MMSE SPEECH SPECTRAL AMPLITUDE ESTIMATORS WITH CHI AND GAMMA SPEECH PRIORS

Ioannis Andrianakis and Paul R. White

Institute of Sound and Vibration Research,
University of Southampton,
Southampton SO17 1BJ, UK
{ia,prw}@isvr.soton.ac.uk

ABSTRACT

We present two novel algorithms for optimal MMSE estimation of the speech spectral amplitude, assuming it has been corrupted by additive and uncorrelated noise. The noise DFT coefficients are modelled as Gaussian random variables, while the speech spectral amplitude is modelled using either a Chi or a Gamma distribution. The influence of the priors' shape parameter is investigated. Results from simulations that demonstrate the performance of the proposed algorithms for different noise types and SNRs, as well as a comparison with previously developed MAP estimators are presented.

1. INTRODUCTION

The portability of digital systems with man-machine voice interfaces allow them to be deployed in environments where background noise conditions can be adverse. Background noise poses a serious problem for both voice-based communication and automated services. Speech enhancing algorithms can restore, to some extent, the noise corrupted speech, increasing its quality and potentially its intelligibility. The success rate of speech recognition engines can also be improved.

Most modern speech enhancement algorithms operate in the frequency domain. The transformation to the frequency domain is performed with the Short Time Fourier Transform (STFT), typically with windows of 30 ms overlapped by 20 ms. An estimator is applied to approximate the clean speech STFT, which is then transformed to the time domain with the inverse DFT and the overlap and add method.

Bayesian estimators are very popular in estimating the clean speech STFT coefficients. The most frequently used estimators are the Minimum Mean Square Error (MMSE) and the Maximum A Posteriori (MAP), which can be used to estimate either the DFT coefficients of the STFT (real and imaginary part) or their amplitude. Some assumptions about the distribution of the noise and speech STFT coefficients must also be made. A typical example is the Wiener filter where noise and speech DFT coefficients are assumed to be Gaussian and the estimator applied is the MMSE [1]. Based on the observation that the speech DFT coefficients are better modelled by a Gamma distribution, Martin in [1] developed an MMSE estimator with Gamma priors for speech.

A well known method of amplitude estimation is given by Ephraim and Malah [2], who developed an amplitude MMSE estimator assuming that the noise DFT coefficients are Gaussian and the speech amplitude coefficients follow the Rayleigh distribution. Observing that speech amplitude coefficients are better modelled by the Gamma or the Chi distribution, Lotter and Vary [3] and Dat et al. [4] developed the MAP estimators under the above assumptions.

In this paper we introduce the MMSE estimators with the Gamma and Chi speech priors. The performance of both MMSE and MAP algorithms with both priors is evaluated for different types of noise at different signal to noise ratios. The evaluation is also focused on the effect of the value of the shape parameter a of the speech priors (eqs. 2,3).

The organisation of the paper is as follows: In section 2 we demonstrate the statistical model and the speech priors. In section 3 we introduce the MMSE estimators and review the MAP. Section 4 shows the simulation results and section 5 concludes this paper.

2. STATISTICAL MODEL

2.1. Problem formulation

Assume that we observe a noisy speech signal $x(t)$ that is a sum of a speech and noise signal $s(t)$ and $n(t)$, which are uncorrelated. Their representation in the STFT domain is given by:

$$\mathbf{X}(k,l) = \mathbf{S}(k,l) + \mathbf{N}(k,l) \tag{1}$$

where $\mathbf{X}(k,l)$, $\mathbf{S}(k,l)$ and $\mathbf{N}(k,l)$ are the $(k^{\text{th}}, l^{\text{th}})$ samples of the noisy speech, the clean speech and the noise signal's STFT correspondingly. The index k corresponds to the frequency bins and the index l to the time frames of the STFT. The above complex quantities can be expressed as a function of their amplitude and phase as: $\mathbf{X} \equiv Xe^{j\psi}$, $\mathbf{S} \equiv Se^{j\phi}$, $\mathbf{N} \equiv Ne^{j\omega}$

Our objective is to estimate the clean speech spectral amplitude S given the noisy speech amplitude X and phase ψ. The spectral amplitude estimate $\hat{S}$ will then be combined with the noisy phase ψ and inversion of the STFT gives the enhanced speech signal.

