14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A NUMERICAL APPROACH FOR ESTIMATING OPTIMAL GAIN FUNCTIONS IN SINGLE-CHANNEL DFT BASED SPEECH ENHANCEMENT

Jesper Jensen and Richard Heusdens

Dept. of Mediamatics
Delft University of Technology
Delft, The Netherlands
E-mail: {Jesper Jensen, Richard Heusdens}@ewi.tudelft.nl

ABSTRACT

We treat the problem of finding minimum mean-square error (MMSE) spectral amplitude estimators for discrete Fourier transform (DFT) based single-channel noise suppression algorithms. Existing schemes derive gain functions analytically based on distributional assumptions with respect to the speech (and noise) DFT coefficients and on mathematically tractable distortion measures. In this paper we propose a methodology to estimate the MMSE gain functions directly from speech signals, without assuming that speech DFT coefficients follow a certain parametrized probability density function. Furthermore, the proposed scheme allows for estimation of MMSE gain functions for pdf/distortion measure combinations for which no analytical solutions are known. Simulation experiments where noisy speech is enhanced using the estimated gain functions show promising results. Specifically, the estimated gain functions perform better than standard schemes, as measured by a range of objective speech quality criteria.

1. INTRODUCTION

With the increased use of mobile digital communication systems, e.g. mobile phones, digital hearing instruments, etc, there is a need for such systems to work well in acoustically noisy environments. One way of improving the noise robustness of these systems is to reduce the noise level in the noisy speech signals using a pre-processing step and then apply the enhanced signals as input to the communication chain.

Many single-channel methods for reducing the noise level in noisy speech signals are based on the discrete Fourier transform (DFT), see e.g. [1, 2, 3], and more recently [4, 5, 6]. Since these methods rely on a stationarity assumption, the noisy signal is divided into short-time signal frames which are transformed to the frequency domain using the DFT. Assuming that the resulting DFT coefficients are statistically independent, scalar gain functions are applied to each DFT coefficient separately in order to compute an estimate of the DFT coefficients of the clean (noise-free) speech signal. Finally, the estimated clean DFT coefficients are transformed back to the time domain using the inverse DFT and the resulting enhanced signal frames are overlap-added to produce the enhanced speech signal.

The wide range of existing DFT based enhancement techniques mainly differ in the underlying statistical assumptions concerning the probability distribution functions (pdfs) of the speech and noise DFT coefficients and in the distortion measures optimized for. Traditionally, clean speech DFT coefficients have been assumed Gaussian, e.g. [2, 3], but more recently estimators based on a laplacian distribution [5], a gamma distribution [4], and a generalized supergaussian distribution [6] have been proposed. Noise DFT coefficients are most often assumed Gaussian, but also here estimators have been derived for other distributions, e.g. [4, 5]. Existing schemes minimize different distortion measures including the mean-square error (MSE) between spectral magnitudes, e.g. [2], log-spectral magnitudes, e.g. [3], and complex-valued DFT coefficients, e.g. [4, 5], see also [7].

Obviously, the different distributional assumptions and distortion measures outlined above lead to different gain functions with which the noisy DFT coefficients are modified. In most cases, however, these gain functions are parametrized by two quantities, namely the a priori signal-to-noise ratio (SNR) and the a posteriori SNR. While the a posteriori SNR can be computed directly from the available noisy data[1], the a priori SNR is a ratio of two expected values and must be estimated from the noisy data. The a posteriori and *estimated* a priori SNR are then substituted into the derived gain functions to find the gain value with which a given noisy DFT coefficient is modified. It is important to note, though, that this procedure is *not* optimal. The gain functions are derived under the condition that the a priori SNR is known with certainty. By replacing the a priori SNR with an estimate, the underlying assumptions have been violated, and there may (and generally does) exist another gain function leading to better performance.

In [4, 5] and [6] this problem was recognized. Here it was argued that the distribution of clean speech DFT coefficients *at a given a priori SNR level* can be approximated as following a Laplacian or Gamma distribution [4, 5], or even a generalized super-gaussian distribution [6], and appropriate minimum MSE (MMSE) and MAP estimators, respectively, were derived under these distributional assumptions.

