Low Delay Robust Audio Coding by Noise Shaping, Fractional Sampling and Source Prediction

Jan Østergaard
Section on Artificial Intelligence and Sound
Department of Electronic Systems
Aalborg University
Denmark

IEEE Data Compression Conference, 2021

Outline

- Motivation
- Key challenges addressed in the work
- Existing work on delta-sigma quantization for audio coding
- Contributions
  - Extension to many descriptions
  - Fractional sampling
  - Creating balanced descriptions (rate & distortion wise)
  - Decoding rules (MMSE versus PEAQ)
- Simulation study
- Conclusions

Motivation

- Interactive streaming of sound is getting more and more popular.
- For example, Zoom, Google Meet, and TEAMS are often used for online teaching.
- State-of-the-art speech/audio coders: BV32, MPEG, AAC-ELD, MPEG-USAC, 3GPP EVS, Opus.
- Interactive music rehearsal or performances require high quality and extremely low latency.
- Even for one way end-2-end delays > 5 ms, som music is hard to play.
- A good solution is JackTrip from CCRMA Stanford
  (no data compression and no efficient solutions towards packet losses)

Motivation and key challenges for music over networks

- Many wireless channels suffer from packet losses – e.g., 5% losses.
- Even wired communications over the internet suffers from jitter, especially when driving the communications near the minimal possible practical latency.
- In music performances both lost and late packets are ”lost”.
- Re-transmissions add latency – and require a feedback channel.
- The playback rate needs to be stable (nearly constant).
- To ensure this, a jitter (playback) buffer is used, which stores a number of packets before being played out.
- The delay is therefore proportional to the number of packets stored in the buffer.
- Packet-loss concealment methods are mainly helpful when interpolating between short gaps and not extrapolating into the future.

Key challenges addressed in this work:

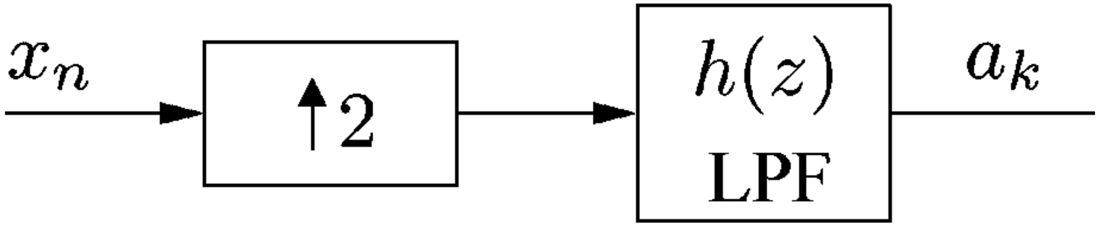
1. very low delay high-quality audio coding
2. robustness to packet losses and packet jitter without introducing further delay

Multiple description audio coding

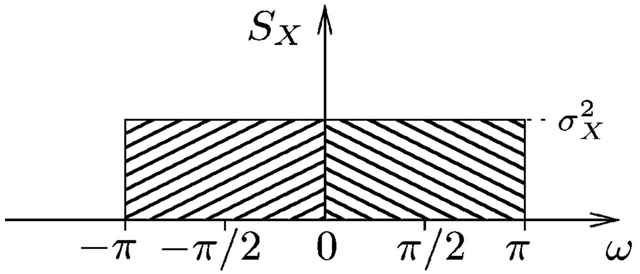
- There are many ways to construct multiple descriptions but less work has applied it to audio coding: (this list is not exhaustive)
  - Multiple description perceptual audio coding with correlating transform.
    Kovacevic, V.K. Goyal. IEEE Trans. Speech and Audio Processing, 2000.
  - Robust low-delay audio coding using multiple descriptions.
    G. Schuller, J. Kovacevic, F. Masson, V.K. Goyal. IEEE Trans. Speech and Audio Processing, 2005.
  - Perceptual audio coding using n-channel lattice vector quantization.
    J. Østergaard, O. Niamut, J. Jensen, R. Heusdens. IEEE ICASSP 2006.
  - Multiple description coding for an mp3 coded sound signal. H. Wey, A. Ito, T. Okamoto, Y. Suzuki. ICA 2010.
  - Real-time perceptual moving-horizon multiple-description audio coding. J. Østergaard, D.E. Quevedo, J. Jensen. IEEE Trans. Signal Processing. 2011.
  - **Practical design of delta-sigma multiple-description audio coding**.
    J. Leegaard, J. Østergaard, S.H. Jensen, R. Zamir. EURASIP Journal on audio, speech, and music. 2014.

Delta sigma quantization

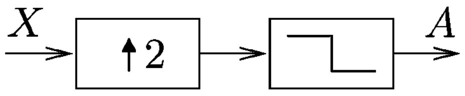
- In delta-sigma quantization, the source is oversampled



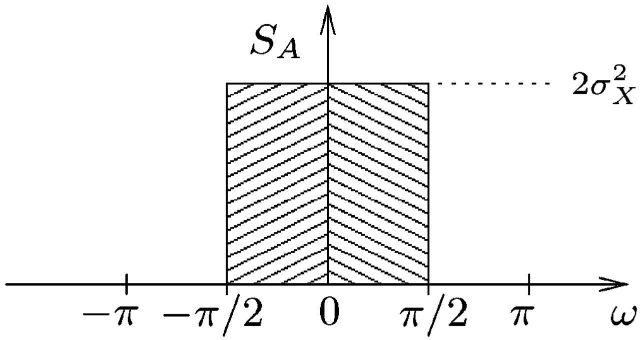
- Consider a white Gaussian source
- Upsample by a factor of 2
- The resulting spectrum covers half the frequency band



