CORRELATING EXTRACTED AND GROUND-TRUTH HARMONIC DATA IN MUSIC RETRIEVAL TASKS

Dylan Freedman
Harvard University
freedmand@post.harvard.edu

Eddie Kohler
Harvard University
kohler@seas.harvard.edu

Hans Tutschku
Harvard University
tutschku@fas.harvard.edu

ABSTRACT

We show that traditional music information retrieval tasks with well-chosen parameters perform similarly using computationally extracted chord annotations and ground-truth annotations. Using a collection of Billboard songs with provided ground-truth chord labels, we use established chord identification algorithms to produce a corresponding extracted chord label dataset. We implement methods to compare chord progressions between two songs on the basis of their optimal local alignment scores. We create a set of chord progression comparison parameters defined by chord distance metrics, gap costs, and normalization measures and run a black-box global optimization algorithm to stochastically search for the best parameter set to maximize the rank correlation for two harmonic retrieval tasks across the ground-truth and extracted chord Billboard datasets. The first task evaluates chord progression similarity between all pairwise combinations of songs, separately ranks results for ground-truth and extracted chord labels, and returns a rank correlation coefficient. The second task queries the set of songs with fabricated chord progressions, ranks each query's results across ground-truth and extracted chord labels, and returns rank correlations. The end results suggest that practical retrieval systems can be constructed to work effectively without the guide of human ground-truthing.

1. INTRODUCTION

Computational algorithms to approximate harmonic content in a song typically output sequences of chord symbols which can be evaluated in terms of accuracy using their recall compared to human-annotated chord progressions. Leading algorithms to extract chord progressions from audio files have an accuracy of around 80% using popular Western music [12, 15]. Though these algorithms can effectively match human chord-labeling intuitions, it is largely unexplored how these approximated chord annotations perform in typical music retrieval tasks relative to human annotations. In this paper, we propose a method for evaluating the correlation of music retrieval task results across extracted and ground-truth datasets corresponding to the same collection of songs. We limit the scope of our exploration to chord labels and a few established similarity methods, but the resulting procedure can be generalized to other musical features such as melody, rhythm, and mid-level representations.

1.1 Contribution

This paper explores an alternative way to evaluate the efficacy of algorithms to extract musical features from songs. Rather than simply calculate accuracy of computationally extracted information relative to a reference, or ground-truth, dataset, we propose the use of *correlational metrics*. Given a set of common music informatics retrieval tasks on a set of songs, correlational metrics quantify to what extent the output results differ between two input sets: computationally extracted and ground-truth features for the same set of songs. Testing this system on a chord labeling algorithm, we design an alignment-based system to calculate harmonic similarity, devise two simple tasks—evaluating similarity between pairs of songs and querying by chord progression—and use a global optimization algorithm over the system's parameters to maximize the resulting correlational metric. The input datasets used and the design of the system are described in the following sections.

2. CHORD PROGRESSION DATASETS

The selection of songs we consider in this paper is motivated by availability. In order to compare ground-truth and computationally extracted chord datasets, it is necessary to have a set of song files, their corresponding ground-truth chord progression data, and a computer algorithm to extract chords from the audio files and create an extracted chord dataset. The number of reliable research-backed human ground-truth chord progression datasets is scarce, thus to maintain a separation of algorithm from data, it is useful to use a chord extraction algorithm that predates the ground-truth dataset such that it could not have been trained against any of its data.

2.1 Chord Extraction

Chordino [1] is an open-source chord extraction software program written by Matthias Mauch based on his winning

[1] http://isophonics.net/nnls-chroma

2009 and 2010 MIREX chord estimation algorithm submissions [4, 15]. Chordino achieves an 80% chord symbol recall and is still considered state-of-the-art [16]. Though an algorithm by Khadkevich [12] currently has the highest chord symbol recall in the 2014 MIREX audio chord estimation task, there is no publicly released source code for his work, whereas Chordino is available as a *VAMP* [2] plugin. The ground-truth dataset we use, as detailed in the following subsection, was compiled in 2011. Unlike Khadkevich's chord identification algorithm released in 2014, there is no possibility that Chordino could have been influenced by or tested against this dataset, maintaining a purity of separation between data and system. Chordino is the only chord extraction algorithm considered in this paper and is used with default settings.

