A data-driven approach to optimizing spectral speech enhancement methods for various error criteria

Jan Erkelens *, Jesper Jensen, Richard Heusdens

Department of Mediamatics, Information and Communication Theory Group, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

Received 31 January 2006; received in revised form 26 June 2006; accepted 27 June 2006

Abstract

Gain functions for spectral noise suppression have been derived in literature for some error criteria and statistical models. These gain functions are only optimal when the statistical model is correct and the speech and noise spectral variances are known. Unfortunately, the speech distributions are unknown and can at best be determined conditionally on the estimated spectral variance. We show that the “decision-directed” approach for speech spectral variance estimation can have an important bias at low SNRs, which generally leads to too much speech suppression. To correct for such estimation inaccuracies and adapt to the unknown speech statistics, we propose a general optimization procedure, with two gain functions applied in parallel. A conventional algorithm is run in the background and is used for a priori SNR estimation only. For the final reconstruction a different gain function is used, optimized for a wide range of signal-to-noise ratios. The gain function providing for the reconstruction is trained on a speech database, by minimizing a relevant error criterion. The procedure is illustrated for several error criteria. The method compares favorably to current state-of-the-art methods, and needs less smoothing in the decision-directed spectral variance estimator.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Speech enhancement; Spectral distortion measures; Speech model

1. Introduction

Single-microphone speech enhancement is important for many applications (Benesty et al., 2005). 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 (Ephraim and Malah, 1984) or log spectral amplitudes (Ephraim and Malah, 1985), based on the assumption of a Rayleigh distribution for the amplitudes, are commonly used, but more general distribution assumptions have been made as well (Lotter and Vary, 2005), and also estimators based on super-Gaussian distributions for the real and imaginary parts of the Fourier coefficients have been proposed, such as Laplace and Gamma distributions (Martin, 2005a). The latter methods are data-driven methods in the sense that optimal estimators are derived for distribution models that fit to observed speech distributions. The analytical derivation of estimators is only possible for some error criteria and certain statistical models. Porter and Boll (1984) used a data-driven method to calculate estimators directly from clean speech. In their work no estimator of spectral variance was used and the estimators were adapted to global speech distributions only.

Perhaps the most commonly used estimator of speech spectral variance is the “decision-directed” variance estimator (Ephraim and Malah, 1984). It greatly reduces an annoying artefact of spectral enhancement methods, called “musical noise” (Cappé, 1994). 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 may lead to smoothing of...
important speech transitions. We will investigate the decision-directed estimator in some detail (Section 2). It will be shown that this estimator is biased at low SNR. The bias makes it converge to a minimum value in low SNR regions, leading to very smooth behavior across time, but causing also significant speech suppression. Correcting fully for the bias is difficult because of the non-linear feedback loop and because the bias depends on the true SNR and speech distributions, which are unknown. To deal with these problems and the mathematical difficulties of deriving optimal estimators, we propose in Section 3 a general data-driven optimization method to improve spectral enhancement methods, using a two-stream structure. A standard algorithm is run in the background, but is only used for spectral variance estimation. 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 proposed method allows for minimization of error criteria for which no closed-form estimators are known. The results of Section 4 show that a large increase in noise reduction and improved perceptual quality can be achieved. 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 estimator

When the speech and noise discrete Fourier transform coefficients are assumed to be independent across time and frequency and distributed according to a complex Gaussian distribution, the MMSE amplitude estimator, \( \hat{A}_k \), for frequency bin \( k \), is given by Ephraim and Malah (1984):

\[
\hat{A}_k = \frac{\sqrt{\pi v_k}}{2\gamma_k} M(-0.5; 1; -v_k)R_k,
\]

where \( R_k \) is the noisy spectral amplitude, \( M(a; c; x) \) is the confluent hypergeometric function (Abramowitz and Stegun, 1965, 6.1.1), and \( v_k \) is defined by

\[
v_k = \frac{1}{1 + \xi_k}.
\]

The a priori and a posteriori SNRs are defined by \( \xi_k = \lambda_s(k)/\lambda_d(k) \) and \( \gamma_k = R_k^2/\lambda_d(k) \), respectively. \( \lambda_s(k) \) and \( \lambda_d(k) \) are the speech and noise spectral variances for frequency bin \( k \). The MMSE log spectral amplitude estimator is (Ephraim and Malah, 1985):

\[
\hat{A}_k = \frac{\hat{\xi}_k}{1 + \hat{\xi}_k} \exp \left\{ \frac{1}{2} \int_{v_k}^{\infty} e^{-t} \frac{dt}{t} \right\} R_k.
\]

Both (1) and (3) can be written in the form \( \hat{A}_k = G(\hat{\xi}_k, \gamma_k)R_k \), where \( G \) is a spectral gain applied to the noisy amplitude \( R_k \). The speech and noise spectral variances are unknown in practice and have to be estimated. The noise spectral variance can be estimated for stationary noise during speech pauses. For non-stationary noise, approaches based on minimum-statistics (Martin, 2001; Cohen, 2003) can be used. The speech spectral variance is commonly estimated by means of the “decision-directed” approach (Ephraim and Malah, 1984), which will be investigated in Section 2.3.

2.2. Maximum a posteriori amplitude estimation

Lotter and Vary (2005) derived a joint MAP estimator of speech phase and amplitude. They assumed a Gaussian model for the noise DFT coefficients and the following parametric distribution for the speech spectral amplitudes

\[
p(A_k) = \frac{\mu^{\gamma_k+1}}{\Gamma(\gamma_k + 1)} \frac{A_k^{\gamma_k+1}}{2^{(\gamma_k+1)/2}} \exp \left\{ -\frac{\mu A_k}{\sqrt{\gamma_k}} \right\},
\]

where \( \Gamma(\cdot) \) is the Gamma-function (Abramowitz and Stegun, 1965, 6.1.1). The estimator for the phase is just the noisy phase, while the estimator for the amplitude is obtained by multiplying the noisy amplitude with the following gain function:

\[
G_{MAP}(\hat{\xi}_k, \gamma_k) = u_k + \sqrt{u_k^2 + \frac{v}{2\gamma_k}}, \quad u_k = 1/2 - \frac{\mu}{4\sqrt{\hat{\xi}_k \gamma_k}}.
\]

In this paper, we will use the frequency-independent parameters \( v = 0.126 \) and \( \mu = 1.74 \), as in (Lotter and Vary, 2005). These parameters have been obtained in (Lotter and Vary, 2005) by fitting the analytical distribution (4) to experimental amplitude distributions that are conditional on a high-value of the estimated a priori SNR \( \tilde{\xi} \). This means that primarily large speech amplitudes are taken into account.

