

# Lightweight measures for timbral similarity of musical audio

## ABSTRACT

Timbral similarity measures based on Mel-Frequency Cepstral Coefficients have been widely reported as the basis for a possible general music similarity function, which would have wide application to searching, browsing and recommendation. Many of the reported methods, however, have computational requirements that make them impractical for searching realistic collections using current hardware. We compare lightweight measures that appear to perform equally well, and introduce a simplification that reduces memory requirements and execution time by a further order of magnitude. This yields a similarity measure that will scale easily to large commercial collections. We give comparative results over two contrasting music collections, one of which has been widely studied, allowing direct comparison with previous work.

## 1. INTRODUCTION

With the wide-scale acceptance of high-capacity personal MP3 players as the music systems of choice, and of download as the mechanism for obtaining new music, new demands and possibilities have emerged for browsing, searching and recommendation from ever larger collections of audio tracks. One particular area of interest has been the development of a music similarity function, capable of capturing directly from audio data the sort of relationships that we express (in so far as words can express them) when we talk of two tracks sharing some particular style, or simply “sounding like one another”.

Initial results exposed in [8, 13] were encouraging, suggesting that an effective music similarity function could be developed, based primarily on timbral features, in particular the Mel-Frequency Cepstral Coefficients (MFCCs) already widely used in the analysis of speech signals. Methods of creating mixture models of MFCCs for a given track, and of comparing the resulting models, were refined in [2, 3] and through the ongoing competitions organised by ISMIR [5, 6]. Recent work such as [1, 19] highlighted serious shortcomings,

however, both in the kind of similarity actually captured by this approach, and in its practical applicability to music collections of realistic size. Although the relationship of a mathematical content-based similarity function to human perception and traditional musical categories remains very much an open research issue [4, 1, 20], simpler methods for comparing MFCCs introduced in [14, 23, 10] appear to work as well or better than the previous heavyweight approaches. Meanwhile commercial search applications apparently based at least in part on timbral similarity, such as MusicSurfer<sup>1</sup> and MusicIP<sup>2</sup>, are already being deployed, although few details of their algorithms, and no useful comparative evaluation of their performance, have been published.

The practical usefulness of any content-based similarity measure, particularly as opposed to increasingly popular social recommenders such as last.fm<sup>3</sup> or MusicStrands<sup>4</sup>, will nonetheless depend on the success of real-world applications rather than laboratory experiments. We therefore expect that low computational cost, in particular fast comparison time to enable searching and browsing large collections, will be an essential attribute of an effective similarity measure. Low memory requirements are also likely to be important, both for online searching of large commercial collections and also to enable embedding search functionality into small portable devices. Just as with fingerprinting today, a compact feature representation will also be important in the design of content-based search interfaces intended to be queried by remote clients. Last but not least, the cost of content-based similarity will have to be low for applications where timbral distance is just one source of information to be combined with textual or community metadata

In this paper we compare a number of lightweight measures for timbral similarity, and show how existing methods can be further simplified to give a 10-fold improvement in both memory requirements and execution time. The organisation of this paper is as follows. Section 2 describes various methods of measuring timbral similarity between music tracks. Section 3 gives an evaluation based on a standard classification task over two different music collections, and Section 4 is a discussion of the results. Section 5 describes a practical search application implementing the lightweight similarity measures evaluated here, and Section 7 outlines future work.

---

<sup>1</sup><http://musicsurfer.iaa.upf.edu>

<sup>2</sup><http://musicip.com>

<sup>3</sup><http://www.last.fm>

<sup>4</sup><http://www.musicstrands.com>

## 2. MEASURING TIMBRAL SIMILARITY

### 2.1 Features for timbral similarity

Throughout this paper we use the first 20 MFCCs to summarise the overall timbre of polyphonic audio. Although other features have been suggested, such as a range of statistics of MFCCs and other well-known features in [16], and temporal envelopes of a gammatone filterbank in [15], we restrict ourselves to MFCCs here for two reasons. Firstly, they are well studied in the literature, allowing comparison with published results, and, secondly, because our focus is primarily on the similarity measure used to define distances between timbral features, rather than the features themselves.

The following subsections describe various distance measures which are then evaluated in Section 3.

