Learning about perception of temporal fine structure by building audio codecs

LARS VILLEMOES[1*], ARIJIT BISWAS[2], HEIKO PURNHAGEN[1], AND HEIDI-MARIA LEHTONEN[1]

[1] *Dolby Sweden AB, Stockholm, Sweden*

[2] *Dolby Germany GmbH, Nürnberg, Germany*

The goal of audio coding is to efficiently describe an auditory experience while enabling a faithful reconstruction to the listener. The subjective quality compared to the original is measured by established psychoacoustic tests (BS.1116, 2015; BS.1534, 2015) and the description cost is measured in number of bits. As it is much cheaper to describe coarse scale signal properties than temporal fine structure (TFS), tools like noise fill, spectral extension, binaural cue coding, and machine learning have increased performance of audio codecs far beyond the first generation based on masking principles (e.g., mp3). In this evolution, implicit knowledge on hearing has been acquired by codec developers, but it has become increasingly difficult to construct tools to predict subjective quality. For example, it is yet unknown which aspects of the TFS that are essential for the listening impression to be preserved. To explore these issues, we study models of auditory representations with the mindset from audio coding. Given a method to solve the inverse problem of creating a signal with a specified representation, evaluating by listening can immediately reveal strengths and weaknesses of a candidate model.

INTRODUCTION

Coarse scale properties of audio signals are cheaper to describe than temporal fine structure (TFS; Moore, 2019). This is exploited in modern audio coding systems. But which aspects of TFS are important to make two signals sound the same to us? In this paper, we walk through current and emerging audio coding methods and suggest an audio coding inspired methodology to improve perceptual modelling. We illustrate this method by an example study regarding tonality which is inspired by research on audio texture synthesis, McDermott *et al.* (2009). For the sake of clarity, we will only discuss mono audio signals.

AUDIO CODING

The goal of audio coding is to convey an auditory experience faithfully while keeping the information rate low (see Fig. 1). A typical source is cinematic content comprising a mix of speech, music, and environmental sounds. Ideally, the decoded content should be perceptually indistinguishable from the original content. This is called transparency. Subjective testing such as BS.1116 (2015) can be used to quantify the

*Corresponding author: lars.villemoes@dolby.com

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

deviation from this ideal case. For the joint evaluation of several codecs and for larger deviations MUSHRA, BS.1534 (2015), is a better choice. Even in a modern scenario where video coding dominates the bit budget, a lower bitrate for a given perceived quality is preferred.

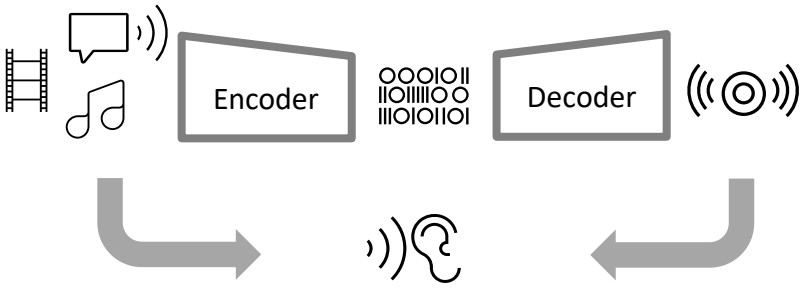

Fig. 1: Audio coding systems encode original sound sources into bits which can be transmitted and decoded into sound again at the receiver end. The combination of an encoder and decoder is called a *codec*.

Currently deployed tools

The concept of a time-frequency (TF) tile is often used to describe a segment of the audio signal in a perceptually motivated frequency band. In practice a filter bank or transform is employed to achieve this. The TF-tiles represent a sufficiently high dimensionality in signals space so that the number of TF-tiles needed to cover the whole signal is much smaller than the number of samples in the signal, see Fig. 2, panel a). Sharing information inside each TF-tile therefore enables bitrate savings. In codecs based on *waveform approximation*, such as mp3, (MPEG-1 layer III), the shared information is a quantization step size which controls the approximation error, and masking principles are employed to make the approximation error inaudible. For high bitrates this method can potentially preserve all aspects of TFS. Significant bitrate savings are obtained by only conveying the energy of the TF-tile, and letting the decoder synthesize a random noise signal in the TF-tile according to this energy target. This method is named *noise fill*. As only a very coarse scale envelope of the signal is preserved, the method rarely offers a high quality. Panel b) of Fig. 2 illustrates the difference between these two methods. *Parametric coding* improves on this situation by adding sinusoids and transients to the repertoire of synthetic signals. Finally, *spectral extension*, illustrated by panel c) of Fig. 2, consists of copying TFS from lower frequencies and adjusting tonal-to-noise ratio with parametric methods. This method is cheap and works surprisingly well. For more details, we refer to the recent tutorials by Brandenburg *et al.* (2013), and Herre and Dick (2019).

