PARAMETERIZED MMSE SPECTRAL MAGNITUDE ESTIMATION FOR THE ENHANCEMENT OF NOISY SPEECH

Colin Breithaupt, Martin Krawczyk, and Rainer Martin

Institute of Communication Acoustics (IKA)
Ruhr-Universität Bochum, 44780 Bochum, Germany
{colin.breithaupt,martin.krawczyk,rainer.martin}@rub.de

ABSTRACT

The enhancement of short-term spectra of noisy speech can be achieved by statistical estimation of the clean speech spectral components. We present a minimum mean-square error estimator of the clean speech spectral magnitude that uses both a parametric compression function in the estimation error criterion and a parametric prior distribution for the statistical model of the clean speech magnitude. The novel parametric estimator has many known magnitude estimators as a special solution and, additionally, affords estimators that combine the beneficial properties of different known solutions. The new estimator is evaluated in terms of segmental SNR, speech distortion, and noise suppression.

Index Terms**—** speech enhancement, MMSE estimation

1. INTRODUCTION

When speech signals are captured in noisy environments, enhancement algorithms increase the listening comfort for users of mobile phones or hearing aids. To reduce the level of noise in a noisy speech signal, adaptive spectral gain functions like the Wiener filter are very effective. The spectral gain is multiplied with the short-time spectrum of a noisy speech signal so that those spectral bins which are dominated by noise are attenuated while spectral components of speech are ideally not affected. The spectral gain functions are either based on optimal estimators of the clean speech spectral coefficients [1, 2] given the observed noisy spectrum or on the estimation of the clean speech spectral magnitudes. For the estimation of the spectral magnitudes of the clean speech, the minimum mean-square error (MMSE) estimation with different compressive weighting functions [3, 4, 5] or different statistical prior models of the clean speech coefficients [6, 7] have been presented. In [8] an MMSE estimator is proposed that combines the two aspects. However, while for the error criterion a compression function is used, the prior of the speech spectral magnitude is found empirically from training data. The empirical prior was chosen as an alternative to the Gaussian model which is known to be an inaccurate model for spectral components of speech [8, 1]. The combination of a variable error criterion and a non-Gaussian speech prior yields a very general form of the MMSE amplitude estimator. Nevertheless, due to the empirical character of the speech prior, the estimator [8] only exists in tabulated form and no analytic description is available.

In this paper, we present the analytic solution to an MMSE estimator of short-time spectral magnitudes of speech in additive and uncorrelated noise that uses both a variable compression function in the error criterion and the chi-distribution as a prior model for speech spectral magnitudes. This novel estimator provides a generic solution to the speech enhancement problem. It also affords many known estimators as special cases.

The paper is organized as follows: In Section 2, speech enhancement based on MMSE amplitude estimation is described and the speech prior introduced. Section 3 gives an analysis of the novel estimator. The evaluation of the estimator is presented in Section 4.

2. MMSE ESTIMATION AND STATISTICAL MODEL

We assume that the spectrum of a noisy speech signal segment is calculated via a short-time Fourier transform resulting in the noisy complex spectral coefficients Y_k, with k the frequency bin index. It is assumed here that the noisy spectrum Y_k is the sum of the clean speech spectrum S_k and the uncorrelated noise spectrum N_k, i.e. $Y_k = S_k + N_k$. Further, N_k is modelled as a complex Gaussian random variable and the phase of S_k is assumed to be uniformly distributed between $-\pi$ and π as was observed in [9].

An optimal estimate $\widehat{A}_k$ of the spectral amplitude $A_k = |S_k|$ can be derived from the minimum mean-square error (MMSE) criterion

$$\widehat{A}_k = \underset{\widehat{A}_k}{\operatorname{argmin}}\, E\left\{ \left| e(A_k, \widehat{A}_k) \right|^2 \,\middle|\, Y_k, P_n(k), \xi_k \right\}, \qquad (1)$$

where the expectation $E\{\cdot\}$ is conditioned on the observed magnitude Y_k, the noise power $P_n(k) = E\left\{|N_k|^2\right\}$ in spectral bin k and the *a priori* signal-to-noise ratio (SNR) given as

The work of C. Breithaupt and M. Krawczyk was funded by the German Research Foundation DFG.

$\xi_k = P_s(k)/P_n(k)$ with $P_s(k) = E\left\{A_k^2\right\}$ the speech power. The estimation error function $e(A_k, \widehat{A}_k) = c(A_k) - c(\widehat{A}_k)$ comprises a compression function $c(\cdot)$, giving a different emphasis on estimation errors of small amplitudes in relation to large amplitudes. Such a non-uniform distortion measure has proven suitable for speech [4].

