SEARCH-EFFECTIVENESS MEASURES FOR SYMBOLIC MUSIC QUERIES IN VERY LARGE DATABASES

Craig Stuart Sapp
CCRMA and CCARH
Dept. of Music
Stanford University
Stanford, CA 94305-3076, USA
craig@ccrma.stanford.edu

Yi-Wen Liu
CCRMA
Dept. of Electrical Engineering
Stanford University
Stanford, CA 94305-3076, USA
jacobliu@stanford.edu

Eleanor Selfridge-Field
CCARH
Depts. of Music, Symbolic Systems
Stanford University
Stanford, CA 94305-3076, USA
esfield@stanford.edu

ABSTRACT

In the interest of establishing robust benchmarks for search efficiency, we conducted a series of tests on symbolic databases of musical incipits and themes taken from several diverse kinds of repertories. The results we report differ from existing studies in four respects: (1) the data quantity is much larger (*c*. 100,000 entries in all); (2) the levels of melodic precision are more refined; (3) anchored and unanchored searches were differentiated; and (4) results from combination pitch-and-rhythm searches were compared with those for pitch-only searches.

The results were evaluated using a theoretical approach which seeks to rank the amount of effort required to achieve "uniqueness." How far into a melody must a search go in order to "find" an item which is unmatched by any other of the available items? How much does the answer depend on the specificity of the query? How much does anchoring the query matter? How much does the result depend on the nature of the repertory? We offer experimental results for all of these questions.

The *Themefinder* Database (http://www.themefinder.org) contains a family of databases encoded in the Humdrum **kern data-format. Unlike MIDI data, in which some pitch-data, vis-à-vis notation, is ambiguous, this format provides explicit pitch and duration descriptions for all notes. This enables evaluation of the importance of distinguishing between enharmonic spellings (e.g., C# vs. Db), where MIDI offers only one designation. The constituent collections (some publicly searchable, others limited to licensed use) each represent a different kind of music (Table 1). Some collections are tonal, some modal, and one is pentatonic. Within the tonal collections, significant range can be found with respect to diatonic, chromatic, and (occasionally) enharmonic usage. Starred items in Table 1 are publicly searchable.

The main purpose of the *Themefinder* website is to enable trained musicians to identify works by their melodies *as remembered.* Users are assumed to be notationally literate, since they are most likely to seek a work-title. That is, they are seeking textual meta-data from a symbolic-data search. Results are viewable in notation and playable as corresponding MIDI files. The length of the queries is at the discretion of the user.

Dataset	Genre(s)	Original code	No. of incipits
Essen European*	Folksongs	EsAC	6,232
Luxembourg*	Folksongs (European)	EsAC	612
Essen Asian*	Folksongs	EsAC	2,241
TF Classical*	Instrumental (sonatas, quartets, concertos, symphonies); Vocal (operas, cantatas)	MIDI (re-edited)	10,718
Renaissance Italian*	Motets, 16th cent.	DARMS	18,946
Polish religious monophony	Devotional songs, 16th, 19th cents.	EsAC	6,060
RISM A II/US	Instrumental and vocal, 17th-18th cents.	Plaine & Easie Code	55,490
Total			100,299

Table 1. Constituent databases in *Themefinder*.

Pitch can be searched by the user at five levels of precision. We give them in the order in which they appear on the *Themefinder* search-form:

[1] **Gross contour**. Intervals are assigned to one of three categories by melodic direction (up, down, or the same);

[2] **Refined contour**: Intervals are assigned to one of five categories: steps (of 1 or 2 semitones) up or down, "skips" (movements of 3 or more semitones) up or down, or unchanged;

[3] **Scale-degree profile**: Pitches are described by their diatonic position in a tonal (major or minor) scale;

[4] **Intervallic profile**: Each interval is specified by melodic direction, diatonic size (3rd, 5th, etc.), and "quality" (perfect, major, minor, augmented, diminished);

[5] **Exact pitch profile**: Notes are named by letter (A..G) and inflection (sharp, flat, natural).

The search-engine available to users contains a meter filter but currently offers no defined description of rhythmic values on an item-per-item basis. We have observed that exact-pitch searches penalize faulty recollection and that gross-contour searches often produce an inappropriately large number of candidate matches. We had come to believe that scale-degree searches are the most robust for tonal music, but we had not attempted to evaluate the search methods statistically prior to this study.