2.2 Ground-Truth Dataset

The *McGill Billboard* annotations collected in [3] and freely available online [3] are a state-of-the-art human-annotated chord dataset. The dataset is comprised of over 1,000 songs sampled from different decades from the 1950s to the early 1990s across different Billboard charts from the United States "Hot 100". [4] The researchers hired music experts and professional jazz musicians to annotate the songs randomly sampled from the Billboard charts. Each song was annotated twice to maintain a standard of accuracy. The resulting dataset is the most comprehensive current ground-truth set of chord annotations and is used in recent MIREX chord annotation competitions. Importantly, the dataset postdates the Chordino chord extraction algorithm, obviating the possibility of training bias.

We were able to locate source audio for 529 of the McGill songs. The corresponding ground-truth annotations for these 529 form the *ground-truth McGill dataset*, or $McGill_g$. We extracted chord annotations for each of these 529 songs using Chordino with default settings, leading to the creation of the *extracted McGill dataset*, or $McGill_e$. To maintain a consistent chord alphabet, we simplify the harmonies used within the ground-truth dataset to match the closest chord within the alphabet of chord qualities used by Chordino. We preserve the root and bass notes of each chord and evaluate the closest simplified chord using the *Harte* metric as described in Subsection 3.2.1.

3. A HARMONIC SIMILARITY SYSTEM

3.1 Smith-Waterman Local Alignment Algorithm

The Smith-Waterman algorithm [17] is a dynamic programming algorithm that searches through two sequences exhaustively, looking for the pair of subsequences with optimal similarity based on the cost of transforming one subsequence into the other using three operators. The sequences are composed of symbols within an alphabet Σ. The first operator, *substitution*, defines the cost of transforming any one symbol into any other and can be represented as a two-dimensional cost matrix S, where $|S| = |\Sigma| \times |\Sigma|$. The second and third operators, *insertion* and *deletion*, quantify the cost of removing or adding a number of elements at a certain position in one of the subsequences, resulting in gaps in the final alignment. These two operators can be represented concisely using a gap function W that assigns costs to gaps of specified lengths. Given a substitution matrix S with a negative expected value but positive values for similar input symbols, the Smith-Waterman algorithm effectively isolates the strongest local regions of similarity corresponding to the highest score.

Smith-Waterman is useful in the context of comparing chord progressions as it has mechanisms to deal well with inexact data, using different gap costs and chord substitution functions that compensate for small errors. To account for songs in different keys, the score that is returned can be the maximum Smith-Waterman score of all twelve transpositions of one sequence relative to the other. Assuming a fixed substitution and gap function, let $sw(s_1, s_2)$ return the Smith-Waterman score for two sequences s_1 and s_2. If $t(s, i)$ is a transpose function that returns a transposed sequence given an input sequence s and a number of semitones i, we can express our final score as a similarity function SW:

$$SW(s_1, s_2) = \max_{t=0}^{11} sw(s_1, t(s_2, i)) \quad (1)$$

Due to its advantages and research that supports its efficacy [6, 10], the Smith-Waterman algorithm will be used to compare chord progressions in this paper and quantify harmonic similarity. There are downsides to the Smith-Waterman algorithm. In its current form, the score returned reflects only the optimal local alignment and does not consider other strong subregions of similarity. Allali et al. [1] describe a process for constructing a 3-dimensional Smith-Waterman algorithm that can account for modulations to a new key signature mid-song. These adaptations leave room for future experimentation. This paper focuses on only returning one optimal local alignment score in the highest scoring transposition.

