LOW-COMPLEXITY POSTFILTER USING MDCT-DOMAIN FOR SPEECH AND AUDIO CODING

Sneha Das, Tom Bäckström

Department of Signal Processing and Acoustics, Aalto University, Finland
sneha.das@aalto.fi, tom.backstrom@aalto.fi

Abstract: Postfilters are used in speech and audio codecs to improve the quality of the decoded signal. Recent studies have shown that postfilters using speech models can considerably improve the output quality of the signal. In our previous work, we proposed a postfilter for frequency-domain speech coding which took advantages of the time-frequency correlation inherent to the speech magnitude spectrum. However, while the proposed approach showed substantial improvement in the output speech quality, it operates in the STFT-domain, whereas state-of-art transform domain coding is implemented in the modified discrete cosine transform (MDCT) domain. As a result, the decoded signal has to be first transformed back to the time-domain and then to the frequency-domain before the postfilter can be applied, thus increasing the algorithmic delay and complexity. In this work, we adapt the context-based postfiltering method to the MDCT-domain, such that the postfilter can be directly applied on the decoded signal within the codec framework. However, we observe that the gains obtained from postfiltering in the MDCT-domain is half of that in the STFT-domain. Further analysis indicates that the correlation between coefficients in the MDCT-domain is lower than the correlations in the STFT-domain. Hence, the modelling methods employed in the STFT-domain postfilter are not directly applicable to a postfilter using the MDCT. To improve performance, we propose to jointly model a pair of MDCT coefficients as a complex-valued pseudo spectrum. The proposed approach shows an average perceptual-SNR (PSNR) improvement of 1.5 dB over a plain MDCT approach, which is similar to the quality with a STFT-domain postfilter.

Introduction

Speech and audio codecs are integral in systems storing, transmitting and processing audio signals. To aid the coding of signals with different characteristics, recent state-of-art speech and audio codecs, like the EVS [1], are hybrids of time-domain coding like the ACELP, and frequency-domain coding. Time-domain coding works on the principle of analysis-by-synthesis, and are particularly efficient in encoding signals with higher temporal variations and with a single dominant fundamental frequency. In contrast, frequency-domain codecs efficiently encode signals with more stationary characteristics, like sustained vowel segments in speech and musical instruments with harmonic sounds.

While codecs perform comfortably in their prime bitrate-range, at bitrates outside this range the performance of the codecs deteriorate. The quality of frequency-domain codecs deteriorates at low bitrates, in particular due to zeros in the spectrum which result in perceptually annoying distortion known as musical noise. Speech codecs often use postfilters to improve the quality of the decoded signals [2]. The most frequently used postfilters are formant enhancement and bass-postfilter methods, which are often used in CELP codecs [2]. They work by reducing energy in

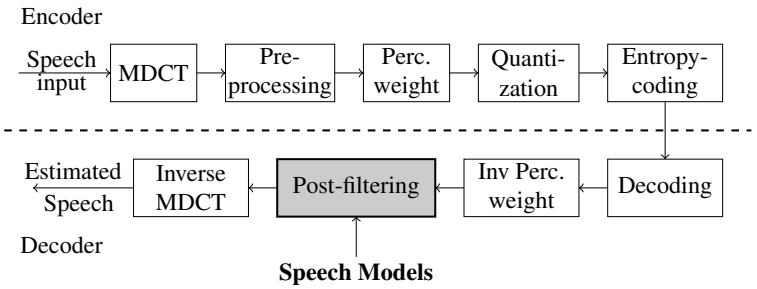

Figure 1 – System overview showing the placement of the post-filtering block.

areas of the signal spectrum which are often noisy, for instance, the regions where the perceptual envelope lies lower, e.g. between formant peaks, and between harmonic peaks. In particular, the bass-postfilter applied in TCX-modes of EVS uses the transmitted LTP parameters to determine a filter which is used to reduce noise between harmonic peaks [1]. Such methods are mostly non-parametric, perceptually motivated, and heuristic in nature with manually tuned parameters for best performance [3]. Intelligent gap filling [4] is a more advanced type of postfilter used in frequency-domain coding, which addresses issues caused by spectral holes at low bitrates [5]. Some methods in this class require the parameters of the postfilter to be transmitted in the bitstream, thus increasing the bit consumption. Also, a few postfilters require operational blocks both at the encoder and decoder, which adds to the computational complexity of the encoder. In contrast, our long-term objective is to address distributed coding scenarios which require a computationally light encoder and an intelligent decoder.

