

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 functional 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 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) [6] are required to understand this article, in particular, concepts such as the Fourier transform (FT), convolution, spectral envelope, pitch, CQT, and MFCCs. In particular, more information about the CQT can be found in [2], [3].

IV. PROBLEM STATEMENT

The multidimensional nature of timbre makes it an attribute that is tricky to quantify in terms of one simple characteristic feature [7]. 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]. Researchers in MIR proposed a number of descriptors to characterize one or more aspects of timbre [8], but they mostly resort to using the MFCCs when they need one simple timbre feature [6]. While the MFCCs were shown to be helpful in some MIR tasks, they were initially designed for speech processing applications [5] and are not necessarily adapted to musical data. In particular, they are derived through an old process which makes use of the mel scale, a perceptual scale experimentally designed 80 years ago to approximate the human auditory system's response [9]. More recently, a number of data-driven approaches attempted to learn some timbral representations from musical data, but generally in terms of implicit embeddings which are tied to a specific trained model [4], [10], and not necessarily as explicit and interpretable features such as the MFCCs, which are still usually preferred as the go-to feature to characterize musical timbre by MIR practitioners.

V. SOLUTION

We propose the CQT-SECs, a novel timbre feature that is well-adapted to musical data, pitch-independent, simple to derive, interpretable, and functional. We will show how to derive it from the CQT, a log-frequency transform which matches the Western music scale, by first decomposing it into a pitch-independent spectral envelope and an energy-normalized pitch component, and then extracting a number of timbral coefficients from the spectral envelope.

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

$$X = E * P \quad (1)$$

We make a first observation that 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 but a shifted pitch component (while two different musical objects at the same pitch would have different envelopes but a similar pitch component). Assuming X , E , P and X' , E' , P' as the log-spectrum, envelope, and pitch component of a musical object and a pitch-shifted version of it, respectively, this is summarized in equation 2.

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

We make a second observation 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 [11]. This implies that the FT of the log-spectrum is equal to the point-wise product of the FT of the envelope and the FT of the pitch component. Given the previous observation that the envelope is not affected by a pitch change in the audio, we conclude that the FT of the envelope is also independent from the pitch. Assuming that $\mathcal{F}(\cdot)$ is the FT function, this is summarized in equation 3.

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

We make a final and third observation that the magnitude FT is shift-invariant [11]. ...

Equation 4.

$$\begin{aligned}\mathcal{F}(X) &= |\mathcal{F}(X)| \cdot e^{j\text{Arg}(\mathcal{F}(X))} \\ \mathcal{F}(X) &= |\mathcal{F}(X')| \cdot e^{j\text{Arg}(\mathcal{F}(X'))} \\ \Rightarrow |\mathcal{F}(X)| &\approx |\mathcal{F}(X')|\end{aligned}\quad (4)$$

Given the ...,

Equation 5.

$$\begin{aligned}\mathcal{F}(E) &\approx |\mathcal{F}(X)| \\ \Rightarrow E &\approx \mathcal{F}^{-1}(|\mathcal{F}(X)|) \\ \Rightarrow P &\approx \mathcal{F}^{-1}(e^{j\text{Arg}(\mathcal{F}(X))})\end{aligned}\quad (5)$$

VI. NUMERICAL EXAMPLE

VII. WHAT WE HAVE LEARNED

We have shown that ...

VIII. AUTHOR

Zafar Rafii (zafarrafi@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 Supérieure 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] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

- [6] M. Müller, *Information Retrieval for Music and Motion*. Academic Press, Springer.
- [7] 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.
- [8] G. Peeters, "The timbre toolbox: Extracting audio descriptors from musical signals," *Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, May 2011.
- [9] S. S. Stevens, J. Volkman, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.
- [10] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *25th European Signal Processing Conference*, Kos, Greece, August 28–September 2 2017.
- [11] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, 1995.