Mood-based navigation through large collections of musical data

Marta Tolos, Raquel Tato, Thomas Kemp
SONY International (Europe) GmbH
Hedelfinger Str. 61
70327 Stuttgart, Germany

Abstract— **The last years have seen a remarkable reduction of cost and size of mass storage devices, like harddisks and flash memory. There is no technological barrier any longer that prevents a user to have several thousand of different music titles stored on his or her mobile device. This calls for new mechanisms that allow users to explore large musical databases. Additionally, mobile 'mp3' players today have the size of a stripe of chewing gum, which gives rise to a yet another problem: how to select a song from a large collection if there is no keyboard and only very little graphical output available?**

In this paper, a mood-based song selection and exploration scheme is proposed, in which the user selects the desired mood of a song in a two-dimensional mood plane. An algorithm that can automatically determine the mood of a music piece is described. An evaluation on a small music collection of 400 music pieces shows the viability of the new approach.

Keywords: song selection; music mood detection

I. Introduction

The new developments in storage technology have led to a fundamentally new problem for the home music consumer: the problem of *effective song selection*. While our parents picked the vinyl record of their choice from the cupboard, put it on the turntable, and selected the song by moving the needle to the appropriate position, this option is no longer available to their grandchildren. Particularly in the case of mobile devices, extremely compact form factors make such an intuitive approach impossible. A modern portable 'mp3' player has roughly the form factor of a stripe of a chewing gum, which does hardly allow for a display that can show more than one or two text lines at once. Therefore, the classical graphical user interface, as it is known from personal computers, is not a possible solution to operate such a player.

The paper is organized as follows. First, some of the existing approaches for the song selection problem are outlined. Then, mood based song selection is introduced, followed by a description of an automatic mood extraction algorithm. Then, we describe a prototypical mood-selection based music player system and give some preliminary results on the user feedback.

II. Existing approaches for song selection

A. Classical libraries

Classical libraries have been among the first to identify the problem, and have come up with a variety of solutions to it. Most of this solutions, however, are based on classical text searches in manually created metadata. Typically, a library index for a music piece contains information about the composer, the performer, the title, the genre, sometimes lyrics, and - for classical music - references to commonly accepted old library catalogues, like e.g. 'BWV' for Johann Sebastian Bachs works. While this allows the knowledgeable user to accurately pinpoint any given piece of music, due to the lack of a keyboard it is not a good solution for small mobile devices. Additionally, the occasional consumer of music often does not know the specific information that is required to find the piece that he or she desires.

B. Genre based hierarchical systems

From the very beginning of the establishment of audio libraries, there have been elaborate schemes how music can be categorized in musical genres which allow a preselection and thereby a reduction of the search space. Since the major genre categories (like 'Pop', 'Jazz', 'Classical' etc) are very coarse, fine-grained granularity has been introduced (e.g. 'Cool Jazz', 'Acid Jazz') as subclasses inside the major categories, thus allowing for a hierarchical ordering system which can in theory be easily traversed by the user. However, the categorization of music into genre classes is ambiguous even on the top level, which will effectively hide the presence of a music piece from the user if this user chooses another category than the transcriber did when the database was created. If music is assigned multiple labels in order to alleviate this problem, the reasoning behind the introduction of the genre categories - reduction of the search space - is jeopardized. Because of this inherent difficulty, genre-based hierarchical systems have not been able to capture significant market share despite the substantial effort that has been spent into their creation and maintenance (for a categorization scheme, see e.g. [8]).

C. Query by Humming

In many cases, the user of a music database knows the *melody*, or at least some phrases of the melody, of the music piece he would like to hear. An obvious solution to the selection problem is therefore to have the user hum, sing or whistle the melody into a microphone. The system then looks up the melody in its database, and typically returns a sorted list of the closest matches. While this approach has the advantage that no keyboard is required for input, it has some severe weaknesses as well. Firstly, many pieces of music - e.g. pure percussion performances, some rap music, some modern

0-7803-8784-8/04/$20.00 © 2004 IEEE

compositions - do not have a melody in the first place. Second, the users ability to accurately reproduce a melody is limited and there is a big variance of the capabilities of individuals how well they can perform music with their voice [10]. Lastly, for large musical databases, the length of the segment that needs to be hummed in order to achieve a unique hit can be quite substantial, which increases the difficulties associated with correct remembrance and correct performance. However, many Query by Humming systems have been described in the literature (e.g. [11], [12], [13]), and where Query by Humming can be applied, it constitutes a natural and particularly easy-to-use access method to music databases.

