1526-1555, 1992.
2. Y. Ephraim, and D. Malah, "Speech enhancement using a MMSE log-spectral amplitude estimations," *IEEE Trans. ASSP*, Vol. 33, No. 2, pp.443-445, 1985.
3. P. Wolfe and S. Godsill, "Simple alternatives to the Ephraim suppression rule for speech enhancement", *IEEE Workshop on Statistical Signal Processing*, 2001.
4. T. Lotter and P. Vary "Noise Reduction by Maximum A Posteriori Spectral Amplitude Estimation with Supergaussian Speech Modeling," *in Proc. IWAENC*, Kyoto, Japan, 2003.
5. R. Martin, B. Colin, "Speech enhancement in DFT domain using Laplacian priors," *in Proc. IWAENC*, Kyoto, Japan, 2003.
6. R. Martin, "Speech enhancement using MMSE Short Time Spectral Estimation with Gamma Speech Prior," *in Proc.ICASSP 02*, Orlando Florida,USA, 2002.
7. T.H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," *in Proc. ICASSP*, Philadelphia, USA,2005.
8. E Parzen, "On estimation of a probability density function and mode," emp Ann. Math. Statist V.33 pp.1065-1076, 1962.
9. R. Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics" *IEEE Trans. ASSP*, Vol. 9, No.5, pp.504-512, 2001.
10. H. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in ISCA ITRW ASR*, 2000.