A SIMPLE AND EFFECTIVE FRAMEWORK FOR A PRIORI SNR ESTIMATION

Johannes Stahl∗, Pejman Mowlaee†

∗Signal Processing and Speech Communication Lab, Graz University of Technology, Graz, Austria
† Widex A/S, Nymøllevej 6, 3540 Lynge, Denmark

ABSTRACT

The problem of estimating the a priori signal-to-noise ratio (SNR) for single-channel speech enhancement is addressed. Similar to the decision-directed approach we linearly combine the maximum likelihood estimate of the a priori SNR with an estimate obtained from the previous frame. Based on the harmonic model for voiced speech we propose to smooth the a priori SNR estimate along harmonic trajectories instead of fixed discrete Fourier transform frequency bins. We interpolate by using a pitch-adaptive zero-padding in order to obtain the spectral coefficients at harmonic frequencies. The resulting pitch-adaptive decision-directed (PADDi) method increases the noise attenuation compared to the classical decision-directed approach and outperforms benchmark methods in terms of speech enhancement performance for several noise types at different SNRs, quantified by objective evaluation criteria.

Index Terms— speech enhancement, a priori snr, decision-directed, pitch-adaptive

1. INTRODUCTION

Speech enhancement algorithms are often formulated and implemented in the discrete short-time Fourier transform (DSTFT) domain by applying a signal-dependent spectral gain function. There exist various gain functions in the literature such as the Wiener filter, the minimum mean square error short time spectral amplitude estimator (MMSE-STSA) [1], or the log spectral amplitude estimator (LSA) [2], to name a few. The vast majority of them have in common that they rely on this parameter. Since the smoothing constant is commonly chosen close to a fixed frequency bin. Since the harmonic frequencies are not necessarily a subset of the set of discrete Fourier transform (DFT) frequencies, the instantaneous SNR may abruptly change from one frame to the next. This renders the DD estimator to be biased and introduces artifacts which are often perceived as artificial reverberation [5].

There exist various approaches that take into account the speech signal’s non-stationarity. For example, Cohen in [6] considers the correlation of successive speech spectral components yielding a time-varying and frequency-dependent smoothing factor. Further, Hendriks et al. proposed to apply an adaptive time segmentation that, based on a sequence of hypothesis tests, selects which segments should contribute to the respective SNR estimate [7]. Also, non-causal estimation of the a priori SNR has been considered as a strategy to better preserve speech onsets [8]. The study in [9] thoroughly analyzes the DD estimator’s capability to preserve speech onsets in transient conditions and to suppress musical noise.

Taking into account specific speech signal models is a promising strategy to improve performance despite the aforementioned difficulties. The cepstro-temporal smoothing (CTS) [10] successfully incorporates knowledge about the harmonic nature of voiced speech into the a priori SNR estimation. The cepstrum of a speech signal can be decomposed into regions representing the spectral envelope (lower quefrency bins) and the harmonic excitation signal, ideally associated to a single peak in higher quefrency regions. Hence, given a fundamental frequency estimate, it is possible to selectively smooth speech related cepstral coefficients with a different smoothing factor than regions that are most likely dominated by noise and spectral outliers resulting from estimation errors. This selective smoothing procedure is applied on the ML estimate of the speech PSD, which is subsequently used to compute the a priori SNR.

Recently, the authors of [11] synthesized the excitation signal in the cepstral domain in order to obtain an instantaneous estimate of the a priori SNR. While weak harmonic structure can be preserved, this approach is also less sensitive to abrupt changes in the acoustic environment compared to the DD estimator. Finally, Plapous et al. proposed to regenerate degraded harmonics by introducing a nonlinearity for refinement of the a priori SNR estimate [5]. To cope with the delay of one frame introduced by the DD algorithm, they initialize the a priori SNR estimate with a two-stage procedure which re-estimates the a priori SNR based on the observation and the gain of the current time step.

Motivated by the aforementioned studies, in this paper, we revisit the DD estimator under the perspective of a harmonic signal model. More specifically, we propose to smooth the a priori SNR along harmonic trajectories instead of fixed frequency bins. Hence, the weighting factor in the DD estimator linearly combines estimates of the a priori SNR that are related to the same harmonic rather than a fixed frequency bin. Since the harmonic frequencies are not necessarily a subset of the set of discrete Fourier transform (DFT)
frequencies, we propose to interpolate to the harmonic frequencies by applying a pitch-adaptive zero-padding in the time domain. The resulting pitch-adaptive decision-directed (PADD) a priori SNR estimator is evaluated in terms of instrumental measures in combination with various gain functions and compared to benchmarks.