(a) Spectrum of X



(b) Oversampling by two

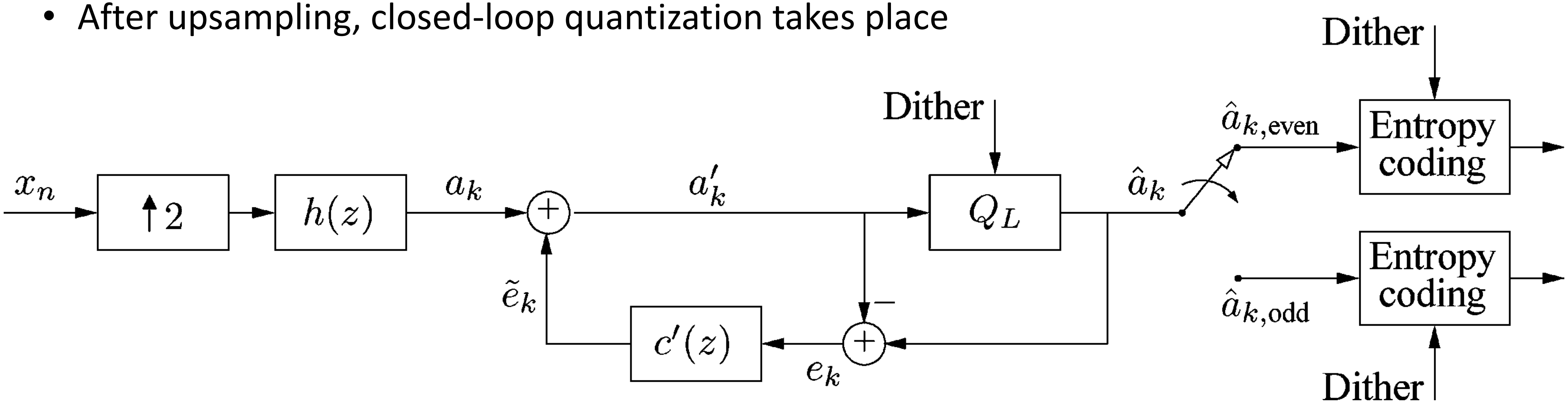


(c) Spectrum of A

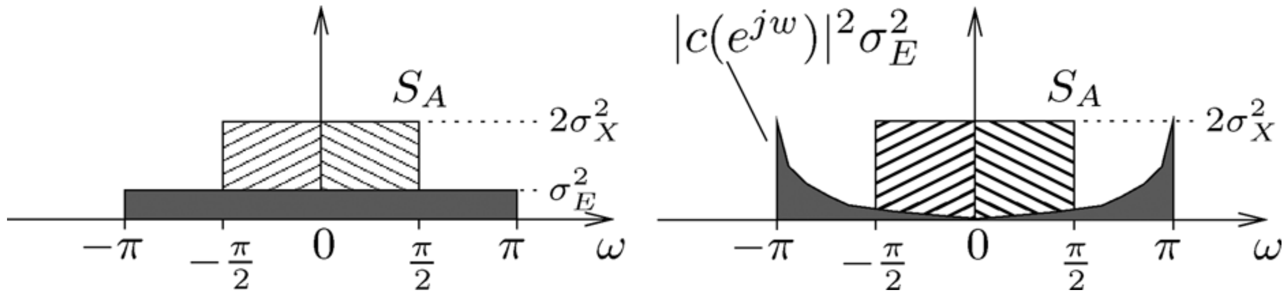
Multiple-Description Coding by Dithered Delta Sigma Quantization.
J. Østergaard, R. Zamir, IEEE Data Compression Conference, 2007.

Delta sigma quantization

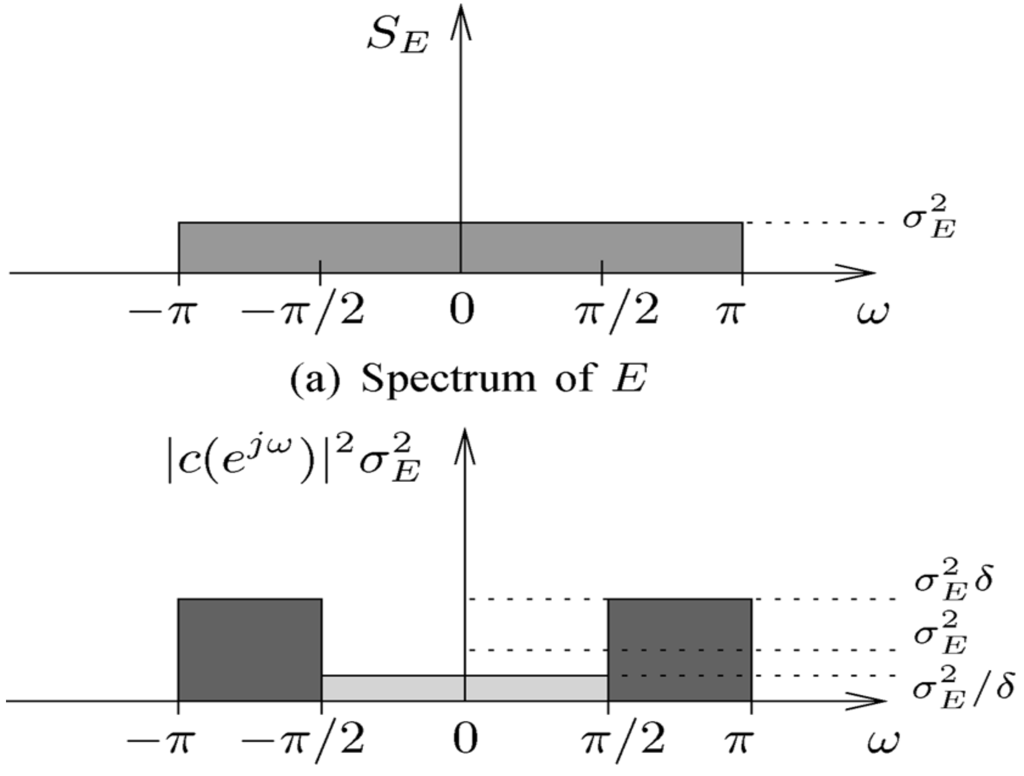
- After upsampling, closed-loop quantization takes place



- The quantization noise covers the full spectrum and is white before being shaped.
- The quantization noise is shaped by a noise-shaping filter, which reduces the energy of the in-band noise spectrum.



