

A Unified Framework for Designing Optimal STSA Estimators Assuming Maximum Likelihood Phase Equivalence of Speech and Noise

Bengt J. Borgström, *Member, IEEE*, and Abeer Alwan, *IEEE Fellow*

Abstract

In this paper, we present a stochastic framework for designing optimal short-time spectral amplitude (STSA) estimators for speech enhancement assuming phase equivalence of speech and noise. By assuming additive superposition of speech and noise, which is implied by the maximum likelihood (ML) phase estimate [6], we effectively project the optimal spectral amplitude estimation problem onto a 1-dimensional subspace of the complex spectral plane, thus simplifying the problem formulation. Assuming generalized Gamma distributions (GGDs) for *a priori* distributions of both speech and noise STSAs, we derive separate families of novel estimators according to either the maximum likelihood (ML), the minimum mean-square error (MMSE), or the maximum a posteriori (MAP) criterion. The use of GGDs allows optimal estimators to be determined in a generalized form, so that particular solutions can be obtained by substituting statistical shape parameters corresponding to expected speech and noise priors. It is interesting to note that several of the proposed estimators exhibit strong similarities to well-known STSA solutions. For example, the magnitude spectral subtractor (MSS) [2] and Wiener filter (WF) [1] are obtained for specific cases of GGD shape parameters. Quantitative analysis of a selected subset of the proposed estimators shows improvement over the traditional log-spectral MMSE estimator of Ephraim and Malah [7], in terms of segmental SNR and the COSH distance measure [34], when applied to the Noizeus database [28]. Although single-channel speech enhancement is offered as an illustrative example, the theory presented here could be applicable to other signals, such as music and images.

Index Terms

This work was supported in part by the NSF

Speech Enhancement, Noise Suppression, Short-Time Spectral Amplitude Estimation, Generalized Gamma Distribution, Spectral Subtraction.

I. INTRODUCTION

Suppressing additive acoustic noise in corrupted speech is an important task in many communication systems since it improves perceptual quality for human listeners. Phase values of observed spectral components are commonly neglected when inferring the underlying clean speech, leading to the short-time spectral amplitude (STSA) estimation framework. Traditional solutions to single channel STSA estimation include the maximum likelihood (ML) estimator proposed by McAulay and Malpass [5], maximum a posteriori (MAP) estimators proposed by Wolfe and Godsil [8] and Lotter and Vary [10], and minimum mean-square error (MMSE) estimators proposed by Ephraim and Malah [6],[7].

In traditional studies, spectral coefficients for both speech and noise components were modelled as Gaussian processes, due in part to the resulting mathematical efficiency. Particularly, following the assumption of Gaussian *a priori* models allows the conditional distribution of observed speech given clean speech to elegantly be reduced to a closed form expression involving the modified Bessel function [5]. Note that Gaussian models in the spectral coefficient (DFT) domain correspond to Rayleigh models in the spectral amplitude domain.

Recently, several studies have considered super-Gaussian *a priori* distributions for speech, since such models may more accurately describe distributions observed in empirical studies [9],[10]. However, super-Gaussian processes in the spectral coefficient domain correspond to mathematically complicated distributions in the spectral amplitude domain. Furthermore, such STSA distributions are generally phase-dependent. To circumvent such issues, Lotter and Vary proposed a phase-invariant STSA model fit to randomly generated data [10]. In [12], the authors model STSAs with the generalized Gamma distribution (GGD), which allows flexibility in capturing the underlying statistical behavior of speech. In this paper, we utilize GGD *a priori* speech STSA distributions, providing solutions in generalized form, as functions of GGD shape parameters.

In [2], the notion of phase equivalence of speech and noise components was studied by Boll for the task of STSA estimation. The assumption of phase equivalence greatly simplifies the inference process by projecting the estimation problem onto a 1-dimensional subspace of the complex plane. Similar theory has since been applied to noise robust automatic speech recognition (ASR) [30]. Motivated by the previously mentioned studies, we present a framework for deriving speech enhancement rules which apply the phase equivalence assumption, along with stochastic modeling of signal and noise components, resulting in a

unified framework for designing families of STSA estimators. We provide STSA solutions corresponding to the ML, MMSE, and MAP optimization criteria. Furthermore, a subtraction factor [3] is used to minimize over-attenuation.

When applied to the Noizeus database [28], a selected subset of the proposed estimators shows improved signal quality in terms of SSNR-related measures and the COSH distance metric [34], relative to the MMSE estimator from [6]. Informal listening tests provided promising results, showing proposed estimators to achieve improved noise suppression relative to [7].

This paper is organized as follows: in Section II we motivate the phase equivalence assumption and expand on the generalized Gamma distribution. ML, MMSE, and MAP solutions to STSA estimation are presented in Sections III, IV, and V, respectively. In Section VI we present experimental results for speech enhancement. Discussions are provided in Section VII.

II. DISTRIBUTIONS OF SPEECH AND NOISE SPECTRAL COMPONENTS

A. Phase Equivalence and Spectral Subtraction

Assuming an additive noise model, an observed speech signal can be expressed as:

$$Y_k = X_k + N_k, \quad (1)$$

where Y_k , X_k , and N_k represent the spectral coefficients of observed speech, clean speech, and noise, respectively, and where k denotes channel index. In this study, speech and noise processes are assumed to be statistically independent, both for amplitude and phase terms. Spectral coefficients can be decomposed into magnitude and phase components:

$$Y_k = R_k \exp(j\eta_k) \quad (2)$$

$$X_k = A_k \exp(j\alpha_k) \quad (3)$$

$$N_k = D_k \exp(j\psi_k) \quad (4)$$

Here, R_k , A_k , and D_k denote the spectral amplitudes of observed speech, clean unknown speech, and noise, respectively, and η_k , α_k , and ψ_k are the corresponding phases.¹ Note that R_k is an observed value and thus is a deterministic value. The values A_k and D_k represent random hidden variables. As in [5] and [6], we define *a priori* and *a posteriori* SNRs, ξ_k and γ_k , as:

¹Note that in this study, the terms amplitude and magnitude are used interchangeably.

$$\begin{aligned}\xi_k &= \frac{E[A_k^2]}{E[D_k^2]} = \frac{\sigma_x^2(k)}{\sigma_n^2(k)} \\ \gamma_k &= \frac{R_k^2}{E[D_k^2]} = \frac{R_k^2}{\sigma_n^2(k)}.\end{aligned}\quad (5)$$

Throughout this study, enhancement solutions are expressed concisely as gain functions: $G(\xi_k, \gamma_k) = A_k/R_k$. Also, second-order statistics of noise amplitudes are approximated as:

$$\sigma_n^2(k) \approx \hat{D}_k^2, \quad (6)$$

where \hat{D}_k^2 denotes a channel-specific noise estimate.

As early as [2], the notion of phase equivalence in spectral magnitude estimation was studied by Boll, leading to the well-known magnitude spectral subtraction (MSS) solution. Since then, multiple variations of spectral subtraction, such as power spectral subtraction (PSS) [29] and nonlinear spectral subtraction (NSS) [30], have been developed for both speech enhancement and noise robust automatic speech recognition (ASR).

In this study, we explore the role of phase equivalence in statistical approaches to STSA estimation. In [2], spectral observations are interpreted as deterministic values, implying underlying spectral speech and noise components to be deterministic as well. In this study, however, we apply a stochastic approach, and interpret spectral observations, and thus the underlying speech and noise components, as random processes.