### 2.2 Mixture modelling and model comparison

In this widely-reported approach, MFCCs from each track are modelled as a mixture of Gaussians, either by training a conventional Gaussian Mixture Model [2] or by clustering MFCCs with k-means clustering and then fitting a Gaussian to each cluster [13]. The distance between two tracks is then calculated as a symmetrised Kullback-Leibler (KL) divergence between their mixture models. No closed form exists for the KL divergence of two mixtures, but various approximate methods have been proposed, including the Earth Mover’s Distance [13] or MonteCarlo sampling [2]. Unfortunately these approximations are very expensive to compute. Even the coarser approximation described in [17] requires 2ms of CPU time for each pair of tracks on a consumer PC, i.e. around 20s for a single scan of a large personal collection, or many minutes to search a commercial database. Similar figures, using a speed-up method based on precision/cpu-time trade-off, are reported in [21]. Because the distance measure used is not a metric, indexing structures that might otherwise improve search times cannot be applied. These methods must therefore be considered impractical for the time being. We give reported performance statistics for comparison in Section 3.

### 2.3 Global modelling and histogram comparison

A simpler alternative to modelling each track individually is to use a Vector Quantisation (VQ) algorithm to partition the global space of MFCCs into indexed regions, where each region is represented by a single vector in a VQ codebook. The MFCCs for each track are then mapped onto a sequence of VQ codebook indices, and the similarity measure becomes a distance between the index sequences for two tracks. Recent work in [12] suggests that there may be some merit in a measure that reflects the ordering of the indices (in contrast to results in [7] in relation to mixture models), but insufficient information is given from which to judge the practical applicability of such a method to online searching of large collections. Neglecting ordering, we can simply histogram the occurrences of VQ indices over each sequence, and then treat the normalised histogram as a probability distribution. The distance between distributions  $p, q$  is again given by the KL divergence:

$$KL(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$

where the sum is over the codebook indices. A practical implementation is reported in [23], where a Self-Organising Map (SOM) [9] is used to generate the VQ sequences. The SOM is trained on MFCCs from the entire collection. The index sequence for each track is then created as the sequence of best-matching SOM units for each MFCC vector.

A possible issue affecting the applicability of this approach is that the VQ codebook may need to be regenerated if the collection of tracks grows or changes, as asserted, for example, in [13]. We investigate this and give comparative results in Section 3 below. We also compare the KL divergence as used in [23] with a symmetrical alternative, the Jensen-Shannon (JS) divergence:

$$JS(p||q) = KL(p||r) + KL(q||r)$$

where  $r = \frac{p+q}{2}$ .

### 2.4 Single Gaussian modelling and comparison

Despite the fact that MFCCs extracted from music are clearly not Gaussian, [14] showed, somewhat surprisingly, that a similarity function comparing single Gaussians modelling MFCCs for each track can perform as well as mixture models. A great advantage of using single Gaussians is that a simple closed form exists for the KL divergence. For two Gaussians  $p(x) = \mathcal{N}(x; \mu_p, \Sigma_p)$  and  $q(x) = \mathcal{N}(x; \mu_q, \Sigma_q)$

$$2KL(p||q) = \log \frac{\det \Sigma_q}{\det \Sigma_p} + tr(\Sigma_q^{-1} \Sigma_p) + (\mu_q - \mu_p)^T \Sigma_q^{-1} (\mu_q - \mu_p) - d$$

where  $d$  is the dimensionality. A symmetrised distance is then given by

$$2KL_s(p||q) = 2KL(p||q) + 2KL(q||p) \\ = tr(\Sigma_q^{-1} \Sigma_p + \Sigma_p^{-1} \Sigma_q) + (\mu_p - \mu_q)^T (\Sigma_q^{-1} + \Sigma_p^{-1}) (\mu_p - \mu_q) - 2d$$

This distance is already several orders of magnitude faster to compute than one between mixture models. We can introduce a further speed-up by using diagonal covariance matrices, reducing both the number of multiplications required for each distance computation, and eliminating the need for matrix inversion. As the covariance matrix is effectively trying to capture the shape of the (typically non-Gaussian) distribution of MFCCs, one would expect that using only a diagonal covariance would have a dramatic impact on the performance of the distance measure. We might similarly expect that a single Gaussian, particularly with only a diagonal covariance, would be more able to model MFCCs from a short segment of a track, especially one that was likely to be internally consistent, than from a whole track, as suggested in [10]. We test both these hypotheses in Section 3 below.