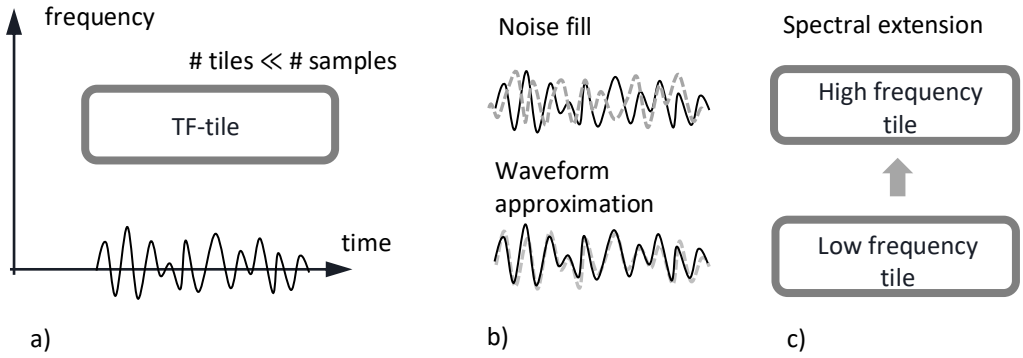

Fig. 2: Currently deployed audio coding tools. Panel a) shows the sharing of information in a TF-tile, panel b) the difference between waveform approximation and noise fill, and panel c) the principle of spectral extension.

Machine learning tools

The application of machine learning to audio synthesis typically consist of training a generative model that maps features or labels into sounds. As depicted in Fig. 3, one can think of these methods as inverse sound classifiers.

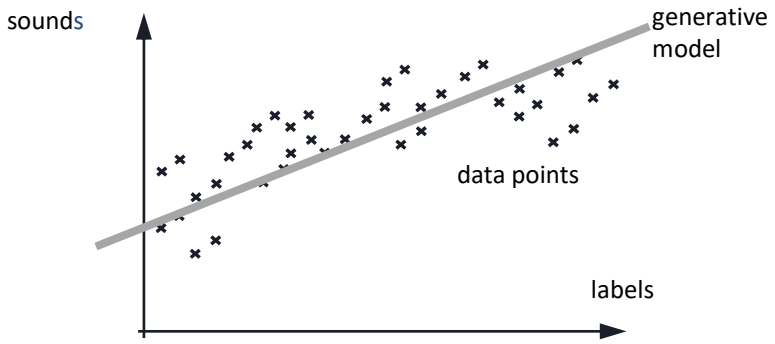

Fig. 3: Conceptual machine learning based sound synthesis.

Recent speech coding examples use vocoder features, such as linear predictive coding (LPC) based spectrum, pitch and degree of voicing (Kleijn *et al.* 2018; Klejsa *et al.* 2019). Autoregressive probability density models are trained on large speech datasets to approximate the distribution of signal samples conditioned on these features. Probabilistic sampling of the resulting model offers a substantial quality improvement over a manually crafted parametric vocoder synthesis. Example results from Klejsa *et al.* (2019) are given in Table 1.

Codec	SILK	SampleRNN	AMR-WB	Vocoder
Bitrate [kb/s]	16	8	23.05	8
MOS-LQO	4.41	3.48	4.39	3.67
MUSHRA	80	78	67	34

Table 1: Bitrates, average of predicted mean opinion scores (MOS-LQO) from the objective tool POLQA, P.863 (2018), and subjective mean MUSHRA scores for four codecs (from Klejsa *et al.*, 2019).

In terms of subjective MUSHRA scores, the machine learning codec based on the autoregressive model SampleRNN performs on par with the waveform codec SILK operating at twice the bitrate, while the parametric vocoder synthesis performs significantly worse. However, the mean opinion scores (MOS-LQO) predicted by POLQA contradict the subjective scores with respect to the comparison between the vocoder and SampleRNN.

