A Scale-Free Distribution of False Positives for a Large Class of Audio Similarity Measures

Jean-Julien Aucouturier [a,*],

[a] *Ikegami Lab, Grad. School of Arts and Sciences, The University of Tokyo, Japan.*

Francois Pachet [b]

[b] *SONY Computer Science Laboratory Paris, France.*

Abstract

The "bag of frames" approach (BOF) to audio pattern recognition models signals as the long-term statistical distribution of their local spectral features, a prototypical implementation of which being Gaussian Mixture Models of Mel-Frequency Cepstrum Coefficients. This approach is the most predominent paradigm to extract high-level descriptions from music signals, such as their instrument, genre or mood, and can also be used to compute direct timbre similarity between songs. However, a recent study by the authors shows that this class of algorithms when applied to music tends to create false positives which are mostly always the same songs regardless of the query. In other words, with such models, there exist songs - which we call *hubs* - which are irrelevantly close to very many songs. This paper reports on a number of experiments, using implementations on large music databases, aiming at better understanding the nature and causes of such hub songs. We introduce 2 measures of "hubness", the number of n-occurrences and the mean neighbor angle. We find that in typical music databases, hubs are distributed along a scale-free distribution: non-hub songs are extremely common, and large hubs are extremely rare - but they exist. Moreover, we establish that hubs are not a property of a given modelling strategy (i.e. static vs dynamic, parametric vs non-parametric, etc.) but rather tend to occur with any type of model, however only for data with a given amount of "heterogeneity" (to be defined). This suggests that the existence of hubs could be an important phenomenon which generalizes over the specific problem of music modelling, and indicates a general structural property of an important class of pattern recognition algorithms.

Key words: Music similarity, timbre, false positives, hubs, Mel-Frequency Cepstral Coefficients, Bag of frames

1 Introduction

The majority of systems extracting high-level music descriptions from audio signals rely on a common, implicit model of the global "sound" or "timbre" of a musical signal. This model represents timbre as the long-term accumulative distribution of frame-based spectral features. This approach has been nicknamed "bag-of-frames" (BOF), in analogy with the "bag-of-words" (BOW) treatment of text data as a global distribution of word occurrences, used in Text Classification (1). The signal is cut into short overlapping frames (typically 50ms with a 50% overlap), and for each frame, a feature vector is computed. Features usually consists of a generic, all-purpose spectral representation such as Mel Frequency cepstrum Coefficients (MFCCs) (2). The physical source of individual sound samples is not explicitely modelled: the features are fed to a statistical model, such as a Gaussian Mixture Model (GMM) (3), which models their global distributions over the total length of the extract. Global distributions can then be used to compute decision boundaries between classes (to build e.g. a genre classification system such as (4)) or directly compared to one another to yield a measure of timbre similarity (5).

1.1 Existence of hubs

The above approach has led to some success, but recent research (6) on the issue of polyphonic timbre similarity shows that BOF seems to be bounded to moderate performance. Notably, thorough exploration of the space of typical algorithms and variants (such as different signal features, static or dynamic models, parametric or non-parametric estimation, etc.) and exhaustive fine-tuning of the corresponding parameters fail to improve the precision above a empirical *glass-ceiling*, around 70% precision (although this of course should be defined precisely and depends on tasks, databases, etc.). Further, traditional means to model data dynamics, such as delta-coefficients, texture windows or Markov modelling, do not provide any improvement over the best static models for polyphonic textures of several seconds length. This is a paradoxical observation, as psychophysical experiments (7) have established the importance of dynamics in the perception of individual instrument notes.

* Corresponding author. Address: Ikegami Laboratory, Department of General Systems Studies, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan. Tel:+81-3-5454-4378. Fax : +81-3-5454-6541.

Email addresses: aucouturier@gmail.com (Jean-Julien Aucouturier), pachet@csl.sony.fr (Francois Pachet).

However, the most intriguing finding of (6) is that the class of algorithms described above tends to create false positives [1] which are mostly always the same songs regardless of the query. In other words, there exist songs, which we call *hubs*, which are irrelevantly close to all other songs.

This paper reports on a number of experiments aiming at better understanding the nature and causes of such hubs. We give a detailed description of this phenomenon, as well as a methodological basis to its study by introducing 2 measures of "hubness", the number of n-occurrences and the mean neighbor angle. We find that in typical music databases, hubs are distributed along a scale-free distribution: non-hub songs are extremely common, and large hubs are extremely rare - but they exist. Moreover, we establish that hubs are not a property of a given modelling strategy (i.e. static vs dynamic, parametric vs non-parametric, etc.) but rather tend to occur with any type of model, however only for data with a given amount of "heterogeneity" (to be defined). We find that the hubness of a given song is not an emerging global property of the distribution of its frames, but rather can be localised to certain parts of the distribution, defined by their statistical weight.

1.2 Why this may be an important problem

The phenomenon of hubs is reminiscent of other isolated reports in other domains than music. Biometric verification systems, such as fingerprints, but also speech and speaker recognition systems typically exhibit striking performance inhomogeneities among users within a population. The statistical significance of such critical classes of users, in the context of Speaker Verification, was formally shown in (8), by analysing population statistics based on the test data used for the NIST 1998 speaker recognition evaluation. The paper established a speaker taxonomy in terms of animal names, notably *goats* (users that are very difficult to recognize), *lambs* (users that are particularly easy to impersonate) and *wolves* (users who are particularly successful at imitating other speakers).

A complete analogy with this taxonomy would call a wolf a song which is constantly closer to a random song S than S is to itself. However, the Speaker Recognition *menagerie* is essentially pointing out the same phenomenon as the hubs observed with our timbre similarity measure: that high false positive rates are not uniformly distributed in the database, but manifests only in a small critical population.

The reason for the appearance of such classes is generally thought to be an

[1] we describe an evaluation framework to practically decide such false positives in section 2.1

intrinsic property of human users. However, a recent study (9) in the context of fingerprint recognition suggests that these properties of wolfiness, goatness, etc.. are rather properties of the algorithms themselves. The observation that we make here of the existence of "wolf songs", in the different context of music pieces, seem to corroborate this hypothesis. This is especially interesting as the techniques used for timbre similarity (namely variations on the GMMs of MFCCs) are typically similar to the ones employed in Speaker/fingerprint recognition systems. We will show in the remaining of this paper that hubs occur for many different algorithms but that the hubness of a given song is algorithmic-dependent.