3.2 Parameters

We chose a number of parameters to alter the nature of the Smith-Waterman algorithm used. These parameters are used with global optimization techniques to find good settings such that ground-truth and extracted chord annotations perform similarly.

3.2.1 Chord Distance Functions

We consider two chord distance metrics. Like Haas et al. [6], we use Lerdahl's Tonal Pitch Space (*TPS*) [14] as a chord distance function to populate the substitution matrix S. *TPS* quantifies the distance between two chords relative to the key signature of a song based on psychological qualities of human chord perception. We utilize the key finding approach in [6] to establish the tonic and mode of each song we are considering and assume no transpositions occur midsong. We additionally consider a metric

[2] http://www.vamp-plugins.org/
[3] http://ddmal.music.mcgill.ca/billboard
[4] http://www.billboard.com/charts/hot-100

proposed in Harte's PhD thesis (*Harte*) [11] which quantifies the fraction of similar pitch classes between two chords over their cumulative set of pitch classes. If $P_c(c)$ returns the set of pitch classes for a given chord c this can be expressed as:

$$Harte(c_1, c_2) = \frac{|P_c(c_1) \cap P_c(c_2)|}{|P_c(c_1) \cup P_c(c_2)|} \quad (2)$$

We denote a variable C_d to correspond to which distance function is used, *TPS* or *Harte*.

We additionally devise two parameters to scale and subtract the function C_d such that a cost matrix S populated by C_d has a negative expected value. We first normalize the chord distance function to a value in [0,1], where 0 indicates no similarity and 1 perfect similarity. In *TPS* this requires a division by 13. m_x represents the amount by which this normalized value is multiplied and m_s the amount it is subtracted. We round this number to the nearest integer out of consideration for the Smith-Waterman implementation we used. We arbitrarily only considered integers from 1 through 30 inclusive for both m_x and m_s as values in this range seemed to achieve a good resolution of scaled chord distance values. Finally, S can be populated based on the final scaled and subtracted value and choice of C_d by iterating over all possible pairs of chords in Σ.

3.2.2 Gap Costs

We only consider one class of gap functions, *affine gap functions* [9], which can be defined by the following equation for gaps of size $i \geq 1$:

$$W(i) = -gap_{open} - gap_{\text{extension}} \cdot (i - 1) \quad (3)$$

The two constants gap_{open} and $gap_{extension}$ are parameters that can be changed to alter the penalty of the initial gap and following gaps in the sequence alignment, a potentially useful feature to model an initial alignment gap being more or less costly than subsequent gaps. In our implementation, we considered integer values ranging from 0 through 127 inclusive for gap_{open} and $gap_{extension}$.

3.2.3 Normalization

A difficulty with comparing Smith-Waterman scores is that they tend to have a positive correlation with increased sequence length. There are approaches to combat this effect using statistical learning techniques [2]. We tested a simpler normalization metric that returns values in [0,1]:

$$SW_{norm}(s_1, s_2) = \frac{SW(s_1, s_2)}{\max\{SW(s_1, s_1), SW(s_2, s_2)\}} \quad (4)$$

We devise a parameter CP_d to represent the chord progression similarity function used, *SW* or SW_{norm}.

4. EXPERIMENTAL DESIGN

This paper tests how similarly common harmonic music retrieval tasks perform using extracted chord data versus human-produced data. We primarily test two tasks across $McGill_g$ and $McGill_e$ datasets, rank both sets of results, and calculate a correlational metric P. We then run a black-box optimization strategy to approximate a maximum for this correlational metric across the different parameters detailed in Section 3.2. This process is outlined in a flowchart in Figure 1.

