Das, Sneha; Bäckström, Tom; Fuchs, Guillaume

Fundamental Frequency Model for Postfiltering at Low Bitrates in a Transform-Domain Speech and Audio Codec

Published in:
Proceedings of Interspeech

DOI:
10.21437/Interspeech.2020-1067

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Fundamental Frequency Model for Postfiltering at Low Bitrates in a Transform-Domain Speech and Audio Codec

Sneha Das[1], *Tom Bäckström*[1], *Guillaume Fuchs*[2]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Fraunhofer IIS, Germany

(sneha.das, tom.backstrom)@aalto.fi

Abstract

Speech codecs can use postfilters to improve the quality of the decoded signal. While postfiltering is effective in reducing coding artifacts, such methods often involve processing in both the encoder and the decoder, rely on additional transmitted side information, or are highly dependent on other codec functions for optimal performance. We propose a low-complexity postfiltering method to improve the harmonic structure of the decoded signal, which models the fundamental frequency of the signal. In contrast to past approaches, the postfilter operates at the decoder as a standalone function and does not need the transmission of additional side information. It can thus be used to enhance the output of any codec. We tested the approach on a modified version of the EVS codec in TCX mode only, which is subject to more pronounced coding artefacts when used at its lowest bitrate. Listening test results show an average improvement of 7 MUSHRA points for decoded signals with the proposed harmonic postfilter.[1]

Index Terms: Speech coding, Postfiltering, Fundamental frequency

1. Introduction

Speech coding is required in all applications which store or transmit speech. While codecs based on the ACELP provide good quality encoding for transmission, it comes at the cost of high complexity [1]. Frequency domain coding algorithms like the TCX mode in the EVS are of lower complexity and perceptually transparent at moderate to high-bitrates [2]. However, the performance degrades at lower bitrates due to spectral holes arising from the shortage of bits. The resulting sparse signals contain annoying artifacts such as musical noise.

In standard codecs pre- and post-filtering methods aid in the removal of spectral holes, thereby improving the quality of the decoded signal. For instance, the LTP-postfilter in the EVS codec improves the harmonic structure in the voiced parts of the decoded signal by attenuating the spectral energy between harmonic peaks [3]. To achieve the desired effect, it utilizes the LTP parameters transmitted in the bitstream to design an IIR comb-filter. Another approach for reducing the effect of spectral-holes is noise-filling by, for example, intelligent gap filling (IGF) which is used to fill spectral gaps at high operating bitrates, and replicate high frequency components using copy-up from lower frequencies, at lower operating bitrates [1, 4]. Further methods used to improve the quality of the decoded signal include formant enhancement and pitch sharpening [1].

While postfiltering methods are effective in improving the quality of the decoded signal, many methods require processing both at the encoder and the decoder, thus adding computational costs at both ends. This can be problematic in resource-constrained devices. Additionally, parametric postfiltering methods transmit the parameters as side information, hence requiring additional overhead in bit consumption. In comparison, conventional non-parametric methods are mostly blind enhancement techniques manually tuned to improve coded speech. Model-based postfiltering approaches, which employ information on the inherent structure of speech, show improvements in decoded speech quality: In our previous work, we have proposed postfiltering methods based on spectral envelope modelling [5] [6]. While the methods demonstrate considerable output gain, their performance could potentially be improved by using a model for the fundamental frequency, F_0.

With the focus on improving the harmonic structure of coded signals, we begin with a simple model: a linear post-filter of low-complexity with an integrated F_0 model. Since this work is part of a larger objective of devising methods suitable for acoustic sensor networks with resource-constrained devices, we design the postfilter such that it is a standalone functional block at the decoder and does not need the transmission of any side information. For realistic and fair evaluation of the postfilter, we need a low-complexity codec optimized for speech. While the TCX mode in codecs like the EVS are low in complexity,