2. SIGNAL MODEL AND NOTATION

Under the assumption of additive noise, the observed, noise corrupted (noisy) speech signal \( y(n) \) is given by \( y(n) = x(n) + d(n) \), where \( x(n) \) is the clean speech signal, \( d(n) \) is the unwanted noise, and \( n \) is the discrete-time index. In practice, \( y(n) \) is divided into frames of length \( N \) and subsequently multiplied with a window function \( w(n) \), i.e., \( y(n, \ell) = y(n + \ell L)w(n) \), where \( w(n) \) is nonzero only within the interval \( n \in [0, N - 1] \), \( \ell \) is the frame index, and \( L \) is the frame shift (in samples). Taking the DFT of each windowed segment yields the well known DSTFT

\[
Y(k, \ell) = \sum_{n=0}^{N_DFT(\ell) - 1} y(n, \ell)e^{-j \frac{2\pi}{N_DFT(\ell)} n} = X(k, \ell) + D(k, \ell),
\]

where \( N_DFT(\ell) \) is the DFT length at frame \( \ell \), the frequency index is given by \( k \in [0, N_DFT(\ell) - 1] \), and capital letters denote the frequency domain representations of the corresponding time-domain signals (represented by lower-case letters). The DFT length is commonly chosen to be constant, i.e., \( N_DFT(\ell) \equiv N_{DFT} \). The dependency of \( N_{DFT}(\ell) \) on the frame index \( \ell \) in Eq. (1) is a key ingredient of our proposal and will be explained in Section 4.

It is common to estimate the clean speech signal by applying a multiplicative gain function \( G(\cdot) \) on the noisy signal in frequency domain. Typically, this gain function is a function of the so-called a priori SNR \( \xi(k, \ell) = \sigma_x^2(k, \ell)/\sigma_d^2(k, \ell) \) as well as the a posteriori SNR \( \zeta(k, \ell) = |X(k, \ell)|^2/\sigma_d^2(k, \ell) \), i.e., \( \hat{X}(k, \ell) = G(k, \ell, \zeta(k, \ell), \xi(k, \ell))Y(k, \ell) \), where \( \sigma_x^2(k, \ell) \) and \( \sigma_d^2(k, \ell) \) denote the speech PSD and the noise PSD, respectively. The hat symbol \( \hat{\cdot} \) denotes estimates in this paper.

3. THE DECISION-DIRECTED A PRIORI SNR ESTIMATOR

Given the a posteriori SNR estimate, the ML estimate of the a priori SNR is given by [1]

\[
\hat{\xi}_ML(k, \ell) = \zeta(k, \ell) - 1.
\]

A second estimate of the a priori SNR is obtained from the preceding frame’s speech estimate [1]

\[
\hat{\xi}_{\ell-1}(k, \ell) = \frac{|\hat{X}(k, \ell - 1)|^2}{\sigma_d^2(k, \ell - 1)}.
\]

The DD estimator linearly combines the two estimates as follows [1]

\[
\hat{\xi}_{DD}(k, \ell) = \alpha_{DD}\hat{\xi}_{\ell-1}(k, \ell) + (1 - \alpha_{DD})\max[\hat{\xi}_ML(k, \ell), 0],
\]

where \( \max[\cdot, \cdot] \) indicates the maximum operator and \( \alpha_{DD} \in [0; 1] \) is the smoothing factor commonly chosen close to one [12].