The estimation of the clean speech spectral amplitude requires some assumptions about the distributions of $\mathbf{S}$ and $\mathbf{N}$. The DFT coefficients of noise are assumed to have a Gaussian distribution. This assumption is supported by the Central Limit Theorem for stationary noise and simulation results suggest that it is a quite reasonable choice for quasi-stationary noise.

Speech on the other hand, is a highly non-stationary signal and the Gaussian distribution is not necessarily the best model for its STFT coefficients. In this paper we use two different distribution functions to model the speech spectral amplitude. These distributions are quite flexible and their form can be controlled by tuning their shape parameter a.

2.2. Speech priors

We now present the models we use for the speech spectral amplitude (speech priors), which are the Chi and the Gamma.

The functional form of the Chi distribution is given by:

$$p(S) = \frac{2}{\theta^a \Gamma(a)} \, S^{2a-1} \, \exp\left[-\frac{S^2}{\theta}\right] \tag{2}$$

This is a Chi distribution with $2a$ degrees of freedom and scale parameter $\sqrt{\theta/2}$ [5]. Sometimes is known as Generalised Rayleigh distribution. The parameter a influences the shape of the prior at the origin, introducing a pole at zero for $a < 0.5$, while $p(0) = 0$ for $a > 0.5$. The second moment $\mathrm{E}[S^2]$ of the prior is controlled by both θ and a and is $\mathrm{E}[S^2] = \theta a$. Well known instances of this distribution are the Half-Gaussian ($a = 0.5$) and the Rayleigh ($a = 1$).

The Gamma distribution is given by [5]:

$$p(S) = \frac{1}{\theta^a \Gamma(a)} \, S^{a-1} \, \exp\left[-\frac{S}{\theta}\right] \tag{3}$$

The parameter a again influences the behaviour of the distribution at the origin, introducing the pole at zero for $a < 1$. The second moment is given by $\mathrm{E}[S^2] = \theta^2 a(a+1)$. The Gamma distribution simplifies to the Exponential distribution for $a = 1$.

The phase ϕ of the speech STFT is assumed to be uniform i.e. $p(\phi) = \frac{1}{2\pi}$, which can be easily verified by simulations. Additionally, Martin in [1] states that the amplitude and phase are statistically less dependent than the real and imaginary parts of the STFT coefficients, whose dependence was found to be weak. This observation allows us to factorise the joint density of the amplitude and phase as: $p(S, \phi) = p(S)p(\phi)$. Accordingly, the speech STFT coefficients are modelled as complex, circular symmetric random variables.

3. DERIVATION OF THE ESTIMATORS

3.1. MMSE estimators

Based on the modelling assumptions made in section 2 we now derive the MMSE speech spectral amplitude estimators, for the two speech priors presented.

The MMSE estimator is known to be equal to the mean of the posterior density [2]. Applying Bayes theorem and integrating with respect to the phase of speech we get the following expression:

$$\begin{aligned} \hat{S} &= \mathrm{E}[S|X,\psi] = \frac{\int_0^\infty S\, p(X,\psi|S)p(S)\, dS}{\int_0^\infty p(X,\psi|S)p(S)\, dS} \\ &= \frac{\int_0^\infty \int_0^{2\pi} S\, p(X,\psi|S,\phi)p(S)p(\phi)\, dS\, d\phi}{\int_0^\infty \int_0^{2\pi} p(X,\psi|S,\phi)p(S)p(\phi)\, dS\, d\phi} \end{aligned} \tag{4}$$

Given the Gaussian assumption for the distribution of the noise STFT coefficients, $p(X, \psi|S, \phi)$ can be written as:

$$p(X,\psi|S,\phi) = \frac{X}{2\pi\sigma_N^2} \, \exp\left[-\frac{X^2 + S^2 - 2XS\cos(\psi - \phi)}{2\sigma_N^2}\right] \tag{5}$$

where $\mathrm{E}[N^2] \equiv 2\sigma_N^2$.