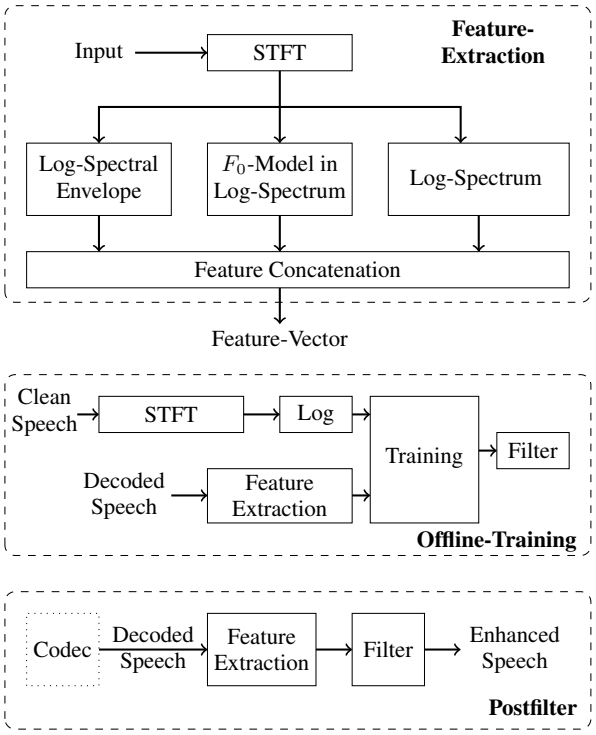

Figure 1: *Block diagram of the system architecture.*

[1]Sample speech files at: *https://harmonicPFSoundSamples*

http://dx.doi.org/10.21437/Interspeech.2020-1067

Table 1: *Complexity of the proposed algorithm.*

Stage	Process	WMOPS
Features	Cepstrum (log-magnitude)	0.3099
	linear → log-magnitude	0.3099
Filtering	Matrix multiplication: $\mathbf{Ad}$	17.163
Postprocessing	Magnitude→Complex	0.0119
	log-magnitude→linear	0.2980
Total		17.88

it is designed to complement speech-oriented algorithms like ACELP. Therefore, we utilize a modified version of the EVS codec in TCX mode, which was optimized to achieve a fair quality for speech signals even at low bit-rates, without the additional complexity of ACELP. Furthermore, since the proposed postfilter is agnostic to the codec, it can be applied on any low bit-rate transform-domain codecs like LC3 and Opus [7] [8].

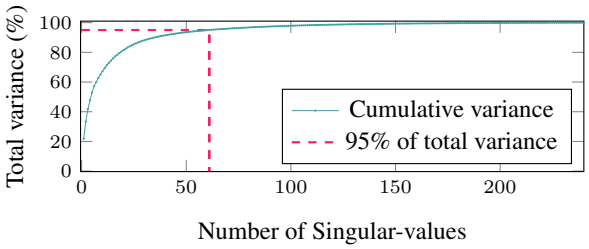

Figure 2: *Plot of the cumulative variance over the singular values of filter* $\mathbf{A}$.

2. Methodology

2.1. Signal Model

Speech is produced when a glottal excitation is shaped by the vocal tract response, whereby the vocal tract can be treated as a filter in the time-domain [1]. Convolution in the time-domain corresponds to multiplication in the frequency-domain, and summation in the log-frequency domain. Hence, in the log-frequency domain, clean speech can be represented as $\mathbf{s} = \mathbf{x_{F_0}} + \mathbf{x_{env}}$, where $\mathbf{x_{F_0}} \in \mathbb{R}^{K\times 1}$ is the excitation and is a function of F_0, the spectral envelope $\mathbf{x_{env}} \in \mathbb{R}^{K\times 1}$ represents the vocal tract response, and K is the number of frequency-bins [1].

