IDIAP RESEARCH REPORT

A MAP APPROACH TO NOISE COMPENSATION OF SPEECH

Philip N. Garner

Idiap-RR-08-2009

JUNE 2009

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

A MAP Approach to Noise Compensation of Speech

Philip N. Garner, *

February, 2009
Revised: June 9, 2009

Abstract

We show that estimation of parameters for the popular Gaussian model of speech in noise can be regularised in a Bayesian sense by use of simple prior distributions. For two example prior distributions, we show that the marginal distribution of the uncorrupted speech is non-Gaussian, but the parameter estimates themselves have tractable solutions. Speech recognition experiments serve to suggest values for hyper-parameters, and demonstrate that the theory is practically applicable.

Keywords: Speech processing, noise.

1 Introduction

An important problem encountered in speech signal processing is that of how to normalise a signal for the effects of additive noise. In speech enhancement the task is to remove noise from a signal to reproduce the uncorrupted signal such that it is perceived by a listener to be less noisy. In Automatic Speech Recognition (ASR), the task is to reduce the effect of additive noise on recognition accuracy.

Many common practical solutions are based on assumption of a simple additive Gaussian model for both speech and noise in the spectral domain. In ASR, the spectral subtraction approach of Boll [1] is well established, and often used as a means to derive a Wiener filter. In speech enhancement, much work is based on the technique of Ephraim and Malah [2].

More recently, there has been much interest in super-Gaussian distributions as models for speech [3, 4]. Given such a model, it has been shown that spectral subtraction can still generate state of the art results in ASR [5].

In this paper, we present a novel analysis approach for speech in noise. We show that simple priors placed on the parameters of a Gaussian model can lead to super-Gaussian marginal distributions. Furthermore, computationally simple maximum a-posteriori (MAP) additive noise removal solutions exist for the two cases considered.

2 Mathematical framework

2.1 Gaussian model

Let us assume that a DFT operation produces a vector, $\mathfrak{x}$, with complex components, $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_F$, where the real and imaginary parts of each $\mathfrak{x}_f$ are i.i.d. normally distributed with zero mean and variance υ_f. That is,

$$f\left(\mathfrak{x}_f \mid \upsilon_f\right) = \frac{1}{\pi\upsilon_f} \exp\left(-\frac{|\mathfrak{x}_f|^2}{\upsilon_f}\right). \tag{1}$$

In the case where we distinguish two coloured noise signals, a background noise, $\mathfrak{n}$, and a signal of interest, $\mathfrak{s}$, typically speech, denote the noise variance as $\boldsymbol{\nu}$ and the speech variance as $\boldsymbol{\varsigma}$. In general, the background noise can be observed in isolation and modelled as

$$f\left(\mathfrak{n}_f \mid \nu_f\right) = \frac{1}{\pi\nu_f} \exp\left(-\frac{|\mathfrak{n}_f|^2}{\nu_f}\right). \tag{2}$$

*This work was supported by the Swiss National Center of Competence in Research on Interactive Multi-modal Information Management. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein. The author is grateful to Fabio Valente for comments on the content and structure of the document.

The speech, however, cannot normally be observed in isolation. It is always added to noise. When both speech and additive noise are present the variances add, meaning that the total signal, $\mathfrak{t}_f = \mathfrak{s}_f + \mathfrak{n}_f$, can be modelled as

$$f\left(\mathfrak{t}_f \mid \varsigma_f, \nu_f\right) = \frac{1}{\pi(\varsigma_f + \nu_f)} \exp\left(-\frac{|\mathfrak{t}_f|^2}{\varsigma_f + \nu_f}\right). \tag{3}$$

The above model is the basis of the Wiener filter and of the widely used Ephraim-Malah speech enhancement technique [2]. The goal is usually formulated as requiring an estimate of $\mathfrak{s}_f$; this proceeds via estimation of ς_f.

2.2 Parameter estimation

We seek an expression for the speech variance, ς_f, in terms of the observable variable $\mathfrak{t}_f$. In the following, we drop the subscript for simplicity, assuming that all expressions apply to a single bin of a single DFT.

