The MAP and cumulative distribution function equalization methods for the 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.naoya-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 two statistical estimation methods of maximum a posterior probability (MAP) and cumulative distribution function equalization (CDFE) for the speech spectral component estimation approaches with the application in the noise suppression filters. In contrast to the histogram equalization approach, the CDFE is developed here based on speech and noise spectral modeling, which is also used in the MAP approach. Both of the conventional Gaussian and general gamma modeling of speech and noise spectral are investigated in this work. For the CDF estimation, we develop a flexible method for the non-Gaussian distribution by using the characteristic function. The advantage of proposed methods is that yields a flexible solution of the speech spectral estimation problem in general case of speech and noise modeling, which should be determined for each particular condition. Finally the systems of noise suppressed filters based on CDFE, MAP and MMSE are investigated via an experimental evaluation on the SNR improvement measurements. The performances of MAP and CDFE based systems are shown to be at least comparable or exceeded the conventional MMSE based in both the cases of Gaussian and gamma modeling. In the hearing test, the CDFE based system products a less musical noise level compared to the MAP and MMSE methods. ...

1 Introduction

Noise reduction is an important problem in speech and audio processing. Among the one-channel approaches, the statistical spectral estimation methods are shown to be effective for the noise reduction and to produce less speech distortion [1]. The statistical modeling and estimation method are two important task of the spectral estimation systems. The Gaussian modeling of speech and noise spectral components has been used in the literatures ans it was successfully combined with the MMSE estimators in the speech enhancement systems [1],[2]. However,

the last experiments on hearing and psychoacoustic [3] show that the speech spectral magnitude and phase are not independent and therefore, the Gaussian modeling on spectral components is not optimal for the speech spectral representation. In last decade, the number of research on non-Gaussian modeling of speech has been increased, where the approaches are carried out in two different ways. In [4], an implementation of Gaussian model based Emphraih-Malah filter is proposed via the spectral magnitude estimation and based on the generalized gamma modeling of speech and the MAP estimator. In [5],[6] authors propose a MMSE spectral components estimation approaches using the Laplacian or a special case of the gamma modeling of speech and noise spectrum. However the estimations presented in [5] and [6] are given just for these particular cases of the gamma modeling, where the distribution parameters are fixed, and therefore it limits the application in general case. Moreover the extension of MMSE for general distribution modeling is in application due its complexity. Beside to the MMSE, the MAP and histogram equalization have been used in signal processing. In this work we develop these techniques for the speech spectral components estimation problem in general case of gamma modeling of the speech and noise spectral components. In contrast to the conventional histogram equalization method, we develop a model based cumulative distribution function equalization (CDFE) approach, where the distribution of speech and noisy spech speech are given parametrically from the prior modeling. The main difficulties of applying CDF is that the CDF of noisy speech, which is presented as a superposition of the speech and noise is highly difficult to estimate with the non-Gaussian modeling. In this work, we develop a CDF estimation method via the characteristic function, which is multiple of speech and noise's and therefore always has a closed analytical forms with the Gaussian and gamma modeling. The organization of this paper is follows. In section 2, we make a briefly review of the statistical modeling of the speech spectral components. Section 3 reviews the MMSE approaches of the speech spectral estimation. Section 4 and 5 develop the MAP and CDFE estimation systems. In both cased the conventional Gaussian modeling and the gamma modeling are investigated. For both estimators, the noise suppressed filters are implemented. Section 6 reports an experimental evaluation and section 7 summaries the work.

2 Statistical modeling of speech and noise spectral components

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: $\mathbf{X}[k,m]$, $\mathbf{S}[k,m]$, $\mathbf{N}[k,m]$ are noisy speech, clean speech and noise complex spectrum and couple $[k,m]$ indicates the frequency and frame indexes. 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 extended form the conventional Gaussian model: (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 indexes $[k, m]$. In this work we investigate the Gaussian model and the double-gamma model of the spectral components for both speech and noise , where the PDFs are noted by

$$p_{gauss}\left(C\left[k,m\right]\right)=\frac{1}{\sqrt{2\pi}\sigma_{C}\left[k,m\right]}\exp\left(-\frac{C^{2}\left[k,m\right]}{2\sigma_{C}^{2}\left[k,m\right]}\right), \tag{3}$$

$$p_{double-gamma}\left(C\left[k,m\right]\right)=\frac{ba^{b-1}}{2\sigma_{C}^{b}\left[k,m\right]\Gamma\left(a\right)}C^{b-1}\left[k,m\right]\exp\left(-b\frac{C\left[k,m\right]}{\sigma_{C}\left[k,m\right]}\right), \tag{4}$$

where $C\left[k,m\right]$ denotes the spectral components and $\sigma_{C}^{2}\left[k,m\right]$ denotes the local power density at each frequency-frame indexes $[k, m]$. For making a comfortable, we will drop the indexes in the notation. Note that the normalization condition $\left\langle C^{2}\right\rangle=\frac{\sigma_{C}^{2}}{2}$ implies a relationship as follows:

