Harmonic Structure Identification with the Cumulative Distribution Transform

Will Ashe Will Xiao

***Abstract*—Harmonic structure identification is a complex problem examining the relationship between perceived sound and the sound's frequency content across time. Harmony depends on consistent relationships in the frequency domain, but the variation in musical analysis presents a significant issue in decoding the harmony based on peak frequencies in a spectral transform. Complicating factors include recording noise, temporal and spectral variation, and resonance variations between instruments. The cumulative distribution transform offers an attractive paradigm for signal matching, where difficult derivations of signal domain mapping functions for positive, normalized signals become trivial optimization problems in the transform domain. The result is a mechanism for signal comparison that is light-weight while maintaining or improving accuracy against typical signal metrics like Euclidean distance. Here, the harmonic identification problem is considered using the cumulative distribution transform paradigm and two classifiers are presented, based on the CDT-nearest subspace and linear discriminant analysis methods.**

I. Introduction

Music is commonly broken down to several analytical components, notably including melody, harmony, rhythm, and timbre. In the digital age, these mammoth topics are being dissected using new datasets and information retrieval techniques. One example is the classification of musical chords, a typical exercise introduced in a beginner's musicianship class which serves as the building block for transcribing harmonies in real music. Due to the infinite potential combinations of musical notes, the actual structures used in even the simplest pop music can become complex. To make the problem tractable for musicians learning theory, the situation is typically restricted to a subset of all possible chords, allowing for easier differentiation. The converse implies the general problem is difficult - indeed, modern transcription annotations are closer to well-presented arguments than definitive prescriptions. The difficulty in transcription translates to additional time and reviewers needed to confirm that a given chord transcription is accurate. New techniques such as long short-term memory (LSTM) recurrent neural networks (RNN) and other transient deep learning allow for the embedding of state into the processing of a typical deep learning/neural net model, which appears promising for temporally-related data like musical signals. Unfortunately, these solutions require large amounts of data and time to train such a model that sees the complexity automatically.

The optimal transport paradigm offers a promising approach to handling certain pattern matching tasks in which signal classes are defined by transforming the domain of a function while leaving the shape equivalent to a scale factor [3]. By converting the data through a lossless function to an equivalent transform space, differences in scale and translation in the signal domain become directly computable, comparable, and usable in the classification of signal to underlying template classes. This paper examines musical signals as time-frequency functions that can be compared using an optimal transport perspective.

II. Background

Music is a temporal art form - the vast majority of symbolic and mathematical understandings of music rely on agreed upon units, termed notes, being played for defined intervals of time. The human understanding of the any sound is proportional to the frequency of the signal played, making music a signal embedded in the time-frequency domain. Fast, repeating sounds are detected as oscillating, while non-repeating sounds are perceived as noise, with the contained frequencies determining the overall sound perceived by a listener. When the sound is repeating, and the frequency falls between 20 Hz and 10 kHz, the ear perceives the sound as having a *pitch*. Pitch is a perceptual quality that allows a listener to decode whether a note is "higher" or "lower" than other notes in the frequency spectrum. Many systems for understanding this pitch spectrum exist, relying on the scalar relationship between different frequencies and some notable quality of sound. The vast majority of these systems organize the octave (i.e. the range between a frequency and its double) into significant subintervals or scales.

While the determination of which notes to be played for a given scale and the nuances between variants is a global topic better saved for another venue, the impressive conclusion is that the harmonic series of a note is widely agreed to be the basis for understandings of *consonance* and *dissonance*, the perceptual qualities of sounds fitting together or clashing. While any repeating signal at a given frequency will sound the same pitch, higher frequencies at scalar multiples of the original, *fundamental* frequency can coexist and change the perceived sound without affecting pitch. The second harmonic is the first octave, which forms the interval of perceptual repetition of pitches, and remaining powers of two harmonics form additional, higher octaves. The intervening frequencies determine the order of consonance for other harmonics; the closer to the original fundamental, the more consonant. Applying these two rules (multiple by power of two to get octave, and by other positive integers then divide by a power of two to get other scale pitches) generates a set of tones within the original octave. Western music commonly chooses twelve, as opposed to other systems that might use five or 24. There

is some complication about how the exact frequencies are determined (either starting from harmonic fractions or from an equal splitting of the octave in logarithmic space), but this discussion can be comfortably shelved, with the knowledge that any system expresses pitch frequencies as scalar multiples of a master frequency with decent consistency. In modern western classical music, this is chosen to be A_4 at 440 Hz (Stuttgard pitch), the standard concert tuning pitch.