The first step is to note that the expression will depend upon the noise variance, and introduce it as a nuisance parameter

$$f\left(\varsigma \mid \mathfrak{t}\right) = \int_0^\infty d\nu\, f\left(\varsigma, \nu \mid \mathfrak{t}\right). \tag{4}$$

The second is to apply Bayes's theorem to the term inside the integral,

$$f\left(\varsigma \mid \mathfrak{t}\right) = \int_0^\infty d\nu\, \frac{f\left(\mathfrak{t} \mid \varsigma, \nu\right) f\left(\varsigma, \nu\right)}{\int_0^\infty d\varsigma' d\nu'\, f\left(\mathfrak{t} \mid \varsigma', \nu'\right) f\left(\varsigma', \nu'\right)}. \tag{5}$$

Before attempting to solve 5, we apply three practical simplifications:

1. We assume that ς is dependent upon ν such that

$$f\left(\varsigma, \nu\right) = f\left(\varsigma \mid \nu\right) f\left(\nu\right). \tag{6}$$

2. We assume that an estimate, $\hat{\nu}$, of ν is available via solution of 2 during non-speech segments of the signal. This reduces $f\left(\nu\right)$ to a delta function, removing the need to integrate.

3. Given that we seek an estimate rather than a distribution, the denominator can be ignored.

The problem then reduces to the maximisation

$$\hat{\varsigma} = \max_{\varsigma} f\left(\mathfrak{t} \mid \varsigma, \hat{\nu}\right) f\left(\varsigma \mid \hat{\nu}\right). \tag{7}$$

3 Three estimates of speech variance

3.1 Maximum Likelihood

If we substitute 3 into 7, and ignore the second term of 7, the solution is

$$\hat{\varsigma} = \begin{cases} |\mathfrak{t}|^2 - \hat{\nu}, & \text{if } |\mathfrak{t}|^2 > \hat{\nu}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This is the well known Maximum Likelihood estimate. It is known to provide a "reasonable" estimate of the speech variance, but always requires regularisation. In the Ephraim-Malah technique, the ML estimate is regularised using the decision-directed estimator, which forms a linear combination of the ML estimate, and an expression derived from the previous frame. In ASR, the ML estimate is known as power spectral subtraction. It is regularised by means of an over-subtraction factor, α, and a flooring factor, β:

$$\hat{\varsigma} = \begin{cases} |\mathfrak{t}|^2 - \alpha\hat{\nu}, & \text{if } |\mathfrak{t}|^2 - \alpha\hat{\nu} > \beta\hat{\nu}, \\ \beta\hat{\nu} & \text{otherwise.} \end{cases} \tag{9}$$

3.2 Maximum a-Posteriori: inverse gamma prior

The inverse gamma distribution, which we parameterise here as

$$f\left(x \mid \alpha, \beta\right) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \tag{10}$$

is a common prior for a variance. This is normally for reasons of conjugacy. In this situation, the inverse gamma is not a conjugate prior because of the extra noise variance term. Nevertheless, it is possible to derive a MAP estimate of the speech variance:

$$I = f\left(\mathfrak{t} \mid \varsigma, \hat{\nu}\right) f\left(\varsigma \mid \hat{\nu}\right) \tag{11}$$

$$= \frac{1}{\pi(\varsigma + \hat{\nu})} \exp\left(-\frac{|\mathfrak{t}|^2}{\varsigma + \hat{\nu}}\right) \tag{12}$$

$$\times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \varsigma^{-\alpha-1} \exp\left(-\frac{\beta}{\varsigma}\right) \tag{13}$$

$$\frac{\partial \log I}{\partial \varsigma} = -\frac{1}{\varsigma + \hat{\nu}} + \frac{|\mathfrak{t}|^2}{(\varsigma + \hat{\nu})^2} - \frac{\alpha + 1}{\varsigma} + \frac{\beta}{\varsigma^2} = 0. \tag{14}$$

$\hat{\varsigma}$ is then a solution of the cubic

$$-(\alpha+2)\varsigma^3 + \left(|\mathfrak{t}|^2 - 2(\alpha+1)\hat{\nu} - \hat{\nu} + \beta\right)\varsigma^2 + \left(2\beta\hat{\nu} - (\alpha+1)\hat{\nu}^2\right)\varsigma + \beta\hat{\nu}^2 = 0. \tag{15}$$