In our previous work [5, 6], we proposed context-based postfilters which model the intrinsic speech characteristics through adaptive speech models employed at the decoder and without the transmission of additional bits. The postfilter for frequency-domain speech coding [6] incorporated the time-frequency correlation in speech magnitude spectrum using context-neighborhood. While the context-based postfilters show considerable improvement in the decoded signal, they operate in the STFT-domain whereas state-of-the-art frequency-domain codecs use the MDCT (modified discrete cosine transform) domain. Hence, to apply the postfilter, the decoded signal has to be transformed to the STFT-domain by first transforming to the time-domain, thus increasing algorithmic delay and complexity. The motivation of this work is to obtain low-complexity and low-delay context-based postfilters which function in the MDCT domain. We begin by directly adapting the context-based postfilter in the STFT-domain [6] to the MDCT-domain and evaluate the performance of the MDCT-domain postfilter with respect to the STFT-domain results. To further improve and equate the performance of the MDCT-domain postfilter to the STFT-based postfilter [6], we propose a novel approach to represent the MDCT-coefficients as a pseudo-complex spectrum and model the context from the energy of the pseudo-complex coefficients. Through this method, we are able to better address the lower correlation between the MDCT coefficients, which can be attributed to the effect of phase differences on the magnitude of the MDCT coefficients.

System Overview

Since the focus of our application is the noisy signal at the output of frequency-domain codecs, to test the proposed postfilters we employ a codec similar to the EVS codec in the TCX mode [1]. A block diagram providing an overview on the system is depicted in Fig. 1; The input speech is transformed to the frequency domain using the MDCT, following which the frames are pre-processed and perceptually weighted. The residual is then quantized and transmitted to the decoder using entropy-coding. At the decoder, we follow the inverse process, except that post-filtering is applied after inverse perceptual weighting, before the inverse MDCT.

Algorithm 1 Estimation of signal from quantized observation (MDCT-domain)

Require: Quantized signal Y, prior-models C_1
  function ESTIMATION(Y, C_1)
    for $frame = 1 : N$ **do**
      for $k = 1 : Length(Y(frame))$ **do**
        $\mu_{up}, \sigma_{up} \leftarrow EstimateStatistics(C_1, \hat{X}_{prev})$
        $pdf \leftarrow GaussianPDF(\mu_{up}, \sigma_{up}, l(k), u(k))$
        if $\text{sign}\{u(k)\} - sign\{l(k)\} == 0$ **then**,
          $\hat{X} \leftarrow Expectation(pdf)$
        else
          $\hat{X} \leftarrow RandomSampling(pdf)$

Magnitude Modelling in the MDCT-Domain

Models which employ time-frequency correlations of speech are estimated in an offline-training phase. These models, together with the estimated speech in the past time-frames and lower frequency bins, are used in the postfilter to infer the speech probability distribution function (PDF) in the frequency-bin under consideration. Subsequently, the obtained PDF is used to estimate the clean speech. Detailed descriptions of the stages are provided in the next sections.