There are an incredible number of tangents and qualifiers to the above, but the most pressing are the concepts of timbre and harmony. Timbre is defined as the perceptual quality of a sound *outside* of pitch, which includes the relationships of the harmonic series coefficients, non-pitch noise over time, time amplitude envelope (so-called attack, decay, sustain, release envelope), and glide and frequency oscillation (vibrato). Timbre is the quality that differentiates between instruments that would otherwise sound the same, caused by variations in instrument resonance and style of play. When a sound is converted into the frequency domain via a Fourier transform or similar, these timbre effects are partially explained the relationship of the fundamental to every other frequency. The temporal effects can be seen in a scaleogram from a wavelet transform. Any processing of sound has to understand that a pitched sound may have frequency content outside the harmonic series, and that all content (harmonic and otherwise) can change in time.

The understanding of harmony is equally important, as a single note played alone is rare. Chords (multiple notes played simultaneously) are built off the idea that if multiple frequencies can make a sound more interesting via the harmonic series, then combining specific pitches (and their harmonics) to either be consonant or dissonant can create a desirable effect. Raw sound synthesis is linear; that is, two analog sound waves can be added together to produce the equivalent sum of their two sounds. Instrumental amplification, however, is non-linear, with the characteristics of the instrument determining what frequencies are exaggerated or attenuated. Still, the resulting content of a chord is strikingly similar to the sum of a harmonic set repeated for each note in the chord, as any pitch content is still the result of a harmonic series. Further, since all notes in the scale are related by a scalar, the pitch content of a single type chord can roughly be described as a sequence of scalar-related frequencies. That is, to move a major chord up a half step, one would simply multiply every frequency in the original harmonic spectrum by the same scale ($2^{1/12}$ under equal temperament) [1].

Given the above, the cumulative distribution transform (CDT) seems an appropriate fit for the underlying problem. The CDT is defined as a function $f(x)$ that maps the domain of a probability density function p_1 to the domain of another p_2, such that integration of p_1 up to point x will map to the integration of p_2 up to domain point $f(x)$. Equivalently,

$$\int_{-\infty}^{x} p_1(u)du = \int_{-\infty}^{f(x)} p_2(u)du \tag{1}$$

Given a template $s_0 \in \mathcal{P}_1$ with antiderivative S_0, and a signal $s \in \mathcal{P}_1$ with antiderivative S, one can define the function $\hat{s} = f(x) = S^{-1}(S_0(x))$ and write $s_0(x) = f'(x)s(f(x))$. Below are important properties of the CDT, repeated from [2]

A. Composition Property of CDT

Let $s_g = g's \circ g$. Then,

$$\hat{s}_g = g^{-1} \circ \hat{s} \tag{2}$$

B. Translation Property of CDT

Let $g(x) = x - \mu$. Then,

$$\hat{s}_g = \hat{s} + \mu \tag{3}$$

C. Translation Property of CDT

Let $g(x) = \alpha x$. Then,

$$\hat{s}_g = \frac{\hat{s}}{\alpha} \tag{4}$$

III. Methods

A. CDT-Nearest Subspace Classifier

The CDT-Nearest Subspace (CDT-NS) classifier [3] offers a promising approach to the problem of classifying signals to match template classes, particularly those that are composed through input scaling, translation, and other domain compositions applied according to (2). The training algorithm applies the CDT to each positive, normalized signal in $\mathbb{R}^n$. The model then constructs a representative orthonormal set from the transformed data; these vectors are taken as the basis for a subspace in $\mathbb{R}^n$. After constructing a basis for each class, the subspaces can be enhanced under the composition theorems, as applicable to the problem. By inspection, such a basis would cover all scaled copies, automatically including all domain-scaled versions of the template signal. Additional transformations can include all shifted versions of the signal in the span of the subspace, and the method could be applied to other domain composition functions.

$$\hat{\mathcal{S}} = \{\hat{s}_j | \hat{s}_j = g_j^{-1} \circ \hat{\phi}, \forall g_j \in \mathcal{G}\} \tag{5}$$

If $\mathcal{G}^{-1}$ is chosen to be a convex set, then the transformed signal set $\hat{\mathcal{S}}$ is also convex. [3]