In this paper our goal is to find optimal gain functions while taking into account that the true a priori SNR is unknown, but only an estimate is available. In contrast to most existing schemes, e.g. [1]–[7], our approach is non-parametric, i.e., we do not assume that pdfs related to the target signal follow any particular parameterized class of pdfs. Rather, we estimate the relevant distributional information prior to run-time from training speech material. In this way our approach bears similarities to the data-driven scheme presented in [8] (which, however, assumed the true

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

[1] We assume here that the power spectral density (psd) of the noise is known with certainty in all signal frames.

a priori SNR to be known with certainty). Further, the proposed scheme allows for estimation of MMSE gain functions for speech pdf/distortion measure combinations for which no analytical solutions are known. In this paper we focus for simplicity on schemes where the a priori SNR is estimated using the maximum likelihood (ML) approach described in [2]. We show that our gain functions significantly outperform other schemes based on the (ML) apriori SNR estimate; in fact, they give better performance scores than schemes based on the much more used decision-directed approach for a priori SNR estimation [2].

2. DEMONSTRATION OF PROBLEM

To set the stage we perform the following initial experiment. Let us consider the zero-mean random signal model

$$X(m,k) = S(m,k) + W(m,k), \tag{1}$$

where the complex random variables $X(m,k)$, $S(m,k)$, $W(m,k) \in \mathbb{C}$ represent the k'th DFT coefficient in frame m of the noisy, clean and noise signal, respectively. We assume that the real and imaginary parts of $S(m,k)$ are independent and identically distributed (iid) Gaussian random variables, each distributed according to $\mathcal{N}(0, \sigma_S^2/2)$, where σ_S^2 denotes the variance of the complex spectral component $S(m,k)$. Thus, realizations of $S(m)$ are iid. In a similar way, we assume that the real and imaginary parts of $W(m,k)$ are iid and distributed according to $\mathcal{N}(0, \sigma_W^2/2)$, where σ_W^2 is the variance of $W(m,k)$. We assume that $S(m,k)$ are $W(m,k)$ are independent. We now synthetically generate a complex-valued time-series of the form in Eq. (1), for $m = 1, \ldots, 10^4$ and for some fixed k. $X(m,k)$ represents the DFT coefficients of a noisy speech signal, but clearly the situation is idealized because the generated time series is completely stationary.

In [2] the MMSE short-time spectral amplitude (STSA) estimator was derived under conditions satisfied by the constructed signal above. This estimator is a function of the a posteriori SNR defined as [1]

$$\gamma(m,k) = \frac{|X(m,k)|^2}{E\{|W(m,k)|^2\}} = \frac{|X(m,k)|^2}{\sigma_W^2(m,k)} \tag{2}$$

and the a priori SNR defined as

$$\xi(m,k) = \frac{E\{|S(m,k)|^2\}}{E\{|W(m,k)|^2\}} = \frac{\sigma_S^2(m,k)}{\sigma_W^2(m,k)}, \tag{3}$$

where $E\{\cdot\}$ denotes the statistical expectation operator.

We apply the MMSE-STSA estimator to the noisy time series $X(m,k)$ (fixed k) in order to estimate the magnitude of $S(m,k)$. We assume that the noise variance $\sigma_W^2(m,k) = \sigma_W^2$ is perfectly known, and consider two different situations: *a*) an ideal (but not practically realizable) situation where the variance of $S(m,k)$, $\sigma_S^2(m,k)$, and thus $\xi(m,k)$, is known as well, and *b*) the situation in practice where $\xi(m,k)$ is unknown and must be estimated from the available noisy data $X(m,k)$. In order to emphasize the problem in b), we estimate $\xi(m,k)$ using the maximum likelihood approach described in [2],

$$\hat{\xi}(m,k) = \left(\frac{1}{M} \sum_{l=m-M+1}^{m} |X(l,k)|^2 \right) / \sigma_w^2 - 1, \tag{4}$$