2 Definition and Measures

This section gives a detailed description of the phenomenon of hubs, as well as the algorithms for which these were observed. Notably, we describe 2 metrics we designed to quantify the "hubness" of a song, which will be used in the experiments in the remaining of the paper.

2.1 Algorithms, databases and groundtruth

We sum up here the timbre similarity algorithm presented in (6). The signal is first cut into frames. For each frame, we estimate the spectral envelope by computing a set of Mel Frequency Cepstrum Coefficients (MFCCs). We then model the distribution of the MFCCs over all frames using a Gaussian Mixture Model (GMM). A GMM estimates a probability density as the weighted sum of $\mathcal{M}$ simpler Gaussian densities, called components or states of the mixture:

$$p(x_t) = \sum_{m=1}^{m=\mathcal{M}} \pi_m \mathcal{N}(x_t, \mu_m, \Sigma_m) \tag{1}$$

where x_t is the feature vector observed at time t, $\mathcal{N}$ is a Gaussian pdf with mean μ_m, covariance matrix Σ_m, and π_m is a mixture coefficient (also called component prior probability). The parameters of the GMM are learned with the classic E-M algorithm ((3)).

We then compare the GMM models to match different signals, which gives a similarity measure based on the audio content of the items being compared. We use a Monte Carlo approximation of the Kullback-Leibler (KL) distance between each duple of models A and B. The KL-distance between 2 GMM

probability distributions p_A and p_B (as defined in Equ.1) is defined by :

$$d(A,B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx \tag{2}$$

The KL distance can thus be approximated by the empirical mean :

$$d(\widetilde{A,B}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_B(x_i)}{p_A(x_i)} \tag{3}$$

(where n is the number of samples x_i drawn according to p_A) by virtue of the central limit theorem.

In this study, we will consider several variations on the above algorithm, inspired by the study in (6). These variations were chosen to be representative of several typical modelling strategies in pattern recognition (as classified e.g. in (3)), namely:

- Static parametric model: 20 MFCCs (incl. 0^{th} coefficient), 50-component GMM, compared with $n = 2000$ Monte-Carlo draws.
- Static non-parametric model: 20 MFCCs (incl. 0^{th} coefficient), vector-quantized to 200 codebook vectors using LVQ (10), modelled by histograms compared by euclidean distance.
- Static parametric modelling of first-order dynamics: 20 MFCCs (incl. 0^{th} coefficient), appended with 20 delta coefficients (11), 50-component GMM, compared by Monte Carlo.
- Static parametric modelling of second-order dynamics: 20 MFCCs (incl. 0^{th} coefficient), appended with 20 first-order delta coefficients and 20 second-order acceleration coefficients, 50-component GMM, compared by Monte Carlo.
- Dynamic modelling with parametric model: 20 MFCCs (incl. 0^{th} coefficient), modelled with 12-state HMM (12), using 4 Gaussian components per state, compared by Monte Carlo.

In all of the above, the specific algorithm settings such as number of GMM components correspond to optimally-performing values found in previous research (6).

This study uses two music databases:

- a large set of 15,460 popular music titles, assembled for the purpose of the Cuidado European IST project (13) (referred to as the "Cuidado database").
- a subset of this database, containing 350 titles, used for the evaluation study in (6). It is organized in 37 clusters of songs by the same artist, encompassing very different genres and instrumentations (from *Beethoven* piano sonata to

The Clash punk rock and *Musette*-style accordion). In the following, we refer to this database as the "test database".

When relevant, we will measure the precision of the above algorithms on the test database by computing their R-precision. It measures the ratio of the number of relevant documents to the number of retrieved documents, when all relevant document have been retrieved (i.e. precision at recall = 1). The set of relevant documents for a given music title is the set of all titles of the same artist cluster than the seed. This is identical to the methodology used in (6). In this framework, we call a "false positive" to a seed song S a song T which is found in the nearest neighbors of S, but not in the cluster of S.

In the following, we will argue that hub songs are close to many songs (according to the algorithmic measure) to which they have "no perceptual similarity". This is judged on the basis of the groundtruth described above, and not on any psychological evaluation using actual human similarity ratings. Note however that recent research (14) has found that typical human ratings are indeed in accordance with groundtruths that are similar to the one used here. Moreover, the fact that certain hubs are found close algorithmic matches to more than a fourth of the very heterogeneous Cuidado database (as seen in Section 3) strongly indicates that similarity measures involving hubs have little perceptual grounding if any.

2.2 Definition of a hub

In this paper, we call *hub* a song which occurs frequently as a false positive according to a given similarity measure. This both implies that

(1) a hub appears in the nearest neighbors of many songs in the database
(2) most of these occurrences do not correspond to any meaningful perceptual similarity.

Each condition in itself is not sufficient to characterize a hub:

(1) A given song may occur very many times in the nearest neighbors of other songs, but this may not be a *false* positive (as defined by the evaluation procedure described above). Depending on the composition of a given databases, some songs may well approximate the perceptual center-of-mass of the database. For instance, it may be found that *A Hard Day's Night* by *The Beatles* is a song that bears close timbre similarity to most of 60's pop music, and therefore could be found to occur very frequently as a nearest neighbor to many songs in a database composed by a majority of Rock and Pop songs. However, in a classical music database, the same song would not be such a common neighbor.

(2) A given song may be a false positive for a given seed song, i.e. be in the first nearest neighbors of the seed without any actual perceptual similarity. However, different songs may have different false positives. For instance, a given *Beethoven* piano sonata may be mismatched to an acoustic guitar piece, but not necessarily mismatched to other songs. A hub is a piece than is irrelevantly close to very many songs, i.e. a bug which is not local to only a few queries.

2.3 Measures of hubness

We propose here 2 measures to quantify the "hubness" of a given song.

2.3.1 Number of occurrences

A natural measure of the hubness of a given song is the number of times the song occurs in the first n nearest neighbors of all the other songs in the database. As discussed in appendix A, the measure of the number of n-occurrences N_n of a song has the property that the sum of the values for all songs is constant given a database.