2.3. Decision-directed a priori SNR estimation

Ephraim and Malah (1984) proposed the following decision-directed estimator for the a priori SNR in time frame \( n \) and frequency bin \( k \):

\[
\hat{\xi}_k(n) = x \frac{\hat{A}_k^2(n-1)}{\hat{\lambda}_d(k, n)} + (1 - x)P[\gamma_k(n) - 1].
\]

\( P[x] \) is the clipping function: it sets negative values to zero. The weighting coefficient \( x \) is usually chosen near one, e.g., \( x = 0.98 \). 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 excessively smoothed. Usually, \( \hat{\xi}_k \) is constrained to be larger than a certain minimum value \( \hat{\xi}_{min} \). This helps in reducing musical noise (Cappé, 1994). We will use \( \hat{\xi}_{min} = -19 \text{dB} \).

2.3.1. Convergence behavior of the decision-directed approach

As indicated by Martin (2005b), 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
\[
\hat{\zeta}_k(n) = \max \left[ \frac{\tilde{A}_k(n-1)}{\hat{\lambda}_d(k,n)} + (1 - \alpha)[\gamma_k(n) - 1], \xi_{\min} \right].
\] (7)

However, there is a bias due to the term \(\tilde{A}_k(n-1)/\hat{\lambda}_d(k,n)\) as well. The expectation of the square of a speech spectral amplitude \(\tilde{A}_k(n)\) equals \(\hat{\lambda}_d(k,n)\) 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 \(\xi_k(n)\), \(\tilde{A}_k(n)\) is nearly equal to \((\pi/4)\xi_k(n)\hat{\lambda}_d(k,n)\), since the hyper-geometric function \(M(-0.5;1;-v_k)\) is then close to one. This means that even if \(\hat{\lambda}_d(k,n)\) were known exactly, i.e., \(\xi_k(n) = \hat{\lambda}_d(k,n)\), we would have \(\tilde{A}_k(n) \approx (\pi/4)\xi_k(n)\hat{\lambda}_d(k,n)\), i.e., a biased estimator of \(\hat{\lambda}_d(k,n)\) at low SNRs, because of the factor \(\pi/4\). Note that the bias is caused by a mismatch between (1) and (7): the square of an estimator of the amplitude is used in (7) instead of an estimator of \(\xi_k(n)\). For high SNRs, \(\hat{\lambda}_d(k,n) \approx R_k(n)\) and there is no significant bias in \(\tilde{A}_k(n)\).

The fact that \(\hat{\lambda}_d(k,n)\) has to be estimated, makes things worse. Because of the factor \(\pi/4\), \(\xi_k(n+1)\) tends to be smaller than \(\hat{\xi}_k(n)\). Therefore, in stationary signals, at low SNR, \(\xi_k\) will converge to \(\xi_{\min}\), and \(\tilde{A}_k\) to \((\pi/4)\xi_{\min}\hat{\lambda}_d\), which generally leads to too much suppression. For \(\alpha < 1\), the term \((1 - \alpha)[\gamma_k(n) - 1]\) counters this to some extent and is therefore necessary,\(^1\) but it has a large variance for values of \(\alpha\) 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 (1994).

In fact, the conclusion is independent of the assumption made for the distribution of speech energy. Because, for any distribution we have:
\[
E_{A,R}[A^2] \geq (E_{A,R}[A])^2, \tag{8}
\]
where \(E_{A,R}[\cdot]\) is the expectation over the conditional amplitude distribution. Eq. (8) expresses the well-known fact that an expectation of a square is always greater than or equal to the square of the expectation, see, e.g., (Ross, 1972, p. 37). An approximation of the right-hand side of this inequality is used in the decision-directed \(a\ priori\ SNR\) estimator, while the left-hand side is better, because \(E_{A}[E_{R}[A^2]] = \hat{\lambda}_d\), where \(E_{A}[\cdot]\) is the expectation over the distribution of the noisy amplitude. At high SNRs, the difference between the terms in (8) is small because then the variance of \(A\) given \(R\) is small, but at low SNRs it can be significant.

\(^1\) For real speech, which is non-stationary, this term helps in reacting more quickly to changes in signal energy as well.

Similar effects also happen with the log-amplitude estimator. For (3) it can be shown that for small values of \(\hat{\xi}_k(n)\), \(\tilde{A}_k(n)\) is approximately equal to \(\exp(-\gamma)\hat{\xi}_k(n)\hat{\lambda}_d(k,n)\), where \(\gamma = 0.57721566\ldots\) is Euler's constant.

We can conclude that even for stationary signals \(\alpha = 1\) is suboptimal, because it generally causes too much suppression. The bias is a function of the true SNR and true speech distribution, which are unknown. It is therefore difficult to correct for it. For a Gaussian model, the bias can be corrected for at low SNR for the amplitude estimator by inserting a factor \(4/\pi\) into (7):
\[
\hat{\xi}_k(n) = \max \left[ \frac{4}{\pi} \frac{\tilde{A}_k(n-1)}{\hat{\lambda}_d(k,n)} + (1 - \alpha)[\gamma_k(n) - 1], \xi_{\min} \right]. \tag{9}
\]

Ephraim and Malah (1984) have investigated the sensitivity of the gain function \(G\) in (1) to errors in \(\zeta\) and concluded that an overestimation of \(\zeta\) is more appropriate than using an underestimate. Therefore, although the correction factor \(4/\pi\) above introduces a bias at high SNRs, its influence is smaller than the bias in (7) at low SNRs.

