OVERCOMING THE STATISTICAL INDEPENDENCE ASSUMPTION W.R.T. FREQUENCY IN SPEECH ENHANCEMENT

Tim Fingscheidt, Christophe Beaugeant, Suhadi Suhadi
Siemens AG, COM Mobile Phones, Grillparzerstr. 10-18, D – 81675 Munich, Germany
{first name}.{last name}@siemens.com

ABSTRACT

In this paper we give a solution on how to overcome the assumption of statistical independence of adjacent frequency bins in noise reduction techniques. We show that under relaxed assumptions the problem results in an a-priori SNR estimation problem, where all available noisy speech spectral amplitudes (observations) are exploited. Any state-of-the-art noise power spectral density (psd) estimation and weighting rule can be used – they do not need to be restated. In order to solve for an estimator well suited for real-time applications, we model the a-priori SNR values as Markov processes w.r.t. frequency. On the basis of the formulation by Ephraim and Malah, this leads to a new a-priori SNR estimator that yields fewer musical tones.

1. INTRODUCTION

A large number of criteria for speech spectral amplitude estimators has been published. They lead to so-called weighting rules that are applied to the noisy speech spectral amplitude in order to get the estimated speech spectral amplitude. Among these are power estimation (equivalent to power spectral subtraction) [1], Wiener filtering [2], maximum likelihood estimation (ML) [3], MMSE estimation of the spectral amplitude [4], MMSE estimation of the log-spectral amplitude [5] (the two latter approaches by Ephraim and Malah), and minimum least square amplitude estimation [6].

All of the approaches listed above assume the statistical independence of adjacent frequency bins. This is of course far from reality, since two effects contribute to the statistical dependence over frequency: First the short-time Fourier transform and its window function. Windowing in the time domain corresponds in the frequency domain to convolution with the Fourier transform of the window impulse response. Although the spectral maximum of the window function usually is narrow, direct neighbors in the frequency domain show a dependence even if the time domain signal is white noise. In practice however, background noises are far from being spectrally white, which additionally contributes to the statistical dependence of adjacent frequency bins. Secondly, the speech signal itself inherits some spectral envelope as it is generated by the vocal tract.

There are several proposals for post-processing the enhanced speech spectrum that work inter- and intra-frame oriented [7–9]. They mainly try to eliminate musical tones. These approaches however are only available as non-causal algorithms requiring additional delay in a real implementation. Additionally, they are quite heuristic, and sometimes focus more on intelligibility rather than the total quality of the speech signal.

Following to some extent the terminology used by Cohen [10], in section 2 we will introduce an estimator (in analogy to a weighting rule) that uses *all* observations available: The neighboring, the past, and possibly even future observations. This approach will then be further developed *without* the assumption of statistical independence of adjacent frequency bins. This means that inter- as well as intra-frame correlations are exploited *during* the estimation process. In analogy to [10], subsection 2.1 will show that our general problem formulation still allows to use all known weighting rules, which is an important result. It directly leads to the formulation of a speech spectral variance estimator in subsection 2.2 and an a-priori SNR estimator in subsection 2.3. Section 3 gives a practical solution for a causal, decision-directed a-priori SNR estimator (like Ephraim and Malah's) exploiting both inter- *and* intra-frame correlations. Finally, in section 4 we discuss the performance of the algorithm found.

2. AN ESTIMATOR BASED ON ALL OBSERVATIONS AVAILABLE

2.1. Spectral Amplitude Estimation

We assume the clean speech signal to be distorted by additive noise, hence the noisy observation at time instant (or frame number) l can be expressed in the frequency domain after short-time Fourier transform (STFT) of length K as

$$Y_l(k) = X_l(k) + D_l(k), \tag{1}$$

with k being the frequency index regarded in the following only from 0 to $K/2$, $X_l(k)$ being the clean speech, and $D_l(k)$ being the noise in frame l. Let's denote then

$$\mathcal{Y}_0^{l'} = \{\mathcal{Y}_0^{l'}(0), \mathcal{Y}_0^{l'}(1), ..., \mathcal{Y}_0^{l'}(K/2)\}$$
$$\text{with} \quad \mathcal{Y}_0^{l'}(k) = \{Y_0(k), Y_1(k), ..., Y_{l'}(k)\}$$

as knowledge about all observations form time instant 0 until l'.