3.3 Maximum a-Posteriori: gamma prior

The gamma distribution, parameterised here as

$$f\left(x \mid \alpha, \beta\right) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \tag{16}$$

is not normally used as a prior for a variance as it is not conjugate. However, it has been suggested (e.g., Gazor and Zhang [6]) that such a distribution is a good fit for speech under some circumstances. We show here that it leads to a MAP solution similar to that of the inverse gamma prior. The MAP solution proceeds as

$$I = f\left(\mathfrak{t} \mid \varsigma, \hat{\nu}\right) f\left(\varsigma \mid \hat{\nu}\right) \tag{17}$$

$$= \frac{1}{\pi(\varsigma + \hat{\nu})} \exp\left(-\frac{|\mathfrak{t}|^2}{\varsigma + \hat{\nu}}\right) \tag{18}$$

$$\times \frac{\beta^{\alpha}}{\Gamma(\alpha)} \varsigma^{\alpha-1} \exp\left(-\frac{\varsigma}{\beta}\right) \tag{19}$$

$$\frac{\partial \log I}{\partial \varsigma} = -\frac{1}{\varsigma + \hat{\nu}} + \frac{|\mathfrak{t}|^2}{(\varsigma + \hat{\nu})^2} + \frac{\alpha - 1}{\varsigma} - \frac{1}{\beta} = 0. \tag{20}$$

$\hat{\varsigma}$ is again the solution to a cubic:

$$-(1/\beta)\varsigma^3 + (\alpha - 2\hat{\nu}/\beta - 2)\varsigma^2 + \left(2(\alpha-1)\hat{\nu} + |\mathfrak{t}|^2 - \hat{\nu}^2/\beta - \hat{\nu}\right)\varsigma + (\alpha-1)\hat{\nu}^2 = 0. \tag{21}$$

3.4 Discussion

For each MAP case, the solution $\hat{\varsigma}$ results from solving a cubic. An analytic solution to the cubic is readily available [7]; the first solution, which is guaranteed to be real, is the appropriate one. One important empirical observation is that the solution in the inverse gamma case is not numerically stable. We find, however, that a Newton-Raphson solution is stable, the larger of $|\mathfrak{t}|^2$ and $\hat{\nu}$ being a suitable starting value for the iteration.

4 Experiments

4.1 Re-parameterisation

In the ML case, the two parameters are somewhat arbitrary numbers. This is also true of the α parameter for the MAP cases. Empirically, however, we find the following arguments helpful to set the prior parameter β:

The expectation of the gamma PDF can be written

$$E(x) = \alpha\beta. \tag{22}$$

The expectation can also be written heuristically as a Signal to Noise Ratio (SNR), ω, times the noise level. The suggests a constraint

$$\beta = \omega\hat{\nu}/\alpha. \tag{23}$$

The expectation of the inverse gamma PDF is not defined for small values of α. However, the mode is

$$\hat{x} = \frac{\beta}{\alpha + 1}. \tag{24}$$

By a similar argument, we can write

$$\beta = (\alpha + 1)\delta\hat{\nu}, \tag{25}$$

where δ is related to SNR.

We note briefly that if we were to maximise, say, $\sqrt{\varsigma}$ or $\log\varsigma$, which may be more appropriate perceptually, the Jacobian is simply a power of ς so the change of variable would only result in an additive offset to α. Optimising α hence affords some independence of such domain. However, the domain also changes the meaning of ω and δ.

4.2 ASR experiments

To show the feasibility of the two new estimators, and to find reasonable values of the hyper-parameters, we give results in the form of ASR performance. Given that spectral subtraction is often used in ASR, we suppose that the speech variance, ς, of which we have shown spectral subtraction to be an estimate, is a suitable feature.

The aurora 2 task [8] is a well known evaluation for noise compensation techniques. It is a simple digit recognition task with real noise artificially added in 5dB increments such that performance without noise compensation ranges from almost perfect to almost random. We used a simple "MFCC" front-end with a 256 point DFT every 10ms, 23 mel bins and 13 cepstral coefficients (including C0) plus first and second order delta coefficients. The MFCCs were normalised with an adaptive cepstral mean subtraction. The (hyper-)parameters of each noise compensation technique were coarsely optimised by hand using the figure of merit for test set "A" from aurora 2: $\alpha = 1.0$, $\beta = 0.1$ for maximum likelihood, $\alpha = -0.99$, $\delta = 1.0$ for the inverse gamma prior and $\alpha = 1.35$, $\omega = 15.4$ for the gamma prior. Results shown are those of test set "B" with these parameters. The noise variance, $\hat{\nu}$, was estimated as the mean of the first 10 frames of each utterance.

Observe that all compensation techniques give an improvement over the baseline, but there is little to choose between them. We stress that these results are not state of the art for this database; they just serve to compare the estimators and suggest values for hyper-parameters.