D. Query by example

There are several different methods to search music databases by presenting an existing piece of music to the system. Some are focused on finding this very piece by listening to a degraded quality, partial version of it [3], and others try to identify songs from the database that are similar in some respect to the presented song (e.g. similar timbre [4], similar rhythm [14]). However, in the application scenario of a mobile music player, the existence of a reference music piece cannot be generally assumed, which makes this methods less suitable for music selection.

E. Combined systems

As the output of the research project CUIDADO, Pachet et al [5] presented the 'Music Browser', which incorporates various metadata sources and combines automatically extracted metadata, metadata supplied with the song, and the results of web-crawling into one system that allows to browse through a large musical collection. While this is clearly one of the most advanced solutions up to date, it requires keyboard input and sophisticated graphical output, which makes it difficult to integrate into small portable devices.

III. Mood based retrieval

In many cases, users do not want to listen to any specific song, but rather to a category of songs that is defined by the users current mood. Such a category can either be defined by a mood description itself, or it could be described by a genre or subgenre. The search criterion to the database is something like 'I'd like to listen to a ballad, very slow' or 'I want something that expresses both happiness and serenity'. Such an approach would has the advantage that it is very natural - [6] found that style and genre are the most preferred selection parameters if bibliographic data is unavailable. However, there are several obstacles to be overcome in order to make mood based selection work. They can be summarized as follows:

1) How can a set of 'moods' be defined that is relatively unambiguous, widely accepted and 'useful' for the average user?
2) How can the problem of efficient song selection be solved without having to use too many mood classes?
3) How can 'mood' metadata be produced in a cheap and consistent way in a large database?
4) Is the perceived mood of music depending on individual preferences, or at least on the cultural background of the listener? If so, can it still be useful for song selection?

In this paper, we try to answer the above four questions and propose our mood based music selection system.

A. Taxonomy of Mood

In the psychological literature, there have been many efforts to create a taxonomy of mood. Among the earliest attempts to classify mood is Hevner's adjective checklist [1], which uses a list of 67 descriptive adjectives from which the suitable ones are chosen in order to describe the mood. However, since a mood description by this method involves 67 individually biased yes/no decisions, Hevners taxonomy is highly ambiguous and not well suited for mood based song selection.
In this work, we employ Thayer's mood model [16].

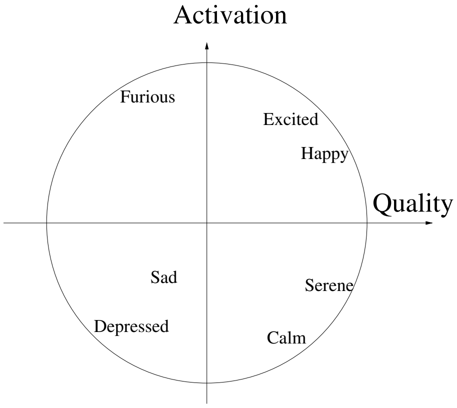

Fig. 1. Thayers model of mood

In this model, an *emotional space* is spanned by two independent emotional dimensions: *activation* and *quality*. Emotions can be viewed as points in the emotional space. By arbitrarily assigning the numerical values $+1$ and -1 for the most extreme possible emotions on the respective dimensions, any possible emotion can be expressed as a two-dimensional vector $\vec{e}$ with $||\vec{e}|| \leq 1$. Following the vector space analogon, similarity between mood can be defined as the euclidean distance between two mood points, and mood detection can be split up into the two independent tasks of activation detection and quality detection [2].

In [7], Liu et al use a hierarchical classification model that directly utilizes the independence of the two dimensions in Thayer's mood model by first classifying along the activation dimension (called 'energy' in their work), and later, using features dependent on the found activity, along the quality dimension. Not surprisingly, the highest confusion rates were measured between the classes 'calm' and 'melancholic' (called 'contentment' and 'depression' in [7]). In our work, we used a different approach, which allows for easier personalization of the mood detection with respect to cultural and individual variations.

