# EXTENDING CONTENT-BASED RECOMMENDATION: THE CASE OF INDIAN CLASSICAL MUSIC

**Parag Chordia**
Georgia Tech
ppc@gatech.edu

**Mark Godfrey**
Georgia Tech
mark.godfrey@gatech.edu

**Alex Rae**
Georgia Tech
arae3@gatech.edu

## ABSTRACT

We describe a series of experiments that attempt to create a content-based similarity model suitable for making recommendations about North Indian classical music (NICM). We introduce a dataset (nicm2008) consisting of 897 tracks of NICM along with substantial ground-truth annotations, including artist, predominant instrument, tonic pitch, *raag*, and parent scale (*thaat*). Using a timbre-based similarity model derived from short-time MFCCs we find that artist R-precision is 32.69% and that the predominant instrument is correctly classified 90.30% of the time. Consistent with previous work, we find that certain tracks ("hubs") appear falsely similar to many other tracks. We find that this problem can be attenuated by model homogenization. We also introduce the use of pitch-class distribution (PCD) features to measure melodic similarity. Its effectiveness is evaluated by *raag* R-precision (16.97%), *thaat* classification accuracy (75.83%), and comparison to reference similarity metrics. We propose that a hybrid timbral-melodic similarity model may be effective for Indian classical music recommendation. Further, this work suggests that "hubs" are a general features of such similarity modeling that may be partially alleviated by model homogenization

## 1 INTRODUCTION AND MOTIVATION

North Indian classical music (NICM) has for the past forty years become an increasingly international phenomenon. A great number of people have had some exposure but otherwise know little about the tradition. This presents an opportunity for a music discovery system that can suggest music based either on known artists or on simple descriptive terms. However, metadata for Indian classical music is often missing or inaccurate, and user-tagging of Indian classical music tracks is uncommon. These problems suggest the use of a content-based recommender. In addition to the goal of exploring models that can be used for content-based recommendation, we hope to explore whether issues observed with the standard timbre-based content-based recommendation (CBR) models, such as the prevalence of many false hits due to a few tracks ("hubs"), are artifacts of the data or appear more generally when novel music is considered. To date, published work on CBR [1, 2, 12] has focused on a few datasets that consist solely of a small slice of Western popular music. Finally we hope to show how properties of the musical genre can be exploited to improve CBR. Because NICM is significantly less polyphonic than most Western music, it relatively easy for us to extract melodic information which can be used in a similarity model.

## 2 BACKGROUND

NICM is one of the oldest continuous musical traditions in the world and it is an active contemporary performance practice. Since the 1960's, due to the emigration of Indians and the popularity of artists such as Ravi Shankar and Zakir Hussain, it has become widely known to international audiences. The repertoire of Indian classical music is extensive, consisting of dozens of styles and hundreds of significant performers. The most prevalent instruments, such as *sitar*, *sarod* and *tabla*, are timbrally quite different from popular Western instruments. NICM is an oral tradition and recordings therefore represent the primary materials.

The performance of Indian classical music typically involves a soloist, either a vocalist or instrumentalist, accompanied by a *tanpura* (drone) throughout and a *tabla* (percussion) in rhythmic sections. Most presentations begin with an extended ametric melodic improvisation (*alap*) and build in intensity throughout the performance. After this section, several compositions are usually presented with *tabla* accompaniment. Here too, the majority of the music is improvised. All NICM is based on *raag*, a melodic concept that defines the melodic parameters of a given piece. There are hundreds of *raags*, of which approximately 150 are widely performed. *Raags* are typically defined by a scale and a set of interrelated phrases. Although highly structured, they allow the performer tremendous scope for improvisation and elaboration. It is this development that forms the core of most NICM performances.

Another important, though quite distinct, performance tradition is solo *tabla*, in which the *tabla* player becomes the soloist. Here timbral and rhythmic patterns form the core material, and the melodic accompaniment is used primarily as a time-keeper.