Means and diagonal covariances of single Gaussians fitted to each song can also simply be concatenated into a single feature vector and compared with Euclidean distance, subject to suitable normalisation. This is the distance measure that has implicitly been used in [22, 16] and other attempts to capture music similarity from a larger feature vector where

means and variances of MFCCs (‘texture windows’) are concatenated with other features. The conventional normalisation is by the variance of each individual feature dimension over the dataset, i.e. a Mahalanobis distance

$$D_M^2(p, q) = (\mu_p - \mu_q)^T \Sigma_\mu^{-1} (\mu_p - \mu_q) + (\Sigma_p - \Sigma_q)^T \Sigma_\Sigma^{-1} (\Sigma_p - \Sigma_q)$$

where  $\Sigma_\mu$  and  $\Sigma_\Sigma$  are the variances of MFCC means and variances respectively over the entire collection of tracks.

As with the VQ approach, the issue arises of whether we need to update the global variances  $\Sigma_\mu$  and  $\Sigma_\Sigma$  as the collection changes (although the cost of doing this will be far less than regenerating the entire VQ codebook). We test this in Section 3.

### 3. CLASSIFICATION EXPERIMENTS

We evaluate these various distance measures with a simple genre classification task over two (in some cases three) different collections. We give classification rates based on a 1-nearest neighbour classifier. Note that these figures are intended only for comparison of the underlying distance measures. One would certainly want to consider a different classifier, such as a Support Vector Machine, to get best results on this task, which is in any event somewhat artificial (in general we have local textual metadata anyway, otherwise a remote fingerprinting interface could be used to look it up). Nonetheless, comparative 1-nn results give a good indication of relative performance on more realistic tasks, they provide a useful sanity check on the underlying measure, and, most importantly, they allow comparison with much previous work.

We also give classification rates using a so-called *artist filter* [19] i.e. classification based on the nearest neighbouring track by a different artist to the query. These figures are of interest both from a practical standpoint (because there are easier ways of finding more tracks by the same artist than content-based similarity), and because they provide evidence of how tracks are clustered in the similarity space defined by each distance measure. We discuss this further in Section 5 below.

In all our experiments MFCCs were extracted from mono audio downsampled to 22050Hz and with a window and hop size of 2048 samples. Results reported from the literature were based on smaller window and hop sizes. We repeated some of our evaluations using smaller window and hop sizes and our results did not change.

#### 3.1 Musical data

We used three music collections in our experiments. In each case, unique genre labels have been assigned to each track. The collections are as follows:

##### 3.1.1 Magnatune (M)

This is a subset of the Magnatune<sup>5</sup> catalogue which is available under the creative commons licence and which has been widely studied, following its release as the training set for the ISMIR 2004 genre classification contest. This collection has a high proportion of classical tracks. The labelled genres

<sup>5</sup><http://www.magnatune.com>

are: classical, electronic, jazz/blues, metal/punk, pop/rock, world. The labels are those given on Magnatune’s own website.

##### 3.1.2 In-House (A)

We selected this collection from a larger in-house collection of some 3600 tracks, with genre labels taken from the Amazon website<sup>6</sup>. We did not inspect individual titles when making our selection, but did choose a set of genres which might be expected to contain characteristic musical content. In some cases we amalgamated similar sub-genres to ensure that each labelled genre contained at least 35 tracks. The final labelled genres are: punk, album rock, jazz, Latin, African, reggae, classical, electronic/techno, R & B.

##### 3.1.3 In-House Larger (A+)

This is a superset of collection A, intended to be more similar to a medium-sized personal collection, and containing several genres that we would not expect to be well-characterised by musical content. To some extent the genre classification task is impossible on this collection, but it serves to highlight some interesting aspects of the similarity space defined by the underlying distance measures. The labelled genres are: punk, indie rock, album rock, classic rock, singer-songwriter pop, jazz, Latin, African, reggae, world, classical, folk, electronic/techno, general dance, R & B, soul, rap.

Statistics of the collections are given in Table 1.

### 3.2 Results

In the tables below we show Leave One Out (LOO) classification rates. We also performed 10 runs of 10-fold cross-validation for each figure reported. In all cases the classification rates were within 1% of the LOO figures, and the cross-validation standard error was between 0.05 and 0.06%.

