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 typically associated with the spectral (but also the temporal) envelope and 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), a log-frequency transform which matches the equal tempered musical scale [2], [3]. 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 practical 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 typically in terms of implicit embeddings which are tied to a specific trained model [4], [10], [11], [12], [13] and not necessarily as explicit and interpretable features such as the MFCCs, which are still usually preferred as the go-to feature to characterize timbre by MIR practitioners.

V. Solution

We present 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 the equal temperament, a musical scale typically used in Western music [2], [3]. We will first show how to decompose the CQT spectrum into a pitch-independent spectral envelope and an energy-normalized pitch component, and then extract a number of timbral coefficients from the spectral envelope.

A. Deconvolution of the CQT

We start with the assumption that a log spectrum S, 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, where $*$ represents the convolution operation.

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

This convolution process can be thought of as a source-filter model [14] 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 S, E, P and S', 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.

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

Observation 2: the FT of the convolution between two functions is equal to the pointwise product between the FTs of the two functions, a property known as the convolution theorem [15].

This implies that the FT of the log spectrum is equal to the pointwise product between 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 and $\cdot$ the pointwise product.

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

Observation 3: The magnitude FT is shift-invariant [15].

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, and j, the imaginary unit.

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

Given the previous observations, we can therefore conclude that the FT of the envelope could be approximated by the magnitude 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.

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

Figure 1 shows an example of deconvolution of a CQT spectrogram into its envelope and pitch component. The CQT spectogram was computed from an audio signal created by concatenating 12 4-second notes of an acoustic bass playing from C1 (32.70 Hz) to B1 (61.74 Hz) in ascending order. The notes come from the NSynth dataset [4] and correspond to instrument id `bass_acoustic_000`, MIDI numbers `024` to `035`, and velocity number `075`. The CQT spectrogram was computed using librosa [16], [17], with a sampling rate of 16 kHz, a hop length of 512 samples, a minimum frequency of 32.70 Hz (corresponding to C1), 95 frequency bins, and 12 bins per octave. As we can see, the envelope looks as if the CQT spectrogram has been normalized in pitch, with all the notes being brought down to the lowest frequency (corresponding to C1); while the pitch component looks as if the CQT spectrogram has been stripped down from all its energy, leaving mostly the fundamental frequencies of the notes. Note that in practice, we use a power CQT spectrogram (i.e., magnitude to the power of 2) and take the real part of the envelope and pitch component to ensure real values. We can also potentially post-process this deconvolution, for example, by zeroing the few negative values in the pitch component and using it to derive a refined envelope from the log spectrum.

This deconvolution process can also be thought of as 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 [18].

B. Extraction of the spectral envelope coefficients

The envelope resulting from the deconvolution of the CQT spectrum can be thought of as a pitch-normalized CQT spectrum where the spectral content, in particular, the harmonics which represent most of the energy of the musical instrument, has been essentially brought down to the same lowest note level. Given the octave resolution which was used when computing the CQT spectrum, i.e., the number of bins per octave, we can then easily infer the locations of those harmonics in the envelope [2], [3]. We can subsequently extract these harmonics or *spectral envelope coefficients* from the envelope and thus obtain a compact and interpretable feature for characterizing the timbre of the musical instrument. Equation 6 shows how to derive the indices of the spectral envelope coefficients given O_r, the octave resolution, and N_c, the number of desired coefficients, and finally extract the CQT-SECs from the envelope E, with $\log_2(.)$ and $round(.)$ representing the binary logarithm and the round function, respectively.

$$\begin{cases} i = round(O_r \log_2(k)) \\ \text{CQT-SEC}_k = E(i) \end{cases} \quad 1 \leq k \leq N_c \tag{6}$$

Figure 2 shows an example of CQT-SECs, on the left plot. 20 coefficients were extracted from the envelope resulting from the deconvolution of the CQT spectrogram of the musical signal shown in Figure 1. These coefficients essentially correspond to the harmonics of the musical instrument which contain most of its spectral energy and can therefore be a reasonable representation of its timbre. For comparison, we also show the MFCCs computed from the same musical signal, on the right plot. 20 coefficients were computed using librosa [16], with a sampling rate of 16 kHz, a window length of 1024 samples, and hop length of 512 samples, matching the

