IMPROVED SPEECH SPECTRAL VARIANCE ESTIMATION UNDER THE GENERALIZED GAMMA DISTRIBUTION

Jan S. Erkelens, Jesper Jensen, and Richard Heusdens

Delft University of Technology, Department of Mediamatics
Information and Communication Theory Group
Mekelweg 4, 2628 CD Delft, The Netherlands
{j.s.erkelens, j.jensen, r.heusdens}@tudelft.nl

ABSTRACT

DFT-based single-microphone speech enhancement methods need an estimate of the clean speech spectral variance. Often the "decision-directed" spectral variance estimator is used, because of its good performance: it strongly reduces the musical noise phenomenon. It has recently been shown that this estimator is severely biased at low SNRs when the smoothing factor approaches one [1]. Here we propose a variance estimator with reduced bias, under the assumption of a generalized Gamma distribution for the clean speech spectral amplitudes. For the reconstruction, the MMSE estimator of the amplitudes themselves is used, derived under the same distribution assumption. For the same speech quality versus noise reduction trade-off, the new variance estimator leads to less musicality.

1. INTRODUCTION

The traditional assumption for speech enhancement in the DFT domain is that the distribution of the complex speech DFT coefficients is Gaussian [2][3]. Consequently, the spectral amplitude distribution is modeled by a Rayleigh distribution. Recently, super-Gaussian models of the DFT coefficients have received quite some attention, because they lead to estimators with better performance than those based on a Gaussian model. Martin [4] derived complex-DFT estimators for Laplacian and Gamma speech priors, and Lotter and Vary [5] proposed a MAP amplitude estimator for a generalized Gamma amplitude distribution. MMSE estimators of the complex DFT coefficients, assuming a two-sided generalized Gamma distribution, have been derived in [6]. MMSE estimators for the amplitudes, assuming a one-sided generalized Gamma distribution, are treated in [7] and [8]. All these estimators need an estimate of the speech spectral variance, for which the "decision-directed" approach is commonly used [2]. The decision-directed estimator contributes much to the good performance of the algorithms, because it strongly suppresses the musical noise [2][9]. However, when the smoothing parameter is very close to one, this estimator has undesirable convergence behavior at low SNRs, leading to over-suppression of the speech [1]. The main reason for this problem is a bias in the square of the MMSE amplitude estimator used in the decision-directed spectral variance estimator. A spectral variance estimator with reduced bias can be found by using the MMSE squared-amplitude estimator instead, as will be shown in this paper. This procedure leads to a decrease of musicality under a generalized Gamma distribution model.

The research is supported by MultimediaN, the Technology Foundation STW applied science division of NWO, and the technology programme of the ministry of Economics Affairs.

2. MMSE SPECTRAL ESTIMATION

2.1. Signal model and assumptions

We consider a signal model of the form

$$X(k,n) = S(k,n) + D(k,n), \tag{1}$$

where $X(k,n), S(k,n)$, and $D(k,n)$ are complex-valued random variables representing the DFT coefficients obtained at frequency index k in signal frame n from the noisy speech, clean speech and noise process, respectively. Applying the standard assumption that $S(k,n)$ and $D(k,n)$ are statistically independent across time and frequency as well as from each other, it turns out that the expressions for the resulting estimators are independent of time and frequency. We drop the time and frequency indices whenever possible for ease of notation. We use capitals for random variables and the corresponding lowercase letters for their realizations. The speech amplitude is $A = |S|$, and the noisy amplitude is $R = |X|$. The noise DFT coefficients D are assumed to follow a complex Gaussian distribution with variance λ_D.
Recent work [5]-[8] has shown that a generalized Gamma distribution assumption for the speech amplitudes can perform much better than the Rayleigh distribution assumption. This generalized Gamma distribution is given by

$$f_A(a) = \frac{\gamma \beta^{\nu}}{\Gamma(\nu)} a^{\gamma\nu-1} \exp(-\beta a^{\gamma}),\ a \geq 0, \tag{2}$$

with the constraints on the parameters $\gamma > 0,\ \nu > 0$. We will consider the cases $\gamma = 1$ and $\gamma = 2$. Because $E\{A^2\}$ equals λ_S by definition, β is related to γ, ν and λ_S. For $\gamma = 1$ we have $\beta^2 = \nu(\nu+1)/\lambda_S$, while $\beta = \nu/\lambda_S$ for $\gamma = 2$. For $\gamma = 2$, the Rayleigh distribution appears as a special case of (2) when $\nu = 1$.