Another way to obtain a less biased \(a\ priori\ SNR\) estimator is to use the MMSE estimator of the square of the clean amplitude of the previous frame in the first term of (7). This estimator is given by (Wolfe and Godsill, 2001; You et al., 2005)
\[
\hat{A}_k^2(n) = \frac{\hat{\xi}_k(n)}{1 + \hat{\xi}_k(n)} \hat{\lambda}_d(k,n)[1 + v_k(n)]. \tag{10}
\]

However, the square root of this estimator is not an MMSE estimator of the clean amplitude and leads to suboptimal amplitude estimation. It would therefore be better to use \(\hat{A}_k^2(n)\) of (10) in (7) for estimation of the \(a\ priori\ SNR\) in the next time-frame and use a different amplitude \(A_k(n)\) for reconstruction, by means of the gain functions in, e.g., (1) or (3). This leads to a consistent \(a\ priori\ SNR\) estimation that is not influenced by the final amplitude estimation.

Note that the correction factor in (9) and the expression (10) are based on the assumption of a Gaussian model for the complex DFT coefficients. This assumption may not be accurate (see, e.g., (Lotter and Vary, 2005; Martin, 2005a) and the discussion in Section 2.4). If a different statistical model is assumed, the expressions should be modified accordingly.

2.3.2. Illustration of bias compensation for stationary signals

Fig. 1 shows the effect of \(\alpha\) when the algorithm with (1) and (7) or (9) is applied to a stationary stochastic signal. The continuous line shows the clean spectrum, the dashed-dotted line the enhanced spectrum when (7) is used, the dashed line the enhanced spectrum with (9), and the horizontal dotted line indicates the noise level. The overall SNR was 10 dB. It can be seen that there is not much noise suppression for low values of \(\alpha\). For larger values of \(\alpha\),
there is a bias in the enhanced spectrum when (7) is used. This bias increases with increasing $a$ and decreasing $a$ priori SNR. The bias-corrected estimator (9) clearly leads to much less signal distortion, although less of the noise is suppressed in very low SNR regions of the spectrum. Speech processed with (7) sounds heavily distorted for $a \neq 1$.

### 2.3.3. Results on speech signals

Table 1 shows the average segmental SNR improvement (SSNR+), scores from the PESQ measure (Goldstein et al., 2003), and a mean-square error (MSE) for the standard MMSE amplitude estimator (1) of Ephraim and Malah, using (7), (9), and (7) + (10) for a priori SNR estimation.

The mean-square error criterion directly measures what MMSE estimators should minimize and is defined as

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k} \left( A_k(n) - \hat{A}_k(n) \right)^2,$$

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. All results in the tables are averages over all test sentences. SSNR+ is expressed in dBs, while MSE uses a linear scale. For Segmental SNR computation, frame SNR values outside the range $-10 \text{ dB}$ to $+35 \text{ dB}$ are clipped. All 30 clean sentences of the NOIZEUS database

| $\alpha$ | SNR (dB) | EM with (7) | | EM with (9) | | EM with (7) + (10) | |
|----------|----------|-------------| |-------------| |-----------------| |
|          | SSNR+ | PESQ | MSE | SSNR+ | PESQ | MSE | SSNR+ | PESQ | MSE |
| 0.5      | 0.5    | 2.49 | 1.29 | 11.5 | 2.33 | 1.27 | 12.2 | 2.38 | 1.28 | 12.0 |
|          | 5      | 2.45 | 1.63 | 3.80 | 2.28 | 1.61 | 4.02 | 2.33 | 1.61 | 3.94 |
|          | 10     | 2.33 | 2.09 | 1.27 | 2.17 | 2.05 | 1.33 | 2.22 | 2.06 | 1.31 |
|          | 15     | 2.12 | 2.55 | 0.44 | 1.96 | 2.51 | 0.45 | 2.02 | 2.52 | 0.45 |
| 0.8      | 0.8    | 3.11 | 1.34 | 8.29 | 2.69 | 1.28 | 9.95 | 2.76 | 1.29 | 9.60 |
|          | 5      | 2.97 | 1.69 | 2.97 | 2.56 | 1.61 | 3.45 | 2.63 | 1.62 | 3.34 |
|          | 10     | 2.75 | 2.15 | 1.07 | 2.36 | 2.04 | 1.21 | 2.43 | 2.05 | 1.17 |
|          | 15     | 2.43 | 2.57 | 0.39 | 2.06 | 2.46 | 0.43 | 2.14 | 2.48 | 0.42 |
| 0.98     | 0.98   | 4.26 | 1.43 | 8.01 | 3.11 | 1.24 | 9.43 | 3.19 | 1.25 | 9.18 |
|          | 5      | 3.66 | 1.82 | 3.91 | 2.66 | 1.57 | 4.00 | 2.73 | 1.58 | 3.95 |
|          | 10     | 3.01 | 2.20 | 1.76 | 2.17 | 1.97 | 1.63 | 2.24 | 1.98 | 1.63 |
|          | 15     | 2.31 | 2.55 | 0.74 | 1.66 | 2.36 | 0.65 | 1.71 | 2.37 | 0.65 |
(Hu and Loizou, 2006) have been used, to which stationary white noise from the Noisex-92 database (Varga et al., 1992), limited to telephone bandwidth (300–3400 Hz), has been added. A randomly chosen section of the noise was added to every sentence at every SNR condition. The noise variance \( \lambda(k) \) was estimated from 0.64 s of noise only, preceding each speech sentence. The bias-corrected \textit{a priori} SNR estimator (9) gives somewhat lower objective performance than (7), but it performs about the same as (7) + (10). There was more residual noise, but the enhanced speech was also free from musical noise, for \( x = 0.98 \), but not for the smaller values of \( x \). This suggests that smoothness across time, i.e., strong correlation in the time series of the \textit{a priori} SNR estimators and corresponding enhanced amplitudes, rather than a low (ensemble) variance, is sufficient for avoiding the musical noise.

