HUBS AND ORPHANS - AN EXPLORATIVE APPROACH

Martin Gasser
Austrian Research Institute
for Artificial Intelligence (OFAI)
martin.gasser@ofai.at

Arthur Flexer
Austrian Research Institute
for Artificial Intelligence (OFAI)
arthur.flexer@ofai.at

Dominik Schnitzer
Austrian Research Institute
for Artificial Intelligence (OFAI)
dominik.schnitzer@ofai.at

ABSTRACT

In audio based music similarity, a well known effect is the existence of hubs, i.e. songs which appear similar to many other songs without showing any meaningful perceptual similarity. We show that this effect depends on the homogeneity of the samples under consideration. We compare three small sound collections (consisting of polyphonic music, environmental sounds, and samples of individual musical instruments) with regard to their hubness. We find that the collection consisting of cleanly recorded musical instruments produces the smallest hubs, wheras hubness increases with inhomogeneity of the audio signals. We also investigate how well the three data sets can be mapped into a 2D visualization space by a dimensionality reduction algorithm based on Multidimensional Scaling.

1. INTRODUCTION

One of the central goals in Music Information Retrieval is the computation of audio similarity. Proper modeling of audio similarity enables a whole range of applications: music classification/recommendation, content-based search, etc. The de facto standard approach to computation of audio similarity is timbre similarity based on parameterization of audio using Mel Frequency Cepstral Coefficients (MFCCs) plus Gaussian mixtures as statistical modeling (see Sec. 3.1). However, it is also an established fact that this approach suffers from the so-called hub problem [1]: sound samples which are, according to the audio similarity function, similar to very many other sound samples without showing any meaningful perceptual similarity to them. The hub problem of course interferes with all applications of audio similarity: hub samples keep appearing unwontedly often in recommendations, they degrade classification performance, etc.

Although the phenomenon of hubs is not yet fully understood, a number of results already exist. Aucouturier and Pachet [2] established that hubs are distributed along a scale-free distribution, i.e. non-hub samples are extremely common and large hubs are extremely rare. This is true for MFCCs modelled with different kinds of Gaussian mixtures as well as Hidden Markov Models, irrespective whether parametric Kullback-Leibler divergence or non-parametric histograms plus Euclidean distances are used for computation of similarity. But is also true that hubness is not the property of a sound sample per se since non-parametric and parametric approaches produce very different hubs. It has also been noted that audio recorded from urban soundscapes, different from polyphonic music, does not produce hubs [3] since its spectral content seems to be more homogeneous and therefore probably easier to model. Direct interference with the Gaussian models during or after learning has also been tried (e.g. homogenization of model variances) although with mixed results. Whereas some authors report an increase in hubness [2], others observed the opposite [4]. Using a Hierarchical Dirichlet Process instead of Gaussians for modeling MFCCs seems to avoid the hub problem altogether [5].

Our main interest in this paper is to explore whether the hub problem also exists in data bases of audio samples (e.g. short recordings of individual notes played on individual instruments) rather than data bases of whole songs. We are also interested in finding out how hubs influence lower dimensional projections of audio data bases. Such projections have been very popular for enabling interactive visualizations of data bases of songs [6–8] as well as samples [9, 10]. Despite the popularity of these interfaces based on lower dimensional projections, it has not yet been clarified how hubs influence and possibly impair these visualizations.

Our contribution to the understanding of the hub problem and its consequences are twofold: (i) We support the assumption that hubness is less predominant in collections of environmental sound textures [3] than in music collections, and show that it is even weaker in musical instrument databases and (ii) we show that hubness is related to the cluster structure of data by inspecting Multidimensional Scaling projections (i.e. hubs tend to lie in the centers of clusters, the more clusters can be found in a data set, the smaller are the hubs).

2. DATA

In this research, we used three sound collections of $size = 1000$. However, we assume that the results remain valid for bigger databases as well.