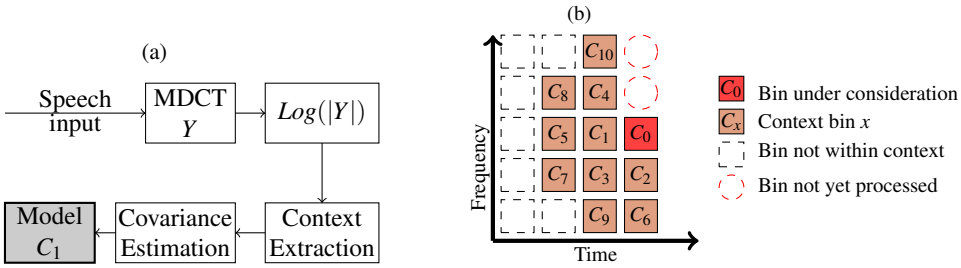

Figure 2 – a) Block diagram of the training process in the MDCT-domain. b) Context neighborhood of size $C = 10$. The previous estimated bins are chosen and ordered based on the distance from the current sample.

Training: To incorporate the time-frequency correlations in the models, we use a context, i.e., neighboring components in the time-frequency representation [5], which characterizes the time-structure of the spectral envelope. We retain the same ordering of context bins as presented in [6] to allow a comparison between the MDCT-domain models implemented in this papers and the models studied previously in the STFT-domain; the structure of the context neighborhood is depicted in Fig. 2 (b). To train the models, the speech signal is first transformed to the MDCT domain, as shown in the block diagram of the training process in Fig. 2 (a). Past studies [6] have shown that the log-spectrum of STFT is approximately Gaussian, whereby we model the MDCT spectrum in the log-domain. After transforming the coefficients to log-domain, we extract the context vector of the desired size from the input time-frequency bins following the ordering adapted from [7] and illustrated in Fig. 2 (b). The context is used to estimate the covariance matrices, which is the model C_1 and is employed in the postfilter at the decoder.

Inference: The inputs to the postfilter are the decoded signal Y and the model C_1 trained offline as described in the previous section. For every frequency-bin in each frame, we use C_1 to obtain the posterior distribution statistics in terms of the conditional mean and variance, μ_{up}, $\sigma_{up} = f[P(X|\hat{X}_{prev})]$. The quantization bin limits l, u, which serve as an accurate noise model, is provided by Y. Thus, $\mu_{up}, \sigma_{up}, l, u$ follow a truncated PDF. We obtain the speech estimate by computing the mean of truncated PDF in bins which are quantized to non-zero values. For bins quantized to zero, we randomly sample a point from the truncated PDF with the

underlying motivation that at regions of low energy, speech has a more random characteristic; this step is conceptually similar to noise-filling [2]. The detailed pseudo-code is presented in Algorithm 1.

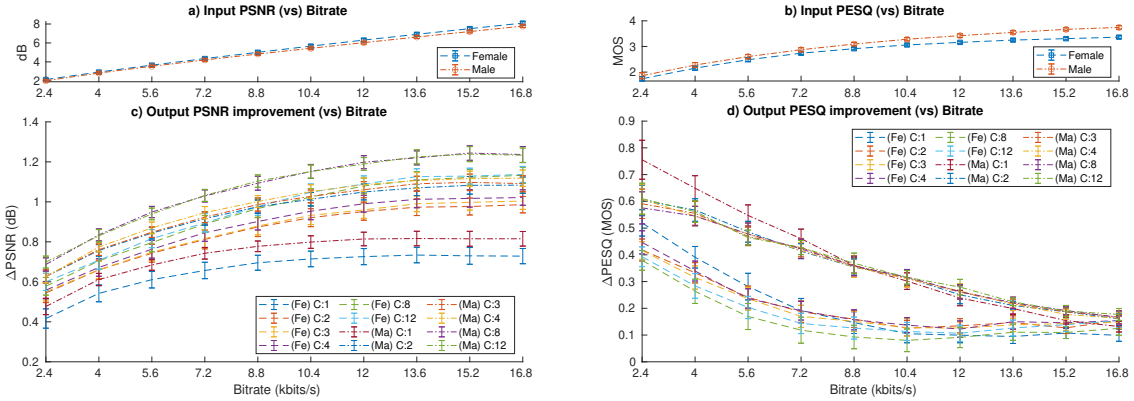

Figure 3 – (a)-(b) PSNR and PESQ scores of the input signal with respect to bitrate. (c)-(d) Δ PSNR and Δ PESQ scores of the output signal over bitrates. Fe: female samples, Ma: male samples.