Table 1 shows the ten songs in the test database having the largest number of occurrences in the first 10 nearest neighbors over all queries (N_{10}). This illustrates the predominance of a few songs that occur very frequently. For instance, the first song, `MITCHELL, Joni - Don Juan's Reckless Daughter` is very close to 1 song out of 6 in the database (57 out of 350), which is more than 6 times more than the theoretical mean value (10). Among these occurrences, many are likely to be false positives.

2.3.2 Neighbor angle

An operational definition of a hub is that it is a song H which is found to be "close" (though not perceptually) to duplets of songs A and B which themselves are (perceptually) distant from one another. Therefore, the hubness of song H can be estimated by comparing its distances to its neighbors $d(H, A)$ and $d(H, B)$ on the one hand, and the distance between the neighbors $d(A, B)$ on the other hand. Equivalently, one can measure the angle θ_H formed by the segments $[H, A]$ and $[H, B]$

$$\cos\theta_H = \frac{d(A,B)^2 - d(H,A)^2 - d(H,B)^2}{2d(H,A)d(H,B)} \tag{4}$$

Table 1
10 Most Frequent False Positives

Song	N_{10}
MITCHELL, Joni - Don Juan's Reckless Daughter	57
MOORE, Gary - Separate Ways	35
RASTA BIGOUD - Tchatche est bonne	30
BRIDGEWATER, DD - What Is This Thing Called Love	30
PUBLIC ENEMY - Cold Lampin With Flavor	27
MOORE, Gary - Cold Day In Hell	27
MARDI GRAS BIG BAND - Funkin'Up Your Mardi Gras	25
GILBERTO, João - Tin tin por tin tin	25
MITCHELL, Joni - Talk To Me	22
CABREL, Francis - La cabane du pêcheur	22

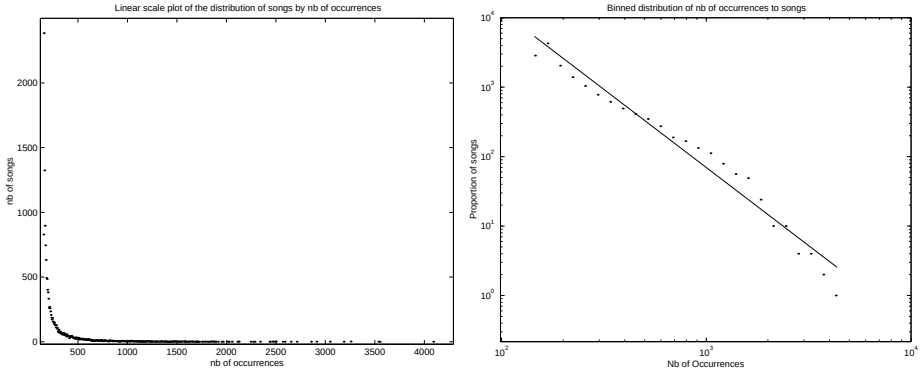

Fig. 1. Distribution of the songs according to their number of 100-occurrences in the Cuidado 15,460-song database, with a GMM-based distance. Left: using a linear scale. Right: using a log-log scale. In this second scale, the distribution is approximately linear, which indicates a power-law.

This is computed for a given song H by drawing a large number of successive duplets of neighbors (A, B) (such that $A \neq B \neq H$), and computing the mean value of θ_H. We use 1000 successive random draws.

An important property of the neighbor-angle value is that, like the number of n-occurrences N_n of a song, the sum of the values for all songs is constant given a database size. (see Appendix A for more details).

3 Hubs form a scale-free distribution

Figure 1-left shows the distribution of songs in the Cuidado database (15,460 titles) according to their number of 100-occurrences for the optimal GMM-based distance. One can observe that while most songs only have around a few hundred occurrences (more than 6,000 songs have between 150 and 160 occurrences), a few songs get upward of 2000 occurrences. This latter songs can reasonably be described as hubs. Moreover, hubness appears to be a continuous variable (with a continuum of intermediate values), rather than a discrete boolean property.

Table 2 shows the 5 biggest hubs in the Cuidado database ranked by their number of 100-occurrences for the baseline GMM-distance. The first song, from French alternative rock band *Noir Désir*, is a close neighbor to more than a fourth of the database. The fifth biggest hub, a folk song by *Joni Mitchell* was the biggest hub of the much smaller test database, as seen previously in Table 1.

Table 2
5 Most Frequent False Positives in the Cuidado database

Song	N_{100}
NOIR DESIR - En Route Pour la Joie	4090
VANNELI, GINO - Stay With Me	3552
OSWALD, JOHN - Explo	3533
ABC - When Smokey Sings	3256
MITCHELL, Joni - Don Juan's Reckless Daughter	3255

Figure 1-right shows the same plot than Figure 1-left, but on a log-log scale the same distribution shows itself to be linear. This is the characteristic signature of a power-law distribution $P[X = x] = x^{-\gamma}$. The nearly linear relationship extends over 4 decades ($[1 - 10^4]$) songs, which is why such distributions have been called "scale-free", or lacking a "characteristic length scale".

Many man-made and naturally occurring phenomena, including city sizes, word frequencies, number of links to a web page, are distributed according to a power-law distribution (15; 16). Similarly, scale-free distributions have been observed in musical data, notably in networks of artists that co-occur in playlists from specialized websites (17).

For all these reasons, the scale-free distribution of networks of timbrally similar songs is a remarkable, but not utterly surprising phenomenon. If all timbre distances were perceptually relevant ("no bugs"), then it would an acceptable conclusion that some songs be more "prototypical" than others, thus translat-

ing the distribution of musical and social influences and communities inherent to possibly every human activity. However, as already noted, what we observe here is a distribution of algorithmic bugs rather than the self-organization of an ideal music space: The most connected songs (extreme hub songs that are close matches to more than a fourth of a given database) typically appear as the nearest neighbors of songs to which they do not bear any perceptual similarity. It is yet unclear whether the scale-free distribution that we observe here is

- the result of a scale-free organisation of an ideal perceptual distance measure, which is being polluted by measurement errors
- the result of a non remarkable ideal distribution, polluted by a scale-free distribution of false-positives
- or both

The influence of measurement errors on scale-free distributions could be studied e.g. in the light of recent results on the robustness of experimental topological analysis of protein interaction networks (18).

