The Constant-Q Transform Spectral Envelope Coefficients: A Timbre Feature Designed for Music

I. Scope

TIMBRE is the attribute of sound which makes, for example, two musical instruments playing the same note sound different. It is generally associated with the spectral (but also temporal) envelope and is typically assumed to be independent from the pitch (but also the loudness) of the sound [1]. In this article, we will show how to derive a simple but well-founded timbre feature from the constant-Q transform (CQT), a log-scaled frequency transform that is well-adapted to musical data [3], [4]. We will show how to decompose the CQT into an energy-normalized pitch component and a pitch-invariant spectral envelope, from which we will extract a number of meaningful coefficients. We will then compare these CQT spectral envelope coefficients (CQT-SEC) with the mel-frequency cepstral coefficients (MFCC) [2], a feature originally designed for speech recognition but liberally used to characterize timbre in music, on the NSynth dataset, a large-scale and publicly-available dataset of musical notes [5].

II. Relevance

III. Prerequisites

Basic knowledge of audio signal processing and music information retrieval is required to understand this article, in particular, concepts such as the Fourier transform, convolution, spectral envelope, pitch, CQT, and MFCC.

IV. Problem Statement

V. Solution

convolution theorem: [6].

A. Observations

Assumption: A log-spectrum, such as the CQT-spectrum, can be represented as the convolution of a pitch-invariant log-specrtal envelope component (= timbre) and a envelope-normalized pitch component.

- A pitch change in the audio translates to a linear shift in the log-spectrum.
- The Fourier transform (FT) of a convolution of two functions is equal to the point-wise product of their FTs (convolution theorem).
- The magnitude FT is shift-invariant.

VI. Numerical Example

VII. What We Have Learned

We have shown that ...

VIII. Author

Zafar Rafii (zafarrafii@gmail.com) received a PhD in Electrical Engineering and Computer Science from Northwestern University in 2014, and an MS in Electrical Engineering from both Ecole Nationale Superieure de l'Electronique et de ses Applications in France and Illinois Institute of Technology in the US in 2006. He is currently a senior research engineer at Gracenote in the US. He also worked as a research engineer at Audionamix in France. His research interests are centered on audio analysis, somewhere between signal processing, machine learning, and cognitive science, with a predilection for source separation and audio identification.

References

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 2004.

[2] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[3] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[4] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.

[5] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *34th International Conference on Machine Learning*, Sydney, NSW, Australia, August 6-11 2017.

[6] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, 1995.