EMOTION-BASED MUSIC RETRIEVAL ON A WELL-REDUCED AUDIO FEATURE SPACE

Maria M. Ruxanda[1], *Bee Yong Chua*[1], *Alexandros Nanopoulos*[2], *Christian S. Jensen*[1]

1. Department of Computer Science
Aalborg University, Denmark
{mmr, abychua, csj}@cs.aau.dk

2. Information Systems and Machine Learning Lab
University of Hildesheim, Germany
nanopoulos@ismll.de

ABSTRACT

Music expresses emotion. A number of audio extracted features have influence on the perceived emotional expression of music. These audio features generate a high-dimensional space, on which music similarity retrieval can be performed effectively, with respect to human perception of the music-emotion. However, the real-time systems that retrieve music over large music databases, can achieve order of magnitude performance increase, if applying multidimensional indexing over a dimensionally reduced audio feature space. To meet this performance achievement, in this paper, extensive studies are conducted on a number of dimensionality reduction algorithms, including both classic and novel approaches. The paper clearly envisages which dimensionality reduction techniques on the considered audio feature space, can preserve in average the accuracy of the emotion-based music retrieval.

Index Terms— audio features, dimensionality reduction, content-based music retrieval

1. INTRODUCTION

Music expresses emotion [1]. An effective emotion-based music retrieval depends on audio extracted features, that can capture the perceived emotional expression of music. Various empirical studies in music psychology investigated the influence of different perceptual music features on the perceived music-emotion. A recent survey [2] envisaged that the eight emotional expression categories proposed by Hevner [3] – *serene, dreamy, sad, solemn, vigorous, exciting, happy, and playful*, are the most appropriate and relevant for 21st century listeners. As well, it was found that the main audio factors apprehended by listeners to have influence on the perceived music-emotion are: tempo, articulation, rhythm motion, pitch, harmony, and loudness. For instance, happy musical pieces usually have fast-perceived tempo, less firm rhythm motion and staccato articulation, and sad musical pieces have slow-perceived tempo, firm motion and legato articulation[1].

The recent research of Chua [4, 5] further built on the above findings and proposed methodologies for the classification of polyphonic music signals, based on the notions of perceived music and music-emotion. By employing various signal processing algorithms, Chua automatically extracted 17 audio features, that relate to the main audio factors mentioned above. Through extensive experimental studies, this work [4, 5] showed that the 17 audio features are efficient for classifying the eight emotional categories of Hevner [3].

Similar work [6, 7] addressed the detection of emotion in acoustic music. Liu et al. [6] extracted intensity, rhythm and timbre features, and projected them into an emotional space of mood and arousal. Li and Ogihara [7] used timbral texture, rhythmic and pitch features, to classify music into an emotional space originating from Hevner's emotional categories.

However, the related work does not address the problem of emotion-based music similarity retrieval in the setup of large music datasets. The explosion of music available on WWW, PCs and MP3 players, calls for efficient methods for music similarity retrieval. The classical way of approaching the retrieval problem "find the *k*-nearest neighbors of the query song", is to perform similarity search by considering the distance between the query song and other songs in a music collection, over the set of features extracted from music.

Various audio features were extracted from music, in order to capture the emotional expression of music [4, 5, 6, 7]. Among these, the 17 audio features of Chua[4] capture effectively both the factors that influence the notion of perceived music [1], and the music-emotion psychological aspects of Hevner [3]. However, the similarity retrieval in such a high-dimensional space is challenging for real-time systems that manipulate large music databases. The multidimensional indexes such as the R-tree[8] and variants, that are implemented in database management systems (Oracle, PostgreSQL, Informix), speed up the retrieval but work well up to ≈10-dim.

Study Directions. These ideas constitute the motivation of the work in this paper. Reducing the audio feature space [4] to a lower dimensionality, allows the efficient use of indexing techniques such as the R-tree, that support k-nearest neighbor (*k*-NN) queries. However, an inadequate dimensionality reduction may negatively affect the retrieval accuracy.

