HUBS AND HOMOGENEITY: IMPROVING CONTENT-BASED MUSIC MODELING

Mark T. Godfrey
Georgia Institute of Technology
Music Technology Group
mark.godfrey@gatech.edu

Parag Chordia
Georgia Institute of Technology
Music Technology Group
ppc@gatech.edu

ABSTRACT

We explore the origins of hubs in timbre-based song modeling in the context of content-based music recommendation and propose several remedies. Specifically, we find that a process of model homogenization, in which certain components of a mixture model are systematically removed, improves performance as measured against several ground-truth similarity metrics. Extending the work of Aucouturier, we introduce several new methods of homogenization. On a subset of the `uspop` data set, model homogenization improves artist R-precision by a maximum of 3.5% and agreement to user collection co-occurrence data by 7.4%. We also explore differences in the effectiveness of the various homogenization methods for hub reduction. Further, we extend the modeling of frame-based MFCC features by using a kernel density estimation approach to non-parametric modeling. We find that such an approach significantly reduces the number of hubs (by 2.6% of the dataset) while improving agreement to ground-truth by 5% and slightly improving artist R-precision as compared with the standard parametric model.

1 INTRODUCTION

Content-based music similarity is a promising but underdeveloped approach to automatic music recommendation. To date, most work in this area has been focused on calculating similarity through comparison of song-level statistical models. However, such systems have thus far yielded only limited results [2], regardless of modeling method. It is thought that this may be connected to the existence of "hubs", songs that are found to be inaccurately similar to a large number of songs in a database. The origin of these hubs has been conjectured, yet no clear strategy for combating them has been established.

We conjecture that some songs are modeled particularly poorly, in effect leaving them far from other songs in the database and thus unlikely to be recommended. These songs, which we call "anti-hubs", are shown to be identifiable from certain properties of their models. In this paper, we propose a method to systematically reduce the incidence of anti-hubs. Another modeling approach suggested by the goal of hub reduction was also explored.

2 METHODOLOGY

2.1 Prior Work

Gaussian mixture models (GMMs) of short-time MFCC frames have been explored extensively and are considered the state-of-the-art approach to content-based song modeling [10, 7, 4, 8, 1]. Typically, the symmetric Kullback-Leibler (KL) divergence is found between these models and is used as a similarity measure. Since there is no closed-form solution for mixture models, this distance must be approximated, usually by a Monte Carlo method [1] or the Earth Mover's distance (EMD) [11].

2.2 Modeling

Following this work, we also use the MFCC-GMM approach for song modeling. While Aucouturier [2] showed 50 components to be optimal, for speed we chose to use 32 components, and empirically deemed these to have sufficient modeling power.

For experiments using non-parametric modeling, we employed kernel density estimation (KDE) [9, 6] as implemented by MATLAB's `ksdensity` routine. Our models therefore consist of a sampled density function for each MFCC dimension, considering each to be independent. We empirically determined a resolution of 1,000 points per density function was sufficient to represent the distributions. The kernel bandwidth, which depends on the number of frames and their median absolute deviation, is scaled to control the smoothness of the density estimate, and this scaling can be varied to explore its effect on model performance.

2.3 Distance

Initial experiments showed the Monte Carlo-based distance to be prohibitively slow for comparing GMMs, and EMD was used instead. For KDE models, we adopted the Bhattacharyya distance [5], a common measure of similarity be-

tween two discrete probability distributions. Note that because each density is sampled over different ranges, we linearly interpolate over the maximum range of the two given models so that each density is defined and compared for common x values.

2.4 Data

These experiments used a subset of the `uspop` collection consisting of 617 songs from 40 artists. This set was intended to match the relative size of Berenzweig's subset [3], while not hand-picking tracks based on intra-class timbral homogeneity and inter-class heterogeneity as with Aucouturier's set [1].

2.5 Hubness

In measuring a kernel's hubness, we adopted the N-occurrences measure used by Aucouturier [1] and Berenzweig [3], choosing N to be 100. This measure is a count of the number of times a song appears in the top-N list of other tracks, in that a large value indicated a hub. Like Aucouturier, we considered a track a hub if its 100-occurrences are greater than 200 (2 times the mean) and an anti-hub if its 100-occurences is less than 20.