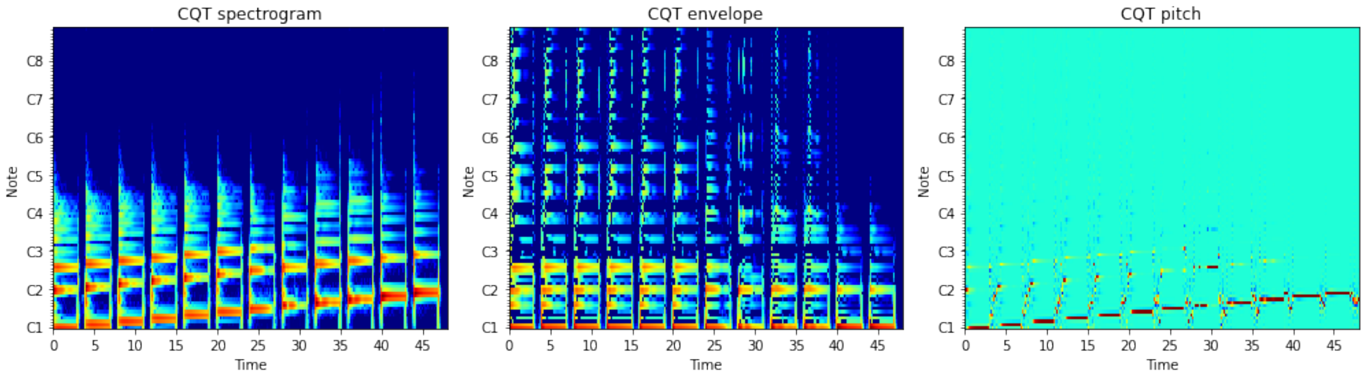

Fig. 1. Deconvolution of the CQT spectrogram (left plot, shown in dB) of 12 acoustic bass notes playing from C1 to B1, into a pitch-independent spectral envelope (middle plot, shown in dB) and an energy-normalized pitch component (right plot, shown in $[0, 1]$).

time resolution of the CQT-SECs so that both features have the same size in time and frequency.

We recall below how the MFCCs are commonly derived. First, a classic spectrogram is computed from the audio signal using the short-time Fourier transform (STFT) and the powers of the frequencies in each frame are then mapped onto the mel scale [9] using triangular overlapping windows. This essentially produces a mel spectrogram. Then, the log of every mel frequency is taken, followed by the discrete cosine transform (DCT) of each frame. Finally, the MFCCs are selected as the lowest coefficients in the resulting spectrum, excluding the very first one (DC component). This process is meant to decouple the envelope from the pitch in the time domain, by extracting the slow-varying time components, which most likely correspond to the envelope and which will become the lowest coefficients, from the fast-varying time components, which most likely correspond to the pitch and which will become the highest coefficients [5].

VI. A Computational Example

We propose to measure the discriminative power of the CQT-SECs by evaluating them on the NSynth dataset, a large-scale and high-quality dataset of annotated musical notes which is publicly available[1] [4]. The NSynth dataset is composed of 305,979 musical notes which were generated from 1,006 instruments as 4-second, monophonic audio signals at a sampling rate of 16 kHz, with pitches ranging over all the note numbers of a standard MIDI piano (21-108) and with 5 different velocities (25, 50, 75, 100, and 127), whenever applicable. The instruments are organized into 11 families (bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, synth_lead, and vocal) and 3 sources (acoustic, electronic, and synthetic). We chose to evaluate the CQT-SECs on the notes with a velocity of 75 only, leading to 60,388 different notes for 945 different instruments.

We derived the CQT-SECs for all the notes in this subset, by first computing the CQT spectrogram using librosa, with a sampling rate of 16 kHz, a hop length of 512 samples, a minimum frequency of 32.70 Hz (corresponding to C1), 95 frequency bins, and 12 bins per octave, and then extracting 20 coefficients from the envelope resulting from the deconvolution of the CQT spectrogram, leading to CQT-SECs of size 20 coefficients and 126 time frames. We used a power CQT spectrogram (i.e., magnitude to the power of 2) and took the real part of the envelope to ensure real values. For comparison, we also computed the MFCCs, using librosa, with a sampling rate of 16 kHz, a window length of 1024 samples, a hop length of 512 samples, and 20 coefficients, leading to MFCCs of size 20 coefficients and 126 time frames, matching the size of the CQT-SECs.

