COMPRESSING MUSIC RECORDINGS INTO AUDIO SUMMARIES

Oriol Nieto
New York University
oriol@nyu.edu

Eric J. Humphrey
New York University
ejhumphrey@nyu.edu

Juan Pablo Bello
New York University
jpbello@nyu.edu

ABSTRACT

We present a criterion to generate audible summaries of music recordings that optimally explain a given track with mutually disjoint segments of itself. We represent audio as sequences of beat-synchronous harmonic features and use an exhaustive search to identify the best summary. To demonstrate the merit of this approach, we evaluate the criterion and show consistency across a collection of multiple recordings of different works. Finally, we present a fast algorithm that approximates the exhaustive search and allows us to automatically learn the hyperparameters of the algorithm for a given track.

1. INTRODUCTION

One of the classic motivations in the field of music informatics is facilitating the navigation of massive digital music collections by human users. Research in this area aims to develop computational methods of organizing and retrieving music recordings —tracks— in the spirit of reducing the amount of effort necessary to find desired content. Ultimately, the user must listen to any unfamiliar track to validate the search results, making the process considerably time consuming.

In digital music storefronts and other kinds of large collections, the traditional solution is to represent a full track with a single, identifiable excerpt. Known as audio thumbnailing, much effort has been invested into the development of automatic systems to these ends; for a partial review, we refer to [1, 2, 6, 7]. For some popular music that is highly repetitive in nature, these methods perform well in identifying useful thumbnails. Regardless, representing a full track with a single excerpt presents one unavoidable deficiency: the defining characteristics of a track are rarely concentrated in one specific section.

Recognizing this shortcoming, we motivate an alternative approach to classical thumbnailing that instead creates a short, listenable audio summary, capturing both the most unique and representative parts of a track. Specifically, this paper presents a novel audio summary criterion and an efficient method of automatically generating these summaries from real music recordings. The criterion is maximal for the set of segments that best explain the overall track while simultaneously exhibiting minimal overlap between them. Via examples and an experimental study we show how this measure yields good audio summaries. Furthermore, we show that it is possible to automatically select the optimal number and length of the selected subsequences specific to a given recording.

The remainder of this paper is organized as follows: Section 2 addresses the topic of feature representation. Section 3 defines the music summary criterion and showcases the measure in practice. Section 4 details a heuristic approximation to the exhaustive evaluation over the free parameters. Section 5 presents a systematic evaluation of the feature representation, heuristic solution and effect of automatically learning hyperparameters. Finally, we discuss our conclusions and observations for future work in Section 6.

2. FEATURE REPRESENTATION

The goal of developing an appropriate representation is to capture the information relevant to a given task while discarding unnecessary attributes. With this in mind, we describe the method of transforming time-domain audio signals into beat-synchronous sequences of harmonic features from which audio summaries can be identified.

2.1 Beat-Synchronicity

As a preprocessing stage, a recording is first analyzed by a beat tracking algorithm adapted from [3] for subsequent beat-synchronous feature extraction. In the interest of mitigating octave errors and producing consistent feature sequences across a variety of content, we impose constraints on the range of possible tempi the system can track. This is achieved by the following modification: periodicity analysis of the novelty function Δ_n is computed at $N\ log_2$ spaced frequencies per octave over the range $[1:8]$ Hz, producing the tempogram $\mathcal{T}$ as defined in [3]. This time-frequency representation is then wrapped to a single tempo octave of N bins and the most likely tempo path is extracted via the Viterbi decoder. In lieu of static transition probabilities, the transition probability matrix p_{trans} is defined as an identity matrix I of rank N convolved with a 1-D, 0-mean Gaussian window $\mathcal{N}$, where the standard deviation σ_n is parameterized by the relative amplitude of the maximum tempogram as a function of time n, as follows:

$$p_{trans}[n] = I_N * \mathcal{N}\left(\mu = 0, \sigma_n = \frac{max(|\mathcal{T}[n]|)}{\mu_{|\mathcal{T}[n]|}}\right) \quad (1)$$

This has the desirable effect of allowing the tempo estimator to adapt when the pulse strength is high, but resist change when the tempo becomes ambiguous. To find the best tempo octave to unwrap the path into, we analyzed a histogram of the chord durations contained in publicly available chord annotations [1] . Having found that approximately 95% of the chord durations are greater than 0.5 seconds in duration, we select 2Hz as a natural upper bound and map the optimal path through the single octave tempogram into the range of 60-120 BPM. At this stage, the remainder of the implementation follows the reference algorithm.