2.6 Ground-truth Agreement

In measuring agreement to ground-truth, we first measured each kernel's artist R-precision. This is the percentage of retrieved the R nearest neighbors with the same artist as the seed, where R is the number of the artist's songs in the data set. This corresponds to the common k-NN classifier with leave-one-out cross-validation, except that k is dependent on the size of each seed's class.

As another measure of ground-truth agreement, we used the OpenNap user collection co-occurrence data accompanying the `uspop` collection [4]. Using the top-N rank agreement score, we found how well our computed kernels' neighbor rankings matched the kernel computed from the OpenNap data.

3 HUBS OR ANTI-HUBS?

Berenzweig discovered that models with very few near neighbors, which we now refer to as anti-hubs (and classified by Pampalk as "always dissimilar" [8]), had certain characteristic properties [3]. It was hypothesized that perhaps the focus in the quest for understanding hubs was on the wrong side of the hub distribution: "... hubs may actually be the only songs behaving nicely, while non-hubs [are] pathologically far away from everything else." Because we base our recommendations and, in result, notions of hubness, on nearest neighbors in kernel space, anti-hubs could actually be considered as problematic as hubs. In other words, anti-hubs are absent from their rightful timbral neighborhoods, leaving their would-be neighbors near other songs that are perhaps not perceptually suitable.

	Correlation
Trace of single Gauss. covar.	-0.2432
Max. intra-comp. dist.	-0.3272
Max. comp. dist from centroid	-0.3156

Table 1. Pearson correlation coefficients between 100-occurences count and measures of model spread

We speculate that these anti-hubs originate not from what would be considered perceptually anomalous timbres, but from a relative small number of frames representing transient timbral sections. Because the algorithms used to train song models are musically agnostic (i.e. silence is as musically valid as a chorus), we have found several components of mixture models are spent modeling these unrepresentative timbral sections.

This section demonstrates that models of anti-hubs tend to contain outlier mixture components that can prove detrimental to their parent models' discriminative power. We also propose that anti-hubs are at least easier to identify through measuring attributes of these components and therefore more easily treatable.

3.1 Model Variance

By measuring the overall "variance" of his GMMs, Aucouturier found no correlation with hubness and this measure of model "spread" [1], disproving his hypothesis that hubs are well-connected to other models simply due to a relatively large distribution of frames. However, using three other measures of model spread, we found a negative correlation between model size and hubness, as seen in Table 1. This suggests hubs actually have small spreads compared to anti-hubs, likely indicating that anti-hubs have largely multi-modal distributions.

3.2 Outlier Components

But, as Berenzweig observed [3], the large spread of anti-hubs can be attributed to relatively few mixture components. Berenzweig observed anti-hubs contain components with very small variance, leading to models that are overly specific to a certain region in feature-space and thus making them less likely to match other models. He also found these components tend to have other common attributes: relatively significant prior probabilities so they cannot be ignored as "mathematical nuisances", large distance from the mixture's centroid meaning they are most likely to blame for anti-hubs' overall "wide diameter" models, and close proximity to the origin, suggesting these components are primarily

	Correlation
Min. log-det. of covar.	0.2247
Max. dist. from centroid	−0.3113
Min. dist. from origin	0.3253
Min. prior probability	0.0908

Table 2. Pearson correlation coefficients between 100-occurences count and measures of component attributes

modeling low-energy frames. We found that components of anti-hubs in general can be characterized with the same attributes. To verify, we calculated the Pearson correlation between each model's 100-occurrences count and measurements of the most extreme component according to these attributes. Table 2 shows these correlations, which were all found to be statistically significant.