IV. Automatic detection of mood from the music signal

In our model, the music data, suitably preprocessed, is a stream of observation vectors $\mathbf{o_t}$ which is assumed to originate from a random, unknown source. Corresponding to the three mood classes happy, melancholic and aggressive there are three random sources S_i (with $i \in [h, m, a]$) with known structure. Then, a hypothesis test can be used to determine whether the music piece originates from one of the three sources (H_0), or not (H_1). This can be expressed by the (log) likelihood ratio

$$L = \log\left(\frac{L(\mathbf{o_{h,m,a}}|\lambda_{h,m,a})L(\mathbf{o_x}|\lambda_x)}{L(\mathbf{o_{h,m,a+x}}|\lambda_{h,m,a+x})}\right) \tag{1}$$

where h, m, a indicates one of *happy, melancholic, aggressive*, and o_h refers to a sequence of observation vectors taken from the random source for mood class 'happy'. o_{h+x} is the concatenated sequence of observation vectors taken from the sources happy and from the unknown (to be determined) source (song). The λ_y are parametric models for the random sources which are trained on some training data (for $\lambda_{h,m,a}$) or on the test song (for λ_x).

The log-likelihood (1) indicates, whether the two signals o_h (and o_m, o_a respectively) and the unknown signal o_x originate from the same random source or not. If the characteristics of the two sources are the same, the models λ_x and λ_h and λ_{h+x} are all identical, the likelihoods $L(.)$ in 1 in the numerator and denominator are identical and hence $L = 0$. If, on the other hand, the sources are different, the likelihood of the numerator term will always be higher than that of the denominator, and $L > 0$. The higher the value of L, the lower is the probability that x and h are taken from the same random source.

The above is only true if the model complexities of the three models λ_x, λ_h and λ_{x+h} are identical, because the likelihood values $L(.)$ depend strongly on the model complexity (the 'strength' of the model). As usual, a compromise must be found between a very general but weak model and a complex model that learns the training set very well but generalizes badly. On typical pop music data, using a multivariate normal distribution was found to work very well. Assuming such a model for λ, the likelihood terms $L(.)$ in eq. 1 can be expressed as

$$L(x_i; \mu_1, \Sigma_1) = \prod_i \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

Taking the negative log yields

$$\log L(x_i; \mu_1, \Sigma_1) = \frac{N}{2}(d\log(2\pi) + \log(|\Sigma|)) \tag{2}$$

$$+1/2\sum_i (x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \tag{3}$$

The ML estimate of Σ is

$$\Sigma = \frac{1}{N}\sum_i (x_i-\mu)(x_i-\mu)^T \tag{4}$$

The term (3) is equal to $\frac{N}{2}$. Therefore, the likelihood $L(x; \mu_1, \Sigma_1)$ depends only on the determinant of the (ML estimate of the) covariance matrix Σ_1 and the feature space dimensionality d, which is the same for all feature vectors. Using this result and $N = N_1 + N_2$, eq. 1 can be written

$$\lambda = N_1\log(|\Sigma_1|) + N_2\log(|\Sigma_2|) - N\log(|\Sigma|) \tag{5}$$

If multivariate normal densities are used for the λ_y in eq. 1, the knowledge of the model parameters is sufficient to compute the likelihoods $L(.)$ in eq. 1. There is no need to store the feature vectors or the musical data itself.

The denominator term in 1 requires the knowledge of the model λ_{h+x}. This model can be computed solely based on the ML models for the sources h and x using

$$\mu_j^{(h+x)} = \frac{1}{N}(N_h\mu_j^{(h)} + N_x\mu_j^{(x)}) \tag{6}$$

for the means and

$$\begin{aligned} N\sigma_{ik}^{(h+x)} &= \\ &N_h\sigma_{ik}^{(h)} + N_x\sigma_{ik}^{(x)} + N_h\mu_i^{(h)}\mu_k^{(h)} + N_x\mu_i^{(x)}\mu_k^{(x)} \\ &-N\mu_i^{(h+x)}\mu_k^{(h+x)} \end{aligned}$$

for the covariance of the joint model $\lambda^{(h+x)}$.