2.2. Amplitude estimators for the generalized Gamma model

The MMSE estimator of some power p of the speech amplitude is (see, e.g., [10]):

$$\widehat{A^p} = E\{A^p|R\} = \frac{\int_0^\infty a^p \exp(-\frac{a^2}{\lambda_D}) I_0(\frac{2aR}{\lambda_D}) f_A(a) da}{\int_0^\infty \exp(-\frac{a^2}{\lambda_D}) I_0(\frac{2aR}{\lambda_D}) f_A(a) da}, \tag{3}$$

where $I_0(\cdot)$ is the zeroth order Bessel function of the first kind. $\widehat{A^p}$ is called the *p-th order* amplitude estimator. MMSE estimators are unbiased [11]. In particular, the expectation of $\widehat{A^2}$ w.r.t. the distribution of R equals the speech spectral variance λ_S, i.e., $\int_0^\infty \widehat{A^2}(r) f_R(r) dr = \lambda_S$. This property will be used in section 3 to improve the spectral variance estimation.

2.2.1. First and second order estimators

MMSE amplitude estimators for the distribution classes $\gamma = 1$ and $\gamma = 2$ have been derived in [8]. For $\gamma = 2$ the derivation is exact, see also [7]. For the case $\gamma = 1$ accurate analytical approximations to the exact MMSE estimators can be made for $\nu > 0.5$[1] [8]. The maximum achievable performance for both classes is about the same. Due to space limitations, we show only the expressions for the estimators of the $\gamma = 2$ class for which the Gaussian speech model occurs as a special case for $\nu = 1$. For $p = 1$, the estimator is

$$\hat{A}_\nu^{(2)} = \frac{\Gamma(\nu+0.5)}{\Gamma(\nu)} \sqrt{\frac{\xi}{\zeta(\nu+\xi)}} \frac{{}_1F_1\left(\nu+0.5; 1; \frac{\zeta\xi}{\nu+\xi}\right)}{{}_1F_1\left(\nu; 1; \frac{\zeta\xi}{\nu+\xi}\right)} R, \tag{4}$$

with the ν-dependency explicitly indicated. The superscript $^{(2)}$ indicates that $\gamma = 2$. ${}_1F_1(a;b;x)$ is a confluent hypergeometric function [12, 13.1.2], $\zeta = R^2/\lambda_D$ is the *a posteriori* SNR, and $\xi = \lambda_S/\lambda_D$ is the *a priori* SNR. For $p = 2$, we arrive at

$$\widehat{A_\nu^2}^{(2)} = \frac{\nu\xi}{\zeta(\nu+\xi)} \frac{{}_1F_1\left(\nu+1; 1; \frac{\zeta\xi}{\nu+\xi}\right)}{{}_1F_1\left(\nu; 1; \frac{\zeta\xi}{\nu+\xi}\right)} R^2. \tag{5}$$

[1] Because of an approximation used in the derivation of MMSE estimators for $\gamma = 1$, only $\nu > 0.5$ is allowed for this class.

For large arguments of the convergent hypergeometric function, and finite ν, (4) and (5) converge to $\xi R/(\nu+\xi)$ and $(\xi R)^2/(\nu+\xi)^2$, respectively [12, 13.5.1].

3. A PRIORI SNR ESTIMATION

3.1. Conventional decision-directed estimator

The *a priori* SNR has to be estimated and a common approach is the "decision-directed" method [2]:

$$\hat{\xi}_1(n) = \max\left[\alpha \frac{\hat{A}^2(n-1)}{\lambda_D(n)} + (1-\alpha)[\zeta(n)-1], \xi_{min}\right], \tag{6}$$

where the subscript 1 refers to the use of the *first* order amplitude estimator $\hat{A}(n-1)$ of the previous time frame. It can be shown that (6) is biased low at low SNRs [1]. The bias propagates in time, affecting future realizations of $\hat{A}$ and $\hat{\xi}$ and causing the latter to converge to ξ_{min} in stationary signals for $\alpha \to 1$, thereby causing too much suppression of weak speech components. This explains why at low SNRs, the estimated *a priori* SNR is a highly smoothed version of the observed *a posteriori* SNR, as was found experimentally by Cappé [9]. For low values of α, the second term $(1-\alpha)[\zeta - 1]$ becomes the most important. This term is unbiased, but has a large variance, causing the residual noise to sound musical [9]. The reason for the bias in (6) for α near one is the following. An MMSE amplitude estimator is a conditional expectation. Its square is used in (6). It is well-known (Jensen's inequality) that the square of an expectation is less than or equal to the expectation of the square. The expectation of the second order MMSE amplitude estimator equals λ_S. Therefore, the expectation of the square of the first order amplitude estimator is smaller than the spectral variance.