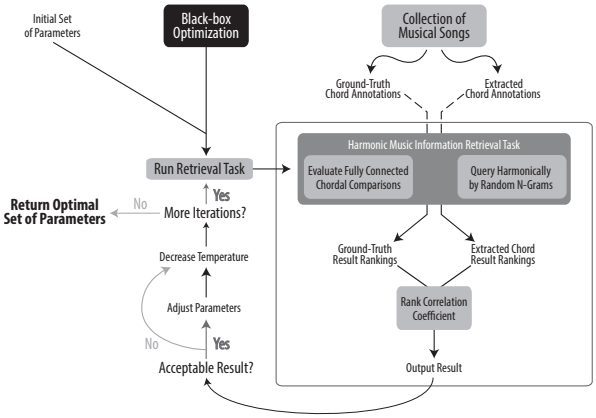

Figure 1. Flowchart of experimental design. This experiment requires a collection of songs with corresponding ground-truth and computationally extracted chord annotations. These different chord datasets describing the same collection of songs are fed into a harmonic retrieval task in isolated experiments, each producing a different result list. These result lists are ranked and correlated to return a correlational metric. Global optimization techniques search for maximum correlational metric scores by running many iterations of the retrieval task with changing parameters based on the performance of the correlational result relative to previous iterations. The returned set of parameters represents an approximate optimal configuration for minimizing algorithmic differences between human and computationally extracted chord inputs.

4.1 Retrieval Tasks

This subsection describes the two high-level tasks that form the substance of the experiments. Inputted with the parameters described in the previous section, these algorithms perform chord progression comparisons over a collection of songs using the harmonic similarity system previously outlined to accomplish a common music retrieval objective. The result is a collection of harmonic similarity scores that can be enumerated in an ordered fashion.

4.1.1 Fully Connected Pairwise Harmonic Comparison

This method (*FCC*), given parameters and a collection of chord annotations, returns the harmonic similarity scores for every pairwise combination of songs. The algorithm proceeds in a well-ordered manner such that no pair of songs is iterated twice and results are consistently positioned across two sets of chord annotations corresponding to the same collection of songs (e.g. $McGill_g$ and $McGill_e$).

4.1.2 Query by N-gram

This retrieval task (*QBN*), given parameters and a collection of chord annotations, involves comparing the collection of annotations with random chord sequence queries to simulate a basic search algorithm. Each query sequence is compared with every song in the database, and a two-dimensional table of harmonic similarity scores is returned.

We initially fabricate 100 query sequences, generated randomly within the alphabet of chord qualities Σ but used consistently across experiments and song collections. Out of the 100 query sequences, four groups of 25 query sequences are generated with lengths of 4, 8, 16, and 32, respectively. Each query sequence is padded in length by repeating itself such that the length is at least that of the longest song in the collection so that the Smith-Waterman scores are not restricted by length of query sequence. We chose this repetition of query sequences to imitate the repetitive structure of musical songs and emphasize the cyclic nature of chord progression perception. For each of the 100 query sequences, *QBN* collects harmonic similarity scores by comparing the query sequence against each of the songs in the given collection. The result is a two-dimensional table of harmonic similarity scores of size 100 by the length of the input collection of songs.

4.2 Correlational Metrics

4.2.1 Ranking and the Spearman Correlation Coefficient

The ranking of a sequence is a mapping of every element of the sequence to its position in the sequence. This ranking is done such that elements with the same value are assigned the average index of their positions. Two equally sized lists of rankings s_1 and s_2 can be assigned a correlation coefficient based on the Spearman correlation coefficient (ρ) [7]. If n is the length of one of the ranked lists, ρ can be calculated:

$$\rho(s_1, s_2) = 1 - \frac{6 \sum_i^n (s_{1_i} - s_{2_i})^2}{n(n^2 - 1)} \quad (5)$$

ρ returns a number in [-1,1], with 1 indicating a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation.

4.2.2 Calculating the Correlational Metric

For each of the two experimental tasks, we compare the resulting harmonic similarity scores across ground-truth and extracted chord annotations corresponding to the same set of songs, $McGill_g$ and $McGill_e$.