2.1 Music collection

We used a subset of a data base comprised of the music of an Austrian music portal. The FM4 Soundpark is an in-

Category	Number of samples
Foreign Towns & Countrysides	10
Nature Ambiences	22
Birds and Animals	73
Wind	38
Water	49
Rain	33
Planes and Trains	52
Cars	99
Traffic	49
Guns	91
Crashes&Impacts	19
Sports and Boats	85
Towns	21
General Ambiences	359

Table 1. Structure of environmental sounds collection

Instrument	Number of samples
Strings Ensemble	185
Double Bass Ensemble	135
Clarinet	191
Tuba	133
Drums	200
Flute	56
Horn	100

Table 2. Structure of musical instruments collection

ternet platform [1] of the Austrian public radio station FM4. This internet platform allows artists to present their music free of any cost in the WWW. All interested parties can download this music free of any charge. This music collection contains about 10000 songs and is organized in a rather coarse genre taxonomy. The artists themselves choose which of the $G_M = 6$ genre labels "Hip Hop, Reggae, Funk, Electronic, Pop and Rock" best describe their music. The artists are allowed to choose one or two of the genre labels. We use a data base of $size = 1000$ songs for our experiments.

2.2 Environmental sample collection

The collection of environmental sounds consists of subsets of the commercially available sound effects libraries *Sound Ideas* [2] and *Hollywood Edge* [3].

2.3 Collection of musical instruments

The musical instruments collection consists of a subset of a commercially available sample collection for professional composers. Tab. 2 gives an overview of the structure of our evaluation dataset.

3. METHODS

In this paper we use Single Gaussian (G1) distributions of Mel Frequency Cepstral Coefficients (MFCCs) to model the spectral content of individual sound samples; based on the bag-of-frames approach, those models represent the average spectral envelope and covariances between the individual coefficients, corresponding to the specific "sound" or "timbre" of a sample. By comparing the statistical models using the symmetric Kullback-Leibler divergence, we calculate a dissimilarity measure between samples.

For the visualization mapping, we arrange sound samples in 2D space, such that the distances between the 2D locations approximate the acoustic distances. To construct the 2D arrangement, we use Multidimensional Scaling [11].

3.1 Mel Frequency Cepstral Coefficients and Single Gaussians (G1)

We use the following approach to compute acoustic similarity. For a given collection of sound samples, it consists of the following steps:

1. for each sample, compute MFCCs for short overlapping frames
2. train a single Gaussian (G1) to model each of the samples
3. compute a distance matrix M_{G1} between all samples using the symmetrized Kullback-Leibler divergence between respective G1 models

The sound samples are resampled to 22050Hz and mixed down to mono audio signals. Then, we divide the raw audio data into overlapping frames of short duration and use Mel Frequency Cepstral Coefficients (MFCC) to represent the spectrum of each frame. MFCCs are a perceptually meaningful and spectrally smoothed parameterization of audio signals and a standard technique for computation of spectral similarity in music analysis (see e.g. [12]). The frame size for computation of MFCCs for our experiments was $23.2ms$ (512 samples), we used a hop size of $11.6ms$ (256 samples). We used the first $d = 20$ MFCCs for all experiments.

A single Gaussian (G1) with full covariance represents the MFCCs of each sample [13]. For two single Gaussians, $p(x) = \mathcal{N}(x; \mu_p, \Sigma_p)$ and $q(x) = \mathcal{N}(x; \mu_q, \Sigma_q)$, the closed form of the Kullback-Leibler divergence is defined as [14]:

$$KL_N(p\|q) = \frac{1}{2}\Bigg(\log\left(\frac{\det(\Sigma_p)}{\det(\Sigma_q)}\right) + Tr\left(\Sigma_p^{-1}\Sigma_q\right) + (\mu_p - \mu_q)'\,\Sigma_p^{-1}(\mu_q - \mu_p) - d\Bigg) \quad (1)$$

where $Tr(M)$ denotes the trace of the matrix M, $Tr(M) = \Sigma_{i=1..n} m_{i,i}$. The divergence is symmetrized by computing:

$$KL_{sym} = \frac{KL_N(p\|q) + KL_N(q\|p)}{2} \quad (2)$$

[1] http://www.soundpark.at
[2] http://www.sound-ideas.com/
[3] http://www.hollywoodedge.com/