If we assume phase equivalence of speech and noise spectral components, i.e. $\alpha_k = \psi_k$, (1) becomes:

$$R_k = A_k + D_k \quad (7)$$

This assumption reduces the complexity required by the estimation process, as it effectively projects the spectral estimation problem onto a 1-dimensional subspace of the complex domain. As discussed in [4], the relationship of (7), which is implied by the assumption of phase equivalence, results in a speech spectral amplitude estimate that is less than or equal to the actual additive clean speech. Specifically, the law of cosines [15] leads to (see Appendix I):

$$A_k = R_k - \rho(\gamma_k, \theta_k) D_k \quad (8)$$

where:

$$\theta_k = \alpha_k + (\pi - \psi_k) \quad (9)$$

and:

$$\rho(\gamma_k, \theta_k) = \sqrt{\gamma_k} - \sqrt{\gamma_k - \sin^2 \theta_k} - \cos \theta_k \quad (10)$$

Note that the subtraction factor assumes the following values at the boundary points of θ_k :

$$\begin{aligned} \rho(\gamma_k, \theta_k) \Big|_{\theta_k=\pi} &= 1 \\ \rho(\gamma_k, \theta_k) \Big|_{\theta_k=0} &= -1 \end{aligned} \quad (11)$$

Thus, the subtraction factor is bounded such that

$$-1 \leq \rho(\gamma_k, \theta_k) \leq 1 \quad (12)$$

Figure 1 illustrates the subtraction factor $\rho(\gamma_k, \theta_k)$ as a function of *a posteriori* SNR and relative angle θ_k . It can be concluded that $\rho(\gamma_k, \theta_k)$ decreases monotonically as the *a posteriori* SNR increases. The subtraction factor shows the asymptotic behavior:

$$\rho(\gamma_k, \theta_k) \Big|_{\gamma_k \gg 1} \approx -\cos(\theta_k) \quad (13)$$

$$\rho(\gamma_k, \theta_k) \Big|_{\gamma_k=1} = 1 - |\cos(\theta_k)| - \cos(\theta_k) \quad (14)$$

In terms of the estimation rule implied by (7), for $\theta_k=\pi$, which corresponds to the MSS solution [2], over-attenuation can be expected. Conversely, for $\theta_k=\pi/2$, under-attenuation can be expected. Finally, for $\theta_k=0$, (7) will tend to amplify noise components.

Motivated by the previous discussion, a subtraction factor, $\tilde{\rho}(\gamma_k)$, is applied:

$$R_k = A_k + \tilde{\rho}_k(\gamma_k) D_k \quad (15)$$

where $\tilde{\rho}_k(\gamma_k)$ is obtained by evaluating (10) at a certain relative angle. This yields modified definitions of *a priori* and *a posteriori* SNR:

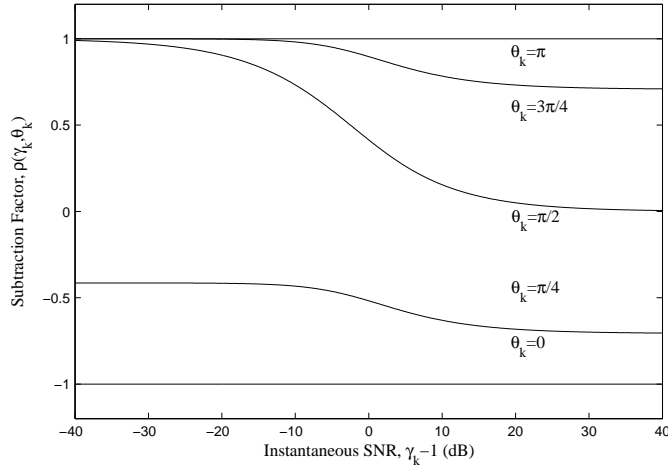


Fig. 1. Subtraction Factors as a Function of γ_k and θ_k

$$\begin{aligned}\tilde{\xi}_k &= \frac{\sigma_x^2(k)}{\tilde{\rho}_k^2(\gamma_k) \sigma_n^2(k)} \\ \tilde{\gamma}_k &= \frac{R_k^2}{\tilde{\rho}_k^2(\gamma_k) \sigma_n^2(k)}\end{aligned}\quad (16)$$

Without prior knowledge of the relative phases of speech and noise components, $\tilde{\rho}_k$ offers the ability to tune the level of suppression applied during enhancement. Note that this term is similar to the concept of over-subtraction in [2] and [30]. In this study, a value of $\theta_k=7\pi/8$ was empirically observed to provide promising results for estimating $\tilde{\rho}_k(\gamma_k)$, when tested on the Noizeus database [28]. Note that in Sections III-V, STSA estimators are derived in terms of standard values of γ_k and ξ_k for the sake of consistency with existing solutions. However, a subtraction factor can be applied by using SNR values from (16).

B. Generalized Gamma Distributions

Statistical modelling of spectral magnitudes has been widely studied for speech processing applications [5], [6], [10], [12].² We utilize the generalized Gamma distribution during the derivation of optimal estimators, resulting in general solutions from which special cases can be obtained. In this way, we

²Note that since this study deals with spectral magnitude estimation, only nonnegative distributions will be discussed. Thus, unless otherwise explicitly noted, no reference to "one-sided" will be made.

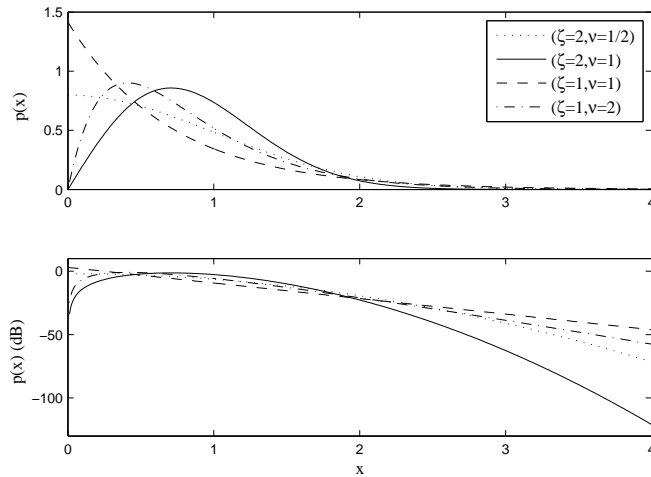


Fig. 2. The Generalized Gamma Distribution for Various Shaping Parameter Pairs: For illustrative purposes, the variance was normalized to unity.

increase the flexibility of the resulting estimators to varying signal types. The generalized Gamma distribution is given by:

$$p(x) = \frac{\zeta \beta^\nu}{\Gamma(\nu)} x^{\zeta\nu-1} \exp(-\beta x^\zeta), \quad (17)$$

for $x \geq 0; \beta, \nu, \zeta > 0$,

where Γ is the Gamma function. ν and ζ are referred to as shape parameters, and determine the overall shape of the distribution $p(x)$, whereas β is referred to as the scale parameter, and is related to the noncentral second moment of the distribution. Specifically, the second moment of $p(x)$ is given by [12]:

$$\sigma^2 = E[x^2] = \begin{cases} \frac{\nu(\nu+1)}{\beta^2}, & \text{if } \zeta = 1 \\ \frac{\nu}{\beta}, & \text{if } \zeta = 2 \end{cases} \quad (18)$$

Figure 2 provides probability distribution functions (pdfs) of the GGD for various shaping parameter pairs. Specifically, the top panel shows $p(x)$ in the linear scale, illustrating the effect of shaping parameters on GGDs. The bottom panel shows $p(x)$ in the log scale, illustrating the relative prominence of distribution tails for each set of shaping parameters. Note that (17) is equivalent to that used in [12]. Further, it is similar to that used in [10], where scale parameters are defined differently.

Studies in [10] and [9] provide empirical *a priori* distributions for speech and noise. Non-monotonic, unimodal distributions are shown for both speech and noise spectral amplitudes, similar to Erlang-2 and Rayleigh random variables. It should be noted, however, that empirical studies may often be greatly dependent on data, and may not convey true statistics. Therefore, we derive estimation solutions as functions of GGD shape parameters when mathematically possible. In this way, specific solutions can be obtained by substituting those shape parameters corresponding to expected speech and noise prior distributions. Alternatively prior distributions can be fitted according to ML or Kullback divergence criteria, as in [10].

In this study distinct distributions are used for noise and speech processes. Subscripts are utilized to denote shape parameters corresponding to noise $(\zeta_n, \nu_n, \beta_n)$ or to speech $(\zeta_x, \nu_x, \beta_x)$. Additionally, the proposed framework is robust to the use of channel-specific shape parameters. However, for the sake of simplicity, global parameters are utilized in this study.

This paper presents a framework for determining optimal STSA estimators using the assumption of phase equivalence for speech and noise. Various combinations of GGD shape parameters for noise and speech processes lead to specific solutions. This study presents gain functions for those combinations which yield closed-form expressions. In the case of MMSE estimation, higher order shape parameters generally require numerical analysis since such expressions rely on integration. In the case of MAP estimation, certain combinations of lower order shape parameters result in a monotonic cost function for which a MAP solution does not exist. A summary of proposed estimators is provided in Table I.