Evaluation: To evaluate the system we trained the models using 250 speech samples from the training set of the TIMIT database [8], each two to four seconds long. For testing, we used the method on 40 speech samples, 20 males and females each, obtained from the test set of the TIMIT-database, for context sizes$=\{1,2,3,4,8,12\}$ and for bitrates between 2.4 to 16.8 kbits/s. The quality of the enhanced signals were measure in terms of perceptual signal-to-noise ratio (PSNR) and the PESQ scores.

The results are shown in Fig.3 in terms of the Δ PSNR$=(output-input)_{PSNR}$ and Δ PESQ$=(output-input)_{PESQ}$, in other words, the difference between the enhanced and the input signals in terms of the PSNR and PESQ scores. The means with 95% confidence intervals are presented for the different context sizes and genders; the distribution of the objective scores resembles a Gaussian, whereby the confidence interval $=1.96\sigma/\sqrt{N}$, σ being the standard deviation and N being the sample length. Plots 3(a)-(b) depict the absolute values of the PSNR and the PESQ of the input signal and Plots 3(c)-(d) present the Δ PSNR and Δ PESQ values, respectively. We observe that the Δ PSNR increases as the context size increases, for both males and females. At higher bitrates the improvements saturate. The Δ PESQ score has a negative slope with respect to bitrate and outperforms the PESQ score of context $=1$ above bitrate 8.8 kbits/s. In addition, the PESQ score of males is higher than the score of females. The spectrogram of a speech sample enhanced using the above method is shown in Fig. 6 (c).

Covariances in the Time-Frequency Representations

For a fair evaluation of the quality in the MDCT- and STFT-domains with respect to the bitrate, we modified the codec to encode only the magnitude spectrum, such that the number of coefficients in both domains remain equal. From the ΔPSNR and ΔPESQ plots in Fig. 4, we observe that the gains in PSNR and PESQ, obtained by post-filtering in the MDCT domain is approximately half of that in the STFT domain.

Analysis: To study the magnitude of correlations in the two domains, we use a sinusoid with a constant frequency and amplitude, which is divided into frames. While the original signal has a constant amplitude and frequency over all time, its phase varies over individual frames. In the STFT domain, we observe that the magnitude spectrum is reasonably uniform over all the time frames in the STFT-domain, as shown in Fig. 5 (a). In contrast, in the MDCT-domain the

