A GENERAL OPTIMIZATION PROCEDURE FOR SPECTRAL SPEECH ENHANCEMENT METHODS

Jan Erkelens, Jesper Jensen, and Richard Heusdens

Department of Mediamatics / Information and Communication Theory Group
Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
phone: +31 15 2785859, fax: +31 15 2781843, email: j.s.erkelens@tudelft.nl
web: http://www-ict.ewi.tudelft.nl/

ABSTRACT

Commonly used spectral amplitude estimators, such as those proposed by Ephraim and Malah, are only optimal when the statistical model is correct and the speech and noise spectral variances are known. In practice, the spectral variances have to be estimated. A simple analysis of the ”decision-directed” approach for speech spectral variance estimation reveils the presence of an important bias at low SNRs. To correct for modeling errors and estimation inaccuracies, we propose a general optimization procedure, with two gain functions applied in parallel. The unmodified algorithm is run in the background, but for the final reconstruction a different gain function is used, optimized for a wide range of signal-to-noise ratios. When this technique is implemented for the algorithms of Ephraim and Malah, a large improvement is obtained (in the order of 2 dB Segmental SNR improvement and 0.3 points increase in PESQ). Moreover, less smoothing is needed in the decision-directed spectral variance estimator.

1. INTRODUCTION

Single-microphone speech enhancement is important for many applications [1]. Techniques in the Short-Time Fourier domain are often used, because they are fast, perform well and the statistical modeling in the frequency domain is simple. Minimum Mean-Square Error (MMSE) estimators of the spectral amplitudes [2] or log spectral amplitudes [3], based on the assumption of a Rayleigh distribution for the amplitudes, are commonly used, but more general distribution assumptions have been made as well [4], and also estimators based on Laplace and Gamma distributions for the real and imaginary parts of the Fourier coefficients have been proposed [5].

Spectral speech enhancement algorithms can suffer from an annoying artefact, called ”musical noise”. In [2], a spectral variance estimator, termed ”decision-directed” variance estimator, was proposed which reduces the musical noise at the expense of smoothing of speech transitions. The decision-directed estimator combines the estimated amplitude of the previous analysis frame with the noisy amplitude of the current frame into one estimator of the spectral variance. Although it reduces the musical noise, it is heuristic in nature, lacking a solid theoretical basis. We will investigate the decision-directed estimator in some detail (Section 2). It will be shown that this estimator can be severely biased at low SNR. Correcting fully for the bias is difficult because of the nonlinear feedback loop and because the bias depends on the true SNR, which is unknown. Instead, we propose a general optimization method to improve spectral enhancement methods, using a two-stream structure, in section 3. The standard algorithm is run in the background, but the final reconstruction is made by applying a *different, separate* gain function to the noisy amplitude. This gain function is obtained from a training procedure and is optimized for a wide range of SNRs. The method is applied to the algorithms of Ephraim and Malah in section 4. The results show a significant increase in noise reduction. At the same time, less smoothing is needed in the decision-directed estimator, which may result in better intelligibility. Section 5 concludes the paper.

2. THE DECISION-DIRECTED SPECTRAL VARIANCE ESTIMATOR

2.1 MMSE (log) spectral amplitude estimation

When the Short-Term Fourier coefficients are assumed to be independent across time and frequency and distributed according to a complex Gaussian distribution, the MMSE amplitude estimator, $\widehat{A}_k$, for frequency bin k, is given by [2]:

$$\widehat{A}_k = \frac{\sqrt{\pi v_k}}{2\gamma_k} M(-0.5; 1; -v_k) R_k, \tag{1}$$

where R_k is the noisy spectral amplitude, $M(a;c;x)$ is the confluent hypergeometric function, and v_k is defined by

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k.$$

The *a priori* and *a posteriori* SNRs are defined by $\xi_k = \lambda_x(k)/\lambda_d(k)$ and $\gamma_k = R_k^2/\lambda_d(k)$, respectively. $\lambda_x(k)$ and $\lambda_d(k)$ are the speech and noise spectral variances for frequency bin k. The MMSE log spectral amplitude estimator is [3]:

$$\widehat{A}_k = \frac{\xi_k}{1+\xi_k} \exp\left\{ \frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t}\, dt \right\} R_k. \tag{2}$$

Both (1) and (2) can be written in the form $\widehat{A}_k = G(\xi_k, \gamma_k) R_k$, where G is a spectral gain applied to the noisy amplitude R_k. The speech and noise spectal variances are unknown in practice and have to be estimated. In the following, we assume that the noise spectral variance can be estimated accurately. It can be estimated for stationary noise during speech pauses. For non-stationary noise, approaches based on minimum-statistics [6], [7] can be used. In this paper, we consider stationary noise only and will focus on the estimation of the speech spectral variance.