PROBLEM

There is a gap in our understanding of auditory perception. The results in Table 1 offer one recent example of the frequently encountered phenomenon that the subjective performance of audio coding is not predicted well by tools such as POLQA and PEAQ (BS.1387, 2001), especially when mechanisms beyond masking are exploited by the codec. Our hypothesis is that TFS aspects are central for explaining this gap. As many model-based predictors of the results of psychoacoustic experiments compare stimuli in an *auditory representation*, (e.g., spectrogram or auditory filter bank, see P.863, 2018; BS.1387, 2001; Dau *et al.* 1996), such auditory representations will be the object of our study.

PROPOSED METHOD

As a complement to the established validation procedures based on targeted psychoacoustic testing, we here propose a method for evaluation and successive improvement of auditory representations based on the idea of building a *mock-up codec*, Fig. 4. For an original sound s having the auditory representation $\theta(s)$, the synthesis process consists of finding a sound u with $\theta(u) \approx \theta(s)$. This "synthesis by analysis" procedure was discussed by Slaney (1995) and is also the basis of spectrogram inversion methods, in which case the representation $\theta(s)$ is a spectrogram (Decorsière *et al.,* 2015).

Synthesis by analysis is an ill-posed inverse problem for which a solution is typically obtained only after many iterations starting from a random noise or manually crafted first guess. Whereas this approach might not be feasible for a deployable codec, off-line synthesis for the purpose of basic research is. Once the synthesis method is constructed, the idea is to run audio signals through the system and evaluate it as a

codec. The machine listener provided by the analysis can then be compared directly to the human listener.

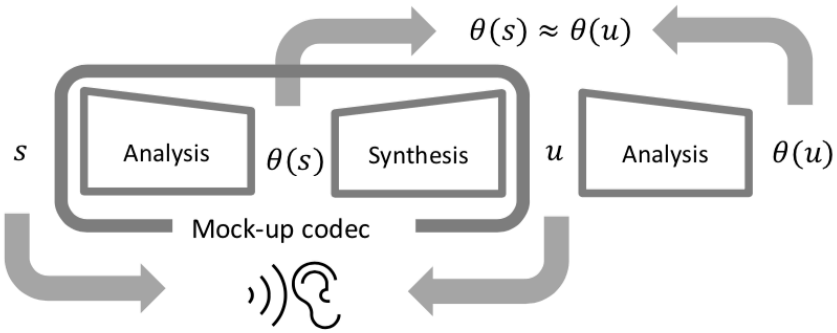

Fig. 4: Synthesis by analysis aims at producing a signal u given the analysis $\theta(s)$ of an original signal s by solving $\theta(u) \approx \theta(s)$.

EXAMPLE STUDY

To illustrate proposed method, we study audio representations derived from a framework which already includes tools for synthesis by analysis and whose signal analysis resembles that of many other models.

McDermott *et al.* (2009) evaluated combinations of time invariant summary statistics for description of stationary audio textures by the method of Fig. 4. For synthesis, an initial white noise signal u_0 was iteratively modified to bring the representation $\theta(u_n)$ closer to $\theta(s)$ than $\theta(u_{n-1})$. Most statistics were derived from envelope values updated every 2.5 ms in a filter bank with 38 bands of perceptually motivated resolution. The quality of textures containing tonal components was not captured well in these experiments. Given our interest in TFS of arbitrary nonstationary signals, and with inspiration from the literature on pitch perception modelling regarding tonality, (Meddis and O'Mard, 1997), we consider two deterministic representations.

A. **Baseline:** measure envelopes every 2.5 ms for all 38 bands as used by McDermott *et al.* (2009).
B. **Extension:** add one lag T and the value $\rho(T)$ of ρ, the normalized autocorrelation function (ACF) for each of the envelope-normalized subband signals every 20 ms.

Fig. 5 depicts the analysis block diagram for both cases. The lag T can be selected in many ways, and the specific steps taken to avoid picking lags related only to the center frequency of the subband are described in Fig. 6. For synthesis, we apply the method of iterative modification of an initial white noise signal. Gradient descent is used for the ACF data.

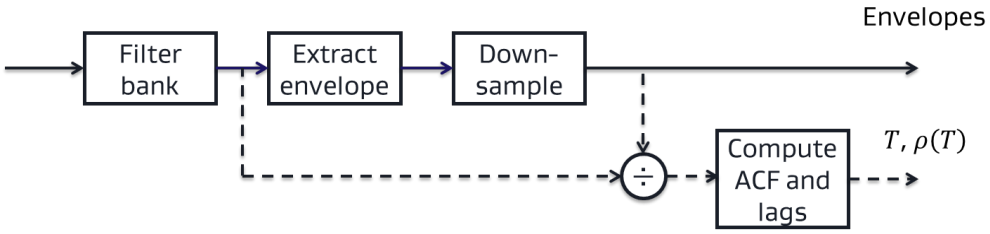

Fig. 5: Analysis algorithm for baseline (solid) and extended (solid and dashed) representations.