We computed the cosine similarity for every pair of CQT-SECs and every pair of MFCCs (without repetition), after flattening the features into one-dimensional vectors of length 2520 (20 times 126). We then averaged these note similarities over every instrument, leading to similarity matrices of size 945 by 945, for both the CQT-SECs and the MFCCs. Figure 5 shows the similarity matrices derived for all the pairs of instruments for the CQT-SECs and the MFCCs. As we can see, the similarities for the CQT-SECs have more variance, while the similarities for the MFCCs are mostly very high (close to 1) showing poor discriminative power. Figure 4 shows the self-similarities (in green) (i.e., the diagonal of the similarity matrix) and the error bars for the cross-similarities (i.e., means in red and standard deviations in yellow) for every instrument, for the CQT-SECs and the MFCCs. As we can see, the self-similarities for the CQT-SECs are noticeably higher that the mean cross-similarities for most of the instruments, and generally higher that the errors too, showing good discriminative power, while the self-similarities and cross-similarities for the MFCCs are all very high (close to 1).

We also averaged the note similarities over every instrument family, leading to similarity matrices of size 11 by 11, for both the CQT-SECs and the MFCCs.

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

[1]https://magenta.tensorflow.org/datasets/nsynth

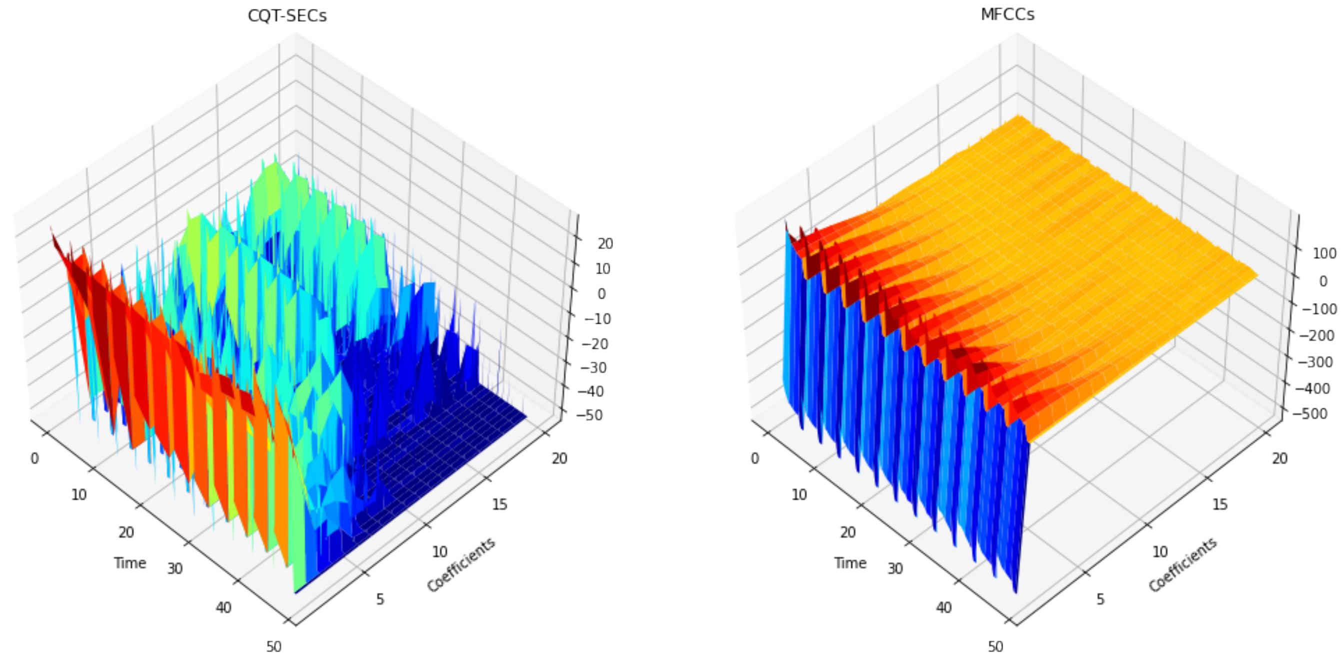

Fig. 2. CQT-SECs extracted from the envelope obtained following the deconvolution shown in Figure 1 (left plot, shown in dB) and MFCCs computed from the same musical signal (right plot).

