Toward Music-Based Data Discrimination for Cybercrime Investigations♦

George R S Weir[1] and Gerry Rossi[2]

Department of Computer and Information Sciences[1]
Department of Music[2]
University of Strathclyde, Glasgow
G1 1XH, UK
george.weir@cis.strath.ac.uk, gerry.rossi@strath.ac.uk

Abstract. In this paper we describe an approach to data interpretation in which 'raw' data is analysed quantitatively in terms of textual content and the results of this analysis 'converted' to music. The purpose of this work is to investigate the viability of projecting complex text-based data, via textual analysis, to a musical rendering as a means for discriminating data sets 'by ear'. This has the potential of allowing non-domain experts to make distinctions between sets of data based upon their listening skills. We present this work as a research agenda, since it is based upon earlier exploration of the underlying concept of mapping textual analyses to music, and explore possible areas of application in the domains of information security and digital forensics.

Keywords: Data interpretation, textual analysis, musical mapping.

1 Introduction

One major issue facing investigators in cybercrime is the need to analyse large quantities of data with a view to uncovering any salient evidence of criminal activity. Such data may comprise terabytes of files from a suspect's computer, or it may consist of captured digital communications between individuals, e.g., as chat logs or email conversations. In such cases, the quantity and complexity of the data is often a significant obstacle to fast turnaround of evidence and this may jeopardise the prospects for securing a conviction.

With such scenarios in mind, there are growing efforts to accelerate the process of data analysis, e.g., through 'triage' software that undertakes to give a speedy indication of suspicious content [1], parallel data analysis [2] or through sophisticated data mining techniques that explore relationships across the data that may otherwise be hard to spot [3].

A second major issue facing law enforcement is the limited availability of suitably skilled digital investigators. Timely analysis of suspect data may also be

♦ Reprinted from: *Cyberforensics: Issue and Perspectives*. Edited by G. R. S. Weir. University of Strathclyde Publishing. 2011.

compromised if there are insufficient experts available. This issue may be addressed either by greater investment in training and manpower or through reduction in the levels of expertise required in order to conduct the digital investigation. Development and research on software tools and analytical methods (such as those cited earlier) may also assist in addressing the required human level of technical sophistication.

Although our work in this area is at an early stage, we believe that our approach to textual analysis combined with mapping of the resultant analysis data to music may afford benefits in addressing each of the two issues described above (data quantity and investigator skill level).

Through textual analysis of cybercrime data and musical articulation of the results, data may be 'processed' in parallel by the human investigator. One channel of analysis employs aural data, which may be heard while a second data channel is explored by conventional digital forensic means. In this fashion, the analyst can explore more data sets in a given time.

Since music is our method of data expression, the analyst may be able to discriminate between data sets in virtue of their musical content, i.e., how they sound, and this may require less technical skill than conventional forensic investigation.

2 Textual Analysis and Music

Our recent research combines two disparate areas of study, textual analysis and musical composition. The main purpose of the former is to apply software tools toward revealing features within collections of text. In this area, we have developed the Posit text profiling tools [4, 5] which produce a quantitative analysis for any given text in terms of its individual words and multi-word units, including part-of-speech data. On this basis, differences between texts, in terms of their lexical content, can be highlighted and the analysis results used to classify or differentiate within or across sets of texts (for example, see [6]).

Such textual analysis already has plausible application in a variety of cybercrime contexts. For instance, syntactic analyses of email may contribute to spam assessment and analysis of textual content may assist in proving or disproving the authorship of a message. Such cases may require a high level of technical expertise on the part of the textual analyst, who may additionally be required to appreciate aspects of digital forensics. This is a major motivation toward production of amenable expressions of the textual analysis data and is where we turn our attention to musical composition.

There are many different ways in which textual characteristics may be mapped to music (and not merely to sound). Clearly, both domains have many different variables and this scope provides us with both an opportunity and a challenge. The opportunity lies in the range of different possibilities that may be explored as a basis for the musical projection of the textual data. We consider this to be an opportunity because the large mapping space in each domain will afford innumerable different possible ways of 'converting' the textual features to musical features. Not all musical renderings are likely to effective, either as acceptable music, or as expressions of significant textual attributes but the extensive scope should increase the likelihood of

one or more effective renderings across these two domains. The challenge that arises from the large mapping space in each domain is to identify mappings that satisfy these two requirements, viz., the result is meaningful (revealing) and is also recognizably musical.

