Incorporating Machine-Learning into Music Similarity Estimation

Kris West
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
kw@cmp.uea.ac.uk

Stephen Cox
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
sjc@cmp.uea.ac.uk

Paul Lamere
Sun Microsystems
Laboratories
Burlington, MA
paul.lamere@sun.com

ABSTRACT

Music is a complex form of communication in which both artists and cultures express their ideas and identity. When we listen to music we do not simply perceive the acoustics of the sound in a temporal pattern, but also its relationship to other sounds, songs, artists, cultures and emotions. Owing to the complex, culturally-defined distribution of acoustic and temporal patterns amongst these relationships, it is unlikely that a general *audio* similarity metric will be suitable as a *music* similarity metric. Hence, we are unlikely to be able to emulate human perception of the similarity of songs without making reference to some historical or cultural context.

The success of music classification systems, demonstrates that this difficulty can be overcome by learning the complex relationships between audio features and the metadata classes to be predicted. We present two approaches to the construction of music similarity metrics based on the use of a classification model to extract high-level descriptions of the music. These approaches achieve a very high-level of performance and do not produce the occasional spurious results or 'hubs' that conventional music similarity techniques produce.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms

Keywords

Music Similarity, Machine-learning, audio

1. INTRODUCTION

The recent growth of digital music distribution and the rapid expansion of both personal music collections and the capacity of the devices on which they are stored has increased both the need for and the utility of effective techniques for organising, browsing and visualising music collections and generating playlists. All of these applications require an indication of the similarity between examples. The utility of content-based metrics for estimating similarity between songs is well-known in the Music Information Retrieval (MIR) community [2][10][12] as they substitute relatively cheap computational resources for expensive human editors, and allow users to access the 'long tail' (music that might not have been reviewed or widely distributed, making reviews or usage data difficult to collect)[1].

It is our contention that content-based music similarity estimators are not easily defined as expert systems because relationships between musical concepts, that form our musical cultures, are defined in a complex, ad-hoc manner, with no apparent intrinsic organising principle. Therefore, effective music similarity estimators must reference some form of historical or cultural context in order to effectively emulate human estimates of similarity. Automatic estimators will also be constrained by the information on which they were trained and will likely develop a 'subjective' view of music, in a similar way to a human listener.

In the rest of this introduction, we briefly describe existing audio music similarity techniques, common mistakes made by those techniques and some analogies between our approach and the human use of contextual or cultural labels in music description. In sections 2 - 5 we describe our audio pre-processing front-end, our work in machine-learning and classification and provide two examples of extending this work to form 'timbral' music similarity functions that incorporate musical knowledge learnt by the classification model. Finally, we discuss effective evaluation of our solutions and our plans for further work in this field.

1.1 Existing work in audio music similarity estimation

A number of content-based methods of estimating the similarity of audio music recordings have been proposed. Many of these techniques consider only short-time spectral features, related to the timbre of the audio, and ignore most of the pich, loudness and timing information in the songs considered. We refer to such techniques as 'timbral' music similarity functions.

Logan and Salomon [10] present an audio content-based method of estimating the timbral similarity of two pieces of music that has been successfully applied to playlist generation, artist identification and genre classification of music. This method is based on the comparison of a 'signature' for each track with the Earth Mover's Distance (EMD). The signature is formed by the clustering of Mel-frequency Cepstral Coefficients (MFCCs), calculated for 30 millisecond frames of the audio signal, using the K-means algorithm.

Another content-based method of similarity estimation, also based on the calculation of MFCCs from the audio signal, is presented by Aucouturier and Pachet [2]. A mixture of Gaussian distributions are trained on the MFCC vectors from each song and are compared by sampling in order to estimate the timbral similarity of two pieces. Aucouturier and Pachet report that their system identifies surprising associations between certain songs, often from very different genres of music, which they exploit in the calculation of an 'Aha' factor. 'Aha' is calculated by comparing the content-based 'timbral' distance measure to a metric based on textual metadata. Pairs of tracks identified as having similar timbres, but whose metadata does not indicate that they might be similar, are assigned high values of the 'Aha' factor. It is our contention that these associations are due to confusion between superficially similar timbres, such as a plucked lute and a plucked guitar string or the confusion between a Folk, a Rock and a World track, described in [2], which all contain acoustic guitar playing and gentle male voice. A deeper analysis might separate these timbres and prevent errors that may lead to very poor performance on tasks such as playlist generation or song recommendation. Aucouturier and Pachet define a weighted combination of their similarity metric with a metric based on textual metadata, allowing the user to adjust the number of these confusions. Reliance on the presence of textual metadata effectively eliminates the benefits of a purely content-based similarity metric.