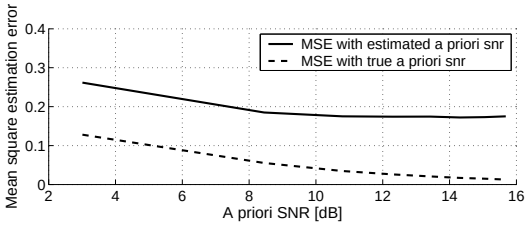

Figure 1: Mean square estimation error for different input SNRs.

using $M = 3$ observations of $X(m,k)$. We then estimate the resulting mean-square estimation error for situations *a*) and *b*). We fix $\sigma_S^2(m,k) = 1$ and repeat this procedure for several noise levels σ_W^2 (i.e. different a priori SNRs) leading to the performance curves shown in Fig. 1. We see that even in this idealized setup where signals are stationary and distributed according to the underlying assumptions, the performance promised by theory (dashed line) is not achieved in practice where a priori SNR is estimated from the available noisy data[2]. The reason is that the theoretically derived MMSE-STSA estimator assumes perfect knowledge of the a priori SNR, an assumption which is not fulfilled in practice.

As mentioned, the problem was addressed in [4, 5] and [6], where gain functions were derived based on speech pdfs measured at a given *estimated* a priori SNR level. While [4, 5, 6] assume that the underlying speech pdfs are members of a parameterized pdf class, we present in the following a data-driven approach which avoids such restrictions, but still takes into account that the apriori SNR is estimated.

3. FINDING OPTIMAL GAIN FUNCTIONS

Consider the random signal model of noisy DFT coefficients

$$X = S + W,$$

where we now have dropped both the frame and frequency bin index because each noisy DFT coefficient is processed independently. Let $X = Re^{j\vartheta}$ and $S = Ae^{j\alpha}$ denote polar representations of the noisy and clean DFT coefficients, respectively. In this paper we focus on minimizing distortion measures which are functions of the spectral magnitude A, i.e., the problem of interest can be stated as

$$\min_{\hat{A}} E\{d(A,\hat{A})\}, \tag{5}$$

where $\hat{A}$ denotes the estimated spectral magnitude, and $d(A,\hat{A})$ is some pre-specified distortion measure. Clearly, for the function $d(A,\hat{A}) = (A-\hat{A})^2$, we have the MMSE-STSA problem addressed in [2], while for $d(A,\hat{A}) = (\log A - \log\hat{A})^2$ the problem was considered in [3] (both assuming Gaussian speech DFT coefficients). Other, more perceptually relevant, choices of $d(A,\hat{A})$ such as the ones proposed in [7] are also possible with our approach.

We assume that the estimator $\hat{A}$ can be written as

$$\hat{A} = g(\gamma, \hat{\xi}) \cdot R,$$

[2] In many state-of-the-art schemes, the the decision-directed approach [2] is prefered over the ML approach for a priori SNR estimation. We choose to use the ML approach in this example for illustration purposes. However, the problem addressed also exists with the decision-directed estimator.

i.e., the gain function $g(\cdot)$ is parameterized by the a posteriori SNR, $\gamma = R^2/\sigma_w^2$, and some *estimate* $\hat{\xi}$ of the a priori SNR $\xi = E\{A^2\}/\sigma_w^2$. This assumption is not very restrictive; many known gain functions can be parameterized like this, see e.g. [2, 3, 6].

3.1 Discretized gain functions

The gain function $g(\gamma,\hat{\xi})$ is a function of two continuous variables, and our goal is to find the function that solves the minimization problem in Eq. (5). To do so we discretize the support C of $(\gamma,\hat{\xi})$ and introduce a piecewise constant approximation of $g(\gamma,\hat{\xi})$. More specifically, we divide C into disjoint cells C_i, $C = \cup_i C_i$, chosen such that the (unknown) gain function can be assumed constant over each cell,

$$g(\gamma,\hat{\xi}) \approx g_i \text{ for } (\gamma,\hat{\xi}) \in C_i. \tag{6}$$

Clearly, the approximation in Eq. (6) can be made arbitrarily accurate by defining suitable small cells C_i and assuming that $g(\cdot)$ is a well-behaved smooth function. With this approximation, it can be shown that the total expected estimation error $E\{d(A,\hat{A})\}$ can be written as a sum of separate distortion terms $E\{d_i\}$ related to each cell C_i,

$$E\{d(A,\hat{A})\} = \sum_i E\{d_i(g_i)\}.$$

In order to minimize this expression, we can minimize each $E\{d_i(gi)\}$ separately with respect to the gain values g_i.

