Interspeech 2018
2-6 September 2018, Hyderabad

Temporal Noise Shaping with Companding

Arijit Biswas[1], *Per Hedelin*[2], *Lars Villemoes*[2], *and Vinay Melkote*[*]

[1]Dolby Germany GmbH, Nürnberg, Germany
[2]Dolby Sweden AB, Stockholm, Sweden
{arijit.biswas, per.hedelin, lars.villemoes}@dolby.com

Abstract

Audio codecs are typically transform-domain based and efficiently code stationary audio signals but they struggle with speech and signals with dense transients such as applause. The temporal noise shaping (TNS) tool standardized in HE-AAC alleviates the issue of noise unmasking in these troublesome cases via signal-adaptive filtering of the transform domain quantization noise, albeit at the cost of significant additional side information in the bitstream. We present a novel alternative referred to as companding that involves QMF domain pre- and post-processing around the core transform-domain coding system: prior to transform encoding, the dynamic range of the signal is reduced locally within a QMF time slot and restored again post decoding, which naturally shapes the coding noise temporally. A primary advantage is that the companding function is fixed and hence enables signal-adaptive noise shaping with just 1-2 bits of side-information per frame. Subjective tests illustrate that the proposed tool improves the quality of hard-to-code applause excerpts compared to TNS while achieving comparable performance on speech signals. The coding tool described in this paper is part of the Dolby AC-4 audio coding system standardized by ETSI and included in ATSC 3.0.

Index Terms: audio coding, pre-echo distortion, MDCT, QMF, transform coding, TNS

1. Introduction

Popular audio coding solutions are lossy [1], i.e. they reproduce audio only with reduced fidelity. Coding artifacts are perceived due to the introduction of quantization noise. Typically, these audio coding systems are based on transform coding in the MDCT-domain. Within a frame of MDCT, audio codecs normally shape the coding noise in the frequency domain to make it least audible: exploiting the so-called simultaneous masking [1], [2] phenomenon. Several successful audio codecs utilize frames of long durations to maximize the coding gain for stationary signals. Since the quantization noise introduced into the signal during the encoding process spreads uniformly throughout the transform block during decoding, coding noise may be most audible during low-intensity passages within an MDCT frame. For example, quantization noise in the quiet region prior to a transient event in a decoded audio signal can be perceived as pre-echoes [1], [3] at the attack portions of transient signals (e.g. castanets) or as reverberation in speech signals [3].

There are two very successful control strategies for reducing pre-echo artifacts [1]. These are MDCT window switching [4] (used in MP3 [5] and AAC [6]) and Temporal Noise Shaping (TNS) [7] (used in the AAC family of codecs [8], [9]). These methods utilize temporal aspects of masking, i.e. noise is masked a short time prior to and after the presentation of a masker signal (pre- and post-masking) [2]. Post-masking is observed for a much longer time period than pre-masking (about 10-50 ms instead of 0.5-2 ms, depending on level and duration of the masker). The adaptive MDCT window switching adjusts the length of the MDCT windows to the characteristics of the input signal. While stationary signal regions are coded using a long window, short windows are used to code the transient parts of the signal. However, speech consists of a series of pseudo-stationary events where window switching does not offer an effective solution. Typically, coding speech with audio codecs introduces time smearing and reverberation artifacts. However, it was shown that TNS improves the subjective quality for speech signals [6], [7]. TNS achieves a temporal shaping of the quantization noise through an open-loop predictive coding along the frequency direction in the MDCT-domain. Another method to reduce pre-echo is to apply a time-domain gain modification [10] to the signal prior to and after transform coding. A recent variant [11] improves the quality of applause by modifying the gains only in the high frequency part of the signal.