## 3 DATABASE

The database was assembled from the author's personal music collection. Recordings encompass both commercial and non-commercial sources. Many of the non-commercial recordings are live concerts that are distributed informally amongst NICM listeners. A substantial number of the most historically important recordings are of this type. The recordings span a range from the early $20^{th}$ century to present with the vast majority of recordings being from the second half of the century. Low fidelity is common due to the quality of the initial recording or the number of intermediate analog copies. Common degradations include hiss and crackle, overloading, missing high and/or low frequency content, artifacts from excessive noise reduction processing, and wow due to tape speed fluctuation. The balance of the accompanying drone and *tabla* varies widely, in some cases barely audible, in other cases overwhelming.

The database consist of 897 tracks. Only the first five minutes of each track was used, leading to a total size of approximately seventy hours. This was done to reduce computation time since many of the tracks were over thirty minutes long. A total of 141 artists are contained in the database as well as 14 different instruments. The instruments include *sitar*, *sarod*, *tabla*, *shenai*, *flute*, *violin*, and *pakhawaj*, with the first three being the most common. There are 171 different *raags*. The distribution of tracks amongst the *raags* was uneven and 71 *raags* are represented by only one recording in the database.

All the features used in this study, including MFCCs and pitch-tracks, along with ground-truth annotation will be made available at `paragchordia.com/data/nicm08/`.

### 3.1 Annotation

For each track the main artist, instrument and *raag* was annotated. In duet tracks, of which there were only fourteen, both instruments were noted. A substantial difficulty in analyzing *raag* recordings is that there is no standard reference scale, and each performer is free to choose any pitch for the tonic. Thus, if frequency values are to be later interpreted as scale degrees, the tonic must be known for each recording. For each track an expert listener tuned an oscillator while listening to the performance. The tracks were divided amongst two experts, and tracks that were challenging or ambiguous were reviewed by both. This annotation was not done for solo *tabla* tracks, as they were excluded from the melodic similarity experiments.

## 4 METHOD

We describe the feature extraction and statistical modeling used to develop the timbral and melodic similarity models.

### 4.1 Timbre Modeling

Timbre modeling was done using MFCCs calculated on 20ms windows overlapped by 50%. Twenty coefficients were used excluding the $0^{th}$ coefficient. As mentioned earlier, features were only calculated on the first five minutes of each song.

A model was built for each track by assuming that each MFCC feature vector was a sample from an underlying distribution for the current track. Because a given track is likely to evolve over time, statistical models that are flexible enough to represent multiple clusters in the feature space are typically used. Following earlier approaches [1], a Gaussian mixture model (GMM) was trained for each track based on the frame-based MFCCs. The GMM was trained by initializing the means using the k-means clustering algorithm and then running the EM algorithm [10]. A total of thirty-two components were used for each GMM, each with a diagonal covariance matrix. The model parameters, namely the means and covariance matrices of the Gaussian components, become our model of each track.

The similarity of two tracks was judged by comparing the distributions that had been learned for each track [3, 12]. Although there are many intuitive ways to measure the distance of points in a feature space, it is less obvious how to compare distributions. Several methods have been proposed such as Kullback-Liebler (KL) divergence and Earth Movers Distance (EMD) [13]. KL divergence measures the relative entropy of two distributions: that is, the reduction in uncertainty for one distribution if the other is known. EMD has been widely applied to the comparison of GMMs. The algorithm considers the minimum cost to "move" the probability mass of the first distribution so it resembles the second. In one dimension it is easy to visualize: each GMM is a set of hills and the hills are moved and probability mass shifted from one to another until they are matched. Another approach that is perhaps the most natural is to compute the likelihood that the features vectors of one track are generated by the other tracks distribution. This last method, while intuitive, is rarely used because of the computational cost. Regardless of the method employed, the distance metric allows us to calculate a scalar value representing the measure of similarity between tracks. If we have $n$ tracks then $n^2$ distances must be calculated, or $(n)(n+1)/2$ if the distance measure is symmetric, meaning that our computation time will increase as a square of the number of tracks we wish to analyze. The arrangement of all such distance pairs forms a similarity matrix. For any given seed song in the database the similarity matrix can then be used to retrieve the $k$ nearest neighbors, which can then be used as candidates for recommendation.

