Categorization of Musical Instrument Sounds Based on Numerical Parameters

2 authors:

Rory A. Lewis
University of Colorado Colorado Springs
170 PUBLICATIONS **8,858** CITATIONS

SEE PROFILE

Alicja Wieczorkowska
Polish-Japanese Academy of Information Technology
103 PUBLICATIONS **984** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   AR in multimedia Communication Design View project

Project   Analytics and Applications of Audio and Image Sensing Techniques View project

Categorization of Musical Instrument Sounds Based on Numerical Parameters

Rory A. Lewis[1] and Alicja Wieczorkowska[2]

[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 this paper we present methodology of categorization of
musical instruments sounds, aiming at the continuing goal of codifying
the classification of these sounds for automating indexing and retrieval
purposes. The proposed categorization is based on numerical parameters.
The motivation for this paper is based upon the fallibility of Hornbostel
and Sachs generic classification scheme, most commonly used for cat-
egorization of musical instruments. In eliminating the discrepancies of
Hornbostel and Sachs' classification of musical sounds we present a pro-
cedure that draws categorization from numerical attributes, describing
both time domain and spectrum of sound, rather than using classification
based directly on Hornbostel and Sachs scheme. As a result we propose a
categorization system based upon the empirical musical parameters and
then incorporating the resultant structure for classification rules.

1   Introduction

Categorization of musical instruments into groups and families, although already
elaborated in a few ways, is still disputable. Basic categorization, commonly used,
is called Sachs and Hornbostel system [7], [14]. It is based on sonorous material
producing sound in each instruments. This system was adopted by the Library of
Congress [1] and the German Schlagwortnormdatei Decimal Classification. They
both use the Dewey classification system [4, 12]; in 1914 Hornbostel and Sachs
devised a classification system, based on the Dewey decimal classification which
essentially classified all instruments into strings, wind and percussion, and later
further into four categories:

1. Idiophones, where sound is produced by vibration of the body of the in-
strument
2. Membranophones, where sound produced by the vibration of a membrane
3. Chordophones, where sound is produced by the vibration of strings
4. Aerophones, where sound is produced by vibrating air.

However, in many cases the sound is produced by vibration of various sonorous
bodies, for example strings, solid body, i.e. body of the instrument (which can
work as resonator amplifying some frequencies) and air contained in the body.
Moreover, this classification does not strictly reflect a natural division of musi-
cal sounds into sustainable and non-sustainable, i.e. containing steady state of

the sound or not. For example, percussive instrument produce non-sustainable sounds, but also plucked string produces such a sound, too. Therefore, in our opinion, categorization of musical instrument and their sounds should take articulation into account, i.e. the way how the sound is performed by a player. Our goal was to elaborate such a categorization, based on numerical sound description (i.e. sound parameters that can be automatically calculated for a sound), which produces a clear classification of musical instrument sounds, leaving no space for doubts.

2 Instrument Classification Based on Numerical Sound Attributes

In order to perform musical instrument sound categorization based on numerical sound parameters, we decided to use a few conditional attributes (LogAttack, Harmonicity, Sustainability, SpectralCentroid and TemporalCentroid) and 2 decision attributes: instrument, and articulation. The attributes are based on MPEG-7 low-level sound description [8]. For purposes of music information retrieval, Hornbostel-Sachs is incompatible for a knowledge discovery discourse since it contains exceptions, since it follows a humanistic conventions. For example, a piano emits sound when the hammer strikes strings hence, strictly speaking, piano is a percussive instrument. Hornbostel-Sachs categorizes piano as a chordophone, i.e. string instrument. We focus on five properties of sound waves that can be calculated for any sound and can differentiate. They are: LogAttack, Harmonicity, Sustainability, SpectralCentroid and TemporalCentroid. The first two 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 [17]. The remaining three attributes are based on temporal observations of sound envelopes for singular sound of various instruments and for various playing method.
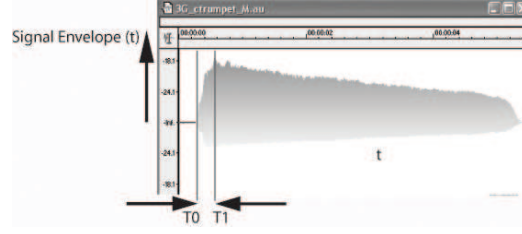
2.1 LogAttackTime (LAT)

Segments containing short LAT periods cut generic sustained and non-sustained sounds into two separate groups [16, 8]. LAT is the logarithm of the time duration between the point where the signal starts to the point it reaches its stable part.[18] Range is the $log_{10}(\frac{1}{samplingrate})$ and is determined by the length of the signal.

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

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

2.2 AudioHarmonicityType (HRM)

AudioHarmonicityType describes the degree of harmonicity of an audio signal.[8] It includes the weighted confidence measure, SeriesOfScalarType that handles

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.