2.2 Harmonic Representations

Conventional approaches to harmonic analysis tasks in music informatics are predominantly built upon the use of chroma features, and we continue that tradition here. We also explore the use of tonal centroids, or Tonnetz features, as a mid-level harmonic representation. Introduced for the purpose of detecting harmonic change by Harte et al [4], the intuition for this decision is motivated as follows. First, typical distance metrics fail to capture musical significance between chroma vectors. In a pitch class representation, for example, the L_2 distance between a C major triad and a C♯ triad is equal to the distance between either triad and the notes B, B♭, and A. Additionally, chroma behaves like a mass function and it is not immediately apparent how to best measure the distance between these vectors. A Tonnetz representation, however, provides a geometric interpretation of pitch collections where distance is better defined as a musical and an Euclidean sense.

To compute both harmonic feature variants, we apply the constant-Q transform to a frame of audio over the range of 110–1760 Hz with 12 bins per octave, producing a pitch vector X. The length of the analysis window is determined by the longest filter, and is set to 0.45 seconds. Inspired by [5], a modified pitch vector Y is produced by standardizing the log-coefficients $log(\lambda X)$ and half-wave rectifying the result. The λ scale factor is heuristically set to 1000, but values within an order of magnitude in either direction produce similar results. Chroma features are derived from Y by wrapping onto a single octave and scaling by the L_2 norm, and Tonnetz features are computed identically to the method presented in [4].

2.3 Feature Quantization

It is computationally advantageous to quantize the feature space into a finite number of discrete values. We perform vector quantization by clustering the feature space via K-means and replacing each feature vector by its cluster's centroid. The pairwise distances between centroids are precomputed to accelerate distance calculations between symbolic feature sequences (see Section 3). Though larger values of K more faithfully reproduce the original features, this might result in an intractable process due to computation limitations as we see in subsection 3.3.

3. DEFINING AN AUDIO SUMMARY CRITERION

Structure and repetition are fundamental characteristics of a musical work, and an audio summary should retain the minimum number of distinct parts that are necessary to describe it. Therefore, a good summary criterion actually synthesizes two opposing notions: we seek to lose as little information as possible, while avoiding overlap between chosen segments. A summary is defined as the set $\Gamma = [\gamma_1^N, \ldots, \gamma_P^N]$ of P, N-length subsequences that maximizes a function Θ over a feature sequence $\mathbf{S}$ of length M, where $\exists m$ s.t. $s_m^N = \gamma_i^N, m \in [1 : M]$, $s_m^N \in \mathbf{S}$, and $i \in [1 : P]$.

3.1 Compression Measure

The goal of describing a sequence in terms of itself with a minimal loss of information is fundamentally a data compression problem. Building upon this idea, we define a compression measure $\mathcal{C}(\Gamma|\mathbf{S})$ that quantifies the extent to which Γ explains a given $\mathbf{S}$, defined as follows:

$$\mathcal{C}(\Gamma|\mathbf{S}) = 1 - \frac{1}{PJ}\sum_{i=1}^{P}\sum_{m=1}^{J} ||\gamma_i^N, s_m^N||_2 \qquad (2)$$

This measure can be interpreted as a normalized, convolutive Euclidean distance, such that there are $J = M-N+1$ element-wise comparisons between a given N-length subsequence γ_i^N and all J N-length subsequences $s_m^N \in \mathbf{S}$. All distances, taken directly from the precomputed pairwise matrix discussed in Subsection 2.3, are then averaged over the J rotations and P subsequences in Γ. Intuitively, the compression measure equals 1 when $\Gamma = \mathbf{S}$ and 0 when $\Gamma \nsubseteq \mathbf{S}$.