We model the decoded signal $\mathbf{y}$ as follows: $\log|\mathbf{y}| = \mathbf{x_{F_0}} + \mathbf{x_{env}} + \mathbf{x_n}$, where $\log|\mathbf{y}|$ is the decoded signal in the log-frequency domain, and $\mathbf{x_n}$ corresponds to the noise components in the decoded speech. To estimate the clean speech signal, we define a linear model $\hat{\mathbf{s}} = \mathbf{A}^T\mathbf{d}$, $\hat{\mathbf{s}} \in \mathbb{R}^{K\times 1}$, where $\mathbf{d} \in \mathbb{R}^{(3K+1)\times 1}$ is the feature vector comprised of the three aforementioned features computed from $\mathbf{y}$ and is composed of the representations of the speech signal: $\mathbf{x_{F_0}}, \mathbf{x_{env}}$, and the coding-noise: $\mathbf{x_n}$. To obtain the harmonic model $\mathbf{x_{F_0}}$, we first compute the fundamental frequency by picking the largest peak, corresponding to the range 50 to 400 Hz, in the cepstral domain,

$$F_0 = \max(\mathcal{F}^{-1}\{\log|\mathbf{y}|\}), \tag{1}$$

and then transform it back to the frequency domain to obtain a modulated sinusoid $\mathbf{x_{F_0}}$, whose frequency matches the F_0 of the input signal. We use linear prediction on the original signal $\mathbf{y}$ to derive an envelope model $\mathbf{x_{env}}$, and utilize the log-magnitude spectrum of $\mathbf{y}$ to model the coding-noise.

Our goal is to define a filter $\mathbf{A}$, which partitioned as $\mathbf{A} = [\mathbf{A_{F_0}}, \mathbf{A_{env}}, \mathbf{A_n}, \mathbf{b}]^T$, such that $\hat{\mathbf{s}} = \mathbf{A_{F_0}}\mathbf{x_{F_0}} + \mathbf{A_{env}}\mathbf{x_{env}} + \mathbf{A_n}\mathbf{x_n}$, where $\mathbf{A} \in \mathbb{R}^{(3K+1)\times K}$, $\mathbf{b} \in \mathbb{R}^{K\times 1}$ is the bias vector, and $\mathbf{A_{F_0}}, \mathbf{A_{env}}, \mathbf{A_n} \in \mathbb{R}^{K\times K}$ are the regions of the filter, constituting information about the harmonic structure, spectral envelope and coding-noise, respectively. We estimate $\mathbf{A}$ from data, and begin by defining the error matrix as:

$$\mathbf{E}^T = \hat{\mathbf{S}} - \mathbf{S} = \mathbf{A}^T\mathbf{D} - \mathbf{S}, \tag{2}$$

$\mathbf{D} \in \mathbb{R}^{(3K+1)\times N}$ is the feature matrix, $\mathbf{S} \in \mathbb{R}^{K\times N}$ is the true log-spectrum of speech, and N is the number of speech frames used for training. To obtain an optimum filter in the minimum mean square error (MMSE) sense, we minimize the mean of squared error and subsequently solve the equation as follows:

$$\frac{\partial \mathrm{Tr}(\mathbf{E}^T\mathbf{E})}{\partial(\mathbf{A})} \triangleq 0 \implies \mathbf{A} = (\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{S}^T. \tag{3}$$

2.2. System Overview

As motivated in Sec. 2.1 we model speech as the sum of the logarithm of the spectral envelope and the F_0-model. Therefore, as input features we have chosen (i) F_0-model (ii) logarithm of the spectral envelope, (iii) log-magnitude spectrum of the decoded speech. The feature vector is computed from the decoded signal, thus jointly modelling the speech and the coding-noise. The pre-processing involves transforming the time-domain signal to the frequency-domain using STFT, with 30 ms window sizes and an overlap of 10 ms. The block-diagram of the system is presented in Fig. 1.

We investigate the modelling approach in both the frequency domain (FD) and the perceptual domain (PD) to determine the domain more suited to the proposed approach. For the PD approach, the signal is transformed to the perceptual-domain by weighting the spectrum with the perceptual envelope, following which it is transformed to the log-domain; the perceptual-domain transformation is omitted for the FD approach. We compute the spectral envelope from the decoded signal using linear prediction [1]. To compute the F_0-model we use the largest cepstral coefficient i as per Eq. 1 and the two adjacent coefficients $i-1, i+1$ to obtain a modulated harmonic signal vector. In this work we apply the proposed postfilter along with the LTP-postfilter [9] since it sharpens the pitch contour of the decoded signal and hence, aids in the extraction of the cepstral peak.

