Frequency-Domain Prediction for Audio Coding

Christian R. Helmrich, Richard Füg, *Member, IEEE*, and Bernd Edler

***Abstract*—Minimization of temporal redundancy in perceptual and lossless audio transform coding using backward- or forward-adaptive linear prediction has been a subject of intensive scientific investigation. The computational complexity of most conventional approaches, however, is considerable. In this paper we present a forward-adaptive predictor which is guided by a single eight-bit periodicity parameter per frame and channel and which, owing to its direct integration into the modified discrete cosine transform (MDCT) domain of the codec, operates with very low complexity. It is shown that our frequency-domain prediction (FDP) method may also be used to improve the performance of state-of-the-art joint-channel coding schemes. The results of a blind listening test indicate that, in the context of the MPEG-H 3D Audio codec, the FDP technique leads to small but significant coding quality gains.**

***Index Terms*—Audio coding, linear prediction, MDCT, MDST**

I. Introduction

CONVENTIONAL perceptual as well as lossless audio codecs (coders/decoders) divide the incoming time-domain (TD) waveform into successive frames which are transformed, quantized, and coded separately and largely independently. For quasi-stationary input signals, however, there naturally remains some temporal redundancy in the transform samples between adjacent frames for a given channel or even between channels. This was found to be especially true for recordings of sustained notes played by isolated acoustic or electronic instruments.

To minimize such residual redundancy, two approaches have been pursued in the past. The first, proposed by Mahieux *et al.* for a discrete Fourier transform (DFT) based system [1], is to apply linear predictive coding (LPC) across time to individual transform coefficients. This method was later adapted for real-valued modified discrete cosine transform (MDCT) coding and extended to include joint-stereo coding [2], [3], [4], [5], [6].

The second approach is to account for temporal redundancy during the entropy coding stage, i. e., after quantization in the transform or frequency domain (FD). Most recently, this was addressed in the Unified Speech and Audio coding standard [7] using arithmetic coding with intra-/inter-frame signal-adaptive probability contexts for each quantized MDCT sample [8].

The latter technique can be implemented with relatively low algorithmic complexity since the quantized MDCT values are readily available at both the encoder (transmitter) and decoder (receiver) side. The LPC based schemes, in turn, typically lead to higher coding (and quality) gains but require much greater encoder-side complexity: given that they operate on the initial *uncoded* MDCT coefficients using previously *decoded* values, a full FD decoding path inside the encoder becomes necessary. This is particularly the case for TD long-term prediction (LTP) [9], [10], which may even require additional MDCT operations in order to prepare the prediction signal. For such a predictor, a reduction to $22 \cdot 672 = 14784$ algorithmic operations per frame and channel was reported [9], which represents a considerable improvement over previous work [3], [11]. For a typical stereo signal sampled at a rate of 48 kHz and coded at a frame length of 1024 samples, however, this still leads to an undesirable 1.4 million operations per second (MOPS) of added complexity.

This manuscript was rejected by the IEEE due to incomplete documentation of the conducted subjective evaluation. The desired information about the test procedure and the audio material used, which was omitted because of the page limit, can be found in the dissertation of the first author at www.ecodis.de.

The authors are with the International Audio Laboratories Erlangen, a joint institution of Fraunhofer IIS and Friedrich-Alexander University of Erlangen-Nürnberg, Germany (e-mail: christian.helmrich@audiolabs-erlangen.de).

A. Contribution and Organization of this Letter

In order to realize very-low-complexity prediction for audio coding, the following structural constraints should be enforced.

- The predictor's computation and application should reside as closely around the codec's quantizer as possible. This reduces the amount of encoder-side coefficient decoding.
- The prediction should be bandlimited to further minimize the workload at both encoder and decoder. Hence, a good trade-off between coding gain and range must be chosen.
- Only those signal components exhibiting temporal correlation, e. g., the individual harmonics of a tonal waveform, should be subjected to prediction. This not only minimizes the complexity but also increases the prediction gain.