Ideal noise shaping

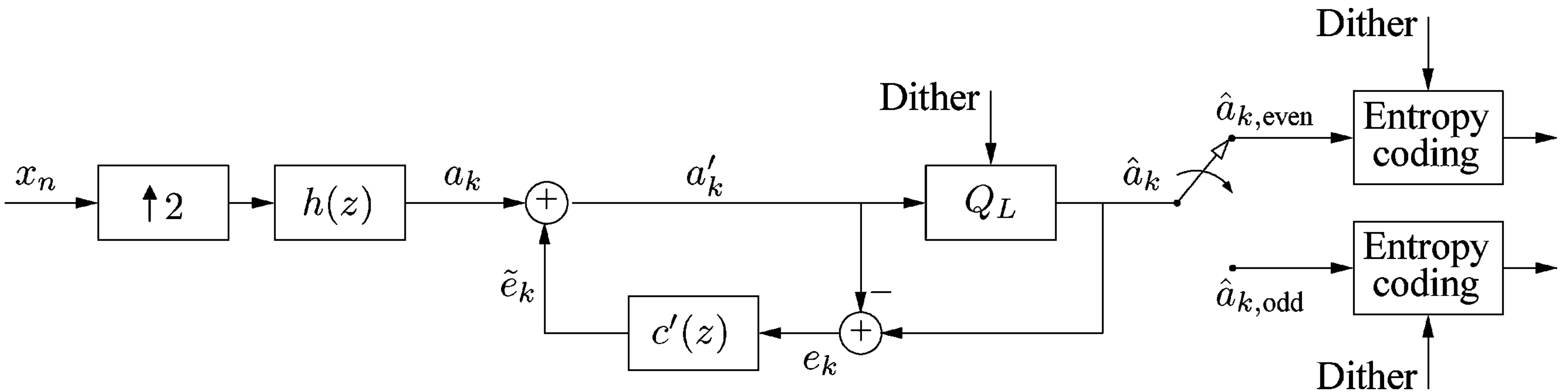


(a) Spectrum of E

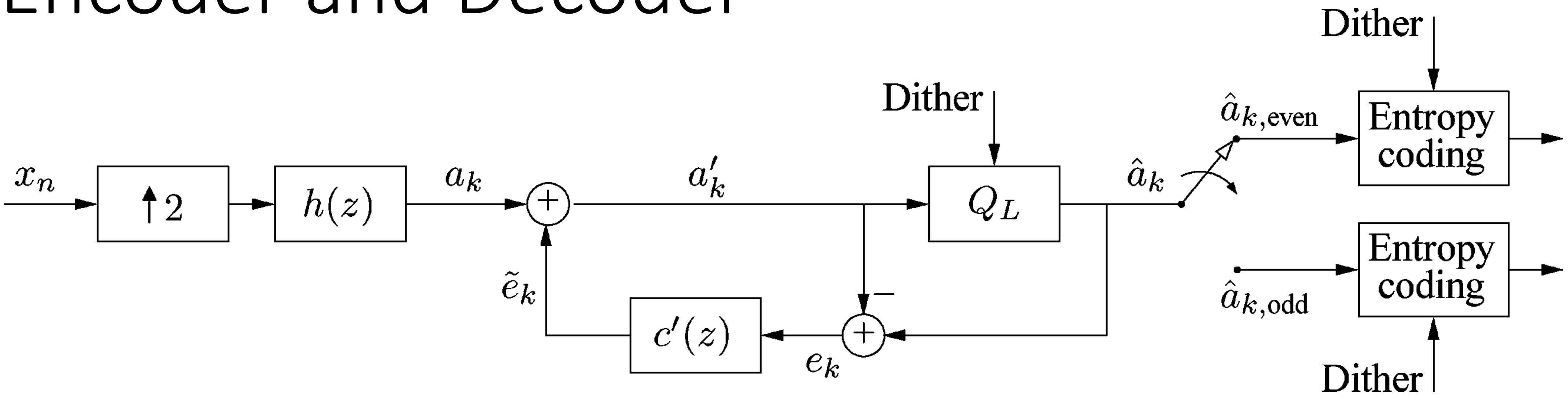
$|c(e^{j\omega})|^2\sigma_E^2$
$\sigma_E^2\delta$
σ_E^2
σ_E^2/δ
$-\pi$ $-\pi/2$ 0 $\pi/2$ π ω