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.

$$H(i) = max\ r(i, k) \tag{2}$$

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

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

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 $K=\omega*sr$, $\omega =$ 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 $c(j)$:

$$a(f_{lim}) = \sum_{f=f_{lim}}^{f_{max}} p'(f) \Bigg/ \sum_{f=f_{lim}}^{f_{max}} p(f) \tag{4}$$

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:

$$UpperLimitOfHarmonicity = log2(f_{ulim}/1000) \tag{5}$$

2.3 Sustainability (S)

We define sustainability into 5 categories based on the degree of dampening or sustainability the instrument can maintain over a maximum period of 7 seconds.
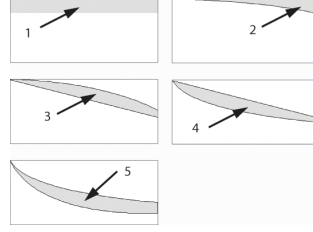
Fig. 2. Five levels of sustainability to severe dampening.

2.4 TemporalCentroid

The TemporalCentroid Descriptor also characterizes the signal envelope, representing where in time the energy of a signal is focused. This Descriptor may, for example, distinguish between a decaying piano note and a sustained organ note, when the lengths and the attacks of the two notes are identical. It is defined as the time averaged over the energy envelope:

$$TC = \sum_{n=1}^{length(SEnv)} n/sr \cdot SEnv(n)) \left/ \sum_{n=1}^{length(SEnv)} SEnv(n)) \right. \tag{6}$$

where $SEnv$ is the is the Signal Envelope and sr is the Sampling Rate

2.5 SpectralCentroid

The SpectralCentroid Descriptor measures the average frequency, weighted by amplitude, of a spectrum. In cognition applications, it is usually averaged over time:

$$c = \sum c_i \left/ i \Rightarrow ...or... \Rightarrow c_i = \sum f_i a_i \right/ \sum a_i \tag{7}$$

where c_i is the centroid for one spectral frame, and i is the number of frames for the sound. A spectral frame is some number of samples which is equal to the size of the FFT where each individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes.

3 Experiments

The data set used in our experiments contains 356 tuples, representing sounds of various musical instruments, played with various articulation. The initial data set (see http://www.mir.uncc.edu) contained parameterized sounds for every instrument available on MUMS CDs [11]. The resulting database contains numerical descriptors for 6,300 segmented sounds representing broad range of musical instruments, including orchestral ones, piano, jazz instruments, organ, etc. The MUMS CD's are widely used in musical instrument sound research [2], [19], [20], [21],[22], [23] so they can be considered as a standard. Only a part of these data were used, namely, single sounds for every instrument/articulation, or one from each octave. The risk is that we may have numerous classes with very few representations.

We decided to use decision trees available via Bratko's Orange software that implements the C4.5 with scripting in Python [13], [3]. We run orange/c4.5 with decision attribute which we called sachs-hornbostel-level-1, including 4 classes: aerophones, idiophones, chordophones, and membranophones. Also, we run the classifier for articulation. We are interested in all objects, which are in the wrong leaves, i.e. misclassified objects. Also, apart from c4.5 algorithm, we decided to use a Baysian classifier.

4 Testing

Observing the database, it is evident in Figure 3that the five descriptors by default separate sound into viable clusters. This becomes quite evident when looking at three attributes: LogAttack, Sustainability and SpectralCentroid in relation to the articulation.

Additionally, Figure ?? which is a Polyviz plot which combines RadViz and barchart techniques. It illustrates 1) the clustering of the data points in the middle of the polygon and 2) the distribution along the different dimensions. This becomes quite evident when looking at two attributes: Sustainability and TemporalCentroid in relation to the articulation.

To induce the classification rules in the form of decision trees from a set of given examples we used Quinlan's C4.5 algorithm and Oranges implementation of naive Bayesian modeling to compare the results.

4.1 c4.5

Quinlan's C4.5 algorithm [13] constructs a decision tree to form production rules from an unpruned tree. Next a decision tree interpreter classifies items which the produces the rules. This produced a 10-level tree with 36 leaves. Of the 36 leaves 17 were 100%. The remaining 19 leaves averaged 69.8% with percussive and string instruments imitating each other's attributes the most.
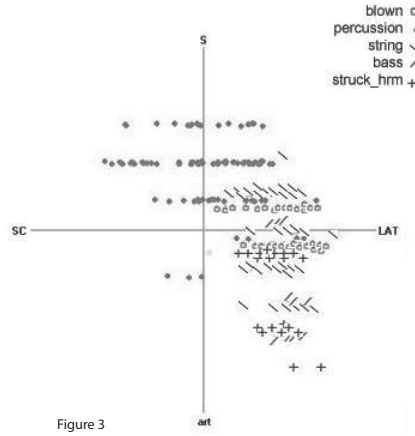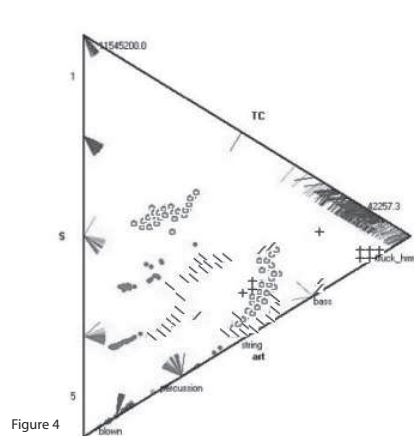
Fig. 3. Illustrations of Linear Projection and Polyviz Clustering. Figure 3 is a 4 dimensional linear projection to illustrate clustering: LogAttack, Sustainability, SpectralCentroid and Articulation, Figure 4 illustrate clustering: Sustainability, TemporalCentroid and articulation