3.2 Disjoint Information Measure

In addition to determining how well Γ describes $\mathbf{S}$, it is necessary to also measure the amount of information shared between each pair of subsequences in a set. Conversely, a disjoint information measure $\mathcal{I}(\Gamma)$ seeks to quantify the uniqueness of each subsequence in Γ relative to the rest, defined as follows:

$$\mathcal{I}(\Gamma) = \left(\prod_{i=1}^{P}\prod_{j=i+1}^{P} D_{min}(\phi(\gamma_i^N), \phi(\gamma_j^N))\right)^{\frac{2}{P(P-1)}} \qquad (3)$$

We achieve shift-invariance by mapping a sequence of features γ_i^N to a sequence of *shingles* ρ_i^K with length $K = N-L+1$ where a shingle is defined as the stacking of L adjacent feature frames into a single feature vector. The function ϕ returns the *shingled* version of a subsequence. A modified Euclidean distance function D_{min} then measures the intersection between sequences of shingles, returning the average minimum distance between the u^{th} shingle in

[1] https://github.com/tmc323/Chord-Annotations

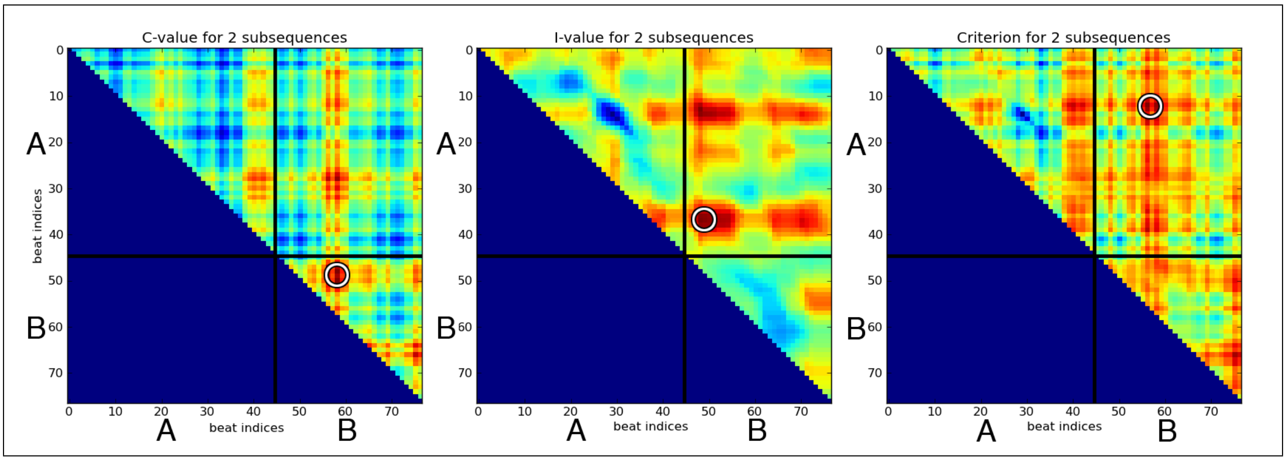

Figure 1. Search space for $\mathcal{C}$, $\mathcal{I}$ and Θ (left, middle, and right respectively) for $P = 2$ subsequences in the first half of a performance of the Mazurka Op. 30 No. 2. Black lines split part A and B. Circles mark the maximum value. Each position in the matrices correspond to a 8-beat subsequence.

ρ_i^K and all v shingles in a different subsequence ρ_j^K, defined as follows:

$$D_{min}(\rho_i^K, \rho_j^K) = \left(\sum_{u=1}^{K} min_v(\rho_i[u] - \rho_j[v])^2 \right)^{1/2} \qquad (4)$$

There are two important subtleties that must be observed when calculating this measure. First, distances between shingles are defined by the element-wise L_2 norm based on the same pairwise distance matrix as before. Additionally, $\mathcal{I}(\Gamma)$ is a geometric mean and only produces large values when all pairwise distances are also large; any small distance in the product forces the overall measure toward zero.

3.3 Criterion Definition and Calculation

Having established measures of compression and disjoint information for some Γ, we capture both of these traits by defining a single criterion Θ as follows:

$$\Theta(\mathcal{C}, \mathcal{I}) = \frac{2\mathcal{C}\mathcal{I}}{\mathcal{C} + \mathcal{I}} \qquad (5)$$

Noting that $\mathcal{C}$ and $\mathcal{I}$ are constrained on the interval [0,1] and converge to one when optimal, computing the criterion as a harmonic mean enforces the behavior that its value is only large when both measures are as well.