The remainder of this letter presents a novel frequency-domain predictor (FDP) supporting both perceptual and lossless coding (although the former aspect will be emphasized herein) which, as will be described in Section II, adheres to all of the above three constraints. Extensions of the FDP design for low-bitrate perceptual coding of monophonic and stereophonic content are discussed in Sections III and IV, respectively. Section V reports on the preparation and results of objective and subjective tests of the basic FDP scheme, and Section VI concludes the paper.

II. Low-Complexity Frequency-Domain Prediction

The fundamental architecture of a two-channel audio codec employing temporal prediction as a pre- and post-processor is shown in Figure 1. The depicted block diagrams, in which the

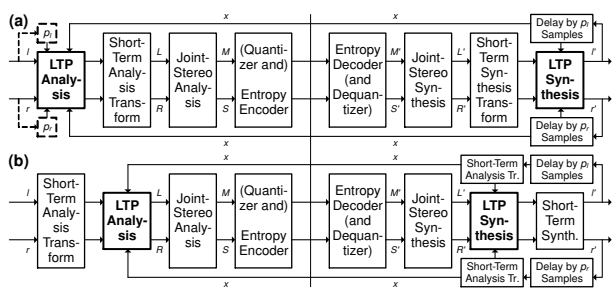

Fig. 1. Long-term predictive coding in the (a) time and (b) transform domain.

lower-case and upper-case letters indicate TD and transformed signals, respectively, apply to both perceptual (e. g. MDCT for transformation) and lossless (e. g. LPC or integer MDCT [12] for transformation, no (de)quantization) coding. The predictor memory x_p, i. e., the left or right source signal subtracted (after scaling) from the respective target signal during LTP analysis and added again (after equivalent scaling) upon LTP synthesis, equals a delayed portion of the previously *decoded* l' or r':

$$x_p(n) = \begin{cases} l'(n-p_l), & p_l \geq N, \quad \text{for left channel,} \\ r'(n-p_r), & p_r \geq N, \quad \text{for right channel,} \end{cases} \tag{1}$$

where $p_{\{l,r\}}$ denotes the per-frame/channel pitch or periodicity lag, $0 \leq n < 2N$, and N is the frame length in samples. Note that, in the encoder (left side of thin vertical line) of a "lossy" perceptual codec, the *uncoded* inputs l and r may also be used (dashed paths) instead of l' and r', yielding an open-loop pitch pre-/post-filter for quantization noise shaping [14] instead of a conventional closed-loop LTP. For the sake of brevity, though, this solution will not be examined in the present publication.

Two issues can be observed here. First, as seen in Fig. 1(b) and noted in Sec. I, transform-domain LTP application for best selectivity [9], [13] necessitates an extra analysis transform for each channel since the (re)construction of l' and r' involves an inevitable overlap-and-add (OLA) process between temporally adjacent synthesis transform results. This is especially true for MDCT-based coding, where N samples of inter-frame overlap are typically used, and is also the reason why the minimum lag in (1) must be larger for MDCT than for LPC coding [10].

The expensive OLA-related demand for auxiliary transforms can be circumvented by moving the calculation and application of x into the MDCT domain [3], [11], as depicted in Fig. 2(a):

$$X_P(k) = \begin{cases} L'_{m-P_L}(k), & P_L \geq 1, \quad \text{for left channel,} \\ R'_{m-P_R}(k), & P_R \geq 1, \quad \text{for right channel,} \end{cases} \tag{2}$$

where $P_{\{L,R\}}$ is the channel-wise transform-domain (subband) delay for the current frame at index m, restricted to an integer value, and $0 \leq k < N$. Given the symmetric MDCT-compliant window $w(n)$ applied in the current frame [12], it follows that $L' = \text{MDCT}(l'w)$, $R' = \text{MDCT}(r'w)$, i. e., $X \equiv \text{MDCT}(xw)$. This approach, however, still exhibits a second issue: for every algorithmic tool, a corresponding decoder is also needed inside the encoder, thereby increasing the complexity of the latter.

In order to minimize such additional encoder-side workload, the LTP may be implemented as closely around the FD entropy coder (and quantizer in case of perceptual codecs) as possible. Fig. 2(b) illustrates that, at most, only the dequantizer and the obligatory LTP decoder must then be added to the encoder. In the example of Figs. 1 and 2, this means that the two predictors now operate in the joint-stereo (e. g., mid-side, M-S) domain.