III. ML ESTIMATION

In this section, we present ML short-time spectral estimators based on the aforementioned phase equivalence assumption. ML estimation offers an efficient framework for inferring unknown parameters when *a priori* distributions of the target signal are not known. From (7), it can be concluded that the conditional probability of observed spectral components, given underlying clean components, is simply dependent on noise statistics:

$$\begin{aligned} p(R_k|A_k) &= p(D_k = R_k - A_k) \\ &= \frac{\zeta_n \beta_n^{\nu_n}}{\Gamma(\nu_n)} (R_k - A_k)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n}\right) \end{aligned} \quad (19)$$

ML estimation provides the estimate of A_k as:

$$\hat{A}_k = \arg \max_{A_k} \mathfrak{F}\{p(R_k|A_k)\} \quad (20)$$

where $\mathfrak{F}\{\cdot\}$ is monotonically increasing. Using $\mathfrak{F}\{\cdot\}=\log(\cdot)$ and (5) and (18), the ML solution is generalized as:

$$G_{ML}(\xi_k, \gamma_k) = 1 - \frac{1}{\sqrt{\gamma_k}} \left(\frac{\zeta_n \nu_n - 1}{\zeta_n \sqrt{\nu_n (\nu_n + 2 - \zeta_n)}} \right)^{1/\zeta_n}, \quad (21)$$

for $\zeta_n \in \{1, 2\}$

Analysis of the second derivative of (19) reveals that for $\zeta_n \nu_n < 1$, the solution of (21) does not exist, since the distribution $p(R_k|A_k)$ is monotonic, and thus includes no maximum. For $\zeta_n=1$ and $\nu_n=1$, the G_{ML} estimator reduces to unity.

It is interesting to note the similarity between the G_{ML} estimator and magnitude spectral subtraction presented by Boll in [2]. Particularly, for large values of ν_n , which correspond to deterministic values of A_k without uncertainty, the proposed ML estimator approaches that given in [2].

$$G_{ML}(\xi_k, \gamma_k) \Big|_{\zeta_n=2, \nu_n \rightarrow \infty} = \frac{\sqrt{\gamma_k} - 1}{\sqrt{\gamma_k}} \quad (22)$$

Figure 3 illustrates gain curves for the proposed ML STSA estimator for $\zeta_n=2$, and for various values of the GGD shape parameter ν_n .

IV. MMSE ESTIMATION

In this section, we derive MMSE short-time spectral amplitude estimators assuming phase equivalence of speech and noise. In the general case, using (1), and assuming A_k and α_k to be statistically independent, the MMSE estimate of A_k is determined as [6]:

$$\begin{aligned} \hat{A}_k &= E[A_k|Y_k] \\ &= \frac{\int_0^\infty \int_0^{2\pi} A_k p(Y_k|A_k, \alpha_k) p(\alpha_k) p(A_k) \partial \alpha_k \partial A_k}{\int_0^\infty \int_0^{2\pi} p(Y_k|A_k, \alpha_k) p(\alpha_k) p(A_k) \partial \alpha_k \partial A_k} \end{aligned} \quad (23)$$

By assuming phase equivalence of speech and noise components, the distribution of α_k and ψ_k are given by:

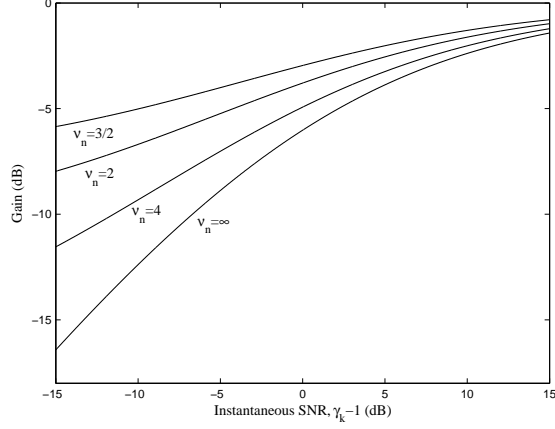


Fig. 3. Gain Curves for the G_{ML} STSA Estimator, with $\zeta_n=2$, and for Various Values of ν_n . Note that for $\nu_n=\infty$, the proposed ML estimator is equivalent to the magnitude spectral subtraction solution from [2].

$$p(\alpha_k) = \delta(\alpha_k - \eta_k) \quad (24)$$

$$p(\psi_k) = \delta(\psi_k - \eta_k) \quad (25)$$

where $\delta(\cdot)$ represents the Kronecker delta function, and (23) simplifies to:

$$\hat{A}_k = \frac{\int_0^\infty A_k p(R_k | A_k) p(A_k) \partial A_k}{\int_0^\infty p(R_k | A_k) p(A_k) \partial A_k} \quad (26)$$

Using (19), and the fact that $A_k, D_k \in [0, R_k]$, (23) reduces to:

$$\hat{A}_k = \frac{\int_0^{R_k} A_k p(D_k = R_k - A_k) p(A_k) \partial A_k}{\int_0^{R_k} p(D_k = R_k - A_k) p(A_k) \partial A_k} \quad (27)$$

Assuming generalized Gamma distributions for speech and noise spectral magnitudes, the solution of (27) becomes:

$$\hat{A}_k = \frac{\int_0^{R_k} A_k^{\zeta_x \nu_x} (R_k - A_k)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n} - \beta_x A_k^{\zeta_x}\right) \partial A_k}{\int_0^{R_k} A_k^{\zeta_x \nu_x - 1} (R_k - A_k)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n} - \beta_x A_k^{\zeta_x}\right) \partial A_k} \quad (28)$$

(28) provides a general solution to MMSE spectral magnitude estimation assuming phase equivalence, and assuming priors from the generalized Gamma family. Particular solutions can be obtained by substituting specific statistical shape parameters corresponding to desired speech and noise distributions.

Due to varying conclusions regarding the true statistical behavior of speech, we have avoided specifying models for speech and noise. Instead, (28) is provided as a function of GGD shape parameters. In the following subsections, we derive MMSE solutions for shape parameters which may be considered realistic for speech and noise.

A. Assuming Gaussian Noise Priors ($\zeta_n=2, \nu_n=1/2$)

In this section, we derive the MMSE spectral magnitude estimator assuming Gaussian noise priors. Specifically, we present separate particular solutions assuming Gaussian and exponential speech priors, referred to as *GGMMSE* and *GEMMSE*, respectively.

1) *Assuming Gaussian Speech Priors* ($\zeta_x=2, \nu_x=1/2$): The GGMMSE solution, which utilizes Gaussian speech and noise priors, is derived from (28):

$$\hat{A}_k = \frac{\int_0^{R_k} A_k \exp\left(-\beta_n (R_k - A_k)^2 - \beta_x A_k^2\right) \partial A_k}{\int_0^{R_k} \exp\left(-\beta_n (R_k - A_k)^2 - \beta_x A_k^2\right) \partial A_k} \quad (29)$$

Using (5) and (18), the integrals in (29) can be solved to reveal the GGMMSE gain function:

$$G_{GGMMSE}(\xi_k, \gamma_k) = \phi_k - \sqrt{\frac{2\phi_k}{\pi\gamma_k}} \left(\frac{\exp\left(-\frac{\gamma_k}{2\xi_k(1+\xi_k)}\right) - \exp\left(-\frac{\gamma_k\phi_k}{2}\right)}{\operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k\phi_k}{2}}\right)} \right) \quad (30)$$

where ϕ_k is the traditional Wiener filter (WF) [1]:

$$\phi_k = \frac{\xi_k}{1 + \xi_k}. \quad (31)$$

Additionally, $\operatorname{erf}(\cdot)$ represents the Gauss error function:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-\tau^2} d\tau. \quad (32)$$

It is interesting to note that for extreme values of the a posteriori SNR, the GGMMSE estimator approximates the Wiener filter:

$$G_{GGMMSE}(\xi_k, \gamma_k) \Big|_{|\gamma| \gg 1} \approx \phi_k. \quad (33)$$

More specifically, the GGMMSE gain function can be interpreted as the Wiener filter with an additive modification factor. To avoid a solution which includes irreducible functions, we approximate the Gauss error function by its truncated Taylor series expansion. The Taylor series expansion of the Gauss error function is given as [15]:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n! (2n+1)}. \quad (34)$$

Using the 1st-order approximation, we derive the \hat{G}_{GGMMSE}^1 solution:

$$\begin{aligned} \hat{G}_{GGMMSE}^1(\xi_k, \gamma_k) = & \quad (35) \\ \phi_k \left[1 - \frac{1}{\gamma_k} \exp\left(-\frac{\gamma_k(1+\xi_k^2)}{4\xi_k(1+\xi_k)}\right) \sinh\left(\frac{\gamma_k(\xi_k-1)}{4\xi_k}\right) \right] \end{aligned}$$

The 1st-order approximation utilized in (35) can be expected to provide high accuracy for small values of γ_k , which correspond to acoustic conditions wherein the gain function is most important.