#### 3.2.1 Impact of audio segment choice

As discussed in Section 2.4, we expect the ability of a single Gaussian to model MFCCs from a whole track to be limited, particularly when using only a diagonal covariance matrix. We therefore compared the performance on the classification task of KL divergence and Mahalanobis distance between single Gaussians with diagonal covariances, using MFCCs drawn from a range of different audio segments. We compared using fixed segments from the centre of each track with whole tracks, and also with ‘thumbnail’ segments. To extract the thumbnails we implemented the automatic method described in [11]. The extracted thumbnail segments were between 7 and 40s in length. By design, we can expect the thumbnails to be internally consistent and to be representative, i.e. unlikely to contain extremes of timbre for each track, such as typical song intro sections. The results are given in Table 2.

Although the raw classification rates drop with the length of the audio segment used, the artist filtered rate for both collections improves as the segment size is reduced to 30s. The raw confusion matrix for collection M also showed a small improvement in classification of rock/pop tracks (which would be most likely to benefit from the removal of intros and outros) using 30s segments or thumbnails, but this was not

<sup>6</sup><http://www.amazon.com>

**Table 1: Music collections used**

collection	genres	artists	tracks	artists/genre		tracks/genre	
				min	max	min	max
Magnatune (M)	6	128	716	5	40	26	312
In-House (A)	9	176	761	7	44	37	210
In-House larger (A+)	17	360	1564	7	55	32	319

supported by the results for collection A. Altogether the results give only limited support to the hypothesis that the distance measures improve when MFCCs are drawn from a segment rather than the whole track.

### 3.2.2 Effect of training set on global measures

To test the significance of the choice of training set for the measures based on VQ index sequences described in Section 2.3, we re-implemented the approach outlined in [23]. The VQ sequences are generated as the sequence of best-matching SOM units for MFCCs for each track, according to a 16 x 16 rectangular SOM (we also tried various alternative topologies but all gave worse classification rates). We trained the SOM on MFCCs extracted from a fixed 30s segment from the centre of each track, where the tracks comprised either the first half or simply the whole of each collection. Results are given in Table 3. In our experiments, collection M was ordered in such a way that the first and second halves both contained tracks from all the albums, while collection A was ordered by genre, so the first half contained tracks only from four genres.

The results do not support the idea that it will be necessary to regenerate the entire VQ codebook as a collection grows. In fact the best classification rates over both collections come from the same training set, suggesting that a single VQ model can be optimised for the global space of MFCCs.

In [23] a simple KL divergence is preferred as the distance measure on psychological grounds favouring asymmetric measures. Table 3 also shows some comparative results for KL and the symmetrical JS divergence given in section 2.3. The classification rates are in fact much higher using the symmetrical distance.

Global variances are required for the Mahalanobis distance described in Section 2.4, and we calculated these from MFCCs extracted from the central 30s of each track. Table 3 gives classification rates for collections M and A using global variances calculated from all of their own and all of each other’s tracks. As with the global VQ measure, the results show little significant benefit from recalculating global variances to fit a particular collection.

### 3.2.3 Choice of similarity measure

Table 4 compares classification rates using the various similarity measures discussed so far. We also include rates for collection M reported in [18] for KL divergence between single Gaussians with full covariance matrices, and for the mix-

ture model approach summarised in Section 2.2 above <sup>7</sup>. For each measure we used the best approaches as described in the preceding subsections.

### 3.2.4 Computational requirements

Table 5 gives figures for the feature size and approximate comparison time for a pair of tracks for each distance measure. The comparison times for the lightweight measures are based on our own un-optimised Java implementations running on a standard consumer laptop and averaged over 500,000 distance computations. The reference comparison time for the GMM measure is given in [19]. When using a full covariance matrix, it may be desirable to precompute the inverse covariance for each track in the collection, in which case the feature length increases to 820.

### 3.2.5 More realistic music collections

Tables 6 and 7 show confusion matrices for collection A and its superset A+, using the best-performing measure, KL divergence with full covariance. Rows correspond to queries, and columns to their nearest neighbours, without artist filter. In contrast to the collections used in most previous experimental work, A+ includes a wide range of genres, but these have not been hand-picked to improve their separability by timbral content.