It is worthwhile at this point to make the observation that this criterion can —at least theoretically— be evaluated at every unique combination of subsequences Γ over an entire sequence $\mathbf{S}$. The output of this exhaustive calculation is a P dimensional tensor where each axis is of length J, and the best summary is given simply by the $argmax$ of the resulting data structure. From here onward, we use *optimal criterion* Θ_{max} to refer to the absolute maximum of this tensor, as would be found through a naïve, exhaustive search of the space. Note that for large J and P however, evaluating every cell in this tensor becomes computationally intractable and efficient approximations are necessary (see Section 4).

3.4 Case Example

Here we illustrate the behavior of the audio summary criterion by analyzing by the first half of Frédéric Chopin's Mazurka Op. 30 No. 2, which exhibits a well-defined *AB* structure. For the sake of demonstration, we select a subsequence length of $N = 8$ and define $P = 2$ such that an exhaustive evaluation of Θ produces a $J \times J$ matrix. The result of computing $\mathcal{C}$, $\mathcal{I}$ and Θ over all pairs of subsequences is shown in Figure 1.

The compression measure $\mathcal{C}$ is shown in the left-most matrix of Figure 1. This measure quantifies the extent to which a set Γ explains the overall track independent of any correlation between subsequences. The optimal $\mathcal{C}$ in this matrix corresponds to the two subsequences at beat indices (48, 59) in the *B*-*B* quadrant. These subsequences correspond to repetitions of the same part, making the information in Γ redundant.

The center matrix in Figure 1 corresponds to the disjoint information measure $\mathcal{I}$. This measure captures the degree of uniqueness between subsequences in Γ. It is clear from the plot that the measure behaves as expected: repeated subsequences in the same section (in quadrants *A*-*A* or *B*-*B*) produce significantly lower values of $\mathcal{I}$ than subsequence pairs in *A*-*B*, where the highest $\mathcal{I}$ is found.

Finally, the previous two matrices combine to yield a third, the criterion Θ. In the example the maximum value of $\mathcal{C}$ corresponds to repetitions of the same part, thus making $\mathcal{I}$ to be small and forcing the overall Θ to also be small. Similarly, the position of the maximum value of $\mathcal{I}$ at the boundary between *A* and *B* results in a low $\mathcal{C}$ value, again producing a smaller Θ. In this example, Θ is maximized by the combination of subsequences in *A*,*B* that best balance the two criteria by capturing the midsections of each part.

4. APPROXIMATING THE OPTIMAL SOLUTION

As mentioned in the previous section, naïve calculation of the optimal criterion can, in certain scenarios, become computationally inefficient, impractical, or worse. More specifically, an exhaustive evaluation and parallel search of the full Θ tensor of size $(J/2)^P$ would result in an algorithm of complexity $O((JN\log J)^P)$. In this section, we present a heuristic approach that approximates the optimal solution using a much faster implementation.

4.1 Heuristic Search Algorithm

The main idea behind the fast approach is to assume that the most relevant parts of a song will most likely be uniformly spread across time. The pseudocode is found in Algorithm 1. The method *EquallySpaced()* initializes all P subsequences into equally spaced time indices and stores them in the array Υ. We then iterate over the P subsequences, fixing all of them except the P_i being processed. We use a sliding window, operating over the region between the endpoint of the previous subsequence and the start of the next one, to find the best local music criterion θ by calling the function *ComputeCriterion()*. At every iteration we check if the sliding window is within the correct bounds with the method *CheckBounds()*, and if it is, we update the best index υ in Υ. Finally, the summary Γ is obtained by concatenating the subsequences at the time indices in Υ. This operation is done inside the method *GetSubseqsFromTimeIdxs()*.