Figure 4 illustrates gain curves for the \hat{G}_{GGMMSE}^1 estimator, for various *a priori* SNRs. The Wiener filter is included for reference. As can be observed in Figure 4, the \hat{G}_{GGMMSE}^1 converges to the Wiener filter for large values of the *a posteriori* SNR. Additionally, for favorable acoustic conditions (eg. $\xi_k=15$ or 5 dB), the \hat{G}_{GGMMSE}^1 estimator provides increasing attenuation as the *a posteriori* SNR decreases, which follows intuitively. For unfavorable conditions (eg. $\xi_k=-5$ or -15 dB), however, increased attenuation is applied for increasing *a posteriori* SNR. As discussed in [6], such behavior is the result of the estimator compromising between *a priori* information in the form of ξ_k , and new information introduced by γ_k .

2) *Assuming Exponential Speech Priors* ($\zeta_x = 1, \nu_x = 1$): Retaining the previous Gaussian model for the noise component and using (5) and (18), the GEMMSE solution can be adapted from (28) by assuming Exponential speech priors:

$$\hat{A}_k = \frac{\int_0^{R_k} A_k \exp\left(-\beta_n (R_k - A_k)^2 - \beta_x A_k\right) \partial A_k}{\int_0^{R_k} \exp\left(-\beta_n (R_k - A_k)^2 - \beta_x A_k\right) \partial A_k} \quad (36)$$

The GEMMSE gain function can be derived from (36) as:

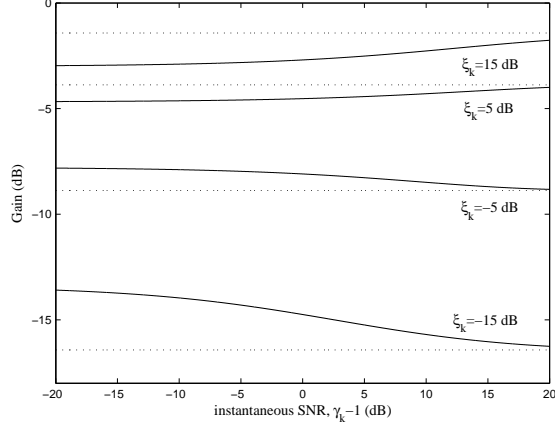


Fig. 4. Gain curves for the \hat{G}_{GEMMSE}^1 STSA estimator (solid line) for various values of ξ_k : The Wiener filter [1] (dotted line) is included for comparison.

$$G_{GEMMSE}(\xi_k, \gamma_k) = 1 - \sqrt{\frac{2}{\xi_k \gamma_k}} - \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{\exp\left(-\frac{1}{\xi_k}\right) - \exp\left(-\frac{\gamma_k}{2} - \frac{1}{\xi_k} + \frac{\sqrt{2\gamma_k}}{\sqrt{\xi_k}}\right)}{\operatorname{erf}\left(\frac{1}{\sqrt{\xi_k}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2}} - \frac{1}{\sqrt{\xi_k}}\right)} \right] \quad (37)$$

As in the previous section, the GEMMSE includes irreducible Gauss error functions, which can be expressed as Taylor series expansions [15]. Following the steps from the previous section, the GEMMSE solution can be approximated as:

$$\hat{G}_{GEMMSE}^1(\xi_k, \gamma_k) = 1 - \frac{1}{2\sqrt{\xi_k \gamma_k}} - \frac{1}{\gamma_k} \left[\exp\left(-\frac{1}{\xi_k} - \frac{\gamma_k}{4} + \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \sinh\left(\frac{\gamma_k}{4} - \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \right] \quad (38)$$

Figure 5 illustrates gain curves for the $GEMMSE^1$ estimator for various values of ξ_k . It is interesting to note the dissimilarities in behavior between the $GEMMSE^1$ and $GGMMSE^1$ solutions. The increased attenuation applied by the former is due to the narrow peak of the exponential distribution, which shifts spectral estimates downward.

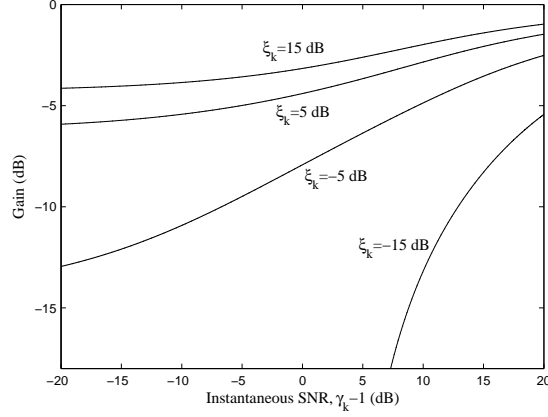


Fig. 5. Gain curves for the \hat{G}_{GEMMSE}^1 STSA estimator for various values of ξ_k

B. Assuming Exponential Noise Priors ($\zeta_n = 1, \nu_n = 1$)

In this section, we derive a general MMSE solution assuming exponential noise spectral magnitude priors, with speech prior distributions constrained by $\zeta_x=1$, but given as a function of $\nu_x \in \mathbb{N}_1$, where \mathbb{N}_1 is the set of natural numbers. Note that for $\zeta_x=2$, the MMSE estimator is irreducible, and numerical approximation must be used to obtain a solution. For $\zeta_x=1$, (28) can be simplified as:

$$\hat{A}_k = \frac{\int_0^{R_k} A_k^{\nu_x} \exp(-(\beta_x - \beta_n) A_k) \partial A_k}{\int_0^{R_k} A_k^{\nu_x-1} \exp(-(\beta_x - \beta_n) A_k) \partial A_k} \quad (39)$$

The following identity, which is proven in Appendix II, is helpful in deriving the current spectral estimator:

$$\int_0^z \tau^m \exp(c\tau) d\tau = \begin{cases} \frac{\exp(cz)}{c} \sum_{h=0}^m [(-1)^h \frac{m!}{(m-h)!} \frac{z^{m-h}}{c^h}] , \text{ for } c \neq 0 \\ \frac{z}{m+1}, \text{ for } c = 0 \end{cases} \quad (40)$$

Using (5), (18), (39), and (40), the MMSE STSA estimator for exponential noise priors can be expressed as:

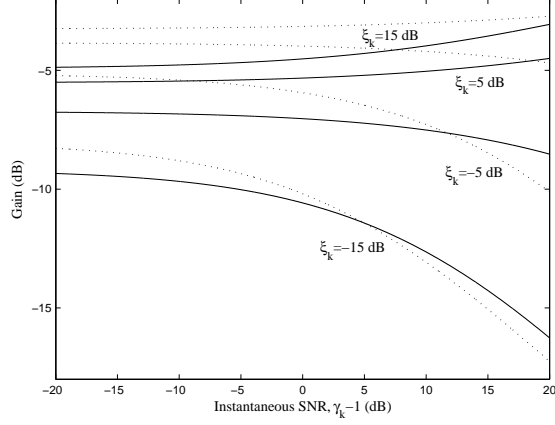


Fig. 6. Gain Curves for the $G_{EMMSE}^{(\nu_x)}$ STSA Estimator for $\nu_x=1$ (solid line) and $\nu_x=2$ (dotted line)

$$G_{EMMSE}^{(\nu_x)}(\xi_k, \gamma_k) = \begin{cases} \frac{1}{\mu_k} \left[\frac{(-1)^{\nu_x} \nu_x! + \exp(\mu_k) \sum_{h=0}^{\nu_x} \binom{\nu_x}{h} \frac{\nu_x!}{(\nu_x-h)!} \mu_k^{\nu_x-h}}{(-1)^{\nu_x-1} (\nu_x-1)! + \exp(\mu_k) \sum_{h=0}^{\nu_x-1} \binom{\nu_x-1}{h} \frac{(\nu_x-1)!}{(\nu_x-h-1)!} \mu_k^{\nu_x-h-1}} \right], & \text{for } c \neq 0 \\ \frac{\nu_x}{\nu_x+1} & \text{for } c = 0 \end{cases} \quad (41)$$

where:

$$\mu_k = \sqrt{\frac{2\gamma_k}{\xi_k}} \left(\sqrt{\xi_k} - \sqrt{\frac{\nu_x(\nu_x-1)}{2}} \right). \quad (42)$$

Note that the gain function in (41) is expressed as a function of ν_x , which allows flexibility in modelling the speech component. We present an example EMMSE solution for $\nu_x=1$, which corresponds to an exponential speech prior:

$$G_{EMMSE}^{(\nu_x=1)} = \frac{1}{\mu_k} \left(\frac{-1 + \exp(\mu_k)(-1 + \mu_k)}{1 + \exp(\mu_k)} \right), \quad (43)$$

Figure 6 provides gain curves for the $G_{EMMSE}^{(\nu_x)}$ estimator, for $\nu_x=\{1,2\}$. As can be observed, with increasing ν_x , the estimator generally provides decreased attenuation. This is due to the corresponding *a priori* speech distribution shifting its mode outward, thereby increasing the expected value of A_k .