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

and therefore, the distribution (4) is defined by only one parameter a. Since the spectral components are identical independent, the additive model of complex noisy speech spectral (3) can be denoted in terms of each spectral component as follows:

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

where each symbol in (5) indicates both the real and imaginary parts of complex spectrum. We investigate the following models of speech and noise:

2.1 Model 1: Gaussian noise and speech

From (5) and (3) the conditional and prior densities can be denoted as follows:

$$p\left(X|S\right)=\frac{1}{\sqrt{2\pi}\sigma_{N}}\exp\left(-\frac{\left(X-S\right)^{2}}{\sigma_{N}^{2}}\right), \tag{7}$$

$$p\left(S\right)=\frac{1}{\sqrt{2\pi}\sigma_{S}}\exp\left(-\frac{S^{2}}{\sigma_{S}^{2}}\right). \tag{8}$$

2.2 Model 2: Gaussian noise and gamma speech

The conditional density is same as in (6) and the prior density is noted by

$$p\left(S\right)=\frac{b_{S}a_{S}^{b_{S}-1}}{2\sigma_{S}^{b}\Gamma\left(a_{S}\right)}\left|S\right|^{a_{S}-1}\exp\left(-b_{S}\frac{\left|S\right|}{\sigma_{S}}\right), \tag{9}$$

where (a_S, b_S) is parameter set of speech spectral modeling.

2.3 Model 3: Gama noise and gamma speech

The conditional density is expressed by

$$p\left(X|S\right)=\frac{b_N a_N^{b_N-1}}{2\sigma_S^{b_N}\Gamma\left(a_N\right)}\left|X-S\right|^{a_N-1}\exp\left(-b_N\frac{\left|X-S\right|}{\sigma_S}\right), \tag{10}$$

where(a_N, b_N) is parameter set of noise spectral modeling.

3 MMSE estimation

In general, the MMSE estimation minimizes the second order moment of the residual and is denoted by

$$\hat{S}=\arg\min_S\left(E\left[\left(\hat{S}-S\right)^2|X\right]\right). \tag{11}$$

Eq. (11) yields the estimation equal to the conditional expectation, which is expressed as follows:

$$\hat{S}=E\left[S|X\right]=\frac{\int_{-\infty}^{\infty}Sp\left(X,S\right)dS}{p\left(X\right)}=\frac{\int_{-\infty}^{\infty}Sp\left(X|S\right)p\left(S\right)dS}{\int_{-\infty}^{\infty}p\left(X|S\right)p\left(S\right)dS}, \tag{12}$$

where the conditional distribution $p\left(X|S\right)$ is given by (7) or (10) and the prior distribution $p\left(S\right)$ is given by (8) or (9). It is well known that [4], the MMSE for the cases of Gaussian modeling of both noise and speech spectral yields the conventional Wiener filtering:

$$\hat{S}=\frac{\sigma_S^2}{\sigma_S^2+\sigma_N^2}X. \tag{13}$$

Recently, the MMSE estimation of non-Gaussian modeling of speech and noise spectral has been studied by R.Martin and his colleagues. The MMSE estimation based on the gamma modeling of speech spectral components (model 2 and 3) are investigated in particularly cases of Laplacian ($a_S=1, b_S=\sqrt{2}$) and a special case where $a_S=1, b_S=\sqrt{\frac{3}{2}}$. The Laplacian modeling of noise spectral component is also taken into account. The closed form solution of these estimation methods are shown in [5], [6]. However, for other set of parameters, the MMSE estimation in (10) becomes to highly complicated and inconvenience even for the numerical methods. At other hand, the prior parameters of speech and noise spectral components must be dependent on each the particular environment condition and noise estimation method. Therefore, the improvement of these approaches is in applicable.

