Music Retrieval as Text Retrieval: Simple Yet Effective

J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St, Champaign IL
(217) 352-7422
jdownie@uiuc.edu

ABSTRACT

This poster reports on the latest findings of the author concerning the development of a simple approach to the storage and retrieval of music information. Using the McNab et al. collection of 9354 folksongs [1], we have developed and tested a series of test databases where monophonic melodies are represented as a collection of interval-only n-grams (i.e., length-n substrings of the signed differences between pitches). These melodic n-grams of length-4, 5 and 6 can be treated like artificial "words". By treating n-grams as "words" we have been able to apply traditional text retrieval methods to the music information retrieval (MIR) problem. The traditional text retrieval system, SMART, was used to test the hypothesis that music information can indeed be treated as text. Randomly selected extracts from the databases were used to simulate potential queries. The results were evaluated using the standard text retrieval normalized precision and recall measures. Several test databases performed very well, confirming the notion that a simple, text-styled, approach to MIR is indeed feasible

1. INTRODUCTION

Stated informally, the factors found in Table 1 were examined in an attempt to answer the following questions, respectively:

1. Does the size of the classificatory set used in the creation of the n-gram representations affect performance (CLASS)? If it does not, then one would prefer to represent the melodies using the smallest classificatory set as this would reduce index size. A reduction in index size is to be preferred *ceteris paribus*.

2. Does the length of the n-gram representations affect performance (NLEN)? If it does not then one would prefer to represent the melodies using the shortest n-gram length, as this would reduce index size. Again, a reduction in index size is to be preferred *ceteris paribus*.

3. Does the location of the query affect retrieval effectiveness? Can users submit queries that represent internal phrases or must they try to match only the beginnings (i.e., *incipits*) of the melodies (QLOC)? The ability to search for internal phrases would greatly enhance the utility of an MIR system.

4. How much of the melody will the users have to remember (QLEN)? The probability of query error increases with query length. An MIR system that minimizes the necessary query length, while still performing adequately, is to be preferred *ceteris paribus*.

5. Do minor query errors affect performance (QQUAL)? An MIR system that minimizes the effect of minor query errors, while still performing adequately, is to be preferred *ceteris paribus*.

Table 1. Experimental factors.

Factor Name	Short Form	Definition	Codes Used	Comments
*C*lassification	CLASS	The number of interval classes used to represent melodies (i.e., size of alphabet)	*CU*	Intervals are taken as given in melody
			C7	7 intervals used
			C15	15 intervals used
N-gram *L*ength	NLEN	Number of contiguous intervals in each n-gram	*L4*	length-4 string
			L5	length-5 string
			L6	length-6 string
*Q*uery Length	QLEN	Number of contiguous intervals in a string used as a query	*Q4*	length-4 string
			Q7	length-6 string
			Q10	length-8 string
Query Location	QLOC	Position in a song from which a query string is extracted	*I*	*Incipit*: query string extracted from a song starting at song's first interval
			R	*Random*: query string extracted from a song starting anywhere but song's first interval
Query Quality	QQUAL	Indicates whether query string represents a perfectly formed query or one with an error present	*P*	*Perfect*: query string is taken as extracted from song then undergoes Classification
			E	*Error*: query string has one interval randomly changed prior to Classification

2. Method

Melodic strings were extracted from the songs in the database (30 Incipit, 30 Random). These strings were used as the basis of the queries, subject to the various treatments. The experimental model was a complex, mixed, five-way factorial design. Thus, retrieval effectiveness was evaluated under 108 experimental conditions with a total of 3240 queries run. Data were analyzed under a Repeated Measures General Linear Model using SPSS. To meet the assumptions of normality required by Analyses of Variance (ANOVA), the data were subjected an ArcSin(Sqrt(x))) transformation

determination of a "best" combination of factors can be made. For example, *CUL6P* was the superior performer while *CUL6E* was the worst. Notwithstanding the inherent complexities of analysis the within-subject tests do show:

CLASS: C15 superior to C7 No significant difference C15–CU

NLEN: L5 superior to L6 No significant difference L4–L5

QLEN: Q10 superior to Q8 Q8 superior to Q6

QLOC: No significant difference I–R

QQUAL: P superior to E

Table 2. Normalized precision (averaged over queries of all lengths)

	NLEN								
	L4			L5			L6		
	QQUAL			QQUAL			QQUAL		
CLASS	P	E	AVE(P&E)	P	E	AVE(P&E)	P	E	AVE(P&E)
C7	.8874	.6319	.7596	.9330	.6109	.7720	.9485	.5114	.7300
C15	.9438	.6909	.8174	.9781	.6481	.8131	.9905	.4883	.7394
CU	.9445	.6952	.8199	.9807	.6489	.8148	.9927	.4869	.7398
AVE(NLEN)	.9252	.6727	.7990	.9639	.6360	.8000	.9773	.4955	.7364

Table 3. Select results of the tests of within-subject effects, normalized precision (ArcSin(Sqrt(x))

Source	CLASS	NLEN	QLEN	QQUAL	df	F	Sig.
CLASS	C7 vs. C15				1	24.8	0.00
NLEN		L5 vs. L6			1	30.85	0.00
QLEN			Q6 vs. Q8		1	21.89	0.00
			Q8 vs. Q10		1	55.44	0.00
QQUAL				P vs. E	1	442.56	0.00
CLASS * NLEN	C7 vs. C15	L5 vs. L6			1	18.75	0.00
CLASS * QQUAL	C7 vs. C15			P vs. E	1	18.02	0.00
NLEN * QQUAL		L4 vs. L5		P vs. E	1	63.82	0.00
		L5 vs L6		P vs. E	1	73.76	0.00
QLEN * QQUAL			Q6 vs. Q8	P vs. E	1	14.84	0.00
			Q8 vs. Q10	P vs. E	1	26.99	0.00

3. Key Findings[1]

With regard to normalized precision (NPREC) (Tables 2) the overall results were quite positive The nominally best performance came from the *CUL6P* condition (Unclassified intervals, Length-6 n-gram, Perfect query) which had a NPREC (averaged over all query lengths) of .9927. The nominally best performance under the Error condition was *CUL4E* (.6952). *CUL4* also returned the best performance when the Perfect and Error conditions were averaged (.8199).

Tests of within-subject effects (Table 3) indicate that analyses of the results have to go beyond selecting the nominally best performers. The results of the within-subject tests bring out two important facts. First, there were situations where the superior results were not significantly different from the next-best (e.g., (*CU* =*C15*) >C7). Second, the persistent interaction of the factors with the QQUAL factor makes it apparent that no clear-cut

[1] Space constraints limit the present discussion to the NPREC results.

4. Summary and Conclusionss

We have shown that useful MIR systems can be constructed using traditional text retrieval methods. This implies that the use of interval-only n-grams to represent monophonic folksongs in a traditional text retrieval environment (e.g., WWW, library catalogue, etc.) would require only ***minor*** modifications to those environments (i.e., interface) to create a useful integrated MIR system.

5. Acknowledgements

We thank Drs. .M. Nelson, Y. Quintana, B. Frohmann, and R. Wood, whose support made this research possible. Roger McNab must be thanked for allowing us to use his collection of folksongs

6. Reference

[1] McNab, Rodger J., Lloyd A. Smith, Ian H. Witten, Clare Henderson, and Sally Jo Cunningham. Towards the digital music library: tune retrieval from acoustic input. In *Digital Libraries '96: Proceedings of the ACM Digital Libraries Conference, Bethseda.*