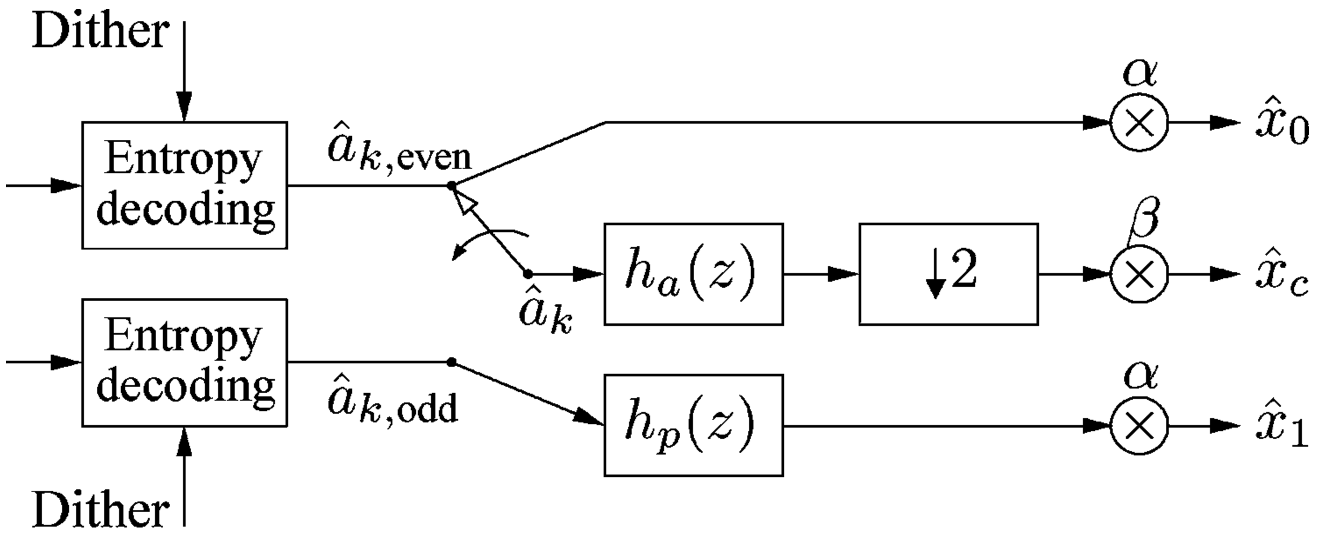
(b) Spectrum of shaped E

- Using approximately ideal noise-shaping filters, the resulting noise spectrum is shaped like a two-step function
- Splitting into even and odd samples, effectively downsamples the signal without first using an anti-aliasing filter
- The noise in each description is therefore aliased