V. MAP ESTIMATION

Maximum a posteriori estimation offers mathematically efficient solutions to problems which may not be feasible via other means. In this section, we present a group of MAP spectral amplitude estimators assuming phase equivalence of speech and noise. We explore Gaussian and Rayleigh noise priors. As with MMSE estimation, higher order shape parameters offer valid solutions; however, higher order GGDs typically result in solutions which can not be expressed in closed form.

MAP estimation determines the solution \hat{A}_k which maximizes the expression $p(R_k|A_k)p(A_k)$, or some monotonic function thereof, given the observation R_k :

$$\begin{aligned}\hat{A}_k &= \arg \max_{A_k} \mathfrak{F}\{p(R_k|A_k)p(A_k)\}, \\ &= \arg \max_{A_k} \mathfrak{F}\{p(D_k = R_k - A_k)p(A_k)\},\end{aligned}\tag{44}$$

where \mathfrak{F} is monotonic. We utilize the logarithmic function during MAP estimation:

$$\hat{A}_k = A_k \text{ such that } \frac{\partial}{\partial A_k} C(A_k) = 0,\tag{45}$$

$$\text{where } C(A_k) = \log(p(R_k|A_k)p(A_k))$$

Assuming generalized Gamma priors for speech and noise components and using (17) and (19), the equality $\frac{\partial}{\partial A_k} C(A_k)=0$ is expressed as:

$$\begin{aligned}\frac{\zeta_x \nu_x - 1}{A_k} - \frac{\zeta_n \nu_n - 1}{R_k - A_k} - \beta_x \zeta_x A_k^{\zeta_x - 1} \\ + \beta_n \zeta_n (R_k - A_k)^{\zeta_n - 1} = 0\end{aligned}\tag{46}$$

Analyzing the 2nd derivative of $C(A_k)$ with respect to A_k leads to:

$$\begin{aligned}\frac{\partial^2}{\partial A_k^2} C(A_k) &= -\frac{\zeta_x \nu_x - 1}{A_k^2} - \frac{\zeta_n \nu_n - 1}{(R_k - A_k)^2} \\ &\quad - \beta_x \zeta_x (\zeta_x - 1) A_k^{\zeta_x - 2} - \beta_n \zeta_n (\zeta_n - 1) (R_k - A_k)^{\zeta_n - 2}\end{aligned}\tag{47}$$

(47) shows that $\frac{\partial^2}{\partial A_k^2} C(A_k)$ will be guaranteed to be negative, and the solution from (46) valid, if each of the following inequalities holds:

$$(i) \zeta_x \geq 1 \quad (48)$$

$$(ii) \zeta_n \geq 1$$

$$(iii) \zeta_x \nu_x \geq 1$$

$$(iv) \zeta_n \nu_n \geq 1$$

(47) can assume negative values even if certain constraints in (48) are not met; however, these inequalities offer a simple check which encompasses the majority of valid speech and noise GGD shape parameters.

A. Assuming Gaussian Noise Priors ($\zeta_n=2, \nu_n=1/2$)

In this section, we present a family of MAP spectral magnitude estimators assuming Gaussian noise prior distributions. We derive separate general solutions for speech distributions constrained by ($\zeta_x=2, \nu_x \in \mathbb{R}$) and ($\zeta_x=1, \nu_x \in \mathbb{R}$).

The *G2MAP* estimator is derived from the assumption of Gaussian noise and speech distributions constrained by ($\zeta_x=2, \nu_x \in \mathbb{R}$). In this case, applying (5) and (18) to (46) leads to:

$$A_k^2 - \frac{\xi_k}{2\nu_x + \xi_k} R_k A_k - \frac{\xi_k (2\nu_x - 1)}{\gamma_k (2\nu_x + \xi_k)} R_k^2 = 0 \quad (49)$$

Applying the quadratic equation, and choosing the nonnegative root, yields:

$$\begin{aligned} G_{G2MAP}^{(\nu_x)}(\xi_k, \gamma_k) & \quad (50) \\ &= \frac{\xi_k + \sqrt{\xi_k^2 + 4(\xi_k/\gamma_k)(2\nu_x - 1)(2\nu_x + \xi_k)}}{2(2\nu_x + \xi_k)} \end{aligned}$$

It is interesting to note that the $G_{G2MAP}^{(\nu_x)}$ estimator of (50) shows striking similarity to the MAP STSA estimator proposed by Wolfe and Godsil in [8]. Furthermore, the case of $\nu_x=1/2$ leads to the Wiener filter

$$G_{G2MAP}^{(\nu_x=1/2)}(\xi_k, \gamma_k) = \phi_k \quad (51)$$

Figure 7 illustrates gain curves for the $G_{G2MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=1$ (solid line), $\nu_x=2$ (dashed line), and $\nu_x=3$ (dotted line).

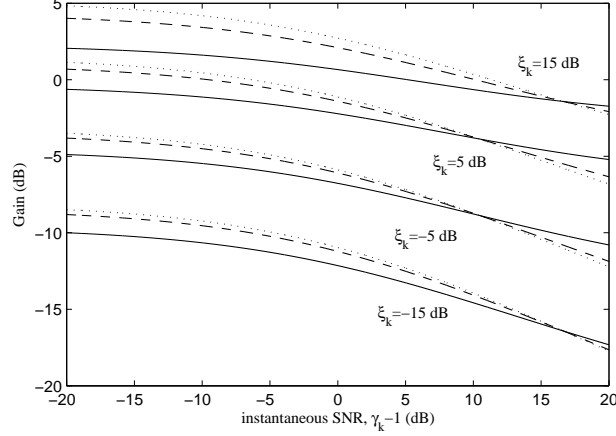


Fig. 7. Gain curves for the $G_{G2MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=1$ (solid line), $\nu_x=1$ (dashed line), and $\nu_x=3$ (dotted line)

The steps involved in deriving MAP spectral estimators, which were previously followed to obtain the $G2MAP$ solution, are summarized below:

- 1) Choose GGD size parameters corresponding to desired speech and noise prior distributions, and check for validity (48).
- 2) Substitute GGD size parameters into the general solution for MAP estimation (46).
- 3) Solve (46) for \hat{A}_k . Note that in certain cases, such as when $\frac{\partial}{\partial A_k} C(A_k)$ is linear or quadratic, this can be done in closed form. Other cases may rely on numerical approximations.
- 4) If multiple roots are obtained for \hat{A}_k choose the root which falls within the desired range $[0, \infty)$.
- 5) Substitute the definitions of the a priori and a posteriori SNRs (Sec. II) into \hat{A}_k .

Following the previously outlined steps, the $G1MAP$ solution is derived from the assumption of ($\zeta_x = 1, \nu_x \in \mathbb{R}$):

$$G_{G1MAP}^{(\nu_x)}(\xi_k, \gamma_k) = \frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}} + \sqrt{\left(\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}}\right)^2 + \frac{(\nu_x - 1)}{\gamma_k}} \quad (52)$$

Figure 8 illustrates gain curves for the $G_{G1MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=1.5$ (solid line), $\nu_x=2.0$ (dashed line), and $\nu_x=2.5$ (dotted line).