5 Analysis

5.1 Comparison with ML solution

The transfer functions of the three estimators are shown in Fig. 2. The abscissa represents an input signal power, $|\mathfrak{t}|^2$, and the ordinate represents an output estimated speech variance, $\hat{\varsigma}$, for a fixed noise variance $\hat{\nu} = 0.2$. The parameters for the curves are (approximately) those found empirically in section 4.2. Notice that both priors give a result qualitatively similar to spectral subtraction, mimicking the concepts of over-subtraction (for high signal power) and flooring (for low signal power). Remarkably for the empirically derived parameters, the inverse gamma prior behaves very like spectral subtraction (a very mild over-subtraction is visible on suitable axes). The gamma prior yields a square-root compression, with much more aggressive over-subtraction and flooring.

5.2 Histogram

Given a large database of speech recorded in a quiet environment, it is possible to plot a detailed histogram representing the distribution of any given DFT bin. In plotting spectral power, we are plotting some estimate of $f\left(|\mathfrak{t}|^2\right)$. We can superimpose upon this plot the same marginal distributions that would be implied by the two choices of prior above. To find these distributions, we proceed in the following stages:

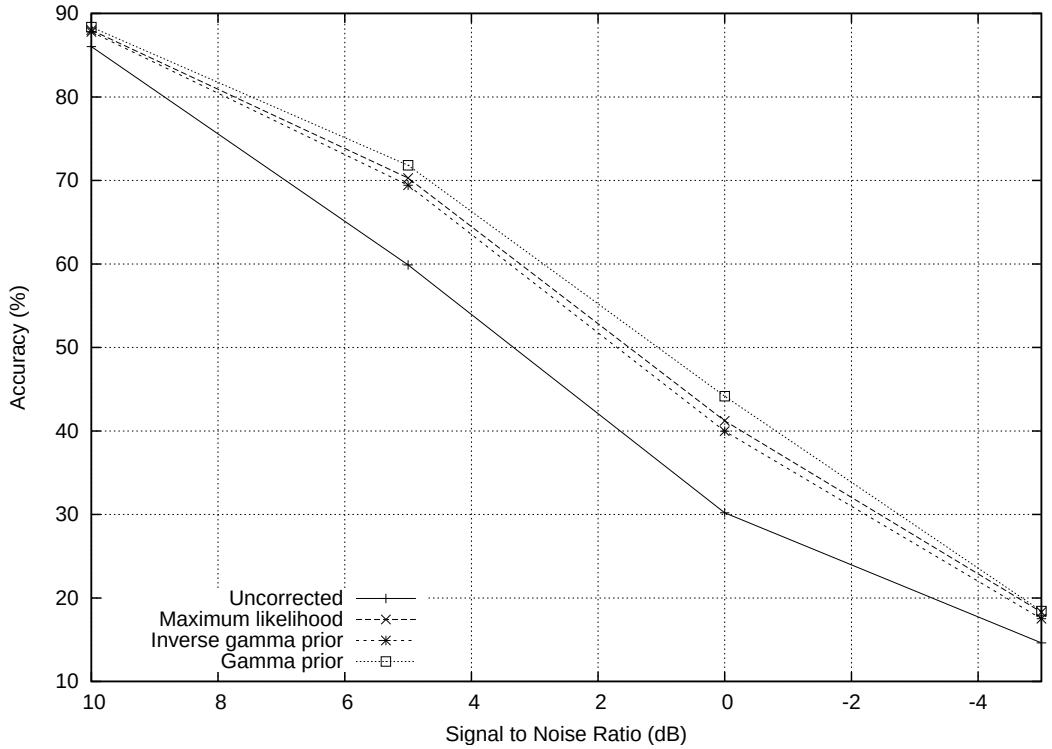

Figure 1: ASR performance of a baseline and the three speech variance estimates on test set B of aurora 2 (clean training). The lines are indistinguishable for SNR $>$ 15dB.