4 Hubs are a consequence of the agglomerative modelling of the features, not of the features themselves

4.1 Hypothesis

In this section, we investigate whether hubs are a consequence of poor featural representation of the frames of audio data. We test the hypothesis that hubs exist on full songs, because hubs also exist on individual MFCC frames, i.e. that there are specific segments of audio data which are close non-perceptive matches to every other possible frames.

4.2 Experiment

We build a database of individual 2048-point hamming-windowed frames of audio data, obtained from the uniform segmentation of a few different songs. The database is made to contain 15,460 frames, so results can be quantitatively compared to the full-song behaviour in the Cuidado database. Each frame is modelled by 20 MFCCs (incl. $0^t h$ order coefficient), which is the feature space used in the best performing full-song measure. A distance measure is implemented using euclidean distance, each dimension being normalized to be between 0 and 1, using the 5% and 95% percentile values. This distance mea-

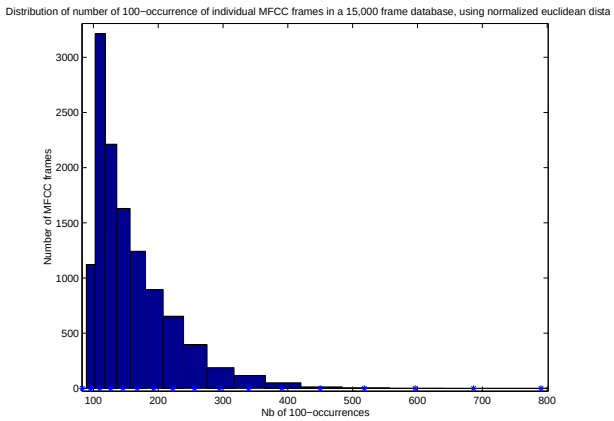

Fig. 2. Distribution of the MFCC frames according to their number of 100-occurrences in a 15,460-frame database, based on normalized euclidean distance.

sure was chosen to yield a behaviour similar to MFCCs comparison in GMM probability estimation (euclidean comparison with mean vector, rescaled by variance coefficients in each dimension). We compute the 100 nearest neighbors of each frame in the database, store them, and compute the number of 100-occurrence of each frame in the database.

4.3 Results

Figure 2 shows the distribution of the MFCC frames according to their number of 100-occurrences. The distribution is exponentially decreasing, with a maximum N_{100} value around 500. Such small numbers do not indicate the presence of hubs, which is confirmed by manual inspection of the neighbors of the most re-occurring frames. These frames typically correspond to sounds that are common to many different songs, such as noise or silence, and thus have more neighbors than more specific frames (harmonic sounds) that tend to be close to frames of the same song only. The maximum N_{100} value of 500 is more than 8 times smaller than the maximum value obtained for full songs in the Cuidado database. This indicates that the hub phenomenon is not a direct consequence of poor featural representation, but rather an effect of the modelling of the agglomeration of the very many frames in full songs.

5 Hubs appear for all algorithms

5.1 Hypothesis

In this section, we investigate whether hubs are a consequence of a specific algorithmic strategy for modelling the agglomeration of frames in full songs. We test the hypothesis that hubs appear only (or in majority) for a given algorithm.

5.2 Experiment

We compare several measures of hubness on our test database for the 5 algorithms described in section 2.1, chosen to be representative of the principal modelling strategies (GMM, Delta, Acceleration, HMM, Histograms).

5.3 Results

Figure 3 shows the distribution of the number of 100-occurrences of songs in the test database, for the 5 algorithmic variants. Since the number of occurrences is a constant-sum measure (see Appendix A), all 5 distributions are centered on the same mean value of 100. However, it appears that the choice of the algorithm has an influence on the shape of the distribution of occurrences. While all algorithms produce extreme hubs having high number of occurrences (e.g. $N_{100} > 300$), hubs tend to be smaller for the GMM-based distance than for both the dynamic-based and the histogram-based ones. Due to the constant-sum effect, algorithms that produce more high-occurrence songs also produce more low-occurrence songs. This results in a skewed distribution (where very many low-occurrence songs compensate a few high-occurrence songs) in the case of the dynamic-based distances, and a bi-modal distribution for the histogram-based distance, for which very few songs actually take the mean occurrence value.

This behaviour is confirmed by Table 3, which shows the number of songs in the test database that exhibit high values for both number of 100-occurrences and number of 20-occurrences. The 5 similarity measures exhibit different proportions of hubs: GMM-based distances produce the fewest, while Histogram-based distance produce 5 times as many. The proportion of hubs produced by each algorithm is in agreement with the precision reported in previous research (6): GMM-based distances perform better than (or equivalent to) dynamics, which perform better than histograms.

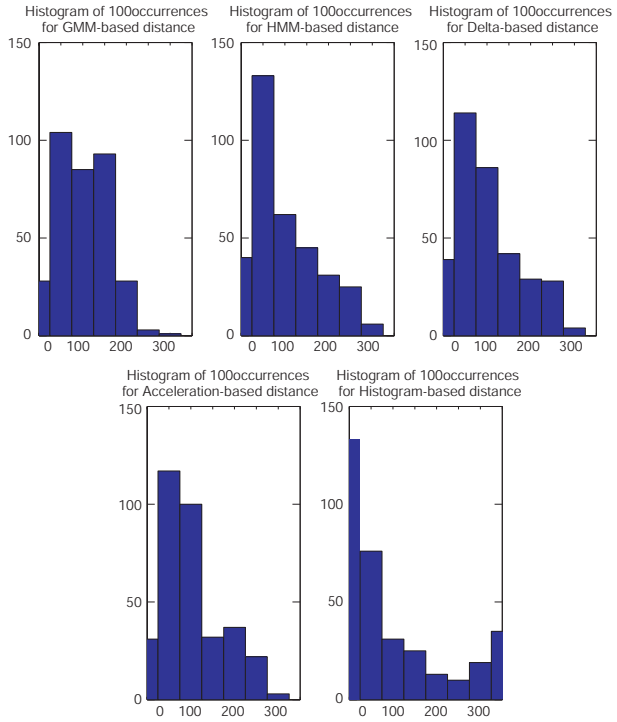

Fig. 3. Distribution of the Number of 100-occurrences of songs in the test database for several distance algorithms.