1. REPERTORY DESCRIPTION

None of the component datasets in *Themefinder* originated as **kern data, which represents pitch, duration, barring, and the global variables of notation music (meter signature, key signature, clef, etc.). All the repertories were originally encoded at a level of detail sufficient to support translation into **kern. The total number of records available used in this study was 100,299.

The use of musical incipits raises fundamental questions of musical identity. All incipits are monophonic as used in *Themefinder*, but some of the underlying repertories are polyphonic. In relation to polyphonic music, Lincoln [5] gives one incipit for each voice (typically five). RISM gives the incipit (usually from the highest-pitched instrumental or voice, *e.g.*, Violin 1 or Soprano) but is generous to giving separate incipits for linked portions of a single movement (*e.g.*, instrumental ritornello and aria).

A qualitative difference distinguishes incipits from themes. "Incipits" *introduce* a song, work, or movement. They serve well for short works that are uncomplicated. "Themes" *represent* a piece of music (usually a longer, more complex one) in some more essentially cognitive way. The mental extrapolation of themes is a human task and therefore vulnerable to subjective variation. As best we know, there is no study of the extent of difference that might be found by multiple subjects in the identification of thematic material.

In constructing this series of tests, we gave consideration to the relationship (or lack thereof) between incipits (in monophonic contexts) and overall pieces (we did not distinguish between incipits and themes *per se*). We subsequently compared search results for incipit data and full-work data for the Essen folksong collection (in which German/Austrian folk-music predominates).

The repertories vary substantially by mode. The Essen-European, Classical, and RISM datasets are overwhelmingly tonal. The Renaissance Italian database employs modes of the period; the Polish data contains two subsets and is almost evenly divided between modal and tonal monophony. The Essen-Asian dataset is pentatonic. Pentatonicism (the use of five tones per octave) makes scale-degree searches ambiguous, since, when mapping five-tone profiles onto seven-tone grids, quantitative differences in "scale"-degree usage are inevitable. Since we were interested in comparing procedures and their effectiveness in different repertories, we did not attempt to correct for this distortion. The Renaissance repertories subscribe to different systems of rhythmic organization than what is conventional in common music notation (CMN). This discouraged the investigation of multiple tiers of precision in rhythmic definition.

2. EXISTING STUDIES

In searching for related literature, we found few studies which were systematic in nature and which addressed substantial quantities of data, although many studies touched on some aspect of this general area of enquiry. Both Dannenberg et al. [2] and Rand and Birmingham [6] explored similar procedures but in relation to a query-by-singing situation. In [6] (the chronologically earlier study) 188 MIDI files were used as a basis for profiling durational change, pitch change, and "note-drop" in an effort to simulate the kinds of user errors anticipated in sung input. They noted that correlation coefficients and count correlations performed equally well in processing and combined them in a modified scoring metric. In [2] one database of 2,844 items was generated from a MIDI collection of Beatles songs and a second, of 8,926 themes (averaging 41 notes), was based on an encoded collection of monophonic folksongs. Here they compared the results of metrics derived from (1) pitch plus inter-onset intervals (IOI), (2) melodic contour matches, and (3) Hidden Markov Models. They found the second and third procedures to be slow in processing time and found the best performance to come from "note-IOI" couplings.

The closest parallels with our own work are found in [4, 7, 8]. The first two were concerned with sorting records stored in related symbolic databases into a musical equivalent of alphabetical order for bibliographical purposes (e.g., finding concordances for works which are anonymous or which are attributed to multiple composers). The authors of [8] sought to determine the feasibility of the query-by-humming approach by simulating some of the known deficiencies in user input. They considered melodic representation at five levels of pitch-resolution, but it is unclear exactly how "intervals" were defined at some levels (in the categories "3, 5, 7, 9, 12") of "pitch resolution." They attempted to simulate different levels of inaccuracy in sung queries. They reported results for different database-sizes (of 0.6, 1.2, 2.5, and 3.6 million notes) and search-key lengths (8-4 tokens). They found that a "three-interval contour" required a 1.7 longer query-length than a semitone ("12 interval") resolution. They reported that 5-state, 7-state, and 9-state representations of the underlying melodies led to similar results but in all cases produced improvement over three-state representations.