Algorithm 1 Heuristic Approach

```
Require: $\mathbf{S} = \{s_1, \ldots, s_M\}, P, N$
Ensure: $\Gamma = \{\gamma_1^N, \ldots, \gamma_P^N\}$
  $\Upsilon \leftarrow$ EquallySpaced($\mathbf{S}, P, N$)
  for $i = 1 \rightarrow P$ do
    $\theta \leftarrow 0$
    for $j = 1 \rightarrow M$ do
      if CheckBounds($\Upsilon$) then
        $\Theta \leftarrow$ ComputeCriterion($\mathbf{S}, \Upsilon, N, P$)
        if $\Theta > \theta$ then
          $\theta \leftarrow \Theta; \upsilon \leftarrow j$
        end if
        $\Upsilon[i] \leftarrow j$
      end if
    end for
    $\Upsilon[i] \leftarrow \upsilon$
  end for
  $\Gamma \leftarrow$ GetSubseqsFromTimeIdxs($\mathbf{S}, \Upsilon$)
  return $\Gamma$
```

The complexity in time of this algorithm is $O(PMJ)$, which makes it linear with respect to P. This approach improves the efficiency dramatically and let us explore different hyperparameter values of P and N, as we will see in subsection 5.4.

5. EVALUATION

We now proceed to evaluate multiple facets of the audio summary criterion. We begin by reviewing the dataset used for evaluation before presenting three different experiments.

5.1 Methodology

In our experimentation, we use a collection of solo piano music compiled by the Mazurka Project[2], comprised of 2914 tracks corresponding to different recorded performances of 49 Mazurkas. For clarity, we use *piece* or *work* when referring to a Mazurka, and reserve *track* or *performance* to describe an instance of the work as audio. The motivation for using this dataset is to leverage the several performances of a single work to measure the consistency of our criterion. Additionally, this collection contains 301 tracks with human-annotated, ground truth beat times, which allows us evaluate the impact of beat tracking on various dimensions of performance. It also provides the added benefit that Chopin's Mazurkas are notoriously difficult to beat-track via computational means [3].

5.2 Parameter Sweep & Selection

In the interest of selecting a feature space with which to proceed, an experiment is designed to sweep across the range of free parameters to identify the optimal configuration. There are three questions to address: Is automatic beat tracking sufficient? Do chroma and Tonnetz features perform equivalently, or is one preferable? Does performance vary significantly as a function of codebook size?

These three decisions can be resolved by observing how the optimal criterion behaves across various performances of the same work, comparing between ground truth and estimated beat annotations. Intuitively, a satisfactory audio summary of the same piece would persist across recorded versions, so the summaries themselves should be substantially similar.

For those 301 recordings with ground truth beat annotations, we stratify the tracks into five folds for cross validation such that all but one are used to train the quantizer and the remaining hold-out is reserved as a test set. Sweeping across the two beat annotation sources (ground truth, automatic), two harmonic representations (chroma, Tonnetz), and three codebook sizes (50, 100, 200) produces 12 possible feature space configurations (see Table 1). Summary sets Γ are identified by exhaustively computing Θ_{max} over all possible combinations of subsequences, where segment length N and number P are fixed at 16 and 4, respectively. Additionally, a stride parameter of $N/2$, analogous to a hop size in frame based audio processing, is applied to make the exhaustive search more computationally tractable.

To measure the degree to which summaries of the same work (intra-class distance) are closer than those from other, dissimilar works (inter-class distance), the pairwise distances between summaries of tracks in each fold are computed and the values are treated as empirical distributions of these two classes. The Fisher ratio, defined by (6), provides an estimate of the separation between intra- and inter-class summary distances.

$$F_{ratio} = \frac{\mu_{intra} - \mu_{inter}}{\sigma_{intra}^2 + \sigma_{inter}^2} \qquad (6)$$

[2] *http://www.mazurka.org.uk*

k	GT-C	GT-T	A-C	A-T
50	3.64	3.97	2.71	3.89
100	3.84	4.29	2.68	4.20
200	4.09	**4.74**	2.87	**4.45**

Table 1. Parameter Sweep. GT: Ground Truth, A: Automatic, C: Chromagram, T: Tonnetz

Intuitively, higher values of F_{ratio} indicate distinct, well-localized distributions where 'similar' items cluster together, and translates to more consistency across performances. Table 1 shows the results of sweeping free parameters in the feature space. There are a few important observations to make about these results. First, a Tonnetz representation produces consistently better results than chroma features. Additionally, Tonnetz features computed from automatically extracted beat times only marginally trail their ground truth equivalent. Finally, the codebook size K has a non-trivial impact on performance and is positively correlated. Therefore, we can conclude that Tonnetz-features computed with a beat tracking front-end are the best choice going forward, and that the parameter K should be large and ultimately based on practical limitations of the implementation.