The resulting correlational metric P is calculated by ranking the result lists for $McGill_g$ and $McGill_e$ separately and returning a rank correlation between both resulting lists. For *FCC*, P is calculated by simply ranking each result list and returning the Spearman correlation coefficient ρ between the two ranked lists. For *QBN*, each result list from $McGill_g$ and $McGill_e$ for each of the 100 queries generated is ranked independently. The correlation coefficients ρ for each of the 100 pairs of ranked lists is averaged and returned as P.

Variable	Notation	Values
Similarity Function	CP_d	$\{SW, SW_{norm}\}$
Gap Open Cost	gap_{open}	$[0, 128]$
Gap Extension Cost	$gap_{extension}$	$[0, 128]$
Chord Distance	C_d	$\{Harte, TPS\}$
Distance Multiplier	m_x	$[1, 30]$
Distance Subtractor	m_s	$[1, 30]$

Table 1. Summary of experimental parameters.

4.3 Global Optimization with Simulated Annealing

To derive optimal parameters to maximize the correlational metric P across tasks, we use a basic implementation of the simulated annealing algorithm [5,13]. Let f refer to one of the retrieval tasks that takes as input a set of parameters s_t and runs over $McGill_g$ and $McGill_e$ to return a correlational metric P.

We try to stochastically search for parameters in s_t to maximize $f(s_t)$. Simulated annealing takes a function, $move(s_t)$, which returns a new state s'_t that is slightly changed from s_t in a random manner. f is recalculated with s'_t to see if the move was beneficial. A temperature variable T stores acceptable deltas between old and new states. If $|f(s'_t) - f(s_t)| > T$, the move is rejected and s_t is left unchanged; otherwise, s_t takes on the new state value, s'_t.

Simulated annealing runs with a fixed number of iterations i_t. In each iteration, we perform $move(s_t)$, and following each iteration, T exponentially decreases. This gives the optimization process more exploratory freedom in initial stages when T is higher. After i_t iterations, the resulting s_t is an approximate maximum of f. This algorithm is useful in search spaces that are sufficiently complex or large, such that exact optimization algorithms are infeasible.

4.3.1 Implementation

Let s_t contain our parameters (see Table 1): $\{CP_d, gap_{open}, gap_{extension}, C_d, m_x, m_s\}$. The $move$ function represents a transition to a nearby state—as each variable in s_t is an integer, the jump must be discrete. Our *move* implementation takes a random step following a normal distribution for each variable in the state, rounding the result to the nearest integer and ensuring the value falls within the bounds of the variable. The standard deviation of this random step for each variable is chosen to be $\frac{1}{3}$ of that variable's range. CP_d and C_d, taking two possible function values each, can be treated as integer variables with values in $\{0, 1\}$. If a move results in a combination of parameters such that the expected value of S is not negative or there are no positive values, the scaling and subtraction factors m_x and m_s are randomized again from their last values following the same normal distribution jump process. This process repeats until S has a negative expected value and some positive values so the Smith-Waterman algorithm can effectively isolate localized chord comparison results.

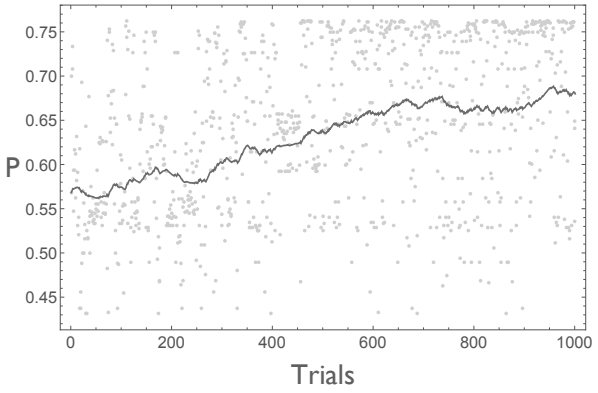

Figure 2. Simulated annealing performance in *FCC*. Each dot represents an iteration of the algorithm and correlational metric P. The jagged line, an exponential moving average, demonstrates the relatively constant increase in performance as iterations progress.

