LEARNING ABOUT PERCEPTION OF TEMPORAL FINE STRUCTURE BY BUILDING AUDIO CODECS

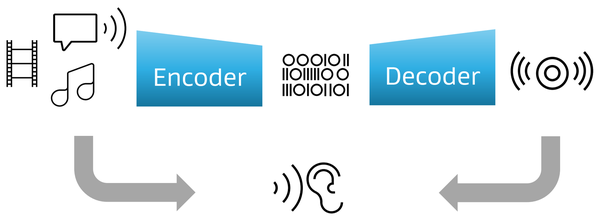
Lars Villemoes[1], Arijit Biswas[2], Heidi-Maria Lehtonen[1], Heiko Purnhagen[1]
[1]Dolby Sweden AB, Stockholm, Sweden, [2]Dolby Germany GmbH, Nürnberg, Germany

Introduction

Coarse scale properties of audio signals are cheaper to describe than temporal fine structure (TFS) [1]. This is exploited in modern audio coding systems. But which aspects of TFS are important to make two signals sound the same to us ? We

- walk through current and emerging audio coding methods
- suggest an audio coding inspired methodology to improve perceptual modeling and illustrate it by an example

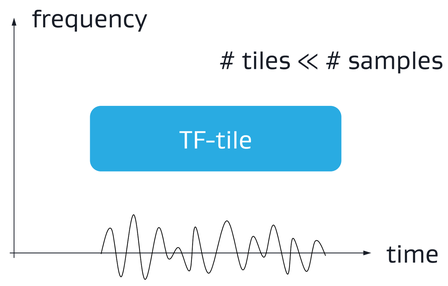
Audio coding



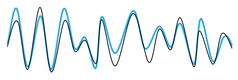
Goals

- Keep information rate low
- Convey auditory experience faithfully

Currently employed tools

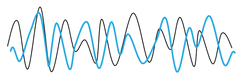


Waveform match



- Use masking principles to make the approximation error inaudible (e.g. mp3)
- Expensive but can keep all TFS

Noise fill

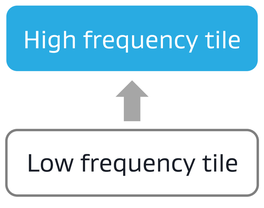


- Replace with random noise of same energy
- Very cheap but rarely offers high quality

Parametric coding

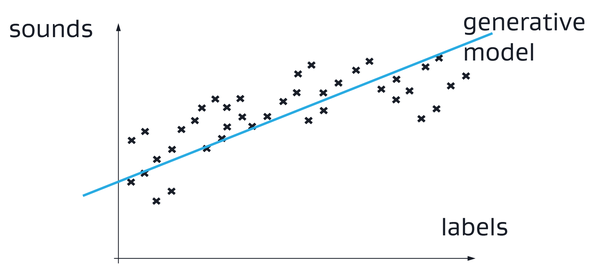
- Sum of sinusoids, noise, prototype transients, etc..

Spectral extension



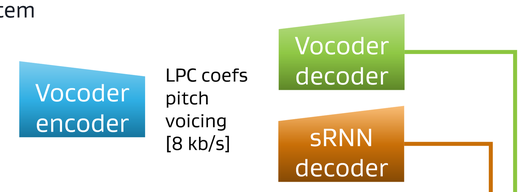
- Copy TFS from lower frequencies
- Adjust tonal to noise ratio with parametric methods
- Cheap and works surprisingly well

Machine learning tools

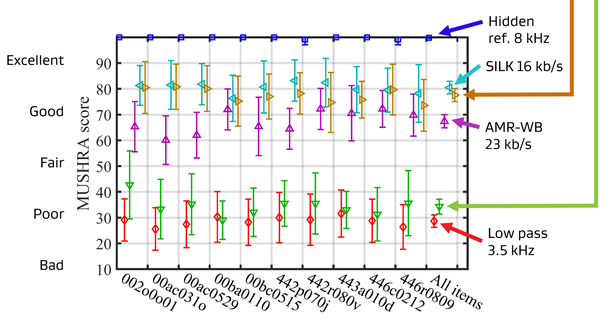


Speech coding example [2]

- System



- MUSHRA (BS.1534) test results



- Standardized predictors of subjective quality rank sRNN below the vocoder, e.g. MOS-LQO (POLQA)