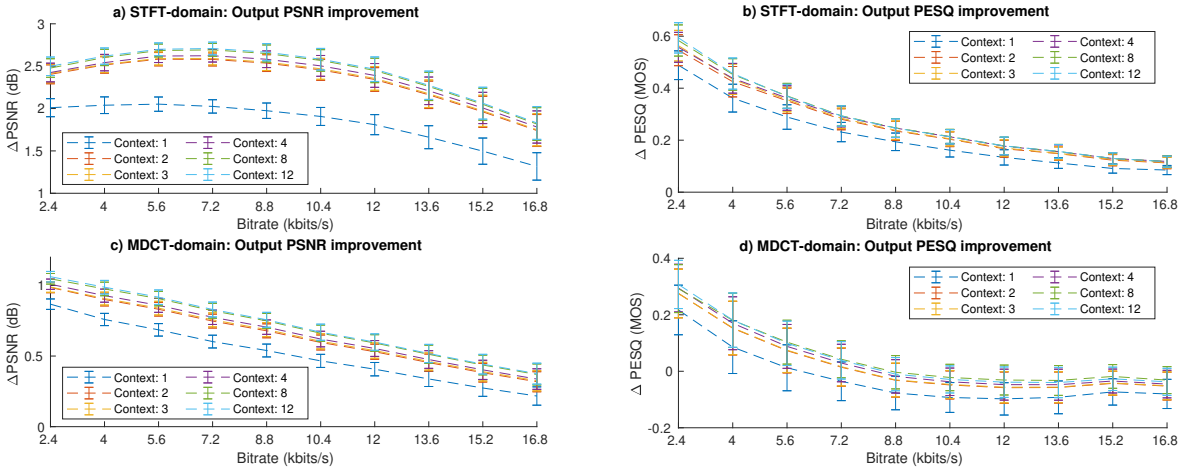

Figure 4 – Δ PSNR and Δ PESQ scores of the output signal over bitrates: (a)-(b) STFT-domain postfilter, (c)-(d) MDCT-domain postfilter.

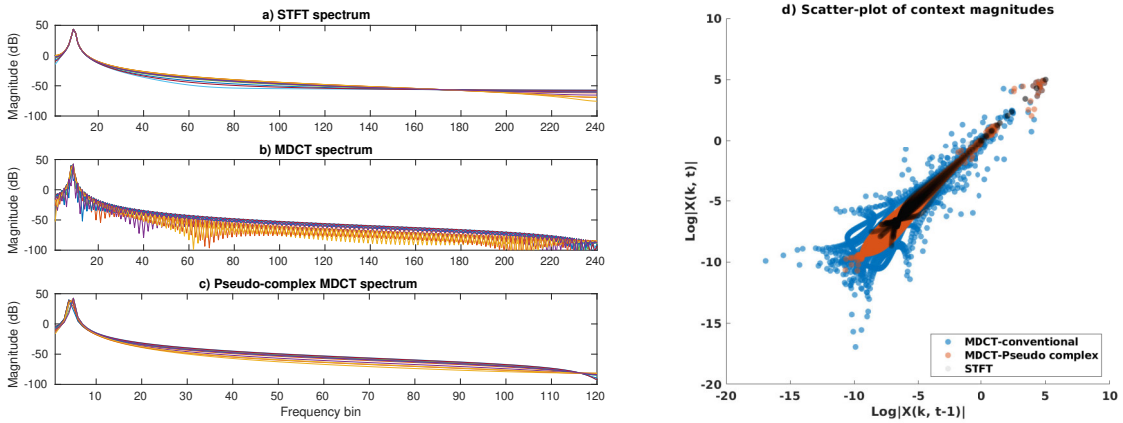

Figure 5 – The a) STFT, b) MDCT and c) Pseudo-Complex MDCT spectra of a sequence of consecutive frames of a single sinusoid, as well as d) the covariance of log-magnitudes over two time frames.

magnitude varies over time, which is indicated by a lower overlap between the MDCT spectra over all time-frames in Fig. 5 (b). This leads to the conclusion that the phase variations in the signal have a direct implication on the magnitude of coefficients in the MDCT-domain. The reason is that the magnitude of MDCT-components depends on the extent with which the basis functions are in phase with the input signal. In the worst case, a frame in the MDCT could be all-zero even for a non-zero input, whereas non-zero frames are always non-zero in the STFT.

To obtain a more stable magnitude and equivalently, improve correlation in the log-MDCT-domain, we propose a *pseudo-complex MDCT* spectrum. We define the pseudo-complex representation of a frequency-bin pair $Y(k,t), Y(k+1,t)$ as $Y_c(k,t) := Y(k,t) + jY(k+1,t)$, where k is the frequency-bin in time-frame t. The magnitude spectrum of the representation is presented in Fig. 5 (c) and clearly indicates improved uniformity and smoothness over time. A scatter-plot representing the variation of the log energy of the pseudo-complex MDCT-, conventional MDCT- and STFT-spectra and their corresponding contexts of size $= 1$ are displayed in Fig. 5 (d). The plot demonstrates that the correlation between the time-frequency coefficients in the pseudo-complex MDCT-domain is higher than in the conventional MDCT-domain, and is similar to the degree of correlation in the STFT-domain. In the next section we therefore implement a postfilter in the pseudo-complex MDCT spectrum to enhance the output quality of the decoded signal in the MDCT-domain.