Substituting eqs. 2 and 5 into 4 and solving the integrals we get the MMSE estimator with the Chi speech priors, which is:

$$\hat{S}_{MMSEChi} = \sqrt{2\sigma_N^2 \, \zeta} \, \frac{\Gamma(a+0.5)}{\Gamma(a)} \, \frac{{}_1F_1(a+0.5;1;\frac{R^2}{2\sigma_N^2}\zeta)}{{}_1F_1(a;1;\frac{R^2}{2\sigma_N^2}\zeta)} \tag{6}$$

where $\zeta = \frac{\theta}{\theta + 2\sigma_N^2}$. For the solution of the above integrals we used eqs. 8.406.3, 8.431.5 and 6.631.1 from [6]. ${}_1F_1(\alpha;\beta;\gamma)$ is the Confluent Hypergeometric Function eq. 9.210.1, [6],[7]. The calculation of ${}_1F_1$ leads to a numerical overflows for large input values. The asymptotic expansion found in eq. 13.5.1 in [8] was used in these cases. The results were then numerically stable for all ranges of input values. The above estimator with $a = 1$ (Rayleigh speech prior) is the one found in the well known Ephraim-Malah algorithm [2].

To obtain the MMSE estimator using the Gamma priors we need to substitute eqs. 3 and 5 into 4 and solve the integrals. The MMSE estimator can be reduced to a form:

$$\hat{S}_{MMSEGamma} = \frac{\Psi(a)}{\Psi(a-1)} \tag{7}$$

where

$$\Psi(\mu) \equiv \int_0^\infty S^\mu \exp\left[-\frac{S^2}{2\sigma_N^2} - \frac{S}{\theta}\right] I_0\left(\frac{SX}{\sigma_N^2}\right) dS \tag{8}$$

where $I_0(x)$ is the modified Bessel function of the first kind and zero order. The above integral has no analytic solution for $\mu \in (-1, \infty)$, which is the range of interest for our problem. To solve this problem we resorted to numerical integration. It turns that the integrand in Ψ is sufficiently smooth to allow convergence in a few iterations of the Adaptive Lobatto Quadrature [9]. Certain values of the parameters in Ψ can still cause numerical issues. In these cases the modified Bessel function was approximated with the formula $I_0(x) = \mathrm{e}^x/\sqrt{2\pi x}$ and then the integral was calculated analytically with eq. 3.462.1 from [6]. The relative error in the value of $\hat{S}_{MMSEGamma}$ due to this approximation was of the order of 10^{-5}

3.2. MAP estimators

Using the statistical model presented in section 2 it is also possible to derive MAP estimators for both speech priors considered here. The MAP estimator is known to be the mode of the posterior density. This can be written as:

$$\hat{S}_{MAP} = \arg\max_S \, \ln\left(p(S|X,\psi)\right) \tag{9}$$

The above estimators can be proven to be [4, 3]