### 4.2 Melody Modeling

NICM, as noted in the introduction, is based on *raag*. In addition to specifying melodic constraints, *raags* are tradition-

ally thought to elicit certain emotions in listeners. Chordia [7] empirically demonstrated that certain *raags* consistently elicit certain emotions, such as joy or sadness, even for listeners with little or no prior exposure to NICM. Thus *raag* identification is an important descriptor of both melodic and emotional content of a track. Chordia [6] demonstrated that pitch-class distributions (PCDs) could be used to recognize *raags* using a variety of classification algorithms. Further, PCDs might reveal connections or perceptual similarities between particular recordings beyond those suggested by *raag* name alone.

These insights are used in the current system to build a melodic similarity model. First, each piece was pitch tracked using the YIN [9] algorithm. Each pitch estimate was then converted to a scale degree using the manually annotated tonic. Given the tonic, the locations of the scale degrees were computed using the ratios that define the chromatic notes of a just intoned scale. The pitch estimate at each frame was compared to the ideal scale values in the log domain and assigned to the nearest scale degree. The octave information was then discarded, giving a sequence of pitch-classes. A histogram was then computed yielding a pitch-class distribution for the track. Because of the consistent presence of the drone, the tonic value usually overwhelms all other scale degrees without providing any useful discriminative power and was therefore discarded. Thus each track in the database was characterized by one eleven dimensional feature vector.

In addition to a pitch estimate, the YIN algorithm returns a pitch aperiodicity measure which was used to weight the pitch estimates. In one case, which we call linear pitch salience weighting, pitch aperiodicity was converted to a pitch salience measure as $1 - pitchAperiodicty$. In a second case, called ratio weighting, the pitch salience was defined as $1/pitchAperiodicity$. Previous work [8] showed improved performance on the *raag* classification task after weighting, particularly for ratio weighting. In practice, such weighting tends to eliminate or de-emphasize regions of the track where the soloist is silent and the pitch track is therefore noisy and uninformative.

A melodic similarity matrix was constructed by evaluating the distance between pairs of PCDs for each of the conditions. Two distance metrics were used, Pearson correlation and Euclidean distance, with each of these yielding a distinct similarity matrix. The similarity models were then used to retrieve the $k$ nearest neighbors for each track.

## 5 EVALUATION

In the end, we are interested in how well such a similarity engine might perform for such tasks as making purchase recommendations, or suggesting songs in playlist. As such, the final evaluation requires building an application and judging its success in terms of user engagement or purchases.

| model type | Accuracy Rates | | |
|---|---|---|---|
| | k = 1 | k = 5 | k = 10 |
| GMM | 81.27% | 71.28% | 65.64% |
| GMM - homogenized | 90.30% | 80.57% | 73.36% |
| non-parametric | 87.96% | 75.96% | 67.97% |

**Table 1**. Instrument classification accuracy using three different modeling techniques with a k-NN classifier. GMMs without homogenization using Earth Mover's Distance, homogenized with log determinant threshold of -150, and non-parametric modeling are shown.

However, before building such a system we would like to have some sense of whether our similarity judgments are appropriate. Various proxy tasks are used that hopefully correlate with the ultimate utility of the system. For example, Berenzweig [2] and Aucouturier [1] have used artist R-precision and classification tasks to judge the quality of the recommendations based on their timbre models.

Although it is true that an artist's sound may change from recording to recording, it is nevertheless likely that many artists will be timbrally consistent and thus distinguishable from each other . Therefore if our timbre model tends to return hits of the same artist we would tend to think it is doing better than if it does not. In addition to tasks based on artist name, we also evaluated R-precision and classification accuracy for the predominant instrument. Similarly, for melodic evaluation, we considered *raag* R-precision, *thaat* classification, and correlation with a ground-truth matrix expressing known relationships between *raags*.

### 5.1 Timbral Evaluation

Following the standard definition, we define R-precision to be the number of relevant hits for a given seed divided by the total possible relevant hits in the database. For example, in the instrument task, we consider the relevant hits to be the number of the $R$ nearest neighbors that have the same instrument label as the seed, where $R$ represents the total number of tracks with the same instrument as the seed. For the artist task, average R-precision was 26.96% using a GMM model and EMD compared with 2.08% for neighbors chosen randomly. A related task is classification, in which the $k$ nearest neighbors are used to classify the seed track. Table 1 gives the classification performance for recognition of the predominant instrument. Nearest neighbor ($k = 1$) classification performs best with an accuracy of 81.27% for fourteen instrument targets. In the rare case (14 tracks) where there are two main instruments, we consider classification successful if either of the main instruments is matched.