The solution to (1) considering the assumptions made above is the conditional expected value (see [8])

$$c(\widehat{A}_k) = E\left\{c(A_k)\middle|\, Y_k, P_n(k), \xi_k\right\} = \frac{\int\limits_0^\infty c(a) \mathrm{e}^{-\frac{a^2}{P_n(k)}} \mathrm{I}_0\!\left(\frac{2|Y_k|a}{P_n(k)}\right) p_{A_k}(a)\,\mathrm{d}a}{\int\limits_0^\infty \mathrm{e}^{-\frac{a^2}{P_n(k)}} \mathrm{I}_0\!\left(\frac{2|Y_k|a}{P_n}\right) p_{A_k}(a)\,\mathrm{d}a}, \quad (2)$$

with $\mathrm{I}_0(\cdot)$ the modified Bessel function of order zero, $p_{A_k}(a)$ the *a priori* probability density function (pdf) of A_k, and a denoting a realization of the random variable A_k.

From the general solution (2), we obtain a specific estimator by choosing $p_{A_k}(a)$ and $c(\cdot)$. Like in [6] we use the chi pdf

$$p_{A_k}(a) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{P_s(k)}\right)^{\mu} a^{2\mu-1}\, \mathrm{e}^{-\frac{\mu}{P_s(k)}a^2}, \quad (3)$$

with $\Gamma(\cdot)$ the complete gamma function. (3) provides the shape parameter μ to fit $p_{A_k}(a)$ to empirical clean speech data or to optimize estimation results. For the compression function we use

$$c(x) = x^{\beta} \quad (4)$$

from [5]. This compression contains the power, the magnitude and the root estimator of [8]. In [5] it was shown for $\mu = 1$ and $\beta \to 0$ that the solution (2) even approaches that for the log–compression giving the well-known log-spectral amplitude (LSA) estimator [4, eqn. (20)].

Substituting (3) and (4) into (2), and using [10, (6.643.2)], [10, (9.220.2)], and [10, (9.212.1)], the estimate $\widehat{A}_k$ of the clean speech magnitude becomes

$$\widehat{A}_k = c^{-1}(c(\widehat{A}_k)) = \sqrt{\frac{\xi_k}{\mu + \xi_k}} \left[\frac{\Gamma(\mu + \frac{\beta}{2})}{\Gamma(\mu)} \frac{\Phi(1 - \mu - \frac{\beta}{2}, 1; -\nu_k)}{\Phi(1 - \mu, 1; -\nu_k)}\right]^{\frac{1}{\beta}} \sqrt{P_n(k)} \quad (5)$$

with $c^{-1}(\cdot)$ the inverse of $c(\cdot)$ and $\Phi(a, c; x) = {}_1F_1(a, c; x)$ the confluent hypergeometric function [10, (9.210.1)]. We have $\nu_k = \gamma_k\, \xi_k/(\mu + \xi_k)$ with $\gamma_k = |Y_k|^2/P_n(k)$ the *a posteriori* SNR. (5) is valid for $\mu > 0$ and $\mu + \beta/2 > 0$ with $\beta \neq 0$. Note that this implies that $\beta < 0$ can be a valid choice. The estimator (5) can be tuned by its two parameters μ and β and yields several known estimators depending on the choice of μ and β (see Table 1).

β	μ	Estimator
1	1	STSA [3, eqn. (7)]
$\beta \to 0$	1	LSA [4, eqn. (20)]
$\beta > 0$	1	[5, eqn. (14)], [4, eqn. (13)]
1	$\mu > 0$	[6, eqn. (6)], [7, eqn. (12)]
2	1	[11, eqn. (20)]

Table 1. List of magnitude estimators that are contained in (5) as special cases.

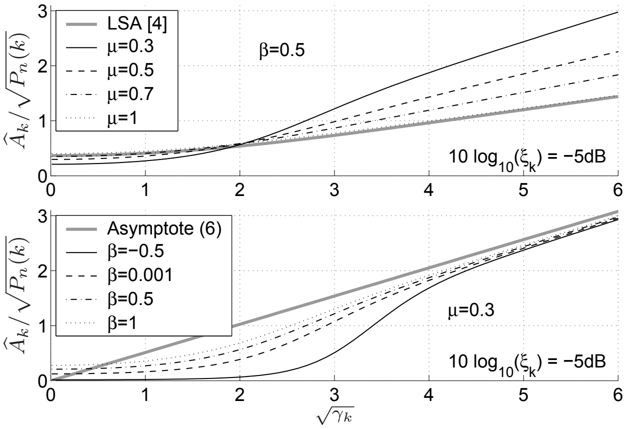

Fig. 1. Input-output mapping characteristics.