Nevertheless, it is difficult to conclude that hubs are a specific property of a given algorithmic strategy to model the MFCC frames. All algorithms create hubs. Moreover, static modelling create more hubs than dynamics in the case of Histograms and HMMs, but not in the case of GMM and HMMs. If anything, it seems that non-parametric (Histograms) create more hubs that parametrics approaches (GMMs, HMMs). This notably rules out possible convergence problems of parametric estimation (local minima) as a source of bugs.

Table 3
Comparison of number of songs exhibiting high number of occurrences in the test database, for several distance algorithms

Measure	GMM	HMM	Delta	Acceleration	Histogram
$N_{100} > 200$	16	48	49	45	69
$N_{20} > 40$	34	41	39	39	42

6 Hubness is not intrinsic to songs

6.1 Hypothesis

In this section, we investigate whether hubs are an intrinsic property of given songs, which will act as hubs independantly of the algorithm used to model them. We test the hypothesis that hub songs are strongly correlated between different algorithmic measures.

6.2 Experiment

We compute the correlation between hubness measures for songs modelled with the same five algorithms as above, using the test database.

6.3 Results

Table 4 reports the correlation of the hubness of all songs between various algorithmic models, using 2 measures of hubness (number of 100-occurrences and the neighbor angle).

Table 4
Correlation of the hubness of all songs between various algorithmic models. The hubness of songs is measured both by the number of 100-occurrences and the neighbor angle (the latter in parenthesis).

	GMM	HMM	Delta	Acceleration	Histogram
GMM	1.0	0.78 (0.67)	0.79 (0.69)	0.79 (0.71)	0.42 (0.17)
HMM	-	1.0	0.95 (0.96)	0.90 (0.96)	0.47 (0.17)
Delta	-	-	1.0	0.97 (0.99)	0.46 (0.15)
Acceleration	-	-	-	1.0	0.43 (0.14)
Histogram	-	-	-	-	1.0

Both measures reveal the same structure:

- Hubs appearing with GMMs are moderately correlated to HMMs, Delta and Acceleration.
- Hubs appearing with HMMs, Delta and Acceleration are very strongly correlated.
- Hubs appearing with Histograms are strongly decorrelated to those appearing with the other algorithms.

In more details, Tables 5 and 6 compare the most frequent hubs for 2 GMM and Histogram-based distances, here measured with their number of 20-occurrences. It appears that some songs act as hubs for both measures, e.g. `MITCHELL, Joni - Dom Juan's Reckless Daughter`. However, a vast majority of the hubs are different. Notably, certain songs are important hubs for one measure and perfectly standard songs for the other. For instance, `SUGAR RAY - Fly` is a hub for the GMM-based distance, but not for the one based on Histograms. Similarly, `CABREL, Francis - Samedi soir sur la Terre` is only a hub for the histogram distance.

Table 5
Most Frequent False Positives for parametric approach with GMMs

Hubs with {MFCC,GMM}	N_{20} ($card(C_S)$)
MITCHELL, Joni - Don Juan's Reckless Daughter	98(9)
BRIDGEWATER, DD - What A Little Moonlight Can do	79(12)
RASTA BIGOUD - Tchatche est bonne	79(7)
MOORE, Gary - Separate Ways	77(9)
SUGAR RAY - Fly	75(13)
...	
CABREL, Francis - Samedi soir sur la Terre	29 (7)

Table 6
Most Frequent False Positives for non parametric approach with Histograms.

Hubs with {VQ,CM}	$N_{20}(card(C_S))$
VOCAL SAMPLING - Radio Reloj	153 (13)
MOORE, Gary - The Hurt inside	126 (9)
CABREL, Francis - Samedi soir su la Terre	122 (7)
CABREL, Francis - Corrida	105
MITCHELL, Joni - Dom Juan's Reckless Daughter	95 (9)
...	
SUGAR RAY - Fly	23(13)

Therefore, we can conclude that:

- The hubness of a given song is not an intrinsic property of the song, but rather a property of a given algorithm.
- Dynamics, both via static modelling of dynamical features (delta, acceleration) or via dynamic modelling (HMMs) seems to have an influence of the songs that act as hubs. All three algorithms tend to create the same hubs.
- Parametric modelling tend to create very distinct hubs from non-parametric

modelling, so the dynamical/static aspect is not the only involved factor in the appearance of hubs

7 Hubs do not appear for any dataset

7.1 Hypothesis

Section 6 establishes that hubness is not an intrinsic property of a given song, but rather is dependent on the modelling algorithm. In this section, we investigate whether hubs are a structural property of pattern recognition-based similarity measures, and that they can be observed in any dataset. This is a relevant question knowing as remarked earlier that hubs have been observerved in this study on timbre similarity, but also in the domain of Speaker and Fingerprint identification.

7.2 Experiment

We apply the same modelling technique (GMMs of MFCCs) to compute the perceptual similarity of another class of audio signals, namely ecological sound textures. We gathered a database of 106 3-minute urban sound ambiances, recorded in Paris using a omni-directional microphone [2]. The recordings are clustered in 4 "general classes" (Boulevard, Neighborhood Street, Street Market, Park) and 11 "detailed classes", which correspond to the place and date of recording of a given environment. For instance, "Parc Montsouris, Paris 14e" is a subclass of the general "Park" class.

Each audio recording is modelled with 50-ms frames, 20-MFCCs and 50-component GMMs. Models are compared to one another with Monte-Carlo distance using 2000 samples.

7.3 Results

Figure 4 shows the histogram of the number of 20-occurrences obtained with the above distance on the database of ecological sound ambiances, compared with the same measure on the test database of polyphonic music. It appears that the distribution of number of occurrences for ambiance sounds is more

[2] This material was collected and kindly made available by Boris Defreville from LASA.

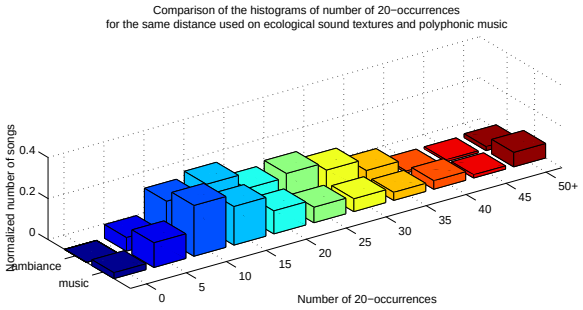

Fig. 4. Comparison of the histograms of number of 20-occurrences for the same distance used on ecological sound ambiances and polyphonic music.

narrow around the mean value of 20, and has a smaller tail than the distribution for polyphonic music. Notably, there are four times as many audio items with more than 40 20-occurrences in the music dataset than in the ambiance dataset. This is also confirmed by the manual examination of the similarity results for the ecological ambiances: none of the (few) false positives re-occur significantly more than random. As we discuss elsewhere (19), this is also revealed when analyzing the precision of the measures, which is significantly better for soundscapes than music.