AC-4 is a state-of-the-art transform-domain based audio codec standardized by ETSI and included in ATSC 3.0 [12], [13]. A high-level overview of the different coding tools used in AC-4 can be found in [12]. Similar to MP3 and AAC, MDCT window switching is also employed in AC-4. In AC-4, there is a dedicated core speech codec [14] designed to operate on speech at very low bitrates, where it offers an advantage over the core audio codec. However, there is a remaining opportunity to enhance the perceptual audio coding for a broader class of nonstationary signals and bitrates. One option is to employ the TNS tool, but it is well-known that TNS introduces artificial pre- and post-aliasing artifacts [15] due to the time-domain aliasing of MDCT. Furthermore, TNS requires a side-info of around 50 bits/frame. The AC-4 system was designed from the start to support an enriched set of metadata across several new categories for an enhanced listening experience [13]. Thus, limiting the side-info rate to a bare minimum was desired.

Our solution to reduce pre-echo, and improve the subjective speech quality, is to combine two add-ons: (1) a pre-processing step prior to the core transform encoder, and (2) a post-processing step affecting the core transform decoder output that performs the inverse operation of the pre-processing step. The solution, referred to as companding

[*] The author was with Dolby Laboratories Inc., San Francisco, USA, at the time of developing this technology but is currently at Qualcomm XR Research, Bangalore, India (e-mail: melkote.v@gmail.com).

10.21437/Interspeech.2018-2096

involves QMF domain pre- and post-processing around the transform-domain coding system: prior to transform encoding, the dynamic range of the signal is reduced locally within a QMF time slot and restored again post decoding, which naturally shapes the coding noise temporally. The post-processor ideally inverts the gains applied by the pre-processor. The proposed system instead is able to estimate the gains required by the post-processor directly from the output of the core decoder, in effect requiring no bits at all. This principle of companding operation is inspired from [16], where adaptive noise shaping in the frequency-domain is accomplished by a companding function.

Novel contributions in this paper are the following. First, using companding in the QMF-domain to achieve temporal noise shaping of core audio codec. Second, encoder control of the desired decoder companding level with a minimal side-info rate. Our approach is similar to the gain modification [10] approach. However, where a time-domain gain modification is prone to introducing artifacts due to non-smooth gain modification, the proposed companding operation in the QMF filter-bank with a short prototype filter [17] results in a smooth gain application according to the shape of the prototype filter. Furthermore, in gain modification a time-varying gain and the modification time interval are transmitted, resulting in a higher side-info rate compared to our proposed system.

In this paper we give details of the companding system as deployed in AC-4. Section 2 describes the companding operation in the encoder and decoder. Section 3 describes the companding control mechanism for maximizing the perceptual quality. Finally, its performance is evaluated in Section 4.

2. Companding

The AC-4 codec architecture is built around core waveform coding in the MDCT-domain and parametric coding tools operating in a complex QMF domain. A similar architecture is also used in HE-AAC [17], but in AC-4 the waveform core coder operates at the same sampling rate as the input signal. In the MDCT-domain, two different spectral front-ends are provided, one tailored for coding arbitrary audio signals (Audio Spectral Frontend, the ASF), and the other tailored for coding speech signals (Speech Spectral Frontend, the SSF [14]). The former is based on a perceptual model of quantization and coding [6], while the latter employs a source model of speech. The Advanced Spectral Extension algorithm (A-SPX) operates in the QMF-domain and performs high-frequency reconstruction from the core waveform coded low-band signal (similar to the familiar MPEG-4 SBR [8], [17]). The companding tool is employed in the QMF-domain to achieve temporal shaping of the core coder (ASF or SSF) quantization noise. In the following sub-sections, we discuss the companding encoder and decoder operations in the context of a mono channel. However, multiple channels can be handled by duplicating the operation separately on each channel.

2.1. Companding – encoder

Companding in the encoder reduces the dynamic range of the input audio signal before the core encoding process. Modification is done per QMF time slot by a broadband gain value. These gain values amplify slots of relatively low intensity and attenuate slots of relatively high intensity. Therefore, the output of the core decoder is a signal with reduced dynamic range perturbed by core coder quantization noise of almost uniform level (time envelope) within each frame. The companding block in the AC-4 encoder is shown in Figure 1. The companding control block is discussed below in Section 3.