3. METHODS

3.1. Levels of Pitch and Rhythmic Resolution

We used the *Themefinder* data collections directly to avoid restrictions of the user interface and to process data more quickly. We therefore did not use the wildcard possibilities offered to users.

The high degree of precision in pitch-specification in the underlying datasets made it possible to investigate levels of precision on which no literature appears to exist. We therefore expanded the number of levels of precision (Table 2). For the purposes of this study we created a (three-state) rhythmic analogue to gross-contour descriptions of pitch in which each new event after the first note was classified as greater than, less than, or the same in duration as its predecessor.

Search type	**No. of states**	**Description**
Pitch/gross contour (**pgc**)	3	Each event after 1st is classified as being up, down, or unchanged in pitch in relation to its predecessor.
Rhythmic/gross contour (**rgc**)	3	Each event after 1st is classified as being longer than, shorter than, or the same duration as its predecessor.
Pitch/refined contour (**prc**)	5	Each event is classified as being a step or skip, up or down, from its predecessor, or as being unchanged.
Scale-degree (**sd**)	7	The diatonic scale-degree of each pitch is evaluated; restricted by octave..
12-tone pitch class (**12p**)	12	The chromatic scale-degree of each pitch is evaluated (limited to one octave).
12-tone pitch interval (**12i**)	12 (24)	The chromatic size of each interval is evaluated; 24 states required if preserving direction.
Base-40 pitch (**pch**)	40	An enharmonic-pitch index is derived.[3].
Melodic-interval (**mi**)	40 (80)	An enharmonic intervallic index is derived [3]; 80 states required if preserving direction.

Table 2. Levels of pitch, rhythmic, and intervallic precision and their associated numbers of theoretical states within the bounds of one octave. The numbers shown in parentheses accommodate directional couplings (i.e., size plus direction, if the next pitch is higher or lower).

One subtlety of pitch representation that is too little discussed in music-query literature is the effect of octave discrimination in the representation of the analyzed data. In the mainframe era it was convenient to fold pitch representations onto a single octave to save storage and speed processing. As processor speed increases and as database-size grows, it is increasingly desirable to leave octaves unfolded. The alternative numbers in Table 2 suggest the extended range when direction-of-change (up or down) is accommodated.

Another critical element in classical-music repertories for large database searches is enharmonic discrimination. Does the pitch-representation distinguish between enharmonic notes (C# vs. Db)? The base-40 system of pitch-representation [3, Appendix 1, Table 8] supports enharmonic discrimination through double sharps and double flats. This elevated level of pitch representation enables the preservation of the nomenclature of intervallic complementarity customary in music theory [Appendix 1, Table 9] using a single integer. The numeral representation of a major third (12), for example, when added to that of a minor sixth (28) equals that of an octave (40); a minor third (11) when added to a major sixth (29), also equals 40. Such discrimination is of value chiefly for repertories that are likely to include double sharps and flats (i.e., music of the 18th-20th centuries). It is also necessary for extracting correct diatonic scale degrees from highly chromatic music, for example, but it is of little value for folk or popular repertories.

Our combined **pgc**/**rgc** searches employ 9 (3*3) states, since either variable may change independently of the other.

In actual computation, not all possible feature states are used. We find that the number of states encountered at higher-orders of pitch precision varies from repertory to repertory. Also, intervallic searches at higher levels of precision are computed with discrete octaves (i.e., the interval of a perfect twelfth is not equated with that of a perfect fifth), so the range of actual feature states varies with the database and the method. We show selected effects in Table 3.

Search-type	**Actual no. of states required**		
	Classical	**Polish**	**All**
12i	70	40	88
pch	29	26	32
mi	95	52	109

Table 3. Actual numbers of states for various database features.

We would point out here that the **pch** figures represent values folded onto one octave, although the systems can be used over an arbitrary numbers of octaves. The classical dataset is overwhelmingly instrumental, while the Polish data is exclusively vocal.