4 HOMOGENIZATION

Aucouturier concurred that a significant amount of modeling power was being occupied by certain outlier frames, as seen through his experiments with "homogenizing" models [1]. His experiments were based on the idea that components with high prior probabilities model statistically important frames, so that we can, in effect, associate these component weights with component "importance". He then removed components whose prior probabilities fell below a given threshold, producing a "homogenized" version of the original model. Through this experiment, he claimed that most of the variance of a GMM is accounted for by the least 5-10% of the statistically weighted components. Also, he argued that the hubness of a song is based primarily on the least statistically "important" components, as the hubness of his collection increased by nearly a factor of 3 when the models were homogenized to just 90%.

Mixture models, however, typically contain components that are highly overlapped. In this way, the prior probability, or "weight", of a particular component may be low, but together with its neighbor components, could comprise a large mode in the overall density function. Therefore, the prior probabilities alone cannot be assumed to correlate with a component's "importance".

Therefore, we claim that component prior probabilities are not a reliable feature to effectively homogenize mixture models. We instead make use of the correlates to hubness highlighted in the previous section. In particular, we propose to base homogenization around procedures aimed at removing the components characterized by the above features. In each case, practically the same algorithm described by Aucouturier is used: components not meeting a certain defined threshold requirement are discarded and the component weights (prior probabilities) are re-normalized.

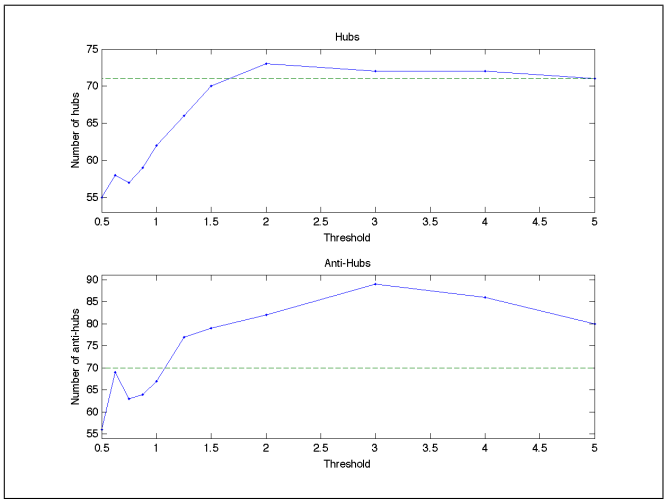

Figure 1. Influence of homogenization by distance from mixture centroid on number of hubs (top) and anti-hubs (bottom) for different thresholds. The un-homogenized amounts are plotted as horizontal lines for reference.

4.1 Homogenization by Distance from Centroid

The first method of homogenization explored was based on the observation that anti-hubs tend to have components that are distant from a main mode of the frame distribution. We therefore discarded components whose Euclidean distance from the component center to the mixture's centroid was greater than a given threshold. The threshold values were determined empirically by observing many activation sequences (showing the likelihood that each frame occurred from each GMM component) of models found from all sections of the hub distribution, as inspired by Berenzweig [3].

4.1.1 Effects on hubness

Figure 1 shows the effects of homogenization on the occurrence of hubs and anti-hubs. Note the symmetry of the un-homogenized distributions: there are approximately the same number of hubs and anti-hubs (71 and 70, respectively). It is clear that the number of hubs and anti-hubs decreased only for severe homogenization levels.

Interestingly, the number of anti-hubs greatly increased after mild homogenization. If anti-hubs become more centralized in distribution space after homogenization as intended, they should attain more neighbors. But would these neighbors be new or are anti-hubs simply getting closer to their previous nearest neighbors? To answer this, we observed where in the hub distribution each song's nearest neighbors existed. It was clear that anti-hubs' only near neighbors tended to be other anti-hubs. Because of this, if we treat some anti-hubs with homogenization, their former neighbors, who were generally unaffected by this procedure, become more severe anti-hubs. Therefore, while several