## 4. DISCUSSION

### 4.1 Applicability of distance measures

The results in Table 4 show that on all our music collections the VQ measure is outperformed by those based on Gaussian models. The remaining lightweight measures all have merits. Within the limits of the widely-observed ‘glass ceiling’ for timbral similarity measures [3], KL divergence and Mahalanobis distance between single Gaussians both perform well on collections M and A. Using diagonal covariances with either measure offers a 10-fold gain in both speed and memory requirements as shown in Table 5, with only a small loss in classification rates compared to using full covariances. The difference in classification rates between the measures is statistically significant, but of course the classification task itself is only a rough and ready guide to their effectiveness for more useful purposes such as similarity search, recommendation, playlist generation, etc. Although KL divergence with diagonal covariance looks to be the best lightweight measure from our experiments, Mahalanobis distance is a straightforward choice if other features are going to be included in the distance measure, as it imposes a natural scaling on the contribution of each feature to the overall distance. Mahalanobis distance also has the advantage

<sup>7</sup>the version of collection M used in [18] contained an additional 13 tracks

**Tables 2-4: Leave One Out genre classification rates (%), af = artist filter over collections M, A, A+**

**Table 2: Varying audio segment**

distance measure	audio segment	M	(af)	A	(af)
KL div (diag cov)	whole track	76	61	-	-
	120s	-	-	72	49
	30s	74	61	70	50
	thumbnail	72	57	61	41
Mahalanobis	whole track	75	59	-	-
	30s	72	61	-	-
	thumbnail	69	57	-	-

**Table 3: Varying training set and VQ distance measure**

distance measure	training set	M	(af)	A	(af)
VQ/JS div	M	64	56	56	43
	M (half)	65	54	-	-
	A	-	-	58	43
	A (half)	72	57	63	47
VQ/KL div	M	59	53	-	-
	A	-	-	48	40
Mahalanobis	M	75	59	69	46
	A	74	59	70	45

VQ distances with SOM trained on MFCCs from whole or first half only of each collection.  
Mahalanobis distances with global variances calculated over each collection.

**Table 4: Varying distance measure**

distance measure	M	(af)	A	(af)	A+	(af)
GMM	79	64	-	-	-	-
KL div (full cov)	78	63	75	48	62	34
KL div (diag cov)	76	61	72	49	52	28
Mahalanobis	75	59	70	45	50	25
VQ/JS div	72	57	63	47	38	23

**Table 5: Space and time requirements for various similarity measures**

distance measure	length of feature vector	comparison time (ms)
GMM	1230	2.000
KL div (full cov)	420	0.050
KL div (diag cov)	40	0.003
Mahalanobis	40	0.001
VQ/JS div	256	0.150

Table 6: Confusion matrix for collection A

	punk	album rock	African	jazz	Latin	reggae	electro	R & B	classical
pnk	115	1	8	9	7	1	1	6	0
aor	9	52	2	0	2	2	0	5	1
afr	2	1	30	0	0	1	0	2	1
jzz	2	1	5	30	1	0	0	6	1
lat	2	1	1	1	25	1	1	3	2
reg	8	1	2	2	2	68	0	13	2
elc	8	3	2	5	2	1	16	5	3
r&b	2	4	7	7	3	1	0	41	2
cls	0	1	2	3	1	3	0	4	196

Table 7: Confusion matrix for collection A+

	punk	indie rock	album rock	classic rock	pop	jazz	Latin	Afric	reggae	world	class- ical	folk	elect	dance	R&B	soul	rap
pnk	110	5	1	0	2	8	6	7	1	0	0	0	1	0	4	2	1
ind	29	165	8	16	6	9	2	5	5	2	5	17	6	0	18	22	4
aor	8	2	52	1	0	0	1	2	2	0	1	0	0	0	4	0	0
clr	10	10	1	79	1	0	0	1	2	0	2	5	0	0	8	5	0
pop	1	5	1	4	21	0	1	0	0	0	1	0	0	0	3	3	1
jzz	2	0	1	0	0	29	1	5	0	0	1	1	0	0	5	1	0
lat	2	0	1	0	0	1	24	0	1	0	2	1	1	0	3	1	0
afr	2	1	1	0	0	0	0	29	1	0	1	0	0	0	1	1	0
reg	8	0	1	1	1	2	2	2	65	2	2	0	0	0	10	1	1
wld	0	2	0	0	1	1	1	0	0	20	1	0	0	0	3	3	0
cls	0	1	1	1	0	2	1	1	1	1	196	0	0	0	2	2	1
flk	3	8	1	3	0	0	2	2	0	0	1	42	0	0	4	3	1
elc	6	2	2	2	1	4	2	2	1	0	2	0	15	0	4	1	1
dan	3	5	1	4	0	3	2	3	0	2	0	2	0	1	6	11	2
r&b	2	2	3	2	0	6	3	5	1	0	1	0	0	0	37	5	1
sol	3	8	2	7	0	3	7	2	3	1	1	1	0	0	13	77	2
rap	2	3	0	3	0	0	0	3	1	1	0	0	0	2	7	4	8

of being a metric, enabling yet further potential speed-up through the use of suitable indexing structures.