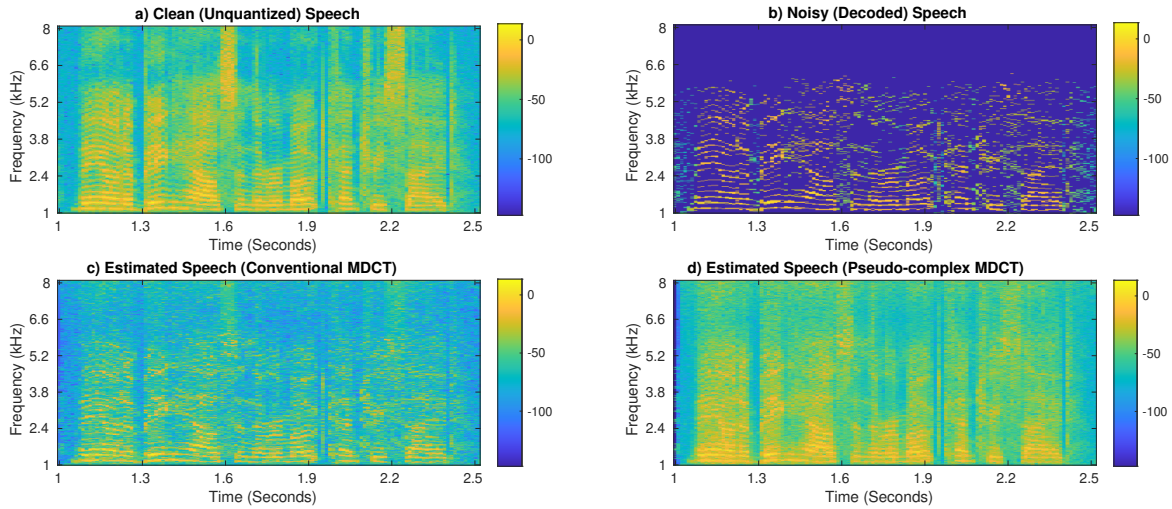

Figure 6 – Spectrograms of the: a) Clean signal, b) Noisy signal, c) Estimated speech using conventional MDCT-domain postfilter, d) Estimated speech using pseudo-complex MDCT postfilter.

Algorithm 2 Estimation of signal from quantized observation (pseudo-complex MDCT)

Require: Quantized signal Y, prior-models C_2, C_3
  function ESTIMATION(Y, C_2, C_3)
    $Y_c = getComplexRepresentation(Y)$
    for $frame = 1 : N$ **do**
      for $k = 1 : Length(Y(frame))$ **do**
        $\mu_{up}, \sigma_{up} \leftarrow EstimateStatistics(C_2, \hat{X}_{c(prev)})$
        $pdf_{c(energy)} \leftarrow GaussianPDF(\mu_{up}, \sigma_{up}, \vec{l}(k), \vec{u}(k))$
        $pdf_{c(joint)} \leftarrow GaussianPDF(C_3(k), \vec{l}(k), \vec{u}(k))$
        $pdf = pdf_{c(energy)} \times pdf_{c(joint)}$
        if $\text{sign}\{\vec{u}(k)\} - sign\{\vec{l}(k)\} == 0$ **then**,
          $\hat{X}_c \leftarrow Expectation(pdf)$
        else
          $\hat{X}_c \leftarrow RandomSampling(pdf)$

Pseudo-Complex MDCT-Domain Modelling

Training: The training process is illustrated in Fig. 7. After transforming the signal to the MDCT domain, the coefficients, $Y_k \in \mathbb{R}$ are converted to the complex spectrum representation, $Y_c = Y_r + jY_i$, where $Y_r = Y(k,t), Y_i = Y(k+1,t)$ and k is the frequency-bin in the time-frame t. The context is modeled in C_2 by following the training process as described in Sec. 3, with the only difference being that the context values are the log of energy of the pseudo-spectrum representation. In addition to the context, in this approach we also compute the joint PDF of the log energies of the real and imaginary parts of the complex spectrum, which we denote as C_3. While C_2 caters to the time-frequency structural relation in speech and is adaptive due to its conditioning on the past estimates, C_3 is static and provides information on the probable speech energy over frequency.