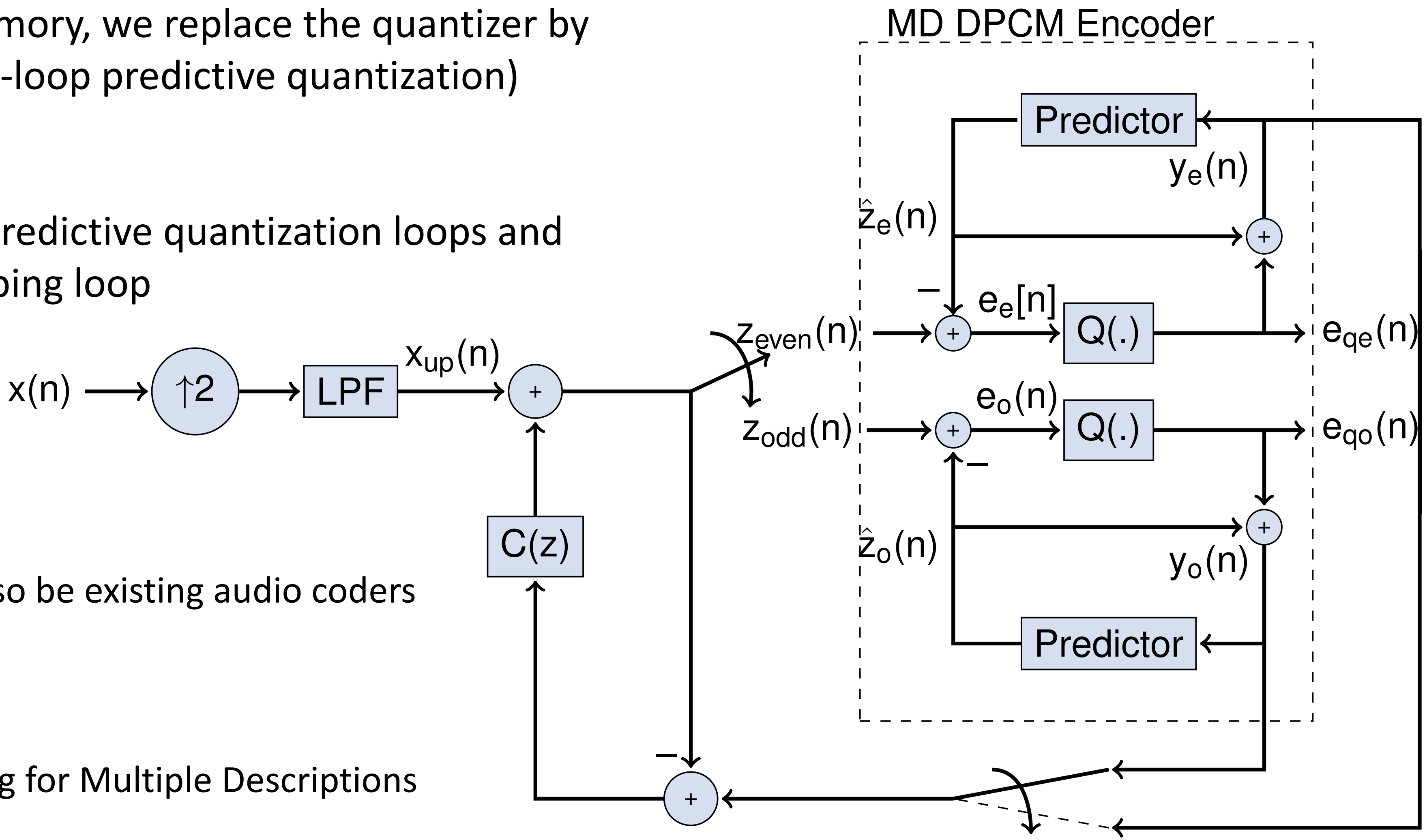
Encoder and Decoder





Noise-shaping & source prediction

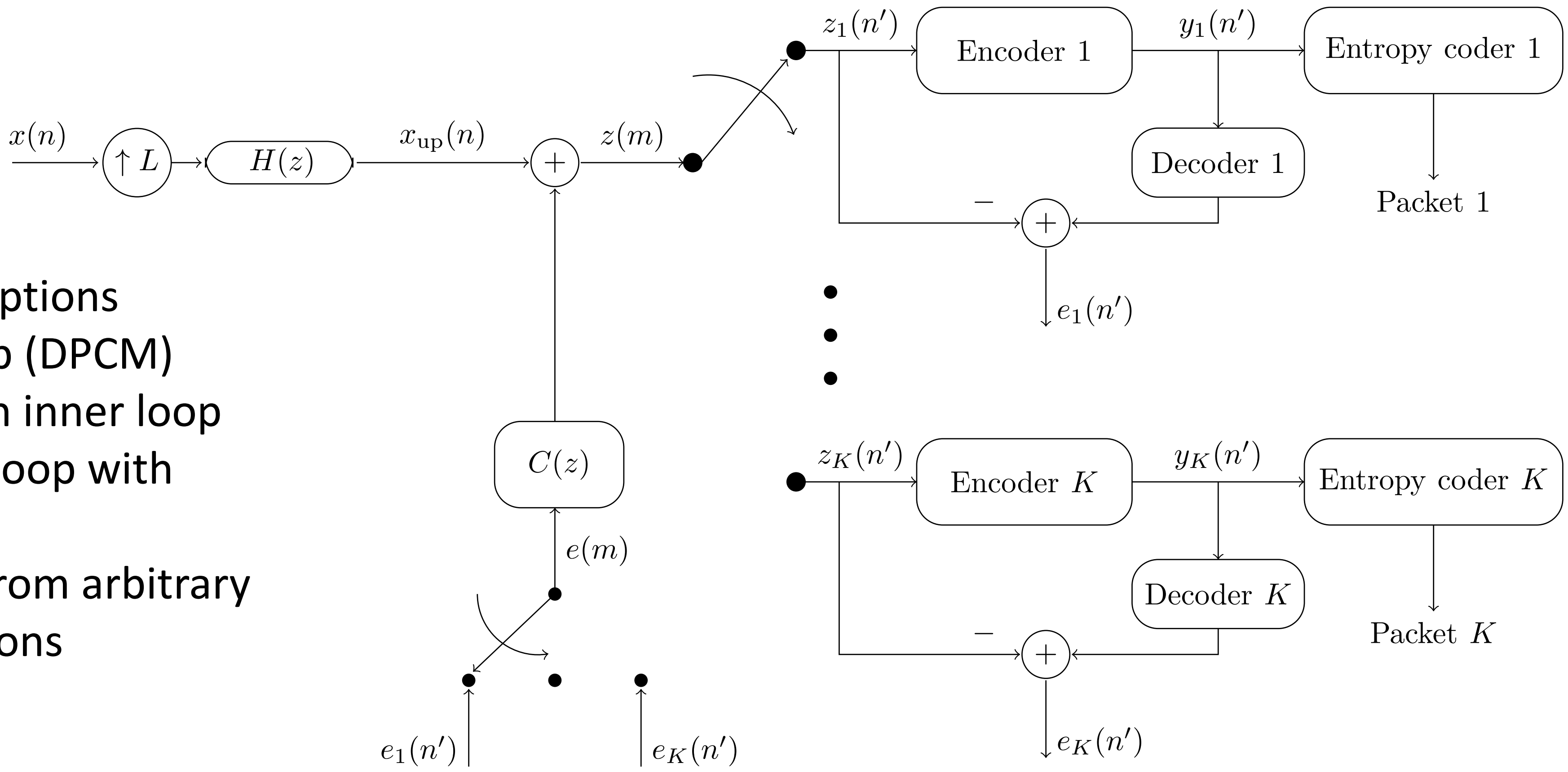
- For sources with memory, we replace the quantizer by a DPCM loop (closed-loop predictive quantization)
- We have two inner predictive quantization loops and one outer noise-shaping loop



The DPCM loop can actually also be existing audio coders

Noise-Shaped Predictive Coding for Multiple Descriptions of a Colored Gaussian Source.
Y. Kochman, J. Østergaard, R. Zamir. IEEE Data Compression Conference, 2008.

Many descriptions by fractional sampling



- Upsample by L ≥ 2
- Create K >= L descriptions
- Perform closed-loop (DPCM) quantization in each inner loop
- Perform one outer loop with noise shaping
- Perform decoding from arbitrary subsets of descriptions

Key research questions

- **Balanced descriptions**: Can we guarantee that the distortion only depends upon the number of received descriptions and not which?
- **Fractional under-sampling**: Is it advantageous to choose $K<L$
- **Decoder**: How do we reconstruct from a given subset of descriptions?