We don't want to restrict ourselves to any specific distortion measure, therefore we assume an arbitrary distortion $D[X_l(k), \hat{X}]$ to be minimized. An estimator for the spectral amplitude $X_l(k)$ that takes into account all observations made until time instant l' is then

$$\hat{X}_l(k) = \arg\min_{\hat{X}} E\{D[X_l(k), \hat{X}] | \mathcal{Y}_0^{l'}\}. \tag{2}$$

Note that if $l' > l$ we derive a non-causal estimator, meaning that the estimate $\hat{X}_l(k)$ is computed after the observations $Y_{l'}(\kappa)$, $\kappa = 0, ..., K/2$, get available. This approach may be very useful for off-line speech enhancement as well as for robust speech recognition, where the delay constraints often times are not as tight as in conversational telephony applications. The usually addressed case $l' = l$ however yields the speech estimates right after availability

of the noisy observations at time instant l, which leads then to a causal estimator.

Instead of (2), e.g., Ephraim and Malah [4, 5] propose weighting rules based on

$$\hat{X}_l(k) = \arg\min_{\hat{X}} E\{D[X_l(k), \hat{X}]|Y_l(k)\}, \tag{3}$$

where only the observation in frequency bin k of frame l is exploited for the derivation of the weighting rule. Accordingly, they assume statistical independence between any two frequency bins $X_l(k_1)$ and $X_l(k_2)$, as well as between any two time instants $X_{l_1}(k)$ and $X_{l_2}(k)$. Cohen's novelty [10] was that he omitted the latter assumption, which leads then to an estimator

$$\hat{X}_l(k) = \arg\min_{\hat{X}} E\{D[X_l(k), \hat{X}]|\mathcal{Y}_0^{l'}(k)\}. \tag{4}$$

We do not require both afore mentioned assumptions: We will use instead a – relaxed – assumption that is implicit to any approach based on a priori SNR estimation. Assumption (A): Given the speech spectral variance $\lambda_{X_l}(k)$, then $X_l(k)$ is statistically independent of any $X_{\tilde{l}}(\tilde{k})$ for $\tilde{l} \neq l$ and $\tilde{k} \neq k$.

We want to solve now (2) in analogy to [10]. Since there is no straightforward solution, we introduce the speech spectral variance $\lambda_{X_l}(k)$ as a help variable that is known. By applying assumption (A) equation (2) becomes

$$\begin{aligned}\hat{X}_l(k) &= \arg\min_{\hat{X}} E\{D[X_l(k), \hat{X}]|\mathcal{Y}_0^{l'}, \lambda_{X_l}(k)\} \\ &= \arg\min_{\hat{X}} \int D[X_l(k), \hat{X}] p(X_l(k)|\mathcal{Y}_0^{l'}, \lambda_{X_l}(k)) d(X_l(k)) \\ &= \arg\min_{\hat{X}} \int D[X_l(k), \hat{X}] p(X_l(k)|Y_l(k), \lambda_{X_l}(k)) d(X_l(k)).\end{aligned} \tag{5}$$

If we compare this to the conventional approach in (3), we find that they are essentially the same, if the conventional approach also assumes knowledge of a speech spectral variance $\lambda_{X_l}(k)$. Thus we can conclude, that due to the relaxed assumption (A) the exploitation of all observations does not require a restatement of a weighting rule. Instead, all weighting rules having (an estimate of) the speech spectral variance as an intermediate step can be used (e.g. the approaches by Ephraim and Malah [4, 5], Cohen [10], but it can be also the Wiener filter [11]).

In reality of course we don't know $\lambda_{X_l}(k)$, but we can estimate it given *all* available observations. This can be expressed as

$$\hat{\lambda}_{X_l|l'}(k) = E\{|X_l(k)|^2|\mathcal{Y}_0^{l'}\}. \tag{6}$$

The estimation of the speech spectral variance according to (6) will make the difference to state-of-the-art speech enhancement. Thus we can fully focus on an improved speech spectral variance estimate $\hat{\lambda}_{X_l|l'}(k)$, given all observations $\mathcal{Y}_0^{l'}$.