4 MAP estimation

On the basic of above considerations, it is reasonable to develop the estimation method for the general case of the gamma modeling with arbitrary distribution parameter set. In this work we develop the MAP estimation for spectral components for both the cases of Gaussian and gamma modeling of noise. First, we denote the MAP estimation in general form:

$$\hat{S} = \arg\max_{S} \log\left(p\left(S|X\right)\right) = \arg\max_{S} \log\left(p\left(X|S\right)p\left(S\right)\right). \tag{14}$$

This expression yields a equation in derivate and is noted by

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right) + \log\left(p\left(S\right)\right)\right] = 0. \tag{15}$$

Since the MAP estimation for the case of Gaussian modeling of both noise and speech (model 1) yields the same results as MMSE in (13) (i.e. Wiener filter), we begin this section with the model 2.

4.1 Model 2: Gaussian/Gamma

For the model 2, the conditional and prior distributions are as follows:

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right)\right] = \frac{X-S}{\sigma_N^2}, \tag{16}$$

$$\frac{\partial}{\partial S}\left[\log\left(p\left(S\right)\right)\right] = \frac{\left(a_S-1\right)}{S} - sign\left(S\right)\frac{b_S}{\sigma_S}. \tag{17}$$

Making an assumption

$$sign\left(S\right) = sign\left(X\right), \tag{18}$$

and substituting (16) and (17) into (15) yields the estimation as follows:

$$\frac{X-S}{\sigma_N^2} + \frac{\left(a_S-1\right)}{S} - sign\left(X\right)\frac{b_S}{\sigma_S} = 0. \tag{19}$$

Equation (19) can be transformed into following:

$$G^2 - G\left(1 - \frac{sign\left(X\right)b_S}{\sqrt{\gamma\xi}}\right) + \frac{\left(a_S-1\right)}{\gamma} = 0, \tag{20}$$

where: $G = \frac{S}{X}$ is gain function, $\gamma = \frac{X^2}{\sigma_N^2}$, $\xi = \frac{\sigma_S^2}{\sigma_N^2}$ are posterior and prior SNRs. Finally, the solution of (14) is given as follows:

$$G = \max\left\{u \pm \sqrt{u^2+v}, 0\right\}, \tag{21}$$

where:

$$u = \left(0.5 - \frac{sign\left(X\right)b_S}{\sqrt{4\gamma\xi}}\right), \tag{22}$$

$$v = \frac{\left(a_S-1\right)}{4\gamma}. \tag{23}$$

4.2 Model 3: Gamma/Gamma

For the model 3, the conditional distributions are expressed by

$$\frac{\partial}{\partial S}\left[\log\left(p\left(X|S\right)\right)\right] = -\frac{\left(a_N - 1\right)}{X - S} - sign\left(X - S\right)\frac{b_N}{\sigma_N}. \tag{24}$$

Substituting (24) and (17) into (15) yields the estimation equation as follows:

$$-\frac{\left(a_N - 1\right)}{X - S} + sign\left(X\right)\frac{b_N}{\sigma_N} + \frac{\left(a_S - 1\right)}{S} - sign\left(X\right)\frac{b_S}{\sigma_S} = 0. \tag{25}$$

After making a assumption

$$sign\left(X - S\right) = sign\left(X\right), \tag{26}$$

the estimation equation is transformed to following:

$$G^2 - G\left(1 - \frac{\left(a_S + a_N - 2\right)}{\sqrt{\gamma}sign\left(X\right)\left(b_N - \frac{b_S}{\sqrt{\xi}}\right)}\right) + \frac{\left(a_S - 1\right)}{\sqrt{\gamma}sign\left(X\right)\left(b_N - \frac{b_S}{\sqrt{\xi}}\right)} = 0. \tag{27}$$

The solution of (27) is given by the same as in the (20).