3.2 Visualization algorithm

It is well known that the symmetric KL divergence between Gaussians does not satisfy the requirements to a proper distance measure [15]. However, it is possible to embed data items in a Euclidean space, such that the Euclidean distances between pairs approximate the original data distances as closely as possible. Such techniques are generally termed *Multidimensional Scaling* [16]. One possible approach to solve the Multidimensional Scaling problem are spring models [17, 18], physically inspired models of spring-connected nodes aiming at finding a node placement that minimizes the cumulative deflection from the springs' resting states. Mapping distances in feature space to spring lengths, a spring model solves the MDS problem. Implicitly, this implements a gradient-descent based solution algorithm, where a placement of the nodes that minimizes the overall stress (the deflection from the springs' resting state) is constructed. Eq. 3 gives a measure for the normalized stress in an MDS mapping.

$$S = \frac{\Sigma_{i<j}(d_{ac}(i,j) - d_{lo}(i,j))^2}{\Sigma_{i<j} d_{lo}(i,j)} \tag{3}$$

(d_{ac}: acoustic distance, d_{lo}: distance in low dimensional space, i, j: data points)

4. RESULTS

4.1 Hubs in sample databases

As a measure of the hubness of a given sample, we use the so-called n-occurrence [2], i.e. the number of times the sample occurs in the first n nearest neighbors of all the other samples in the data base. Please note that the mean n-occurrence across all samples in a data base is equal to n. Any n-occurrence significantly bigger than n therefore indicates existence of a hub. For every sample in the data bases, we computed the first n nearest neighbors. The first n nearest neighbors are the n samples with minimum Kullback-Leibler divergence (Equ. 2) to the query sample.

The results given in Tab. 3 show results calculated from the three data sets. We give the number of nearest neighbors n, the absolute number of the maximum n-occurrence $maxhub$ (i.e. the biggest hub), the percentage of samples in whose nearest neighbor lists this biggest hub can be found $maxhub\% = maxhub/size$ and the percentage of hubs $hub3\%$ (i.e. the percentage of samples of which the n-occurrence is more than three times n).

When looking at the results, it should be immediately clear that music collections are more prone to hubness than collections of environmental textures or musical instruments. This can be intuitively justified because the spectral structure of music is much less homogeneous than that of sound textures or even individual musical instruments. Thus, it is quite likely that a given musical piece is similar to a higher-than-average number of other pieces.

4.2 MDS mappings of the data sets

Fig. 1(a), 1(b), and 1(c) show 2D embeddings of the music, textures, and instruments data sets, respectively. We ran 1500 iterations of the MDS algorithm on the data. Additionally to the location of the samples, we also plotted the 50 largest hubs (samples that appear most often in the nearest neighbor lists of all other samples) and the smallest orphans (samples that rarely appear in the nearest neighbor lists of all other samples). Hubs are marked with a green dot, orphans with a red 'X'-sign.

data set	n	maxhub	maxhub%	$hub3\%$
MUSIC	50	358	35.80	5.20
TEXTURES	50	190	19.04	0.90
INST	50	172	17.18	0.29

Table 3. Hub analysis results for the three data sets

data set	stress
MUSIC	0.0327
TEXTURES	0.0434
INST	0.0556

Table 4. Stress values after 1500 MDS iterations

The music data in fig. 1(a) seems to contain only one cluster and all hubs are located within the center of this cluster. Orphans are located at the border of the point cloud.