We estimate the gain values using a data-driven approach, where we collect a large number of realizations (a,r) of the random variable pair (A,R) using synthetically mixed noisy speech signals (see Sec. 3.2 for further details). For each such observed (a,r) we compute a $(\gamma,\hat{\xi})$ pair which in turn falls within a cell C_i. Define the set $\mathscr{O}_i$ containing the observed (a,r) pairs associated in this way with cell C_i. To find g_i we must minimize the distortion $E\{d_i\}$ related to cell C_i:

$$E\{d_i(g_i)\} \approx \frac{1}{|\mathscr{O}_i|} \sum_{(a,r)\in\, \mathscr{O}_i} d(a,g_i r),$$

where $|\mathscr{O}_i|$ is the number of (a,r) pairs in cell C_i.

The optimal gain value, say g_i^*, for cell C_i is found by solving $\partial E\{d_i\}/\partial g_i = 0$ for g_i. For example, for the MMSE-STSA problem, $d(A,\hat{A}) = (A-\hat{A})^2$, we find

$$g_i^* = \frac{\sum_{(a,r)\in\, \mathscr{O}_i} ar}{\sum_{(a,r)\in\, \mathscr{O}_i} r^2}, \tag{7}$$

and for the log-spectral distortion measure, $d(A,\hat{A}) = (\log A - \log\hat{A})^2$ we have

$$g_i^* = \exp\left(\frac{1}{|\mathscr{O}_i|} \sum_{(a,r)\in\, \mathscr{O}_i} \log(a/r)\right). \tag{8}$$

In a similar way, it is easy to derive expressions for g_i^* for more complicated distortion measures e.g. the ones considered in [7].

3.2 Estimation of gain functions

We collect realizations (a,r) of the random variable (A,R) from a large quantity of clean and corresponding, synthetically mixed, noisy speech signals. The clean speech material is taken from the Timit data base [9] and consists of approximately 400 speech signals (roughly 25 minutes of speech) from 14 female and 24 male speakers. The speech signals are appropriately lowpass filtered and downsampled to obtain a sample rate of 8 kHz. The initial and trailing silence regions are discarded. Noisy signals are generated by adding white Gaussian noise to the clean signals, scaled to obtain global SNRs ranging from -14 to 25 dB in steps of 3 dB. The result of this procedure is a total of approximately 5500 noisy signals (for which the underlying clean signal is known).

In order to generate (a,r) pairs, we process the signals in a standard front-end for a DFT based enhancement algorithm as follows. The noisy and clean signals are divided into frames of 256 samples with an overlap of 50%. The frames are weighted by a Hann window, the DFT is applied, and noisy and corresponding clean DFT coefficient magnitudes are combined to form observations (a,r). For each noisy DFT coefficient, we estimate the a priori SNR $\xi(m,k)$ using the maximum likelihood approach, Eq. (4) $(M=3)$, and compute the a posteriori SNR $\gamma(m,k)$ using Eq. (2). We define the cells C_i by quantizing $\hat{\xi}(m,k)$ and $\gamma(m,k)$ uniformly in the logarithmic domain in steps of 0.5 dB in the range $[-40;40]$ dB. With these definitions of C_i, each and every observed pair (a,r) is associated with a cell C_i (i.e., the sets $\mathscr{O}_i$ are constructed), and the optimal gain value g_i^* corresponding to that cell may be calculated using e.g. Eq. (7) or (8).