3. INPUT-OUTPUT MAPPING CHARACTERISTICS

In order to describe the mapping of input values to the output, we use the input-output characteristics [8] defined as the normalized clean speech spectral estimate $\widehat{A}_k/\sqrt{P_n(k)}$ given the normalized input $|Y_k|/\sqrt{P_n(k)} = \sqrt{\gamma_k}$. This makes the analysis independent of the absolute signal amplitudes. Especially, for Gaussian noise the mean value of noise amplitudes will be $E\left\{\sqrt{\gamma_k}|Y_k = N_k\right\} = \sqrt{\pi}/2$.

In Figure 1 the input-output mapping characteristics are shown for several values of μ and β. The two parameters control different aspects of the mapping characteristics.

For large input values with $\nu_k \gg 1$, the mapping characteristics can be shown with [12, (eqn. (2.17))] to asymptotically approach

$$\left.\frac{\widehat{A}_k}{\sqrt{P_n(k)}}\right|_{\nu_k \gg 1} = \frac{\xi_k}{\mu + \xi_k} \sqrt{\gamma_k}\,. \quad (6)$$

For $\mu = 1$, this is the Wiener solution. A value $\mu < 1$ results in a mapping with a steeper slope than that of the Wiener solution (see Figure 1 top) and $\mu = 0$ gives the identity mapping. Note that in [8, 13] it was observed that a lesser attenuation of large input values is an important property of clean speech spectral estimators. Therefore, values $\mu < 1$ give better estimation results as was already shown in [7].

For $|Y_k| = 0$, the output is

$$\eta_k = \left.\frac{\widehat{A}_k}{\sqrt{P_n(k)}}\right|_{Y_k = 0} = \sqrt{\frac{\xi_k}{\mu + \xi_k} \left[\frac{\Gamma(\mu + \frac{\beta}{2})}{\Gamma(\mu)}\right]^{1/\beta}}\,. \quad (7)$$

For $\mu = 1$ and $\beta \rightarrow 0$, when the LSA estimator is asymptotically approached, we get $\eta_k = \sqrt{\xi_k/(1+\xi_k)}\,\mathrm{e}^{-\mathbf{c}/2}$, with $\mathbf{c} = 0.5772\ldots$ the Euler constant [4]. From Figure 1 we can see that (7) is the lower limit of the output indicating the degree of noise suppression for given ξ_k. From (6) and (7) we find that η_k can be controlled by β without influencing the mapping characteristics for large input values (6).

4. EVALUATION

In the following evaluation, we will show that the parameters μ and β allow to tune the novel estimator (5) in a way that yields a better SNR improvement than the known estimators.

For the experiments, we use (5) in a speech enhancement framework for audio applications. The sampling rate of the system is $f_s = 16\,\text{kHz}$. The noisy signal is segmented in half-overlapping frames of length $M = 512$ taps. Each frame λ is weighted with a M-tap Hann-window and transformed with a discrete Fourier transform (DFT) of length M resulting in the noisy spectral bins $Y_k(\lambda)$. An estimate $\widehat{P}_n(k,\lambda)$ of the noise power for each bin k of a frame λ is obtained by using the method from [14]. An estimate $\widehat{\xi}_k(\lambda)$ of the *a priori* SNR is obtained from the *decision-directed* approach in the form presented in [15], with parameters $\alpha = 0.98$ and $10\log_{10}(\xi_{\min}) = -25\,\text{dB}$. The gain $G_k(\lambda) = \widehat{A}_k(\lambda)\,/|Y_k(\lambda)\,|$ of the filter is bound between values $G_{\min} \le G_k(\lambda) \le 1$. The lower value is set to $20\log_{10}(G_{\min}) = -20\,\text{dB}$ and helps to mask *musical noise* [16]. The upper bound has no audible effect and is used for numerical reasons: As we normally have $\eta_k > 0$ in (7), the gain function has a pole at $|Y_k(\lambda)\,| = 0$. The estimate $\widehat{S}_k(\lambda)$ of clean speech spectral coefficients is obtained from (5) and the noisy phase factor $\mathrm{e}^{j\psi_k(\lambda)} = Y_k(\lambda)/|Y_k(\lambda)\,|$ as [17, 3]:

$$\widehat{S}_k(\lambda) = \widehat{A}_k(\lambda)\ \mathrm{e}^{j\psi_k(\lambda)}\,. \qquad (8)$$