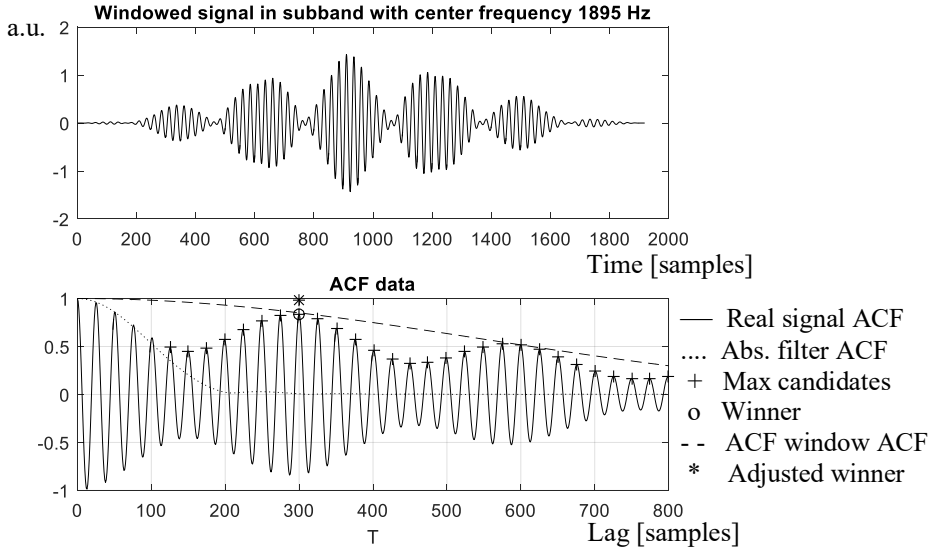

Fig. 6: Example of lag search for sawtooth signal with period 300 samples at 48 kHz sampling. The top panel depicts the windowed subband signal for a band with center frequency 1895 Hz. In the bottom panel, we consider the local maxima (crosses) of the windowed signal's ACF (solid curve), discarding maxima below the ACF of the absolute value of the impulse response of the subband filter (dotted curve). For a better usage of the interval between 0 and 1 where $\rho(T) = 1$ denotes maximum tonality, the maximum value (circle) is divided by the value of the ACF (dashed curve) of the subband ACF window leading to the final selection (star).

RESULTS

Informal listening to inputs and synthesized signals for speech, music, and environmental sounds reveals that the relatively detailed envelope representation (Fig. 7A) alone is not sufficient to capture tonality, while a clear improvement is obtained in voiced parts of speech and tonal parts of music by using the extension (Fig. 7B) including one lag and the corresponding ACF value per band per 20 ms. Spectrograms for an example signal are depicted in Fig. 7.