For each task, *FCC* and *QBN*, we run 1,000 iterations of simulated annealing to optimize the correlational metric P with a temperature T that starts at 1 and decreases exponentially to 0.005 at the final iteration.

5. RESULTS

In this section, we detail the results of the optimization procedures across the two retrieval tasks (*FCC* and *QBN*) as detailed in Subsection 4.1.

5.1 Optimizing Fully Connected Comparison

Across 1,000 iterations of simulated annealing for the *FCC* task, the correlational metric P at each iteration generally increased (see Figure 2). The maximal P returned by the simulated annealing was 0.7619, indicating a strong correlation. The parameters s_t resulting in this correlation first occurred at iteration 472 with values $\{CP_{\mathbf{d}} : SW, gap_{open} : 0, gap_{extension} : 28, \mathbf{C_d} : TPS, \mathbf{m_x} : 5, \mathbf{m_s} : 9\}$. The correlation between the ranked result lists for ground-truth and extracted chord data with these parameters can be visualized with a 3-dimensional histogram in Figure 3.

A common measure for accuracy in music retrieval is the Average Dynamic Recall (ADR) [18], which has been used to evaluate similarity assessments in MIREX competitions since 2005. In the context of retrieval results, ADR assesses at all given position how many songs have occurred up to that position that should have occurred relative to ground-truth rankings, returning an average in [0,1], with 1 indicating perfect similarity. We calculated the ADR of the *FCC* results list of extracted chord data relative to the generated ground-truth results list, deriving a result of 0.7664. As a warning, this measure is not particularly applicable to our work as the output ground-truth results list does not demonstrate an actual ground-truth similarity assessment, but its use here nonetheless illustrates the correlation of this parameter set in the context of music retrieval.

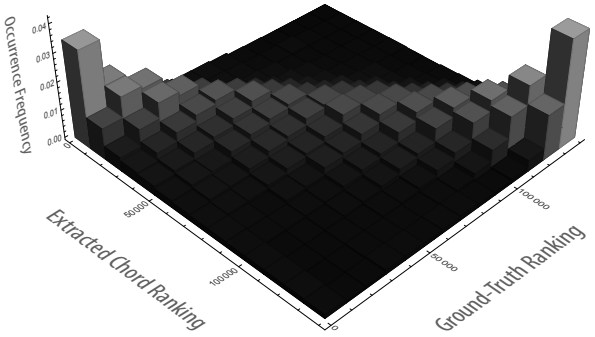

Figure 3. 3-dimensional histogram of the optimal fully connected comparison (*FCC*) rankings. The correlation (ρ=0.76) is visible through the elevated diagonal band. The density of points along this band is greatest at the corners as evidenced by bin heights—this means salient strongly and weakly ranked chord progression results are most preserved by the parameters that led to this result.

5.2 Optimizing Query by N-grams

Like *FCC*, the correlational metric P also generally increased across iterations in *QBN* (see Figure 4). The maximal P returned by simulated annealing was 0.7790, occurring singularly with the parameters $s_t = \{CP_{\mathbf{d}} : SW_{norm}, gap_{open} : 1, gap_{extension} : 82, \mathbf{C_d} : TPS, \mathbf{m_x} : 1, \mathbf{m_s} : 10\}$. The average ADR across each of the 100 queries with these parameters was 0.7900.

Task	P	ADR
FCC	0.7619	0.7664
QBN	0.7790	0.7900

Table 2. Summary of experimental results.

5.3 Parameter Optimization

The harmonic retrieval tasks presented in this paper, *FCC* and *QBN*, rely on a common set of parameters s_t. Though generalizations on effective values for the parameter set cannot be fully founded, it can still be useful to future experimentation to detail average correlational metric values associated with ranges of parameter values from the simulated annealing experiments.