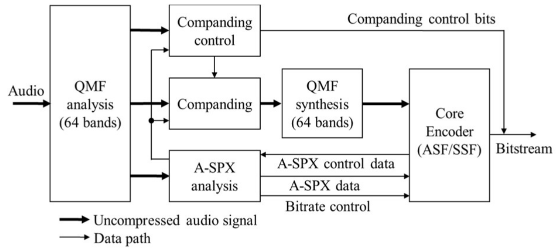

Figure 1: *Companding in AC-4 encoder.*

In an example configuration of AC-4, the length of the core coder frame is 4096 samples, and it has an overlap of 2048 samples with the neighboring frames. At 48 kHz such a frame is 85.3 ms long. In contrast, the employed QMF has a stride of 64 samples, providing a fine temporal resolution of 1.3 ms for applying the gains. Furthermore, the QMF has a smooth prototype filter that is 640 samples long ensuring that the gain application varies smoothly across time. Analysis with this QMF filter-bank provides a time-frequency tiled signal representation, where each QMF time-slot nominally corresponds to a stride of samples and has 64 uniformly spaced sub-bands of width 375 Hz. Depending on the signal characteristics, the actual number of QMF time-slots within an A-SPX frame can be different from $2048/64 = 32$, and is generated by the A-SPX frame information generator (inside the A-SPX Analysis box in Figure 1).

Let $S_t(k)$ be a complex valued filter-bank sample at time slot t and frequency bin k. The pre-processing step is to scale the core codec input to become

$$SC_t(k) = S_t(k) g_t, \tag{1}$$

where g_t is the *normalized slot mean* or referred to as the *gain*, with $g_t = (SM_t)^{\alpha-1}$ and $\alpha = 0.65$. Here $SM_t = \frac{1}{K}\sum_{k=1}^{K}|S_t(k)|$ is the *mean absolute level*, where K could be equal to the total number of sub-bands in the filter-bank, or lower. A generic p-norm could be defined in this context as $SM_t = \frac{1}{K}\left(\sum_{k=1}^{K}|S_t(k)|^p\right)^{1/p}$. Gain calculation using a p-norm of the spectral magnitudes with $p < 2$ has been found to be more effective in shaping quantization noise, than basing it on energy ($p = 2$). Hence, in AC-4, mean absolute level ($p = 1$) has been chosen. An explanation for this choice is given in the next section.

2.2. Companding – decoder

Companding in the decoder restores the core decoder outputs back to the original dynamic range by applying the inverse of the encoder gain values per QMF time slot. In this manner, core coder quantization noise is concurrently shaped to approximately follow the temporal envelope of the original signal. This has the desired effect of rendering the quantization noise less audible during quiet passages. Although the noise is amplified during passages of high intensity, it is still inaudible, due to the temporal masking effect of the loud signal. The companding block in AC-4 decoder is shown in Figure 2.

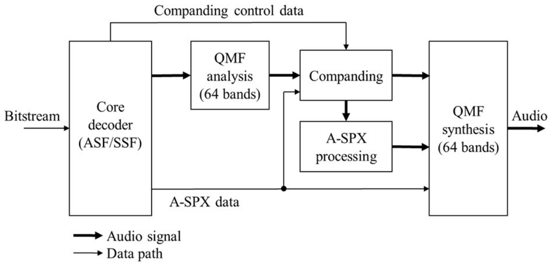

Figure 2: *Companding in AC-4 decoder.*

In the companding decoder, the dynamic range of the core codec output is expanded by an inverse of the companding encoder. This requires an exact replica of the QMF filter-bank employed in the encoder. Furthermore, they need to be time synchronous in that any effective delays between the output of the pre-processor and input to the post-processor must be in multiples of the stride of the filter-bank.

Let $R_t(k)$ be a complex valued sample of this decoder filter bank. The post-step is to scale the core codec output to become

$$RC_t(k) = R_t(k)h_t, \tag{2}$$

where h_t is a normalized slot mean with $h_t = (RM_t)^{\frac{1-\alpha}{\alpha}}$ and $\alpha = 0.65$. Here $RM_t = \frac{1}{K}\sum_{k=1}^{K}|R_t(k)|$ is the mean absolute level. In AC-4, mean absolute level has been chosen because a gain calculation that is based on the short-term energy (*2*-norm) tends to be biased towards the stronger low frequencies and hence is dominated by the stationary sources, and exhibits little variation across time. Thus it is ineffective in temporally shaping the noise introduced by the core coder. However, computing $|S_t(k)|$ and $|R_t(k)|$ in the companding encoder and decoder can be computationally expensive. Hence, a lower complexity approximation (based on [18]) of calculating the magnitude of a complex number is standardized [19].

