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 design a simple but practical pitch-independent timbre feature which is well-adapted to musical data, by deriving it from the constant-Q transform (CQT) [2], [3], a log-scaled frequency transform which matches the notes of the Western music scale. We will show how to decompose the CQT spectrum into an energy-normalized pitch component and a pitch-independent spectral envelope, the latter from which we will extract a number of timbral coefficients. We will then evaluate the discriminative power of these CQT spectral envelope coefficients (CQT-SEC) on the NSynth dataset [4], a large-scale dataset of musical notes which is publicly available, comparing them with the mel-frequency cepstral coefficients (MFCCs) [5], features originally designed for speech recognition but commonly used to characterize timbre in music.

II. Relevance

A timbre feature which is well-adapted to musical data, pitch-independent, and with high discriminative power can find uses in a number of applications, such as similarity detection, sound recognition, and audio classification, in particular, of musical instruments. Additionally, the ability to decompose the spectrum of a sound (here, the CQT spectrum) into a pitch-independent spectral envelope and an energy-normalized pitch component can be useful for audio analysis, transformation, and resynthesis. The energy-normalized pitch component can also potentially be used for pitch identification and melody extraction.

III. Prerequisites

Basic knowledge of audio signal processing and some knowledge of music information retrieval (MIR) are required to understand this article, in particular, concepts such as the Fourier transform (FT), convolution, spectral envelope, pitch, CQT, and MFCCs.

IV. Problem Statement

The multidimensional nature of timbre makes it an attribute that is tricky to quantify in terms of one single characteristic feature [6]. While it is assumed to be independent from pitch and loudness, it is not really feasible to fully disentangle timbre from those qualities, as timbre is inherently dependent on the spectral content of the sound, which is also defined by its pitch and loudness [1]. While researchers in MIR proposed a number of descriptors to characterize one or more aspects of timbre [7], they mostly adopted the MFCCs as the go-to single timbre feature.

[4]

V. Solution

We start with the assumption that a log-spectrum X, in particular, the CQT spectrum, can be represented as the convolution of an energy-normalized pitch component P (which mostly contains the pitch information) and a pitch-independent spectral envelope component E (which mostly contains the timbre information), as shown in Equation 1.

$$X = E * P \tag{1}$$

Property 1: A pitch change in the audio translates to a linear shift in the log-spectrum [2], [3].

Assuming that pitch and timbre are independent, this implies that the same musical object at different pitches would have a similar envelope component but a shifted pitch component (while two different musical objects at the same pitch would have different envelope components but a similar pitch component). Assuming X', E' and P' as the log-spectrum, envelope, and pitch component of the

Equation 2.

$$\begin{aligned} X &= E * P \\ X' &= E' * P' \\ \Rightarrow E &\approx E' \end{aligned} \tag{2}$$

We show in a second property that the FT of a convolution of two functions is equal to the point-wise product of their FTs, also known as the convolution theorem [8].

Equation 3.

$$\begin{aligned} \mathcal{F}(X) &= \mathcal{F}(E) \cdot \mathcal{F}(P) \\ \mathcal{F}(X') &= \mathcal{F}(E') \cdot \mathcal{F}(P') \\ \Rightarrow \mathcal{F}(E) &\approx \mathcal{F}(E') \end{aligned} \tag{3}$$

in a third property that the magnitude FT is shift-invariant [8].

Equation 4.

$$\begin{aligned} \mathcal{F}(X) &= |\mathcal{F}(X)| \cdot e^{jArg(\mathcal{F}(X))} \\ \mathcal{F}(X) &= |\mathcal{F}(X')| \cdot e^{jArg(\mathcal{F}(X'))} \\ \Rightarrow |\mathcal{F}(X)| &\approx |\mathcal{F}(X')| \end{aligned} \tag{4}$$

Given the ...,

Equation 5.

$$\begin{aligned} &\mathcal{F}(E) \approx |\mathcal{F}(X)| \\ &\Rightarrow E \approx \mathcal{F}^{-1}(|\mathcal{F}(X)|) \\ &\Rightarrow P \approx \mathcal{F}^{-1}(e^{jArg(\mathcal{F}(X))}) \end{aligned} \tag{5}$$

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, no. 5, p. 1270–1277, May 1977.

[7] G. Peeters, "Digital signal processing: Principles, algorithms and applications," IRCAM, Tech. Rep., 2003.

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