3.2. Measures of Effectiveness

Entropy is a measure of diversity used in comparisons of information. In mathematic terms, *entropy*, $H(X)$, evaluates the randomness of a variable in terms of how widely spread the probability distribution, $P(X)$, is. Mathematically, it can be shown that $2^{H(X)}$ is never larger than the actual number of states. $H(X)$ is measured in bits. Equation 1 gives the definition of *first-order entropy* [1],

$$H(X) = \Sigma -P_i \log_2 P_i \qquad (1)$$

where P is the probability distribution of X, and the sum of all P_i is 1.

On the other hand, the *entropy rate*, $G(X)$, describes the unpredictability of a random process X^n. A random process X^n with a large entropy does not necessarily have a large entropy rate. Mathematically, it can be shown that $G(X) \leq H(X)$. G is measured in bits/symbol.

If X varies in time and forms a sequence, then its entropy rate, G, is defined as

$$G = H(X^n) / n \qquad (2)$$

where X^n denotes the random vector ($X1$, $X2$, …, X_n), and its entropy can be calculated by summing over all possible sequences.

When combining two features in a search, *joint entropy*, $H(a,b)$, refers to the combined entropy of the features. The joint entropy is always less than or equal to the sum of the individual features. *Mutual information*, $I(a;b)$ is that portion of the joint entropy which is shared by both features (see Figure 1).

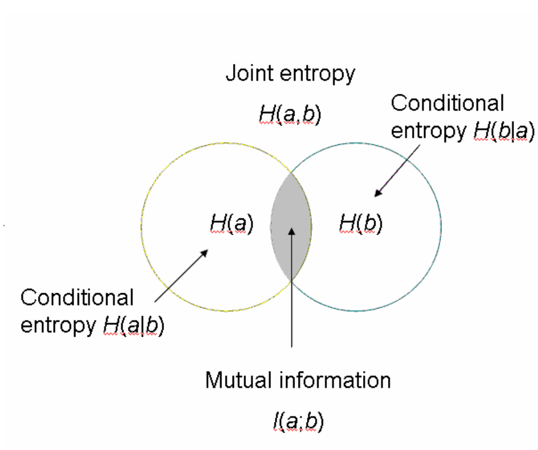

Figure 1. Venn diagram showing relationship of joint entropy, conditional entropy, and mutual information.

We now define two experimental measures related to entropy rate:

Time-to-uniqueness (**TTU**): a measure to establish the minimum number of tokens needed in the search-string to produce a single match candidate.

Time-to-sufficiency (**TTS**): a measure to establish the minimum number of tokens needed to produce not more than K match candidates.

We use TTS to calculate G:

$$G = \log2\ (M/K)/\mathrm{TTS}_K \qquad (3)$$

where M is the total number of unique incipits in a dataset, K defines the cut-off number of match for sufficiency (K = 10), and T is the average time-to-sufficiency of all anchored search.

TTS determines entropy rate more accurately than TTU does, provided that the search-decay rate remains exponential up to the TTS point. We note that the match-rate begins with an exponential decay which flattens out when the match-count become small. This typically happens between the TTS and the TTU points (see Figure 2).

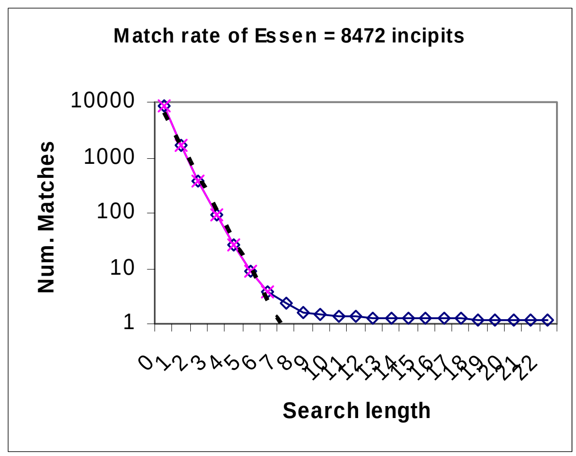

Figure 2. Match-count vs. query length for the complete Essen datasets.

4. PROCEDURES

4.1. Series 1 (Data Features)

We ran tests on the seven levels of pitch-resolution and the one of rhythmic resolution explained in Table 2. With those figures in hand, we reran the first seven tests in combination with rhythmic gross contour (**rgc**). Thus the number of search types run in this series was 15.