2.2. Speech Spectral Variance Estimation

The speech spectral estimate after (6) is

$$\begin{aligned}\hat{\lambda}_{X_l|l'}(k) &= E\{|X_l(k)|^2|\mathcal{Y}_0^{l'}\} \\ &= \int |X_l(k)|^2 p(X_l(k)|\mathcal{Y}_0^{l'}) d(X_l(k)).\end{aligned} \tag{7}$$

In order to suitably exploit the knowledge about all observations $\mathcal{Y}_0^{l'}$ we follow a similar approach as before. We assume however *all* speech spectral variances in frame l to be known and apply assumption (A) so that the pdf in (7) becomes

$$\begin{aligned}&p(X_l(k)|\mathcal{Y}_0^{l'}, \lambda_{X_l}(0), ..., \lambda_{X_l}(K/2)) \\ &\quad = p(X_l(k)|Y_l(k), \lambda_{X_l}(0), ..., \lambda_{X_l}(K/2)).\end{aligned} \tag{8}$$

Again, in a practical system we do not know the speech spectral variances in frame l, but we assume having *preliminary* estimates $\hat{\lambda}'_{X_l|l'}(\kappa)$, $\kappa = 0, ..., K/2$, according to any state-of-the art approach, available. While each single $\hat{\lambda}'_{X_l|l'}(\kappa)$ represents the observation sequence $\mathcal{Y}_0^{l'}(\kappa)$ (over time) without making use of the statistical dependence over frequency, the whole set of preliminary speech spectral variance estimates $\hat{\lambda}'_{X_l|l'}(\kappa)$, $\kappa = 0, ..., K/2$ allows us to exploit statistical dependence also over frequency.

The preliminary speech spectral variance estimate is the basis for our final speech spectral variance estimator formulated as

$$\hat{\lambda}_{X_l|l'}(k) = E\{|X_l(k)|^2|\hat{\lambda}'_{X_l|l'}(0), ..., \hat{\lambda}'_{X_l|l'}(K/2)\}. \tag{9}$$

Note that we have omitted the observation $Y_l(k)$. This can be done since $\hat{\lambda}'_{X_l|l'}(k)$ sufficiently represents the observation at frequency bin k for the purpose of a speech spectral variance estimation.

2.3. Noise psd Estimation and SNR Estimation

As in many approaches to speech enhancement, we assume some noise psd estimation to be performed in a first step. In particular, we assume a noise spectral variance $\lambda_{D_l|l'}(k) = E\{|D_l(k)|^2|\mathcal{Y}_0^{l'}\}$ to be estimated from the observations[1]. In the rest of the paper we focus on the causal case ($l' = l$). An example for a simple causal noise psd estimation during speech pause is given by

$$\lambda_{D_l|l}(k) = \alpha\lambda_{D_{l-1}|l-1}(k) + (1-\alpha)|Y_l(k)|^2 \tag{10}$$

with a smoothing factor $\alpha = 0.9$. During speech activity the noise spectral variance is kept constant $\lambda_{D_l|l}(k) = \lambda_{D_{l-1}|l-1}(k)$.

Taking the noise spectral variances $\lambda_{D_l|l}(\kappa)$ as deterministic and known variables, and looking for a causal estimator solution to (9), we can restate (9) as a so-called *a-priori SNR* estimator

$$\hat{\xi}_{l|l}(k) = E\{|X_l(k)|^2|\hat{\xi}'_{l|l}(0), ..., \hat{\xi}'_{l|l}(K/2)\}/\lambda_{D_l|l}(k) \tag{11}$$