SILK	sRNN	AMR-WB	Vocoder
4.41	3.48	4.39	3.67

Problem

A gap in our understanding

- Audio coding sometimes works better than predicted by measures like PEAQ and POLQA, especially when exploiting mechanisms beyond masking
- A hypothesis is that TFS aspects are central

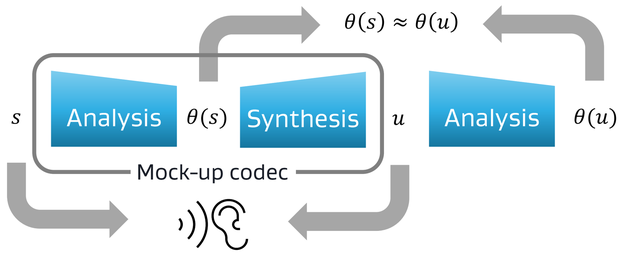
Better auditory representations could

- increase efficiency of audio coding
- advance our understanding of the role of TFS in hearing

Proposed method

Evaluate auditory representations by building a mock-up codec

1. Use synthesis by analysis [3] to generate sounds from a given auditory representation $\theta(s)$

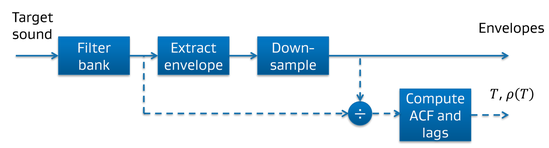


2. Run **general audio** through the system
3. Evaluate as for an audio codec
4. Do we agree with the machine listener?

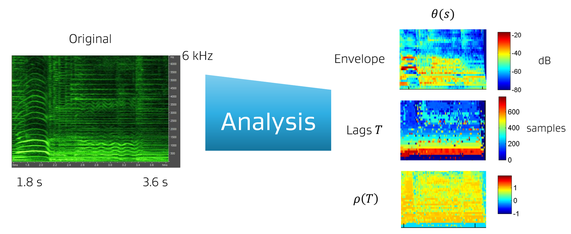
Example experiment

Comparing two auditory representations based on nonstationary audio textures

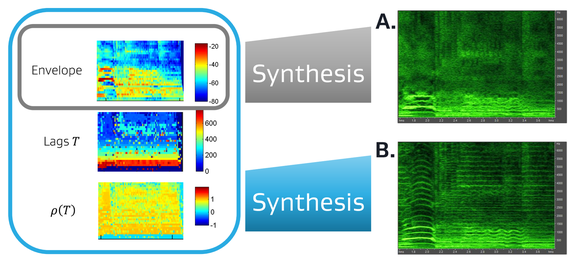
- Use the front end of [4] with 38 bands



A. **Baseline**: measure envelopes every 2.5 ms
B. **Extension**: add a lag T and the value $\rho(T)$ of the autocorrelation function (ACF) per band every 20 ms



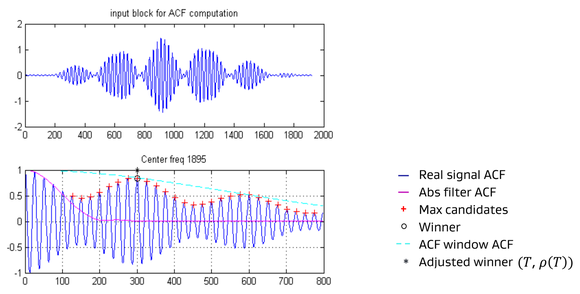
- Synthesis based on iterations and gradient descent



Informal listening reveals that

- the envelope description alone is not sufficient for tonality
- adding one lag and the corresponding ACF value per band per 20 ms gives a dramatic improvement

Details of lag search for the bandpass signals



Conclusion

- Building a mock-up codec based on a candidate auditory representation is a stress test providing immediate insights
- Such experiments could be key to understanding perception of TFS

References

[1] B. Moore, "The roles of temporal envelope and fine structure information in auditory perception". Acoustical Science and Technology. 40. 61-83. (2019).

[2] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality Speech Coding with Sample RNN", IEEE ICASSP 2019 pp. 7155-7159. (Demo samples: https://sigport.org/documents/high-quality-speech-coding-sample-rnn)

[3] M. Slaney, "Pattern playback from 1950 to 1995", Proc. IEEE Int. Conf. on Systems, Man and Cybernetics (1995), vol 4, pp. 3519-3524

[4] J. H. McDermott, A. J. Oxenham and E. P. Simoncelli, "Sound texture synthesis via filter statistics", IEEE WASPAA 2009, pp. 297-300