4.2. Series 2 (Search Features)

We ran tests in two modes: *anchored* to the beginning of each example and *unanchored*. Anchored searches are suitable for incipits. Unanchored searches may be more useful for themes. They are also potentially useful for searches of fully encoded electronic scores, in which thematic repetition is likely to occur.

For the first two series, we calculated the differences between repertories where appropriate.

4.3. Series 3 (Representational Adequacy)

We tested the results for incipits against the results for full-score searches in the Essen dataset. Our purpose was to determine the extent to which the tests we ran on incipits would be valid for complete works. Since the Essen data is monophonic, it does not fully represent a polyphonic database but gives a best-case notion of the relationship.

5. RESULTS: DATA FEATURES

The most noteworthy preliminary finding was that searches combining pitch and rhythmic features significantly reduced the both TTS and TTU (Table 4).

	TTS	TTU
Uniformly random **pgc** + **rgc** (theoretical)	3.7	4.9
pgc + **rgc** (actual)	4.5	6.9
pgc (actual)	7.2	10.2
rgc (actual)	8.7	10.2

Table 4. Preliminary results from Essen data for separate and combined **pgc** and **rgc** effectiveness measures.

The exercise of running these tests raises a new question about joint-feature searches in general: Should the features be searched in parallel (that is, should coincident matches be assessed for every feature-pair?) or should they be searched serially? We hope to explore this question at a later time.

No TTU could be retrieved for highly generic themes, but the details vary by dataset size and feature-set. (For example, for pgc in the classical set, 11% did not achieve TTU within the full incipit size.) Because certain incipits did not have a unique match, their TTUs were excluded from averages.

The convergence pattern we show in Figure 3 is representative of what we found in all of the databases, although there are small differences in the details from repertory to repertory. The average TTS, given the greatest precision in pitch resolution and an anchored search, was found to be less than 6 [data tokens], but the performance differences between the four most precise levels were minimal. This confirms and refines previous results reported by Schlichte [7] and Howard [4].

Because they were working exclusively with RISM data, we chose the table representing RISM data to show here. The scope of their datasets, which were not identical (nor was ours to either of theirs), are discussed in [4].

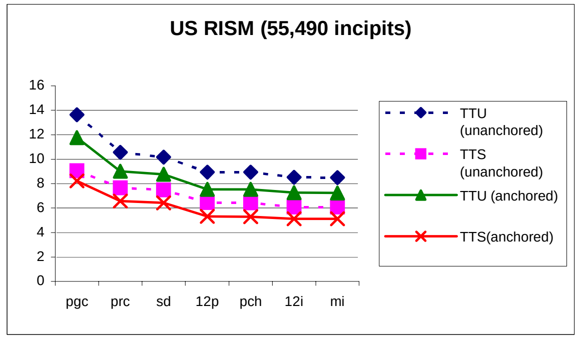

Figure 3. Average search length required for unique (TTU) and sufficient (TTS) matches in the US RISM dataset (55,470 incipits). The X axis identifies the seven levels of pitch (or intervallic) precision. The Y axis indicates the average number of events required to match each type. TTU = time-to-uniqueness. TTS = time-to-sufficiency.

Perhaps the most important finding in this set of tests was that the most significant increases in search-effectiveness come from the progressions in data-precision from pitch/gross-contour (**pgc**) to **prc**, and from scale-degree (**sd**) to 12-tone pitch (**12p**). However, the improvement in performance of **sd** (7-states) over **prc** (5 states) is slight, a finding which is generally similar to [8]. We found insufficient evidence of interdependency between pitch and rhythm (Table 5) to warrant further investigation at this time.

Using conditional entropy, $H(a|b) = H(a,b) - H(b)$, the results are shown in Table 5.

Entropy type	Definition	Value
pgc first-order entropy	H(pgc)	1.5325
rgc first-order entropy	H(rgc)	1.4643
joint first-order entropy	H(pgc, rgc)	2.9900
conditional pitch entropy given the rhythm	H(prc|rgc)	1.5256
conditional rhythmic entropy given the pitch	H(rgc|pgc)	1.4575
mutual information	H(pgc) – H(pgc | rgc)	0.0068

Table 5. Calculation of mutual information from first-order pitch and rhythm entropies. (values based on Essen datasets.)