5 Cumulative distribution function equalization

5.1 Cumulative distribution function equalization

The cumulative distribution function equalization (CDFE) method was originally named as histogram equalization and was developed for the speech recognition adaptation approaches [7], where the cumulative histograms derived from available data is used. The principle of this method is finding a non-linear transform from the noisy signal to the clean features, which provide a zero (absolute minimum) distance between their cumulative distribution function. Denoting the equalization in general note as below:

$$\widehat{s} = g\left(x\right), \tag{28}$$

the criteria to estimate here is expressed as follows:

$$F_{g(x)}\left(g\left(x\right)\right) = F_s\left(s\right). \tag{29}$$

The key point of the method is that, the cumulative distribution function (CDF) of an arbitrary function of a variable is equal to it's deft CDF i.e.

$$F_{g(x)}\left(g\left(x\right)\right) = F_x\left(x\right). \tag{30}$$

From (29) and (30), the non-linear transform is given by the equalization of the prior CDFs of noisy and clean signals.

$$g\left(x\right) = F_s^{-1}\left(F_x\left(x\right)\right). \tag{31}$$

In this works the CDF of noisy speech and speech spectral components are used in (31) and the CDFE is carried out for each frame-frequency index. In contrast to the histogram equalization we develop a model based approach by using the distribution modeling in section 2.

5.2 Model 1: Gauss/Gauss

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

$$X \sim \mathrm{N}\left(0, \sigma_S^2 + \sigma_N^2\right) \tag{32}$$

In that case, both CDF $F_X\left(.\right), F_S\left(.\right)$ is Gaussian and therefore the CDFE operation is carried out without any difficulties.

5.3 CDF estimation via the characteristic function

The main problem of CDF approaches is that the CDF of noisy speech which is presented as a superposition of the speech and noise is highly difficult to estimate with the non-Gaussian modeling. Here we develop a CDF estimation method via the characteristic function (CF), which is defined as the Fourier transform of PDF and noted by

$$f\left(u\right) = \int_{-\infty}^{\infty} p\left(x\right) e^{iux} dx \tag{33}$$

Taking into account the property of Fourier transform the cumulative distribution function can be denoted by

$$F\left(x\right) = \frac{1}{2} - sign\left(x\right) \frac{2}{\pi} \int_{0}^{\infty} \frac{R\left(u\right)}{u} \sin\left(A\left(u\right) - ux\right) du. \tag{34}$$

The numerical integration following the next expression is used for the cumulative distribution function estimation:

$$F\left(x\right) = \begin{cases} 1 & x \geq m + 4\sigma \\ \frac{1}{2} - sign\left(x\right) \frac{2}{\pi} \int_{\varepsilon}^{2\pi} \frac{R(u)}{u} \sin\left(A\left(u\right) - ux\right) du & m + 4\sigma > x > m - 4\sigma \\ 0 & x \leq m - 4\sigma \end{cases} \tag{35}$$

The main point here is that the CF of additive model (5) is multiple and therefore, it has closed form when the noise and speech are modelled by Gaussian or gamma distribution and consequently the CDF is always defined. Note that, since the PDF of spectral components are assumed to be symmetrical, their CF is always real i.e. the phase part is equal to zero.

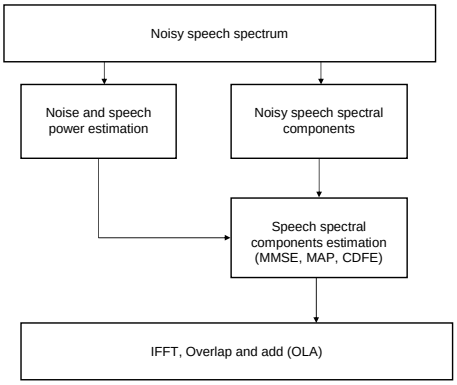

Fig. 1. Noise suppression filter

5.4 Model 2: Gauss/Gamma

Denote the CF of Gaussian noise and double gamma speech spectral components as follows:

$$f_N(u) = e^{-\frac{u^2\sigma_N^2}{2}}, \tag{36}$$

$$f_S(u) = Re\left[\left(\frac{a_S}{a_S - iu}\right)^{b_S}\right] = \cos\left(b_S a \cos\left(\frac{a_S^2}{a_S^2 + u^2}\right)\right), \tag{37}$$

the CF of noisy speech spectral component is given by multiplying (36) to (37). Substituting it into (35) and consequently into (31) the CDFE is given.

5.5 Model 3: Gamma/Gamma

Analogously, the CF of the noisy speech in the case of model 3 is given by following:

$$f_X(u) = \cos\left(b_N a \cos\left(\frac{a_N^2}{a_N^2 + u^2}\right)\right)\cos\left(b_S a \cos\left(\frac{a_S^2}{a_S^2 + u^2}\right)\right). \tag{38}$$

The CDF and consequently the CDFE is given by (33) and (31).