As point of departure, we hypothesize that a feasible dimensionality reduction in our case, would be one that *discards the dimensions along which the data is statistically*

This research was supported by the Danish Research Council for Technology and Production Sciences project no. 26-04-0092 Intelligent Sound.

varying the most. The intuition is to eliminate the undesired property of a high-dimensional space — *distances between songs can be negatively affected by only few dimensions with high dissimilarity and partial similarities remain uncovered*.

To validate our hypothesis, we investigated on several dimensionality reduction techniques and on their performance measured through the *k*-NN retrieval accuracy. The latter was evaluated relative to music-emotion labels, collected from human annotations. We performed thorough studies on two music datasets and we evaluated the retrieval accuracy: "objectively" on a small dataset (1,000 songs) fully and carefully annotated by musicians, and "subjectively" on a big dataset (41,446 songs) by measuring users'satisfaction on a random sample annotated by casual online users.

Contributions. The paper presents extensive studies on dimensionality reduction on the audio feature space, that captures the perceptual emotion of music. A large variety of dimensionality reduction algorithms is investigated, including both classic and novel approaches ([11, 12]). The obtained results reveal that dimensionality reduction can better preserve the average *k*-NN retrieval accuracy of the full audio space, if it eliminates the dimensions along which the data is statistically varying the most. These results are highly relevant for real-time systems, that can achieve order of magnitude increase in retrieval performance, by applying available multidimensional indexes on a well-reduced audio feature space.

The rest of the paper is organized as follows. Section 2 introduces aspects and techniques further employed in the paper. Section 3 elaborates on the dimensionality reduction of the music-emotion feature space. Section 4 concludes the paper.

2. PRELIMINARIES

Section 2.1 introduces the audio features we extract, in order to project the songs in a music-emotion feature space. Section 2.2 reviews various dimensionality reduction techniques.

2.1. Emotion-based Audio Features

By applying signal processing algorithms, Chua[4] extracted 17 audio features from polyphonic music, that can capture the notion of perceived music-emotion, quantified relative to important music psychology and psychoacoustic research findings[1]. Further, these audio features are grouped in six sets, each relating to one of the main factors (tempo, articulation, rhythm motion, loudness, pitch and harmony) that have influence on the perceived emotional expression of music [1].

The emotion-based audio features [4] are: the *perceptual tempo* (capturing the perceived fastness or slowness) and the *perceptual tempo variation*; the *articulation* (capturing the perceived staccato or legato) and the *articulation variation*; the *rhythm motion* (capturing the perceived firm or flowing) and the *rhythm motion variation*; the *loudness variation* (representing the variation of the perceived signal's intensity); the *spectral flux* (representing the variation of spectrum's energy between successive time frames) and the *spectral flux variation*; the *low pitch* (capturing the relative perceived intensity of low frequencies) and the *low pitch variation*; the *pitch density* (capturing the relative number of audible frequency sub-bands) and the *pitch density variation*; the *harmonicity* (representing the multiple integers of dominant frequencies) and the *harmonicity variation*; the *roughness* (representing the small frequency differences) and the *roughness variation*.

We shall further use these audio features as basis to perform emotion-based similarity retrieval of music.

2.2. Dimensionality Reduction on the Feature Space

There exist two main categories of methods that perform dimensionality reduction: (a) by selecting some features among the entire set of features, and (b) by transforming the original space into a new feature space of lower dimensionality.

From the first category, we shall investigate on:
1) the "maximum likelihood common factor analysis" [9], that returns the maximum likelihood estimates of the specific variances along each dimension, combined with the ranking of the dimensions in ascending order of the variance. We denote this method as "FA";
2) the method that ranks the dimensions by individual evaluation, denoted as "Ranker". We use infoGain evaluator, that evaluates the worth of a single feature (dimension) by measuring the information gain with respect to the class: $InfoGain(Class, Feature) = H(Class) - H(Class|feature)$.
3) the genetic-search method that implements the genetic algorithm of Goldberg [10], used in conjunction with the evaluator method. The latter evaluates the worth of a subset of dimensions, by considering the individual predictive ability of each data dimension along with the degree of redundancy between them. We denote this as "GA";
4) the K-N-Match algorithm [11], denoted as"K-N-Match". The algorithm, proposed by recent research in the database field, returns the first *k* objects that are closest to the query object in any *n* dimensions, *n* is the reduced dimensionality.