4.2 Naive Bayes

Bratco's Naive Bayes based classifier requires a small amount of training data to estimate the parameters necessary for classification. It calculates the probability between input and predictable columns and *naively* assumes that the columns are independent and, by making this assumption, it ignores possible dependencies. This produced a 10-level tree with 61 leaves. Of the 61 leaves 44 were 100%. The remaining 17 leaves averaged 83.2% again with percussive and string instruments imitating each other's attributes the most.

4.3 Rough Sets

Using RSES we discretized and generated rules by using rough set based classifiers and the LEM2 algorithm. Here we dad results comprising 124 rules but paired it down to 33 after setting parameters of minimum of 90% confidence with support of no less than 6. Comparing this to the c4.5 17 leaves 44 at 100% with 17 leaves averaging 83.2%, it happens that both c4.5 and rough sets convey the same rules when operating at 90% confidence with support of no less than 6.

5 Summary and Conclusion

The experiments have proven that implementing temporal attributes that focus on the machine-level view of a signal, and interacting it with MPEG-7 descrip-

tors has proven to yield remarkably strong results. However, keeping this in mind, it has also revealed music information's most problematic differentiation of instruments, first the string/percussive cluster and second the blown/bass clusters.

In our future experiments we plan to run cauterization algorithms, to see how to solve the string/percussive and blown/bass clusters. In essence, the results are strong and this is pleasing but, more importantly, it has revealed an Achilles heal that MIR will no doubt focus on with new intent.

We plan to continue our experiments, using more of our MPEG-7 features and applying clustering algorithms in order to find probably better classification scheme for musical instrument sounds.

6 Acknowledgements

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. 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.
3. Demsar, J., Zupan, B. and Gregor Leban: http://www.ailab.si/orange
4. Doerr, M.: Semantic Problems of Thesaurus Mapping: *Journal of Digital Information.* Volume 1, issue 8, Article No. 52, 2001-03-26, 2001–03, (2001).
5. 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.
6. Fujinaga, I. and McMillan, K. (2000). Realtime recognition of orchestral instruments. *Proceedings of the International Computer Music Conference* (141–143).
7. Hornbostel, E. M. V., Sachs, C. (1914). Systematik der Musikinstrumente. Ein Versuch. *Zeitschrift fur Ethnologie*, Vol. 46, No. 4-5, 1914, 553-90, available at http://www.uni-bamberg.de/ppp/ethnomusikologie/HS-Systematik/HS-Systematik
8. 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)
9. Kaminskyj, I. (2000). Multi-feature Musical Instrument Classifier. *MikroPolyphonie* 6 (online journal at http://farben.latrobe.edu.au/).
10. 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.

8

11. Opolko, F. and Wapnick, J. (1987). MUMS – McGill University Master Samples. CD's.

12. 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).

13. Quinlan, J.R. 2pMU9. Bagging, boosting, and C4. 5. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Volume 725, 730, (1996).

14. SIL International (1999). LinguaLinks Library. Version 3.5. Published on CD-ROM, 1999. The Internet: http://www.silinternational.org/LinguaLinks/Anthropology/ ExpnddEthnmsclgyCtgrCltrlMtrls/MusicalInstrumentsSubcategorie.htm

15. Wieczorkowska, A. (1999). 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.

16. 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).

17. Wieczorkowska, A., Wróblewski, J., Synak, P. and Slęzak, D.: Application of temporal descriptors to musical instrument sound recognition: in *Proceedings of the International Computer Music Conference* (ICMC'00), Berlin, Germany, (2004).

18. 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).

19. 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.

20. Fujinaga, I. and McMillan, K. (2000). Realtime recognition of orchestral instruments. Proceedings of the International Computer Music Conference (141–143).

21. Kaminskyj, I. (2000). Multi-feature Musical Instrument Classifier. *MikroPolyphonie* 6 (online journal at http://farben.latrobe.edu.au/).

22. 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.

23. Wieczorkowska, A. (1999). 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.