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 the *a priori* SNR, defined as the ratio of the speech power spectral density (PSD) and noise PSD, as a key parameter [3, 4]. Since the *a priori* SNR is not known in practice, its estimation is a crucial step in every speech enhancement algorithm that relies on this parameter.

Since its proposal in 1984, the decision-directed (DD) *a priori* SNR estimator [1] has been widely used in speech enhancement methods. The DD estimator linearly combines the maximum likelihood (ML) estimate of the *a priori* SNR of the current signal segment with an estimate obtained from the preceding estimated speech coefficients by a smoothing factor. A thorough analysis of its mechanism is given in [3], where the smoothing factor is identified to play a key role in suppressing audible distortions due to the recursive smoothing. Choosing it too small results in unwanted spectral outliers due to the high variance in the ML *a priori* SNR estimate, often perceived as musical noise [4]. On the contrary, if chosen too large, it may cause distortions of the speech signal [3].

Since the smoothing constant is commonly chosen close to one, the DD approach introduces one frame delay and the resulting estimate strongly relies on the speech spectrum estimation in the previous frame. However, as a result of onsets, offsets, and generally time-varying signal characteristics, such as the fundamental frequency, 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)

This work was supported by the Austrian Science Fund (FWF): P28070-N33.

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 (PADDi) *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 non-zero only within the interval $n \in [0, N-1]$, ℓ is the frame index, and L is the frame shift (in samples). Taking the DFT of each windowed segment yields the well known DSTFT

$$Y(k,\ell) = \sum_{n=0}^{N_{\text{DFT}}(\ell)-1} y(n,\ell) e^{-j\frac{2\pi k}{N_{\text{DFT}}(\ell)}n} = X(k,\ell) + D(k,\ell), \quad (1)$$

where $N_{\text{DFT}}(\ell)$ is the DFT length at frame ℓ, the frequency index is given by $k \in [0, N_{\text{DFT}}(\ell) - 1]$, 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., $N_{\text{DFT}}(\ell) \triangleq N_{\text{DFT}}$. The dependency of $N_{\text{DFT}}(\ell)$ on the frame index ℓ 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 $\zeta(k,\ell) = |Y(k,\ell)|^2/\sigma_d^2(k,\ell)$, 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]

$$\hat{\xi}_{\text{ML}}(k,\ell) = \hat{\zeta}(k,\ell) - 1. \quad (2)$$

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

$$\hat{\xi}_{\ell-1}(k,\ell) = \frac{|\hat{X}(k,\ell-1)|^2}{\hat{\sigma}_d^2(k,\ell-1)}. \quad (3)$$

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

$$\hat{\xi}_{\text{DD}}(k,\ell) = \alpha_{\text{DD}}\hat{\xi}_{\ell-1}(k,\ell) + (1-\alpha_{\text{DD}})\max[\hat{\xi}_{\text{ML}}(k,\ell), 0], \quad (4)$$

where $\max[\cdot,\cdot]$ indicates the maximum operator and $\alpha_{\text{DD}} \in [0;1]$ 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 $\ell = \ell'$, $\hat{\xi}_{\text{DD}}(k',\ell')$ 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,

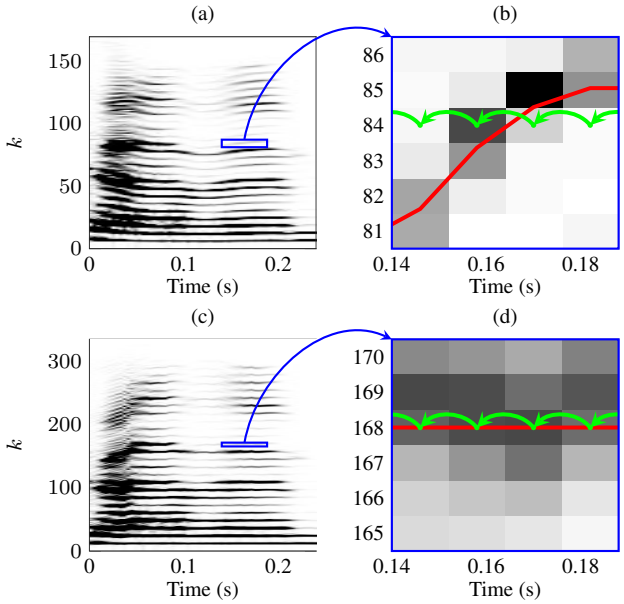

Fig. 1. (a): Spectrogram of a speech snippet uttered by a female speaker taken from [13]. (b) Zoom into a voiced time-frequency region where the fundamental frequency changes over time. The red solid line indicates the trajectory of harmonic 14. The green arrows indicate the DD smoothing path. (c) The same speech snippet as in (a) analyzed with the PADSTFT ($K = 12$ in Eq. (8)). (d) Now the DD smoothing path at frequency bin $k = 168$ and the trajectory of harmonic 14 coincide.