In the ideal case where $R_t = SC_t$, it is easy to check that $h_t g_t = 1$. In reality, the gain applied in the encoder is a close approximation of the inverse of the gain applied in the decoder. Furthermore, as made clear in Section 3, the companding encoder and decoder are not completely blind. Encoder control of the desired decoder expanding level is signaled in the bitstream so that inverse of the encoder gain values are applied to the corresponding QMF time slots in the decoder.

3. Companding control

Initial experiments showed that companding provides benefit for speech and transient signals. However, companding degraded the quality of some stationary music signals. A reason for this: modifying each QMF time slot individually with a gain value, results in discontinuities. Figure 3 shows the variation of companding gain across QMF time slots from a chord excerpt. Such discontinuities at the companding decoder result in discontinuities in the envelope of the shaped noise, something which causes audible artifacts. Hence, a companding control mechanism is desired.

With a control, companding is activated for transient signals and switched off for stationary signals. In addition, instead of switching off companding abruptly, a constant gain is applied to an audio frame resembling the gains of adjacent active companding frames. Such a gain factor is calculated by averaging the mean absolute levels over the slots in a frame. This mode of companding is called *average companding* (see, Figure 3). The companding on/off/averaging decision is detected in the encoder and transmitted to the decoder. A temporal peakiness measure, computed over QMF frequency bands, is used as detector for controlling the compander in the encoder. For highly correlated multi-channel signals, equal companding gains are applied to all the channels.

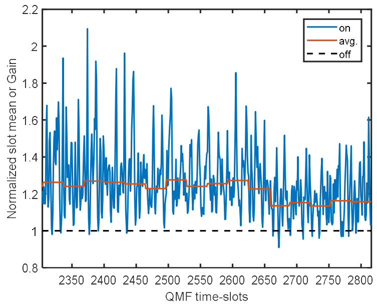

Figure 3: *Variation of companding gain across QMF time slots from a chord excerpt with companding enabled (blue), disabled (black-dashed), and with average companding (red).*

Thus, for switching companding on or off, plus an intermediate average companding require transmitting 1 or 2 bits/frame. Another bit needs to be transmitted to signal if all channels of a multichannel configuration use common companding gain factors; or if they should be calculated independently for each channel.

4. Listening results

In the following, we evaluate the impact of companding on perceptual quality with MUSHRA [20] listening tests. Listening tests below included a hidden reference, two low-pass anchors (7 kHz and 3.5 kHz), and the codecs under evaluation. Audio files with a sampling frequency of 48 kHz were used as input to the codec. All tests were performed by experienced listeners using headphones.

4.1. Performance in AC-4

In this section, we evaluate the performance of AC-4 at stereo 64 kbit/s without and with companding. For the latter, our companding detector was used to control the companding. MDCT window switching was active in both cases. The test is done on well-known MPEG test set and on a stereo version of an applause excerpt. The MPEG test set consists of: es01 (Suzanne Vega), es02 (male speech, German), es03 (female speech, English), sc01 (trumpet), sc02 (orchestra), sc03 (pop music), si01 (harpsichord), si02 (castanets), si03 (pitch pipe), sm01 (bagpipe), sm02 (glockenspiel), and sm03 (plucked strings). Surround channels of a critical 5.1 applause excerpt from the EBU multichannel test set [21] were used as a test signal for applause (appLsRs) because distinct clapping sounds are present in these channels. As can be seen from the differential MUSHRA scores in Figure 4, companding improves the quality of the speech excerpts, applause, and castanets. Improvements in the range of 8 and 3-7 MUSHRA

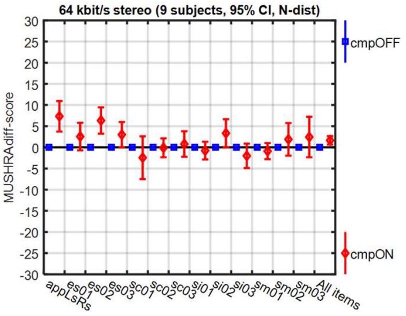

Figure 4: *MUSHRA differential results of 64 kbit/s stereo AC-4 coding with companding control (red); with respect to AC-4 coding without companding (blue).*

points are observed for applause and speech, respectively. Overall, none of the items are significantly degraded, indicating the desired functioning of the companding control mechanism.