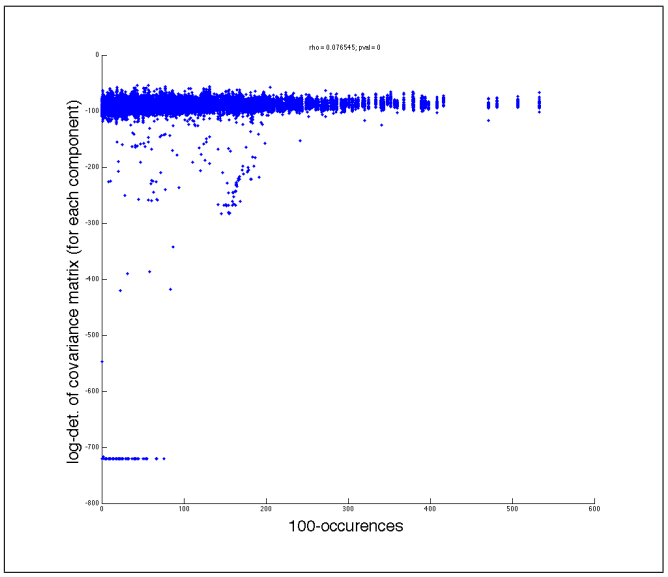


**Figure 1**. Log of the determinant for the components of GMMs. It can be seen that certain tracks have outlier components where the log-determinant is very small.

### 5.2 Hubness

It has been previously observed that a substantial problem with the standard timbre modeling strategies for CBR is the existence of hubs, certain tracks which inappropriately appear as hits for many different seeds [1]. More precisely, we define a hub to be a track that is a top one hundred hit for at least 200 of the seeds in the database (twice the average). In our dataset a total of 105 tracks, or 11.71% of the database, fit this definition. Recent work by Godfrey [11] has shown that these hubs may arise because certain tracks, termed "anti-hubs", are effectively taken out of the pool of possible neighbors due to poor modeling by the GMM. 121 tracks in our dataset (13.49%) were anti-hubs, which we defined as those that match less than twenty seeds. Figure 1 shows the log-determinant of the covariance matrix for each GMM component of each track. The determinant of the covariance matrix can be thought of as a measure of the volume of the Gaussian data cloud. We see that certain components have nearly zero volume. By looking at an activation matrix, which indicates when a GMM component is active in a track, Godfrey found that such components are often active only for very short segments of the track and are otherwise unused.

A possible solution to this problem is the removal of degenerate components of the GMM. This homogenized model may then have a more evenly distributed hub histogram. To see the effect of this, we reduced each GMM by removing components where the log of the determinant was less than a specified threshold. This was done for three levels as shown in Table 2. The similarity matrix was then recomputed using

| Threshold | R-Precision |
|---|---|
| -300 | 0.3040 |
| -200 | 0.3215 |
| -150 | 0.3269 |

**Table 2**. Artist R-Precision results obtained using homogenization

these new models. We find that artist R-precision increases by 5.73 percentage points to 32.69% for the maximum level of homogenization, and instrument classification accuracy increases by 9.03 percentage points to 90.30%. This is consistent with results on the uspop2002 dataset, and suggests that the problems of hubs is general and may be ameliorated by model homogenization [11].

An alternative approach to the hub problem is to use alternative modeling strategies. One such approach is to model feature vectors non-parametrically rather than using a GMM. In kernel density estimation, the points in the feature space representing observations (e.g. MFCC feature vectors) are essentially convolved with a window such as a Gaussian, and summed to yield the density estimate [10]. The estimated densities are then compared using a metric such as the Bhattacharya distance [5], defined as

$$B(p,q) = -log \sum_{x \in X} \sqrt{p(x)q(x)}. \quad (1)$$

Using such an approach, we found that instrument R-precision increased by 3.14 percentage points to 30.10% and instrument classification accuracy improved 6.69 percentage points to 87.96% compared with the non-homogenized GMM model. Again, this is consistent with experiments performed on the uspop2002 dataset [11], suggesting that this modeling approach may be broadly applicable. A further advantage of this non-parametric approach is that computation time is greatly reduced, primarily because the iterative EM algorithm step can be skipped. The bandwidth of the kernel was varied but was not found to have a significant effect over a range of reasonable values.