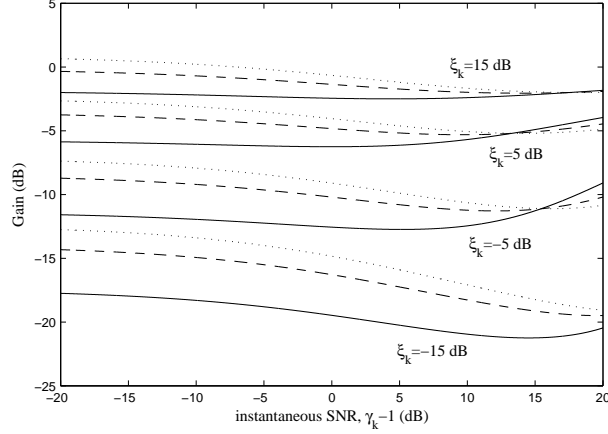


Fig. 8. Gain curves for the $G_{G1MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=1.5$ (solid line), $\nu_x=2.0$ (dashed line), and $\nu_x=2.5$ (dotted line)

B. Assuming Rayleigh Noise Priors ($\zeta_n = 2, \nu_n = 1$)

In this section, we present a family of MAP spectral magnitude estimators assuming Rayleigh noise prior distributions. (46) now reduces to:

$$(\zeta_x \nu_x - 1) - \beta_x \zeta_x A_k^{\zeta_x} - \frac{A_k}{R_k - A_k} + 2\beta_n A_k (R_k - A_k) = 0. \quad (53)$$

It can be observed that the expression in (53) is a $(\zeta_x + 1)^{th}$ -order polynomial. However, if the speech distribution shape parameters are constrained such that $\zeta_x \nu_x = 1$, the expression can be reduced to a ζ_x^{th} -order polynomial, and its roots can be obtained more efficiently. If Gaussian speech prior distributions are assumed, the *RGMAP* solution can be derived according to the steps outlined in Section V-A:

$$G_{RGMAP}(\xi_k, \gamma_k) = \frac{1}{2} \left(\frac{4\xi_k + 1}{2\xi_k + 1} \right) + \frac{1}{2} \sqrt{\left(\frac{4\xi_k + 1}{2\xi_k + 1} \right)^2 - \frac{4\xi_k}{\gamma_k} \left(\frac{2\gamma_k - 1}{2\xi_k + 1} \right)} \quad (54)$$

Instead if exponential speech priors are assumed, the *REMAP* estimator is derived as:

$$G_{REMAP}(\xi_k, \gamma_k) = 1 - \frac{1}{2\sqrt{2\xi_k\gamma_k}} + \sqrt{\left(1 - \frac{1}{2\sqrt{2\xi_k\gamma_k}}\right)^2 + \left(\frac{1}{\sqrt{2\xi_k\gamma_k}} + \frac{1}{2\gamma_k} - 1\right)} \quad (55)$$

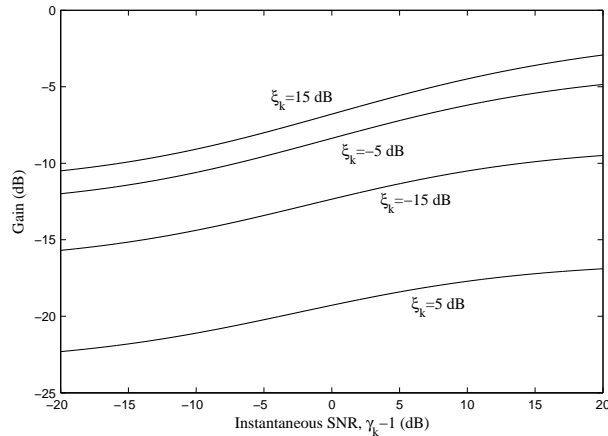


Fig. 9. Gain curves for the G_{RGMAP} STSA estimator for various values of ξ_k

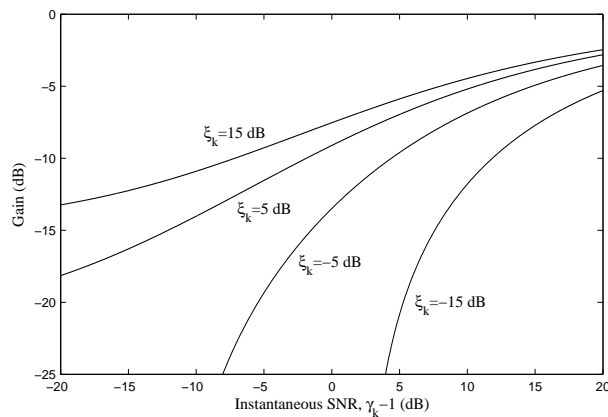


Fig. 10. Gain curves for the G_{REMAP} STSA estimator for various values of ξ_k

It is interesting to note the similarity between the proposed $G_{G1MAP}^{(\nu_x)}$ and $G_{REMAP}(\xi_k, \gamma_k)$ solutions, and the joint-MAP (JMAP) STSA estimator proposed by Lotter and Vary in [10]. Figures 9 and 10 illustrate gains curves for the G_{RGMAP} and G_{REMAP} estimators, respectively.

Table I provides a summary of the generalized STSA estimators derived in Sections III, IV, and V.

VI. EXPERIMENTAL RESULTS

In order to assess the success of the proposed spectral estimators, they are tested on the Noizeus database [28]. The Noizeus database is comprised of phonetically balanced utterances, and includes 8 types of non-stationary additive noise. Proposed STSA estimators were embedded into code from [32],

TABLE I

SPECTRAL MAGNITUDE ESTIMATORS DERIVED IN SECTIONS III, IV, AND V: PARTICULAR SOLUTIONS ARE OBTAINED BY SUBSTITUTING INTO GENERAL SOLUTIONS THOSE STATISTICAL PARAMETERS CORRESPONDING TO DESIRED NOISE AND SPEECH PRIORS. NOTE THAT ζ_n AND ν_n REFER TO GGD SHAPE PARAMETERS CORRESPONDING TO THE NOISE PROCESS, AND ζ_x AND ν_x REFER TO GGD SHAPE PARAMETERS CORRESPONDING TO THE SPEECH PROCESS.

Criterion	(ζ_n, ν_n)	(ζ_x, ν_x)	Name	$G(\xi_k, \gamma_k)$
ML	$(\zeta_n \in \{1, 2\}, \nu_n \geq 1/\zeta_n)$	-	G_{ML}	$1 - \frac{1}{\sqrt{\gamma_k}} \left(\frac{\zeta_n \nu_n - 1}{\zeta_n \sqrt{\nu_n (\nu_n + 2 - \zeta_n)}} \right)^{1/\zeta_n}$
MMSE	$(2, 1/2)$	$(2, 1/2)$	\hat{G}_{GMMSE}^1	$\phi_k \left[1 - \frac{1}{\gamma_k} \exp\left(-\frac{\gamma_k(1+\xi_k^2)}{4\xi_k(1+\xi_k)}\right) \sinh\left(\frac{\gamma_k(\xi_k-1)}{4\xi_k}\right) \right]$
		$(1, 1)$	\hat{G}_{GEMMSE}^1	$1 - \frac{1}{2\sqrt{\xi_k \gamma_k}} - \frac{1}{\gamma_k} \left[\exp\left(-\frac{1}{\xi_k} - \frac{\gamma_k}{4} + \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \sinh\left(\frac{\gamma_k}{4} - \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \right]$
	$(1, 1)$	$(1, \nu_x \in \mathbb{N}_1)$	$G_{EMMSE}^{(\nu_x)}$	$\frac{1}{\mu_k} \left(\frac{(-1)^{\nu_x} \nu_x! + \exp(\mu_k) \sum_{k=0}^{\nu_x} (-1)^k \frac{\nu_x!}{(\nu_x-k)!} \mu_k^{\nu_x-k}}{(-1)^{\nu_x-1} (\nu_x-1)! + \exp(\mu_k) \sum_{k=0}^{\nu_x-1} (-1)^k \frac{(\nu_x-1)!}{(\nu_x-k-1)!} \mu_k^{\nu_x-k-1}} \right)$ where $\mu_k = \sqrt{\frac{2\gamma_k}{\xi_k}} \left(\sqrt{\xi_k} - \sqrt{\frac{\nu_x(\nu_x-1)}{2}} \right)$
MAP	$(2, 1/2)$	$(2, \nu_x \in \mathbb{R})$	$G_{G2MAP}^{(\nu_x)}$	$\frac{1}{2(2\nu_k + \xi_k)} \left(\xi_k + \sqrt{\xi_k^2 + 4(\xi_k/\gamma_k)(2\nu_x - 1)(2\nu_x + \xi_k)} \right)$
		$(1, \nu_x \in \mathbb{R})$	$G_{G1MAP}^{(\nu_x)}$	$\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x+1)}}{2\sqrt{\gamma_k \xi_k}} + \sqrt{\left(\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x+1)}}{2\sqrt{\gamma_k \xi_k}}\right)^2 + \frac{(\nu_x-1)}{\gamma_k}}$
	$(2, 1)$	$(2, 1/2)$	G_{RGMAP}	$\frac{1}{2} \left(\frac{4\xi_k+1}{2\xi_k+1} \right) + \frac{1}{2} \sqrt{\left(\frac{4\xi_k+1}{2\xi_k+1} \right)^2 - \frac{4\xi_k}{\gamma_k} \left(\frac{2\gamma_k-1}{2\xi_k+1} \right)}$
		$(1, 1)$	G_{REMAP}	$1 - \frac{1}{2\sqrt{2\xi_k \gamma_k}} + \sqrt{\left(1 - \frac{1}{2\sqrt{2\xi_k \gamma_k}}\right)^2 + \left(\frac{1}{\sqrt{2\xi_k \gamma_k}} + \frac{1}{2\gamma_k} - 1\right)}$

which applies a gain floor of -18 dB and performs improved minima controlled recursive averaging (IMCRA) noise estimation [33]. The enhancement system applies 20 ms analysis windows with 15 ms overlap. A fast Fourier transform (FFT) of length 256 is used during spectral decomposition.