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

2.2 Speech spectral variance estimation

Ephraim and Malah [2] proposed the following decision-directed estimator for the *a priori* SNR in time frame n in frequency bin k:

$$\hat{\xi}_k(n) = \alpha \frac{\widehat{A}_k^2(n-1)}{\lambda_d(k)} + (1-\alpha)P[\gamma_k(n)-1]$$

$P[x]$ is the clipping function: it sets negative values to zero. The weighting coefficient α is usually chosen near one, e.g. 0.98. A value near one gives the highest noise reduction, while avoiding the musical noise. However, it comes at the expense of a reduction in intelligibility, because important speech transitions are smoothed. Usually, $\hat{\xi}_k$ is constrained to be larger than a certain minimum value ξ_{min}. This helps in reducing musical noise [8]. We will use $\xi_{min} = -19$ dB.

2.2.1 *Convergence behavior*

As indicated by Martin [5], the clipping causes a bias, which can be reduced by letting the clipping operator work on both terms together. The spectral variance estimator thus becomes:

$$\hat{\xi}_k(n) = \max\left[\alpha \frac{\widehat{A}_k^2(n-1)}{\lambda_d(k)} + (1-\alpha)[\gamma_k(n)-1], \xi_{min}\right]. \quad (3)$$

However, there is a bias due to the term $\widehat{A}_k^2(n-1)/\lambda_d(k)$ as well. The expectation of the square of a speech spectral amplitude $A_k^2(n)$ equals $\lambda_x(k,n)$ by definition. Suppose the algorithm with $\alpha = 1$ is applied to a *stationary* stochastic signal. From (1) we can see that for small values of $\hat{\xi}_k(n)$, $\widehat{A}_k^2(n)$ is nearly equal to $(\pi/4)\hat{\xi}_k(n)\lambda_d(k)$, since the hypergeometric function $M(-0.5;1;-v_k)$ is then close to one. This means that even if $\lambda_x(k,n)$ were known exactly, i.e., $\hat{\xi}_k(n) = \lambda_x(k,n)/\lambda_d(k)$, we would have $\widehat{A}_k^2(n) \approx (\pi/4)\lambda_x(k,n)$, i.e., a biased estimator of $\lambda_x(k,n)$ at low SNRs, because of the factor $\pi/4$. Note that the bias is caused by an *inconsistency* between (1) and (3): the square of an estimate of the amplitude is used in (3) instead of an estimate of the square. At very high SNRs, $\widehat{A}_k(n) \approx R_k(n)$ and there is no significant bias in $\widehat{A}_k^2(n)$.

The fact that $\lambda_x(k,n)$ has to be estimated, makes things worse. Because of the factor $\pi/4$, $\hat{\xi}_k(n+1)$ tends to be smaller than $\hat{\xi}_k(n)$. Therefore, in stationary signals, at low SNR, $\hat{\xi}_k$ will converge to ξ_{min}, and $\widehat{A}_k^2$ to $(\pi/4)\xi_{min}\lambda_d$, which generally leads to too much suppression. For $\alpha < 1$, the term $(1-\alpha)[\gamma_k(n)-1]$ counteracts this to some extent and is therefore necessary, but it has a large variance for values of α that are too small, causing musical noise and less than optimal suppression of the noise. For $\alpha \to 1$, the *estimated* spectral amplitude will have a very low variance, much lower than the variance of the true spectral amplitude A_k. This explains why at low SNRs, the estimated *a priori* SNR is a highly smoothed version of the *a posteriori* SNR, as was observed experimentally by Cappé [8]. We can conclude that $\alpha = 1$ is not optimal, not even for stationary signals, because it generally causes too much suppression. The bias is a function of the true SNR, which is unknown. It is therefore difficult to correct for it. Ephraim and Malah [2] have pointed out that an overestimation of ξ is more appropriate than using an underestimate, because the gain function G is less sensitive to an overestimate of ξ than to an underestimate. We therefore choose to correct for the bias for low SNR, by inserting a factor $4/\pi$ into (3):

$$\hat{\xi}_k(n) = \max\left[\alpha \frac{4}{\pi} \frac{\widehat{A}_k^2(n-1)}{\lambda_d(k)} + (1-\alpha)[\gamma_k(n)-1], \xi_{min}\right]. \quad (4)$$

