Overcoming the Statistical Independence Assumption w.r.t. Frequency in Speech Enhancement

Tim Fingscheidt, Christophe Beaugeant, Suhadi Suhadi

Siemens AG, COM Mobile Devices, Grillparzerstr. 10-18, D – 81675 Munich, Germany, E-Mail: tim.fingscheidt@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

Conventional approaches to speech enhancement usually 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: (1) The short-time Fourier transform and its window function, and (2) the characteristic of background noises (which are not spectrally white) and speech signals (which inherit some spectral envelope as they are generated by the vocal tract).

Following to some extent the terminology used by Cohen [1], in section 2 we will introduce an estimator (in analogy to a weighting rule) that uses *all* observations available: The neighboring, and all past ones. *Without* the assumption of statistical independence of adjacent frequency bins, this approach exploits inter- as well as intra-frame correlations *during* the estimation process. In analogy to [1], we will show that our general problem formulation still allows to use all known weighting rules, and directly leads to the formulation of a speech spectral variance estimator and an a-priori SNR estimator. Section 3 gives a practical solution for a 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

After short-time Fourier transform (STFT) of length K, the noisy observation spectral magnitude with additive noise assumption at time instant (or frame number) l can be expressed as follows

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

with k being the frequency index regarded in the following only from 0 to $K/2$, $X_l(k)$ and $N_l(k)$ being the clean speech and noise spectral magnitudes, respectively. Let's denote then $\mathcal{Y}_0^l = \{\mathcal{Y}_0^l(0), ..., \mathcal{Y}_0^l(K/2)\}$ with $\mathcal{Y}_0^l(k) = \{Y_0(k), ..., Y_l(k)\}$ as knowledge about all observations from time instant 0 until l.

Without restriction to any specific distortion measure, we assume an arbitrary distortion $D[X_l(k), \hat{X}]$ to be minimized. An estimator for the spectral amplitude $X_l(k)$ given all observations until time instant l is then

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

Assumption: Given 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 $\tilde{k} \neq k$. With this relaxed assumption, (2) can be solved in analogy to [1] yielding as result, that the exploitation of all observations does not need to restate 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 [2, 3], Cohen [1], and the Wiener filter [4]).

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

$$\hat{\lambda}_{X_l}(k) = E\{|X_l(k)|^2 | \mathcal{Y}_0^l\}. \tag{3}$$

We assume having *preliminary* estimates $\hat{\lambda}'_{X_l}(\kappa)$, $\kappa = 0, ..., K/2$, according to any state-of-the art approach, available. While each single $\hat{\lambda}'_{X_l}(\kappa)$ 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 $\hat{\lambda}'_{X_l}(\kappa)$, $\kappa = 0, ..., K/2$ allows us to exploit statistical dependence also over frequency in a second step.

On the basis of the preliminary speech spectral variance estimates, our speech spectral variance is estimated as

$$\hat{\lambda}_{X_l}(k) = E\{|X_l(k)|^2 | \hat{\lambda}'_{X_l}(0), ..., \hat{\lambda}'_{X_l}(K/2)\}. \tag{4}$$

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

3 New Approach to A-priori SNR Estimation

Subsequent to noise estimation, we assume that a noise spectral variance $\lambda_{N_l}(k) = E\{|N_l(k)|^2|\mathcal{Y}_0^l\}$ being estimated from the observations $\mathcal{Y}_0^l$ is available. Hence, we can restate (4) as a so-called *a-priori SNR* estimator

$$\hat{\xi}_l(k) = E\{|X_l(k)|^2|\hat{\xi}_l'(0), ..., \hat{\xi}_l'(K/2)\}/\lambda_{N_l}(k) \qquad (5)$$

with the a-priori SNR [2] defined as $\xi_l(k) := \frac{\lambda_{X_l}(k)}{\lambda_{N_l}(k)}$.

A practical solution to (5) consists of preliminary a-priori SNR estimates $\hat{\xi}_l'(k)$ by the well-known decision-directed estimator of Ephraim and Malah [2]. Assuming them having Markov property over frequency, and introducing a correlation parameter $\beta_l(k)$, eq. (5) can be formulated as

$$\hat{\xi}_l(k) = \beta_l(k)\hat{\xi}_l'(k) + \tfrac{1-\beta_l(k)}{2}\left[\hat{\xi}_l'(k-1) + \hat{\xi}_l'(k+1)\right] \qquad (6)$$

with $\beta_l(k) = f(\hat{\xi}_l'(0), ..., \hat{\xi}_l'(K/2))$.

The rationale behind making $\beta_l(k)$ dependent on the preliminary a-priori SNR estimates is the following: The higher the SNR, the more likely it is a spectral speech harmonic (which should be preserved), the less it should be modified by its neighbors. At low SNRs however, especially during speech pause, (6) performs a smoothing between adjacent frequency bins which avoids local single spectral harmonics, typically called musical noise.

Therefore, we successfully employed a speech presence probability estimator $\beta_l(k) = 1-\hat{q}(k,l)$ with $\hat{q}(k,l)$ being the speech *absence* probability as defined in [5, eq. (16)].

4 Experimental Results

In our experiments, we employed the Wiener filter in [4] as weighting rule for the spectral amplitudes

$$G_l'(k) = \tfrac{\hat{\xi}_l'(k)}{1+\hat{\xi}_l'(k)}, \quad \hat{X}_l(k) = G_l'(k) \cdot Y_l(k). \qquad (7)$$

The a-priori SNR values are computed twice: via Decision-directed approach [2] (*baseline system*) and according to eq. (6) (*proposed algorithm*).

In Fig. 1, an exemplary part of a signal is shown as an example of the result (sampling frequency $f_s = 8$ kHz, additive street noise at about 0 dB). It is depicted that our new approach can significantly reduce the musical tones in speech absence (see especially frame index 1...15), and the signal quality during speech presence is kept almost equal to the reference system. Hence, we can conclude that using our a-priori SNR estimator (6) with a correlation parameter being a measure of speech presence, mainly the statistical dependence of the noise in different frequency bins is exploited.

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.

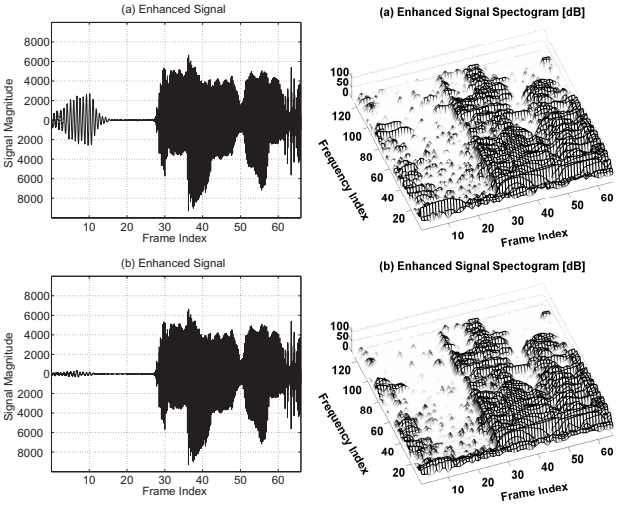

Figure 1: The enhanced signal and its STFT spectra of (a) baseline system, (b) the proposed algorithm.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.