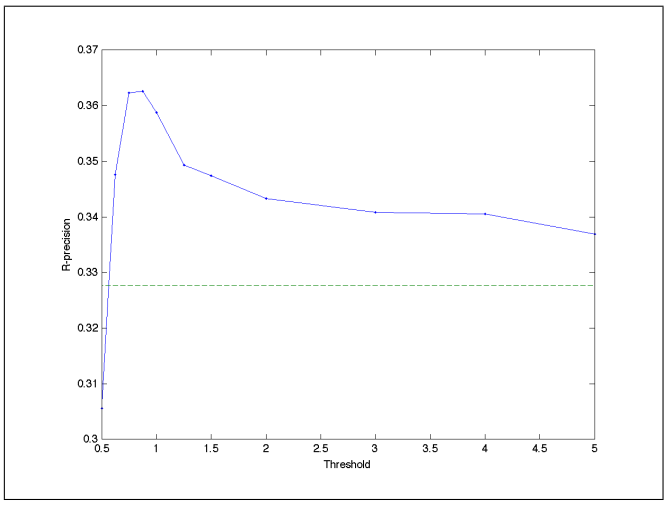

Figure 2. Influence of homogenization by distance from mixture centroid on artist R-precision. The R-precision level before homogenization is plotted as a horizontal line for reference.

anti-hubs are clearly being introduced into the song pool, increasing not only their similarity accuracy but also that of their new neighbors, we see many borderline anti-hubs dropping into the anti-hub region after homogenization.

4.1.2 Effects on agreement to ground-truth

Figure 2 shows the artist R-precision computed for each homogenization by distance from centroid threshold. R-precision increased monotonically until a threshold of 0.875 with a maximum increase of 3.50% (absolute) above the un-homogenized baseline (R-precision with a random kernel was found to be 0.03). This was likely due to anti-hubs being brought nearer to their appropriate perceptual neighbors, as songs by the same artist are generally similar timbrally. After this threshold, the R-precision drops dramatically. The decrease with severe homogenization was no doubt due to song models disintegrating into generic distributions with little discriminating information.

We then computed the top-39 (our dataset contains 40 artists) rank agreement scores against the OpenNap user co-occurrence kernel for each level of homogenization. Each score was averaged over 1,000 runs to smooth out inconsistencies resulting from ties. A Wilcoxon signed-rank test showed that agreement scores for homogenization distance thresholds from 3.0 to 0.875 were significantly higher than the un-homogenized score, where a maximum increase of 5.95% (absolute) over the GMM-EMD kernel was found.

4.2 Homogenization by Variance

We next examined homogenization by variance. With this method, we removed components whose log-determinant did not meet a minimum threshold. The determinant of a covariance matrix can be thought of as a measure of a component's volume, so we were in effect removing components that cover a small region of the timbral space. Again, the existence of such components was shown to be negatively correlated with a model's hubness, so we again expected homogenization to primarily affect models on the lower end of the hub distribution.

Note we were unable to affect all models with this approach without removing all of certain models' components. Therefore, our most severe homogenization level with this method affected only 70% of the models. This highlights an advantage in the use of relative component features (e.g. distance from centroid) in defining the homogenizing function, as opposed to absolute cut-offs (e.g. log-determinant).

4.2.1 Effects on hubness

Unlike the previously discussed homogenization method, this method did not show improvement in hubness at any level. The number of hubs increased by 2 for weak homogenization and remained unchanged for more severe thresholds. The number of anti-hubs, in fact, increased by 16 (2.6% of the dataset) for most homogenization levels. This was assumed to be a result of the aforementioned abandonment of borderline anti-hubs.

4.2.2 Effects on agreement to ground-truth

Despite its apparent detriment to hubness, this homogenization improved agreement to ground-truth. All levels except for the most severe were found to significantly increase over the un-homogenized level, with a maximum increase of 1.26% at the -110 threshold. Note this is only about half of the improvement seen with homogenization by distance from mixture centroid.

Significant improvement was also seen in agreement to the OpenNap user co-occurrence data, fairly consistently across homogenization levels. The maximum increase of 6.18% (absolute) was seen at a log-determinant threshold of -105.

4.3 Homogenization by Distance from Origin

The last homogenization method explored was based on the observation that anti-hubs tend to have components near the origin. These are likely modeling frames with low energy (e.g. silence) and can reasonably be considered not perceptually relevant in relation to the song's timbre. As before, several thresholds were found empirically, and components less than this distance away from the origin were discarded from the model. Like with homogenization by variance, we were only able to treat at most 60% without fulling collapsing certain models whose components were all fairly near the origin.

