

Empirical Classification for Musical Information Retrieval

Rory A. Lewis¹ and Alicja Wieczorkowska²

¹ University of North Carolina, 9201 University City Blvd. Charlotte, NC 28223, USA

² Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008
Warsaw, Poland

Abstract. In the continuing goal of codifying the classification of musical sounds and rules for data mining we present the following classification methodology that produces a hierarchical granulation structure. The motivation for this paper is based upon the fallibility of Hornbostel/Sachs, the generic classification scheme used in Music Information Retrieval for instruments. In eliminating the redundancy and discrepancies of Hornbostel/Sachs' classification of musical sounds we present a procedure that reduces the musical attributes empirically and defines a robust rule discovery. Rather than using classification rules based directly on Hornbostel/Sachs, we rely on the empirical data of the log attack, sustainability and harmonicity. We propose a classification system based upon the empirical musical parameters and then incorporating the resultant hierarchical granulation structure for classification rules.

1 Instrument Classification

Information retrieval of musical instruments and their corresponding sounds has invoked a need to constructive cataloguing conventions with specialized vocabularies and other encoding schemes. For example the Library of Congress subject headings [1] and the German Schlagwortnormdatei Decimal Classification both use the Dewey classification system [2, 3] In 1914 Hornbostel-Sachs devised a classification system, based on the Dewey decimal classification which essentially classified all instruments into strings, wind and percussion. Later it went further and broke instruments into four categories: 1.1 Idiophones, where sound is produced by vibration of the body of the instrument and 2.2 Membranophones, where sound produced by the vibration of a membrane. 3.3 Chordophones, where sound is produced by the vibration of strings and 4.4 Aerophones, where sound is produced by vibrating air.

For purposes of music information retrieval, the Hornbostel-Sachs cataloguing convention is problematic, since it contains numerous exceptions and follows a subjective humanistic conventions making it incompatible for a knowledge discovery discourse. For example, a piano emits sound when the hammer strikes strings. The striking of an object is percussive yet its the string that emits the sound vibrations which also classifies it possibly as a chordophone. Similarly, the tamborine comprises a membrane and bells making it both an membranophone

and an idiophone. Considering this paper presents a basis for an empirical music instrument classification system conducive for music information retrieval, specifically for automatic indexing of music instruments.

2 A three-level empirical tree

We focus on three properties of sound waves that a machine can differentiate and they are log-attack, harmonicity and sustainability. These properties are part of the set of descriptors for audio content description provided in the MPEG-7 standard and have aided us in musical instrument timbre description, audio signature and sound description.[4]

2.1 LogAttackTime

The motivation for using the MPEG-7 temporal descriptor, LogAttackTime (LAT), is because segments containing short LAT periods cut generic percussive and harmonic signals into two separate groups.[5, 6] The *attack* of a sound is the first part of a sound, before a real note develops where the LAT is the logarithm of the time duration between the point where the signal starts to the point it reaches its stable part.[7] The range of the LAT is defined as $\log_{10}(\frac{1}{\text{samplingrate}})$ and is determined by the length of the signal. Struck instruments, such as most percussive instruments have a short LAT whereas blown or vibrated instruments contain LATs of a longer duration.

$$LAT = \log_{10}(T1 - T0), \quad (1)$$

where $T0$ is the time the signal starts; and $T1$ is reaches its sustained part (harmonic space) or maximum part (percussive space).

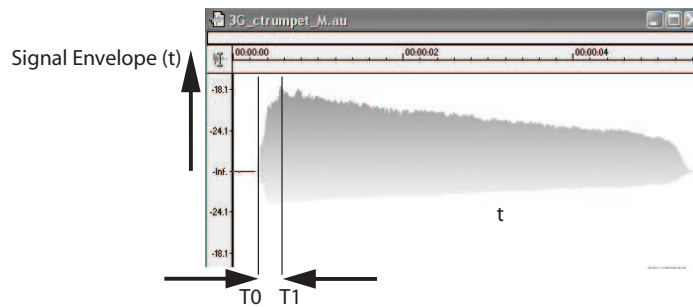


Fig. 1. Illustration of log-attack time. $T0$ can be estimated as the time the signal envelope exceeds .02 of its maximum value. $T1$ can be estimated, simply, as the time the signal envelope reaches its maximum value.