A similar method is applied to the estimation of similarity between tracks, artist identification and genre classification of music by Pampalk, Flexer and Widmer [12]. Again, a spectral feature set based on the extraction of MFCCs is used and augmented with an estimation of the fluctuation patterns of the MFCC vectors over 6 second windows. Efficient classification is implemented by calculating either the EMD or comparing mixtures of Gaussian distributions of the features, in the same way as Aucouturier and Pachet [2], and assigning to the most common class label amongst the nearest neighbours. Pampalk, Pohle and Widmer [13] demonstrate the use of this technique for playlist generation, and refine the generated playlists with negative feedback from user's 'skipping behaviour'.

1.2 Contextual label use in music description

Human beings often leverage contextual or cultural labels when describing music. A single description might contain references to one or more genres or styles of music, a particular period in time, similar artists or the emotional content of the music, and are rarely limited to a single descriptive label. For example the music of Damien Marley has been described as "a mix of original dancehall reggae with an R&B/Hip Hop vibe"[1]. There are few analogies to this type of description in existing content-based audio music similarity techniques: these techniques do not learn how the feature space relates to the 'musical concept' space.

Purely metadata-based methods of similarity judgement have to make use of metadata applied by human annotators. However, these labels introduce their own problems. Detailed music description by an annotator takes a significant amount of time, labels can only be applied to known examples (so novel music cannot be analysed until it has been annotated), and it can be difficult to achieve a consensus on music description, even amongst expert listeners.

1.3 Challenges in music similarity estimation

Our initial attempts at the construction of content-based 'timbral' audio music similarity techniques showed that the use of simple distance measurements performed within a 'raw' feature space, despite generally good performance, can produce bad errors in judgement of musical similarity. Such measurements are not sufficiently sophisticated to effectively emulate human perceptions of the similarity between songs, as they completely ignore the highly detailed, non-linear mapping between musical concepts, such as timbres, and musical contexts, such as genres, which help to define our musical cultures and identities. Therefore, we believe a deeper analysis of the relationship between the acoustic features and the culturally complex definition of musical styles must be performed prior to estimating similarity. Such an analysis might involve detecting nuances of a particular group of timbres, perhaps indicating playing styles or tunings that indicate a particular style or genre of music.

The success of music classification systems, implemented by supervised learning algorithms, demonstrates that this difficulty can be overcome by learning the complex relationships between features calculated from the audio and the metadata classes to be predicted, such as the genre or the artist that produced the song. In much of the existing literature, classification models are used to assess the usefulness of calculated features in music similarity measures based on distance metrics or to optimise certain parameters, but do not address the issue of using information and associations, learnt by the model, to compare songs for similarity. In this paper we introduce two intuitive extensions of a music classification model to audio similarity estimation. These models are trained to classify music according to genre, as we believe this to be the most informative label type for the construction of 'macro' (general) -similarity metrics. Other more specific label sets, such as mood or artist, could be used to build more 'micro' (specific) similarity functions.

2. AUDIO PRE-PROCESSING

A suitable set of features must be calculated from the audio signal to be used as input to our audio description techniques. In this paper, we use features describing spectral envelope, primarily related to the timbre of the audio, which define a 'timbral' similarity function. The techniques we introduce could be extended to other types of similarity function, such as rhythm or melody, by simply replacing these

[1]http://cd.ciao.co.uk/

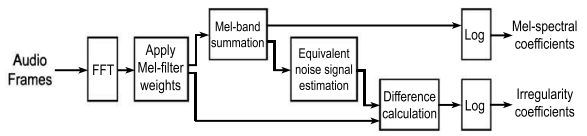

Figure 1: Overview of the Mel-Frequency Spectral Irregularity caclculation.

features with other appropriate features. The audio signal is divided into a sequence of 50% overlapping, 23ms frames, and a set of novel features, collectively known as Mel-Frequency Spectral Irregularities (MFSIs) are extracted to describe the timbre of each frame of audio, as described in West and Lamere [15]. MFSIs are calculated from the output of a Mel-frequency scale filter bank and are composed of two sets of coefficients: Mel-frequency spectral coefficients (as used in the calculation of MFCCs, without the Discrete Cosine Transform) and Mel-frequency irregularity coefficients (similar to the Octave-scale Spectral Irregularity Feature as described by Jiang et al. [7]). The Mel-frequency irregularity coefficients include a measure of how different the signal is from white noise in each band. This helps to differentiate frames from pitched and noisy signals that may have the same spectrum, such as string instruments and drums, or to differentiate complex mixes of timbres with similar spectral envelopes.