DFT bin k', which is dominated by speech at frame ℓ', is not necessarily dominated by speech at frame $\ell'-1$ and vice versa.

Basically, this has two reasons. First, onsets and offsets induce a change in speech presence/absence from one frame to the next (which is addressed in, e.g., [6, 7]). Second, considering voiced speech as the summation of harmonically related sinusoids has similar consequences. As illustrated in Fig. 1 (a) and (b), one harmonic does not dominate the same frequency bin for every frame, since the fundamental frequency changes over time. Hence, assuming that the *a priori* SNR is approximately constant along harmonic trajectories, it is not necessarily $\hat{\xi}_{\ell-1}(k',\ell')$ which approximates $\xi(k',\ell')$ best but potentially any other $\hat{\xi}_{\ell-1}(k,\ell')$ with k close to k'. In this work, we are interested in taking this relation into account.

4. THE PROPOSED METHOD

Following the above discussion, we propose to recursively smooth the *a priori* SNR estimates along harmonic trajectories instead of fixed frequency bins. Hence, for successive frames, we seek for frequency bins that are dominated by the same harmonic. In order to find potential candidates, we define $k_h(\ell)$, representing the frequency bin k which is closest to the h^{th} harmonic with frequency $f_h(\ell) = hf_0(\ell)$, i.e.,

$$k_h(\ell) = \arg\min_k \left| k - N_{\text{DFT}}(\ell)\frac{hf_0(\ell)}{f_s} \right|. \quad (5)$$

4.1. The pitch-adaptive decision-directed approach

We can simplify Eq. (5) by choosing $N_{\text{DFT}}(\ell)$ dependent on the fundamental frequency

$$N_{\text{DFT}}(\ell) = \text{round}\left[K\frac{f_s}{f_0(\ell)}\right], \quad (6)$$