2.2 AudioHarmonicityType

The motivation for using the MPEG-7 descriptor, AudioHarmonicityType is that it describes the degree of harmonicity of an audio signal.[6] Most "percussive" instruments contain a latent indefinite pitch that confuses and causes exceptions to parameters set forth in Hornbostel-Sachs. Furthermore, some percussive instruments such as a cuico or guido contain a weak LogAttackTime and therefore fall into non-percussive cluster while still maintaining an indefinite pitch. The use of the descriptor AudioHarmonicityType theoretically should solve this issue. It includes the weighted confidence measure, SeriesOfScalarType that handles portions of signal that lack clear periodicity. AudioHarmonicity combines the ratio of harmonic power to total power: HarmonicRatio, and the frequency of the inharmonic spectrum: UpperLimitOfHarmonicity.

First: We make the Harmonic Ratio $H(i)$ the maximum $r(i, k)$ in each frame, i where a definitive periodic signal for $H(i) = 1$ and conversely white noise = 0.

$$H(i) = \max r(i, k) \quad (2)$$

where $r(i, k)$ is the normalised cross correlation of frame i with lag k :

$$r(i, k) = \frac{\sum_{j=m}^{m+n-1} s(j) s(j-k)}{\left(\sum_{j=m}^{m+n-1} s(j)^2 * \sum_{j=m}^{m+n-1} s(j-k)^2 \right)^{\frac{1}{2}}} \quad (3)$$

where s is the audio signal, $m=i*n$, where $i=0, M-1$ =frame index and M = the number of frames, $n=t*sr$, where t = window size (10ms) and sr = sampling rate, $k=1, K=lag$, where $K=\omega*sr$, ω = maximum fundamental period expected (40ms)

Second: Upon obtaining the i) DFTs of $s(j)$ and comb-filtered signals $c(j)$ in the AudioSpectrumEnvelope and ii) the power spectra $p(f)$ and $p'(f)$ in the AudioSpectrumCentroid we take the ratio f_{lim} and calculate the sum of power beyond the frequency for both $s(j)$ and $c(j)$:

$$a(f_{lim}) = \frac{\sum_{f=f_{lim}}^{f_{max}} p'(f)}{\sum_{f=f_{lim}}^{f_{max}} p(f)} \quad (4)$$

where f_{max} is the maximum frequency of the DFT.

Third: Starting where $f_{lim} = f_{max}$ we move down in frequency and stop where the greatest frequency, f_{ulim} 's ratio is smaller than 0.5 and convert it to an octave scale based on 1 kHz:

$$UpperLimitOfHarmonicity = \log_2(f_{ulim}/1000) \quad (5)$$

2.3 Sustainability

Initially, we are defining the sustainability into whether the instrument can maintain a particular pitch without dampening for more than 10 seconds. For example, a flutist, horn player and violinist can maintain a singular note for more than ten seconds. Conversely a plucked guitar or single drum note typically cannot sustain that one sound for more than ten seconds. It is true that a piano with pedal could maintain a sound after ten seconds but a damping factor would be present. Consider a given sound sample, referred to as an object $u \in U$. We can split it onto, say, 7 intervals of equal width. Average values of amplitudes within these intervals are referred to as $Amp.1, \dots, 7$. Sequence [8]

$$\overrightarrow{Amp}(u) = \langle Amp.1(u), \dots, Amp.7(u) \rangle \quad (6)$$

corresponds to a kind of envelope, approximating the behavior of amplitude of each particular u in time. We can consider, e.g., Euclidean distance over the space of such approximations. Then we can apply one of basic clustering or grouping methods to find the most representative envelopes. In Figure 2 we show representative envelopes as centroids obtained from the algorithm dividing data onto 6 clusters.

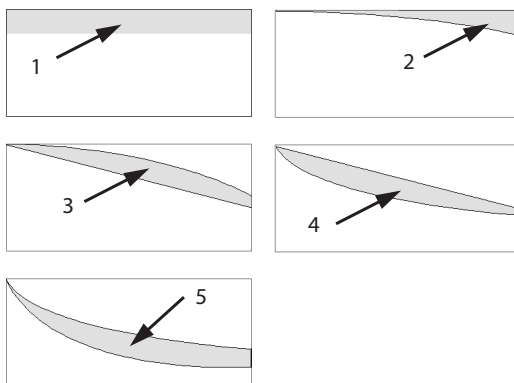


Fig. 2. Centroids (the most typical shapes) of sound envelopes, used in clustering.