Results were averaged across 30 utterances, and 8 noise types. In order to analyze the effect of enhancement on speech and noise components separately, we use SSNR-related metrics defined in [10]. Specifically, along with the overall SSNR measure, we consider separately the SSNR of active speech frames ($SSNR_X$) and inactive speech frames ($SSNR_N$). For an enhanced speech signal $\hat{x}(n)$ and clean reference signal $x(n)$, the global SSNR is defined as:

$$SSNR = \tag{56}$$

$$\frac{1}{N_f} \sum_{k=1}^{N_f} \left[10 \log_{10} \left(\frac{\sum_{i=1}^{N_w} x^2(i + kN_h)}{\sum_{i=1}^{N_w} (x(i + kN_h) - \hat{x}(i + kN_h))^2} \right) \right]$$

where N_w is the length of the analysis window, N_h is the shift in samples for each successive window, and N_f is the number of frames analyzed. The $SSNR_X$ is determined similarly to (56), although only computed for speech frames wherein the clean signal energy is ≥ -30 dB [10]. Conversely, the $SSNR_N$ is computed only for frames exhibiting energy < -30 dB.

Table II provides quantitative results for a subset of the proposed STSA estimators, when applied to the Noizeus database [28]. The subset of proposed estimators was chosen to represent a diverse group. Note that most often the proposed MMSE- and MAP-based estimators outperform the ML-based estimator. However, for estimation of general signals, the ML solution is applicable when *a priori* statistics of the signal of interest are unknown. Δ , Δ_X , Δ_N refer to improvements in $SSNR$, $SSNR_X$, $SSNR_N$, respectively, relative to the unprocessed signal. Bold entries denote the best score for each metric at each noise condition. Results in Table II were obtained with a subtraction factor $\rho_k(\gamma_k, \theta_k)$ evaluated at $\theta_k=7\pi/8$, as discussed in (15) and (16).

Table II shows the proposed STSA estimators to generally provide improved overall signal quality ($SSNR$) and noise suppression ($SSNR_N$), relative to the LMMSE solution from [7], the MAP solution from [8], and the JMAP solution from [10]. The proposed methods provide generally improved active speech quality ($SSNR_X$) relative to [7]. The $GGMMSE^1$ estimator, which provides the best scores, outperforms that from [7] by approximately 1-2.5 dB in terms of $SSNR$ across noise levels, and by approximately 5-6.5 dB in terms of $SSNR_N$. It should be noted that for some noise levels, the solutions from [8] and [10] lead to slightly better $SSNR_X$ scores, relative to the proposed estimators. In these cases, the proposed solutions may be attenuating speech components while attempting to suppress noise components. In the proposed approach, the subtraction factor $\tilde{\rho}_k$ can be tuned to control the tradeoff between noise suppression and speech distortion.

STSA estimators are also evaluated using a discretized version of the COSH distance from [34]:

$$d_{COSH}(x(n), \tilde{x}(n)) = \frac{1}{2N_{ch}} \sum_{k=1}^{N_{ch}} \left(\frac{A_k}{\hat{A}_k} + \frac{\hat{A}_k}{A_k} - 2 \right) \tag{57}$$

where N_{ch} is the number of channels used during spectral analysis. Table III provides COSH measures for a subset of the proposed STSA estimators, when applied to the Noizeus database [28]. Bold entries

TABLE II

SEGMENTAL SNR SCORES FOR SELECTED STSA ESTIMATORS: Δ , Δ_X , Δ_N REFER TO IMPROVEMENTS IN SSNR, SSNR_X , SSNR_N , RESPECTIVELY, RELATIVE TO THE UNPROCESSED SIGNAL. RESULTS WERE OBTAINED ON THE NOIZEUS DATABASE [28]. THE SUBSET OF ESTIMATORS INCLUDED WERE SELECTED FROM THE PROPOSED FRAMEWORK TO REPRESENT A DIVERSE MIX. BOLD ENTRIES DENOTE THE BEST SCORE FOR EACH METRIC AT EACH NOISE CONDITION.

Estimator	(ζ_n, ν_n)	(ζ_x, ν_x)	Input SNR											
			15 dB			10 dB			5 dB			0 dB		
			Δ	Δ_X	Δ_N	Δ	Δ_X	Δ_N	Δ	Δ_X	Δ_N	Δ	Δ_X	Δ_N
<i>LMMSE</i> [7]	–	–	4.8	2.5	10.6	5.6	3.4	10.8	6.5	4.5	11.1	7.3	5.7	11.0
<i>MAP</i> [8]	–	–	4.8	2.5	10.6	5.6	4.5	10.8	6.5	5.0	11.1	7.3	6.4	11.0
<i>JMAP</i> [10]	–	–	4.9	2.5	10.8	5.7	4.6	11.1	6.5	5.0	11.3	7.3	6.4	11.3
\hat{G}_{GGMMSE}^1	(2, 1/2)	(2, 1/2)	5.2	1.8	13.7	6.3	3.0	14.4	7.4	4.4	15.0	8.6	6.0	15.0
$G_{EMMSE}^{(\nu_x=1)}$	(1, 1)	(1, 1)	4.7	2.5	10.0	5.4	3.4	10.1	6.2	4.4	10.2	6.9	5.5	10.1
$G_{G2MAP}^{(\nu_x=1/2)}$	(2, 1/2)	(2, 1/2)	5.2	2.5	11.9	6.1	3.5	12.3	7.0	4.7	12.6	7.9	5.9	12.6
G_{RGMAP}	(2, 1)	(2, 1/2)	5.1	2.5	11.6	5.9	3.5	11.8	6.8	4.7	12.1	7.7	5.8	12.0
G_{REMAP}	(2, 1)	(1, 1)	4.9	2.5	10.7	5.6	3.4	10.9	6.4	4.5	11.0	7.2	5.6	10.9

denote the best score for each metric at each noise condition. It can be observed that several of the proposed estimators provide lower distortion measures than the previously mentioned baseline solutions, across all noise levels. Furthermore, the best results are achieved by the \hat{G}_{GGMMSE}^1 estimator.

Informal listening tests show the proposed estimators included in Table II to provide a noticeable improvement in noise suppression, across noise types and levels, relative to the solutions from [7], [8], and [10]. Speech enhancement at low noise levels (5 dB and 0 dB), did result in some apparent musical noise, especially for highly non-stationary, speech-shaped noise types, such as *babble* and *restaurant*. However, this musical noise was generally not more noticeable than that produced by the baseline methods.

VII. SUMMARY AND DISCUSSION

In this paper, we present a stochastic framework for designing optimal spectral magnitude estimators for speech enhancement assuming equivalent phase of speech and noise components. By assuming phase equivalence, we effectively project the optimal spectral amplitude estimation problem onto a 1-dimensional subspace of the complex plane, thus simplifying the problem formulation. We derive separate families of novel estimators assuming generalized gamma distributions (GGDs) for both speech and noise

TABLE III

COSH DISTORTION MEASURES [34] FOR THE STSA ESTIMATORS INCLUDED IN TABLE II. RESULTS WERE OBTAINED ON THE NOIZEUS DATABASE [28]. BOLD ENTRIES DENOTE THE BEST SCORE FOR EACH METRIC AT EACH NOISE CONDITION.