Surprisingly, (7) achieves somewhat better objective scores than (9) and the combination (7) + (10). There are three reasons for this behavior: Firstly, (7) may suppress more noise in very low SNR regions. But this may lead to more suppression of weak speech components as well. In other words, (7) has a different noise reduction versus speech distortion trade-off than (9) and (7) + (10). Secondly, as mentioned earlier, the bias-corrected SNR estimators used here assume a Gaussian model for the complex DFT coefficients. This is a good model for the synthetic signals used for Fig. 1, but it may not be the best model for real speech. Thirdly, the \textit{a priori} SNR estimator and the gain function are linked in a non-linear feedback loop. Therefore, the performance of the one depends on the other in a complicated way. A different \textit{a priori} SNR estimator needs another gain function for optimal performance. The latter two points are discussed in more detail in the next section.

The optimal value of \( x \) depends on the input SNR. For the lowest SNRs, the optimal value of \( x \) is near 0.98, while for the highest SNR it is near 0.8, according to SSNR+, PESQ, and MSE. However, there is much more musical noise for \( x = 0.8 \), so there seems to be a disagreement between these objective measures and subjective quality at these high SNRs.

2.4. Other statistical models

We can conclude so far that the performance of the amplitude estimators (1) and (3) depends on the properties of the particular \textit{a priori} SNR estimator used. Their performance also depends on the accuracy of the statistical model. The estimators (1), (3) and (10) assume a complex Gaussian distribution for the DFT coefficients of speech and noise. It has been shown that the distribution of speech spectral amplitudes at very high SNR, conditional on a small range of (high) values of the estimated \textit{a priori} SNR, deviates from Gaussianity and a Laplacian, Gamma, or more general distribution model for the amplitudes (Lotter and Vary, 2005) or real and imaginary parts (Martin, 2005a) of the Fourier coefficients can lead to improved speech enhancement. Although the authors of these papers give compelling arguments for a deviation from Gaussianity, there is also support for the Gaussian model. Different speech sound classes, such as vowels, plosives, fricatives, etc., have different spectral variances. The spectral variance can be modeled as a random variable itself. Ephraim and Cohen (2006) show that the Gaussian and other models are not necessarily contradictory. The Gaussian assumption for spectral components is conditioned on knowledge of the variance of that component, and they show that the marginal pdf of, e.g., the real part of a DFT coefficient is a continuous mixture of Gaussian densities. Constraining the estimated \textit{a priori} SNR to lie in a small range of values does not change this, because the estimated \textit{a priori} SNR is also a random variable, differing from the true \textit{a priori} SNR. This means that when its value lies in a selected small range of values, the true variance may lie outside that range. Therefore, the distribution of the real and imaginary parts, or amplitudes, of speech DFT coefficients, conditional on the estimated \textit{a priori} SNR, is again some mixture of distributions with different true spectral variances and can deviate from Gaussianity even if the contributing distributions follow the Gaussian model. On the experimental side: Cohen (2006) has shown that for a different estimator of the \textit{a priori} SNR based on GARCH models, a Gaussian model can lead to better speech enhancement 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.) (Jensen et al., 2005). One has to conclude that the true local speech distributions remain unknown and can at best be determined conditional on certain parameters such as estimated \textit{a priori} SNR.

MMSE estimators of the speech amplitudes, log amplitudes or complex speech DFT coefficients have been derived for a few distribution models only (Ephraim and Malah, 1984, 1985; Martin, 2005a). Lotter and Vary (2005) proposed a general parametric distribution function for the speech spectral amplitudes. Because of mathematical difficulties, they had to resort to MAP estimators instead of MMSE estimators. There are many more error criteria which could be relevant for speech enhancement. Loizou (2005) derives estimators for several distortion measures under a Gaussian statistical model and reports good performance for a weighted-Cosh and a weighted-Euclidean distortion measure, for example. It is not possible to derive optimal estimators analytically for many relevant distortion measures under more general distribution assumptions. Furthermore, 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. Analytical gain functions are derived for \textit{known a priori} SNR, but in practice \textit{a priori} SNR has to be estimated. In addition, the \textit{a posteriori} SNR and the estimated \textit{a priori} SNR are correlated, as is evident from (7). This has also not been taken into
account in the analytical derivation of known gain functions. To solve all these difficulties we propose a data-driven optimization method to improve speech enhancement algorithms. We propose to apply two gain functions in parallel. A conventional algorithm is run in the background, and performs the estimation of a priori SNR. For the final reconstruction a different gain function is applied, which corrects for some of the modeling errors and estimation inaccuracies. This 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. This procedure is general in the sense that it can be applied to many speech enhancement algorithms, regardless of their distribution assumptions. The gain function used for reconstruction can also be optimized for many error criteria, such as those in (Loizou, 2005). The procedure is illustrated for the MMSE amplitude and log-amplitude criteria, and the weighted-Euclidean and weighted-Cosh distortion measures in the following sections.