This indicates that hubs are not an intrinsic property of the class of algorithm used here, but rather appear only for a certain classes of signals, among whom polyphonic music, but not ecological sound ambiances. As we will see now, the two classes of signals can notably be distinguished in terms of their homogeneity.

8 Hubness can be localized to certain frames

8.1 Hypothesis

This section investigates whether the hubness of a given song is a emerging global property of the distribution of its frames, or rather can be localised e.g. to certain frames that are less discriminant then others.

We have already shown that MFCC frames intrinsically don't exhibit hub behaviours, i.e. one cannot find a specific frame of audio which is close to any other frame, in an euclidean framework. However, this doesn't make any statement about the discriminative power of MFCC frames: it is well possible that most MFCC frames be globally close to one another, which has notably been observed in the domain of speech sounds in (20). It is therefore possible to imagine that a large part of the distribution of MFCCs is composed of non-discriminative frames, and that what is perceptually salient for a human

listener may not be statistically predominant when comparing models of the frame distribution.

8.2 Experiment

We describe here an experiment to assess whether there exists such a small portion of the frame distribution which is responsible in majority for the discrimination between non-perceptually close songs. We propose to explore the distribution of MFCC frames by ranking them by statistical importance. We define a statistical homogeneity transform $h_k : \mathcal{G} \mapsto \mathcal{G}$ on the space $\mathcal{G}$ of all GMMs, where $k \in [0, 1]$ is a percentage value, as:

g_2 = $h_k(g_1)$
  $(c_1, \ldots, c_n) \leftarrow$ sort(components(g_1), decreasing w_c)
  define $\mathcal{S}(i) = \sum_{j=1}^{i} weight(c_j)$
  $i_k \leftarrow \arg\min_{i \in [1,n]} \{\mathcal{S}(i) \geq k\}$
  $g_2 \leftarrow$ newGMM(i_k)
  define d_i =component(g_2,i)
  $d_i \leftarrow c_i$, $\forall i \in [1, i_k]$
  weight(d_i) $\leftarrow$ weight(c_i)/$\mathcal{S}(i_k), \forall i \in [1, i_k]$
  return g_2
end h_k

From a GMM g trained on the total amount of frames of a given song, the transform h_k derives an homogenized version of g which only contains its top k% components. Frames are all the more so likely to be generated by a given gaussian component c than the weight w_c of the component is high (w_c is also called prior probability of the component). Therefore, the homogenized GMM accounts for only a subset of the original song's frames: those that amount to the k% most important statistical weight. For instance, $h_{99\%}(g)$ creates a GMM which doesn't account for the 1% least representative frames in the original song.

We apply 11 transforms h_k for $k \in [20, 40, 60, 80, 90, 92, 94, 96, 98, 99, 100]$ to the GMMs corresponding to the optimal measure described above. Each transform is applied to 2 datasets, the test database containing polyphonic music and the database of urban soundscapes used in Section 7. This yields 11 similarity measures per dataset, the properties of which we study below.

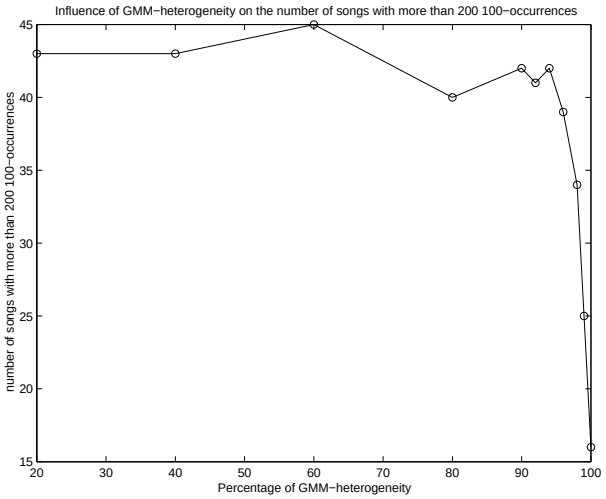

Fig. 5. Influence of the percentage of statistical homogenization on the number of songs with more than 200 100-occurrences

8.3 Influence on hubs

Figure 5 and 6 show the influence of the homogenization transform on the number of hubs in the database of polyphonic music. The database of soundscapes is not used in this comparison, as we found in Section 7 that soundscapes did not engender hubs. Hubness is measured in the case of Figure 5 by the number of songs in the test database having a number of 100-occurrences greater than 200, and in the case of Figure 6, by the number of songs with a mean neighbor angle greater than 65°.

Both metrics indicate that GMM homogenization critically increases the number of big hubs in the music database: homogenization with $k = 30\%$ creates more than twice as many hubs with more than 200 occurrences, and more than 5 times as many hubs with angles greater than 65°. It seems reasonable to interpret the increase of hubness when k decreases as a consequence of reducing the amount of discriminative information in the GMMs (i.e. from representing a given song, down to a more global style of music, down to the even simpler fact that it *is* music).

However, the increase in hubness is not monotonic. Both figures clearly show a very important increase in the number of hubs in the first few percent of homogenization. The extreme number of hubs obtained with $k = 30\%$ is reached as early as $k = 92\%$ in the case of the occurrence metric and $k = 96\%$ in the case of the mean angle metric. This is a strong observation: the hubness (or rather non-hubness) of a song seems to be controlled by an extremelly small amount of critical frames, which represent typically less than 5% of whole distribution. Moreover, these frames are the least statistically significant ones, i.e. are modelled by the least important gaussian components

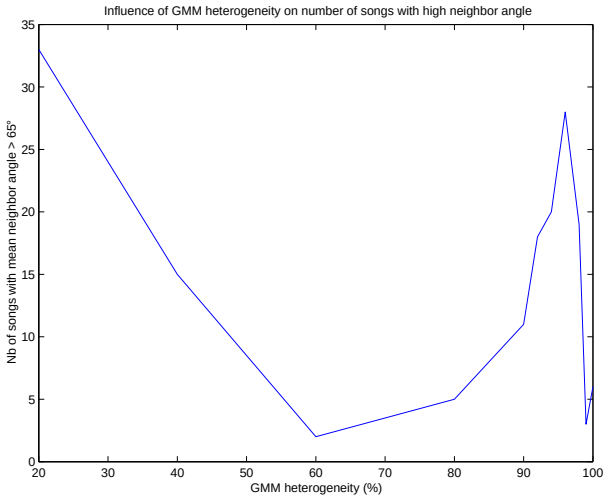

Fig. 6. Influence of the percentage of the homogenization on the number of songs with a mean neighbor angle greater than 65°

in the GMMs. This indicates that the majority (more than 90%) of the MFCC frames of a given song are a poor representation of what discriminates this song from other songs.