2.3. Computational Complexity

The algorithmic complexity is provided in Table 1 in terms of weighted million operations per second (WMOPS) [10]. We mainly present the complexity of the post-filtering part in the algorithm and the standard pre- and post-processes comprising of STFT and transformations between the frequency- and perceptual-domains are excluded in the analysis. The main contribution of the complexity is from the filtering operation, at ≈17.88 WMOPS, which is close to the magnitude of complexity of the underlying encoder and decoder. However, since 25% of the singular values of $\mathbf{A}$ represent 95% of the variance as shown in Fig. 2, we can correspondingly reduce the dimensions of $\mathbf{A}$ to obtain ≈75% reduction in complexity, with only minor reduction in accuracy.

3. Evaluation and Results

We use the TIMIT database for training and evaluation of the postfilter. The filter was trained over 1000 randomly chosen

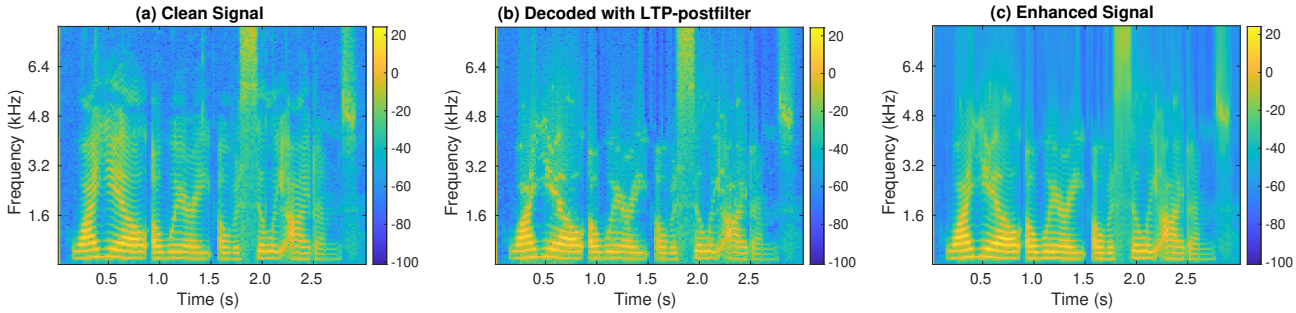

Figure 3: *Spectrograms of a test sample: (a) Clean signal, (b) Decoded signal with LTP-postfilter, (c) Enhanced signal.*

speech sentences from the training-set, comprising of 340 sentences spoken by females and the rest spoken by males; this distribution of male vs. female samples is due to the composition of the dataset, containing 70% male and 30% female samples [11]. Since we observed a difference in performance between male and female samples, we train the system separately for males and females and the results presented here are using gender specific filters. However, later experiments with gender blind filters produced identical objective results and further analysis suggests that the difference in performance between genders is mostly connected to the harmonic model, which was found to be more accurate for higher F_0 values.

The spectrograms of the clean, decoded and estimated signals of a test sample are shown in Fig. 3. As seen in the example, the harmonic structure of the estimated signal is more pronounced and closer to the clean speech than that of the decoded signal with the LTP-postfilter, which contains considerable spectral-holes and temporal discontinuity.

3.1. Objective Results

We tested the system over 118 test samples with 40 female and 78 male samples; each sample is approximately 4 seconds long and is randomly selected from the test set of the TIMIT database. The evaluation results are described separately for females (F) and males (M) and, the FD and PD approaches. We evaluate the system in terms of the Perceptual SNR (PSNR), PESQ and POLQA [1].