3. Data-driven optimization of gain functions

The Ephraim–Malah suppression rules are functions of the a priori and a posteriori SNRs. This remains true for suppression rules based on 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

\[ \tilde{A}_k(n) = F \left( \frac{\hat{A}_k^2(n - 1)}{\hat{\lambda}_d(k, n)}, \frac{\hat{R}_k^2(n)}{\hat{\lambda}_d(k, n)} \right) R_k(n), \]  

where \( F \) is a complicated non-linear function which, of course, also depends on \( x \) and \( \xi_{\text{min}} \). 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 can be implemented as a correction to the algorithms, using a two-stream procedure, as follows: \( \tilde{A}_k \) or \( \left( \tilde{A}_k^2 \right)^{0.5} \) 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 another amplitude \( \tilde{A}_k \) obtained by multiplying the noisy amplitude \( R_k \) by a separate gain function \( \tilde{G}(\xi_k, \gamma_k) \), differing from the gain function \( G(\xi_k, \gamma_k) \). In other words, some algorithm performing a priori SNR estimation 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 \( \tilde{G} \) corrects for some of the modeling errors and estimation inaccuracies. We invoke the standard assumption that DFT coefficients are independent across time and frequency. Consequently, for each frequency bin \( k \), the corrective gain function \( \tilde{G} \) is a function of the two parameters \( \xi_k \) and \( \gamma_k \) only, and we can write

\[ \tilde{A}_k = \tilde{G}(\xi_k, \gamma_k)R_k. \]  

\( \tilde{G} \) is implemented as a look-up table: the support of \( \xi_k \) and \( \gamma_k \) is discretized onto a grid. The grid points range from \(-19 \text{ dB} \) to \(-40 \text{ dB} \) in steps of \(1 \text{ dB} \). Each parameter cell contains the values of \( \xi \) and \( \gamma \) closest to the grid point and has its corresponding value of \( \tilde{G}(\xi, \gamma) \) stored in a matrix. Assuming that each gain value is represented by 16 bits, the look-up table can be stored in less than 7.2KB of memory. Ephraim and Malah found that there was essentially no difference in performance between an analytical calculation of their gain function and an implementation by means of a look-up table, for a 1 dB step size (Ephraim and Malah, 1984). This step size is also large enough for our gain tables: decreasing cell dimensions to 0.5 dB increased neither objective nor subjective quality significantly.

The idea to use look-up tables to implement estimators has been proposed earlier by Porter and Boll (1984). However, in their work no estimator of a priori SNR was used and the estimators were adapted to global speech distributions only. The method presented here can be looked at as a generalization of the work of Lotter and Vary (2005). Instead of deriving analytical gain functions for speech distributions conditioned on a high estimated a priori SNR, we find optimized gain values for all values of \( \xi \) and \( \gamma \). Furthermore, we can optimize for many error criteria, which is not possible analytically.

3.1. The training procedure

Our aim is to find the function \( \tilde{G}(\xi_k, \gamma_k) \) that minimizes the average distortion for each cell for a wide range of SNR conditions. We consider the range \(-15 \text{ dB} \) to \(+25 \text{ dB} \) of overall SNRs, because this covers the range of practical interest. \( \tilde{G} \) is found by means of a training procedure using the speech from the TIMIT–TRAIN database (Garofolo et al., 1990), converted to telephone bandwidth. To the clean signals, bandpass-filtered white noise is added at the various overall SNRs. Then the a priori SNR estimator consisting of the combination of (7) and (10) is run. It is also possible to run a different algorithm, such as the standard Ephraim–Malah algorithm (Ephraim and Malah, 1984). It will turn out, however, that the results after optimization are almost the same for both algorithms. We prefer to use the combination of (7) and (10) for a priori SNR estimation, because they are more consistent with each other, as was pointed out in Section 2.3.1. Only stationary noise is used for training. The algorithm will be tested for non-stationary noise as well, where the minimum statistics method (Martin, 2001) provides estimates of the noise variance.

In each frame, for each frequency bin, we have a \( (\xi_k, \gamma_k) \) pair that falls into one of the parameter cells. \( (\xi_k, \gamma_k) \) 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 \((i, j, k)\) 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 \(\bar{G}_{ij}\) for parameter cell \((i, j)\) is found by minimizing a distortion measure of interest. We consider the following distortion measures:

1. **Weighted-Euclidean distortion measure** (Loizou, 2005):

   \[
   D_{WE} = \sum_{m=1}^{M_{ij}} \left( A_{ij}(m) - \frac{G_{ij}R_{ij}(m)}{A_{ij}(m)} \right)^2, \tag{14}
   \]

   \[
   \bar{G}_{ij} = \sum_{m=1}^{M_{ij}} A_{ij}^p(m)R_{ij}(m) / \sum_{m=1}^{M_{ij}} A_{ij}^p(m), \tag{15}
   \]

2. **Log-Euclidean distortion measure** (Ephraim and Malah, 1985):

   \[
   D_{LE} = \sum_{m=1}^{M_{ij}} \left( \log[A_{ij}(m)] - \log[G_{ij}R_{ij}(m)] \right)^2, \tag{16}
   \]

   \[
   \bar{G}_{ij} = \prod_{m=1}^{M_{ij}} \frac{A_{ij}(m)}{R_{ij}(m)}. \tag{17}
   \]

3. **Weighted-Cosh distortion measure** (Loizou, 2005):

   \[
   D_{WC} = \sum_{m=1}^{M_{ij}} \left[ \frac{A_{ij}(m)}{G_{ij}R_{ij}(m)} + \frac{G_{ij}R_{ij}(m)}{A_{ij}(m)} - 1 \right] A_{ij}^p(m), \tag{18}
   \]

   \[
   \bar{G}_{ij} = \sqrt{\frac{\sum_{m=1}^{M_{ij}} A_{ij}^{p+1}(m)/R_{ij}(m)}{\sum_{m=1}^{M_{ij}} A_{ij}^{p+1}(m)}}. \tag{19}
   \]

\(R_{ij}(m)\) is the \(m\)th noisy amplitude that fell into parameter cell \((i, j)\) and \(A_{ij}(m)\) the corresponding clean amplitude. We consider \(p = 0\) and \(p = -1\) for the Weighted-Euclidean distortion measure. \(p = 0\) corresponds to an MMSE criterion, while with \(p = -1\), smaller amplitudes are given more weight. For the Weighted-Cosh distortion measure we will use \(p = -0.5\), following (Loizou, 2005).