The complexity can be further reduced by limiting the LTP to a bandwidth lower than the 15 kHz utilized in [3], [9], [11]. The tonality and/or harmonicity of most natural sound sources as well as the ability for phase locking in human hearing [15] vanish around 4–5 kHz, which coincides with the region of the highest musical note, C8 $\approx$ 4.19 kHz on a piano or piccolo. A prediction range from 70 Hz (to avoid DC offset or AC hum) to 4.2 kHz should, therefore, suffice. In fact, even a bandwidth of 3.1 kHz does not appear to notably restrict the performance.

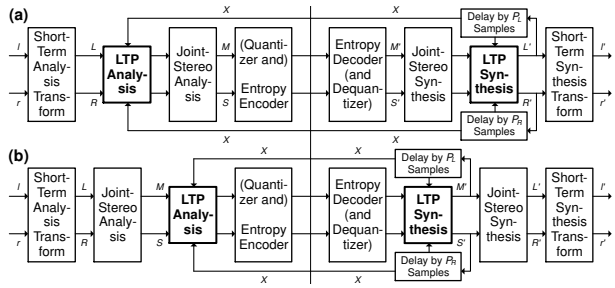

Fig. 2. FD prediction without TD components: (a) conventional, (b) proposed.

A. Frequency-Domain Prediction on a Harmonic Grid

In order to address the remaining third bullet point in Sec. I.A, we follow the forward-adaptive scheme of [9], where LTP parameters and activation flags are transmitted for every frame and channel as part of the coded bit-stream (instead of deriving them at both the encoder and decoder, as in backward-adaptive designs), and apply a spectral line-wise second-order predictor

$$\hat{Y}_s(k) = \dot{X}(k)\, t_1 + \ddot{X}(k)\, t_2, \quad 0 \leq k < K, \tag{3}$$

as in [11], with k as an indicator of the MDCT line index and, using (2), the abbreviations $\dot{X} = X_1$ (last frame) and $\ddot{X} = X_2$ (second-to-last frame). This linear combination of two consecutive instances of X_P, with weights t_1 and t_2, allows for high temporal resolution in the prediction since pitch lags including fractions of the frame length can be used. However, the costly spectral band-wise activation/signaling as well as the line-wise prediction for any k below some limit $K < N$ is undesirable.

To determine t_1 and t_2 in (3), consider the MDCT spectrum

$$Y(k) = \sum_{n=0}^{2N-1} w(n)\, y(n) \cos\Big(\omega_k (n + \tfrac{N+1}{2})\Big), \quad 0 \leq k < N,$$

where the commonly employed normalization factor $\sqrt{2/N}$ is included in w and the modulation frequency is $\omega_k = \frac{\pi}{N}(k + \frac{1}{2})$ [12]. Assume, further, that the input waveform y is a harmonic signal composed of multiple sinusoids at frequencies $\omega_s = \frac{\pi}{N} s$:

$$y(n) = \sum_{h=1}^{\lfloor N/s_0 \rfloor} y_s(n) = \sum_{h=1}^{\lfloor N/s_0 \rfloor} \cos\Big(\omega_s (n + mN) + \varphi_s\Big), \tag{4}$$

where φ_s is an arbitrary phase offset and where each harmonic with index $h > 1$ lies at an integer multiple of the fundamental frequency $s_0 = \frac{s}{h}$, for which $h = 1$. This s_0 is a FD equivalent of the LTP pitch lag $p_{\{l,r\}}$ in (1) and, conveniently, indicates the spectral spacing between the individual harmonics in units of line indices k. As long as the minimum value for s_0 exceeds the main-lobe width exhibited by w for sufficient FD harmonic separation, which is the case for $s_0 \geq 3$ (70.3 Hz for $N = 1024$ and 48 kHz sampling rate), an efficient line-selective FDP can be realized. Specifically, for the above example, only the lines at $k < K$ whose ω_k are close enough to the nearest ω_s, e. g.,