A. Preprocessing

After selecting the desired part of the music, the corresponding segment is downsampled into a 22.05 kHz, 16 bit, mono channel PCM signal. 19 Cepstral features are computed for every 32 ms window with a shift of 32 ms. First, for each frame, the power spectrum is computed using a preemphasis $z - 0.97z^{-1}$ and a Hamming-window. On the power spectrum, spectral centroid, spectral rolloff ($w = 0.85$) and spectral flux are computed. The power spectrum is further condensed into 24 mel spaced bins in the frequency range from 200 to 7500 Hz and logarithmized. A discrete cosine transform and sine-liftering is applied, resulting in 19 cepstral coefficients per 32 ms frame. C0 is discarded, and first order derivatives of the cepstral coefficients with an analysis window of ±2 frames are computed. The 19 cepstral coefficients and the 19 delta cepstral parameters are concatenated to form a 38-dimensional intermediate feature vector, which is then LDA-transformed to a 24-dimensional vector. Three spectral features directly computed from the power spectrum are appended to this 24-dimensional vector yielding the final 27-dimensional features.

V. Experiments

A. Selection of mood classes for music

For the application of Thayer's model to music, ten subjects were asked to label a database of music according to their impression of the mood that was conveyed by the respective

0-7803-8784-8/04/$20.00 © 2004 IEEE

pieces. In an initial experiment, 4 mood classes corresponding to the four quadrants of the two-dimensional mood space of figure 1 were provided: *Aggressive* for the upper left, *Happy* for the upper right, *Calm* for the lower right and *Melancholic* for the lower left quadrant. However, after the initial labeling of the entire database with several different subjects, a high inter-labeler disagreement was noticed on some well-known pieces. This led to the design of a small experiment where the same set of 34 randomly chosen songs was judged by a set of 10 human listeners, 5 of which came from a Western European cultural background and the other 5 from an Asian one (Chinese and Japanese). The total distribution of the found emotions, averaged over all labelers and split by cultural background, is summarized in table I.

	Western	Asian	Total
Aggressive	24.7%	32.4%	28.5%
Happy	33.5%	31.8%	32.6%
Calm	18.2%	19.4%	18.8%
Melancholic	23.5%	16.5%	20.0%

TABLE I

RESULT OF MANUAL MOOD ANNOTATION

There is a statistically significant ($\alpha > 0.94$) cultural difference in the perception of mood in music: People from an Asian cultural background perceive the same music as more aggressive (and less melancholic) as compared to their Western background counterparts.
Interestingly, there was not a single case where all ten listeners agreed to one of the four emotion classes. Therefore, the evaluation was repeated with the two dimensions defined by Thayer's model of mood. As a result, people tended to agree much better on the active / nonactive dimension (14 cases of total agreement of all ten listeners, and an average entropy of 0.125) than on the positive / negative dimension (still no case of total agreement of all ten listeners, and an average entropy of 0.23).

Since there was no case of total agreement between the listeners, the cases where 9 or 8 of the 10 listeners agreed on one perceived mood were analyzed. This occurred for a total of 14 songs. From this 14 songs, 7 were labeled as 'aggressive', 6 were labeled as 'happy', one as 'melancholic' and none as 'calm'. Therefore, it can be concluded that agreement on the positive / negative dimension in Thayer's mood model can be achieved much more easily for the songs with a 'high' activity value than for songs with a 'low' activity value (corresponding to the emotions 'calm' and 'melancholic' in our taxonomy). In general, the expected error rate for human transcription (as evaluated on a leaving-one-out scenario) is 34.7% (22.9% if the two classes 'calm' and 'melancholic' are fused).

From the above experiments, the following conclusions may be drawn:

1) There is a strong inter-human variance in the perception of mood in music
2) Cultural background influences the perception of mood - on average, Asians perceive Western pop music as more aggressive than Westerns do
3) It is in general easier to classify musical mood along the 'active/nonactive' dimension than on the 'positive/negative' dimension
4) 'Positive/negative' is easy to classify for songs that are also rated 'active'

Therefore, we chose three mood classes for our experiments: 'happy', 'aggressive', and 'melancholic/calm', dropping the distinction between positive and negative in the case of low activity.

B. Performance of the automatic mood detection

Table II shows the result of the automatic mood detection algorithm on a test set of 616 pop songs taken from the period from 1950 until 1995. Although the total error rate is around one third, it has to be taken into account that the expected human transcriber "error rate" is around 20..25%, which means that the system is not that far from human performance.