4.3.1 Effects on hubness

Hubness was not improved for any homogenization level with this method. In fact, like with homogenization by variance, we saw a slight increase in hubs and a considerable increase in anti-hubs, by more than 15 (2.4% of the dataset) for each threshold.

4.3.2 Effects on agreement to ground-truth

We saw that artist R-precision also improved with this homogenization, increasing monotonically with distance from origin threshold. The maximum increase of 3.09% (absolute) was found when discarding components less than 5.5 units from the origin (in MFCC space). All changes were found to be significant under the Wilcoxon signed-rank test.

Agreement with the OpenNap data significantly increased as well with this type of homogenization, increasing monotonically and reaching a maximum of 7.38% (absolute) above the un-homogenized baseline at a threshold of 5.5.

5 NON-PARAMETRIC MODELING

We discussed in Section 3 that using algorithms such as Expectation-Maximization to train parametric mixture models such as GMMs can result in mixture components that are devoted to modeling timbral frames that are not related to perceptually salient sections. This tends to result in models with poor representative power that in turn leads to inaccurately low similarity scores with other models. However, instead of iteratively training a parametric model to fit a given distribution, non-parametric approaches can be used to explicitly model these complex distributions of MFCC frames. In particular, using kernel density estimation (KDE), we are given some control over the effect spurious frames have on a model by increasing the kernel bandwidth. Wider bandwidths yield smoother density functions, effectively reducing the multi-modal behavior shown to be consistent with songs containing outlier frames (i.e. anti-hubs).

Aucouturier compares a type of non-parametric modeling to other modeling algorithms in his thesis [1]. Using three methods (independent histograms and two vector quantization methods), he shows that each performs much worse than the GMM approach, in both R-precision and hubness. Interestingly, our approach here is similar to his independent histograms modeling (which scored 24% lower in R-precision than GMM), in that we treat each MFCC dimension independently, but since we use estimated density functions, we use the Bhattacharyya distance or the Monte Carlo approximated divergence to compare these models instead of Euclidean distance.

We verified that our KDE models were consistent with the GMM models by computing the top-N rank agreement between kernels. We chose N to be 616 and slowly decaying rank weights to allow for a large set of neighbors to impact the scores. The agreement scores are shown in Table 3 and show KDE kernels from both distance metrics agree well with the GMM kernel.

	Agreement
Bhattacharyya	0.8940
Monte Carlo	0.8866
Random	0.3218

Table 3. Top-N rank agreement scores for KDE kernels and the standard GMM kernel using different distance metrics and unity bandwidth scaling

5.1 Hubness

Looking at hubness, however, here was a large discrepancy between the GMM kernel and the KDE-BD kernels. A large decrease was found in both the number of hubs and anti-hubs, as high as 16 and 29 respectively or 23% and 41% of the un-homogenized GMM levels. It seems there is no strong relationship between hubness and the smoothness of the density function. After examining this in more detail, it was shown that the anti-hub region is unaffected by smoothing of the density functions. This goes against our earlier hypothesis that anti-hubs are severely multi-modal, which first led to our experiments with homogenization. We speculated that the smoother the density functions (i.e. the more homogenized the underlying distribution), the more hub-like the model would become. We did see the amount of hubs decrease 9.8% (61 to 55) with increased smoothing, suggesting, if anything, we were decreasing hubs. This could be a result of more models from the middle of the hub distribution moving nearer hubs, thus splitting the former hubs' neighbors amongst the new hubs. In this way, a song simply occupying a centralized region in space (or as Aucouturier calls a "center of mass") does not make it a hub; the song must be relatively alone in this region.

It was also shown that there is a strong correlation (0.788) between the hubness values of GMM and KDE models. In other words, songs that appear as hubs in the GMM kernel are likely to appear as hubs in the KDE kernel. This is contrary to Aucouturier's experiment [1] where he finds a much weaker correlation between hubs appearing from GMMs and his non-parametric histogram models.