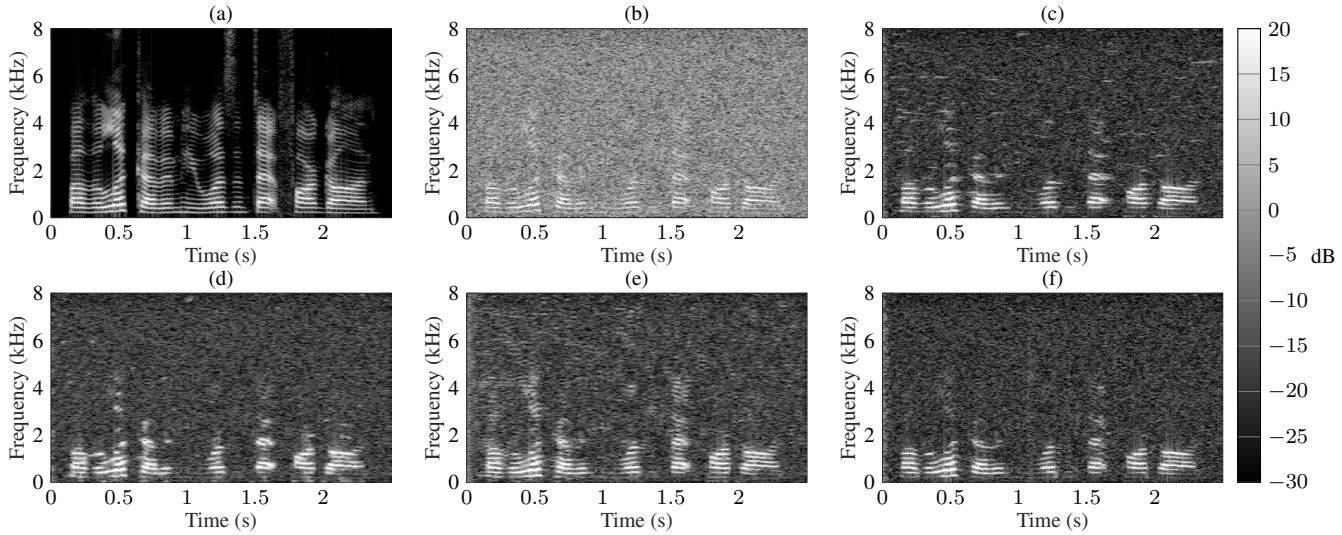

Fig. 2. Proof-of-concept: Demonstrating PADDi on a noisy speech signal mixed at 0 dB global SNR corrupted with white noise. The clean speech signal is a recording of a female speaker taken from TIMIT saying "By the look of him he wasn't that far gone.". (a) Clean speech signal, (b) noisy speech, (c) Wiener Filter + DD [1], (d) Wiener filter + HRNR [5], (e) Wiener filter + CTS [10], (f) Wiener filter + PADDi.

where K is an integer constant and $\mathrm{round}[\cdot]$ denotes the rounding operator. The factor K controls the amount of zero-padding in the DFT. Inserting (6) into (5) renders $k_h(\ell)$ to be independent of $N_{\mathrm{DFT}}(\ell)$. Further, using $K\frac{f_s}{f_0(\ell)} \approx \mathrm{round}\left[K\frac{f_s}{f_0(\ell)}\right]$, we obtain

$$\begin{aligned} k_h(\ell) &= \arg\min_k \left| k - \mathrm{round}\left[K\frac{f_s}{f_0(\ell)}\right]\frac{hf_0(\ell)}{f_s}\right| \\ &\approx \arg\min_k \left| k - Kh\right| \\ &= Kh. \end{aligned} \tag{7}$$

By applying a pitch-adaptive zero-padding, $k_h(\ell)$ becomes a constant that does not depend on the frame index ℓ anymore. Hence, $k_h(\ell)$ is consistently dominated by the same harmonic h and its argument ℓ becomes redundant, which is why we drop it in the rest of the paper. The resulting time-frequency representation is pitch-adaptive (PA) and we refer to it as PADSTFT.

As a result of spectral leakage, harmonic h not only impacts on frequency bin k_h but on all other frequency bins as well. Under the assumption that the speech signal is perfectly harmonic and we know its fundamental frequency, the amount of leakage depends on the chosen window function only. This means that ideally not only all frequency bins k_h, but also those in-between harmonics, are affected similarly by the harmonics at all time instances.

By applying the decision-directed approach as defined in Eq. (4) in the PADSTFT framework, obtained by using Eq. (6) in Eq. (1), we automatically smooth along harmonic trajectories instead of fixed frequencies as illustrated in Fig. 1, panels (c) and (d). The resulting estimator is termed pitch-adaptive decision-directed (PADDi).

4.2. The factor *K*

In principle, the larger we choose the integer factor K, the finer the resolution of the resulting DFT. However, it is evident that we cannot 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 $[f_{0,\min}, f_{0,\max}]$ this maps to the following bounds for $N_{\mathrm{DFT}}(\ell)$

$$\max\left[N, K\frac{f_s}{f_{0,\max}}\right] \leq N_{\mathrm{DFT}}(\ell) \leq K\frac{f_s}{f_{0,\min}}. \tag{8}$$

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

$$K = \lceil \frac{f_{0,\max}}{f_s} N \rceil, \tag{9}$$

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 $m \in [f_s/f_{0,\max}; f_s/f_{0,\min}]$ is applied on $r_{yy}(m,\ell)$ to obtain an estimate of the fundamental period

$$\hat{T}_0(\ell) = \frac{1}{f_s}\arg\max_m r_{yy}(m,\ell). \tag{10}$$

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 $\hat{f}_0(\ell) \notin [\hat{f}_0(\ell-1) - 30\,\mathrm{Hz}; \hat{f}_0(\ell-1) + 30\,\mathrm{Hz}]$.

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 $N_{\mathrm{DFT}}(\ell') > N$. 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., $\sum_{\ell=-\infty}^{\infty} w^2(\ell L - n) = 1$.