with the a-priori SNR [4] defined as

$$\xi_{l|l}(k) := \frac{\lambda_{X_l|l}(k)}{\lambda_{D_l|l}(k)}. \tag{12}$$

Furthermore, an *a-posteriori SNR* can be defined as

$$\gamma_l(k) := \frac{|Y_l(k)|^2}{\lambda_{D_l|l}(k)}. \tag{13}$$

[1] It is important to note that in our formalism also the noise estimator may exploit all observations $\mathcal{Y}_0^{l'}$. While statistical dependence over frequency is often times exploited already in state-of-the-art noise psd estimation techniques (e.g. those operating on a Bark scale), the development of non-causal techniques ($l' > l$) could give better results especially in low-energy word endings, since a few frames look-ahead into the next speech pause helps finding an interpolated instead of an extrapolated noise estimate. We will leave this however as an outlook for further work to be done.

Each of the preliminary a-priori SNR estimates $\hat{\xi}'_{l|l}(\kappa)$ with $\kappa = 0, ..., K/2$, already incorporates knowledge about all the past observations in frequency bin κ. The estimator we are looking for shall of course not be the trivial (state-of-the-art) solution $\hat{\xi}_{l|l}(k) = \hat{\xi}'_{l|l}(\kappa)$, but a solution that *also* takes advantage of the statistical dependence of adjacent a-priori SNR values. What remains to be formulated now is a practical solution to (11).

Once the noise spectral variance (and thereby the a-posteriori SNR) and the a-priori SNR are estimated properly by exploiting all available observations, most of the known weighting rules can be applied:

$$\hat{X}_l(k) = G(\xi_{l|l}(k), \gamma_l(k)) \cdot Y_l(k). \tag{14}$$

In summary, our proposed system is composed of the four steps noise psd estimation ($\lambda_{D_l|l}(k)$), preliminary estimation of the a-priori SNR $\hat{\xi}'_{l|l}(k)$, estimation of the a-priori SNR $\hat{\xi}_{l|l}(k)$ taking into account also neighboring preliminary a-priori SNR values, and finally application of a weighting rule $G(\hat{\xi}_{l|l}(k), ...)$.

3. NEW APPROACH TO A-PRIORI SNR ESTIMATION

This section presents a practical solution to (11). The preliminary a-priori SNR estimates can be found using the well-known decision-directed estimator by Ephraim and Malah [4] with a subsequent limiting function

$$\begin{aligned}\hat{\xi}''_{l|l}(k) &= w\frac{\hat{X}^2_{l-1}(k)}{\lambda_{D_{l-1|l-1}}(k)} + (1-w)\max\{\gamma_l(k) - 1, 0\} \\ \hat{\xi}'_{l|l}(k) &= \max\{\hat{\xi}''_{l|l}(k), \xi_{\min}\}.\end{aligned} \tag{15}$$

The factor $w = 0.98$ controls the trade-off between residual speech distortions and musical noise. The a-priori SNR threshold is chosen to be -25 dB for an agreeable level of residual noise.

Assuming the preliminary a-priori SNR estimates having Markov property over frequency, (11) becomes a function[2]

$$\hat{\xi}_{l|l}(k) = F(\hat{\xi}'_{l|l}(k-1), \hat{\xi}'_{l|l}(k), \hat{\xi}'_{l|l}(k+1)). \tag{16}$$

We introduce a correlation parameter $\beta_l(k)$, that reflects the amount of intra-frame correlation of the a-priori SNR estimates. Assuming equal statistical dependence to the right and to the left neighbor, the final estimator $F(.)$ reads then

$$\begin{aligned}\hat{\xi}_{l|l}(k) = \quad & \beta_l(k) \cdot \hat{\xi}'_{l|l}(k) \\ & + \frac{1-\beta_l(k)}{2} \cdot \left[\hat{\xi}'_{l|l}(k-1) + \hat{\xi}'_{l|l}(k+1)\right]\end{aligned} \tag{17}$$

with

$$\beta_l(k) = f(\hat{\xi}'_{l|l}(0), ..., \hat{\xi}'_{l|l}(K/2)). \tag{18}$$

The rationale behind making $\beta_l(k)$ dependent on the preliminary a-priori SNR estimates is the following: The higher the SNR, the less it should be modified by its neighbors, since it is likely to be a spectral speech harmonic, which should be preserved as much as possible. At low SNRs however, especially during speech pause,

[2] As justification for (16) some analogy to channel robust speech transmission can be stated, namely scalar quantized line-spectral frequency (LSF) parameters, which show also inter- and intraframe correlation. A 1st order Markov modeling over time and LSF index (here: frequency) was shown to give large improvements [12, section VII].