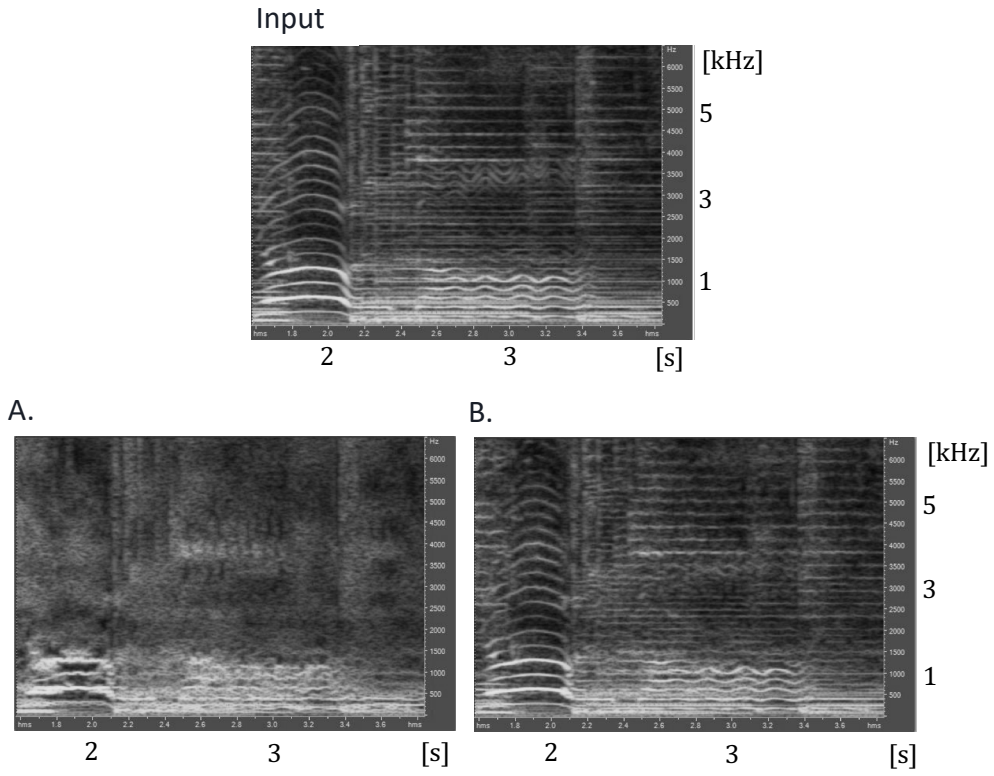

Fig. 7: Spectrograms of input and synthesis from (A) baseline and (B) extended representation for a segment of singing over guitar.

DISCUSSION

The example study shows that informal listening to the outputs of a mock-up codec for general audio provides immediate guidance on the construction of audio representations with the ambition to capture perceptually relevant aspects of audio. A MUSHRA test could be used to verify the shortcomings of the baseline representation (A) relative to the extended representation (B) whose own shortcomings would then also be revealed. For example, we expect both representations to fail for aspects requiring analysis of cross-frequency band coherence of TFS. We believe that the proposed method could be used in the construction and refinement of more well-developed models of TFS aspects of hearing, as well as improved predictors of subjective quality for pairs of perceptually similar signals as those available in the samples link of Klejsa *et al.* (2019).

REFERENCES

Brandenburg, K., Faller, C., Herre, J., Johnston, J. D., and Kleijn, W. B., (**2013**). "Perceptual Coding of High-Quality Digital Audio," Proc. IEEE, **101**, 1905 - 1919, doi: 10.1109/JPROC.2013.2263371

BS.1116 (**2015**). "Methods for the subjective assessment of small impairments in audio systems," Recommendation ITU-R BS.1116-3. Retrieved from: https://www.itu.int/rec/R-REC-BS.1116/en

BS.1387 (**2001**). "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, https://www.itu.int/rec/R-REC-BS.1387

BS.1534 (**2015**). "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3. Retrieved from: https://www.itu.int/rec/R-REC-BS.1534

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," J. Acous. Soc. of Am., **99**, 3615-3622, doi: 10.1121/1.414959

Decorsière, R., Søndergaard, P. L., MacDonald, E. N., and Dau, T., (**2015**). "Inversion of Auditory Spectrograms, Traditional Spectrograms, and Other Envelope Representations," IEEE-ACM T. Audio Spe., **23**, 46-56, doi: 10.1109/TASLP.2014.2367821

Herre, J., and Dick, S., (**2019**), "Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review," Appl. Sci., **9**, 2854, doi: 10.3390/app9142854

McDermott, J. H., Oxenham, A. J., and Simoncelli, E. P., (**2009**). "Sound texture synthesis via filter statistics," IEEE WASPAA, 297-300 doi:10.1109/aspaa.2009.5346467

Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am., **102**, 1811-20, doi:10.1121/1.420088

Moore, B., (**2019**). "The roles of temporal envelope and fine structure information in auditory perception," Acoust. Sci. Technol., **40**, 61-83, doi: 10.1250/ast.40.61

Kleijn, W. B., Lim, F. S. C., Luebs, A., Skoglund, J., Stimberg, F., Wang Q., and Walters, T. C., (**2018**). "Wavenet Based Low Rate Speech Coding," IEEE ICASSP, 676-680, doi: 10.1109/icassp.2018.8462529

Klejsa, J., Hedelin, P., Zhou, C., Fejgin, R., and Villemoes, L., (**2019**). "High-quality Speech Coding with Sample RNN," IEEE ICASSP, 7155-7159. doi: 10.1109/icassp.2019.8682435 (Samples retrieved from: https://sigport.org/documents/high-quality-speech-coding-sample-rnn)

P.863 (**2018**). "Perceptual objective listening quality prediction," Recommendation ITU-T P.863. Retrieved from: https://www.itu.int/rec/T-REC-P.863

Slaney, M. (**1995**). "Pattern playback from 1950 to 1995," Proc. IEEE Int. Conf. Syst. Man. Cybern., **4**, 3519-3524, doi:10.1109/icsmc.1995.538332