5.3 Heuristic Approximation

We evaluate the performance of the heuristic approach by comparing the summaries it produces with the optimal solution obtained through exhaustive computation. A second comparison is made with the expected performance a random algorithm, obtained by averaging across all results observed in the course of computing Θ_{max}. This establishes the upper (optimal) and lower (random) bounds of performance and allows us to determine where on this continuum our heuristic solution lives. We measure the discrepancy between the optimal Θ_{max}, random Θ_{rand}, and heuristic Θ_{heur} solutions by computing the averaged Mean-Squared Error (MSE) across all tracks in the full dataset. To account for local variance resulting for a given track, we normalize the range of $\mathbf{\Theta}$ such that $\Theta_{max} = 1$ and $\Theta_{min} = 0$. The normalized MSE can be expressed formally as follows:

$$\text{MSE}(\mathbf{\Theta}) = \frac{1}{\mathcal{S}} \sum_{i}^{\mathcal{S}} (1 - \mathbf{\Theta}_i)^2 \qquad (7)$$

Here, a normalized Θ_{max} is always equals 1, $\mathbf{\Theta}$ represents a vector of normalized criteria obtained by some search strategy, and $\mathcal{S}$ is the number of songs in the Mazurka data set.

Setting the hyperparameters to $P = 4$ and $N = 16$, the MSE of the random baseline is approximately 21%, whereas our heuristic approximation is nearly two orders of magnitude better, achieving a MSE of slightly over 1%. It is evident from this contrast that the heuristic search very closely approximates the results of exhaustive computation, significantly outperforming the random baseline. Therefore we offer the preliminary conclusion that the heuristic approach is a sufficient approximation, allowing a more thorough exploration over the space of hyperparameters.

5.4 Automatically Selecting Hyperparameters

Having gained the efficiency to perform a search across hyperparameters P and N, we can compute Θ_{max} for different combinations and define the maximum over this set as the optimal summary. In this experiment we explore 9 pairs of $P \in [2 : 5]$ and $N \in [16 : 64]$ (constraining N to powers of two), avoiding (P, N) combinations such as $(5, 64)$ or $(2, 16)$ that would produce summaries that are too long or short, respectively. These ranges incorporate prior musical knowledge, as there are typically a small number of distinct parts in a work and meter is predominantly binary. It is worthwhile to mention though the best choice of P and N is signal-dependent and that, in reality, there is no universally optimal combination for all music.

In light of this, the combination of P and N that yields Θ_{max} for a given track provides another statistic that should persist across multiple performances of the same work, as structure and meter are generally invariant to interpretation. We evaluate the criterion further by measuring consistency of the optimal (P, N) pair using the entire Mazurka dataset, and provide qualitative examples of the observed behavior.

5.4.1 Quantitative Evaluation

A consistency distribution resulting from a sweep across combinations of P and N is given in Figure 2. The x-axis represents the proportion of performances for a given Mazurka that produces the most frequent (P, N) pair at Θ_{max}, where a value of 1 indicates complete agreement and 0 complete disagreement. The y-axis represents the number of works that produce a given consistency value, and there are 49 in total.

As illustrated by the plot, there is very high consistency ($\geq$ 90%) for more than half of the data set, resulting in an average consistency of 87%. This shows that our criterion is able to capture high-level information about the structure of a work across various performances, validating its capacity to produce informative audio summaries. Despite a high average overall, it is of special interest to qualitatively analyze the Mazurkas that yield different optimal configurations of the hyperparameters.