2.2.2 *Illustration for stationary signals*

Figure 1 shows the effect of α when the algorithm with (1) and (3) or (4) is applied to a stationary stochastic signal. The blue continuous line shows the clean spectrum, the green dashed line the enhanced spectrum when (3) is used, the black dash-dotted line the enhanced spectrum with (4), and the horizontal red dotted line indicates the noise level. The overall SNR was 10 dB. It can be seen that there is not enough noise suppression for low values of α. For larger values of α, there is a bias in the enhanced spectrum when (3) is used. This bias increases with increasing α and decreasing *a priori* SNR. The bias-corrected estimator (4) clearly leads to much less signal distortion, although less of the noise is suppressed in very low SNR regions of the spectrum. Similar effects happen with the log-amplitude estimator (2). Speech processed with (3) sounds heavily distorted for $\alpha \to 1$.

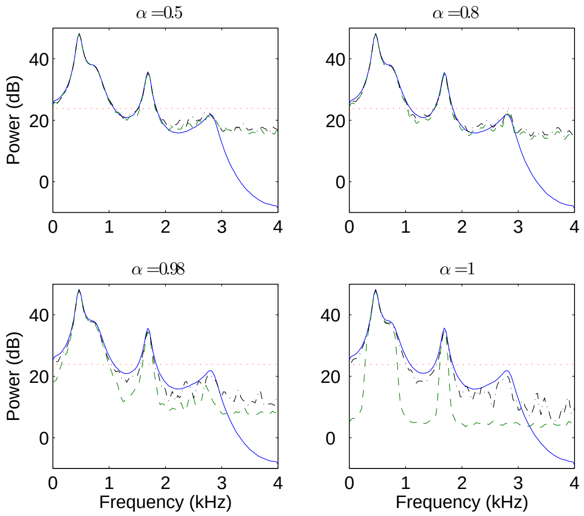

Figure 1: Influence of α on enhancement result for stationary stochastic signals. The blue continuous line is the clean spectrum, the green dashed line is the average enhanced spectrum when (3) is used, the black dash-dotted line results with (4), and the red dotted line is the noise level.

2.2.3 *Results on speech signals*

Tables 1 and 2 show the average Segmental SNR improvement (SSNR+) and scores from the latest PESQ measure [9] for the standard algorithm (EM with (1); EMlog with (2)), for the *a priori* SNR estimators (3) and/or (4). All 30 clean sentences of the NOIZEUS database [10] have been used, to which white or car noise from the Noisex-92 [11] database, limited to telephone bandwidth (300-3400 Hz), has been added. A randomly chosen section of the noise was added to

Table 1: Segmental SNR improvement (SSNR+) and PESQ scores on the NOIZEUS sentences for the algorithm of Ephraim and Malah (EM/EMlog), with the conventional *a priori* SNR estimator and with a bias-corrected version, as a function of α and overall SNR for telephone-bandwidth-filtered white noise.

	SNR	EM with (3)		EM with (4)		EMlog with (3)	
α	(dB)	SSNR+	PESQ	SSNR+	PESQ	SSNR+	PESQ
0.5	0	2.49	1.29	2.33	1.27	3.01	1.34
	5	2.45	1.63	2.28	1.61	2.96	1.71
	10	2.33	2.09	2.17	2.05	2.83	2.19
	15	2.12	2.55	1.96	2.51	2.59	2.66
0.8	0	3.11	1.34	2.69	1.28	3.81	1.42
	5	2.97	1.69	2.56	1.61	3.64	1.82
	10	2.75	2.15	2.36	2.04	3.40	2.31
	15	2.43	2.57	2.06	2.46	3.03	2.74
0.98	0	4.26	1.43	3.11	1.24	4.85	1.49
	5	3.66	1.82	2.66	1.57	4.28	1.92
	10	3.01	2.20	2.17	1.97	3.62	2.31
	15	2.31	2.55	1.66	2.36	2.83	2.69

Table 2: Segmental SNR improvement (SSNR+) and PESQ scores on the NOIZEUS sentences for the algorithm of Ephraim and Malah (EM/EMlog), with the conventional *a priori* SNR estimator and with a bias-corrected version, as a function of α and overall SNR for telephone-bandwidth-filtered car noise.