These tests gave useful insights into the results of the more extensive series of measures we derived from coupling **rgc** with each level of pitch precision (Figure 2). The gains from the addition of **rgc** caused the requisite number of events to decline by 40% (with **pgc**) and by smaller amounts with more precise levels of pitch. (See Tables 4a, b.)

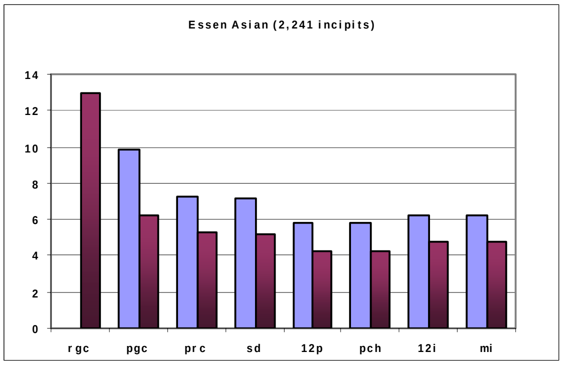

Figure 4a. Pitch-only (seven levels) vs. joint searches with **rgc**. Dark bars show **rgc** alone and successively with each level of pitch precision plotted against the number of events. This view shows results for the Essen-Asian (pentatonic, full-score) dataset. Light bars show seven levels of pitch/interval precision without rhythmic information.

One notable finding is that **rgc** is nearly as effective as **pgc** in the classical repertory (4b) but less so the folksong repertory (4a). The additional gain in combining **rgc** with other levels of pitch-precision is modest.

6. RESULTS: SEARCH FEATURES

In the realm of anchored matches, we attempted to compute the relative gain which might be expected for each single-note addition to a query. Here the dataset contained 8,473 fully encoded folksongs. We found that after the processing of the first five items, the curve representing incremental gain for the most recently added item flattened considerably (Figure 2). Table 6 gives comparative rates for anchored (to the first note) and unanchored searches (i.e., the match could begin anywhere in the song). The average length of the songs in the dataset was 52 notes.

Anchored search		Unanchored search	
Query length	**Matches**	**Query length**	**Matches**
1	1,719.55	1	7,194.33
2	369.83	2	3,179.93
3	91.18	3	883.09
4	26.07	4	221.86
5	8.92	5	55.75
6	3.67	6	15.46
7	2.28	7	5.61
8	1.64	8	2.58
9	1.47	9	1.77
15	1.24	15	1.25
22	1.18	22	1.19

Table 6. Comparison of anchored and unanchored searches, events 1-9, 15, 22 (Essen folksong full scores).

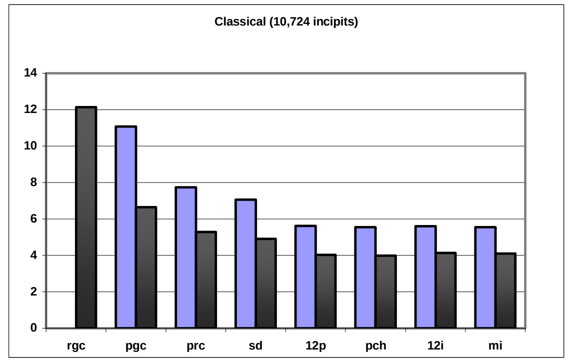

Figure 4b. Pitch-only vs. joint searches with **rgc** as in Figure 4a for the classical theme database.

7. RESULTS: ENTROPY

We attempted to determine how "complex" each of the repertories is. Figure 5 compares the first-order entropy with the experimentally measured entropy rate for the various datasets. The first-order entropy indicates the maximum expected complexity of the music, and the entropy rate quantifies the actual complexity. Figure 5 demonstrates that the classical-theme dataset is the most "complex" from the perspective of informational feature (here **12p**) variety, while the Polish religious-song corpus is the least "complex." The Renaissance dataset has the lowest first-order entropy. However, its entropy-rate is higher with respect to its entropy than any other dataset. This means that even though the Renaissance incipits use fewer twelve-tone pitches than the classical theme set, the incipits are just as "complex."