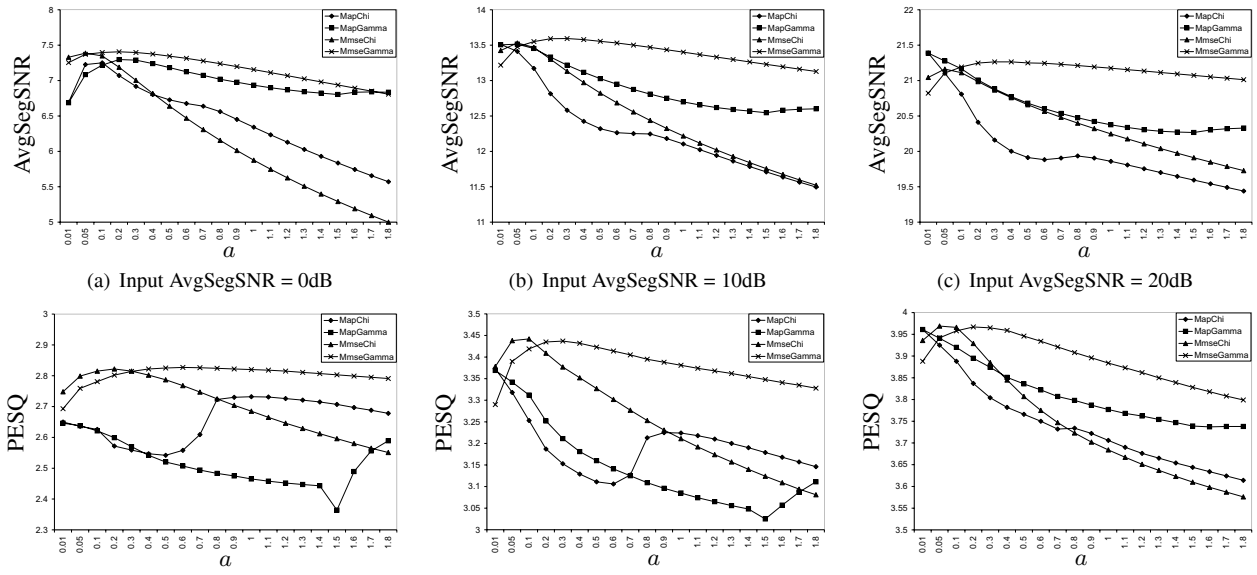
$$\hat{S}_{MAPChi} = \zeta\frac{X}{2} + \left[\left(\zeta\frac{X}{2}\right)^2 + (a - 0.75)\, 2\sigma_N^2\, \zeta\right]^{1/2} \tag{10}$$

where $\zeta = \frac{\theta}{\theta + 2\sigma_N^2}$ and

$$\hat{S}_{MAPGamma} = \zeta + \left[\zeta^2 + (a-1.5)\sigma_N^2\right]^{1/2} \tag{11}$$

where $\zeta = \frac{X}{2} - \frac{\sigma_N^2}{2\theta}$. The MAPGamma algorithm was presented in [3] with $a = 2$, and Dat et al. in [4] used the MAPChi and MAPGamma with $a = 0.5$ and $a = 1.5$ respectively.

There are two issues concerning the MAP estimators. The first is that in maximising analytically the expression in (9) the modified



(a) Input AvgSegSNR = 0dB (b) Input AvgSegSNR = 10dB (c) Input AvgSegSNR = 20dB

Fig. 1. AvgSegSNR and PESQ results from enhancing speech corrupted with white noise.

Bessel function is approximated, possibly introducing some error in the estimation. Secondly, when $a < 0.75$ or $a < 1.5$ for the Chi and Gamma priors correspondingly, the problem of finding the global maximum is no longer well defined. This is because the posterior density with the approximation of the modified Bessel function has a pole at zero. The strategy we follow in this case is to set $\hat{S}_{MAP} = S_{lm}$, where S_{lm} is the value of S where the posterior density with the Bessel approximation has a local maximum, while if the local maximum does not exist we suppress the noisy speech amplitude sample by a fixed amount i.e. $\hat{S}_{MAP} = KX$. In the simulations K was equivalent to 25dB of attenuation.

4. RESULTS

4.1. Simulation setup

To evaluate the performance of the presented algorithms we used 48 sentences from the TIMIT database, uttered by 3 male and 3 female speakers and downsampled at 8KHz. The sentences were corrupted with white noise and car noise at 3 different levels of input average segmental SNR (AvgSegSNR). The performance of the algorithms was evaluated by the output AvgSegSNR and the PESQ, which is the International Telecommunication Union (ITU) recommendation P.862. The STFT transformation was performed with Hamming windows of 256 samples and 75% overlap.