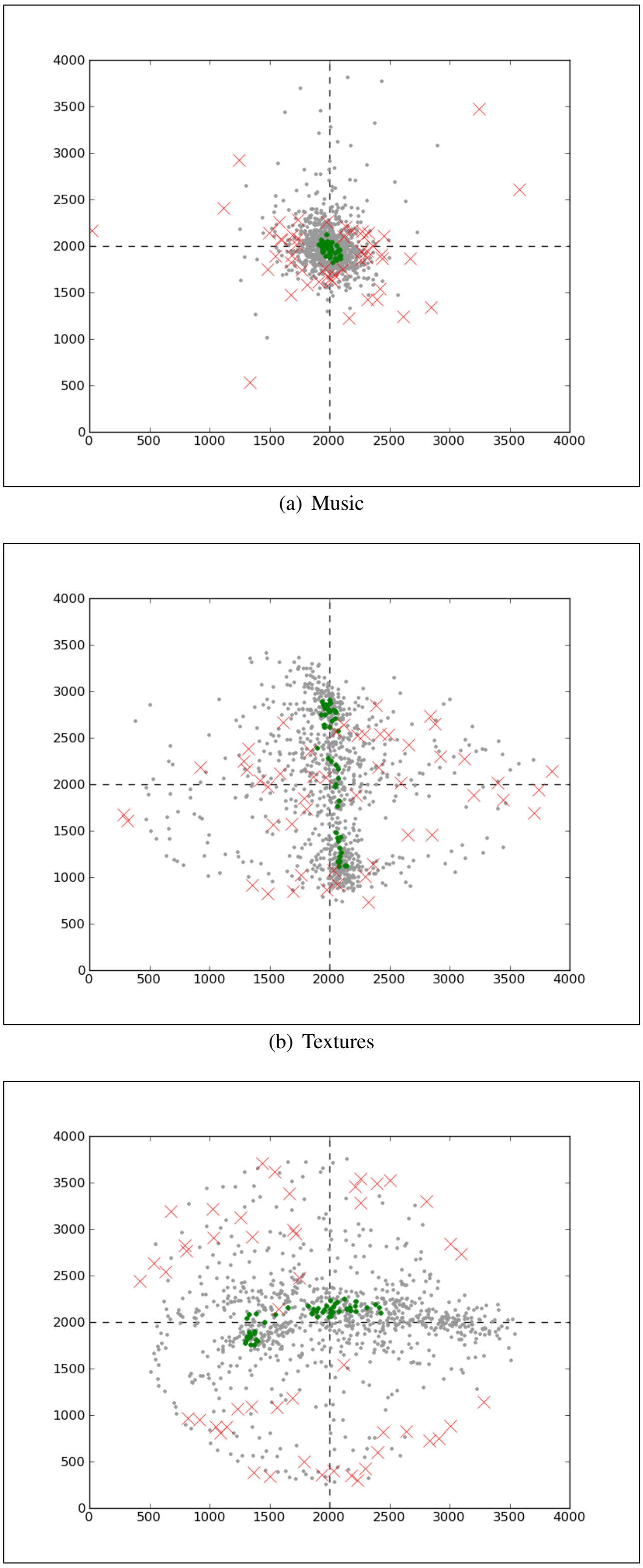
Fig. 1(b) and fig. 1(c) show a different picture. Here we can clearly see the existence of a more complex clustering in the data (e.g., sounds of engines vs. wind/sea textures, flutes vs. double bass) and the hubs are distributed more evenly over the clusters. We can also see that orphans do not strictly lie at the border of the point cloud anymore. Since MDS aims at minimizing the sum of squared differences between high- and low-dimensional distances, it is clear that in the music data set, hubs (i.e. samples with small distance to a large number of other samples) should go to the center of the point cloud, whereas orphans are pushed to the borders.

To measure the performance of MDS, we evaluate the stress formula 3 for all three data sets. Tab. 4 shows the stress values after 1500 iterations of the gradient descent-based MDS algorithm. Apparently, it is more difficult for the algorithm to cope with data sets containing multiple clusters, whereas the data set containing one cluster yields the lowest stress value, although it also contains the largest hubs.