The distributions of the objective results are shown in Fig. 4 using the violin-plots, which depicts 1. the summary statistics of the results using the median and inter-quartile range (IQR), 2. the density trace of the results [12]. Note that we use the decoded signal with LTP-postfilter as the baseline reference, represented as dec-F and dec-M in Fig. 4. The PSNR is the ratio of the signal to noise in the perceptual domain and the perceptual weights are computed from the spectral envelopes of the clean and decoded signals. Fig. 4 (a), (d) show the absolute PSNR and ΔPSNR, i.e., difference between the estimated and decoded signals. We observe that the PSNR of the estimated signal is higher than the decoded signal by $\approx$1 dB and this improvement is consistent for both the domains and genders. However, PSNR of the estimate for males using the PD approach has a higher IQR, i.e., higher variability in the PSNR.

The absolute and ΔPESQ scores are presented in Fig. 4 (b), (e), respectively. For females, the PESQ score of the estimated signal is higher than the PESQ score of the decoded signal, both in the FD and PD. However, on average the PESQ score of the estimated signal for males is lower than that of the decoded signal. Furthermore, the POLQA and ΔPOLQA scores presented in Fig. 4 (c), (f) demonstrate a positive POLQA improvement, on average, for both genders. In addition, female samples show higher POLQA improvement in contrast to male samples; the IQR for females lies between 0.5 to 0.75 ΔMOS.

3.2. Subjective Results

For subjective evaluation of the proposed postfiltering method, we conducted a MUSHRA listening test. The test, comprising of 12 items with 6 conditions each, were presented to every participant. The conditions are (1) Lower-anchor: signal low-pass filtered at 3.5 kHz, (2) LTP postfilter: decoded signal with LTP postfiltering, (3) Postfilter-FD: proposed postfilter using FD approach, (4) Postfilter-PD: proposed postfilter using PD approach, (5) Hidden Reference, (6) Decoded: noisy signal.

Out of the total 12 samples, 6 were male and 6 were female. The samples were selected as follows: (a) 4 samples, 2 female and male each, were randomly picked from the test set of 118 samples (items F5, F6, M5, M6). (b) 4 samples, 2 female and males each which showed the highest POLQA improvements were included in the test (items F1, F2, M1, M2). (c) The remaining 4 samples comprised of the 2 male and 2 female samples showing the least POLQA improvement. We adopted this approach to select samples in order to obtain an indication of the subjective performance bounds. Additionally, the extreme samples could potentially show conditions under which the system performs optimally or under-performs.

The results of the test with 10 naive listeners are depicted in Fig. 5 in absolute scores (a) and difference scores (b) with respect to the decoded signal. The LTP-posfiltered condition is rated higher than the decoded condition for all the items and conditions. Also, the proposed postfilter in both the domains is rated higher than decoded signal on average. Postfilter-PD is scored highest for most of the items and the improvement is generally largest for female items. However, the scores of postfilter-FD relative to the LTP-postfiltered condition is inconsistent over the items. The PD-postfiltering approach is rated higher than the decoded signal by 7 MUSHRA points, on average. Hence, we can conclude that the proposed postfiltering method with F_0-model improves the quality of the decoded signal both objectively and subjectively and this improvement is consistently observed in the PD-approach.

4. Conclusions

In this paper, we proposed a postfiltering method for speech and audio coding which incorporates the inherent information in speech, specific to its harmonic structure. Since the features of the postfitler are computed from the decoded signal, it adapts to the temporal changes in speech signals. Objective evaluations of the method demonstrate positive improvement in the PSNR,