4.2. Performance against TNS

In this section, we benchmark the companding performance against TNS, and compare the decoded speech quality at 32 kbit/s mono. For a fair comparison, we compared their performance within a common coding framework. We chose the coding framework to be HE-AAC because TNS is operating in the core codec. We implemented companding in the available QMF in HE-AAC and modified the HE-AAC bitstream with additional companding control bits. We used this HE-AAC codec augmented with the companding encoder-decoder system to compare the performance against TNS.

In the first test, augmented HE-AAC at 32 kbit/s mono was used to compare the performance with neither TNS nor companding, with TNS but no companding, and without TNS but with companding. Solely speech items were tested, since TNS is known to improve the speech quality [6], [7]. Figure 5 depicts the differential MUSHRA scores. In the item mnemonics, the first letter indicates a language: English, German, Korean or French; the second letter indicates female or male speakers. As expected, all speech items are improved with TNS. With companding enabled, it can be observed that companding is also able to match the performance of TNS. With either TNS or companding, improvements as high as 11 points are observed. Overall, there is a significant average improvement of 5 points while none of the items are significantly degraded.

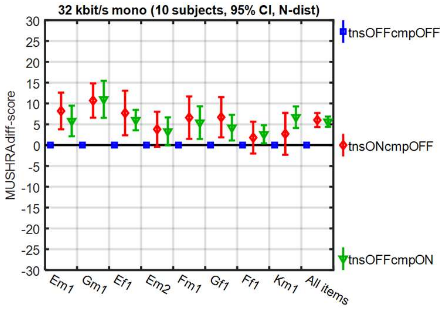

Figure 5: *MUSHRA differential results of 32 kbit/s mono HE-AAC augmented with companding. Differential results of either TNS enabled (red) or companding enabled (green) is shown with respect to augmented HE-AAC codec without both companding and TNS (blue).*

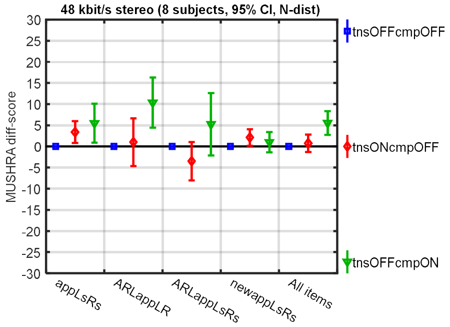

Figure 6: *MUSHRA differential results of 48 kbit/s stereo HE-AAC augmented with companding. Differential results of either TNS enabled (red) or companding enabled (green) is shown with respect to augmented HE-AAC codec without both companding and TNS (blue).*

In the second listening test, the test setup remains the same as above, except that we chose only applause excerpts and compare the performance at 48 kbit/s stereo. The four items chosen in the test included the applause excerpt (mentioned in sub-section 4.1), left and right channels of ARLapplause, and left- and right-surround channels of ARLapplause and new_applse_15. The last applause excerpt is noisy and lacks distinct claps. Figure 6 depicts the differential MUSHRA scores. Now it can be observed that companding provides a significant improvement over TNS. With companding, improvements as high as 10 MUSHRA points can be observed over TNS. Overall, there is a significant average improvement of 6 points while none of the items is significantly degraded. Similar results are expected for other signals containing dense transient events (e.g., rain, crackling fire, etc.).

We attribute the improvements with companding due to the following. First, companding relaxes the bit rate demand imposed on the core encoder by reducing short-time dynamics of the input signal. Thus, improvements are most observed on applauses with distinct claps. Second, speech and applauses have traditionally been very difficult to code with perceptual audio codecs, particularly at low bit rates. Due to bare minimum side-info rate, remaining bits are better utilized in core encoding of the signal with reduced dynamics.