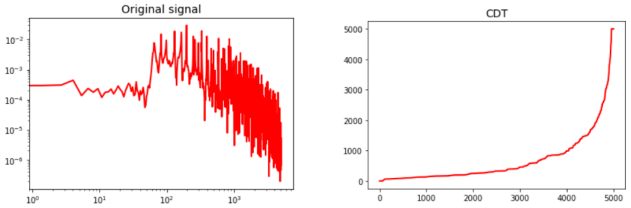


Fig. 1: CDT of a chord

B. Datasets

The primary dataset for our experiment is composed of more than 150 labeled chords with varying triad chord type (major, minor, diminished) and root note on different software instruments, such as piano, organ, or strings that were home-recorded via MIDI. Additional data collected include recorded, labeled strummed chords from early Beatles albums to enrich and add noise to our dataset that could be used to test in the future, once the classifiers achieve higher accuracy on less complex data.

The final dataset did not use all chord files, as some were pitched too high or low for the samples to process correctly during creation (muddy or missing content apparent as incorrect peak structures in the FFT).

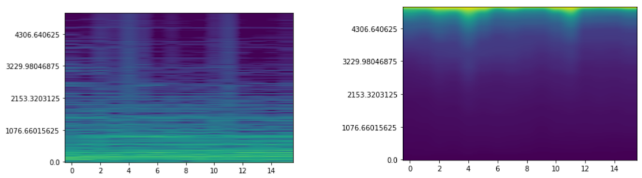
All chords are recorded at a sampling rate of 44.1kHz, the standard CD format, and have a two-second duration.

C. Preprocessing

The raw sound files are time-domain signals, and they needed to be processed into frequency signals in order for a CDT to extract frequency scaling. Originally, the discrete wavelet transform (DWT) and short-time Fourier transform (STFT) were considered due to their ability to separate frequencies in time, as opposed to considering the whole signal. To approximate the STFT, the signal was partitioned into subintervals of samples, and a fast Fourier transform (FFT) was run on each subvector. The magnitude of each complex FFT output was normalized to produce valid probability density functions, which were passed through a CDT. This process is demonstrated in Fig. 2. After some experimentation with the number of partitions, including relatively disastrous classification attempts with performance well below random guessing, the idea of time-domain classification was scrapped in favor of increasing the frequency resolution.

The relationships contained in the FFT data are complex. Pure frequencies appear as steep peaks, rather than perfect spikes, and the noise between peaks and below the fundamental presented significant problems for comparing two signals. Multiple strategies were implemented in an attempt to improve the results, including reducing spectral variation (limiting samples to a piano), expanding spectral variation (adding more instruments to increase harmonic color), changing the template signal s_0 used, and hardcoded reductions in spectral range (cutting the FFT at the same frequency on each chord). While each of these approximated a solution to a single issue, none of them accurately addressed the issues of varying timbre in decoding frequencies.

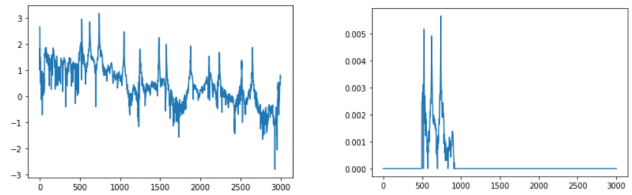
In an effort to make the problem more tractable for the relatively small amount of data, FFTs were cut such that the peak frequencies remaining represented the relationships between the fundamental frequencies of the notes and the total bandwidth of the frequency domain was limited to just higher than the highest harmonic (roughly 3500 Hz). To achieve the effect of harmonic filtering, the bass frequency of each chord was extracted from a predetermined label. Based on this note, a range of harmonics was selected, and frequencies outside the band were cut to zero amplitude. After experimentation, the second octave proved to be the most reliable in detecting three peaks with spacing corresponding to the chord's relationships.



(a) Spectrogram of a chord (b) TF-CDT of the chord

Fig. 2: Spectrogram and TF-CDT of a chord

To improve the consistency and quality of the resulting peaks, the log amplitude was taken, and the highest peak was set to log amplitude 2.5. The resulting signal was half-wave rectified, and resulting zero values were set to a small constant before the signal was re-normalized.



(a) Original signal (b) Filtered signal

Fig. 3: Harmonic filtering of a chord

IV. Results

Two classifiers were tested: the CDT-NS classifier and Linear Discriminant Analysis (LDA), using Leave-One-Out (L1O) cross validation on our dataset. Each classifier accepted the Time-Frequency CDT (TF-CDT) of the signals as input. Furthermore, seeking to improve the classifiers' accuracies, experiments were conducted to filter subharmonic content and higher order harmonics (overtones), chopping the chords' Fourier transform in the frequency space to limit the extent of the undesirable effects of these harmonic series, making the input signal more interpretable (described above). Classification is performed in the CDT space, deciding between major, minor, and diminished chords.