In section 5 we describe a search application which we have implemented to allow direct comparison of these and other music similarity measures in a realistic context. This will enable us to use practical user tests as our main evaluation tool in future work.

## 4.2 Hubs

The phenomenon of *hubs*, tracks that, according to a particular distance measure, are close to a disproportionately large number of others, was identified in [1], in relation to distances based on GMMs. As many as 15% of the tracks in a collection of some 360 were found to occur amongst the 100 nearest neighbours of 200 or more other tracks. Using a similar criterion, we found far fewer hubs with our lightweight measures. With KL divergence using diagonal covariance, for example, we identified around 2% of collection M and 6% of collection A as hubs. Detailed studies in [1] relate hubs to Gaussian mixture components that are highly non-discriminatory with respect to genre, despite being statistically significant, i.e. accounting for large numbers of audio frames within a given track. Using single Gaussians, we therefore speculate that much of the discriminatory information expressed by MFCCs is captured in the extrema of the covariance matrices. This would explain why, contrary to our expectations, modelling whole tracks gives better classification rates in general than modelling single consistent segments. It would also caution against the practice of removing outliers before modelling, as recommended in [16].

## 4.3 The nature of timbral similarity

Collections M and A, as with most used in previous studies, have been chosen so that the labelled genres are reasonably well-separated by timbre. This is a reasonable prerequisite for the use of genre classification rates as an evaluation measure. On the other hand, for application purposes we are equally - if not more - interested in the performance of our similarity measures on collections such as A+. The classification task is clearly very artificial for A+, as we would not expect genres such as indie rock, album rock and classic rock, or R & B and soul, to be clearly separated by timbre. The confusion matrix shown in Table 7 suggests, however, that the problem is worse than this, as tracks of almost any genre can be confused with one or other kind of rock music. This implies that, according to a timbral distance measure, rock music occurs throughout the similarity space. If A+ is typical of real collections, then this observation has important consequences for the practical application of timbral similarity. The very low classification rates with artist filter on this collection, comparable to those observed in [19], are also striking, even allowing for the natural fuzziness of genre labels. These suggest that clustering in the similarity space consists principally in small clusters corresponding to artists, or even albums.

We can imagine a continuum of existing and as yet hypothetical music search methods, from fingerprinting to find exact audio matches at one extreme, to an idealised ‘sounds something like this one’ recommendation-search at the other. These observations suggest that timbral similarity is closer to fingerprinting and further from recommendation-search

than we would ideally like for many applications. In real applications, of course, we can constrain similarity search by using textual metadata that already includes genre labels, publication date, etc. The value of timbral similarity in this context remains an open research question, although one that may be answered in the open marketplace more readily than in the laboratory.

## 5. SOUNDBITE: A CONTENT-BASED MUSIC SEARCH APPLICATION

In order to let users experience and evaluate a variety of similarity measures in a realistic application context, we have implemented a content-based music search engine. The system also incorporates automatic segmentation and thumbnailing, enabling easier and faster user evaluation of search results by presenting a short representative thumbnail audio clip for each track returned. SoundBite manages a database of tracks, allowing the entry of simple metadata (album title, trackname, artist, genre, etc.) as each new track is added, while automatically extracting and saving segmentation information and a thumbnail segment, as well as MFCCs and other features required to support various similarity measures. Tracks in the database can be browsed and searches can be made by similarity to a chosen track: in all cases the results are presented as a search engine-style list, with thumbnail audio immediately available for playback for each track in the list. The overall system architecture is shown in Figure 1 and a screenshot in Figure 2. The software design defines a simple interface to be implemented for each similarity measure. The measure then becomes available to the user in a pull-down menu. Different similarity measures can be selected at will for each search, and searches can also be constrained by textual metadata, for example to search only within the labelled genre of the query track.