The algorithms were examined for a range of values of the a parameter of the prior densities. The influence of this parameter on the algorithms' performance was then observed. The optimal values of a for this database are then inferred by the algorithms' performance. The value of θ was determined using the a priori SNR $\xi = \mathrm{E}[S^2]/\mathrm{E}[N^2]$, and the expressions that relate the second moment of each prior with θ and a, given in subsection 2.2. This method of estimation of θ enabled the incorporation of the Decision-Directed method [2] for the estimation of the a priori SNR into the algorithms, which helped towards the reduction of the musical noise.

4.2. Evaluation

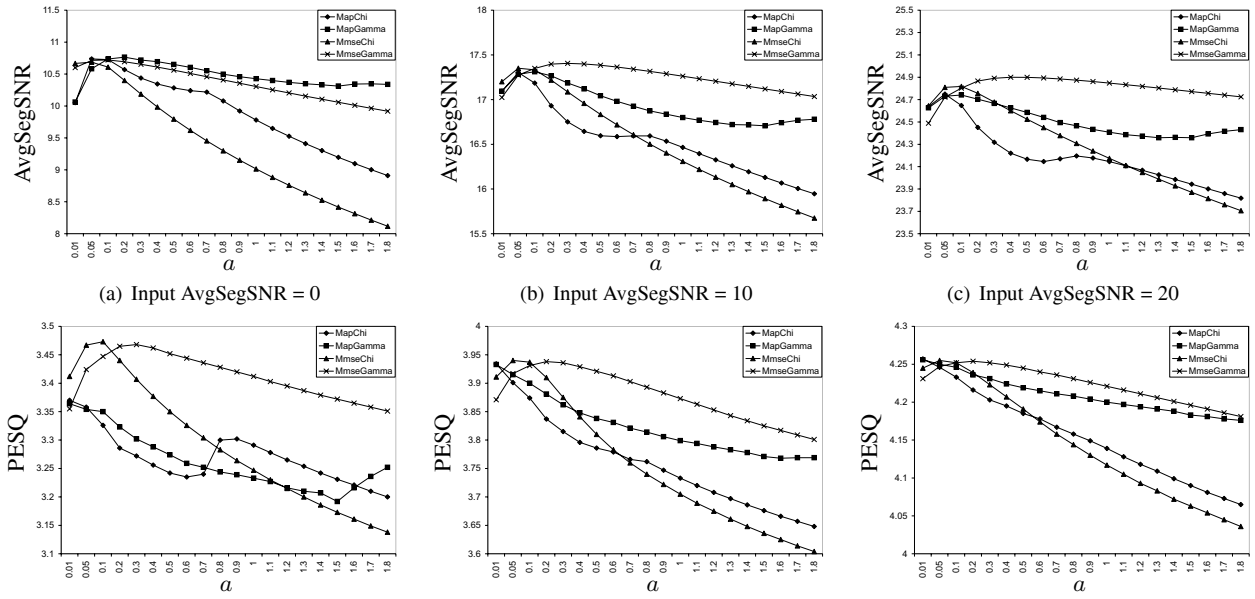
Figure 1 shows the AvgSegSNR and PESQ scores for each algorithm, for different input AvgSegSNRs and values of a. The corrupting noise was white and Gaussian. Figure 2 shows the same results for car noise.

For small values of a the MAPChi algorithm produces musical noise but some weak speech spectral components are preserved. As a increases the musical noise peaks reduce but the weaker speech spectral components are lost. As a increases beyond 0.5 some broadband background noise is added, which is believed to cause the increase in the PESQ scores at low input AvgSegSNR. At high input AvgSegSNR, accurate restoration of the weaker speech spectral components probably weighs more than the nature of the residual noise, which could explain the almost monotonic drop of both AvgSegSNR and PESQ with increasing a.