	SNR	EM with (3)		EM with (4)		EMlog with (3)	
α	(dB)	SSNR+	PESQ	SSNR+	PESQ	SSNR+	PESQ
0.5	0	2.33	1.49	2.18	1.48	2.78	1.55
	5	2.16	1.93	2.01	1.90	2.59	2.01
	10	1.98	2.41	1.84	2.37	2.39	2.50
	15	1.87	2.88	1.73	2.84	2.28	2.99
0.8	0	2.83	1.55	2.46	1.48	3.40	1.63
	5	2.55	1.99	2.20	1.90	3.10	2.12
	10	2.29	2.44	1.95	2.34	2.82	2.59
	15	2.11	2.87	1.78	2.76	2.65	3.04
0.98	0	3.47	1.56	2.64	1.40	3.88	1.61
	5	2.84	2.01	2.14	1.81	3.27	2.10
	10	2.33	2.43	1.70	2.24	2.84	2.56
	15	1.95	2.82	1.38	2.64	2.45	3.00

every sentence at every SNR condition. The noise variance $\lambda_d(k)$ was estimated from 0.64 seconds of noise only, preceding each speech sentence. The bias-corrected *a priori* SNR estimator (4) gives somewhat lower objective performance than (3). There was more residual noise, mainly because $\hat{\xi}$ of (4) does not reach ξ_{min} for very low SNRs. The enhanced speech was also free from musical noise, for $\alpha = 0.98$, but not for the smaller values of α. This suggests that *smoothness across time*, i.e., strong correlation in the time series of *a priori* SNR estimators and corresponding enhanced amplitudes, rather than a low (ensemble) *variance*, is sufficient for avoiding the musical noise. The optimal value of α depends on the input SNR. For the lowest SNRs, the optimal value of α is near 0.98, while for the highest SNR it is near 0.8, according to SSNR+ and PESQ. However, there is much more musical noise for $\alpha = 0.8$, so there seems to be a disagreement between these objective measures and subjective quality at high SNRs.

2.2.4 *Other statistical models*

We can conclude sofar that the performance of the estimators (1) and (2) depends on the properties of the particular *a priori* SNR estimator used. It has been shown that the distribution of clean speech spectral amplitudes, *conditional* on a small range of (high) values of the *estimated a priori* SNR deviates from Gaussianity and a Laplacian, Gamma, or more general distribution model for the real and imaginary parts or the amplitudes of the Fourier coefficients can lead to improved speech enhancement [4] [5]. However, for a different estimator of the *a priori* SNR based on GARCH models [12], a Gaussian model leads to better results than Gamma or Laplacian models. A slight preference for complex Gaussian distributions has also been found for the DFT-coefficients from short analysis frames of individual speech sound classes (vowels, plosives, fricatives, etc.) [13].

The gain function and the decision-directed variance estimator are linked in a non-linear feedback loop. Changing either one of them will affect the performance of the other. We propose a general optimization method to improve speech enhancement algorithms for a wide range of SNR conditions. Two gain functions are applied in parallel. The unmodified algorithm is run in the background, but for the final reconstruction a different gain function is applied, which corrects for some of the modeling errors and estimation inaccuracies in the spectral variance estimator *and* the gain function of the original method. The corrective gain function is optimized for a wide range of SNRs by means of a training procedure on a speech database, described in section 3.1.