The first stage in the calculation of Mel-frequency irregularity coefficients is to perform a Discrete Fast Fourier transform of each frame and to the apply weights corresponding to each band of a Mel-filterbank. Mel-frequency spectral coefficients are produced by summing the weighted FFT magnitude coefficients for the corresponding band. Mel-frequency irregularity coefficients are calculated by estimating the absolute sum of the differences between the weighted FFT magnitude coefficients and the weighted coefficients of a white noise signal that would have produced the same Mel-frequency spectral coefficient in that band. Higher values of the irregularity coefficients indicate that the energy is highly localised in the band and therefore indicate more of a pitched signal than a noise signal. An overview of the Spectral Irregularity calculation is given in figure 1.

As a final step, an onset detection function is calculated and used to segment the sequence of descriptor frames into units corresponding to a single audio event, as described in West and Cox [14]. The mean and variance of the Mel-frequency irregularity and spectral coefficients are calculated over each segment, to capture the temporal variation of the features, outputting a single vector per segment. This variable length sequence of mean and variance vectors is used to train the classification models.

3. MUSIC CLASSIFICATION

The classification model used in this work was described in West and Cox [14] and West and Lamere [15]. A heavily modified Classification and Regression Tree is built and recursively split by transforming the data at each node with a Fisher's criterion multi-class linear discriminant analysis, enumerating all the combinations of the available classes of data into two groups (without repetition, permutation or reflection) and training a pair of Gaussian distributions to reproduce this split on novel data. The combination of classes that yields the maximum reduction in the entropy of the classes of data at the node (i.e. produces the most 'pure' pair of leaf nodes) is selected as the final split of the node.

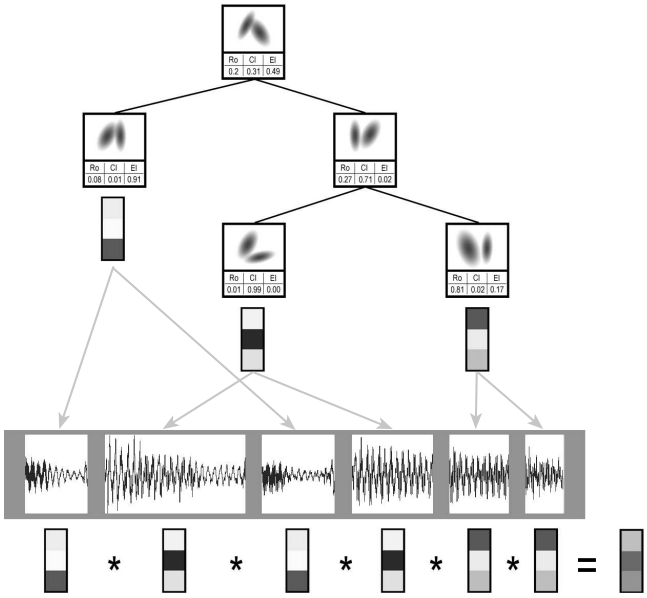

Figure 2: Combining likelihood's from segment classification to construct an overall likelihood profile.

A simple threshold of the number of examples at each node, established by experimentation, is used to prevent the tree from growing too large by stopping the splitting process on that particular branch/node. Experimentation has shown that this modified version of the CART tree algorithm does not benefit from pruning, but may still overfit the data if allowed to grow too large. In *artist filtered* experiments, where artists appearing in the training dataset do not appear in the evaluation dataset, overfitted models reduced accuracy at both classification and similarity estimation. In all unfiltered experiments the largest trees provided the best performance, suggesting that specific characteristics of the artists in the training data had been overfitted, resulting in over-optimistic evaluation scores. The potential for this type of over-fitting in music classification and similarity estimation is explored by Pampalk [11].