The enhanced spectrum is transformed back with the inverse DFT and the enhanced signal is constructed using the overlap-add method. Note that the use of the tapered Hann-window sufficiently suppresses audible cyclic convolution effects, as $G_k(\lambda)$ is real–valued, thus having zero phase, and the extent of the filter's impulse response is sufficiently small.

As the confluent hypergeometric functions $\Phi(a,c;x)$ in (5) lead to filter implementations of high numerical complexity [12], a function table was used for the term that is taken to the power of $1/\beta$ in (5). For values $\nu_k > 50$, (5) is replaced by its asymptote (6). The maximum relative error between the exact and the tabulated or approximated value was less than 2 percent. Note that while the estimator in [8] uses the function tables to actually describe the mapping characteristics, our approach principally allows an exact analysis using (5), whereas the tabulated values are only used for a more efficient implementation. As for the choice of parameters, we use $\beta = 0.001$ to approximate the LSA-case $\beta = 0$ (see [5]).

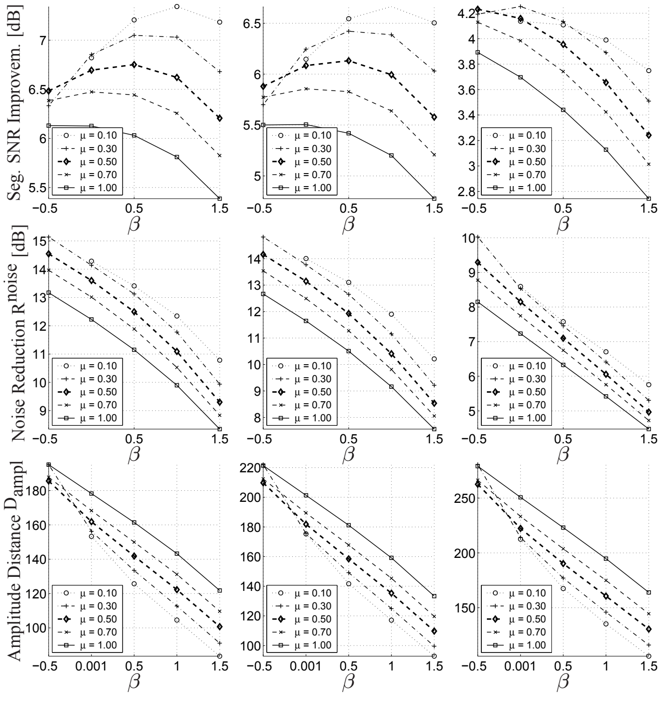

Fig. 2. Average segmental SNR improvement, noise reduction $\mathrm{R}^{\text{noise}}$, and speech amplitude distance D_{ampl} for 12 TIMIT sentences and different noise types. From left to right we have stationary white and pink Gaussian noise, and babble noise.

The enhancement of noisy signals is measured in terms of the segmental SNR, the speech amplitude distance D_{ampl} from [7, eqn. (26)], and the segmental noise reduction $\mathrm{R}^{\text{noise}}$ [9]. Note that less speech distortion is indicated by lower values of D_{ampl}, while higher values of $\mathrm{R}^{\text{noise}}$ indicate a better noise reduction during speech presence.

In Figure 2, the results are shown as the average over 12 speech samples from the TIMIT database [18] for several combinations of the parameters μ and β. The speech samples are disturbed by different noise types at a segmental SNR of 0 dB. As both speech distortion and noise reduction are considered in the segmental SNR, the graphs for this measure exhibit an optimum for the best trade-off. In terms of speech distortion expressed by D_{ampl}, lower values of β increasingly attenuate low–energy speech components, as the range of input values that are strongly attenuated is widened (see Figure 1, bottom graph). Nevertheless, this comes along with an increased noise suppression $\mathrm{R}^{\text{noise}}$. The advantage of a steeper mapping characteristics obtained for lower values of μ is reflected in lower values for D_{ampl}. As the parameter μ also influences the output for low input values (see (7)), lowering μ therefore also improves $\mathrm{R}^{\text{noise}}$. For very low values $\mu = 0.3$ and $\beta = -0.5$, this even results in an increase in the distortion of low–energy speech components as can be seen from the Figures for D_{ampl}.