3. AN IMPROVED MAPPING

The Ephraim-Malah suppression rules are functions of the *a priori* and *a posteriori* SNRs. This remains true for non-Gaussian assumptions about the distribution of DFT-coefficients. The decision-directed estimator of the *a priori* SNR is a function of the estimated amplitude of the previous frame and the noisy amplitude of the current frame. This means that for spectral speech enhancement algorithms that use the decision-directed variance estimator, we can symbolically write:

$$\widehat{A}_k(n) = F\left(\frac{\widehat{A}_k^2(n-1)}{\lambda_d(k)}, \frac{R_k^2(n)}{\lambda_d(k)}\right) R_k(n), \tag{5}$$

where F is a complicated nonlinear function which, of course, also depends on α, and ξ_{min}. Our goal is to find a function F which leads to better speech enhancement performance (in terms of a suitable objective error criterion). This is a difficult problem. As a first step towards this goal, a training procedure is used to find an improved mapping from the estimated *a priori* SNR and the *a posteriori* SNR to the enhanced amplitudes. This is implemented as a correction to the conventional algorithm, using a two-stream procedure, as follows: $\widehat{A}_k$ of (1) or (2) will not be used for reconstruction of the speech signal, but only for estimation of the *a priori* SNR in the next frame. For reconstruction, we use an amplitude $\widetilde{A}_k$, obtained by multiplying the noisy amplitude R_k by a *separate* gain function $\widetilde{G}(\hat{\xi}_k, \gamma_k)$, different from the gain function $G(\hat{\xi}_k, \gamma_k)$. In other words, the conventional algorithm is run in the background, forming one stream, and the final reconstruction is another stream. This procedure guarantees an improvement in terms of the error criterion, because the first stream is left untouched, while the second-stream gain function $\widetilde{G}$ corrects for some of the modeling errors and estimation inaccuracies. For each frequency bin k, the corrective gain function $\widetilde{G}$ is a function of the two parameters, $\hat{\xi}_k$ and γ_k, so we can write:

$$\widetilde{A}_k = \widetilde{G}(\hat{\xi}_k, \gamma_k) R_k.$$

$\widetilde{G}$ is implemented as a look-up table: the support of $\hat{\xi}_k$ and γ_k is discretized in a grid. The grid points range from -19 dB to 40 dB in steps of 1 dB. Each *parameter cell* contains the values of $\hat{\xi}$ and γ closest to the grid point and has its corresponding value of $\widetilde{G}(\hat{\xi}, \gamma)$ stored in a matrix.

3.1 The training procedure

Our aim is to find the function $\widetilde{G}(\hat{\xi}_k, \gamma_k)$ that minimizes the mean-square error in $\widetilde{A}_k$ or $\log[\widetilde{A}_k]$ for a wide range of SNR conditions. We have optimized over the range -15 dB to 25 dB overall SNR. This covers the range of practical interest.

$\widetilde{G}$ is found by means of a training procedure. We have trained on the TIMIT-TRAIN database [14]. To the clean signals, noise is added at the various overall SNRs. Then the Ephraim-Malah algorithm is run. In each frame, for each frequency bin, we have a $(\hat{\xi}_k, \gamma_k)$ pair that falls into one of the parameter cells. $(\hat{\xi}_k, \gamma_k)$ pairs from different frequency bins and different frames can fall into the same parameter cell during the course of the training. To each of those $(\hat{\xi}_k, \gamma_k)$ pairs corresponds a clean amplitude A_k and a noisy amplitude R_k. Those are collected and after all the train signals are processed, the optimal value of $\widetilde{G}_{ij}$ for parameter cell (i, j) is found by minimizing

$$\sum_{m=1}^{M_{ij}} \left\{ A_{ij}(m) - G_{ij} R_{ij}(m) \right\}^2$$

with respect to G_{ij}. $R_{ij}(m)$ is the m-th noisy amplitude that fell into parameter cell (i, j) and $A_{ij}(m)$ the corresponding clean amplitude. The optimal gain $\widetilde{G}_{ij}$ is given by:

$$\widetilde{G}_{ij} = \sum_{m=1}^{M_{ij}} A_{ij}(m) R_{ij}(m) / \sum_{m=1}^{M_{ij}} R_{ij}^2(m). \qquad (6)$$

The corresponding expressions for the logarithmic case are:

$$\sum_{m=1}^{M_{ij}} \left\{ \log\left[A_{ij}(m) / G_{ij} R_{ij}(m) \right] \right\}^2, \; \widetilde{G}_{ij} = \sqrt[M_{ij}]{\prod_{m=1}^{M_{ij}} \frac{A_{ij}(m)}{R_{ij}(m)}}.$$

Some combinations of $\hat{\xi}_k$ and γ_k are highly unlikely and may not, or not often enough, have occured during the training. This means that M_{ij} for that cell is too small to have a reliable $\widetilde{G}_{ij}$. In such cases, $\widehat{A}_k(n)$ is used for reconstruction.