A feature vector follows a path through the tree which terminates at a leaf node. It is then classified as the most common data label at this node, as estimated from the training set. In order to classify a *sequence* of feature vectors, we estimate a degree of support (probability of class membership) for each of the classes by dividing the number of examples of each class by the total number of examples of the leaf node and smoothing with Lidstone's method [9]. Because our audio pre-processing front-end provides us with a variable length sequence of vectors and not a single feature vector per example, we normalise the likelihood of classification for each class by the total number of vectors for that class in the training set, to avoid outputting over-optimistic likelihoods for the best represented classes with high numbers of audio

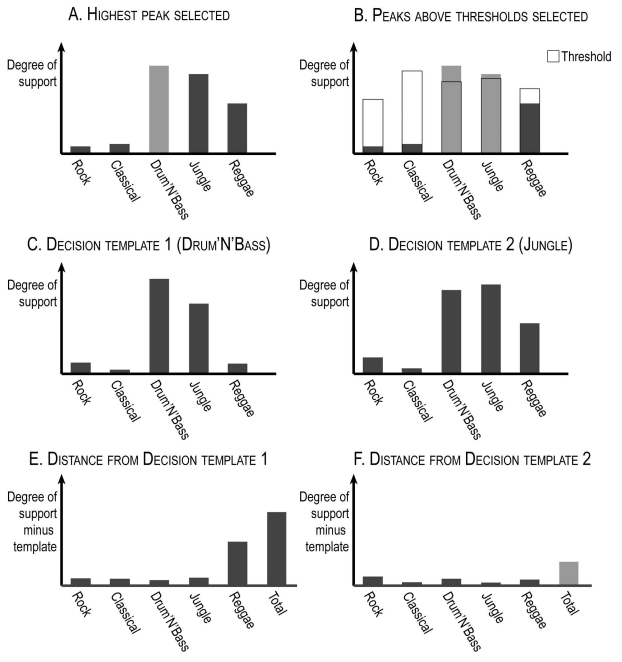

Figure 3: Selecting an output label from the classification likelihood profiles.

events. The logarithm of these likelihoods is summed across all the vectors (as shown in figure 2), which is equivalent to taking the product of the raw probabilities, to produce the log classification likelihood profile from which the final classification will be determined.

The output label is often chosen as the label with the highest degree of support (see figure 3A); however, a number of alternative schemes are available, as shown in figure 3. Multiple labels can be applied to an example by defining a threshold for each label, as shown in figure 3B, where the outline indicates the thresholds that must be exceeded in order to apply a label. Selection of the highest peak abstracts information in the degrees of support which could have been used in the final classification decision. One method of leveraging this information is to calculate a 'decision template' (see Kuncheva [8]) for each class of audio (figure 3C and D), which is the average profile for examples of that class. A decision is made by calculating the distance of a profile for an example from the available 'decision templates' (figure 3E and F) and selecting the closest. Distance metrics used include the Euclidean, Mahalanobis and Cosine distances. This method can also be used to combine the output from several classifiers, as the 'decision template' is simply extended to contain a degree of support for each label from each classifier. Even when based on a single classifier, a decision template can improve the performance of a classification system that outputs continuous degrees of support, as it can help to resolve common confusions where selecting the highest peak is not always correct. For example, Drum and Bass always has a similar degree of support to Jungle music (being very similar types of music); however, Jungle can be reliably identified if there is also a high degree of support for Reggae music, which is uncommon for Drum and Bass profiles.

4. CONSTRUCTING SIMILARITY FUNCTIONS

In this section we detail two methods of extending the CART-based classification model to the construction of music similarity functions.

4.1 Comparison of Likelihood profiles

The real-valued likelihood profiles output by the classification scheme described in section 3 can be used to assign an example to the class with the most similar average profile in a decision template system. We speculate that the same comparison can be made between two examples to estimate their *musical* similarity. For simplicity, we describe a system based on a single classifier; however, it is simple to extend this technique to multiple classifiers, multiple label sets (genre, artist or mood) and feature sets/dimensions of similarity by simple concatenation of the likelihood matrices, or by early integration of the likelihoods (for homogenous label sets), using a constrained regression model combiner.

Let $P_x = \{c_1^x, \ldots, c_n^x\}$ be the profile for example x, where c_i^x is the probability returned by the classifier that example x belongs to class i , and $\sum_{i=1}^{n} c_i^x = 1$, which ensures that similarities returned are in the range [0:1]. The similarity, $S_{A,B}$, between two examples, A and B can be estimated as one minus the Euclidean distance, between their profiles, P_A and P_B , or as the Cosine distance. In our tests the Cosine distance has always performed better than the Euclidean distance.