5.4.2 Qualitative Evaluation

Importantly, Figure 2 fails to capture is the degree of contrast between Θ_{max} and values for other combinations of P and N. Upon closer inspection, we find that the structure of some works is not clearly defined leading to multiple, equally reasonable interpretations. This manifests explicitly in the data, leading to more than one (P, N) with large Θ values. One such instance of multiple interpretations occurs for Op. 7 No. 2. The form of this work is *ABCA*, but– depending on performance – parts *B* and *C* can be interpreted as one longer part, resulting in an *ABA* structure. Consequently, 62% of these performances produced

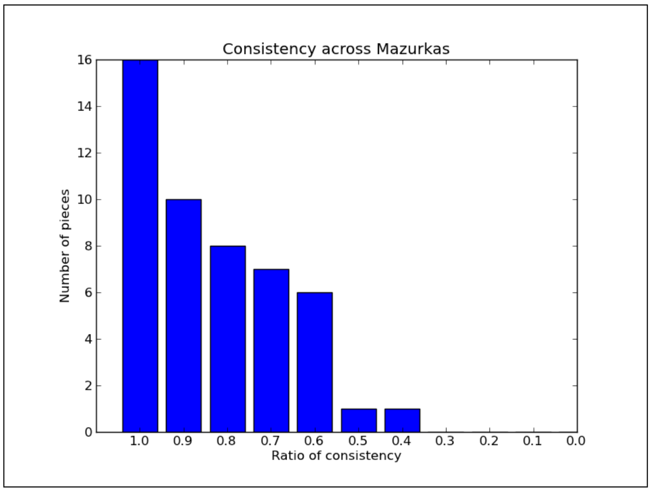

Figure 2. Evaluating consistency across different performances of the same song for the entire Mazurka data-set

a Θ_{max} for $P = 2$, while 31% of performances occurred at $P = 3$.

The other primary cause of inconsistency is due to tempo modulations and the resulting errors and artifacts caused by the beat tracker. An example of this is Op. 41 No. 1, producing the lowest consistency ratio of 49%. Here we observe a lack of well-defined onsets and liberal rhythmic interpretations, both within and between performances. This causes the beat tracker to behave erratically, producing misaligned feature sequences that ultimately yield Θ_{max} values for different pairs of (P, N).

Alternatively, Op. 24 No. 3, which exhibits a clear *ABA* structure and a more stable tempo, achieves 100% consistency for $P = 2$ and $N = 32$. The more noteworthy observation though is that this particular piece is in a ternary meter. Therefore a better summary would likely be obtained with N being a power of 3, and exploring other values of N could potentially improve consistency.

6. DISCUSSION & CONCLUSIONS

We have presented a novel audio summary criterion and established the merit of this approach through data-driven evaluation and qualitative inspection. We have illustrated how our criterion consistently produces informative summaries that capture both meaningful harmonic and high-level structural information. Finally, we have presented a heuristic approach capable of producing audio summaries that closely approximates the absolute maximum.

Complementary to the main body of work itself, the unexpected observation that Tonnetz features definitively yield better results warrants discussion. One possible explanation, as Tonnetz features live in a continuous-valued geometric space, is that any beat estimation errors result in a smooth interpolation of the feature space. Chroma features, acting as a time-varying probability distribution, cannot resolve timing errors in the same way. As a result, a beat tracker does not need to be perfect to be useful given a suitable feature representation.

As part of future work, we identify the potential of audio summaries to be used for various application where the data needs to be time normalized. More related to this work, the next logical step would be to explore the use of variable length subsequences to generate summaries. Finally, several example summaries are made available online [3] .

7. ACKNOWLEDGEMENTS

This work was supported by the Caja Madrid Foundation and the National Science Foundation under grant IIS-0844654, and the authors would like to sincerely thank them.

8. REFERENCES

[1] Mark Bartsch and Gregory Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[2] Matthew Cooper and Jonathan Foote. Summarizing popular music via structural similarity analysis. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130. IEEE, 2003.

[3] Peter Grosche and Meinard Muller. Extracting predominant local pulse information from music recordings. *Ieee Transactions On Audio Speech And Language Processing*, 19(6):1688–1701, 2011.

[4] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia AMCMM 06*, volume C, page 21. ACM Press, 2006.

[5] Matthias Mauch and Simon Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proc. of the International Society of Music Information Retrieval*, volume 11, pages 135–140, 2010.

[6] Geoffroy Peeters and Xavier Rodet. Toward Automatic Music Audio Summary Generation from Signal Analysis. In *Proc. of the International Society of Music Information Retrieval*, 2002.

[7] Xi Shao, NC Maddage, and Changsheng Xu. Automatic Music Summarization Based On Music Structure Analysis. In *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1169–1172. IEEE, 2005.

[3] https://files.nyu.edu/onc202/public/ismir2012