Some combinations of \(\bar{G}_{ij}\) and \(\gamma_k\) are highly unlikely and may not often enough have occurred during the training. This means that \(M_{ij}\) for that cell is too small to have a reliable \(\bar{G}_{ij}\). In such cases, the corresponding gain function for a Gaussian model is used for reconstruction, for which the expressions are given below:

1. **Weighted-Euclidean distortion measure:**

   \(p = 0\) corresponds to an MMSE criterion and the gain function is Ephraim and Malah’s gain function (Ephraim and Malah, 1984):

   \[
   G_{EM} = \frac{\sqrt{R_{ij}}}{2\gamma_k} M(-0.5; 1; -v_k). \tag{20}
   \]

   For \(p = -1\), the gain function is (Loizou, 2005):

   \[
   G_{WE} = \frac{1}{\sqrt{\pi\gamma_k}} M(0.5; 1; -v_k). \tag{21}
   \]

2. **Log-Euclidean distortion measure** (Ephraim and Malah, 1985):

   \[
   G_{LE} = \frac{1}{1 + \frac{\xi_k}{\gamma_k}} \exp \left\{ \frac{1}{2} \int_{v_k}^{\xi_k} \frac{e^{-r}}{r} \, dr \right\}. \tag{22}
   \]

3. **Weighted-Cosh distortion measure,** with \(p = -0.5\) (Loizou, 2005):

   \[
   G_{WC} = \frac{1}{\gamma_k} \sqrt{v_k M(-0.25; 1; -v_k)} \tag{23}
   \]

For training we used the entire TIMIT–TRAIN database (Garofolo et al., 1990), which consists of about 900,000 frames of speech. The speech signals were limited 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 corresponding gain function for a Gaussian model was substituted from (20)–(23), as a function of \(\bar{G}_{ij}(n)\) and \(\gamma_k(n)\).

### 3.2. Examples of gain curves

Figs. 2–5 show examples of the resulting gains \(\bar{G}\) for different **a priori** SNR estimators, different values of \(z\), and different error criteria as a function of **a posteriori** SNR, for the two values of **a priori** SNR \(\bar{\xi} = -10\) dB and \(\bar{\xi} = +10\) dB. Fig. 2 shows the gain curves optimized for the unweighted-Euclidean criterion \((G_{WE} \text{ opt.})\), along
with the gain curve of Ephraim and Malah $G_{EM}$, and the joint MAP gain curve $G_{MAP}$ (5). The optimized gain gives more suppression than Ephraim and Malah’s gain, but less than the joint MAP gain. This indicates that the joint MAP gain is not MMSE optimal at low a posteriori SNRs and may suppress too much of low-energy speech components.

At high $c$ and low $\alpha$ the gains $G_{EM}$ and $G_{WE opt}$ amplify the noisy amplitude. This is because given that a priori SNR is high, very small amplitudes are improbable, as was already pointed out by Ephraim and Malah (1984).

In this figure, the optimized gains are set equal to the analytical gains for $\xi = 10$ dB for $\gamma$ larger than about 8 dB. The reason is that because $\gamma$ is used in the estimated a priori SNR, it can never become larger than about 8 dB when $\xi = 10$ dB and $\alpha = 0.98$. In other words, because $\xi$ and $\gamma$ are dependent, certain combinations of $\xi$ and $\gamma$ never occur in practice, and the values of any gain function are irrelevant for those combinations. A unique feature of our optimization procedure is that it automatically takes into account the dependency of the parameters $\xi$ and $\gamma$.

Fig. 3 shows analytical ($G_{WE}$) and optimized ($G_{WE opt}$) gain curves for the weighted-Euclidean distortion measure with $p = -1$, along with Ephraim–Malah’s gain $G_{EM}$, which corresponds to $p = 0$. $p = -1$ suppresses more than $p = 0$, and the optimized gain is again lower than the analytical gain.

The fact that the optimized gains differ from the analytical gains means that the assumption of a complex Gaussian distribution of the DFT coefficients (conditional on the estimated a priori SNR) is not valid. As expected, gains optimized for the other a priori SNR estimators (Fig. 4) and other values of $\alpha$ (Fig. 5) look slightly different, because the optimization procedure takes into account the statistics of the a priori SNR estimator, as well as the actual speech distribution. The final enhancement results, though, are very similar (see Section 4 and Tables 2, 4–6). Fig. 4 shows that for $\xi = 10$ dB, the gain optimized for the a priori SNR estimator (7) is higher than the gain optimized for the other two a priori SNR estimators. This makes sense, because we have seen in Section 2.3.1 that (7) is generally biased low at low SNR.

4. Experimental results

In this section we evaluate the optimized gain functions described in the previous section. Both in training and testing we used 50%-overlapping frames of 32 ms (256 samples at 8 kHz sampling frequency). We used a cosine-squared data window, which has the perfect reconstruction property. In order to demonstrate that our optimized gain functions are not specifically tailored to the speech corpus used...
for training, we used a completely different database for testing. More specifically, while TIMIT is used for training, all 30 clean sentences of the NOIZEUS database (Hu and Loizou, 2006) are used for testing. To show the potential of our optimization method, Table 2 shows the average segmental SNR improvement, PESQ (Goldstein et al., 2003) scores and MSE values (11) after optimizing the algorithms of Table 1. That is, for \( a \) priori SNR estimation (7), (9), or the combination of (7) and (10) are used with \( G_{EM} \) (first stream). For the final reconstruction (second stream) we use the optimized gain functions: the gain value corresponding to the values of the estimated \( a \) priori SNR and the \( a \) posteriori SNR are looked up. This gain value is multiplied by the noisy amplitude to give the final estimated clean speech amplitude. The testing conditions are exactly the same as for Table 1: The same speech contaminated by white noise from Noisex was used. Results are shown for three values of \( a \) at four overall SNRs. The error criterion is the unweighted-Euclidean (corresponding to MMSE) criterion (14), with \( p = 0 \).