This approach is somewhat similar to the 'anchor space' described by Berenzweig, Ellis and Lawrence [4], where clouds of likelihood profiles, for each vector in a song, are compared with the KL divergence, EMD or Euclidean distance between sets of centroids. We believe the smoothed product of likelihoods may outperform comparison of the centroids of the mixture of distributions and comparison of likelihood profiles with either the cosine of euclidean distances is less complex than calculating either the KL divergence or EMD.

4.2 Comparison of 'text-like' transcriptions

The comparison of likelihood profiles abstracts a lot of information when estimating similarity, by discarding the specific leaf node that produced each likelihood for each frame. A powerful alternative to this is to view the Decision tree as a hierachical taxonomy of the audio segments in the training database, where each taxon is defined by its explicit differences and implicit similarities to its parent and sibling (Differentialism). The leaf nodes of this taxonomy can be used to label a sequence of input frames or segments and provide a 'text-like' transcription of the music. It should be stressed that such 'text-like' transcriptions are in no way intended to correspond to the transcription of music in any established notation and are somewhat subjective as the same taxonomy can only be produced by a specific model and training set. An example of this process is shown in figure 4. This transcription can be used to index, classify and search music using standard retrieval techniques. These transcriptions give a much more detailed view of the timbres appearing in a song and should be capable of producing a similarity function with a finer resolution than the 'macro' similarity function produced by the comparison of likelihood profiles.

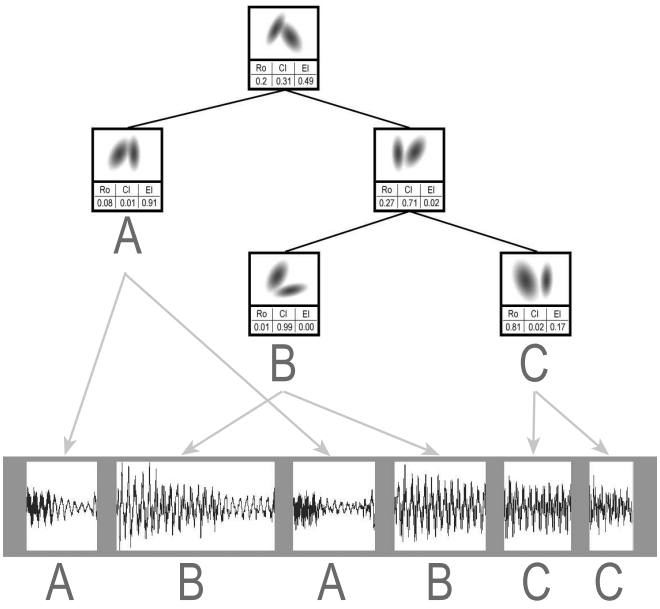

Figure 4: Extracting a 'text-like' transcription of a song from the modified CART.

To demonstrate the utility of these transcriptions we have implemented a basic Vector model text search, where the transcription is converted into a fixed size set of term weights and compared with the Cosine distance. The weight for each term t_i can be produced by simple term frequency (TF), as given by:

$$tf = \frac{n_i}{\sum_k n_k} \tag{1}$$

where n_i is the number of occurences of each term, or term frequency - inverse document frequency (TF/IDF), as given by:

$$idf = \log \frac{\mid D \mid}{\mid (d_i \supset t_i) \mid} \tag{2}$$

$$tfidf = tf \cdot idf \tag{3}$$

where $\mid D \mid$ is the number of documents in the collection and $(d_i \supset t_i)$ is the number of documents containing term t_i. (Readers unfamiliar with vector based text retrieval methods should see [3] for an explanation of these terms). In our system the 'terms' are the leaf node identifiers and the 'documents' are the songs in the database. Once the weights vector for each document has been extracted, the degree of similarity of two documents can be estimated with the Cosine distance.

5. COMPARING MUSIC SIMILARITY FUNCTIONS

5.1 Data set and classification model

The experiments in this paper were performed on 4911 mp3 files from the Magnatune collection [5], into 13 genres, 210 artists and 377 albums. 1379 of the files were used to train the classification model and the remaining 3532 files were used to evaluate performance and the same model was used in each experiment. To avoid over-fitting in the results, *no artist that appeared in the training set was used in the test set*. The final CART-tree used in these experiments had 7826 leaf nodes.