Informal listening reveals that for low values of μ, *musical noise* is not masked any more by $20\log_{10}(G_{\min}) = -20\,\text{dB}$.

We find that $\mu = 0.5$ is a good compromise between the amount of *musical noise* and clarity of speech. Additionally, we find that a value of $\beta = 0.5$ yields good noise reduction without audible speech distortions. This value also corresponds to the optimal value for the segmental SNR improvement in the case of white and pink noise. As (4) with $\beta = 0.5$ is the root–compression and as the speech amplitude pdf (3) with $\mu = 0.5$ is a super–Gaussian pdf, i.e. its Kurtosis is higher than that of a Gaussian process, we refer to the estimator as super–Gaussian amplitude root (SuGAR) estimator.

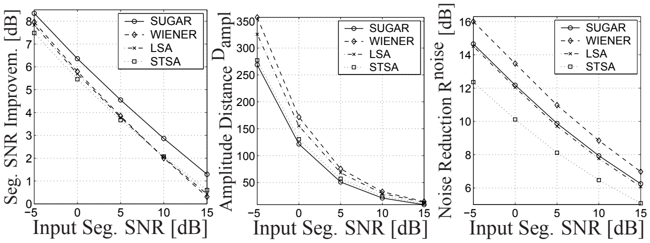

Fig. 3. Averages of segmental SNR improvement, speech amplitude distance $\mathrm{D_{ampl}}$, and noise reduction $\mathrm{R^{noise}}$ for 12 TIMIT sentences and stationary white Gaussian noise at different input segmental SNR. Pink and babble noise yield the same relative results.

In Figure 3, we compare the SuGAR estimator with three well-known estimators for different noise levels. The SuGAR estimator yields a speech distortion as low as the short-time spectral amplitude (STSA) estimator [3, eqn. (7)]. At the same time, it reduces the noise level as well as the LSA estimator [4, eqn. (20)]. For the combined measure, i.e. the segmental SNR improvement, we find that the SuGAR estimator is thus superior by about 0.5 dB. The Wiener filter $\widehat{A}_k = \xi_k/(1+\xi_k)|Y_k|$ achieves the highest noise reduction, as its mapping characteristics is not limited by η_k.

5. CONCLUSION

In this paper we have presented a novel MMSE estimator of speech spectral amplitudes for speech enhancement in noisy environments. The estimator can be varied in its estimation error function and in the shape of the speech prior. This results in an estimator that offers two parameters for an optimization in terms of *musical noise*, speech distortion and noise reduction. In the evaluation, we found that the parameter values $\mu = 0.5$ and $\beta = 0.5$ give a good trade-off between these aspects for high input noise levels. This combination of parameters cannot be realized with any of the estimators in Table 1.

6. REFERENCES

[1] R. Martin, "Speech enhancement based on minimun mean square error estimation and supergaussian priors," *IEEE TSAP*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[2] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 87–90, 2003.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE TASSP*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE TASSP*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE TSAP*, vol. 13, no. 4, pp. 475–486, July 2005.

[6] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," in *IEEE ICASSP*, vol. 3, 2006, pp. 1068–1071.

[7] 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 TASLP*, vol. 15, no. 6, pp. 1741–1752, 2007.

[8] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE ICASSP*, vol. 9, no. 1, pp. 53–56, 1984.

[9] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal of Applied Signal Processing*, vol. 7, pp. 1110–1126, 2005.

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

[11] A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," *IEEE ICASSP*, pp. 201–204, 1999.

[12] K. E. Muller, "Computing the confluent hypergeometric function, M(a,b,x)," *Numerical Mathematics*, vol. 90, pp. 179–196, 2001. [Online]. Available: http://www.bios.unc.edu/∼muller/2001.pdf

[13] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, 2005.

[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE TSAP*, vol. 9, no. 5, pp. 504–512, July 2001.

[15] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, R. C. Dorf, Ed. CRC Press, 2005. [Online]. Available: http://ece.gmu.edu/∼yephraim/ephraim.html

[16] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *IEEE ICASSP*, vol. 2, pp. 789–792, 1999.

[17] P. Vary, "Noise suppression by spectral magnitude estimation – mechanism and theoretical limits –," *EURASIP Signal Processing*, vol. 8, pp. 387–400, 1985.

[18] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.