$$|\, \omega_k - \omega_s| < \tfrac{3\pi}{2N}, \tag{5}$$

are to be subjected to the FDP of (3), with the appropriate s of (4). Note that the coefficients $t_{\{1,2\}}$ only need to be computed once for each ω_s (not for each ω_k), as demonstrated hereafter. For all k with (5) and assuming adequate side-lobe attenuation

due to w at $|\omega_k - \omega_s| \geq \frac{3\pi}{2N}$ as well as the ideal, distortion-free case $Y(k) = X_0(k)$, the predictor coefficients for each ω_s can be obtained from a system of equations making use of the following dependencies, which take into account the hop size N:

$$X_0(k) = \mathrm{MDCT}\big(w(n)\, y_s(n)\big)(k) = A_k \cos\big(mN\omega_s + \varphi_k\big)$$

with $\varphi_k = \varphi_s - \omega_s(N - \frac{1}{2}) + \omega_k(\frac{5N}{2})$ and A_k depending on w and the difference $\omega_s - \omega_k$. From this $X_{\{1,2\}}(k)$ can be derived:

$$X_1(k) = X_0(k)\cos(N\omega_s) + A_k \sin(mN\omega_s + \varphi_k)\sin(N\omega_s),$$
$$X_2(k) = X_0(k)\cos(2N\omega_s) + A_k \sin(mN\omega_s + \varphi_k)\sin(2N\omega_s).$$

Solving the FDP condition $\hat{Y}_s(k) \stackrel{!}{=} Y(k)$ leads to the equations

$$t_1 \cos(a_s) + t_2 \cos(2a_s) = 1,\ t_1 \sin(a_s) + t_2 \sin(2a_s) = 0 \quad (6)$$

with the abbreviation $a_s = N\omega_s$ and, thus, to the final solution

$$t_1 = 2\, g_{\mathrm{opt}} \cos(N\omega_s),\ t_2 = -(g_{\mathrm{opt}})^2, \quad (7)$$

where, luckily, both t_1 and t_2 are independent of A_k and φ_k.

The factor $0 \leq g_{\mathrm{opt}} \leq 1$ in (7) denotes the optimal predictor gain, which shall not be confused with the predict*ion* gain G, i. e., the ratio of the MDCT variances before and after the FDP,

$$G(s_0) = 10 \log_{10}\Big(\sum_{k=0}^{K-1} \big(Y(k)\big)^2 \Big/ \sum_{k=0}^{K-1} \big(Y(k) - \hat{Y}_s(k)\big)^2 \Big), \quad (8)$$

in units of dB. s_0 and $g = g_{\mathrm{opt}}$ can be chosen such that $G(s_0)$ is maximized, which coincides with a minimum prediction error (denominator of the above fraction) in a least-squares sense.

It is worth noting that, for natural tonal signals, a fixed gain of $g \approx 0.9$ decreases the FDP analysis workload and parameter rate because g_{opt} does not need to be calculated or transmitted. Nevertheless, it works almost as well as g_{opt}, especially when g depends on ω_s. For instance, solving the constraint definition

$$\begin{aligned} &\big(1 - t_1 \cdot \cos(N\omega_s) - t_2 \cdot \cos(2N\omega_s)\big)^2 \\ &\quad + \big(t_1 \cdot \sin(N\omega_s) + t_2 \cdot \sin(2N\omega_s)\big)^2 \stackrel{!}{=} \tfrac{1}{256} \end{aligned} \quad (9)$$

for $g(\omega_s)$ as g_{opt} in t_1 and t_2 yields an expression which can be approximated quite closely by the even-exponent polynomial

$$g(\omega_s) \approx \tfrac{1983}{2048} - \tfrac{87}{2048}\cos(N\omega_s)^2 - \tfrac{360}{2048}\cos(N\omega_s)^6 \quad (10)$$

and which, in $\hat{Y}_s$, results in a constant *application* of g, with a gain near 24.1 dB at each instance of ω_s, instead of a constant *value* of g but a varying (and often low) gain at said instances. The values of $0.75 \leq g(\omega_s) \leq 0.9683$, like those of $\cos(N\omega_s)$, are known a priori and can be tabulated for low complexity.