5.2 Objective statistics

The difficulty of evaluating music similarity measures is well-known in the Music Information Retrieval community [12]. Several authors have presented results based on statistics of the number of examples bearing the same label (genre, artist or album) amongst the N most similar examples to each song (*neighbourhood clustering* [10]) or on the distance between examples bearing the same labels, normalised by the distance between all examples (*label distances* [15]). It is also possible to evaluate *hierarchical organisation* by taking the ratio of artist label distances to genre label distances: the smaller the value of this ratio, the tighter the clustering of artists within genres. Finally, the degree to which the distance space produced is affected by hubs (tracks that are similar to many other tracks) and orphans (tracks that are never similar to other tracks) has been examined [11].

Unfortunately, there are conflicting views on whether these statistics give any real indication of the performance of a similarity metric, although Pampalk [11] reports a correlation between this objective evaluation and subjective human evaluation. Subjective evaluation of functions which maximise performance on these statistics, on applications such as playlist generation, shows that their performance can, at times, be very poor. MIREX 2006 [16] will host the first large scale human evaluation of audio music similarity techniques, and may help us to identify whether the ranking of retrieval techniques based on these statistics is indicative of their performance. In this work, we report the results of all three of the metrics described above, to demonstrate the difference in behaviour of the approaches, but we reserve judgement on whether these results indicate that a particular approach outperforms the other. To avoid over-optimistic estimates of these statistics, self-retrieval of the query song was ignored.

The results, as shown in table 1, indicate that the neighbourhood around each query in the transcription similarity spaces is far more relevant than in the space produced by the likelihood models. However, the overall distance between examples in the transcription-based models is much greater, perhaps indicating that it will be easier to organise a music collection with the likelihoods-based model. We believe that the difference in these statistics also indicates that the transcription-based model produces a much more detailed (micro-similarity) function than the rather general or cultural (macro-similarity) function produced by the likelihood model, i.e. in the transcription system, similar examples are very spectrally similar, containing near identical vocal or instrumentation patterns.

Our own subjective evaluation of both systems shows that they give very good performance when applied to music search, virtually never returning an irrelevant song (a 'clanger') in the top ranked results. This property may be the result of the low number of hubs and orphans produced by these metrics; at 10 results, 9.4% of tracks were never similar and the worst hub appeared in 1.6% of result lists in the transcription-based system, while only 1.5% of tracks

Table 1: Objective statistics of similarirty scores

	Evaluation Metric				
Model		Likelihood - Euc	Likelihood - Cos	Trans - TF	Trans - TF/IDF
	Album % in top 5	17.68%	17.99%	29.58%	29.74%
	Artist % in top 5	25.42%	25.76%	37.45%	37.70%
	Genre % in top 5	61.33%	61.58%	64.01%	64.05%
	Album % in top 10	14.81%	15.24%	24.79%	24.98%
	Artist % in top 10	22.17%	22.66%	32.83%	33.13%
	Genre % in top 10	60.22%	60.38%	62.29%	62.39%
	Album % in top 20	11.80%	12.02%	18.96%	19.04%
	Artist % in top 20	18.68%	18.98%	27.08%	27.36%
	Genre % in top 20	58.97%	59.06%	60.52%	60.60%
	Album % in top 50	7.90%	8.04%	11.41%	11.50%
	Artist % in top 50	13.71%	13.88%	18.65%	18.79%
	Genre % in top 50	57.08%	57.21%	57.61%	57.51%
	Avg. Album distance	0.3925	0.1693	0.6268	0.6438
	Avg. Artist distance	0.4632	0.2509	0.6959	0.7145
	Avg. Genre distance	0.6367	0.3671	0.9109	0.9217
	Artist/Genre Ratio	0.7275	0.6833	0.7640	0.7753
	% never similar, 5 results	4.50%	4.41%	9.75%	9.36%
	% never similar, 10 results	1.59%	1.51%	3.57%	2.61%
	% never similar, 20 results	0.48%	0.51%	0.81%	0.04%
	% never similar, 50 results	0.20%	0.14%	0.00%	0.02%
	Max # times similar, 5 results	19	19	27	29
	Max # times similar, 10 results	29	30	65	58
	Max # times similar, 20 results	52	53	98	106
	Max # times similar, 50 results	125	130	205	216

were never similar and the worst hub appeared in 0.85% of result lists in the likelihood-based system. These results compare favourably with those reported by Pampalk [11], where, at 10 results, the system designated G1 found 11.6% of tracks to be never similar and the worst hub appeared in 10.6% of result lists, and the system designated G1C found only 7.3% of tracks to be never similar and the worst hub appeared in 3.4% of result lists. This represents a significant improvement over the calculation of a simple distance metric in the raw feature space and we believe that whilst more descriptive features features may reduce the effect and number of hubs on small databases, it is likely that they will reappear in larger tests. Similar problems may occur in granular model-based spaces, making the model type and settings an important parameter for optimisation.