For training we used the entire TIMIT-TRAIN database [14], which consists of about 900,000 frames of speech. The speech signals were bandpass filtered to telephone bandwidth (300-3400 Hz). Bandpass-filtered computer-generated white noise was added to the train data at overall SNRs ranging from -15 dB to +25 dB, in steps of 5 dB (all the train data were subjected to all SNR conditions). Whenever a certain parameter cell was hit less than 10^4 times in total for all noise conditions during training, the gain function (1) or (2) was substituted, as a function of $\hat{\xi}_k(n)$ and $\gamma_k(n)$, i.e., $\widehat{A}_k(n)$ would be used for reconstruction. Figure 2(a) shows the resulting function $\widetilde{G}$ of (6) for $\alpha = 0.8$. Figure 2(b) shows the gain function of (1) for comparison. Particularly noticeable is that there is more suppression with the second-stream gain function $\widetilde{G}$ than with the first-stream gain function G for small values of $\hat{\xi}$ and γ.

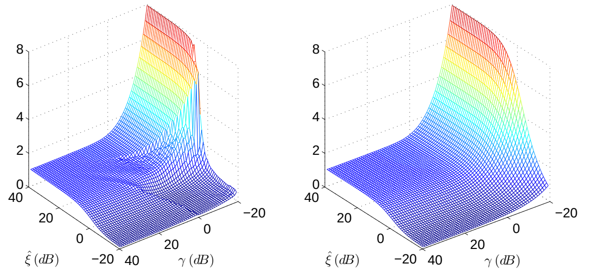

Figure 2: (a) Trained second-stream gain function (6) for $\alpha = 0.8$ and (b) analytical first-stream gain function [2].

4. EXPERIMENTAL RESULTS

Both in training and testing we used 50%-overlapping frames of 32 ms (256 samples at 8 kHz sampling frequency). The data window used was a cosine-squared window, which has the perfect reconstruction property. For testing, we used all 30 clean sentences of the NOIZEUS database [10]. Table 3 shows the average Segmental SNR improvement and PESQ [9] scores for the *Two-Stream* algorithm based on the MMSE amplitude estimator (TS) and MMSE log amplitude estimator (TSlog).[1] The speech was contaminated by white noise from the Noisex database [11], limited to telephone bandwidth. Results are shown for three values of α at four overall SNRs. For Segmental SNR computation, frame SNR values outside the range -10 dB to $+35$ dB are clipped. Table 4 shows the results for contamination by telephone-bandwidth car noise, again from Noisex.

The two-stream method performs much better than the standard algorithm (compare with tables 1 and 2). Note that (3) and (4) perform equally well now. Improvements in SSNR+ are in the order of 1.5 to 3 dB, and in the range of 0.2 to 0.5 points for PESQ. Also, the optimum performance is at a lower value of α now. This is important, since it means that important speech transitions are less smoothed, which should result in less intelligibility reduction. The enhanced signals had much less residual noise, less speech distortion, but some musical noise was introduced. The amount of musical noise was almost independent of α, but increased with decreasing SNR. PESQ scores are almost independent of α as well and of the error criterion used. The MMSE log-amplitude estimator resulted in more noise suppression than the MMSE amplitude estimator. This was clearly audible.

[1] We used the same error criterion in both streams, although this is not strictly necessary.

Table 3: Segmental SNR improvement (SSNR+) and PESQ scores on the NOIZEUS sentences for the Two-Stream algorithm (TS/TSlog) as a function of α and overall SNR for telephone-bandwidth white noise.

		TS with (3)		TS with (4)		TSlog with (3)	
α	SNR	SSNR+	PESQ	SSNR+	PESQ	SSNR+	PESQ
0.5	0	5.96	1.66	5.96	1.66	6.36	1.64
	5	5.55	2.12	5.55	2.12	5.91	2.14
	10	5.00	2.60	5.00	2.60	5.37	2.61
	15	4.30	3.07	4.31	3.07	4.63	3.06
0.8	0	5.96	1.69	5.96	1.69	6.42	1.67
	5	5.53	2.14	5.52	2.14	5.93	2.15
	10	4.97	2.60	4.96	2.60	5.39	2.61
	15	4.28	3.07	4.27	3.07	4.64	3.06
0.98	0	5.85	1.69	5.82	1.69	6.53	1.67
	5	5.35	2.12	5.35	2.12	5.97	2.11
	10	4.74	2.55	4.77	2.56	5.33	2.55
	15	4.03	3.00	4.07	3.02	4.52	3.01

Table 4: Segmental SNR improvement (SSNR+) and PESQ scores on the NOIZEUS sentences for the Two-Stream algorithm (TS/TSlog) as a function of α and overall SNR for telephone-bandwidth car noise.