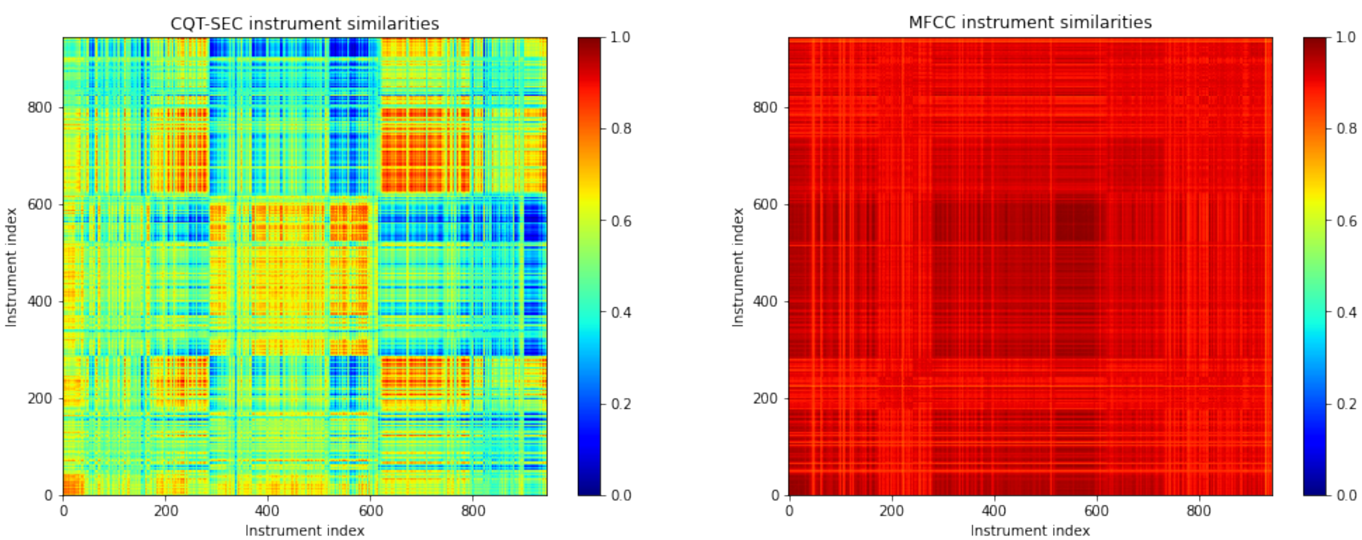

Fig. 3. Similarity matrices computed for every pair of notes and averaged over every instrument of the NSynth subset, for the CQT-SECs (left plot) and the MFCCs (right plot).

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 research engineer manager 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. 4. Self-similarities and cross-similarities derived from the similarity matrices shown in Figure 5, for the CQT-SECs (top plot) and the MFCCs (bottom plot).