For example, we have discretion on how we map the number of words in a document. This could be mapped to the tempo of the music. Similarly, instances of each part of speech could be mapped to different notes on the major scale, and so on. Whatever specific mappings are selected should accommodate the 'significant' information available from the textual analysis.

In addition to the requirement that selected mapping is able to accommodate the salient textual characteristics (or rather, the textual analysis data characteristics) is our further constraint that the result from the selected set of mappings should be acceptable as music (i.e., with all the main musical characteristics: tempo, key signature, and note variations within the same scale) and should sound pleasant. This is a further challenging aspect of this work.

2.1 Text to Music Prototype

A component in this work was the creation of a software application to convert textual analysis data to music [7]. This prototype, called PlayText, takes as input several data files generated as a result of analyzing the text using the Posit tools. From these data files, PlayText creates a midi file according to the selected text to music mapping. Presently, the system supports only two sets of mapping rules.

The first mapping set gives precedence to parts-of-speech in the analysed text and maps the words to notes on this basis. At its simplest, nouns can be mapped to the tonic note of the scale; verbs can be mapped to the supertonic note, and so on. In our adopted approach, we took values based upon the proportion of the different parts of speech in the sampled text and used this to determine the note value. Part of speech ratios were used rather than the absolute frequency values for parts of speech because the ratio affords a clearer distinction between documents.

In order to ensure that the musical result was sufficiently 'interesting', we added variable durations to notes and rests. Note duration was based on variations of the parts of speech (e.g., proper nouns get quarter duration, plural nouns get half duration and so on). Musical rests and their duration were based upon the punctuation marks within the analysed texts. Full stops were mapped to a one beat rest; colons to a half beat rest; commas to a quarter beat rest, and so on. The idea behind this was that the flow of the music should follow the flow of the document (i.e. when we read something we stop at full stops longer than we stop at commas, hence the music should pause longer for full stops than for commas).

The second mapping set derives the melody notes from the ratio of the frequencies of the words (i.e. the value of the ratio is the value of the note). The mapping is based on the ratios and not the absolute values for the reasons explained above. This second mapping set also adds rests to the music according to punctuation marks in the original text document. For some punctuation marks though, such as question mark and exclamation mark instead of putting a rest, we put a note and changed its attack velocity (how 'hard' the note will sound). This was motivated by the way that we

change our voice when we read; for question marks we alter the tone of our voice, for exclamation marks we alter it even more. Accordingly, the notes for the exclamation marks have a bigger attack velocity than the notes for the question marks.

Using a set of sample textual analysis data generated from text documents by the Posit tools, we applied the prototype Playtext system to produce sets of musical interpretations in each of our two mappings. Using these examples, we designed a small experiment to gauge whether these melody mappings would affect users' ability to discriminate textual data on the basis of its musical interpretation. In this experiment, eight subjects were given six midi files from each mapping, and a form to complete for each mapping. Each of the midi files in a set was generated from different documents and the experiment had two components. In the first component, the testers, listened to the midi files and tried to 'guess' the genre of the document they came from. For example, midi file 1 is fiction; midi file 2 is poetry and so on. Six genres were listed in the form provided to the subjects: fiction, poetry, mythology, drama, history, and horror tales. On the basis of our initial choice of text to music mappings, we were interested to determine whether the resultant music suggested a document genre to the listener. In other words, could the mapping convey such information about the document?

The second exercise required that the subjects listen again to the same twelve midi files only this time try to 'guess' which files, if any, were from the same author. In this case, there was no prepared list of suggested authors and subjects were not required to propose specific authors for any document. Our aim was to test whether documents from the same author had any discernable musical similarity. The documents used with the 1st mapping are shown in Table 1 while the documents used with the 2nd mapping are shown in Table 2.

Table 1. Documents used with Mapping 1.

Author	**Title**	**Genre**
John Keats	Endymion	Poetry
Edgar Alan Poe	The Fall of the House of Usher	Horror Tales
Charles Dickens	A Christmas Carol	Fiction
Bram Stoker	Dracula	Horror Tales
Edgar Alan Poe	The Raven	Poetry
Oscar Wilde	The Picture of Dorian Gray	Fiction

Table 2. Documents used with Mapping 2.