5.3 Visualization

Another useful method of subjectively evaluating the performance of a music similarity metric is through visualization. Figures 5 and 6 show plots of the similarity spaces (produced using a multi-dimensional scaling algorithm [6] to project the space into a lower number of dimensions) produced by the likelihood profile-based model and the TF-based transcription model respectively.

These plots highlight the differences between the similarity functions produced by our two approaches. The likelihood profile-based system produces a very useful global organisation, while the transcription-based system produces a much less useful plot. The circular shape of the transcription visualisations may be caused by the fact that the similarities tend asymptotically to zero much sooner than the likelihood-based model similarities and, as Buja and Swayne point out,

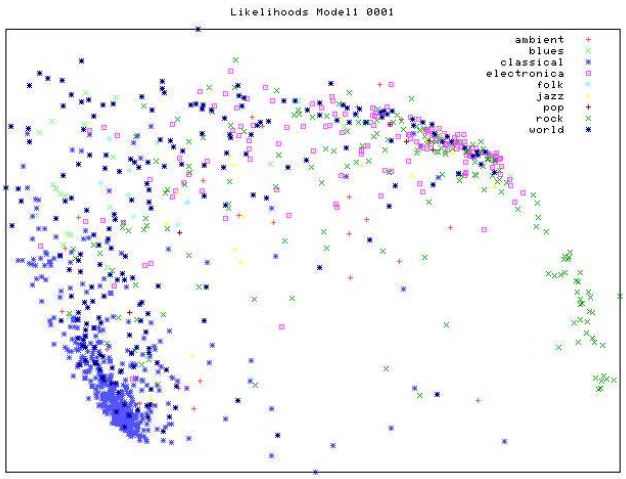

Figure 5: MDS visualization of the similarity space produced by comparison of likelihood profiles.

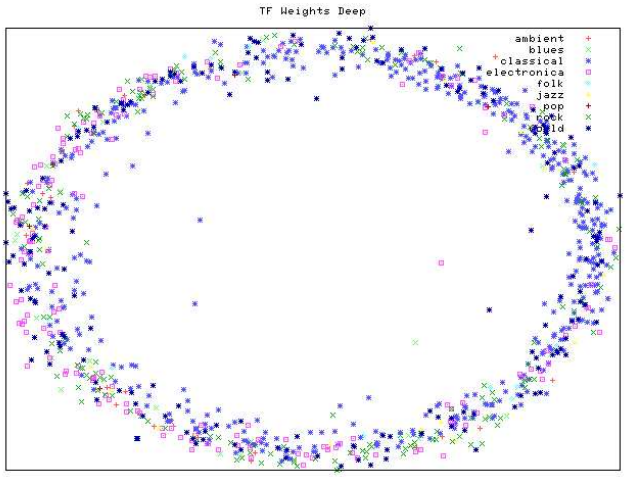

Figure 6: MDS visualization of the similarity space produced by comparison of CART-based transcriptions.

Table 2: Residual stress in MDS visualizations

Plot	Residual stress
Likelihood - Cosine	0.085
Transcription - TF	0.280

'the global shape of MDS configurations is determined by the large dissimilarities' [6]. This perhaps indicates that MDS is not the most suitable technique for visualizing music similarity spaces and a technique that focuses on local similarities may be more appropriate, such as Self-Organising Maps (SOM) or MDS performed over the smallest x distances for each example.

Given sufficient dimensions, multi-dimensional scaling is roughly equivalent to a principal component analysis (PCA) of the space, based on a covariance matrix of the similarity scores. MDS is initialised with a random configuration in a fixed number of dimensions. The degree to which the MDS plot represents the similarity space is estimated by the residual stress, which is used to iteratively refine the projection into the lower dimensional space. The more highly stressed the plot is, the less well it represents the underlying *dissimilarities*. Table 2 shows that the transcription plots are significantly more stressed than the likelihood plot and require a higher number of dimensions to accurately represent the similarity space. This is a further indication that the transcription-based metrics produce more detailed (micro) similarity functions than the broad (macro) similarity functions produced by the likelihood-based models, which tend to group examples based on a similar 'style', analogous to multiple genre descriptions, e.g. instrumental world, is clustered near classical, while the more electronic world music is closer to the electronic cluster.