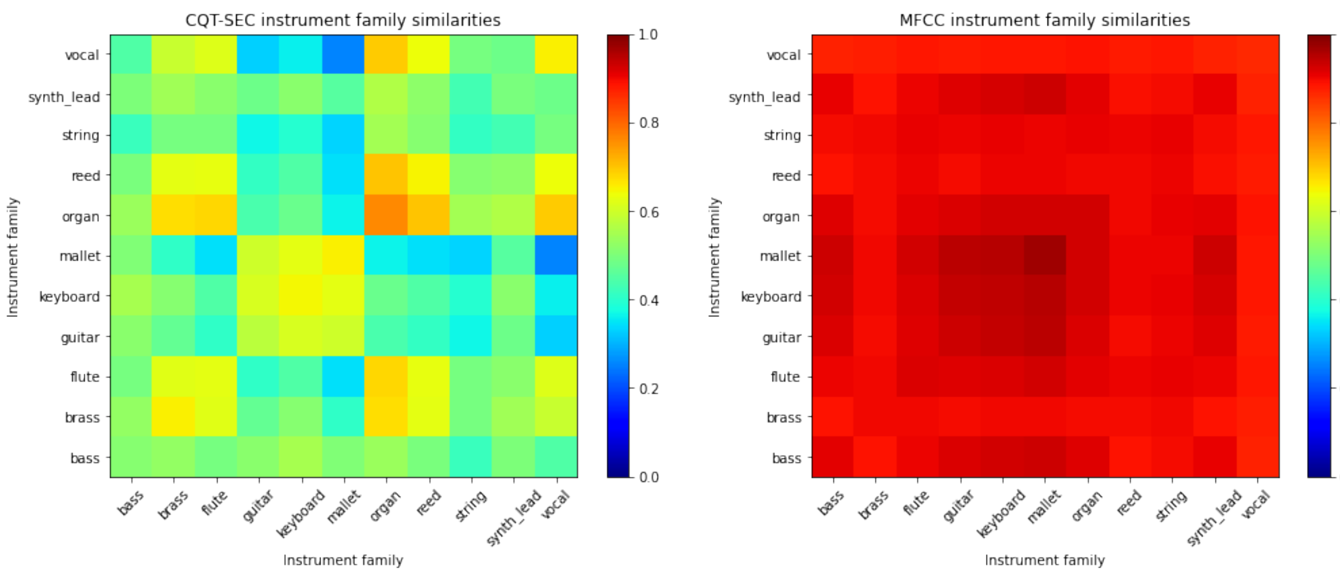

Fig. 5. Blah.

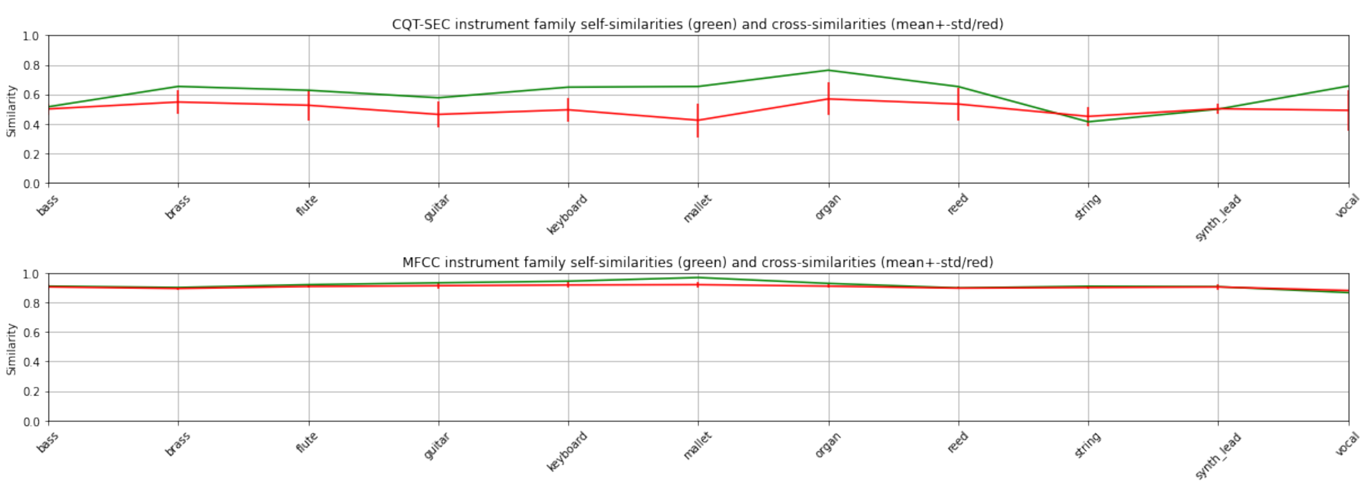

Fig. 6. Blah.

[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] Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang, "Musical composition style transfer via disentangled timbre representations," in *28th International Joint Conference on Artificial Intelligence*, Macao, China, August 10-16 2019.

[12] Y.-J. Luo, K. Agres, and D. Herremans, "Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders," in *20th Conference of the International Society for Music Information Retrieval*, Delft, The Netherlands, November 4-8 2019.

[13] J. Lee, H.-S. Choi, J. Koo, and K. Lee, "Disentangling timbre and singing style with multi-singer singing synthesis system," in *45th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 4-8 2020.

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

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

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *14th Python in Science Conference*, Austin, TX, USA, July 6-12 2015.

[17] C. Schoerkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference*, Barcelona, Spain, July 21-24 2010.

[18] 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.