In Fig. 2a we compare the gain functions estimated for $d(A,\hat{A}) = |A-\hat{A}|^2$ using the proposed procedure with the gain functions derived in [2] assuming Gaussian speech DFT coefficients; the authors are not aware of analytically derived MMSE gain functions for this distortion measure for other distributions. We see that the proposed gain functions generally deliver significantly more suppression than the ones derived in [2]. For high a priori and a posteriori SNRs, the gain functions approach unity, as expected. The noisy appearance of the estimated gain function for $\hat{\xi} = 15$ dB is due to the fact that the expression for g_i^* in Eq. (7) is easily dominated by outliers[3]. We believe that increasing the amount of speech data on which the estimation is based will result in smoother curves. The estimated gain function for $\hat{\xi} = 0$ dB is set to zero for high a posteriori SNRs because these combinations of low a priori SNR and high a posteriori SNR was never observed in the offline estimation procedure.

Fig. 2b compares our estimated gain functions with the ones derived in [3] for the log-spectral distortion measure $d(A,\hat{A}) = |\log A - \log\hat{A}|^2$. Again, it appears that no analytically derived MMSE gain functions exist for this distortion measure for e.g. super-gaussian speech distributions. As before, the estimated gain functions give more suppression than the analytically derived functions. Also, using the log-spectral distortion criterion leads to higher suppression both for the gain functions derived in [2, 3] and for the ones estimated here (compare Figs. 2a and 2b).

It may appear that estimating the gain functions based on a training set consisting of speech signals degraded by

[3] Nevertheless, this gain function was used in all simulation experiments reported in the following.

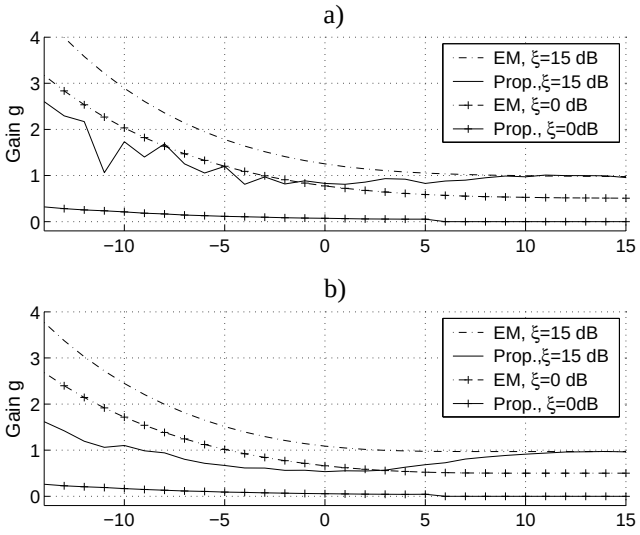

Figure 2: Proposed ('Prop') and analytically derived ([2, 3]) ('EM') gain functions. a) the MMSE-STSA case, $d(A,\hat{A}) = |A-\hat{A}|^2$, b) the log-spectral case, $d(A,\hat{A}) = |A-\hat{A}|^2$.

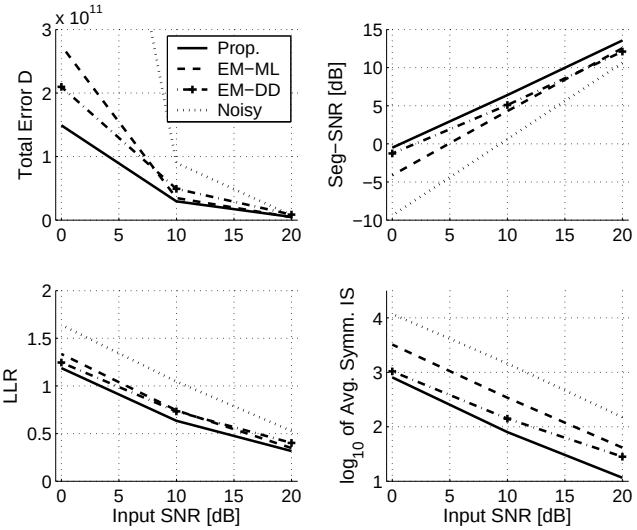

Figure 3: Average objective performance scores as function of input SNR for the criterion $d = |A-\hat{A}|^2$ and signals degraded by additive white Gaussian noise.