Distortion of different subsets of descriptions – a noise shaping strategy

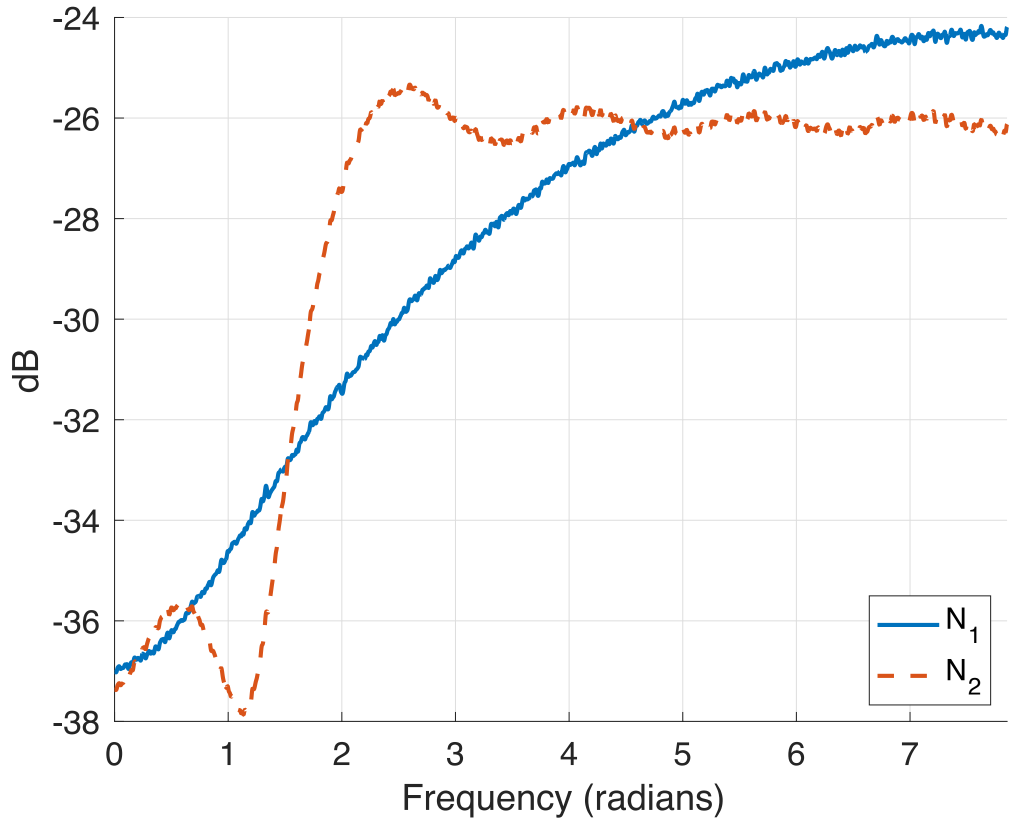
- Assume we upsample a white Gaussian source X by L=5
- Let N be white Gaussian noise
- Let Y1 = X + N1, and Y2 = X + N2
- Use any two descriptions (out of 5)

$$c_1 = [1, -0.6200].$$

$$c_2 = [1.0000, -0.4685, -0.2586, -0.0735, 0.0520, 0.1040, 0.0909, 0.0385, -0.0200, -0.0557, -0.0526].$$

$$S_N(\omega) = \begin{cases} \delta^{1-K}, & |\omega| \le \frac{\pi_L}{K}, \\ \delta, & \pi > |\omega| > \frac{\pi_L}{K}. \end{cases}$$

δ≥1



Subsets (i, j)	1,2	1,5	2,3	3,4	4,5
$Y_1^{(i,j)}$:	-5.37	-5.38	-5.35	-5.31	-5.34
$Y_2^{(i,j)}$:	-4.16	-4.18	-4.15	-4.11	4.00

Subsets (i, j)	1,3	1,4	2,4	2,5	3,5
$Y_1^{(i,j)}$:	-3.02	-3.02	-3.01	-3.02	-3.02
$Y_2^{(i,j)}$:	-4.00	-4.00	-4.00	-4.00	-4.00

Optimal decoder for non-stationary signals

- From an MMSE point of view, a two-stage approach is optimal:
  - First phase-shift each received description to achieve coherence with source
  - Average phase-shifted descriptions to obtain final estimate of source

[Machiach, Østergaard, Zamir, ITW 2013]

- Optimality was established for $L=K$ but not for $L<K$ or $L>K$

- We propose a heuristic decoding rule as a two-stage approach:
  - First replace lost decriptions by the ”nearest” received description
  - Lowpass filter and downsample by L to source sampling frequency

”MMSE” decoder versus Heuristic decoder - MSE

- Source is 10 sec. of Celine Dion music, sampled at 48 kHz
- Framesize is 120 samples, corresponding to 2.5 ms delay
- We upsample by L=2 and downsample by K=5 (descriptions)
- The ratio ƛ of the in-band and out-ot-band spectra of the shaped noise is varied, which control the side versus central distortion ratio.