Figure 3 depicts the frame-wise value of $G(s_0)$ (solid line) in a single-channel perceptual codec operating at 0.5–1 bit per TD sample (variable bit-rate), 48 kHz sampling rate, $N = 1024$ samples, and $K = 132$ (for a prediction bandwidth of 3.1 kHz), with $g(\omega_s)$ as in (10). For this evaluation, s_0 was quantized to an 8-bit index on the following ERB-like [15] nonlinear scale:

$$s_0 = \frac{298 \cdot 3}{298 - i_{\mathrm{opt}}}, \quad 0 \leq i_{\mathrm{opt}} < 2^8, \quad (11)$$

with i_{opt} representing the index which would be transmitted to the FDP decoder. Despite such quantization, it is evident from Fig. 3 that gains of 7 dB or more are achieved on the depicted exemplary input, namely, throughout the three-tone pitch pipe signal and the first seven tones of the harpsichord arpeggio.

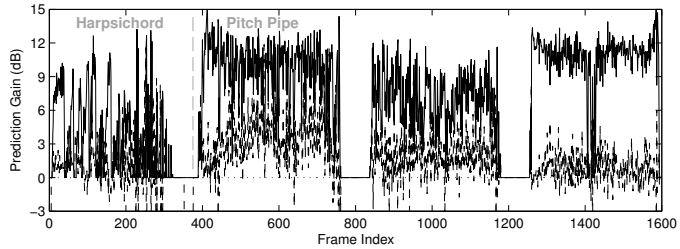

Fig. 3. Relative gain of (—) ω_s-based FDP of Sec. II vs. coding without LTP, (- -) low-rate enhanced FDP of Sec. III vs. basic FDP of Sec. II, on tonal input.

III. Enhanced FDP for Low-Rate "Lossy" Coding

The spectral quantization in a perceptual encoder causes an inevitable loss of information, often modeled as an added noise signal, which also reaches (1) and (2). For $|\cos(N\omega_s)|$ in (7) approaching 1 and a typical gain of $g \approx 0.9$, the FDP design of Sec. II amplifies the noise variance of the predictor memory by a factor of 4, which reduces the achievable prediction gain.

For the two center coefficients $Y(k)$ near each harmonic at ω_s, the predicted MDCT value $\hat{Y}_s(k)$ can, alternatively to (3), be derived from the last frame's MDCT neighbors $\dot{X}(k, k\pm1)$:

$$\hat{Y}_s(k) = \begin{cases} \dot{X}(k)\, f_1^+ + \dot{X}(k+1)\, f_2^+, & \omega_k < \omega_s, \\ \dot{X}(k)\, f_1^- + \dot{X}(k-1)\, f_2^-, & \omega_k \geq \omega_s. \end{cases} \quad (12)$$

Given the real-valued $W_0(\omega) = \mathcal{F}(w(n)) \cdot \mathrm{e}^{j\omega(N - \frac{1}{2})}$ for the current frame and channel and solving the prediction condition $\hat{Y}_s(k) \stackrel{!}{=} Y(k)$ for f_1^+, f_2^+ and f_1^-, f_2^-, respectively, leads to

$$\hat{Y}_s(k) = \dot{X}(k)\, g \cos(N\omega_s) + \dot{X}(k+1)\, g \sin(N\omega_s)\, F_k^+ \quad (13)$$

with $F_k^+ = W_0(\omega_s - \omega_k)/W_0(\omega_s - \omega_{k+1})$ for $\omega_k < \omega_s$ (taking upper neighbor) and, accordingly for $\omega_k \geq \omega_s$ (lower neighbor),

$$\hat{Y}_s(k) = \dot{X}(k)\, g \cos(N\omega_s) - \dot{X}(k-1)\, g \sin(N\omega_s)\, F_k^- \quad (14)$$

with $F_k^- = W_0(\omega_s - \omega_k)/W_0(\omega_s - \omega_{k-1})$. Utilizing (13), (14) instead of (3) whenever $(f_1^\pm)^2 + (f_2^\pm)^2 < (t_1)^2 + (t_2)^2$ reduces the maximum noise variance amplification to a factor of about 2 for $g \approx 0.9$. In practice, however, this algorithmic extension only results in a best-case—but still barely audible—prediction gain increase of about 4 dB (on the first pitch pipe tone, dashed line in Fig. 3) even for coding rates as low as 0.5 bit per TD sample and channel. It also comes at the cost of computational and implementational complexity overhead (more comparison operators and memory usage, frame-wise dependency on W_0).