## 6. CONCLUSIONS

We evaluated various lightweight timbral similarity measures based on MFCCs according to their performance on a simple classification task over two different music collections. Measures based on single Gaussians fitted to MFCCs performed significantly better than one based on VQ sequences. We found that diagonal covariances could be used for a distance based on KL divergence with little loss in classification rates and a 10-fold gain in memory requirements and computation time. In a practical search application implementing these and other similarity measures, this method can search a collection of 50,000 tracks in under 4ms on a standard consumer laptop, and with a memory footprint for the features of only 8MB.

## 7. REFERENCES

- [1] J.-J. Aucouturier. *Ten experiments on the modelling of polyphonic timbre*. PhD thesis, University of Paris 6, 2006.
- [2] J.-J. Aucouturier and F. Pachet. Music similarity measures: What’s the use? In *Proc. ISMIR*, 2002.
- [3] J.-J. Aucouturier and F. Pachet. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 2004.

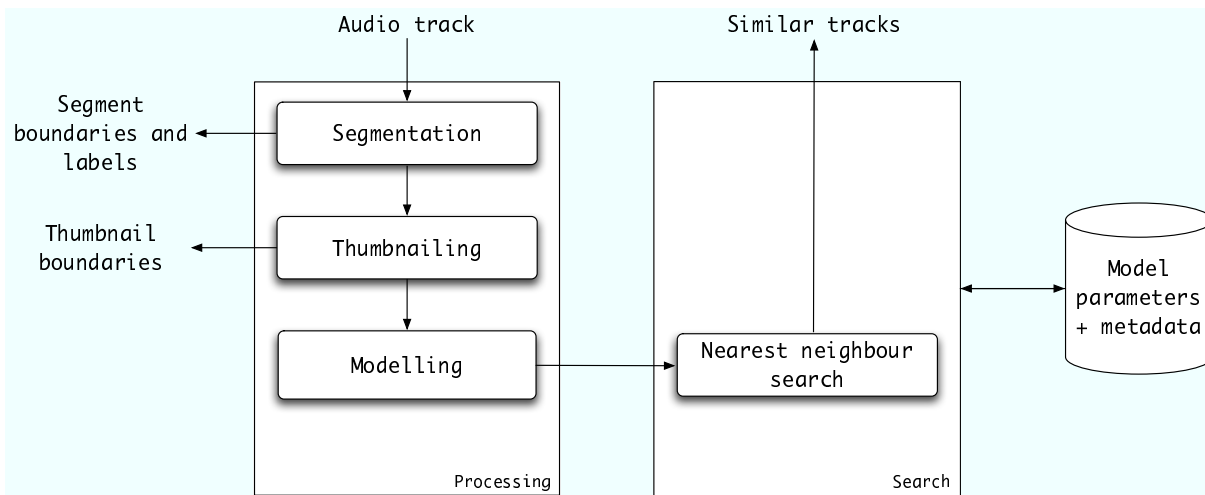


Figure 1: SoundBite framework

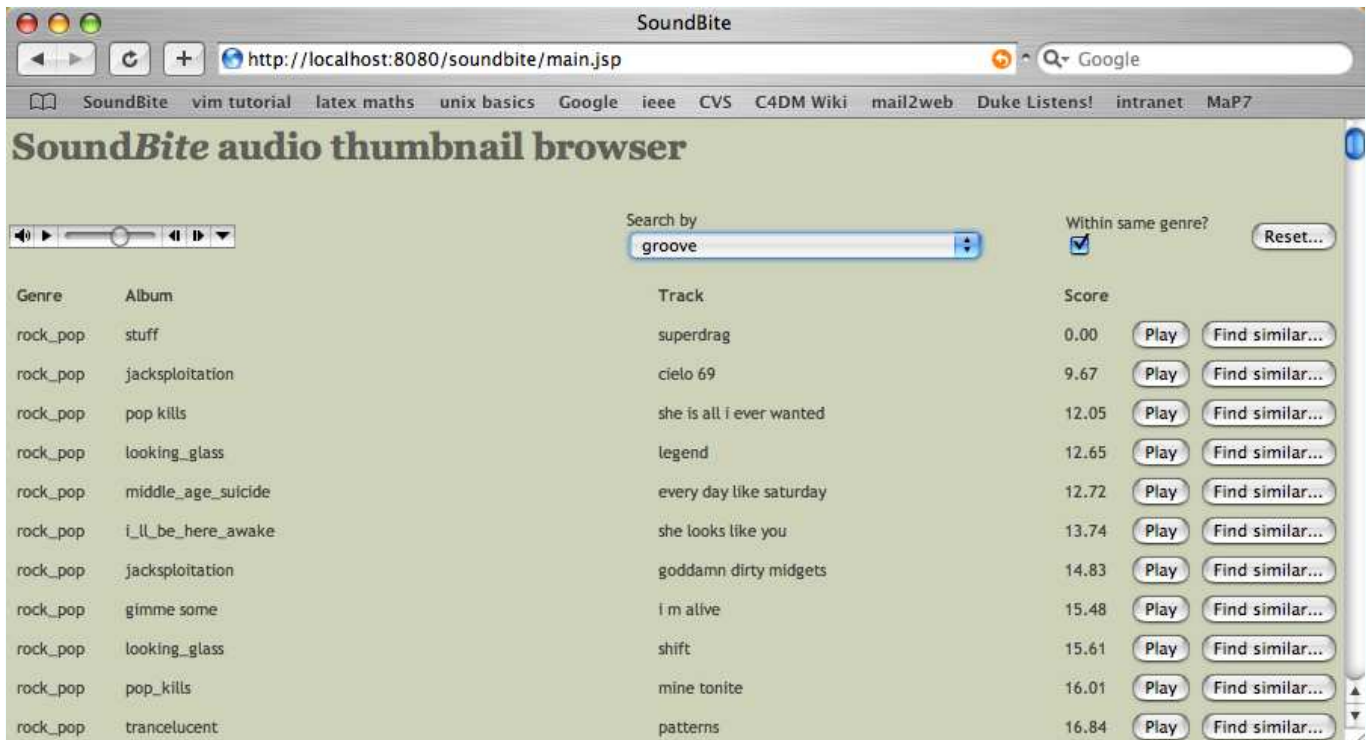


Figure 2: SoundBite screenshot.

- [4] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proc. ISMIR*, 2003.
- [5] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack. Ismir 2004 audio description contest. Technical report, Music Technology Group, Universitat Pompeu Fabra, 2006.
- [6] J. S. Downie, K. West, A. Ehmann, and E. Vincent. The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. In *Proc. ISMIR*, 2005.
- [7] A. Flexer, E. Pampalk, and G. Widmer. Hidden Markov models for spectral similarity of songs. In *Proceedings of the 8th International Conference on Digital Audio Effects*, 2005.
- [8] J. Foote. Content-based retrieval of music and audio. In *Proc. SPIE Multimedia Storage and Archiving Systems*, 1997.