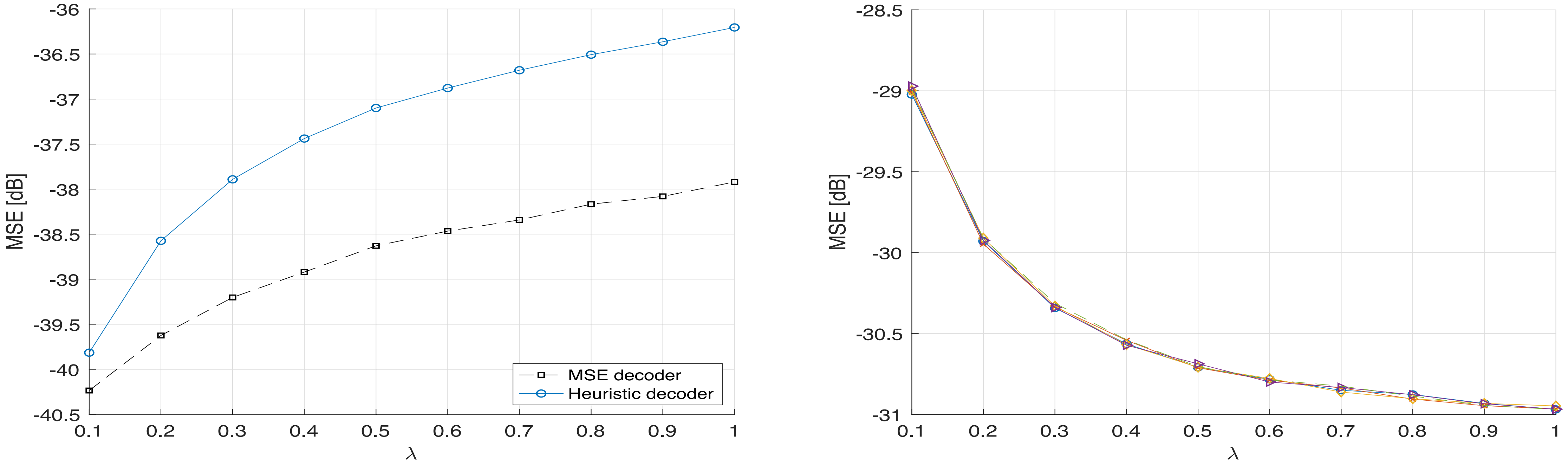


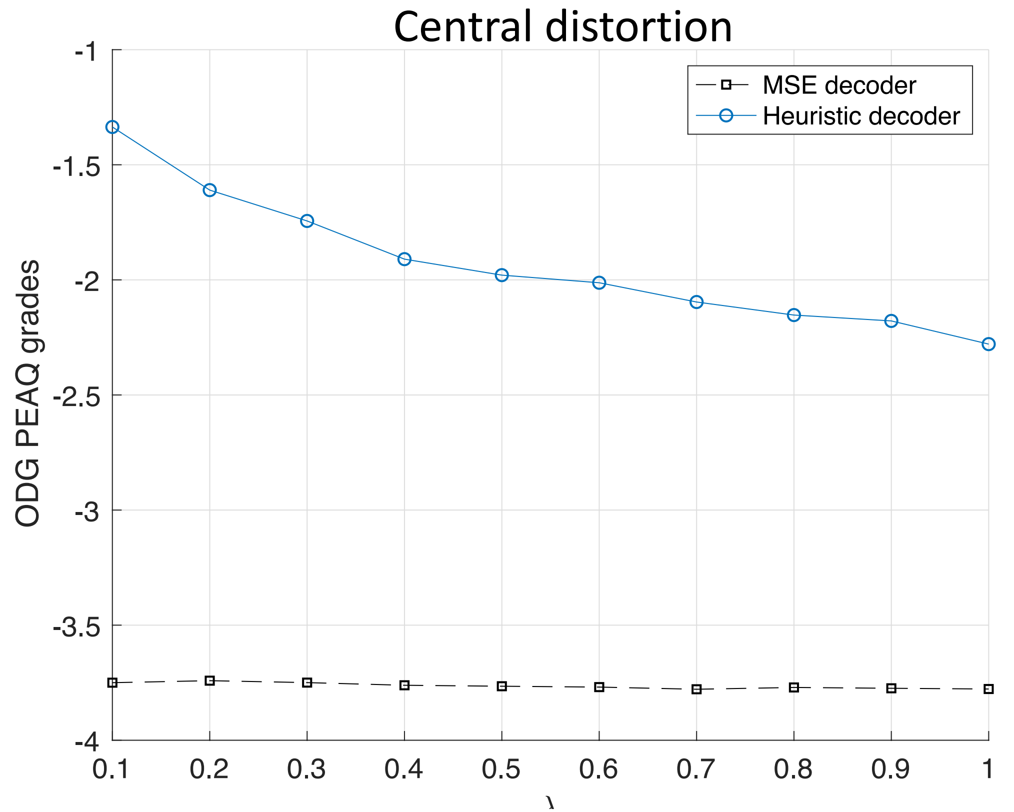
Figure 2: Central (left) and side (right) MSE distortion as a function of λ using two different techniques for central reconstruction. Here $K = 5$, $L = 2$, and the frame size is 120 samples.

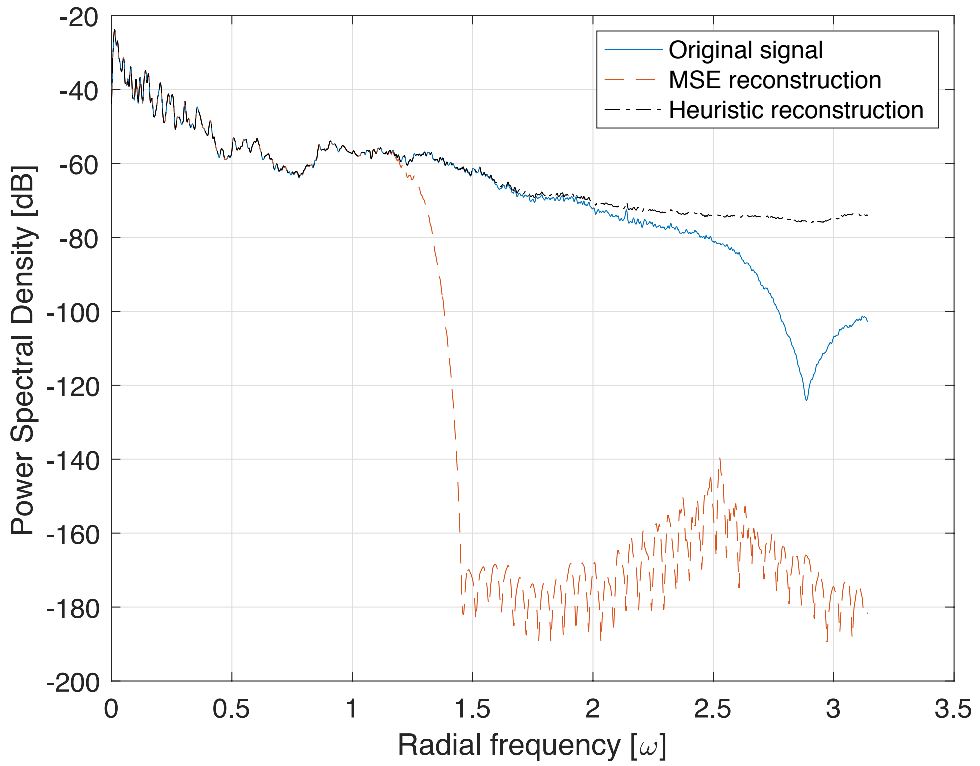
”MMSE” vs Heuristic decoder: Objective Difference Grade (ODG)