5. CONCLUSION & FUTURE WORK

We have shown that music collections are indeed more prone to the *hubness* problem than collections of sound textures or musical instruments. We suppose that this is due to the high degree of inhomogeneity in individual musical pieces, whereas environmental sounds or individual instruments have a relatively stationary spectral structure.

We have also shown that hubs are generally located inside clusters of data, thus, the more clusters are present in the data, the lower is the degree of hubness, since hubs are distributed across clusters.



(a) Music

(b) Textures

(c) Instruments

Figure 1. Distribution of hubs and orphans in MDS projections of the data sets.

For sound textures and samples of musical instruments, the G1 similarity measure produced far fewer hubs than for the music collection; the MDS visualizations of textures and instruments also reflected the structure of the data sets visually, which could be an important argument for successful navigation inside sample libraries.

We have also developed a demonstration application that can be used to interactively explore the data sets we used by zooming/panning in the MDS embeddings and by playing back individual audio samples. It can be used to easily retrieve similar-sounding material from sample libraries without using any metadata.

Next steps in our research will include experiments with other features that also reflect time-dependent properties of the signal and the development of more advanced ways to visualize data in multiple feature spaces.

6. ACKNOWLEDGEMENTS

This research is supported by the Austrian Science Fund (FWF, P21247) and the Vienna Science and Technology Fund (WWTF, project Audiominer)

7. REFERENCES

[1] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky?,” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[2] J.-J. Aucouturier and F. Pachet, “A scale-free distribution of false positives for a large class of audio similarity measures,” *Pattern Recognition*, vol. 41, no. 1, pp. 272–284, 2007.

[3] J.-J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,” *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[4] M. Godfrey and P. Chordia, “Hubs and homogeneity: Improving content-based music modeling,” in *Proceedings of the 9th International Conference on Music Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR’08)*, (Philadelphia, USA), 2008.

[5] M. Hoffman, D. Blei, and P. Cook, “Content-based musical similarity computation using the hierarchical dirichlet process,” in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR’08)*, (Philiadelphia, USA), 2008.

[6] E. Pampalk, A. Rauber, and D. Merkl, “Content-based organization and visualization of music archives,” in *Proceedings of the 10th ACM International Conference on Multimedia*, (Juan les Pins, France), pp. 570–579, 2002.

[7] P. Knees, M. Schedl, T. Pohle, and G. Widmer, “Exploring music collections in virtual landscapes,” *IEEE MultiMedia*, vol. 14(3), pp. 46–54, 2007.

[8] M. Gasser and A. Flexer, "FM4 soundpark audio-based music recommendation in everyday use," in *Proceedings of the 6th Sound and Music Computing Computing Conference (SMC 09)*, (Porto, Portugal), 2009.

[9] E. Pampalk, P. Hlavac, and P. Herrera, "Hierarchical organization and visualization of drum sample libraries," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)*, (Naples, Italy), October 5-8 2004.

[10] S. Heise, M. Hlatky, and J. Loviscach, "Aurally and Visually Enhanced Audio Search With Soundtorch," in *CHI '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, (New York, NY, USA), pp. 3241–3246, ACM, 2009.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience Publication, 2000.

[12] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proceedings of the 1st International Conference on Music Information Retrieval*, (Plymouth, Massachusetts), 2000.

[13] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proceedings of the 6th International Conference on Music Information Retrieval*, (London, United Kingdom), 2005.

[14] W. Penny, "Kullback-leibler divergences of normal, gamma, dirichlet and wishart densities," tech. rep., Wellcome Department of Cognitive Neurology, 2001.

[15] E. Pampalk, *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Vienna University of Technology, Vienna, Austria, March 2006.

[16] M. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, 2001.

[17] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

[18] A. Morrison, G. Ross, and M. Chalmers, "Fast multidimensional scaling through sampling, springs and interpolation," *Information Visualization*, vol. 2, no. 1, pp. 68–77, 2003.