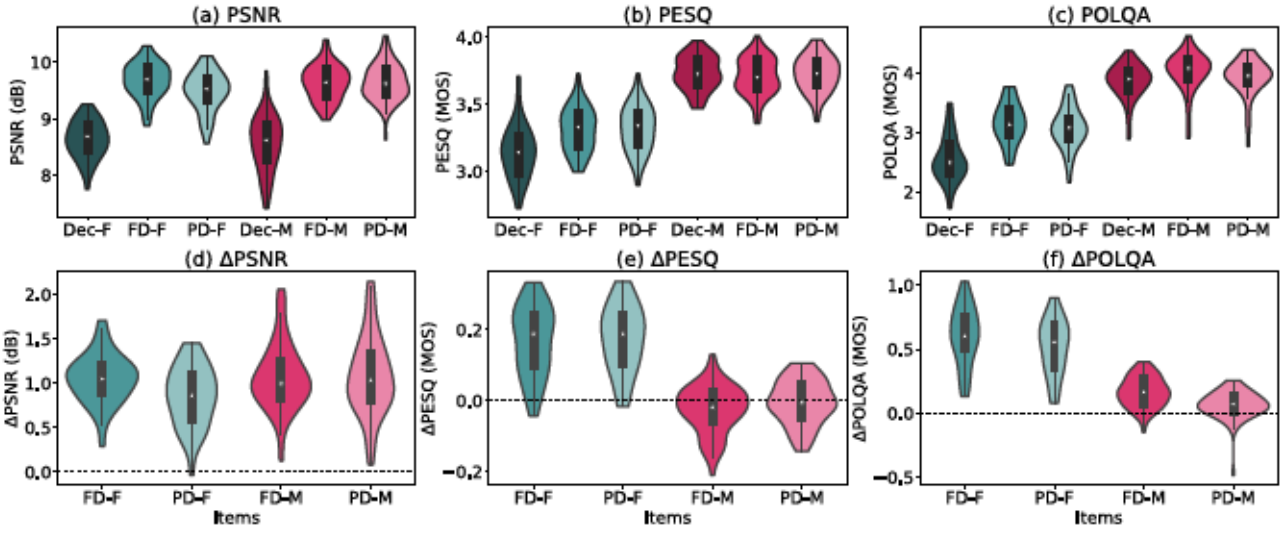

Figure 4: *Distribution of the objective-measures via summary statistics and density trace of the absolute and Δ scores of the PSNR, PESQ, POLQA, for female (F) and male (M) samples in the Frequency domain (FD) and Perceptual domain (PD).*