6. CONCLUSIONS AND FURTHER WORK

We have presented two very different, novel approaches to the construction music similarity functions, which incorporate musical knowledge learnt by a classification model, and produce very different behavior. Owing to this significant difference in behavior, it is very hard to estimate which of these techniques performs better without large scale human evaluation of the type that will be performed at MIREX 2006 [16]. However, the likelihoods-based model is clearly more easily applied to visualization while superior search results are achieved by the transcription[]-based model.

Many conventional music similarity techniques perform their similarity measurements within the original feature space. We believe this is likely to be a sub-optimal approach as there is no evidence that perceptual distances between sounds correspond to distances within the feature space. Distributions of sounds amongst genres or styles of music are culturally defined and should therefore be learnt rather than estimated or reasoned over. Both of the techniques presented enable us to move out of the feature space (used to define and recognise individual sounds) and into new 'perceptually-motivated' spaces in which similarity, between whole songs, can be better estimated. It is not our contention that a timbral similarity metric (a 'micro' similarity function) will produce a perfect 'musical' similarity function (a 'macro' similarity function), as several key features of music are ignored, but that machine-learning is essential in producing 'perceptually' motivated micro-similarity measures and perhaps in merging them into 'perceptually' motivated macro-similarity measures.

Ongoing work is exploring comparison of these techniques with baseline results, the utility of combinations of these techniques and smoothing of the term weights used by the transcription-based approach, by using the structure of the CART-tree to define a proximity score for each pair of leaf nodes/terms. Latent semantic indexing, fuzzy sets, probabilistic retrieval models and the use of N-grams within the transcriptions may also be explored as methods of improving the transcription system. Other methods of visualising similarity spaces and generating playlists should also be explored. The automated learning of merging functions for combining micro-similarity measures into macro music similarity functions is being explored, for both general and 'per user' similarity estimates.

Finally, the classification performance of the transcriptions extracted is being measured, including classification into a different taxonomy from that used to train the original CART-tree. Such a system would enable us to use the very compact and relatively high-level transcriptions to rapidly train classifiers for use in likelihoods-based retrievers, guided by a user's organisation of a music collection into arbitrary groups.

7. REFERENCES

[1] C. Anderson. The long tail. http://www.thelongtail.com, April 2006.

[2] J.-J. Aucouturier and F. Pachet. Music similarity measures: Whats the use? In *Proceedings of ISMIR 2002 Third International Conference on Music Information Retrieval*, 2002 September.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, 1999.

[4] A. Berenzweig, D. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2003.

[5] J. Buckman. Magnatune: Mp3 music and music licensing. http://magnatune.com, April 2006.

[6] A. Buja and D. Swayne. Visualization methodology for multidimensional scaling. Technical report, 2001.

[7] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast feature. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[8] L. Kuncheva. *Combining Pattern Classifiers, Methods and Algorithms*. Wiley-Interscience, 2004.

[9] G. J. Lidstone. Note on the general case of the bayes laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, November 1920.

[10] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, August 2001.

[11] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Johannes Kepler University, Linz, March 2006.

[12] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proceedings of ISMIR 2005 Sixth International Conference on Music Information Retrieval*, September 2005.

[13] E. Pampalk, T. Pohle, and G. Widmer. Dynamic playlist generation based on skipping behaviour. In *Proceedings of ISMIR 2005 Sixth International Conference on Music Information Retrieval*, September 2005.

[14] K. West and S. Cox. Finding an optimal segmentation for audio genre classification. In *Proceedings of ISMIR 2005 Sixth International Conference on Music Information Retrieval*, September 2005.

[15] K. West and P. Lamere. A model-based approach to constructing music similarity functions. *[accepted for publication] EURASIP Journal of Applied Signal Processing*, 2006.

[16] K. West, E. Pampalk, and P. Lamere. Mirex 2006 - audio music similarity and retrieval. http://www.music-ir.org/mirex2006/index.php/Audio_Music_Similarity_and_Retrieval, April 2006.