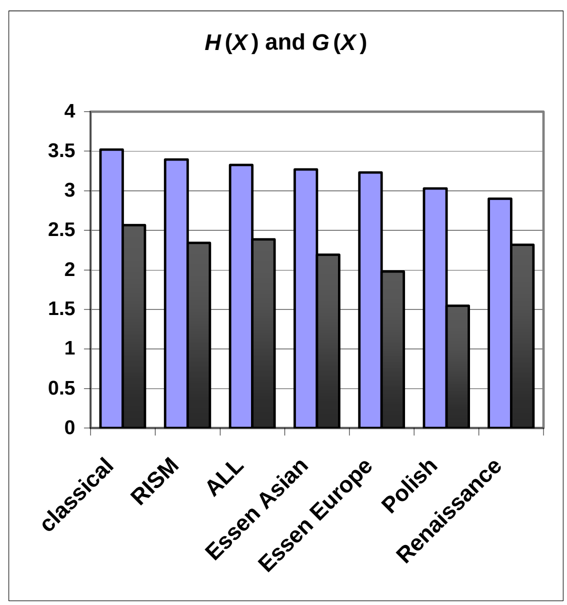

Figure 5. Twelve-tone pitch (**12p**) randomness per repertory, sorted by first-order entropy. Light bars show first-order entropy. Dark bars show entropy rate.

8. RESULTS: INCIPIT ADEQUACY

In another series of tests, we tried to determine what the relationship was between searches of a musical incipit and searches of the work from which the incipit was derived. This question can be paraphrased in two ways: (1) How representative is the incipit of the work as a whole, or (2) How efficient is it to search one in relation to the other? Here our tests delved further into the question of anchored vs. unanchored searches.

Answers to (2) are provided in Table 7. The results suggest that an unanchored search requires, on average, the addition of one event to the query. The repertories used were those of the Essen datasets, which were the simplest ones available. The average number of notes per incipit was 18; the average number of notes per song was 52. This 1:3 relationship would obviously not hold for longer pieces or for polyphonic works in which all parts merit searching.

		Incipit only	**Full work**
Anchored search	TTU (**mi**)	6.87619	8.74826
	Failure rate	0.669%	0.0354%
Unanchored search	TTU (**mi**)	7.926	10.2858
	Failure rate	1.07%	0.0472%

Table 7. Comparison of TTU (computed by interval) for incipits and for the full works from which they originate (results are based on the Essen folksong datasets.)

9. PRACTICAL OBSERVATIONS

Processing time proved to be a significant constraint with the compilation of results for the larger datasets. The primary search engine for the web-based version of *Themefinder* is a PERL script which interfaces ultimately with the Linux command-line program grep with an $O(N)$ search efficiency, where each search takes approximately one second.

Raw search statistics for the complete set of 100,000 incipits require 1.76 million searches for each type of feature, so using the pre-existing search engine was not practical. A search program was therefore written in C which kept the sorted datasets in memory between searches. This program had the ability to perform anchored searches on the dataset of 100,000 incipits at a rate of 487 searches/second for a 1.5 GHz computer.

Obviously, the greater the number of states considered, the longer the processing time required. For anchored searches the computational complexity is $O(\log N)$ where N is the number of incipits in the dataset. For unanchored searches which are searched without presorting, the complexity is $O(N)$, where N is the number of features in the dataset. The processing time for unanchored searches on 100,000 incipits was approximately three days for each feature set.

10. CONCLUSIONS

We cannot report the full range of our findings in the space available, but we call attention to a few of the most significant ones. Other results will be posted in due course at www.ccarh.org.

Combining **rgc** and **pgc** is an effective search strategy since there is little mutual information between the two features. Both features are vague by themselves, but together they are as effective as pitch-only scale-degree searches. Interestingly, incorporating **rgc** with other more specific pitch features yields only marginally better TTS/TTU times (about 1.5 fewer notes required in the search query, as compared to 4.5 fewer notes with **pgc**).

Octave information does not significantly improve search results, since the **pch** and **mi** sets had nearly identical entropy rates. The **pch** set did not encode octave, while the **mi** set preserved octave information. Also, enharmonic spellings are not more effective than twelve-tone search types, since **12p** and **pch** have very similar entropy and entropy-rates.

We refrain from making fine-grained comparisons between our results and those cited above both because of insufficient information about some details of previous work and also because the studies we cited all had motives somewhat different from our own.