### 5.3 Melody Evaluation

*Raag*, as noted above, is the most essential melodic description of NICM. Our first evaluation task was therefore *raag* R-precision. Accuracy was 16.97% compared to a random baseline of 1.18%. We excluded *tabla* and *pakhawaj* solos as well as a few semi-classical tracks that did not have a clearly defined *raag*. The sparseness of the data, with respect to certain *raags*, led us to additionally classify each track according to parent scale, or *thaat*. The system of abstract parent scale-types, developed by Bhatkande in the early $20^{th}$ century [4], consists of a mapping of the wide variety of scale-types used in *raags* to a small set of seven

note scales that were considered to represent basic melodic spaces. This step reduced the number of melodic categories to ten. Classification accuracy for *thaats* was 75.83% using the nearest neighbor. Surprisingly, neither of the pitch-salience weighted vectors performed better than the unweighted PCDs.

Although R-precision and classification tasks are informative, we would like to be able to compare the similarity matrix to some ideal ground-truth similarity matrix. Although no such data exist based on perceptual experiments, reasonable references may nevertheless be built.

A ground-truth similarity matrix was built by converting each *raag* to a binary PCD vector that indicated whether each of the twelve scale degrees was used in that *raag*. Distances between *raags* were then computed using the Hamming distance, which counts the number of mismatches between the vectors. It should be noted that a match is found when both *raags* use a certain scale degree and also when both omit a certain scale degree. We also experimented with the inner product, which counts the number of scale degrees that are shared between the *raags*. However, we found that the Hamming distance matched intuitive notions of distance better since the absence of scale degree in a *raag* is as important as its presence.

One approach to comparing similarity matrices is to generate lists of hits for each seed using both the empirical similarity matrix and the ground-truth similarity matrix. The average distance of the hits from the seed would be computed for all seeds in the empirical similarity matrix, for example based on the Euclidean distance between each retrieved track and the seed. Likewise the distances of the neighbors fetched according to the ground-truth matrix could be calculated. The averages over each model might then be compared. The problem with such an approach is that the distances cannot be directly be compared, since they use two different distance metrics. For example, in our case the ground-truth distances are based on Hamming distances between binary PCDs while the empirical distances are based on correlation between the continuous-valued PCDs. One approach that has been proposed to solve this problem is to use the rank order rather than the distance. Taking inspiration from measures used in text retrieval, Berenzweig et al. [3] define what they call the Top-N Ranking agreement:

$$s_i = \sum_{r=1}^{N} (\alpha_r)^r (\alpha_c)^{k^r}, \qquad (2)$$

where $k_r$ is the ranking according to the empirical similarity matrix of the $r^{th}$-ranked hit using the ground-truth matrix. The score is computed for each seed and averaged to give the overall agreement. The $\alpha_c$ and $\alpha_r$ parameters determine the sensitivity of the score to ordering in each of the lists generated by the two metrics.

In order to compare the distances directly we used Hamming distance on the empirical results. This was done by converting each track into a binary PCD vector based on its *raag* label so that the Hamming distance could be applied. This made it easy to compare empirical performance to an upper bound, whereas the Top-N Ranking agreement can be difficult to interpret.

| similarity matrix | Average Distances | | |
|---|---|---|---|
| | k = 1 | k = 5 | k = 10 |
| ground-truth | 0.0331 | 0.1444 | 0.2770 |
| empirical | 1.831 | 2.3669 | 2.6235 |
| random | 4.682 | 4.668 | 4.6442 |

**Table 3**. Average distances from seeds calculated with binary PCDs for ground-truth, empirical, and random similarity matrices.