3.2. Bias-corrected estimator

The above reasoning suggests to use instead in (6) the second order amplitude estimator (i.e., (3) with $p = 2$), leading to:

$$\hat{\xi}_2(n) = \max\left[\alpha \frac{\widehat{A^2}(n-1)}{\lambda_D(n)} + (1-\alpha)[\zeta(n)-1], \xi_{min}\right]. \tag{7}$$

However, $\sqrt{\widehat{A^2}}$ is not used also for amplitude reconstruction, because it is not the MMSE first order amplitude estimator. We compute *both* the amplitude and second order amplitude estimates for the same prior distribution (2). The second order amplitude estimate is used only for *a priori* SNR estimation, while the amplitude estimate is used only for the speech reconstruction. Note that the bias correction makes it possible to use $\alpha = 1$ in (7), without very severe speech distortions, in contrast to (6).

It can be shown that for $\nu \to \infty$, the first and second order amplitude estimates approach $\sqrt{\hat{\lambda}_S}$ and $\hat{\lambda}_S$, respectively,

where $\hat{\lambda}_S = \hat{\xi}\hat{\lambda}_D$. This is also true for the $\gamma = 1$ case. The reason for this behavior is that for $\nu \to \infty$, (2) tends to a delta-function centered around $\sqrt{\lambda_S}$ for a given λ_S. Consequently, both decision-directed *a priori* SNR estimators (6) and (7) behave like an ordinary exponential smoother for $\nu \to \infty$, resulting in a very reverberant character of the enhanced speech, without any musicality. Of course, these estimators are also identical for $\alpha = 0$.
We will show that the use of (7) leads to a better perceptual quality than the use of (6) for a proper choice of the parameters ν and α.

4. EXPERIMENTAL RESULTS

4.1. Experimental set-up

In the enhancement system, we use 50%-overlapping frames of 32 ms (256 samples at 8 kHz sampling frequency). The data window used was a cosine-squared window. We use all 30 clean sentences of the NOIZEUS database [13]. Noisy signals were generated by adding white and nearly stationary car noise from the Noisex-92 database [14] to the clean signals, at 5 and 15 dB overall SNR. The noise was limited to telephone bandwidth (300-3400 Hz). The noise spectral variance was estimated from 0.64 seconds of noise only, preceding speech activity. Objective quality was measured in two different ways. We measure mean-square error (MSE), because it is the quantity that MMSE estimators should minimize on average. We compute MSE as

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\sum_{k}\left\{a(k,n) - \hat{a}(k,n)\right\}^2, \qquad (8)$$

where $a(k,n)$ and $\hat{a}(k,n)$ are the clean speech spectral amplitude and the estimated amplitude of frequency bin k and time frame n, and N is the number of frames in a speech sentence. To exclude silence intervals, frames with a clean energy more than 40 dB below the maximum clean frame energy of a speech sentence are not taken into account. All results at a given SNR are averages over all test sentences. Furthermore, to quantify the speech distortion versus noise reduction trade-off, we also measure separately segmental Speech Quality (SQ) and Noise Reduction (NR) as in [5], and plot these quantities against each other while varying ν.

4.2. Performance evaluation

We evaluated the estimators for both the $\gamma = 1$ and $\gamma = 2$ classes. For each class, we compared the results of using either (6) or (7) for *a priori* SNR estimation. Figures 1 and 2 compare the standard *a priori* SNR estimator (6) with the corrected one (7), for $\gamma = 1$ and $\gamma = 2$, respectively. We used $\alpha = 0.98$ and $\xi_{min} = -19\ dB$. For a proper choice of ν, about a 10% lower MSE is possible with (7)

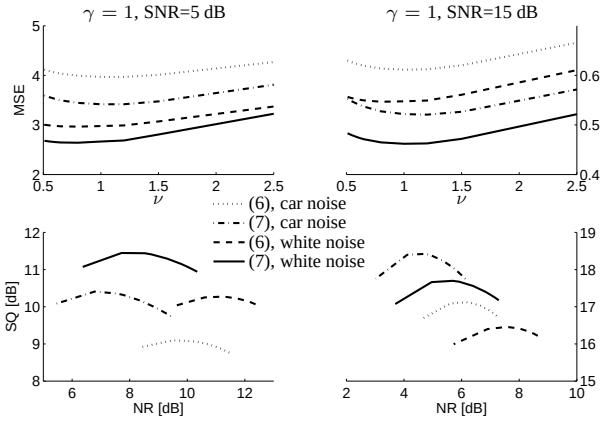

Figure 1: MSE versus ν and Speech Quality versus Noise Reduction for $\gamma = 1$ for the standard *a priori* SNR estimator (6) and the corrected one (7). White and car noise have been used at overall SNRs of 5 and 15 dB.