In the second category, belong the widely-applied Principal Component Analysis (PCA), and pivots-based techniques. The latter map the original space into a lower-dimensional space by using a number of chosen objects, called "pivots". An object in the reduced space is then represented by its distances to the pivots. We shall also investigate on:
5) PCA, henceforth denoted as "PCA" and
6) the pivot-based algorithm [12] – the variant "local optimum selection of pivots". This algorithm selects good pivots based on an efficiency criterion. We denote this as "Pivot-Sel.".

3. RETRIEVAL ACCURACY ON A REDUCED SPACE OF EMOTION-BASED AUDIO FEATURES

This section investigates on the performance of the dimensionality reduction techniques on the emotion-based audio feature space[4]. The performance is evaluated by measuring the *k*-NN retrieval accuracy relative to the perceived music-emotion by listeners, and is done from two perspectives – the objective evaluation is necessary as a possible "*ground-truth*" measure, the subjective evaluation is a "*must*" due to the inherent human-subjective notion of perceived emotion.

3.1. Objective Evaluation of the Retrieval Accuracy

We used a dataset of 1,000 songs downloaded from two music websites: *www.songpeddler.com* and *www.allmusic.com*. The music style varies and includes rock, metal, jazz, electronic, pop, classical, hip-hop,r&b, country, folk, and ambient music. Each song was *carefully* labeled by musicians with emotional terms, consisting of adjectives that were either directly found in Hevner's proposed categories (see Fig. 1) or were synonyms (based on the thesaurus in Microsoft Word 2002) with Hevner's emotional terms. Finally, each song was labeled in a simplified manner with one number from 1 to 8, representing the eight emotional categories of Hevner.

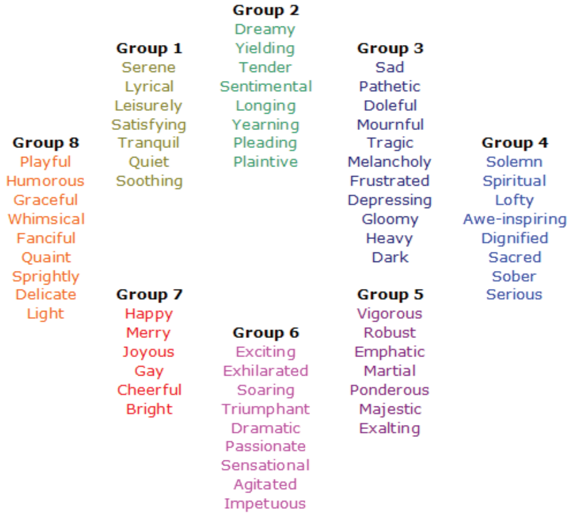

Fig. 1. Hevner's emotional expression categories of music.

We tested the dimensionality reduction methods introduced in Section 2.2, and as baseline the full-match approach (all 17-dim). We implemented the K-N-Match and Pivot-Sel. algorithms in Matlab. We used the implementations provided by the Matlab environment for FA and PCA, while for Ranker and GA we used the implementations of the Weka tool [1].

In all experiments, we measured the average difference in emotion labels between a seed song and its k-NN retrieved songs. The k-NN songs were retrieved by applying the Euclidean distance over the considered feature space. Due to the circular arrangement of the eight emotional categories, the maximum difference between the emotion labels of any two songs is 4. Thus, we computed the average difference in emotion labels between a seed song and its k-NN songs as:

$$Avg_{diff} = \frac{1}{k}\sum_{i=1}^{k}[4 - abs(abs(label_{seed} - label_{i}) - 4)]$$

We queried with each of the 1,000 songs, and we averaged over all songs the values obtained by the above formula. We shall refer to this measurement as "*the average difference of emotion labels*" for k-NN queries. The smaller this value is, the better the dimensionality reduction technique captures the notion of emotion-based similarity of music.

We first investigated which is the optimal number of dimensions in the reduced space for the various methods – see Fig. 2a. (note: GA is fixed to a 6-dim reduced space). The figure reports for 5-NN queries, but different values of k were tried. The optimal reduced dimensionality is preserved over varying k, and is for each method: FA 11-dim, Ranker 6-dim, GA 6-dim, K-N-Match 13-dim, PCA 10-dim, Piv-Sel 10-dim.