- [9] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1995.
- [10] M. Levy and M. Sandler. Application of segmentation and thumbnailing to music browsing and searching. In *Proc. 120th AES Convention*, 2006.
- [11] M. Levy, M. Sandler, and M. Casey. Extraction of high-level musical structure from audio data and its application to thumbnail generation. In *Proc. ICASSP*, 2006.
- [12] M. Li and R. Sleep. Genre classification via an lz78-based string kernel. In *Proc. ISMIR*, 2005.
- [13] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proc. ICME*, 2001.
- [14] M. Mandel and D. Ellis. Song-level features and SVMs for music classification. In *Proc. ISMIR*, 2005.
- [15] M. F. McKinney and J. Breebaart. Features for audio and music classification. In *Proc. ISMIR*, 2003.
- [16] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken. Modelling timbre distances with temporal statistics from polyphonic music. *IEEE Trans. Audio, Speech and Language Processing*, 14(1):81–90, 2006.
- [17] E. Pampalk. Speeding up music similarity. In *Proc. ISMIR*, 2005.
- [18] E. Pampalk. *Computational models of music similarity and their application to music information retrieval*. PhD thesis, Vienna University of Technology, 2006.
- [19] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proc. ISMIR*, 2005.
- [20] T. Pohle, E. Pampalk, and G. Widmer. Evaluation of frequently used audio features for classification of music into perceptual categories. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing*, 2005.
- [21] P. Roy, J.-J. Aucouturier, F. Pachet, and A. Beurivé. Exploiting the trade-off between precision and cpu-time to speed up nearest neighbour search. In *Proc. ISMIR*, 2005.
- [22] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Acoustics, Speech and Signal Processing*, 10(5), July 2002.
- [23] F. Vignoli and S. Pauws. A music retrieval system based on user-driven similarity and its evaluation. In *Proc. ISMIR*, 2005.