Estimator	$(\hat{\zeta}_n, \nu_n)$	$(\hat{\zeta}_x, \nu_x)$	Input SNR (dB)			
			15	10	5	0
<i>LMMSE</i> [7]	–	–	1.43	2.65	4.55	8.13
<i>MAP</i> [8]	–	–	1.45	2.78	4.55	8.10
<i>JMAP</i> [10]	–	–	1.44	2.61	4.46	7.93
\hat{G}_{GGMMSE}^1	(2, 1/2)	(2, 1/2)	1.33	2.07	3.26	5.48
$G_{EMMSE}^{(\nu_x=1)}$	(1, 1)	(1, 1)	1.47	2.74	4.81	8.63
$G_{G2MAP}^{(\nu_x=1/2)}$	(2, 1/2)	(2, 1/2)	1.39	2.37	3.97	6.93
G_{RGMAP}	(2, 1)	(2, 1/2)	1.39	2.41	4.10	7.23
G_{REMAP}	(2, 1)	(1, 1)	1.45	2.58	4.48	7.95

spectral magnitudes, according to the ML, MMSE, or MAP criterion. Solutions are provided in general form, so that estimators can be obtained by substituting statistical shape parameters corresponding to desired speech and noise priors.

Table I provides an overview of the novel estimators derived in this paper. It is interesting to note the similarity between certain proposed estimators, and traditional solutions such as MSS [2], the Wiener filter [1], and MAP estimators presented in [8] and [10].

Quantitative analysis of the proposed estimators generally shows improvement in terms of segmental SSNR and COSH distance, when applied to the Noizeus database [28]. Informal listening tests showed proposed estimators to provide apparent improvement in noise suppression, relative to [7]. Though some musical artifacts were present at low SNRs, it was generally less noticeable than for [7]. Additionally, the subtraction factor was applied to decrease the over-attenuation common with spectral subtraction-based methods, improving both quantitative and perceptual results.

Traditionally, Gaussian pdfs were used to model real and imaginary components of speech and noise DFT coefficients. This corresponds to Rayleigh-distributed spectral magnitudes. More recent work has suggested super-Gaussian pdfs for spectral magnitudes. In our paper, the best results are obtained for the one-sided Gaussian distribution for spectral magnitudes of speech and noise. This case can in many respects be considered to lie between the two previously mentioned cases.

The flexibility of the proposed framework makes it applicable to estimating arbitrary signals corrupted by additive noise. Given the expected shape parameters of the target and noise STSA distributions, an optimal spectral magnitude estimator can be derived. Possible extensions of the presented work include enhancement of general audio signals, such as music enhancement [8],[25], and image denoising [26],[27].

APPENDIX I

DERIVATION OF (10)

Using a geometric approach to spectral subtraction, and applying the law of cosines leads to:

$$R_k^2 = A_k^2 + D_k^2 - 2A_k D_k \cos \theta_k \quad (58)$$

Solving for A_k using the quadratic equation results in:

$$A_k = D_k \cos \theta_k + \sqrt{D^2 \cos^2 \theta_k - D_k^2 + R_k^2} \quad (59)$$

Grouping terms in (59) to match those in (8), with $\gamma_k = R_k^2 / D_k^2$ yields:

$$A_k = R_k - \left(\sqrt{\gamma_k} - \sqrt{\gamma_k - \sin^2 \theta_k} - \cos \theta_k \right) D_k \quad (60)$$

APPENDIX II

DERIVATION OF THE IDENTITY IN (40)

Define the integral I_m as:

$$I_m = \int_0^z x^m \exp(cx) dx. \quad (61)$$

For $c=0$, the integral can be easily solved as:

$$I_m \Big|_{c=0} = \frac{z^{m+1}}{m+1} \quad (62)$$

For $c \neq 0$, the integral can be expressed recursively as:

$$I_m \Big|_{c \neq 0} = \begin{cases} \frac{1}{c} z^m \exp(cz) - \frac{m}{c} I_{m-1}, & \text{for } m > 0 \\ \frac{1}{c} (\exp(cz) - 1), & \text{for } m = 0 \end{cases}. \quad (63)$$

Using (63), I_m can be grouped concisely into a summation:

$$I_m = \begin{cases} \frac{\exp(cz)}{c} \sum_{h=0}^m \left[(-1)^h \frac{m!}{(m-h)!} \frac{z^{m-h}}{c^h} \right], & \text{for } c \neq 0 \\ \frac{z}{m+1}, & \text{for } c = 0 \end{cases} \quad (64)$$

REFERENCES

- [1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Wiley, 1949.
- [2] S. F. Boll, *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 27, pp. 113-120, 1979.
- [3] M. Berouti, R. Schwartz, and J. Makhoul, *Enhancement of speech corrupted by acoustic noise*, ICASSP, pp. 208-211, 1979.
- [4] Y. Lu, and P. Loizou, *A geometric approach to spectral subtraction*, Speech Communication, Vol. 50, pp. 453-466, 2008.
- [5] R. J. McAulay and M. L. Malpass, *Speech Enhancement Using a Soft-Decision Noise Suppression Filter*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 28, No.2, pp. 137-145, 1980.
- [6] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. on Acoustics Speech, and Signal Processing, Vol. 32, No. 6, pp. 1109-1121, 1984.
- [7] Y. Ephraim and D. Malah, *Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985.
- [8] P. J. Wolfe and S. J. Godsil, *Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement*, EURASIP J. Appl. Signal Processing, No. 10, pp. 1043-1051, 2003.
- [9] R. Martin, *Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors*, IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 845-856, 2005.
- [10] T. Lotter and P. Vary, *Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model*, EURASIP Journal on Applied Signal Processing, Vol. 7, pp. 1110-1126, 2005.
- [11] I. Cohen, *Speech Enhancement Using Super-Gaussian Speech Models and Noncausal a Priori SNR Estimation*, Speech Communication, Vol. 47, pp. 336-350, 2005.
- [12] 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 Processing, Vol. 15, No. 6, pp. 1741-1752, 2007.
- [13] T. H. Dat, K. Takeda, and F. Itakura, *Generalized Gamma Modelling of Speech and its Online Estimation for Speech Enhancement*, ICASSP, Vol. IV, pp. 181-184, 2005.
- [14] B. Chen and P. C. Loizou, *A Laplacian-based MMSE Estimator for Speech Enhancement*, Speech Communication, Vol. 49, pp. 134-143, 2007.
- [15] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, 1972.
- [16] E. D. Solomenstev, *Cauchy Inequality*, in *Encyclopedia of Mathematics*, Kluwer Academic Publishing, 2001.
- [17] P. Vary, *Noise Suppression by Spectral Magnitude Estimation- Mechanism and Theoretical Limits*, Signal Processing, Vol. 8, pp. 387-400, 1985.
- [18] D. L. Wang and J. S. Lim, *The Unimportance of Phase in Speech Enhancement*, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 30, No. 4, pp. 679-681, 1982.
- [19] ITU-T Rec P.862 Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

- [20] H.-G. Hirsch and D. Pearce, *The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions*, ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium, 2000.
- [21] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [22] D. Pearce, *Enabling New Speech Driven Services For Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends*, AVIOS 2000: Speech Appl. Conf., Vol. 5, May 2000.
- [23] D. Middleton and R. Esposito, *Simultaneous optimum detection and estimation of signals in noise*, IEEE Trans. on Information Theory, vol. 14, pp. 434-444, 1968.
- [24] J. Sohn, N. S. Kim, and W. Sung, *A Statistical Model-Based Voice Activity Detection*, IEEE Signal Processing Letters, Vol. 6, No. 1, pp. 1-3, 1999.
- [25] S.-H. Oh, W.-J. Yoan, Y.-H. Cho, and K.-S. Park, *A New Spectral Enhancement Algorithm in MP3 Audio*, ICCE, pp. 285-286, 2006.
- [26] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, 1989.
- [27] R. Okten, L. Yaroslavsky, K. Egiazarian, and J. Astola, *Transform Domain Approaches for Image Denoising*, Journal of Electronic Imaging, 11(2), pp. 149-156, 2002.
- [28] Y. Hu and P. Loizou, *Subjective evaluation and comparison of speech enhancement algorithms*, Speech Communication, vol. 49, pp. 588-601, 2007.
- [29] S. Kamath and P. Loizou, *A multi-band spectral subtraction method for enhancing speech corrupted by colored noise*, ICASSP, 2002.
- [30] P. Lockwood, J. Boudy, and M. Blanchet, *Non-linear spectral subtraction (NSS) and Hidden Markov Models for robust speech recognition in car environments*, ICASSP, vol. I, pp. 265-268, 1992.
- [31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, 2000.
- [32] <http://webee.technion.ac.il/Sites/People/IsraelCohen/>
- [33] I. Cohen, *Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator*, IEEE Signal Processing Letters, Vol. 9, Issue 4, pp. 113-116, 2002.
- [34] A. H. Gray and J. D. Markel, *Distance Measures for Speech Processing*, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 24, No. 5, pp. 380-391, 1976.