For each frame l do:
- For all $k = 0, 1, ..., K/2$ compute preliminary estimates $\hat{\xi}'_{l|l}(k)$, (15).
- Initialize $\hat{\xi}'_{l|l}(-1) := \hat{\xi}'_{l|l}(0)$ and $\hat{\xi}'_{l|l}(K/2+1) := \hat{\xi}'_{l|l}(K/2)$.
- For all $k = 0, 1, ..., K/2$ compute prediction parameters $\beta_l(k)$, (20), and final estimates $\hat{\xi}_{l|l}(k)$, (17).

Table 1. Summary of the new a-priori SNR estimation approach.

(17) performs a smoothing between adjacent frequency bins which avoids local single spectral harmonics, typically called musical noise. Therefore, we get the following constraints for designing the function (18):

$$\beta_l(k) = \begin{cases} 1 & \text{if} \quad \hat{\xi}'_{l|l}(k) \to \infty \quad \text{(speech only)} \\ 0 & \text{if} \quad \hat{\xi}'_{l|l}(k) = 0 \quad \text{(noise only).} \end{cases} \tag{19}$$

Signal-to-noise ratio and speech presence/absence probability are very closely related. Therefore we successfully employed a speech presence probability estimator

$$\beta_l(k) = 1 - \hat{q}(k, l) \tag{20}$$

with $\hat{q}(k, l)$ being the speech *absence* probability as defined in [13, equation (16)]. An alternative realization of a speech absence/presence estimator can be found e.g. in [14]. In summary, the proposed a priori SNR estimation is listed in Table 1.

4. EXPERIMENTAL RESULTS

For experimental evaluation we employed the simple noise spectral variance estimation as described in section 2.3.

As a *baseline system* we employed a Wiener-type of filter as weighting rule [11] for the spectral amplitudes

$$G'_l(k) = \frac{\hat{\xi}'_{l|l}(k)}{1 + \hat{\xi}'_{l|l}(k)}, \quad \hat{X}_l(k) = G'_l(k) \cdot Y_l(k), \tag{21}$$

formulated using a-priori SNR values that are computed via Ephraim and Malah's approach (15).

Our *new scheme* employed the improved a-priori SNR estimation following Table 1 and applied the weighting rule

$$G_l(k) = \frac{\hat{\xi}_{l|l}(k)}{1 + \hat{\xi}_{l|l}(k)}, \quad \hat{X}_l(k) = G_l(k) \cdot Y_l(k). \tag{22}$$

Simulating the proposed new algorithm it turns out that musical tones are significantly reduced, while the speech signal distortion remains the same as with the reference system. In Fig. 1 an exemplary part of a signal is shown (sampling frequency $f_s = 8$ kHz, additive street noise at about 0 dB). It can be seen that our new approach significantly reduces the musical tones in those signal segments, where only background noise is present (see especially frame index 1...15). The signal quality during active speech periods remains almost equal. Therefore we can conclude that using our a-priori SNR estimator (17) with a correlation parameter being a measure of speech presence, mainly the statistical dependence of the noise in different frequency bins is exploited. The