TABLE I: L10 CV Accuracy

	CDT-NS	LDA
Piano chords C_4 template	0.3157	0.1578
All chords C_4 template	0.5100	0.3624
All chords uniform template	0.5234	0.3959

Accuracy of the models is displayed in Table I. Further analysis of the confusion matrices produced by the L1O CV process are shown in Tables II through V. Interestingly, the classification profiles change drastically between CDT-NS and LDA with the same template, and between the two CDT-NS classifiers, but the LDA is consistent across the two. The sample size for this experiment (n = 148) is likely too small

TABLE II: Confusion Matrix, CDT-NS/C_4 Template

	Maj	Min	Dim
Maj	20	12	9
Min	14	25	6
Dim	18	14	31

TABLE III: Confusion Matrix, LDA/C_4 Template

	Maj	Min	Dim
Maj	26	17	20
Min	10	18	16
Dim	16	16	10

TABLE IV: Confusion Matrix, CDT-NS/Uniform Template

	Maj	Min	Dim
Maj	33	13	7
Min	6	21	15
Dim	13	17	24

TABLE V: Confusion Matrix, LDA/Uniform Template

	Maj	Min	Dim
Maj	27	12	17
Min	10	20	17
Dim	15	19	12

for meaningful conclusions, but the result is striking enough that it is worthy of additional investigation.

V. Discussion

The results show that while harmonic classification is possible under controlled circumstances, it remains difficult even when stripped down to the simplest experiments. In an effort to simplify, chords were reduced from time-frequency content to a representation in the frequency domain. The spectral range was limited, as powerful harmonics occur less frequently in the higher range. When these strategies were insufficient, more care was taken in modifying the amplitude spectrum to an appropriate shape reflecting an underlying template. To this end, the FFT amplitudes were filtered such that the only non-zero amplitudes fell in a single octave containing the highest peaks for the three notes. The peaks were taken to the log amplitude space, where differences between peaks were somewhat normalized, and the whole signal was shifted such that the highest peak of each was consistent, in an effort to control the amount of noise below the peaks. The result was a set of recognizable signals - both classifiers outperformed a random guess across datasets and template choices.

One interesting finding was that once the signals were regularized, the CDT-NS classifier consistently outperformed LDA by 5% or more. This is likely due to the built-in scaling potential that the subspace includes - for scaling signals, the CDT-NS should outperform LDA for scaled samples that it has never seen before, as they are within the span of the (reduced) training set. This hypothesis was confirmed by the present experiments. Additionally, two very different templates were used (a uniform distribution and the FFT of a C_4 piano note), which demonstrated that both were effective against the three-peak signals. The takeaway from the confusion matrices is that different templates identify different characteristics significantly. An study of ensemble approaches using multiple templates would be fascinating.

There are many limitations to the results given here. The FFT amplitudes were improved, most notably by controlling the peak frequencies to have similar amplitude and be structurally limited; all chords used in the training were root position and occurred in the same octave. Improving accuracy and expanding the structural scope would both require more than the roughly 50 samples per class used in the final training cycles. Still, the method of detecting scaled chords is fundamentally sound, and the ideas of this method could aid in future chord identifying experiments.

VI. Conclusion

In this study, the problem of harmonic structure identification was examined as an application of optimal mass transport and the cumulative distribution transform. Two classifers were able to decode the small amount of training data, though the CDT-NS demonstrated the most potential. Future studies would need to expand the dataset to include more chords and instruments, especially if working with additional chord types. A more generalized classifier would have to look at the dispersion of fundamentals and harmonics across the spectrum and deal with the disorganization of structure as notes are cycled and spread across different octaves. The present study also did not address bass or root note, which would both aid in the processing of real audio into a harmonic structure.

References

[1] M. Müller, D. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," in *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 5, NO. 6*, 2011.

[2] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde, "The cumulative distribution transform and linear pattern classifcation," in *Applied and Computational Harmonic Analysis*, 2015.

[3] M. Shifat-E-Rabbi, X. Yin, A. Rubaiyat, S. Li, S. Kolouri, A. Aldroubi, J. M. Nichols, and G. K. Rohde, "Radon cumulative distribution transform subspace modeling for image classification," 2020.