Results are substantially better than random. The rank list score was .1085 compared with .0087 for the random, with $N = 10$, $\alpha_c = .5^{1/3}$, and $\alpha_r = .5^{2/3}$. Using our method for direct distance comparison, the average distance for $k = 5$ was .1444 for the ground truth matrix and 4.668 for the random case. The value for the empirical similarity matrix fell in the middle of this range (2.367). These results are summarized in table 3.

We also found that hubness occurred with PCD vectors, although to a lesser extent than with timbral modeling. Hub and anti-hub percentages using the unweighted PCD feature and Hamming distance were 8.9% and 4.7% respectively, using the same definition as above. Not surprisingly this was less than for the timbre models, most likely due to the lower dimensionality of the feature space (eleven vs. twenty). Berenzweig demonstrated that hubness increases with the dimensionality of the feature space [2].

## 6 DISCUSSION

Although the artist R-precision is relatively low (32.69%), this is expected because in NICM instrumentation is quite similar for many artists. Without further annotation indicating higher-level artist clusters, we would expect relatively low precision values. That instrument classification accuracy was over 90% suggests that the timbre model is capturing essential information. The instrument classification result is particularly encouraging since it is likely that connecting tracks with the same main instrument would be important for a recommendation system. Similarly, although *raag* R-precision was low, *thaat* identification was correct in more than three out of four cases. Comparisons to random and best-case scenarios suggest that despite the difficulties of pitch tracking real recordings, PCDs are sufficiently robust to provide useful melodic information.

The results presented suggest that MFCC based timbre modeling is effective for NICM and generalizes beyond Western popular music. Further we find support for the idea that hubs may be a general problem when models are constructed using GMMs on frame-based MFCC features. This work also supports the observation that model homogenization may lead to improved retrieval precision and classification accuracy.

## 7 APPLICATION

We anticipate that the similarity modeling presented here could be used as the basis for a music recommendation system based on both timbral and melodic characteristics. One way of combining them would be to create a global distance metric. In the simplest case, one could weight timbre and melody equally and simply sum the distances in the respective similarity matrices. More likely we might imagine giving users control over the extent to which recommendations were controlled by one parameter or the other. This could be done explicitly, for example through a slider interface, or implicitly in a live application by tracking the perceived quality of the recommendations, for example by allowing users to rate the suggested tracks. We conjecture that this combined model approach may also allow adaptation to different patterns of user preferences; timbre might dominate the quality judgments of some users, while others might be more responsive to melodic content.

## 8 FUTURE WORK

Although encouraging, the current models are clearly quite simplistic and the results suggest there is ample room for improvement. On the timbral side, although including more features may improve performance incrementally, as is often noted, significant improvement will require a more perceptually grounded model. For melodic modeling we intend to generalize PCDs to include sequential structure though $n$-gram modeling. Earlier work has shown that this markedly improves *raag* classification performance [6]. Pitch tracking could also be improved by suppression of accompaniment.

We hope to use the techniques discussed here to create a CBR system for Indian classical music in which listeners will be able to generate playlists based either on artists or tracks, or alternatively based on simple emotional descriptors. This would allow us to replace evaluation based on measures such as R-precision and artificial ground-truth matrices with more objective measures of the success of these models in generating music streams.

## 9 REFERENCES

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

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

[3] 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.

[4] V.N. Bhatkande. *Hindusthani Sangeet Paddhati*. Sangeet Karyalaya, 1934.

[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] Parag Chordia and Alex Rae. Raag recognition using pitch-class and pitch-class dyad distributions. In *Proceedings of International Conference on Music Information Retrieval*, 2007.

[7] Parag Chordia and Alex Rae. Understanding emotion in raag music. In *Proceedings of International Computer Music Conference*, 2007.

[8] Parag Chordia, Alex Rae, and Jagadeeswaran Jayaprakash. Automatic carnatic raag classification. In *Proceedings of the Digital Audio Effects Conference (sumbitted)*, 2008.

[9] Alain de Cheveigne and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917 – 1930, 2002.

[10] R. Duda, P. Hart, and D. Stork. *Pattern Recognition and Scene Analysis*. John Willey, 2001.

[11] Mark Godfrey. Hubs and homogeneity: Improving content-based modeling. Master's thesis, Georgia Institute of Technology, Atlanta, Georgia, April 2008.

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

[13] 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.