The optimized algorithms perform much better than the standard algorithms (compare with Table 1). Almost equally large improvements have been obtained for telephone-bandwidth car noise from Noisex (Erkelens et al., 2006), indicating that gain functions trained for white noise may also be applied to other types of noise. Note that all first-stream SNR estimators perform equally well now. This shows that the difference in performance of the different \( a \) priori SNR estimators in Table 1 is mainly a consequence of an inappropriate statistical model and corresponding gain functions. Compared to Table 1, Table 2 shows improvements in SSNR+ in the order of 1.5–3 dB, in the range of 0.2–0.5 points for PESQ, and a large decrease in MSE as well. Also, the optimum performance is at a lower value of \( a \) now. This is important, since it means that speech transitions are less smoothed, which should result in better intelligibility. 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 \( a \), but increased with decreasing SNR. PESQ scores are almost independent of \( a \) as well.

We will now apply our optimization technique to speech contaminated with non-stationary noise and use the other distortion measures listed in Section 3.1 as well. We will compare the results to the joint MAP estimator (5) of Lotter and Vary (2005), which is currently one of the best performing amplitude estimators adapted to the actual speech statistics. The minimum statistics method (Martin, 2001) is used to estimate the noise variance for the non-stationary noise types.

### 4.1. Joint MAP estimator

Table 3 shows the results of this algorithm on the NOIZEUS database for three noise types: stationary white noise (from Noisex), street noise and babble noise. The decision-directed estimator of (7) was used.

### 4.2. Other distortion measures

Tables 4–6 shows the results of the optimized algorithms for the log-Euclidean, weighted-Euclidean \( (p = -1) \) and weighted-Cosh \( (p = -0.5) \) measures for white, street, and babble noise, respectively. The combination of (7) and (10) was used for \( a \) priori SNR estimation.

### 4.3. Discussion and comparisons

The optimized results of Table 2 and 4–6 are much less dependent on \( a \) than those of Table 1 and 3. This is due mainly to the following reason: Conventional gain functions are derived for \( a \) priori SNR \( \xi \). In practice, the bias and variance of the estimated \( a \) priori SNR depend quite strongly on \( a \), leading to an optimum value for \( a \) near 0.98. The optimized gain function is trained to take into...
Table 3
Segmental SNR improvement (SSNR+), PESQ scores, and mean-square error (MSE) on the NOIZEUS sentences for the joint MAP estimator of Lotter and Vary (2005) as a function of $\alpha$ and overall SNR for white, street, and babble noise

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>SNR (dB)</th>
<th>White</th>
<th>Street</th>
<th>Babble</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR+</td>
<td>PESQ</td>
<td>MSE</td>
<td>SNR+</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>3.43</td>
<td>1.41</td>
<td>8.82</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3.35</td>
<td>1.82</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3.19</td>
<td>2.35</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2.89</td>
<td>2.83</td>
<td>0.43</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>4.36</td>
<td>1.53</td>
<td>6.39</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.17</td>
<td>1.99</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3.91</td>
<td>2.54</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.47</td>
<td>3.01</td>
<td>0.40</td>
</tr>
<tr>
<td>0.98</td>
<td>0</td>
<td>6.84</td>
<td>1.60</td>
<td>6.83</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.24</td>
<td>2.05</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5.51</td>
<td>2.46</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.60</td>
<td>2.89</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 4
Segmental SNR improvement (SSNR+), PESQ scores, and mean-square error (MSE) on the NOIZEUS sentences for optimization of the log-Euclidean measure as a function of $\alpha$ and overall SNR for white, street, and babble noise

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>SNR (dB)</th>
<th>White</th>
<th>Street</th>
<th>Babble</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR+</td>
<td>PESQ</td>
<td>MSE</td>
<td>SNR+</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>6.36</td>
<td>1.65</td>
<td>5.92</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.91</td>
<td>2.15</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5.39</td>
<td>2.62</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.65</td>
<td>3.06</td>
<td>0.51</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>6.41</td>
<td>1.68</td>
<td>5.85</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.91</td>
<td>2.16</td>
<td>2.72</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5.38</td>
<td>2.61</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.66</td>
<td>3.06</td>
<td>0.50</td>
</tr>
<tr>
<td>0.98</td>
<td>0</td>
<td>6.36</td>
<td>1.67</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5.79</td>
<td>2.11</td>
<td>2.80</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5.18</td>
<td>2.56</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.44</td>
<td>3.04</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 5
Segmental SNR improvement (SSNR+), PESQ scores, and mean-square error (MSE) on the NOIZEUS sentences for optimization of the weighted-Euclidean measure with $p = -1$ as a function of $\alpha$ and overall SNR for white, street, and babble noise

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>SNR (dB)</th>
<th>White</th>
<th>Street</th>
<th>Babble</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR+</td>
<td>PESQ</td>
<td>MSE</td>
<td>SNR+</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>5.20</td>
<td>1.43</td>
<td>6.75</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.89</td>
<td>1.88</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4.39</td>
<td>2.33</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.75</td>
<td>2.81</td>
<td>0.60</td>
</tr>
<tr>
<td>0.8</td>
<td>0</td>
<td>5.24</td>
<td>1.45</td>
<td>6.62</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.91</td>
<td>1.90</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4.42</td>
<td>2.35</td>
<td>1.38</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.78</td>
<td>2.83</td>
<td>0.58</td>
</tr>
<tr>
<td>0.98</td>
<td>0</td>
<td>5.20</td>
<td>1.45</td>
<td>6.76</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.86</td>
<td>1.90</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4.39</td>
<td>2.35</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.77</td>
<td>2.83</td>
<td>0.57</td>
</tr>
</tbody>
</table>
account the properties of the estimated \( a \) \textit{priori} SNR, \( \hat{\xi} \). One of the parameters of the optimized gain functions is the \( a \) \textit{posteriori} SNR, which is present in \( \hat{\xi} \) because of the term \((1 - \hat{\xi})\gamma_k(n - 1)\). Therefore, the \textit{extra} information provided by the parameter \( \hat{\xi} \) about the clean speech amplitude is in the term with \( \hat{\xi}^2(n - 1)/\lambda_d(k, n) \), which is less dependent on \( a \) than \( \hat{\xi} \) itself.