IV. Extended FDP for Joint-Channel Prediction

Analogously to inter-channel extensions of traditional intra-channel predictors [2], [3], the periodicity-based line-selective FDP principle can also be applied to improve a state-of-the-art joint-channel coding scheme such as complex stereo prediction [6]. More precisely, the approximation of the modified discrete sine transform (MDST) of downmix D—the imaginary part of the predictor—from the current and last frame's MDCT downmixes as in [6], [7] can be replaced with a variant of the FDP extension described in Sec. III for the spectral lines residing on the determined (joint-stereo) harmonic grid. Utilizing the past MDCT downmix $\dot{D}_R$, the current MDST downmix coefficients

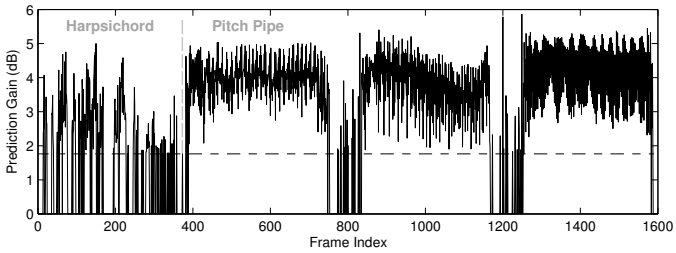

Fig. 4. Relative improvement in MDST approximation error variance for two-frame complex prediction with vs. without joint-stereo FDP for tonal signals.

$D_I(k)$ with frequencies ω_k closest to ω_s can be obtained by

$$D_I(k) = \dot{D}_R(k)\sin(N\omega_s) - \dot{D}_R(k+1)\cos(N\omega_s)\,F_k^+ \quad (15)$$

for $\omega_k < \omega_s$ (upper) and, respectively for $\omega_k \geq \omega_s$ (lower line),

$$D_I(k) = \dot{D}_R(k)\sin(N\omega_s) + \dot{D}_R(k-1)\cos(N\omega_s)\,F_k^-. \quad (16)$$

Note the similarity to (13), (14), aside from g and the reversed sine/cosine terms. All other (non-harmonic) coefficients of D_I can be computed according to [6], [7]. Figure 4 visualizes the benefit of (15) and (16), when applied for the two lines at each $\omega_s < 3.1$ kHz, on the MDST approximation accuracy for that frequency range in case of two harmonic signals. Although the gain in accuracy consistently exceeds the threshold of 1.76 dB (dashed line), this has to be regarded as a best-case result. For no other natural tonal input known to the authors, the threshold was notably exceeded, indicating a lack of practical advantage.

V. Objective and Subjective Evaluation

To assess the performance of the fundamental FDP method described in Sec. II, both in terms of computational complexity and algorithmic benefit (now with respect to perceptual coding scenarios), a version largely realized in integer arithmetic was integrated into Fraunhofer's implementation of the MPEG-H 3D Audio core-codec [16] in the course of the standardization toward a "phase 2" amendment [17]. This codec is an evolution of [7] with [6], [8] and $N = 1024$ samples. For each frame and channel a 1-bit FDP indicator, followed by the 8-bit index i_{opt} of (11) upon FDP activation, was included in the bit-stream for a worst-case parameter rate increase of only 422 bit per second and channel at a codec sampling rate of 48 kHz. Transmission of the optimal gain g_{opt} alongside the periodicity index i_{opt} was omitted since it did not improve the performance significantly.