Moreover, Figure 6 shows that after the extremely rapid peak of hubs when removing the first 5% frames, the number of hub songs tend to decrease when k decreases from 90% to 60%, and then increases again for k smaller than 60%. The minimum value reached at $k = 60\%$ is equivalent to the original value at $k = 100\%$. A similar decreasing behaviour is observable to a smaller extent with the other metric in Figure 5 (with a local minimum at $k = 80\%$), although it is difficult to establish that this is a statistically significant trend.

The behaviour in Figure 6 suggests that there is a population of frames in the range $[60\%, 95\%]$ which is mainly responsible for the hub behaviour. While the hubness of songs diminishes as more frames are included when k increases from 20% to 60% (such frames are increasingly specific to the song being modelled), it suddenly increases when k gets higher than 60%, i.e. this new 30% information is detrimental for the modelling and tend to diminish the discrimination between songs. The continuous degradation from 60% to 95% is only eventually compensated by the inclusion of the final 5% critical frames.

8.4 Influence on precision

Figure 7 shows the influence of homogenization on the precision of the resulting similarity measure, for both datasets. The precision for urban soundscapes is measured with the 10-precision using the detailed classes as ground truth, and with the R-precision for polyphonic music. For both dataset the precision is

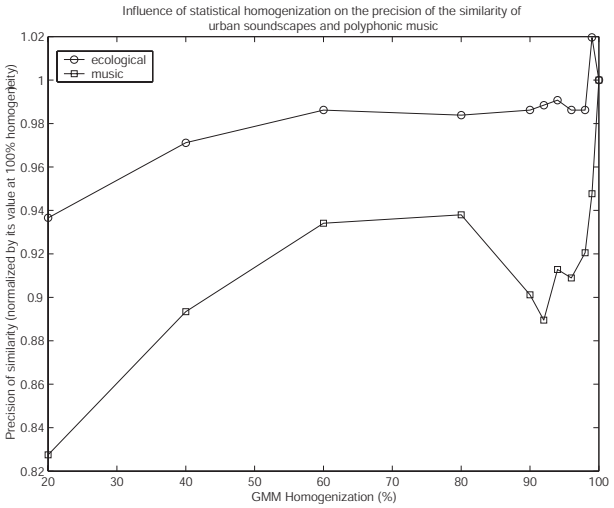

Fig. 7. Comparison of the influence of the homogeneity transform on the precision of the similarity measure for soundscape and music signals.

measured by reference to the baseline precision corresponding to $k = 100\%$, which is different for environmental and music, as we discuss in (19).

For polyphonic music, the figure closely mimics the (inverse) behaviour seen in Figure 6, with precision plummeting when k decreases from 100% to 92%, and then reaching a local maximum again between 60% and 80%. This gives further support to the observation that not all frames are equally discriminative, and that there exists a population of frames in the range $[60\%, 95\%]$ which is detrimental to the modelling of perceptual similarity.

We notice a very different behaviour in the case of urban soundscape signals. It appears that 99% homogenization is slightly beneficial to the precision. This suggests that the 1% less significant frames are spurious frames which are worth smoothing out. Further homogenization down to 60% has a moderate impact on the precision, which is reduced by about 1% (absolute). This suggests that the frame distribution is very homogeneous, and doesn't exhibit critical populations of frames which are either extremely discriminative (such as the $[95\%, 100\%]$ region for polyphonic music), or non-discriminative (such as the $[60\%, 95\%]$ region for polyphonic music). Ecological ambiances can be discriminated nearly optimally by considering only the most significant 50% of the frames.

The greater heterogeneity of polyphonic music data for pattern recognition purposes may explain the appearance of hubs, and their non-existence for other, more homogeneous classes of signals. It would be worth investigating the feature-homogeneity of other hub-prone classes of signals, such as speaker data or fingerprints, to give further support to this hypothesis.

Note that a possible further experiment to validate the existence of a critical

region in $[60\%, 95\%]$ would be to study the properties of models built using only these frames on the one hand, and models using all frames except these ones on the other hand. Models of the latter type would be expected to not generate as many hubs as the standard type. This will form the basis of future work.

9 Conclusion

This study shows that the class of algorithms predominently used to extract high-level music descriptions from music signals tends to create false positives which are mostly always the same songs regardless of the query. In other words, there exist songs, which we call *hubs*, which are irrelevantly close to all other songs.

We studied the nature and properties of such hub songs in a series of experiments, and established that:

- hubs are distributed according to a scale-free distribution.
- hubs are not a consequence of poor feature representation of each individual frame, but rather an effect of the modelling of the agglomeration of the many frames of a sound texture.
- hubs are not a property of a given modelling strategy (i.e. static vs dynamic, parametric vs non-parametric, etc.) but rather tend to occur with any type of model.
- hubs are not an intrinsic property of certain songs, but that different algorithms distribute the hubs differently on the whole database.
- hubs are not a property of the class of algorithms studied here which appears regardless of the data being modelled, but only for data with a given amount of heterogeneity, e.g. for polyphonic music, but not for ecological sound ambiances.
- the hubness of a given song is not an emerging global property of the distribution of its frames, but rather can be localised to certain parts of the distribution, notably a population of non-discriminative frames corresponding to the $[60\%, 95\%]$ region of statistical weight.