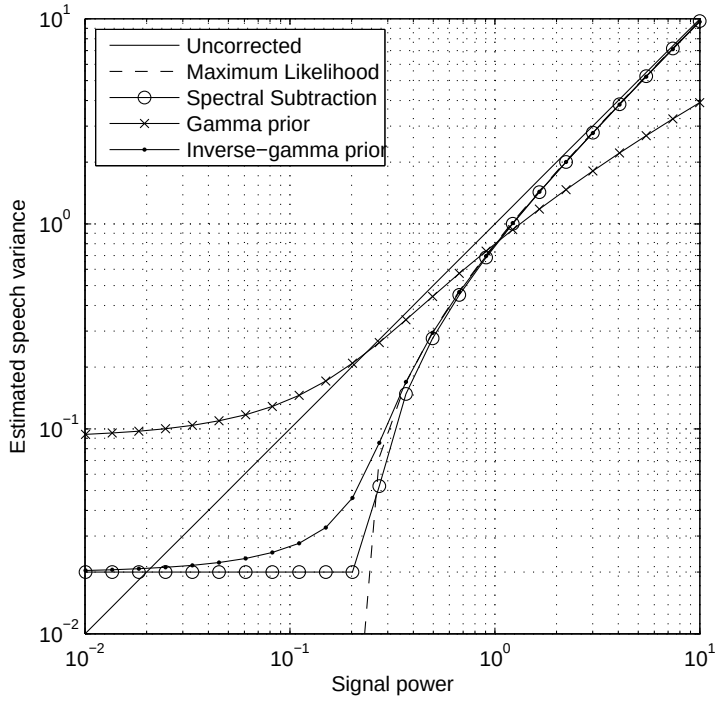

Figure 2: Transfer functions of estimators of $\hat{\varsigma}$ for a fixed noise level $\hat{\nu} = 0.2$.

1. We make a change of variable $p = |\mathrm{t}|^2$ in the expression for the spectral observation 3. This well known operation yields the exponential distribution

$$f\left(p \mid \varsigma, \nu\right) = \frac{1}{\varsigma + \nu} \exp\left(-\frac{p}{\varsigma + \nu}\right). \tag{26}$$

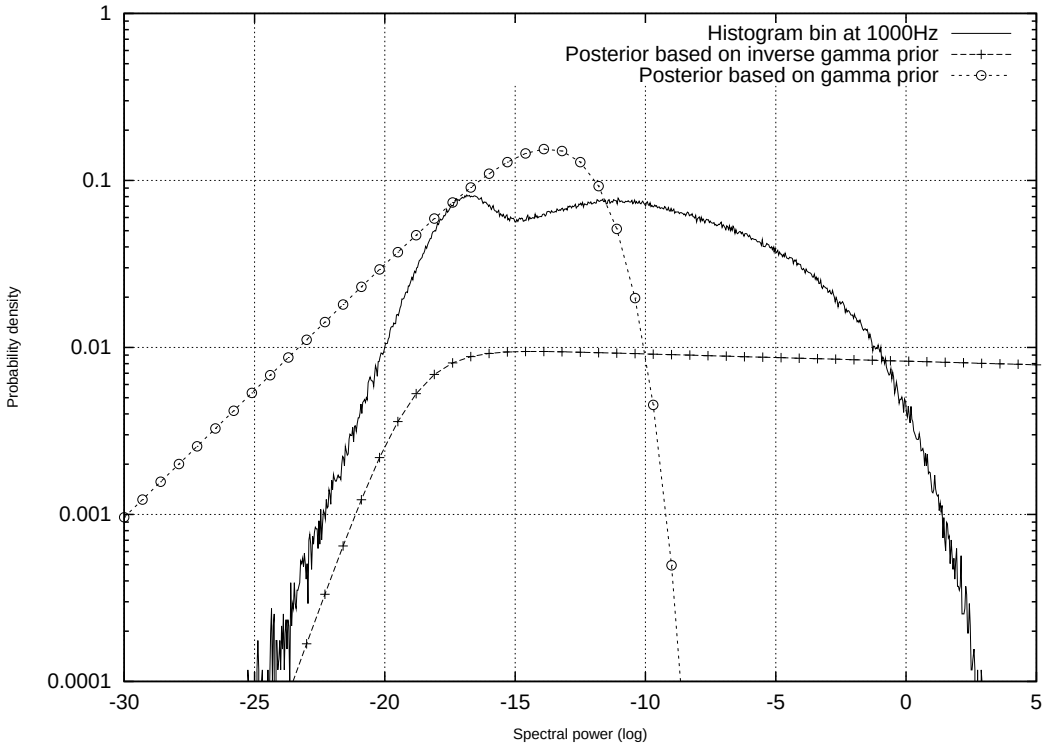

Figure 3: Histogram of a DFT bin for a large database of speech, plus implied marginal distributions using the given priors.

2. We assume that ν is close to zero, as in uncorrupted speech.

3. We write

$$f(p) = \int_0^\infty d\varsigma \, f(p \mid \varsigma) \, f(\varsigma) \, . \tag{27}$$

Notice the similarity with the denominator of 5; the approach is related to the evidence framework of MacKay [9].

