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 pitch-independent timbre feature which is well-adapted to musical data, by deriving it from the constant-Q transform (CQT) [2], [3], a log-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 analysis, transformation, and resynthesis of music signals. The energy-normalized pitch component can also potentially be used for tasks such as pitch identification, melody extraction, and chord recognition.

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. More information about the CQT can also 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 well-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 85 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 here the CQT-SECs, a novel timbre feature which is well-adapted to musical data, pitch-independent, simple to compute, interpretable, and functional. We will show how to derive it from the CQT, a frequency transform with a logarithmic resolution which matches the notes of Western music scale [2], [3], by first decomposing the CQT log-spectrum into a pitch-independent spectral envelope and an energy-normalized pitch component, and then extracting a number of timbral coefficients from the spectral envelope.

A. Deconvolution of the CQT

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 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.

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

This convolution process can be thought of a source-filter model [11] which is here not applied in the time domain but in the frequency domain, with the source and the filter being the pitch component and the envelope, respectively.

Observation 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 but a shifted pitch component (while two different musical objects at the same pitch would have different

envelopes but a similar pitch component). This is summarized in Equation 2, where X, E, P and X', E', P' represent the log-spectrum, envelope, and pitch component for a musical object and for a pitch-shifted version of the same musical object, respectively.

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

Observation 2: the FT of the convolution of two functions is equal to the point-wise product of the FTs of the two functions, a property also known as the convolution theorem [12].

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 first observation, this further implies that the FT of the envelope for a musical object and for a pitch-shifted version of it would be equal. This is summarized in Equation 3, where $\mathcal{F}(.)$ represents the FT function.

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

Observation 3: The magnitude FT is shift-invariant [12].

This implies that the magnitude of the FT of the log-spectrum for a musical object and for a pitch-shifted version of it would be equal. This is summarized in Equation 4, where $|.|$ and $Arg(.)$ represent the modulus and argument, respectively, for a complex array.

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

Given the previous observations, we can therefore conclude that the FT of the envelope could be approximated by the magnitude of the FT of the log-spectrum, while the FT of the pitch component could be approximated by the phase component. This finally gives us the estimates for the envelope and the pitch component, after taking their inverse FTs, as shown in Equation 5, where $\mathcal{F}^{-1}(.)$ represents the inverse FT function.

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

Figure 1 shows and example of deconvolution of the (power) CQT-spectrogram into its envelope and pitch component. The CQT-spectogram was computed from a 48-second audio signal created by concatenating 12 4-second acoustic bass notes, playing successively from C1 (32.70 Hz) to B1 (61.74 Hz). The notes come from the NSynth dataset [4], for instrument `bass_acoustic_000`, MIDI numbers `024` to `035`, and velocity number `075`. As we can see, the envelope gets pitch-normalized, shifting all the frequency components to the origin of the log-spectrum (essentially bringing all the notes to C1), while the pitch component gets energy-normalized, leaving mostly the fundamental frequency of the notes.

This deconvolution process can also be thought of the normalization of the log-spectrum by the magnitude of its FT (which here would correspond to the FT of the envelope) leading to a sharper log-spectrum (which here would correspond to the pitch component), in the manner of the generalized cross-correlation phase transform (GCC-PHAT) method which aims at normalizing a cross-correlation function by its magnitude spectrum to sharpen the cross-correlation peaks [13].

B. Extraction of the spectral envelope coefficients

Once the CQT-spectrum has been decomposed into a spectral envelope and a pitch component, we can then assumed that most of the pitch information has been removed from the envelope which now mostly contains the timbre information. We can think of the spectral envelope as a pitch-normalized log-spectrum where the frequency components have been shifted down to the origin of the spectrum.

Equation 6 O_r the octave resolution, or the number of frequency channels per octave (typically, 12, 24, 36, etc.), N_c the number of coefficients

$$i = O_r \log_2(k) \text{ for } 1 \leq k \leq N_c \tag{6}$$

Briefly explain what how the MFCCs are computed [5] mel scale: [9]

VI. Numerical Examples

A. Analysis on an Example

B. Comparison on a Dataset

VII. What We Have Learned

We have shown that we can derive a simple but function timbre feature which is more adapted to musical data ...

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

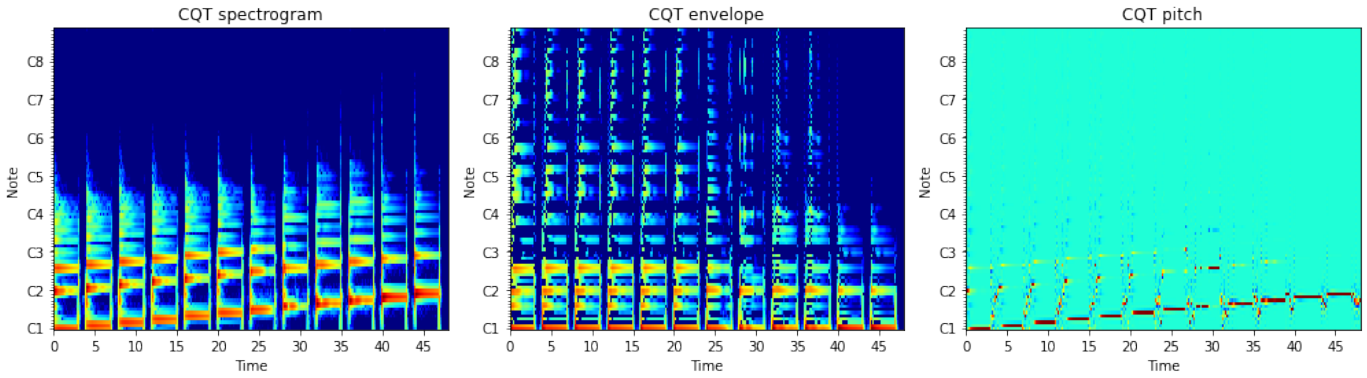

Fig. 1. Deconvolution of the (power) CQT-spectrogram of 12 acoustic bass notes playing from C1 to B1, into a pitch-independent envelope and an energy-normalized pitch component.

[6] M. Müller, *Information Retrieval for Music and Motion*. Springer, 2007.

[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. Volkmann, 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] G. Fant, *Acoustic Theory of Speech Production*. Mouton De Gruyter, 1970.

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

[13] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.