CP_d and C_d are the variables that perhaps change the nature of the Smith-Waterman function the most fundamentally. Average output correlational metric values for inputted choices of CP_d and C_d are as follows:

	FCC		**QBN**	
	$Harte$	TPS	$Harte$	TPS
SW	0.59	<u>0.68</u>	0.40	0.49
SW_{norm}	0.56	0.62	0.35	<u>0.53</u>

where maximum values are underlined. According to these observational results, *TPS* outperforms the *Harte* chord distance metric in both experiments in terms of maximizing correlation.

gap_{open} and $gap_{extension}$ take a wider range of values, thus it is more useful to look at variable ranges and their

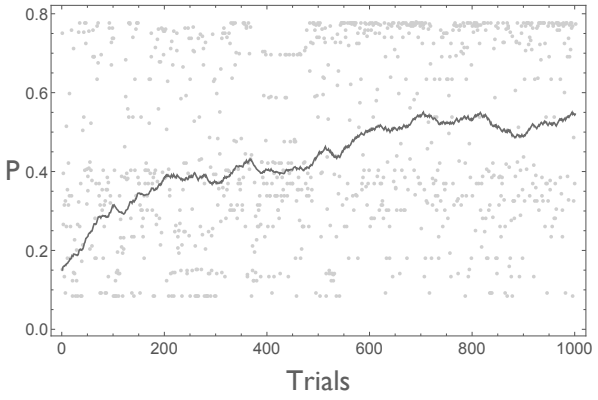

Figure 4. Simulated annealing performance in *QBN*.

average outputs. Following are correlational metrics corresponding to ranges of gap variable values:

	FCC		**QBN**	
Range	gap_{open}	$gap_{extension}$	gap_{open}	$gap_{extension}$
0	0.71	0.64	0.69	<u>0.49</u>
1-8	0.67	0.62	0.46	0.49
> 8	0.61	<u>0.64</u>	0.34	0.47

These results suggest that gap opening penalties of 0 influence higher correlational harmonic metric score.

Finally, we chart which scaling and subtraction factors, m_x and m_s, produced the highest average correlation metric scores:

		FCC			**QBN**		
		m_x			m_x		
		1-9	*10-19*	*20+*	*1-9*	*10-19*	*20+*
m_s	1-9	0.67	0.64	<u>0.69</u>	0.62	0.51	<u>0.65</u>
	10-19	0.68	0.65	0.65	0.51	0.43	0.49
	> 20	0.58	0.58	0.57	0.39	0.38	0.30

These results are both consistent in assigning higher correlational metric scores to large multiplication factors and small subtraction factors. A possible explanation for this behavior and favoritism towards gap penalties of 0 is that these factor choices result in the highest Smith-Waterman expected values and result scores. Though this expected value is ensured to be negative, a value close to 0 will more frequently match chords positively by chance and result in longer local alignment scores that resemble global alignment scores. It is possible that global sequence alignment techniques used in *FCC* and *QBN* have strong correlational harmonic metric scores. Further research in global sequence alignment could present promising correlational metric results.

6. DISCUSSION

This paper suggests a new class of similarity assessments in music information retrieval (MIR), *correlational metrics*, and outlines an experimental procedure for assessing these metrics. Correlational metrics capture the degree to which ground-truth and extracted features perform similarly through retrieval tasks. It is possible that similar results in a retrieval task do not necessarily imply correct or good results. The experimental choices made in this paper, such as using local alignments and the chord distance metrics, are demonstrated in MIR research as strong choices for matching human intuitions of similarity [8, 11]; however, these experimental choices in this paper reflect one possible use case. In the context of chord progressions, there does not exist any reliable ground-truth similarity assessments, which motivated this work.