In the case of the inverse gamma prior, it is shown in appendix A.1 that

$$f(p) = \alpha \beta^\alpha (p + \beta)^{-1-\alpha}, \tag{28}$$

i.e., the conjugate prior yields a simple result. For the gamma prior, the equivalent result is shown in appendix A.2 to be

$$f(p) = \frac{1}{\beta^{\frac{\alpha+1}{2}} \Gamma(\alpha)} p^{\frac{\alpha+1}{2}-1} K_{1-\alpha}\left(2\sqrt{\frac{p}{\beta}}\right). \tag{29}$$

Both the above marginal distributions are super-Gaussian.

A plot of the form described above is shown in Fig. 3, using values of α appropriate for the log domain as discussed in section 4.1. The histogram is taken from the clean part of the aurora 2 training corpus, described above. The hyper-parameter values are as determined above, and $\hat{\nu}$ is an average for the bin at 1000Hz.

Notice that the inverse gamma based marginal distribution models the noise floor quite well, but incorrectly attaches too much probability mass to high values of speech. The fact that $\delta = 1$ suggests that the turning point on the graph is the noise level, and the prior is simply saying that the speech should be above the noise.

Conversely, the gamma based marginal distribution appears to be modelling the speech signal itself, although not particularly well, suggesting that another choice of prior may be better. The bias towards small values explains the higher noise floor observed in Fig. 2.

6 Conclusion

We have derived two new analytic expressions for the variance parameter of a speech power spectrum in noise. One is consistent with a prior model of the minimum power imposed by the background noise level in speech. The other arises from a simple prior model of the speech spectral power itself. In a noise reduction setting, both expressions have been shown to behave qualitatively in a similar manner to power spectral subtraction.

The prior distributions lead to super-Gaussian marginal distributions for the speech power; this is in keeping with recent thinking about the nature of the speech PDF. However, both implied marginal distributions have shortcomings when compared to a measured PDF of real speech. This implies that future work should focus on finding a better prior for the speech.

A Derivations of marginal distributions

A.1 Marginal of exponential variate with inverse-gamma prior

Say we have a variable that is exponentially distributed with parameter θ, and we put an inverse-gamma prior on the parameter. We can then integrate out θ. In doing so, the exponentials disappear giving a completely polynomial distribution.

$$f(x) = \int_0^\infty d\theta \, \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} \exp\left(-\frac{\beta}{\theta}\right) \tag{30}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty d\theta \, \theta^{-\alpha-2} \exp\left(-\frac{x+\beta}{\theta}\right) \tag{31}$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(x+\beta)^{\alpha+1}} \tag{32}$$

$$= \alpha \beta^\alpha (x+\beta)^{-1-\alpha}. \tag{33}$$

A.2 Marginal of exponential variate with gamma prior

Following the above argument, but with a gamma distributed prior, the result is more involved, but still analytic.

$$f(x) = \int_0^\infty d\theta \, \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) \frac{1}{\beta^\alpha \Gamma(\alpha)} \theta^{\alpha-1} \exp\left(-\frac{\theta}{\beta}\right) \tag{34}$$

$$= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty d\theta \, \theta^{\alpha-2} \exp\left(-\frac{\theta}{\beta} - \frac{x}{\theta}\right). \tag{35}$$

This integral is the integral representation of the modified Bessel function ([10], 8.432.7)

$$K_n(yz) = \frac{z^n}{2} \int_0^\infty \exp\left[-\frac{y}{2}\left(t + \frac{z^2}{t}\right)\right] t^{-n-1} dt, \tag{36}$$

with

$$\frac{y}{2} = \frac{1}{\beta}, \qquad z^2 = \beta x, \qquad n = 1 - \alpha. \tag{37}$$

So,

$$f(x) = \frac{1}{\beta^{\frac{\alpha+1}{2}} \Gamma(\alpha)} x^{\frac{\alpha+1}{2}-1} K_{1-\alpha}\left(2\sqrt{\frac{x}{\beta}}\right). \tag{38}$$

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[3] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. III. IEEE, May 2006, pp. 1068–1071, Toulouse, France.

[4] 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 Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, August 2007.

[5] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard, "Unsupervised spectral substraction for noise-robust ASR," in *Proceedings of the 2005 IEEE ASRU Workshop*, December 2005, San Juan, Puerto Rico.

[6] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, July 2003.

[7] "Cubic function," Wikipedia, `http://en.wikipedia.org/wiki/Cubic_function`.

[8] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millenium"*, September 2000, Paris, France.

[9] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, no. 11, pp. 1035–1068, 1999.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. Academic Press, 2000, Alan Jeffrey, Editor.