Secondly, we investigated the performance of the various methods for different values of k (used in k-NN), while considering their optimal reduced dimensionality – see Fig. 2b. Ranker (6-dim), GA (6-dim) and FA (11-dim) are clearly the best performing and they appear not only to preserve the average retrieval accuracy of the full audio space but also slightly improve it, for the 1,000 songs dataset.

The results in Fig. 2 empirically validate our launched hypothesis, and show that: 1) the most effective dimensionality reduction techniques are those that focus on eliminating those dimensions with a high data-variance; and 2) it is possible to apply dimensionality reduction on a high-dimensional space of emotion-based audio features, while in average preserving the quality of the k-NN similarity retrieval.

Among the best performing methods, Ranker and GA exhibit similar performance, but Ranker is slightly better. Moreover, the feature selectivity of Ranker appears well tuned for the emotion-based audio feature space. It selects perceptual tempo, rhythm motion, spectral flux, roughness, articulation, and pitch density, each of these relating to exactly one of the main perceptual factors [1]. FA reduces the space to 11-dim only, but these include the six dimensions selected by Ranker.

3.2. Subjective Evaluation of the Retrieval Accuracy

We further investigated on the behavior of Ranker (6-dim) and FA(11-dim) on a larger music dataset of 41,446 songs [2], that is similar in terms of the spanned musical styles with the 1,000 songs dataset. An "objective" evaluation on this large dataset would require the collection of a tremendous number of human-labeled emotional terms. To tackle this issue, we instead selected a representative random sample, that was annotated with emotional terms through a user survey.

We chose 24 random seed songs, but to represent diversified musical styles. For each seed song, we queried the dataset (applying the Euclidean distance) and we retrieved the 5-NN songs (excluding the seed song) in the three feature spaces: 17-dim, FA (11-dim), and Ranker (6-dim). The random sample thus composed, resulted in a distinct set of 333 songs.

We devised a simple online survey. A user was listening to a song at a time, meanwhile he/she could see the image in Fig. 1. The user was asked to classify the song in the group that better described the perceived emotion expressed by the song. The survey was available for 4 weeks to 105 users from

[1] Open-source software tool: http://www.cs.waikato.ac.nz/ml/weka/

[2] Due to practical reasons concerning the online user survey, we used only 30 seconds of a song.

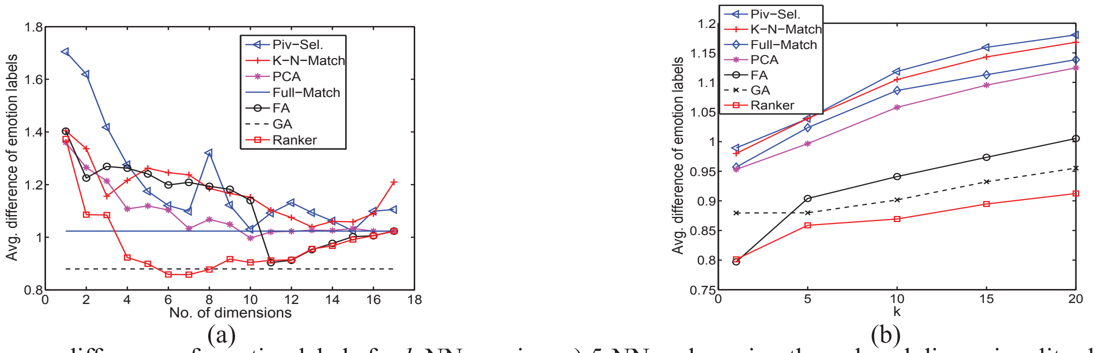

Fig. 2. Average difference of emotion labels for *k*-NN queries: a) 5-NN and varying the reduced dimensionality; b) varying *k* for the optimal reduced dimensionality: Piv-Sel 10-dim,K-N-Match 13-dim,PCA 10-dim,FA 11-dim,GA 6-dim,Ranker 6-dim

8 European countries. To ensure the collection of reliable emotion labels, the survey complied with the conditions: **(a)** a user could rate a song once; **(b)** 5 different users rated each song; **(c)** a song was played repeatedly until classified.