From Eq. (4), we see that the estimate of the a priori SNR for specific \( k = k' \) and \( \ell = \ell' \), \( \hat{\xi}_{DD}(k', \ell') \) strongly relies on \( \hat{\xi}_{\ell-1}(k', \ell') \), which is obtained from the speech estimate \( \hat{X}(k', \ell' - 1) \) of the preceding frame. However, especially in the case of larger frame shifts,
where $K$ is an integer constant and $\text{round}[:]$ denotes the rounding operator. The factor $K$ controls the amount of zero-padding in the DFT. Inserting (6) into (5) renders $N$ automatically smooth along harmonic trajectories instead of fixed $k$ all frequency bins the chosen window function only. This means that ideally not only we know its fundamental frequency, the amount of leakage depends on the assumption that the speech signal is perfectly harmonic and we select the DFT lengths arbitrarily long if we want to keep the computational effort reasonable. On the contrary, the DFT length needs to be at least $N$ samples long to assure no non-zero samples in $y(n, \ell)$ are neglected for the computation of $Y(k, \ell)$. If we constrain the fundamental frequency of a speech signal to lie within the interval $[f_{0,\text{min}}, f_{0,\text{max}}]$ this maps to the following bounds for $N_{\text{DFT}}(\ell)$:

$$\max \left[ N, K \frac{f_s}{f_{0,\text{max}}} \right] \leq N_{\text{DFT}}(\ell) \leq K \frac{f_s}{f_{0,\text{min}}}.$$  

(8)

Given $f_{0,\text{max}}$ and $N$, for the sake of computational efficiency, we select the minimum possible value for the factor $K$, given by

$$K = \left\lceil \frac{f_{0,\text{max}}}{f_s} N \right\rceil,$$  

(9)

where $\lceil \cdot \rceil$ denotes the ceiling operation.

4.3. Fundamental frequency estimation

Clearly, the proposed algorithm relies on a fundamental frequency estimate. Any estimation procedure may be applied, we implemented a simple autocorrelation based $f_0$-estimator [14] which works on a frame-by-frame basis, as explained in the following.

First, the autocorrelation sequence $r_{yy}(m, \ell)$ (with lag $m$) of $y(n, \ell)$ is computed. In a second step, a peak-picking within the range $m \in [f_s/f_{0,\text{max}}; f_s/f_{0,\text{min}}]$ is applied on $r_{yy}(m, \ell)$ to obtain an estimate of the fundamental period

$$\hat{T}_0(\ell) = \frac{1}{f_s} \arg \max_m r_{yy}(m, \ell).$$  

(10)

Given the fundamental period, we easily compute the fundamental frequency estimate by $\hat{f}_0(\ell) = 1/\hat{T}_0(\ell)$. To avoid abrupt changes in the DFT lengths (which yield audible artifacts), we set $\hat{f}_0(\ell) = \hat{f}_0(\ell-1)$ if $\hat{f}_0(\ell) \notin [f_0(\ell-1)-30 \text{ Hz}; f_0(\ell-1)+30 \text{ Hz}].$

4.4. Signal reconstruction

At frame $\ell = \ell'$, due to the circular convolution of $y(n, \ell')$ with the inverse DFT of $G(k, \ell')$, the inverse DFT of $\hat{X}(k, \ell')$ may result in a time domain signal $\hat{x}(n, \ell')$ with support $N_{\text{DFT}}(\ell') > N$. By applying the window function $w(n)$ of length $N$ we neglect all non-zero samples of $\hat{x}(n, \ell')$ for $n \geq N$. As a consequence, we can apply the MMSE synthesis routine for signal reconstruction from [15] if the window is chosen adequately, i.e., $\sum_{n=-\infty}^{\infty} w^2(\ell L - n) = 1.$
The speech samples for the evaluation were taken from the test set of the TIMIT core database [13], which consists of 192 utterances. The speech signals were mixed with white and babble noise taken from the NOISEX-92 database [17] and rain noise (representing an impulsive noise type) taken from [18]. We chose SNRs between –10 dB and 15 dB in 5 dB steps and followed the mixing convention recommended in [19]. All signals were sampled at 16 kHz.

For noise PSD estimation we used the estimator from [20] for all algorithms. In order to make the resulting noise PSD estimate applicable in the PADSTFT framework we linearly interpolate it to the frequency bins of the PADSTFT. In the implementation, we initialized all frequency domain vectors as zero vectors of dimension \( \text{round}(K f_\text{s}/f_{\text{0, min}}) \times 1 \), i.e., all entries with index \( k \geq N_{\text{DFT}}(\ell) \) are zero. The parameters of the \( f_0 \)-estimator were set to \( f_{0, \text{min}} = 90 \) Hz and \( f_{0, \text{max}} = 350 \) Hz, resulting in \( K = 12 \) according to Eq. (9). As window function we chose a square-root hamming window for all algorithms and set the frame length to 32 ms \((N = 512)\), with 50% overlap. All gain functions were floored to have a minimum value of \( G_{\text{min}} = -20 \) dB. The weighting factor of the DD based methods was set to \( \alpha_{\text{DD}} = 0.98 \).