A. Analysis of Algorithmic Complexity

Due to a mostly fixed-point realization, The FDP processing and update loops (across MDCT lines) in both the encoder and decoder can be implemented via inexpensive load, add, binary-shift, binary-and, multiply(-accumulate), absolute-value, store, and compare operations. At 48 kHz sampling rate, their worst-case combined complexity (for $i_{\text{opt}}=0$ and predictor coefficient updates every three MDCT lines) equals 0.54 MOPS for stereo audio. This is only 39 % of the value obtained in [9], as noted in Sec. I, and, according to measurements on a Nexus 7 tablet [18], represents just 2 % of the total core-decoder workload. It is also worth repeating that, by virtue of its design, the FDP is much cheaper at the encoder side than the LTPs in [9], [10]; it increases the encoder complexity by only 11% on an x86 PC.

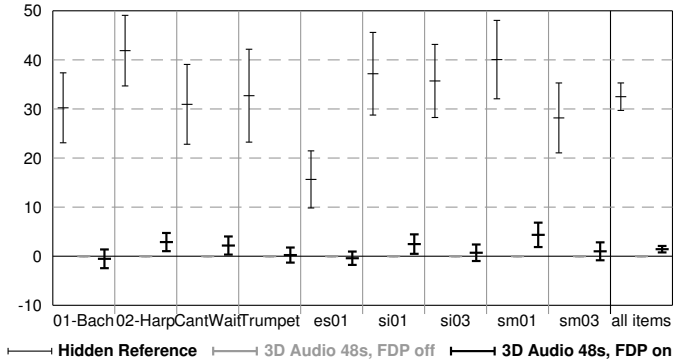

Fig. 5. Zoomed view of listening test results showing mean differential scores.

B. Perceptual Assessment at 48 kbps Stereo

For subjective evaluation, a blind listening test following the MUSHRA (*mu*ltiple *s*timuli with *h*idden *r*eference and *a*nchor) principle [19] was conducted. Both 3D Audio variants tested—with and without FDP coding—were operated at a total bit-rate of 48 kilobit per second (kbps, including the added per-channel periodicity indices in case of FDP coding) and a sampling rate of 48 kHz. FDP processing was enabled whenever its gain, as given by (8), exceeded 1 dB. The remaining codec parameters were configured as described in [20]. 17 experienced listeners, aged 38 or younger, performed the experiment in a quiet room using a silent computer and STAX SR Lambda headphones.

Figure 5 illustrates the results of the listening test as overall and per-stimulus mean differential scores (relative to the codec condition without FDP processing) along with their associated 95 % confidence intervals. For four of the nine signals assessed (tonal vocal or instrumental recordings already utilized in past evaluations [21], [22]) as well as the overall score, the entire confidence interval of the condition with FDP lies above zero, indicating a statistically significant quality increase due to the FDP. Interestingly, the score for the a-cappella solo *CantWait* improves significantly (the FDP can track varying fundamental frequencies sufficiently well), while no significant quality gain can be observed for the highly stationary and harmonic pitch pipe *si03* (most likely since the context-adaptive entropy coder [8] already eliminates most of the spectrotemporal redundancy and, thus, operates less efficiently on "noisy" FDP residuals).

VI. Conclusion

We presented a novel frequency-domain predictor design for both perceptual and lossless coding of mono- and stereophonic audio, unifying the benefit of low side-information rate known from LTP approaches (a single pitch/gain parameter per frame and channel) and the advantage of very low complexity (only low- and mid-frequency harmonic components are subjected to prediction) into one algorithmic solution. Objective assessment of the exhibited prediction gain and complexity as well as subjective performance evaluation in the perceptual MPEG-H 3D Audio core-codec demonstrates the usefulness of our proposal, especially for low-delay applications [14], [22] where a longer transform such as the MELT for tonal input [23] is undesirable.

Acknowledgment

The authors thank the participants of the blind listening test.

REFERENCES

[1] Y. Mahieux, J. P. Petit, and A. Charbonnier, "Transform coding of audio signals using correlation between successive transform blocks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Glasgow, UK, vol. 3, pp. 2021–2024, May 1989.

[2] H. Fuchs, "Improving joint stereo audio coding by adaptive inter-channel prediction," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, USA, pp. 39–42, Oct. 1993.

[3] H. Fuchs, "Improving MPEG audio coding by backward adaptive linear stereo prediction," in *Proc. AES 99th Conv.*, New York, NY, USA, preprint 4086, Oct. 1995.