The emotion label of a song was aggregated from the 5 users ratings per song, as the majority rated emotional category (group). We obtained that for ≈60% of songs *at least* 3 out of 5 users rated the same, while only for 1.5% songs we could not obtain a meaningful emotion label (each of the 5 users rated a different emotion). Overall, the obtained emotion labels spanned evenly the eight emotional categories.

We then measured the average difference of emotion labels for 5-NN queries, in each of the three feature spaces. We considered two alternatives: (a) 21 seed songs – by eliminating the seed songs corresponding to the songs without a label; (b) 24 seed songs – by using the maximum possible difference (that is 4) of the unlabeled songs to their seed songs. We performed the standard statistical t-test, checking if the retrieval accuracy values on the full audio space were produced by a distribution with equal mean to the distribution of the retrieval accuracy values on the reduced spaces. The hypothesis was accepted at confidence level 0.05, with 0.8 probability for the case of 21 seed songs and with 0.4 probability for the case of 24 seed songs. We thus judge that FA and Ranker are comparable in average retrieval accuracy with the full-match.

3.3. Indexing a Reduced Space of Audio Features

In the context of similarity retrieval over large music collections, the indexing of a reduced audio feature space can considerably speed up the search performance. As a proof of concept, we indexed 2,280,760 musical pieces (obtained by splitting the 41,446 songs into excerpts of various length) into the R-tree [8]. We used the index implementation provided by the open-source XXL java library, and a machine configuration of 1GB of RAM, 1.86GHz processor and 4KB disk page. The obtained results – see Table 1, show an order of magnitude increase of performance for a 6-dim audio feature space.

	6-dim (Ranker)	**11-dim (FA)**	**17-dim**
5-NN query	161	1495	3235
10-NN query	181	1675	3548

Table 1. No. of disk accesses when using the R-tree index

4. CONCLUSION

The paper introduced a practical approach to emotion-based music retrieval in large music collections. The proposed approach comprises of an effective methodology that projects the music into an audio feature space that captures the music-emotion. This audio feature space is dimensionally reduced while preserving the average retrieval accuracy, so it can be indexed to significantly improve the retrieval performance.

5. REFERENCES

[1] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*, Oxford University Press, 2001.

[2] E. Schubert, “Update of the Hevner Adjective Checklist,” *Perceptual and Motor Skills*, vol. 96, pp. 1117–1122, 2003.

[3] K. Hevner, “Experimental Studies of the Elements of Expression in Music,” *American Journal of Psychology*, vol. 48, pp. 246–268, 1936.

[4] B. Y. Chua, *Automatic Extraction of Perceptual Features and Categorization of Music Emotional Expressions from Polyphonic Music Audio Signals*, Ph.D. thesis, Monash University, Australia, 2007.

[5] B. Y. Chua and L. Guojun, “Perceptual Rhythm Determination of Music Signal for Emotion-based Classification,” in *Proc. of MMM*, 2006.

[6] D. Liu, L. Lu, and H. J. Zhang, “Detecting Emotion in Music,” in *Proc. of ISMIR*, 2003.

[7] T. Li and M. Ogihara, “Automatic Mood Detection from Acoustic Music Data,” in *Proc. of ISMIR*, 2003.

[8] A. Guttman, “Rtrees: Dynamic Index Structure for Spatial Searching,” in *Proc. of SIGMOD*, 1984, pp. 47–57

[9] R. Reymont and K.G. Joreskog, *Applied Factor Analysis in the Natural Sciences*, Cambridge Univ. Press, 1993.

[10] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989

[11] A. K. H. Tung, R. Zhang, N. Koudas, and B. C. Ooi, “Similarity search: A matching Based Approach,” in *Proc. of VLDB*, 2006.

[12] B. Bustos, G. Navarro, and E. Chavez, “Pivot Selection Techniques for Proximity Searching in Metric Spaces,” *Pattern Recognition Letters*, vol 24, pp. 23(57–66), 2003