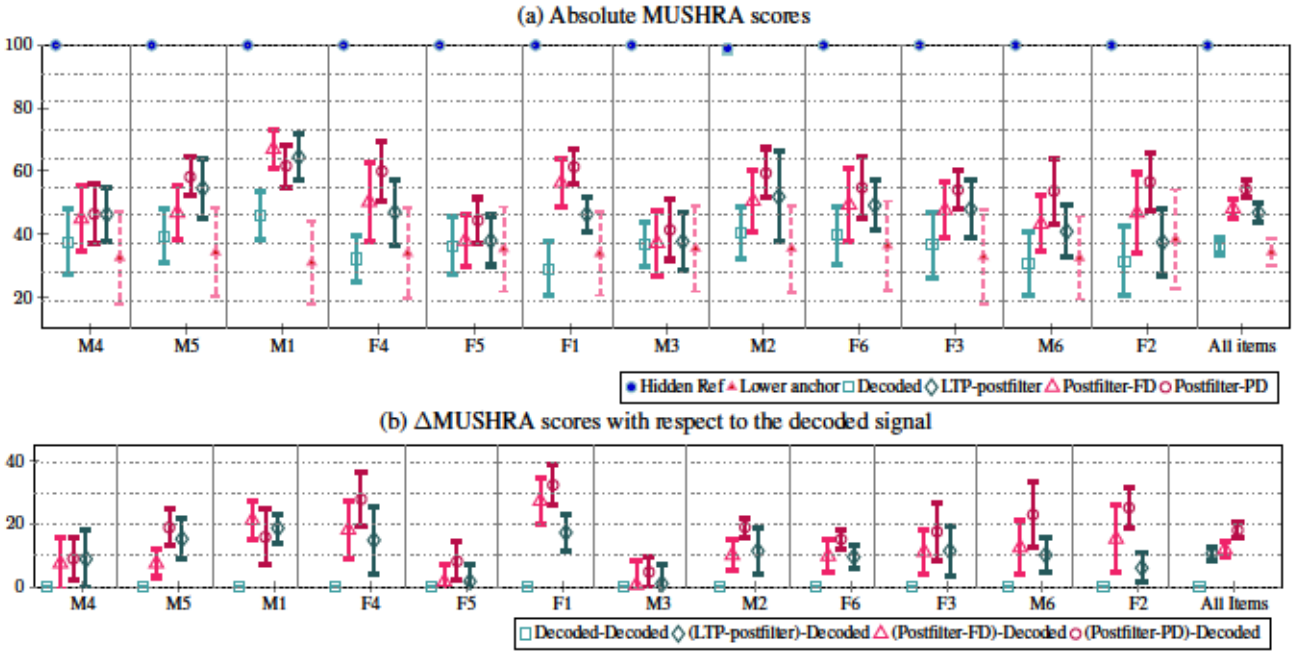

Figure 5: *Results of the MUSHRA listening test: (a) Absolute MUSHRA scores, (b) ΔMUSHRA scores with respect to decoded signal.*

PESQ and POLQA scores of the samples postfiltered with the proposed method, with respect to the decoded signal. However, the improvement is higher for females than for males. In terms of the subjective evaluation of the method, the MUSHRA listening test indicates a higher rating for the samples postfiltered using the proposed PD approach. The MUSHRA score of the PD approach is on average 7 MUSHRA points higher than the decoded signal with LTP-postfiltering. Furthermore, from the perceptual analysis of the samples we conclude that the proposed postfilter aids in removing the artefacts caused due to the discontinuities in the harmonic structure in decoded signal. In addition, at mid- to high-frequency ranges the enhanced signal is marginally biased towards higher energies, thereby imparting a faintly rough characteristic to the signal. This issue will be addressed by a more accurate model of the harmonic structure in coherence with the coding-noise in future work. The proposed method is complementary to the postfilters which model the spectral envelope of speech, whereby the models together can comprehensively enhance the decoded signal [5]. Unifying the envelope and harmonic modelling approaches along with phase modelling is left for future work. Further on, while the complexity of the postfilter is close to the complexity of the underlying codec, 95% of the variance is contained in 25% of the filter coefficients, whereby a follow-up dimensionality reduction on the postfilter can reduce the complexity by 75%.

5. References

[1] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017.

[2] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the EVS codec architecture," in *ICASSP*. IEEE, 2015, pp. 5698–5702.

[3] "EVS codec detailed algorithmic description; 3GPP technical specification," http://www.3gpp.org/DynaReport/26445.htm.

[4] S. Disch, A. Niedermeier, C. R. Helmrich, C. Neukam, K. Schmidt, R. Geiger, J. Lecomte, F. Ghido, F. Nagel, and B. Edler, "Intelligent gap filling in perceptual transform coding of audio," in *Audio Engineering Society Convention 141*. Audio Engineering Society, 2016.

[5] S. Das and T. Bäckström, "Postfiltering using log-magnitude spectrum for speech and audio coding," in *Interspeech*, 2018.

[6] S. Das and T. Bäckström, "Postfiltering with complex spectral correlations for speech and audio coding." in *Interspeech*, 2018.

[7] E. T. . . V. (2018-09), "Study of super wideband codec in DECT for narrowband, wideband and super-wideband audio communication including options of low delay audio connections," https://www.etsi.org/deliver/etsi_tr/103500_103599/103590/01.01.01_60/tr_103590v010101p.pdf.

[8] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," 2012.

[9] G. Fuchs, C. R. Helmrich, G. Marković, M. Neusinger, E. Ravelli, and T. Moriya, "Low delay LPC and MDCT-based audio coding in the evs codec," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5723–5727.

[10] ITU, "ITU-T software tool library 2009 user's manual," https://www.itu.int/rec/T-REC-G.191/en.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[12] J. L. Hintze and R. D. Nelson, "Violin plots: a box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.