Since the actual smoothing characteristics of the overall speech estimator strongly depend on the spectral gain function applied [9], we compared the classical DD and PADDi for various gain functions. We analyzed the segmental noise attenuation \( \Delta \text{SNR}_{\text{seg}} \) together with the segmental speech-to-speech distortion ratio \( \Delta \text{SSDR}_{\text{seg}} \) as explained in [21]. These measures give an insight into details of the respective suppression mechanism. The gain functions applied are the Wiener Filter (WF), the log-spectral short time spectral amplitude estimator (LSA) [2], the super-Gaussian joint maximum a posteriori amplitude and phase estimator (jMAP) [16], and the MMSE-STSA estimator [1].

In order to assess PADDi compared to other approaches for a priori SNR estimation, we report segmental SNR \( \Delta \text{SNR}_{\text{seg}} \) and perceptual evaluation of speech quality (PESQ) [22] as benchmarks we include cepstro-temporal smoothing (CTS) [10] and the harmonic regeneration noise reduction (HRNR) algorithm [5]. Both approaches also consider a harmonic model for speech, yet incorporating it in a different fashion.

5.1. Proof-of-concept

Fig. 2 illustrates the mechanism of our proposal in terms of a proof-of-concept. While the DD approach yields spurious spectral peaks that can be associated to musical noise [12], the PADDi method does not produce such artifacts\(^1\). The CTS algorithm [10] preserves the harmonic structure of the original speech signal very well, however, this is at the expense of reduced overall noise reduction compared to PADDi. The HRNR algorithm [5] also successfully suppresses isolated spectral peaks. Compared to CTS and PADDi, the spectral fine structure appears to be smeared along frequency.

5.2. Objective evaluation

Fig. 3 displays the outcome for the \( \Delta \text{SNR}_{\text{seg}} \) and the \( \Delta \text{SSDR}_{\text{seg}} \) analysis. PADDi consistently brings more noise suppression while the speech distortion level is preserved compared to the DD approach. Fig. 4 shows the comparison to other a priori SNR estimation approaches. Across all SNRs and noise types, PADDi yields an increased or similar \( \Delta \text{SNR}_{\text{seg}} \) compared to the benchmark methods. Except for the impulsive rain noise (where HRNR performs worse than the other benchmarks), all methods perform similar in terms of PESQ.

6. CONCLUSION

In this paper we proposed a new alternative to the well known decision-directed a priori SNR estimator. The core of our proposal is to change the smoothing path from fixed frequencies to harmonic trajectories. Since this requires interpolation to harmonic frequencies, we apply a pitch-adaptive zero-padding in the time domain. Applying the decision-directed approach in the so-obtained PADSTFT framework automatically yields a smoothing path along frequency bins that are dominated by the same harmonics. Compared to the classical decision-directed approach, the resulting pitch-adaptive decision-directed (PADDi) approach comes with more noise suppression while preserving the level of speech distortions. The effectiveness of PADDi in terms of speech enhancement performance is demonstrated by instrumental metrics. While the current study examines the idea of a priori SNR estimation in a pitch-adaptive framework, future work should be directed towards extending it to other parameter estimation tasks that arise in speech enhancement algorithms such as noise PSD estimation.

\(^1\)Listening examples of the proposed method can be found on [23].

Fig. 3. \( \Delta \text{SSDR}_{\text{seg}} \) versus \( \Delta \text{NA}_{\text{seg}} \) averaged over all noise types for the DD a priori SNR estimator and PADDi combined with (\( \times \)) Wiener filter, (\( \ominus \)) LSA [2], (\( \triangle \)) jMAP [16] and (\( \square \)) MMSE-STSA [1].

Fig. 4. \( \Delta \)-improvement of \( \Delta \text{SNR}_{\text{seg}} \) and PESQ for the different a priori SNR estimators. Reported as improvement over the decision-directed approach for \( \Delta \text{SNR}_{\text{seg}} \) and over the noisy observation for PESQ.
7. REFERENCES