		TS with (3)		TS with (4)		TSlog with (3)	
α	SNR	SSNR+	PESQ	SSNR+	PESQ	SSNR+	PESQ
0.5	0	4.89	1.75	4.91	1.75	5.32	1.77
	5	4.46	2.32	4.46	2.32	4.83	2.32
	10	4.13	2.84	4.14	2.84	4.51	2.86
	15	3.86	3.32	3.86	3.32	4.19	3.33
0.8	0	4.91	1.78	4.91	1.78	5.39	1.79
	5	4.46	2.35	4.45	2.35	4.88	2.35
	10	4.13	2.85	4.12	2.85	4.54	2.87
	15	3.83	3.32	3.83	3.32	4.19	3.33
0.98	0	4.75	1.76	4.73	1.77	5.40	1.76
	5	4.28	2.32	4.29	2.33	4.83	2.31
	10	3.87	2.81	3.91	2.83	4.41	2.84
	15	3.54	3.26	3.61	3.29	4.02	3.29

5. CONCLUDING REMARKS

Errors in the statistical models and the estimated model parameters decrease the performance of suppression rules. We have located a large bias in the decision-directed approach of spectral variance estimation, which causes serious speech distortion when the weight factor approaches one.

We have shown for the standard MMSE speech spectral amplitude and log-amplitude estimators with the decision-directed approach for spectral variance estimation, that the performance can be much improved by a two-stream structure. The procedure can be used to optimize for other, perceptually more relevant error criteria, such as those in [15], as long as frequency bins are treated independently. Other *a priori* SNR estimators may also be used. Complex DFT estimators can be handled easily, when the real and imaginary parts are assumed independent and identically distributed [5]. The resulting optimized gain functions for the real and imaginary parts will be the same.

In our two-stream approach, a conventional analytical gain function and a trained corrective gain function are used. We will investigate whether it is possible to optimize the single function F of (5). With the parameters shown in this equation, there might be no need for a separate spectral variance estimator, since that would be included in the function F automatically.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, Berlin: Springer, 2005.

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

[3] ———, ”Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 33, no. 2, pp. 443-445, Apr. 1985.

[4] T. Lotter, and P. Vary, ”Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model,” *EURASIP Journ. Appl. Signal Proc.*, vol.7, pp. 1110-1126, 2005.

[5] R. Martin, ”Statistical methods for the enhancement of noisy speech,” pp. 43-64 in [1].

[6] R. Martin, ”Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Proc.*, vol. 9, no. 5, pp. 504-512, July 2001.

[7] I. Cohen, ”Noise estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 5, pp. 466-475, Sept. 2003.

[8] O. Cappé, ”Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Proc.*, vol. 2, no. 4, pp. 345-349, Apr. 1994.

[9] J. G. Beerends, ”Extending P.862 PESQ for assessing speech intelligibility,” White contribution COM 12-C2 to ITU-T Study Group 12, October 2004 (equivalent to TNO Information and Communication Technology report 33392).

[10] *NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms.*
http://www.utdallas.edu/∼loizou/speech/noizeus/

[11] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, ”The noisex-92 study on the effect of additive noise on automatic speech recognition,” Technical report, DRA Speech Research Unit, 1992.

[12] I. Cohen, ”Supergaussian GARCH models for speech signals,” *Proc. Interspeech*, pp. 2053-2056, Lisbon, Portugal, Sept. 2005.

[13] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, ”A study of the distribution of time-domain speech samples and discrete fourier coefficients,” *Proc. IEEE SPS-DARTS*, pp. 155-158, 2005.

[14] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus*, National Institute of Standards and Technology. NTIS order no. PB91-505065, 1990.

[15] P. C. Loizou, ”Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum,” *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 5, pp. 857-869, Sept. 2005.