This phenomenon of hubs is reminiscent of other isolated reports in different domains, such as Speaker Recognition or Fingerprint Identification, which intriguingly also typically rely on the same features and pattern-recognition algorithms. This suggests that this could be an important phenomenon which generalizes over the specific problem of timbre similarity, and indicates a general structural property of the class of algorithms examined here, for a class of signals which is probably defined by the heterogeneity of their feature distribution. This of course would require further investigation, for which this study

provides a methodological basis, notably by introducing metrics to quantify hubness. Therefore, this paper can be thought as a "witness call" to the community to identify similar effects in different application contexts.

The phenomenon of hubs, and notably the evidence of its important sensibility to certain critical frames, illustrates one deep discrepancy between human perception of timbre and all its computation models. Namely, that all frames are not of equal importance, and that these weights does not merely result of their long-term frequencies(i.e. the corresponding component's prior probability π_m). Some timbres (i.e. here sets of frames) are more *salient* than others: for instance, the first thing than one may notice while listening to a given singer's music is his/her particular timbre of voice, independently of the instrumental background (guitar, synthesizer, etc...). This saliency may depend on the context or the knowledge of the listener and is obviously involved in the assessment of similarity. These experiments open the way for more careful investigations of the perceptive paradoxes proper to polyphonic music timbre, in which listeners "hear" things that are not statistically significant in the actual signal, and that the low-level models of timbre similarity studied in this work are intrinsically incapable of capturing.

Acknowledgment

The authors would like to thanks Anthony Beurivé for helping with the implementation of signal processing algorithms and database metadata management. This research has been partially funded by the Semantic Hifi IST European project (http://shf.ircam.fr/).

A Appendix: Some properties of hubness measures

A.1 n-occurrence is constant-sum

An important property of the number of n-occurrences N_n of a song is that the sum of the values for all songs is constant given a database. A query for n neighbors only gives the opportunity for n occurrences to the set of all the other songs, such that the total number of n-occurrences in a given $\mathcal{N}$-size database is $n * \mathcal{N}$. Therefore, the mean n-occurrence of a song is equal to n, independantly of the database and the distance measure.

A.2 Neighbor-angle is constant-sum

An important property of the neighbor-angle value is that, like the number of n-occurrences N_n of a song, the sum of the values for all songs is constant given a database size. This directly derives from the fact that the angles of a triangle sum to 2π radians (in a euclidean geometry - which is only approximated here in the general case). Given a set of $\mathcal{N}$ points, the number of angles whose vertex is a given point X, and are formed by the lines from X to the $\mathcal{N}-1$ other points, is equal to the number of combinations of 2 points within $\mathcal{N}-1$, i.e. $C^2_{\mathcal{N}-1}$. There are $\mathcal{N}$ possible vertices X for such angles, thus there are a total of $\mathcal{N}C^2_{\mathcal{N}-1} = \frac{n(n-1)(n-2)}{2}$ angles formed between the $\mathcal{N}$ points. It is easy to see that $n(n-1)(n-2)$ is divisible by 3 $\forall n$. Hence, these angles can be clustered by triplets, so that their supporting lines form a triangle, and thus sum to 2π. Therefore, the sum of all angles formed between $\mathcal{N}$ points equals $\frac{2\pi}{3}\mathcal{N}C^2_{\mathcal{N}-1}$.

A.3 Neighbor-angle is distance-dependant

The neighbor angle is dependant on the discrimination capacity of the distance, i.e. the typical distance ratio between what can be considered a close distance, and what can be considered a large distance. Therefore it can't be used to compare different algorithms, but to compare the hubness of different songs within the same distance measure.

A.4 Correlation between measures

Further studies show that there is a nearly logarithmic dependency between the number of occurrences of a given song and its mean neighbor angle. This logarithmic behaviour is observed independently of the modelling algorithm (GMMs, HMMs, histograms, etc.). In all cases, it appears that hub songs tend to be associated to higher values of neighbor angle. However, the logarithmic dependency makes it difficult to distinguish songs with number of occurrences in the range 100–200 using their value of neighbor angle. Therefore, in this paper, the former measure is preferred when comparing different settings in the same database.

References

[1] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[2] L. Rabiner and B. Juang, *Fundamentals of speech recognition.* Prentice-Hall, 1993.

[3] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford Press, 1995.

[4] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in *proceedings ISMIR*, 2001.

[5] J.-J. Aucouturier, F. Pachet, and M. Sandler, "The way it sounds: Timbre models for analysis and retrieval of polyphonic music signals," *IEEE Transactions of Multimedia*, vol. 7, no. 6, pp. 1028–1035, December 2005.

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

[7] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, pp. 177–192, 1995.

[8] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves, a statistical analysis of speaker performance," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), Sydney (Australia)*, December 1998.

[9] A. Hicklin, C. Watson, and B. Ulery, "The myth of goats: How many people have fingerprints that are hard to match?" National Institute of Standards and Technology, USA" (Patriot act Report 7271), 2005.

[10] T. Kohonen, *Self-Organizing Maps.* Springer-Verlag, Heidelberg, 1995.

[11] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[12] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, 1989.

[13] F. Pachet, A. LaBurthe, A. Zils, and J.-J. Aucouturier, "Popular music access: The sony music browser," *Journal of the American Society for Information (JASIS), Special Issue on Music Information Retrieval*, 2004.

[14] E. Pampalk, "Computational Models of Music Similarity and their Application to Music Information Retrieval", *Doctoral Thesis*, Vienna University of Technology, Austria, 2006.

[15] P. Bak, *How Nature Works: The science of self-organized criticality.* Springer-Verlag, New York, 1996.

[16] R. Albert, H. Jeoung, and A.-L. Barabasi, "The diameter of the world wide web," *Nature*, vol. 401:130, 1999.

[17] P. Cano, O. Celma, M. Koppenberger and M. Martin-Buldu, 'Topology

of music recommendation networks," *Chaos*, Vol.16, 2006.

[18] N. Lin and H. Zhao, "Are scale-free networks robust to measurement errors?" *Bioinformatics*, vol. 6:119, 2005.

[19] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frame approach to audio pattern recognition: Why this works for urban soundscapes and not for polyphonic music," *Journal of the Acoustical Society of America (submitted)*, 2006.

[20] T. Kinnunen, I. Krkkinen, and P. Frnti, "Is speech data clustered? - statistical analysis of cepstral features," in *European Conf. on Speech Communiation and Technology, (EUROSPEECH'2001), Aalborg, Denmark*, vol. 4, September 2001, pp. 2627–2630.