5.2 Agreement to ground-truth

No significant difference was found for artist R-precision scores on the KDE kernels as compared to the GMM kernels. However, KDE modeling improved agreement to the OpenNap data, where we saw an improvement of about 5% (absolute) for all bandwidth sizes. Similar results were seen

with the Monte Carlo distance, with a maximum increase over the GMM-EMD baseline of 6% (absolute).

5.3 Computation Time

Aside from apparently better discriminative power, KDE-BD models also showed advantages in necessary computation time. The total computation time to train the KDE models on all 617 songs of the `uspop` subset was found to decrease exponentially with kernel bandwidth. For very small bandwidths, we saw modeling time increase by over a factor of 15 over GMMs, with no apparent detriment to modeling power.

As far as distance computation time, the Bhattacharyya distance (with linear interpolation and 2,000-point density functions) took on average 83 ms. per pair of KDE models, compared to 30 ms. for finding the distance between two GMMs via the Earth Mover's distance [1]. This means computing a KDE-BD kernel took about 2.7 times longer (262 minutes) than the GMM-EMD kernel (95 minutes) on our 617-song `uspop` subset. Speed of the BD computation could of course be improved by employing lower order interpolation and more sparsely sampling the density functions.

The Monte Carlo distance, on the other hand, took significantly longer, averaging 586 ms. per pair (generating 2,000 samples per model), meaning the entire kernel took 31 hours to compute, which is entirely unacceptable in a real-world scenario. Granted, measures could be taken to increase its efficiency, but since the results on the above performance tasks were comparable to the BD kernel, no reason is seen to further use the computationally expensive Monte Carlo-based distance.

6 CONCLUSION

Homogenization of GMMs was shown to improve the hubness of several models, particularly anti-hubs. While the overall amounts of hubs and anti-hubs generally increased after this procedure, this was assumed to be a result of the abandonment of anti-hub neighbors who were themselves untreatable by the given homogenization method. Each method showed significant improvements, however, in agreement to ground-truth data, as shown in Table 4. This is encouraging, since the improved representative power of the models affected by homogenization seems to outweigh the expected loss in models left untreated.

Non-parametric modeling by kernel density estimation proved to offer not only significant reduction in hubness but considerable improvement in computation time and ground-truth agreement.

[1] Computed on a MacPro with 2 2.66 GHz Dual-Core Intel Xeon processors and 2 GB of RAM

Homogenization Method	Artist R-precision	OpenNap Agreement
Dist. from centroid	3.50%	5.95%
Variance	1.26%	6.15%
Dist. from origin	3.10%	7.40%

Table 4. Maximum percent improvement (absolute) of agreement to ground-truth data for different homogenization methods.

Overall, the work presented here suggests approaches to solutions to fundamental problems found in content-based music modeling.

7 REFERENCES

[1] J.-J. Aucouturier. *Ten Experiments on the Modelling of Polyphonic Timbre*. PhD thesis, University of Paris 6, Paris, France, May 2006.

[2] J.-J. Aucouturier and F. Pachet. Improving timbre similiarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[3] A. Berenzweig. *Anchors and Hubs in Audio-based Music Similarity*. PhD thesis, Columbia University, New York City, USA, May 2007.

[4] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. *Computer Music Journal*, 28(2):63–76, June 2004.

[5] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probablity distributions. *Bulletin of the Calcutta Mathematics Society*, 35:99–110, 1943.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[7] Z. Liu and Q. Huang. Content-based indexing and retrieval-by-example in audio. In *IEEE International Conference on Multimedia and Expo.*, pages 877–880, 2000.

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

[9] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[10] D. Pye. Content-based methods for the management of digital music. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, volume 4, pages 2437–2440, Istanbul, Turkey, June 2000.

[11] Y. Rubner, C. Tomasi, and L. J. Guibasg. A metric for distributions with applications to image databases. In *Proc. of the IEEE International Conference on Computer Vision*, pages 59–66, Bombay, India, 1998.