Considering Tables 4–6, our optimized method again achieves the best results for \( a \) equal to 0.5 or 0.8. The results are almost independent of \( a \) in that range, both objectively and perceptually. Importantly, there was no increase in musical noise compared to \( a = 0.98 \). The largest noise reduction is achieved for the log-Euclidean distortion measure. The quality of the residual noise, however, was clearly better for the weighted-Cosh and weighted-Euclidean distortion measures. There was very little difference in speech quality between these three distortion measures.

The joint MAP estimator gives maximum noise suppression for \( a = 0.98 \). However, \( a = 0.8 \) has highest PESQ scores and lowest MSE, indicating that there is more speech distortion for \( a = 0.98 \). However, the quality of the residual noise was much better for \( a = 0.98 \).

The optimized method performs best for \( a \) between 0.5 and 0.8, while the joint MAP estimator performs best near \( a = 0.98 \). The optimized log-Euclidean measure and the joint MAP estimator perform about equally well for their respective optimal \( a \) value.

The tables show that although the gain functions have been trained under white noise conditions, performance for other noise types is also good. Tests for nearly stationary car noise from Noisex have shown about equally large improvements as for white noise (Erkelens et al., 2006). Of course, if the noise color is known beforehand, optimizing for that noise type will give the maximum performance. However, our aim was to develop a general method, that does not make specific assumptions about noise type or color, because for many applications the noise characteristic can vary. In a general method, one cannot rely on \( a \) \textit{priori} knowledge of the noise. A sensible approach seems to assume that the average noise spectrum is that of telephone-bandwidth white noise. Furthermore, under the standard assumption that DFT coefficients are statistically independent across frequency, the only difference between training with white noise and training with colored noise for the same range of overall SNRs, is in the range of \( a \) \textit{priori} SNRs that each individual frequency bin encounters during training. Therefore, we have used a wide range of overall SNRs, so that the resulting gain tables become robust to application under colored-noise conditions as well.

### 5. Concluding remarks

Optimal gain functions for spectral noise suppression can be derived for various error criteria under various statistical models. These suppression rules are no longer optimal when parameters of the assumed statistical model have to be estimated, even if that statistical model itself is accurate. We have located a large bias in the decision-directed approach of spectral variance estimation, which may cause serious speech distortion when the weight factor approaches one. We have shown for the standard MMSE speech spectral amplitude and log-amplitude estimators, that the performance is much improved by a data-driven optimization approach. The method can be applied to optimize for many different, perceptually relevant, distortion measures. We considered the log-Euclidean, the weighted-Euclidean, and the weighted-Cosh distortion measures. The resulting speech quality was very similar for these measures. The main difference was in the character of the residual noise, which was less tonal and of a more broadband character for the latter two distortion measures than for the log-Euclidean distortion measure.

The optimized gains are only functions of \( a \) \textit{posteriori} SNR and estimated \( a \) \textit{priori} SNR and are not frequency dependent. Frequency dependent gain curves can lead to further improvements. However, the storage complexity

| \( a \) | SNR (dB) | White | | Street | | Babble |
|---|---|---|---|---|---|
| | | SNR | PESQ | MSE | SNR | PESQ | MSE | SNR | PESQ | MSE |
| 0.5 | 0 | 5.12 | 1.46 | 6.10 | 2.50 | 1.25 | 19.8 | 2.54 | 1.31 | 17.3 |
| | 5 | 4.83 | 1.92 | 2.84 | 2.27 | 1.64 | 7.18 | 2.38 | 1.71 | 6.79 |
| | 10 | 4.37 | 2.37 | 1.29 | 1.69 | 2.13 | 3.03 | 1.93 | 2.19 | 2.78 |
| | 15 | 3.48 | 2.81 | 0.75 | 1.24 | 2.61 | 1.20 | 1.54 | 2.78 | 1.15 |
| 0.8 | 0 | 5.16 | 1.47 | 6.04 | 2.49 | 1.25 | 19.8 | 2.53 | 1.31 | 17.2 |
| | 5 | 4.85 | 1.94 | 2.81 | 2.25 | 1.65 | 7.19 | 2.37 | 1.72 | 6.74 |
| | 10 | 4.38 | 2.38 | 1.29 | 1.68 | 2.13 | 3.02 | 1.93 | 2.20 | 2.73 |
| | 15 | 3.52 | 2.83 | 0.69 | 1.26 | 2.61 | 1.15 | 1.56 | 2.78 | 1.08 |
| 0.98 | 0 | 5.13 | 1.47 | 6.34 | 2.37 | 1.24 | 20.8 | 2.46 | 1.30 | 17.7 |
| | 5 | 4.79 | 1.92 | 2.90 | 2.09 | 1.62 | 7.64 | 2.24 | 1.71 | 7.18 |
| | 10 | 4.23 | 2.35 | 1.86 | 1.34 | 2.10 | 3.92 | 1.60 | 2.18 | 3.84 |
| | 15 | 2.85 | 2.75 | 2.64 | 0.53 | 2.56 | 2.83 | 0.70 | 2.74 | 3.15 |
will increase accordingly. When a small number of subbands is used, the storage complexity can remain manageable. In principle, the amount of training data needed will increase. However, experiments have shown that the amount of training data used in this paper is much more than needed for frequency independent gain tables, and will still be sufficient for a small number of subbands.

In our two-stream implementation, two different gain functions are used. The gain function that provides for reconstruction is a function of estimated \( a \) priori SNR and \( a \) posteriori SNR. Other parameterizations are possible. For example, the parameters shown in (12), where \( \hat{A}_k(n - 1)/\sigma_k(n) \) is used instead of \( a \) priori SNR. This parameter contains almost the same information as \( a \) priori SNR, but has the additional advantage that some of the loss of information due to clipping is avoided. Optimizing for the single gain function \( F \) of (12) may lead to further improvements, and it may remove the need for a separate spectral variance estimator, which would be included in the function \( F \) automatically.

Acknowledgements

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.

References

Martin, R., 2005b. Statistical methods for the enhancement of noisy speech, pp. 43–64 in [Benesty et al., 2005].