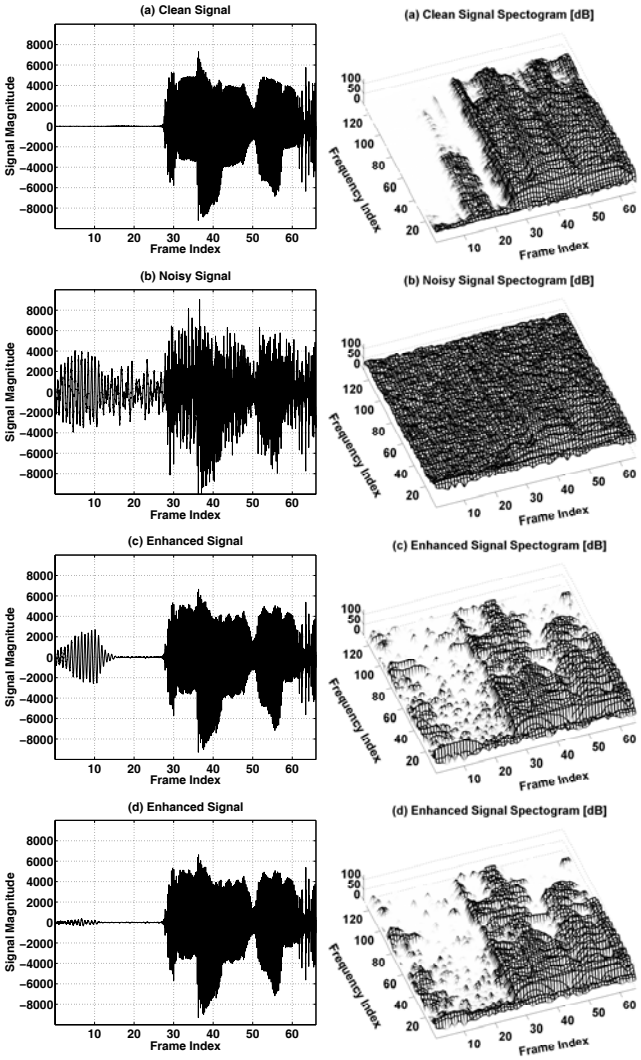

Fig. 1. Signal frame and short time Fourier spectrum of the (a) clean signal, (b) noisy signal, (c) enhanced signal of baseline system, (d) enhanced signal of proposed algorithm.

solution obtained shows a similar performance as the MMSE LSA weighting rule by Ephraim/Malah [5] which is improved by our a-priori SNR estimator. It is left open to the reader to derive different practical solutions to (11) that focus more on reducing the speech signal distortion.

5. CONCLUSIONS

In this paper we have shown how to overcome the commonly used assumption of statistical dependence of frequency bins in speech enhancement. It turned out that state-of-the-art weighting rules can still be used under the new, relaxed assumptions. On the basis of these fundamental findings we proposed a new a-priori SNR estimator closely related to the one by Ephraim and Malah, however with some *post-processing* based e.g. on a speech presence measure, that represents the correlations of adjacent frequency bins. This new technique yields state-of-the-art quality during active speech segments, however significantly reduces musical noise in segments where only noise is present. Apart from the specific proposed estimator, our general approach now opens the door for a variety of solutions to an improved computation of the a-priori SNR.

6. REFERENCES

[1] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, Dec. 1979.

[3] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

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

[6] C. Beaugeant and P. Scalart, "Speech Enhancement Using a Minimum Least Square Amplitude Estimator," in *Proc. of IWAENC*, Darmstadt, Germany, Sept. 2001, pp. 191–194.

[7] G. Whipple, "Low Residual Noise Speech Enhancement Utilizing Time-Frequency Filtering," in *Proc. of ICASSP'94*, Adelaide, Australia, Apr. 1994.

[8] Z. Goh, K.-C. Tan, and B.T.G. Tan, "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Smoothing," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[9] J. Jensen and J.H.L. Hansen, "Speech Enhancement Using a Constrained Iterative Sinusoidal Model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–739, Oct. 2001.

[10] I. Cohen, "On the Decision-Directed Estimation Approach of Ephraim and Malah," in *Proc. of ICASSP'04*, Montreal, Canada, May 2004.

[11] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP'96*, Atlanta, GA, May 1996, pp. 629–632.

[12] T. Fingscheidt, T. Hindelang, R.V. Cox, and N. Seshadri, "Joint Source-Channel (De-)Coding for Mobile Communications," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 200–212, Feb. 2002.

[13] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 112–116, Apr. 2002.

[14] D. Malah, R.V. Cox, and A.J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," in *Proc. of ICASSP'99*, Phoenix, AZ, Mar. 1999, pp. 201–204.