[4] T. Liebchen, "Lossless audio coding using adaptive multichannel prediction," in *Proc. AES 113th Conv.*, Los Angeles, USA, preprint 5680, Oct. 2002.

[5] H. Krüger and P. Vary, "A new approach for low-delay joint-stereo coding," in *Proc. ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, Oct. 2008.

[6] C. R. Helmrich, P. Carlsson, S. Disch, B. Edler, J. Hilpert, M. Neusinger, H. Purnhagen, N. Rettelbach, J. Robilliard, and L. Villemoes, "Efficient transform coding of two-channel audio signals by means of complex-valued stereo prediction," in *Proc. IEEE ICASSP*, Prague, Czech Republic, pp. 497–500, May 2011.

[7] ISO/IEC, Int. Standard IS 23003-3, "Information technology – MPEG audio technologies – Part 3: Unified speech and audio coding," Geneva, Switzerland, 2012.

[8] G. Fuchs, V. Subbaraman, and M. Multrus, "Efficient context adaptive entropy coding for real-time applications," in *Proc. IEEE ICASSP*, Prague, Czech Republic, pp. 493–496, May 2011.

[9] J. Ojanperä, M. Väänänen, and L. Yin, "Long term predictor for transform domain perceptual audio coding," in *Proc. AES 107th Conv.*, New York, NY, USA, preprint 5036, Sep. 1999.

[10] J. Song, C.-H. Lee, H.-O Oh, and H.-G. Kang, "Harmonic Enhancement in Low Bitrate Audio Coding Using an Efficient Long-Term Predictor," *EURASIP J. Adv. in Sig. Process.*, vol. 2010, ID 939542, Aug. 2010.

[11] L. Yin, M. Suonio, and M. Väänänen, "A new backward predictor for MPEG audio coding," in *Proc. AES 103rd Conv.*, New York, NY, USA, preprint 4521, Sep. 1997.

[12] R. Geiger, J. Herre, J. Koller, and K. Brandenburg, "IntMDCT – A link between perceptual and lossless audio coding," in *Proc. IEEE ICASSP*, Orlando, USA, vol. 2, pp. 1813–1816, May 2002.

[13] K. Kjörling, J. Rödén, M. Wolters, J. Riedmiller, A. Biswas, P. Ekstrand, A. Gröschel, P. Hedelin, T. Hirvonen, H. Hörich, J. Klejsa, J. Koppens, K. Krauss, H-M. Lehtonen, *et al.*, "AC-4 – The next generation audio codec," in *Proc. AES 140th Conv.*, Paris, France, preprint 9491, Jun. 2016.

[14] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the Opus codec," in *Proc. AES 135th Conv.*, New York, NY, USA, preprint 8942, Oct. 2013.

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed., London, UK: Academic Press, 2003.

[16] ISO/IEC, Int. Standard IS 23008-3, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," Geneva, Switzerland, 2015.

[17] ISO/IEC, SC29/WG11, N15849, "Text of ISO/IEC 23008-3:201x/DAM 3, MPEG-H 3D Audio Phase 2," Nov. 2015.

[18] ISO/IEC, M37167, "Proposal for profiles and levels for 3DAudio," 2015.

[19] ITU-R, "Recommendation BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," Oct. 2015.

[20] C. R. Helmrich, A. Niedermeier, S. Bayer, and B. Edler, "Low-complexity semi-parametric joint-stereo audio transform coding, in *Proc. EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, Nice, France, Sep. 2015.

[21] K. Brandenburg and M. Bosi, "Overview of MPEG Audio: Current and future standards for low bit-rate audio coding," in *Proc. AES 99th Conv.*, New York, NY, USA, preprint 4130, Oct. 1995.

[22] C. R. Helmrich and M. Fischer, "Low-delay transform coding using the MPEG-H 3D Audio codec," in *Proc. AES 139th Conv.*, New York, NY, USA, preprint 9355, Oct. 2015.

[23] C. R. Helmrich and B. Edler, "Audio Coding Using Overlap and Kernel Adaptation," *IEEE Signal Process. Letters*, vol. 23, no. 5, pp. 590–594, May 2016.