5. Conclusions

This paper presents a new coding tool called companding. Companding is a tool in the Dolby AC-4 coding system for improved perceptual coding of signals that predominantly consist of speech and dense transient events such as applause signals. The benefits of applying companding are two-fold: companding relaxes the bit rate demand imposed on the encoder by reducing short-time dynamics of the input signal; additionally, companding ensures proper temporal noise shaping in the decoder. These benefits are achieved with negligible side-info rate. Subjective tests of companding have shown significant improvement for speech and applause.

6. References

[1] A. Spanias, T. Painter, and V. Atti, *Audio signal processing and coding*. John Wiley & Sons, 2006.

[2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. New York: Academic Press, 1997.

[3] "Perceptual audio coders, what to listen for," CD-ROM, AES Technical Committee on Coding of Audio Signals, 2002.

[4] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz,* vol. 43, pp. 252–256, 1989.

[5] K. Brandenburg and G. Stoll, "ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[6] M. Bosi et al., "ISO/IEC MPEG-2 Advanced Audio Coding," *Journal of the Audio Engineering Society*, vol. 45, no. 10, pp. 789–814, 1997.

[7] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *101st Audio Engineering Society Convention*, November 8-11, Los Angeles, CA, USA, Proceedings, 1996, preprint #4384.

[8] J. Herre and M. Dietz, "MPEG-4 high-efficiency AAC coding [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137–142, 2008.

[9] S. Quackenbush, "MPEG Unified Speech and Audio Coding," *IEEE MultiMedia*, vol. 20, no. 2, pp. 72-78, 2013.

[10] M. Link, "An attack processing of audio signals for optimizing the temporal characteristics of a low bit-rate audio coding system," in *95th Audio Engineering Society Convention*, October 7-10, New York, NY, USA, Proceedings, 1993, preprint #3696.

[11] F. Ghido, S. Disch, J. Herre, F. Reutelhuber, and A. Adami, "Coding of fine granular audio signals using High Resolution Envelope Processing (HREP)," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 5-9, New Orleans, LA, USA, Proceedings, 2017, pp. 701–705.

[12] K. Kjörling et al. "AC-4–The Next Generation Audio Codec," in *140th Audio Engineering Society Convention*, June 4-7, Paris, France, Proceedings, 2016, preprint #9491.

[13] J. Riedmiller et al., "Delivering Scalable Audio Experiences using AC-4," *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 179–201, 2017.

[14] L. Villemoes, J. Klejsa and P. Hedelin, "Speech coding with transform domain prediction," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 15-18, New Paltz, NY, USA, Proceedings, 2017, pp. 324–328.

[15] C. M. Liu, H. W. Hsu and W. C. Lee, "Compression Artifacts in Perceptual Audio Coding," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 681–695, May 2008.

[16] R. Lefebvre and C. Laflamme, "Shaping coding noise with frequency-domain companding," in *IEEE Workshop on Speech Coding For Telecommunications Proceeding*, September 7-10, Pocono Manor, PA, USA, Proceedings, 1997, pp. 61–62.

[17] A. C. den Brinker, J. Breebaart, P. Ekstrand, J. Engdegård, F. Henn, K. Kjörling, W. Oomen, and H. Purnhagen, "An overview of the coding standard MPEG-4 Audio Amendments 1 and 2: HE-AAC, SSC and HE-AAC v2," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, 2009, Article ID 468971.

[18] M. Allie and R. Lyons, "A root of less evil [digital signal processing]," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 93–96, 2005.

[19] "Digital audio compression (AC-4) standard; part 1: Channel based coding," ETSI TS Standard 103 190-1 V1.2.1, European Telecommunications Standards Institute, June 2015.

[20] "Method for the subjective assessment of intermediate quality levels of coding systems," ITU-Recommendation BS.1534-3, 2015.

[21] "EBU Evaluations of Multichannel Audio Codecs," EBU-Tech. 3324. Geneva, September 2007.