Further experimentation is necessary with different chord extraction algorithms and settings. The chord extraction algorithm used in this paper is highly accurate, which may imply stronger correlational metric scores. Testing a variety of chord extraction algorithms would render a comparison of correlational metric scores associated with a gradient of extraction algorithm accuracies, giving statistical significance to the resulting scores and potentially uncovering other salient observations. Once there exist research-backed ground-truth similarity assessments for chord progressions, this work can be enriched with direct comparisons to human intuitions. In its current form, this paper is limited to Western harmonies, and more specifically, pop songs from the 1950s onwards. Many other features could be investigated within our experimental design, from those directly supplemental to harmony, such as chord duration and melody, to external factors, such as song popularity or artist. Incorporating and testing more chord distance metrics and different parameters and ranges would additionally benefit this class of research. Modifying the retrieval tasks and implementing additional tasks could extend this work, as well. For instance, randomized query sequences in the *QBN* task could be generated according to probabilistic n-gram models to match more likely search inputs and limit bias in the resulting correlational metric score as a result of purely random queries being unnatural and distant to the input datasets.

Assuming the parameter choices that resulted in the optimal correlational metrics in this paper resulted in a harmonic similarity metric that matches human intuitions of similarity, this paper suggests that effective MIR systems can be constructed without the need for ground-truth chord annotations and provides a framework for conducting such experiments. As there are few research-backed ground-truth chord datasets, this could massively expand the possible realm of chord datasets to reliably harmonically compare. Correlational metrics can also be used in future research across other musical features. The potential implications of this paper suggest that with proper algorithms and parameters that currently exist in the literature, practical MIR systems can be constructed and optimized to work without the guide of human ground-truthing in similarity assessments.

7. REFERENCES

[1] Julien Allali, Pascal Ferraro, Pierre Hanna, and Costas Iliopoulos. Local transpositions in alignment of poly-

phonic musical sequences. In *String Processing and Information Retrieval*, pages 26–38. Springer, 2007.

[2] Eric Breimer and Mark Goldberg. Learning significant alignments: An alternative to normalized local alignment. In *Foundations of Intelligent Systems*, pages 37–45. Springer, 2002.

[3] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An Expert Ground-Truth Set for Audio Chord Recognition and Music Analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 633–638, 2011.

[4] Chris Cannam, Matthias Mauch, Matthew EP Davies, Simon Dixon, Christian Landone, Katy Noland, Mark Levy, Massimiliano Zanoni, Dan Stowell, and Luıs A Figueira. Mirex 2013 entry: Vamp plugins from the centre for digital music, 2013.

[5] Vladimír Černỳ. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

[6] W Bas De Haas, Matthias Robine, Pierre Hanna, Remco C Veltkamp, and Frans Wiering. Comparing approaches to the similarity of musical chord sequences. In *Exploring Music Contents*, pages 242–258. Springer, 2011.

[7] Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268, 1977.

[8] Pascal Ferraro and Pierre Hanna. Optimizations of local edition for evaluating similarity between monophonic musical sequences. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 64–69. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, 2007.

[9] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705 – 708, 1982.

[10] Pierre Hanna, Matthias Robine, and Thomas Rocher. An alignment based system for chord sequence retrieval. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries*, pages 101–104. ACM, 2009.

[11] Christopher Harte. *Towards automatic extraction of harmony information from music signals*. PhD thesis, Department of Electronic Engineering, Queen Mary, University of London, 2010.

[12] Maksim Khadkevich and Maurizio Omologo. Time-frequency reassigned features for automatic chord recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference*, pages 181–184. IEEE, 2011.

[13] Scott Kirkpatrick et al. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[14] Fred Lerdahl. Tonal pitch space. *Music Perception*, pages 315–349, 1988.

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

[16] Matt McVicar, Raúl Santos-Rodríguez, Yizhao Ni, and Tijl De Bie. Automatic chord estimation from audio: A review of the state of the art. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(2):556–575, 2014.

[17] Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[18] Rainer Typke, Remco C Veltkamp, and Frans Wiering. A measure for evaluating retrieval techniques based on partially ordered ground truth lists. In *Multimedia and Expo, 2006 IEEE International Conference*, pages 1793–1796. IEEE, 2006.