3 Experiments

The sound data are taken from our online database at <http://www.miruncc.edu> which contains 4,461 segmented sounds mostly from MUMS audio CD's that contain samples of broad range of musical instruments, including orchestral ones, piano, jazz instruments, organ, etc. [9]. These CD's are widely used in musical

instrument sound research [10–15], so they can be considered as a standard. The database consists of 188 samples each representing just one sample from group that make up the 4,461 files in the database. Mums divides the database into the following 18 classes: violin vibrato, violin pizzicato, viola vibrato, viola pizzicato, cello vibrato, cello pizzicato, double bass vibrato, double bass vibrato, double bass pizzicato, flute, oboe, b-flat clarinet, trumpet, trumpet muted, trombone, trombone muted, French horn, French horn muted, and tuba. However, this is the point of the paper, we show a novel, empirical methodology of dividing sounds conducive to automatic retrieval of music using rough sets.

4 Testing using rough sets

We use Bratko’s Orange on 188 files consisting of one sound of the 3600 files in the data base. Scripting is in Python.

5 Results

The ...

5.1 Resulting Trees

The ...

6 Conclusion

The ...

References

- [1] Brenne, M.: Storage and retrieval of musical documents in a FRBR-based library catalogue: Thesis, Oslo University College Faculty of journalism, library and information science,(2004).
- [2] Doerr, M.: Semantic Problems of Thesaurus Mapping: *Journal of Digital Information*. Volume 1, issue 8, Article No. 52, 2001-03-26, 2001–03, (2001).
- [3] Patel, M. and Koch, T. and Doerr, M. and Tsinaraki, C.: Semantic Interoperability in Digital Library Systems. IST-2002-2.3.1.12 *Technology-enhanced Learning and Access to Cultural Heritage*. UKOLN, University of Bath, (2005).
- [4] Wiczorkowska, A.A. and Ras, Z.W. and Tsay, L.S.: Representing Audio Data by FS-Trees and Adaptable TV-Trees: Foundations of Intelligent Systems, *Proceedings of ISMIS Symposium*, Maebashi City, Japan, LNAI, Springer-Verlag, Volume 2871, 135–142, Springer, (2005).
- [5] Gomez, E. and Gouyon, F. and Herrera, P. and Amatriain, X.: Using and enhancing the current MPEG-7 standard for a music content processing tool, *Proceedings of the 114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, March, (2003).

- [6] Information Technology Multimedia Content Description Interface Part 4: Audio. ISO/IEC JTC 1/SC 29, Date: 2001-06-9. ISO/IEC FDIS 15938-4:2001(E) ISO/IEC J/TC 1/SC 29/WG 11 Secretariat: ANSI, (2001)
- [7] Peeters, G., McAdams, S. and Herrera, P.: Instrument sound description in the context of MPEG-7: in *Proceedings of the International Computer Music Conference (ICMC'00)*, Berlin, Germany, (2000).
- [8] Wieczorkowska, A., Wróblewski, J., Synak, P. and Słezak, D.: Application of temporal descriptors to musical instrument sound recognition: in *Proceedings of the International Computer Music Conference (ICMC'00)*, Berlin, Germany, (2004).
- [9] Opolko, F. and Wapnick, J. (1987). MUMS – McGill University Master Samples. CD's.
- [10] Cosi, P., De Poli, G., and Lauzzana, G. (1994). Auditory Modelling and Self-Organizing Neural Networks for Timbre Classification *Journal of New Music Research*, 23, 71–98.
- [11] Martin, K. D. and Kim, Y. E. (1998). 2pMU9. Musical instrument identification: A pattern-recognition approach. 136-th meeting of the Acoustical Soc. of America, Norfolk, VA.
- [12] Wieczorkowska, A. (1999b). Rough Sets as a Tool for Audio Signal Classification. In Z. W. Ras, A. Skowron (Eds.), *Foundations of Intelligent Systems* (pp. 367–375). LNCS/LNAI 1609, Springer.
- [13] Fujinaga, I. and McMillan, K. (2000). Realtime recognition of orchestral instruments. *Proceedings of the International Computer Music Conference* (141–143).
- [14] Kaminskyj, I. (2000). Multi-feature Musical Instrument Classifier. *MikroPolyphonie* 6 (online journal at <http://farben.latrobe.edu.au/>).
- [15] Eronen, A. and Klapuri, A. (2000) Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2000* (753–756). Plymouth, MA.
- [66] Wieczorkowska, A., Synak P., Lewis, R., Ras, Z.: Extracting Emotions from Music Data: In *Proceedings of Foundations of Intelligent Systems: 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, May 25-2* Volume 3488, 456 – 460, (2005).
- [67] Wieczorkowska, A., Synak P., Lewis, R., Ras, Z.: Creating Reliable Database for Experiments on Extracting Emotions from Music: *Intelligent Information Processing and Web Mining IIPWM*, : Gdansk, Poland, June 13-16 **Advances in Soft Computing - Springer-Verlag**. (2005) 395 – 402 Springer-Verlag.