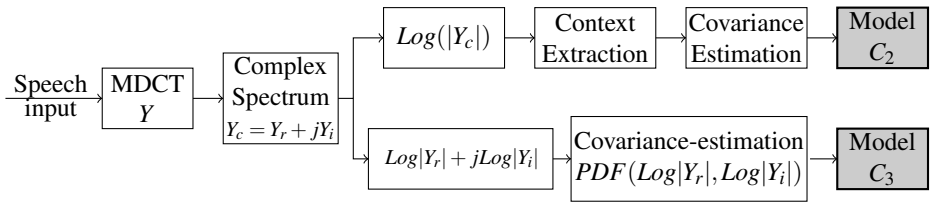

Figure 7 – Modelling process in the pseudo-complex MDCT domain for postfiltering.

Inference: Algorithm 2 presents the inference process. The context model, C_2 and the joint model C_3 along with the decoded noisy signal Y are inputs to the postfilter. After the complex spectrum is obtained from the MDCT coefficients, based on the past estimates, $\hat{X}_c$, and the upper and lower limits obtained from the quantized signal $\vec{l}, \vec{u}$, we obtain the posterior distribution of the sum of energy of the pair of bins, P_{energy}, which is represented as $pdf_{c(energy)}$ in algorithm 2. Note that while the obtained posterior distribution is a one-dimensional truncated distribution describing the distribution of $Y_r^2 + Y_i^2$, we have to estimate two coefficients: Y_r, Y_i from it. We deploy C_3 at this stage to acquire information on the distribution of Y_r and Y_i. The combined models from C_2 and C_3 yield a two-dimensional PDF, which is used to estimate the clean speech in the frequency-bin pair in consideration. Finally, the estimate of the complex spectrum, $\hat{X}_c$ is obtained by computing the joint mean in bins quantized to non-zero values, and random-sampling in the frequency-bins quantized to zero. The spectrogram of the signal enhanced by the postfilter employing this pseudo-complex MDCT approach is shown in Fig. 6 (d). On visually inspecting the spectrograms, we observe that the output of this method 1) looks closest to the clean speech spectrogram in Fig. 6 (a), 2) is less biased towards zero compared to the decoded signal spectrogram in Fig. 6 (b), and the output of the postfilter using conventional MDCT-domain models in Fig. 6 (c).

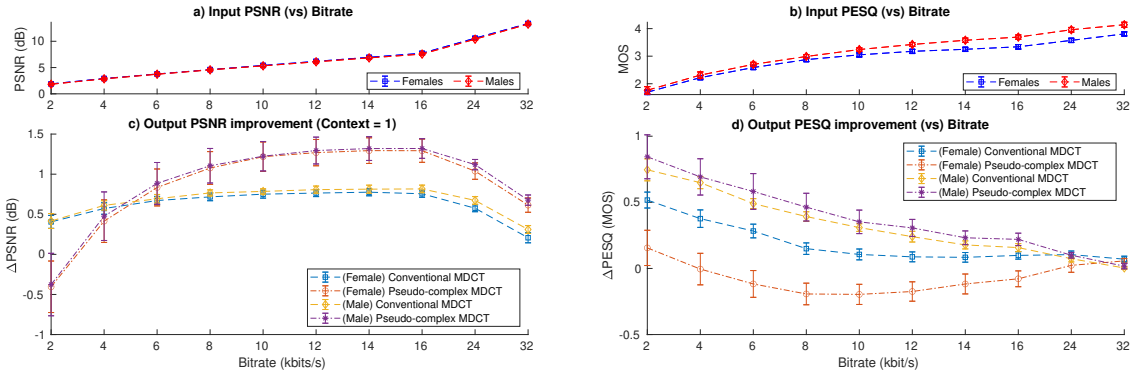

Figure 8 – PSNR and PESQ scores of the input signal with respect to bitrate in (a)-(b). Δ PSNR and Δ PESQ scores of the output signal over bitrates in (c)-(d) for context size = 1.