Impairment	ITU-R 5-grade scale	ODG
Imperceptible	5.0	0.0
Perceptible but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

Due to source aliasing in individual descriptions, the ”MMSE” decoder does not necessarily guarantee a smooth transitions between blocks and it will course a low pass filtering of the signal.

(note that it is not the true MMSE decoder)





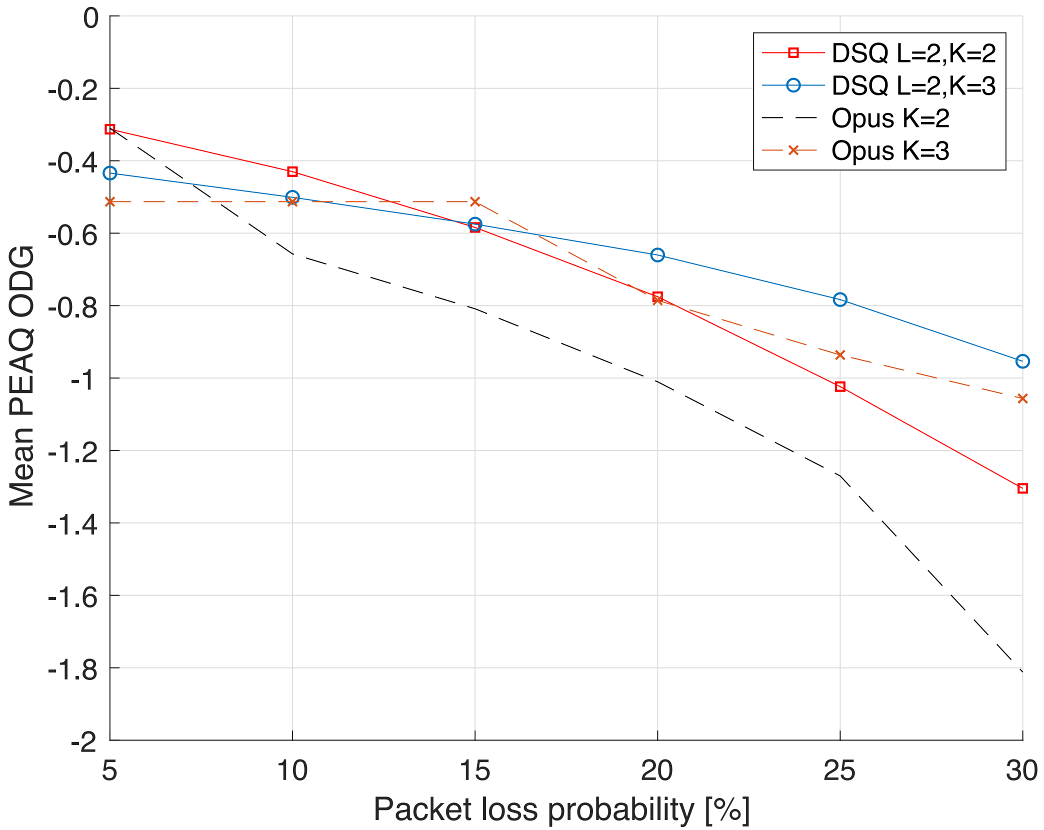
Simulation study: 300 kbps, 2.5 ms delay, i.i.d. packet losses. Music files: 10 excerpts each of 20 sec. duration

DSQ coder

- Oversample with L=2 and make K=2 or K=3 descriptions
- Total coding sumrate is 300 kbps
- Total delay is 120 samples (2.5 ms at 48 kHz)
- 10 dim. LSF vectors: 10 kbps per packet.
- Gain factors: 2 kbps per packet.

Opus coder

- Framesize set to 2.5 ms.
- Encoding at 100 and 150 kbps
- Repeating packets K=3 or K=2 times
- Total sumrate is 300 kbps.
- The effective packet loss rate is p^K
- Note that: round(100*0.15^3) = 0.
- Opus demo implementation with: ”-loss”



Conclusions and discussion

- A flexible multiple-description low-delay audio coder is proposed
  - Coding rate, latency, number of descriptions, and side-to-central distortion ratio can be arbitrarily chosen.
- The coder consists only of simple signal processing blocks
  - Fractional sampling, linear prediction, scalar quantization, and noise-shaping.
- The main application envisioned is very low delay high-quality interactive audio
- At 2.5 ms delay, the performance was better than perceptually optimized coders such as Opus (followed by repetition coding)
- Open source Matlab code is available, see paper for details.