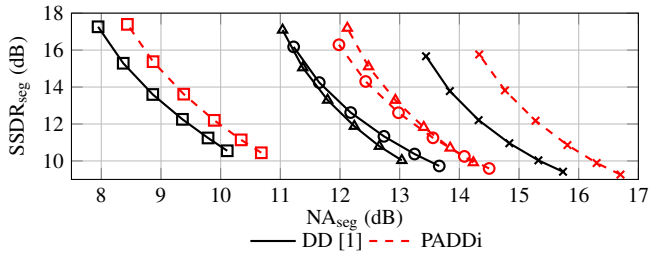

Fig. 3. SSDR$_{seg}$ versus NA$_{seg}$ averaged over all noise types for the DD *a priori* SNR estimator and PADDi combined with (×) Wiener filter, (○) LSA [2] , (△) jMAP [16] and (□) MMSE-STSA [1].

5. EXPERIMENTS

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 $\mathrm{round}[Kf_s/f_{0,\min}] \times 1$, i.e., all entries with index $k \geq N_{\mathrm{DFT}}(\ell)$ are zero. The parameters of the f_0-estimator were set to $f_{0,\min} = 90$ Hz and $f_{0,\max} = 350$ 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 $G_{\min} = -20$ dB. The weighting factor of the DD based methods was set to $\alpha_{\mathrm{DD}} = 0.98$.

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 (NA$_{seg}$) together with the segmental speech-to-speech distortion ratio (SSDR$_{seg}$) 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 (SNR$_{seg}$) 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.

[1]Listening examples of the proposed method can be found on [23].

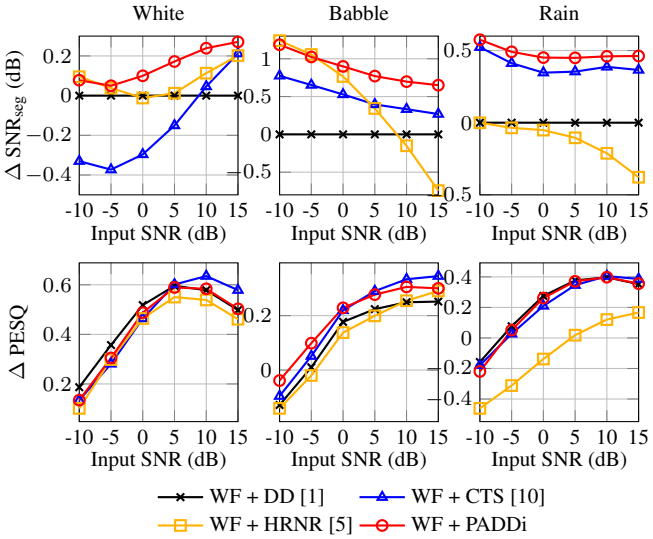

Fig. 4. Δ-improvement of SNR$_{seg}$ and PESQ for the different *a priori* SNR estimators. Reported as improvement over the decision-directed approach for SNR$_{seg}$ and over the noisy observation for PESQ.

5.2. Objective evaluation

Fig. 3 displays the outcome for the NA$_{seg}$ and the SSDR$_{seg}$ 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 SNR$_{seg}$ 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.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[3] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[4] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art*. Morgan & Claypool Publishers, 2013, vol. 9, no. 1.

[5] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2098–2108, Nov 2006.

[6] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 6, pp. 2064–2074, Nov 2006.

[8] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, Sept 2004.

[9] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-snr and transient conditions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 2, pp. 277–289, Feb 2011.

[10] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar 2008, pp. 4897–4900.

[11] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug 2017.

[12] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[14] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[15] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr 1984.

[16] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. on Advances in Signal Processing*, vol. 2005, no. 7, p. 354850, May 2005.

[17] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX–92 Study on the Effect of Additive Noise on Automatic Speech Recognition," *Technical Report, DRA Speech Research Unit*, 1992.

[18] M. Haberkorn, "Raindrops on plastic." [Online]. Available: https://www.freesound.org/people/mmorast/sounds/192149/

[19] "ITU-T P.835 Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Tech. Rep., 2011.

[20] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[21] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," Tech. Rep., 2001.

[23] J. Stahl and P. Mowlaee, "A simple and effective framework for a priori SNR estimation." [Online]. Available: http://www2.spsc.tugraz.at/people/pmowlaee/PADDi