Evaluation: The description of the training and test files and the objective measures is similar to that presented in the evaluation part in Sec. 3. We perform evaluation on 20 speech samples, 10 males and females each, and for context size $= \{1\}$ since for a context size > 1 the context neighborhood shown in Fig. 2 may not be applicable. The evaluation is performed on bitrates between 2 to 32 kbits/s and the results are presented in terms of the Δ PSNR and Δ PESQ plots in Fig. 8 (c),(d).

Plots in Fig. 8 (a)-(b) show the absolute objective measure values for the input signal over different bitrates; we observe that while the input PSNR at an arbitrary bitrate is identical for males and females, the input PESQ score is lower for females and this trend is also observed in the Δ PSNR and Δ PESQ plots. From the Δ PSNR plot in Fig. 8 (c), we observe that the Δ PSNR of the pseudo-complex MDCT approach is, on an average, 0.8 to 1 dB higher than that of the conventional MDCT-domain postfilter, for both males and females. Also, this magnitude of improvement in the output PSNR is close to the improvement achieved with context = 1 by the STFT-domain postfilter. In terms of the Δ PESQ scores in Fig. 8 (d), the pseudo-complex method for females seems to deteriorate the quality of the input signal with negative Δ PESQ scores between 4 to 16 kbits/s. However, for males not only is the Δ PESQ score for the pseudo-complex MDCT postfilter between 0.1 to 0.9 MOS, it consistently outperforms the Δ PESQ

score of the conventional MDCT approach by 0.1 MOS.

Conclusions

In this paper we presented methods for low-delay postfiltering in the MDCT-domain. First, we adapted the context-based STFT-domain postfilters to the MDCT-domain and the evaluation of the performance showed half of the output gain that was achieved in the STFT-domain. We observed that the reason is that the magnitudes of real-valued spectra are less uniform over time than those of complex-valued spectra and proposed a pseudo-complex model to retain quality. We observed that the gain in perceptual-SNR obtained by post-filtering in the pseudo-complex MDCT domain almost matches the gain obtained in the STFT-domain. Analysis of the perceptual effect of the proposed method is left for future work.

References

[1] *EVS codec detailed algorithmic description; 3GPP technical specification*. `http://www.3gpp.org/DynaReport/26445.htm`, 2014.

[2] BÄCKSTRÖM, T.: *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017.

[3] GRANCHAROV, V., J. H. PLASBERG, J. SAMUELSSON, and W. B. KLEIJN: *Generalized postfilter for speech quality enhancement*. *IEEE transactions on audio, speech, and language processing*, 16(1), pp. 57–64, 2007.

[4] DISCH, S., 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] DAS, S. and T. BÄCKSTRÖM: *Postfiltering with complex spectral correlations for speech and audio coding.* In *Interspeech*. 2018.

[6] DAS, S. and T. BÄCKSTRÖM: *Postfiltering using log-magnitude spectrum for speech and audio coding*. In *Interspeech*. 2018.

[7] FUCHS, G., V. SUBBARAMAN, and M. MULTRUS: *Efficient context adaptive entropy coding for real-time applications*. In *ICASSP*, pp. 493–496. IEEE, 2011.

[8] ZUE, V., S. SENEFF, and J. GLASS: *Speech database development at MIT: TIMIT and beyond*. *Speech Communication*, 9(4), pp. 351–356, 1990.