Author	**Title**	**Genre**
William Shakespeare	Hamlet	Drama
Jules Verne	Around the World in 80 Days	Adventure
Homer	Iliad	Mythology
William Shakespeare	Othello	Drama
Edgar Alan Poe	A Dream within a Dream	Poetry
Gaston Leroux	The Phantom of the Opera	Fiction

In the first part of the experiment (genres), mapping 1 allowed the subjects to correctly classify 4 out of the 6 midi files. Mapping 2 allowed the subjects to correctly classify 3 out of 6 midi files. In the second part of the experiment (authors), using mapping1, one subject correctly identified the midi files that came from documents of the same author. For mapping 2, no one found the midi files that came from documents of the same author. Overall, mapping 1 seemed to be more 'successful' in distinguishing the documents.

This small test set does not fully establish the viability of document discrimination via musical interpretation but it does hint at the viability of such an approach. This is all the more plausible when considering the small range of mapping features that were employed for this experiment. Countless variations in mapping sets remain to be explored.

3 Applications in Cybercrime

The research agenda proposed in this paper is to investigate the feasibility of using musical renderings of textual data analyses as a means of discriminating between data sets in the cybercrime domain. We anticipate a range of potential application areas. Most readily amenable to such treatment will be contexts in which the data sets are primarily or entirely textual in nature. If we consider data sets such as chat logs or email data, a range of questions come to mind, such as 'Can we identify the gender of participants in a chat session through musical rendering of the textual content?', 'Are there characteristics of textual content in on-line grooming activities that may allow us to identify such behaviour through musical projection?'.

Other digital forensics issues may also be amenable to such treatment. For instance, all of the textual content from a suspect's hard drive may be subjected to musical projection as a means of discriminating within the data set.

Our approach holds the potential for a novel variety of data exploration. As yet, we cannot predict the data features that will most effectively lend themselves to this treatment nor which data to music mapping will prove most accessible to the listener. However, this approach is motivated by the belief that such discrimination may allow us to move away from reliance upon cybercrime or forensic expert interpretation of source data toward a generalist aural approach that simply considers and contrasts the music that results from 'interpretation' of the analysis data.

We are presently developing a new 'test-bench' software application (Textaural) to afford experimentation with the 'mapping scope' and enable us to explore the strengths and weaknesses of different text to music mappings. A central consideration in this work is the range of possible mappings from textual characteristics to musical characteristics (the mapping scope). This is to be explored through extrapolation of measures available from existing textual analysis tools (such as type/token ratio, proportion or ratios of different parts-of-speech, and internal/external word or n-gram frequency) and the elucidation of available musical features (such as pitch, tempo, attack, mode, key signature, harmony). We expect that different mapping sets of textual data to musical features will be required for optimal aural discrimination of different document classification contexts. This can only be

validated experimentally through use of the features proposed for the Textaural application. On the basis of these experiments, we can work toward optimised data discrimination by music and explore a range of application contexts associated with cybercrime, security and digital forensics.

References

1. Rogers, M., Goldman, J., Mislan, R. and DeBrota, S. Computer Forensics Field Triage Process Model. Conference on Digital Forensics, Security and Law, Arlington, VA, 2006.
2. Lodovico, M., Richard III, G. and Roussev, V. Massive threading: Using GPUs to increase the performance of digital forensics tools. Digital Investigation, 1:73–81, September 2007.
3. Beebe, N. L. and J. G. Clark, "Dealing with Terabyte Datasets in Digital Investigations," in Research Advances in Digital Forensics, M. Pollitt and S. Shenoi, Eds. Norwell: Springer, 2005, pp. 3-16.
4. Weir, G. R. S. The Posit Text Profiling Toolset. Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics. Pan-Pacific Association of Applied Linguistics. December 2007.
5. Weir, G. R. S., Corpus Profiling with the Posit Tools. Proceedings of the 5th Corpus Linguistics Conference. University of Liverpool. July 2009.
6. Weir, G. R. S. and Ozasa, T. Learning from Analysis of Japanese EFL Texts. Educational Perspectives, Journal of the College of Education, University of Hawai‘i at Manoa. 43 (1 & 2). 2010. pp.56-66.
7. Weir, G. R. S. and Livitsanou, M. Playing Textual Analysis as Music. Proceedings of ICTATLL 2010, Kyoto, Japan.. ICTATLL. September 2010.