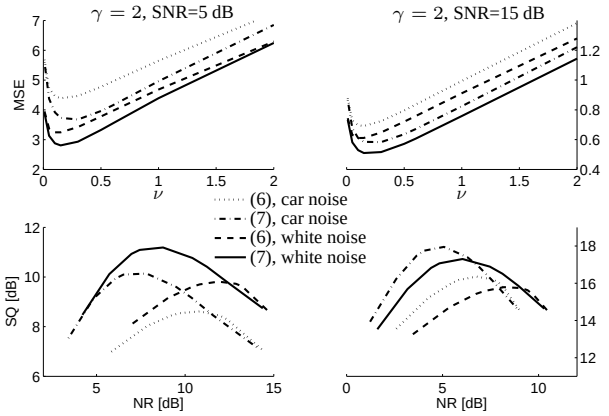

Figure 2: As Figure 1, but for $\gamma = 2$.

than with (6). With (6), for $\gamma = 1$, larger noise reductions are obtained for a given ν, but the maximum achievable speech quality is lower.

It is possible to increase SQ (and decrease MSE) for (6), by lowering the value of α. This will decrease NR and increase the musicality. The question arises whether the gain in objective quality shown for (7) comes at the cost of an increase in musicality. We have found that this is not the case from the following experiment. The quality measures considered are influenced by both parameters ν and α. We evaluated the estimators for a wide range of values. We used fifteen values of α between 0 and 1 and eleven values of ν between 0.01 and 4 for $\gamma = 2$, and between 0.51 and 4.5 for $\gamma = 1$. Figure 3 shows scatter plots of SQ versus NR for all values of ν and α, for $\gamma = 2$, white noise at (a) 5 dB and (b) 15 dB SNR. The dots are for (6) and the pluses for (7). The dots at low SQ that are separated from the main cluster correspond to $\alpha = 1$ in

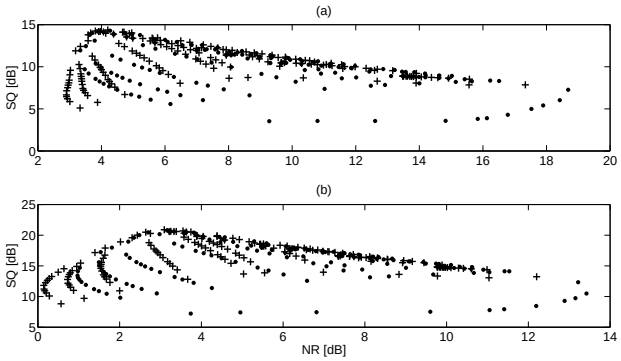

Figure 3: SQ versus NR for $\gamma = 2$. White noise has been used at overall SNRs of (a) 5 dB and (b) 15 dB.

(6). The upper boundary of each cluster corresponds to (ν, α)-pairs that achieve the best possible SQ versus NR trade-off. We see that (6) and (7) have the same upper boundary. However, to a plus and a dot at the same position on this boundary correspond different (ν, α)-pairs, and we have observed that these also achieve almost identical MSEs. However, the value of α for (6) is generally lower and the value of ν higher or the same, compared to the values for (7). Listening to the corresponding enhanced signals revealed more musicality for (6) as a result of the lower value of α. This was especially clear for the white noise and $\gamma = 2$. We may therefore conclude that the bias correction in the decision-directed estimator leads to a decrease in musicality for a given objective quality.

5. SUMMARY AND CONCLUSIONS

We have shown that the bias in the decision-directed spectral variance estimator can be reduced by using the second order amplitude estimator in its definition instead of the square of the first order estimator. With this improved spectral variance estimator, the enhanced speech sounds less musical for the same speech quality versus noise reduction trade-off.

6. REFERENCES

[1] J. S. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement algorithms for various error criteria," *Speech Communication, Special Issue on Speech Enhancement*, 2007, accepted for publication.

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

[3] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech, Audio Processing*, vol. 13, no. 5, pp. 857–869, Sept. 2005.

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

[5] 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, May 2005.

[6] J. Jensen, R. C. Hendriks, J. S. Erkelens, and R. Heusdens, "MMSE estimation of complex-valued discrete Fourier coefficients with generalized Gamma priors," in *Proc. Interspeech*, Sept. 2006.

[7] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, vol. III, pp. 1068 – 1071, May 2006.

[8] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized Gamma distribution," in *Proc. IWAENC*, Sept. 2006.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 2, pp. 345–349, April 1994.

[10] C. H. You, S. N. Koh, and S. Rahardja, "β-order mmse spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 13, pp. 475–486, July 2005.

[11] P. Ishwar and P. Moulin, "On the equivalence of set-theoretic and maxent MAP estimation," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 698–713, March 2003.

[12] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, ninth Dover printing, 1964.

[13] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, vol. I, pp. 153 – 156, May 2006, www.utdallas.edu/∼loizou/speech/noizeus/.

[14] A. Varga and H. J. M. Steeneken, "Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–253, 1993.