white Gaussian noise leads to gain functions tailored for the white noise condition. We can argue, however, that this is not the case, if we make the standard assumption that DFT coefficients are statistically independent (across time and frequency), e.g. [2, 3]. In this case, the noisy DFT coefficients in the training set are simply realizations of independent random variables, each of which is a sum of a speech DFT coefficient (drawn from some underlying pdf) and a noise realization drawn from a zero-mean (complex) Gaussian distribution. We note that, under the given assumption, this would also be the case had the training set been generated using *coloured* Gaussian noise. Thus, by observing the training set we cannot determine whether it was produced using coloured or white noise, and, consequently, the optimal gain functions are the same in the two cases. While the gain functions are independent of the noise colour, they will, however, be tailored for *Gaussian* noise processes.

4. SIMULATION RESULTS

We study the performance of the proposed method in simulation experiments based on approximately 100 speech signals taken from the Timit data base [9]; the speakers and signals are different from the ones used for estimating the gain functions. We construct noisy speech signals synthetically by adding (quasi-)stationary noise sources to the clean signals. The noise source are scaled in order to obtain a prescribed input SNR. The input SNR as well as the objective speech quality criteria introduced below are measured on signals where the initial and trailing silence regions have been discarded.

The noisy speech signals are enhanced using the DFT based noise suppressor described in the previous section. We assume that an ideal voice activity detector is available and estimate the noise psd (which is assumed time-invariant) from a noise-only region of roughly 350 ms preceding speech activity. To evaluate the quality of the enhanced signals, we use a number of objective speech quality measures. First, we estimate the criterion which the particular noise suppression scheme aims at minimizing; for example, for the MMSE-STSA distortion measure $E\{|A-\hat{A}|^2\}$, we compute

$$D = \sum_{l,m} |a(l,m) - \hat{a}(l,m)|^2,$$

where $a(\cdot)$ and $\hat{a}(\cdot)$ denote clean and estimated spectral amplitudes, respectively, and the summation is performed across all frequency and frame indeces. We also evaluate the quality in terms of segmental SNR (Seg-SNR), the Itakura distance measure (log-likelihood ratio, LLR) [10] and the symmetrized Itakura-Saito measure [11].

Fig. 3 considers the MMSE-STSA case, i.e. $d(A,\hat{A}) = |A-\hat{A}|^2$, for signals degraded by additive white Gaussian noise. The figure shows objective quality scores averaged across the test signals, as a function of input SNR for the proposed gain function ('Prop.'), and for the estimator derived in [2] using for a priori SNR estimation the maximum-likelihood approach with $M = 3$ ('EM-ML') and the decision-directed approach with smoothing constant $\alpha = 0.98$ ('EM-DD'), respectively. For EM-ML and EM-DD, $\hat{\xi}$ was limited to values larger than -15 dB. We see that the proposed gain function is superior to the EM-ML estimator for all input SNRs; this is expected since the EM-ML estimator relies on a Gaussian speech assumption, while the proposed gain function is based on real speech DFT coefficients. Remarkably, the proposed gain function performs better than the classical EM-DD estimator, eventhough it uses a ML ($M = 3$) based a priori SNR estimate.

Fig. 4 considers $d(A,\hat{A}) = |\log A - \log\hat{A}|^2$, i.e., the log-spectral amplitude case, for signals degraded by additive white Gaussian noise. We compare here the proposed gain function with the analytically derived estimators in [3], using as before two different methods for a priori SNR estimation. We see that generally, the proposed gain functions lead to superior performance.