The behaviour of the MAPGamma algorithm is quite similar to that of the MAPChi but the scores in the objective measures are somewhat better. Although it is not shown in the figures, the PESQ score of the MAPGamma algorithm increases further after $a = 1.8$ for low input AvgSegSNR, reaching the level of the peak of the PESQ curve of the MAPChi, which is found approximately at $a = 0.9$. The MAPGamma algorithm preserves better the weakest speech spectral components than the MAPChi, but the musical noise spectral peaks can be somewhat sharper.

For small values of a both MMSE algorithms produce a residual noise that is much more broadband than the residual noise of the MAP algorithms for the same a. Additionally, this does not affect the preservation of weaker speech spectral components, which is as good as that of their MAP counterparts. A number of musical noise spectral peaks can be present for small a, but significantly less than those generated by the MAP algorithms, and of less intensity. As a increases the few musical noise spectral peaks are eliminated and replaced by broadband noise. The background noise level increases with a and the increase is faster for the MMSEChi algorithm.

MMSEChi and MMSEGamma algorithms consistently produce their best results for $a \in [0.05, 0.2]$ and $a \in [0.2, 0.5]$ respec-



(a) Input AvgSegSNR = 0 (b) Input AvgSegSNR = 10 (c) Input AvgSegSNR = 20

Fig. 2. AvgSegSNR and PESQ results from enhancing speech corrupted with car noise.

tively. The AvgSegSNR scores obtained for these values are among the highest obtained with any algorithm and any input AvgSegSNR. An exception to this is found for the car noise at the 0 dB input AvgSegSNR. The MAP algorithms in this case suppress more the background indeed, but they also suppress the few first speech harmonics, where most of the noise energy was concentrated. The MMSE algorithms on the other hand, retain slightly more background noise, but they also keep the first harmonics intact. This difference in the quality of the resulting speech is demonstrated in the PESQ scores which clearly favour the MMSE enhanced signals. The MMSE algorithms also obtain the best PESQ scores for values of a in the above mentioned ranges, for all noise types and input AvgSegSNRs. This should be attributed to their ability to preserve the weaker speech spectral components, while the residual noise is more broadband than that of their MAP counterparts.

We should finally mention the good performance of the MAP algorithms with very small values of a (~ 0.01) at high input AvgSegSNRs. Their success is due to their ability to preserve the weak speech components, while the musical residual noise is very low to be perceptually harmful at these low input noise levels. However, for low input AvgSegSNR, MAP algorithms with very narrow speech priors produce a lot of musical noise and they do not consist a good option.

5. CONCLUSION

In this paper we presented four Bayesian speech enhancement algorithms, that included the MAP and MMSE estimators with Chi and Gamma speech priors. The newly introduced MMSE algorithms resulted to an increase in the AvgSegSNR at least as good as that of their MAP counterparts while they demonstrated a further increase in the PESQ scores. This reflects their ability to preserve the weaker speech spectral components, while the residual noise has a much more broadband character compared to the residual noise of the MAP algorithms.

6. REFERENCES

[1] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.

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

[3] T. Lotter and P. Vary, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with supergaussian speech modelling," in *in Proc. of EUSIPCO-04 (Vienna, Austria),*, 2004, pp. 1457–1460.

[4] T. H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on,*, 2005, vol. 4, pp. 181–184.

[5] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1, John Wiley and Sons, New York, second edition, 1994.

[6] I. S. Gradshteyn and I. W. Ryzhik, *Tables of Integrals Series and Products*, New York: Academic Press, fourth edition, 1965.

[7] MATLAB routines for computation of Special Functions, "$http://ceta.mit.edu/comp\_spec\_func/$," Oct. 2005.

[8] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York, 1965.

[9] W. Gander and W. Gautschi, "Adaptive quadrature - revisited," *BIT*, vol. 40, no. 1, pp. 84–101, 2000.