ACKNOWLEDGMENTS

The nucleus of the *Themefinder* database was designed by David Huron. Its original implementation on the web was designed by Andreas Kornstaedt. Much of its subsequent development, including data translation, was undertaken by Craig Stuart Sapp, who also maintains it. The bulk of the Essen data was made available originally by the late Helmut Schaffrath and subsequently by Ewa Dahlig. It was translated into the **kern format by David Huron, with later revisions by Jane Singer. The Chinese data was encoded by Baoqiang Han. The Luxembourgian addition was encoded and made available by Damien Sagrillo. The Italian motet data was provided by Harry Lincoln. RISM data was made available by the US RISM Committee. The “classical” data was provided in large part by David Huron and revised by CCARH staff and Stanford University students. The authors extend their cordial thanks to these contributors and to those who wish to remain anonymous.

REFERENCES

[1] Cover, Thomas, and Joy A. Thomas, *Elements of Information theory*, New York: Wiley, 1991.

[2] Dannenberg, Roger B., William P. Birmingham, George Taznetakis, Colin Meek, Ning Hu, Bryan Pardo, “The MuseArt Testbed for Query-by-

Humming Evaluation," ISMIR2003: Fourth Annual International Symposium on Music Information Retrieval, ed. Ed. Holger Hoos and David Bainbridge, Baltimore, 2003.

[3] Hewlett, Walter B. "A Base-40 Numberline Representation of Musical Pitch Notation," *Musikometrika* 4 (1992), 1-14; www.ccarh.org/ publications/reprints/base40/

[4] Howard, John. "Strategies for Sorting Musical Incipits," *Melodic Similarity: Concepts, Procedures, and Applications* (*Computing in Musicology*, 11), ed. Walter B. Hewlett and E. Selfridge-Field. Cambridge, MA: The MIT Press, (1998-99), 119-128.

[5] Lincoln, Harry. *The Latin Motet: Indexes to Printed Collections, 1500-1600*. Ottawa: Institute of Mediaeval Music, 1993.

[6] Rand, William, and William Birmingham, "Statistical Analysis in Music Information Retrieval," *ISMIR 2001: Second Annual International Symposium on Music Information Retrieval*, ed. J. Stephen Downie and David Bainbridge, Bloomington, IN, 2001.

[7] Schlichte, Joachim. "Der automatische Vergleich von 83,243 Musikincipits aus der RISM-Datenbank: Ergebnisse—Nutzen—Perspektiven," *Fontes artis musicae*, 37 (1990), 35-46.

[8] Sora, Timo, and Jyri Huopaniemi, "Melodic Resolution in Music Retrieval, " *ISMIR 2001: Second Annual International Symposium on Music Information Retrieval*, ed. J. Stephen Downie and David Bainbridge, Bloomington, IN, 2001.

APPENDIX 1.

Note name	Value	Note name	Value
Cbb	1	F#	21
Cb	2	F##	22
C	3	.	23
C#	4	Gbb	24
C##	5	Gb	25
.	6	G	26
Dbb	7	G#	27
Db	8	G##	28
D	9	.	29
D#	10	Abb	30
D##	11	Ab	31
.	12	A	32
Ebb	13	A#	33
Eb	14	A##	34
E	15	.	35
E#	16	Bbb	36
E##	17	Bb	37
Fbb	18	B	38
Fb	19	B#	39
F	20	B##	40

Table 8. Numerical values representing note-name and inflection in the base-40 system. The number 40 represents one octave. Multiples of 40 represent octave. (Alternatively, to preserve the octave number for diatonic pitch names, the values can be arrayed from 0 to 39). Five positions are null.

Interval name	Value		Interval name	Value		Total
P1	0	+	P8	40	=	40
aug1	1	+	dim8	39	=	40
		+				
dim2	4	+	aug7	36	=	40
min2	5	+	maj7	35	=	40
maj2	6	+	min7	34	=	40
aug2	7	+	dim7	33	=	40
		+				
dim3	10	+	aug6	30	=	40
min3	11	+	maj6	29	=	40
maj3	12	+	min6	28	=	40
aug3	13	+	dim6	27	=	40
		+				
dim4	16	+	aug5	24	=	40
P4	17	+	P5	23	=	40
aug4	18	+	dim5	22	=	40

Table 9. Interval states used in **mi** feature searches showing their equivalent base-40 numeric values. The table demonstrates the intervallic complementarity of the base-40 system.