6 Experiment

The diagram of processing is shown in Figure 1. The noisy speech spectral components are given from the complex spectrum. The noise power is estimated following the conventional minimum statistic method [8]. The prior SNR is calculated by "decision directed" method presented in [2]. The car and babble noise conditions are investigated in this work. The speech data are given from AURORA 2 under four SNR conditions 0dB, 5dB, 10dB and 15 dB. The gamma distribution parameters of speech and noise are estimated offline using the training data for both noise conditions. The following set of parameters are given from our experiments: $a_S = 1.5$ and $a_N = 0.6$ for the car noise condition and $a_S = 2$

Table 1. Car noise condition-Segmental SNR improvements evaluation.

Car noise	0dB	5dB	10dB	15dB
Gauss/Gauss				
MMSE-MAP	7.13	5.22	4.09	2.12
CDFE	7.43	5.49	4.26	2.72
Gauss/Gamma				
MMSE	7.74	5.45	4.68	3.05
MAP	8.86	6.81	5.26	3.21
CDFE	8.55	6.72	5.33	3.65
Gamma/Gamma				
MMSE	7.24	5.12	3.89	2.04
MAP	8.12	6.71	5.13	3.07
CDFE	8.34	6.55	5.54	3.43

$a_N = 1.2$ for the babble noise condition. These parameter sets are being used in model 2 and model 3 of our proposed estimation methods. For the reference, the MMSE proposed in [5],[6] is also implemented, where the Laplacian modeling is used. The results of the segmental SNR improvements for the car noise and babble noise conditions are in tables 1-2.The silent durations are deleted in the segmental SNR evaluations by setting a threshold of -10dB at each frame SNR index. From the tables, the Gaussian/Gamma model for the car noise condition and the Gamma/Gamma model for the Gamma/Gamma model perform best. By these results, the Gaussian model is more suitable for car noise when the gamma model is more suitable for the babble noise condition. However in both cases, the gamma model of speech performs better than conventional Gaussian. In side each group, the performances of MAP and CDFE methods overcome the MMSE, where the relative improvements up to 1,8dB is obtained. The CDFE based method is best for the SNR condition 10dB and 15dB and in all cases when produces less musical like noise level.

7 Conclusion

We develop the MAP and cumulative distribution equalization techniques for the speech spectral estimation problems with application in the noise suppression filtering. The advantage of proposed method is that yields the solution in general case of gamma modeling of both speech and noise spectral components and therefore it is applicable to combine with the optimal prior modeling, which should be determined in each particular condition, to improve the performance of system. The experimental evaluation of the noise suppression filtering systems based on proposed methods are shown to be better than conventional MMSE method, where the fixed parameters are used. Among the proposed methods, the MAP is more simple and yields in general better results in SNR improvement

Table 2. Babble noise condition-Segmental SNR improvements evaluation.

Babble noise	0dB	5dB	10dB	15dB
Gauss/Gauss				
MMSE-MAP	5.14	4.01	3.75	1.12
CDFE	5.31	4.52	4.26	1.26
Gauss/Gamma				
MMSE	5.34	4.34	3.98	1.54
MAP	6.12	5.74	4.25	2.43
CDFE	5.75	5.70	4.76	3.01
Gamma/Gamma				
MMSE	5.04	4.41	3.95	1.86
MAP	7.45	6.02	5.26	2.89
CDFE	7.23	6.21	5.48	3.12

indexes. However, the CDFE method products less musical noise effect. In the future works, we will investigate the evaluation of proposed systems in the speech recognition indexes.

References

1. Y. Ephraim, and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP*,vol. ASSP-32, pp. 1109-1121,1984.
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. K. Paliwal, and L. Alsteris,"Usefulness of Phase in Speech Processing," *JNRSAS*, Acrticle 2, 2004.
4. T.H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," *Accepted to be presented at ICASSP2005.*
5. R. Martin, B. Colin, "Speech enhancement in DFT domain using Laplacian priors," *in Proc. IWAENC*,Kyoto, Japan, 2003.
   egura
6. R. Martin, "Speech enhancement using MMSE Short Time Spectral Estimation with Gamma Speech Prior,"*in Proc.ICASSP 02*,Orlando Florida,USA, 2002.
7. C. S, C. Bentez, A. de la Torre, A. Rubio and J. Ramrez, "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Signal Processing Letters,* 11(5), May 2004.
8. R. Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics" *IEEE Trans. ASSP*, Vol. 9, No.5, pp.504-512, 2001.