Finally, in order to demonstrate that the gain functions

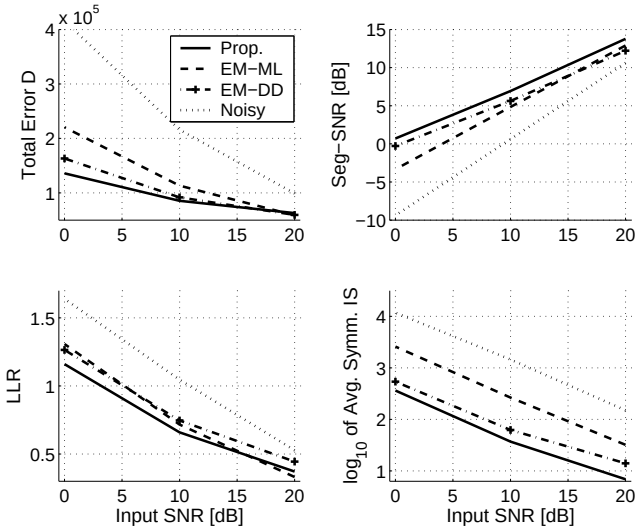

Figure 4: Average objective performance scores as function of input SNR for the criterion $d(A,\hat{A}) = |\log A - \log\hat{A}|^2$ and signals degraded by additive white Gaussian noise.

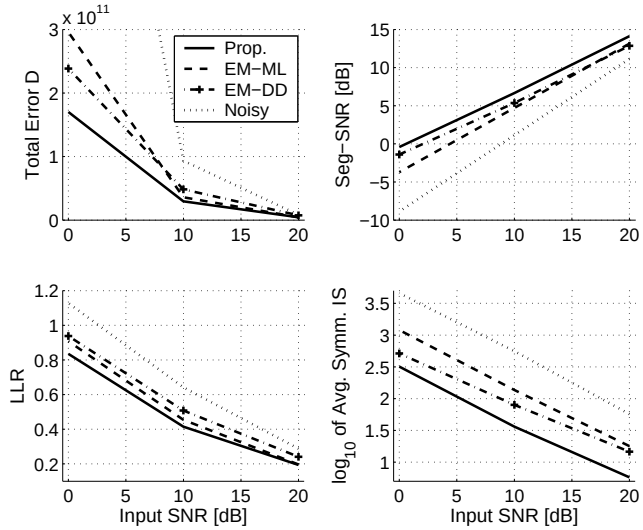

Figure 5: Average objective performance scores as function of input SNR for the criterion $d(A,\hat{A}) = |A-\hat{A}|^2$ and signals degraded by additive F16 cockpit noise.

estimated based on signals degraded by white noise also perform well when applied in coloured noise environments, we use the estimated gain functions to enhance the test speech signals degraded by additive f16 cockpit noise taken from [12]. Fig. 5 shows enhancement performance for $d(A,\hat{A}) = |A-\hat{A}|^2$. As before, the proposed gain functions perform better than the standard estimators.

5. CONCLUSION

We have presented a scheme for estimating gain functions which minimize MSE based distortion criteria for single-channel DFT based speech enhancement. Unlike most existing schemes, our method does not assume that speech DFT coefficients follow a certain parameterized class of pdfs. Furthermore, the proposed scheme allows estimation of gain functions for speech pdf/distortion measure combinations, for which no analytically derived estimators exist. Our gain functions perform better than existing standard schemes, as measured by a range of objective speech quality criteria.

REFERENCES

[1] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, April 1980.

[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] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. ASSP-2, pp. 443–445, April 1985.

[4] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Florida, USA, 2002, pp. 253–256.

[5] R. Martin and C. Breithaupt, "Speech enhancement in the dft domain using laplacian speech priors," in *Int. Workshop, Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 87–90.

[6] T. Lotter and P. Vary, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-gaussian speech modelling," in *Proc. XII European Signal Processing Conference*, Vienna, Austria, September 2004, pp. 1457–1460.

[7] 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, September 2005.

[8] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1984, pp. 18A.2.1–18A.2.4.

[9] DARPA, "Timit, Acoustic-Phonetic Continuous Speech Corpus," NIST Speech Disc 1-1.1, October 1990.

[10] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[11] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

[12] 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 Commun.*, vol. 12, no. 3, pp. 247–253, 1993.