	Aggressive	Happy	Melancholic+Calm	#
Aggressive	51.7%	32.9%	15.3%	170
Happy	8.1%	60.2%	31.8%	236
Melancholic+Calm	4.3%	15.2%	80.5%	210

TABLE II

RESULT OF AUTOMATIC MOOD DETECTION

C. System setup

In our current system, the user injects his music into our music box inserting a CD or by simply copying a music data file to the system directory. The system then computes the three distances to the three mood models (happy, melancholic and aggressive), using eq. (5), resulting in a 3-dimensional distance vector.
Since the mood models are just ML-estimated normal distributions, they can easily be recomputed. This can be used to allow user adaptation. The user is asked to select a few (5-15) songs as representative songs for any of the moods, and the system computes the corresponding mood model on the fly and replaces the precomputed mood model with the new, user-defined one. By this mechanism, both individual and cultural differences in the perception of mood can be compensated. Of course, it is also possible to adapt an existing model rather than to compute a completely new one. For example, if a song is misclassified by the system and the user corrects the system, this song can be taken into the training set of the corresponding emotional class, and the mood model can be adapted accordingly.
Since there exist two independent dimensions of mood, the system needs a two-dimensional input. This can, e.g., be a joystick which has two degrees of freedom, or alternatively,

two sliders on a standard graphical user interface might be used.

The three values from the automatic mood computation are log-likelihood ratios and not distances, and hence they do not constitute a metric. Therefore there is no easy way to transform the three-dimensional distance vectors into something that can be used for mood navigation. We used the following approximation for the mapping. Since the desired output space should reflect the dimensions of Thayer's mood model, it can be concluded that the distances between the four emotional classes that are prototypical for the four quadrants should be maximized in the target space. This is equivalent to maximizing the between-class scatter matrix for the four centroids while keeping the total scatter matrix of all data constant (i.e. not 'blowing up' the feature space), which can be achieved linearly by applying an LDA transformation with the prototypical emotions (three in our case, 'melancholic', 'happy' and 'sad') as classes. The resulting projection is shown in figure 2. As a comparison, figure 3 shows the result of a projection to the first two PCA principal components. It can be clearly seen that the LDA based solution achieves both better class separability and a more uniform distribution of the song samples over the transformed space.

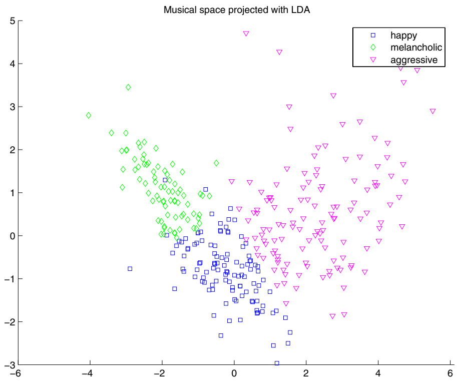

Fig. 2. Transformation into the navigation space: LDA

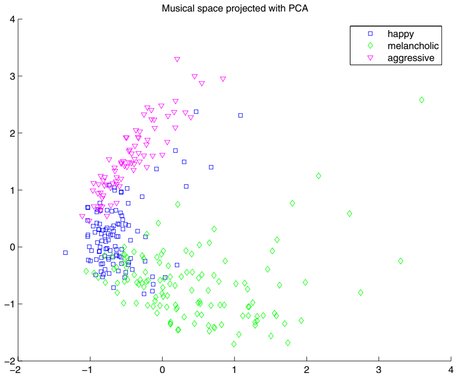

Fig. 3. Transformation into the navigation space: PCA

VI. Conclusions

In this paper, we have proposed a system that allows navigation inside a music database by navigating a two-dimensional mood plane. The system works fully automatically and can be easily adapted to accommodate cultural or individual differences in the mood perception. Since the method does only require two one-dimensional inputs and no keyboard, it is well suited for use in small portable devices. In preliminary experiments, the feedback from our test users was positive.

VII. Acknowledgements

The authors wish to thank all members of the Speech and Sound Group, especially Toni Abella, Franck Giron, Yin Hay (Vicky) Lam, Krzysztof Marasek, Thomas Schaaf, Juergen Schimanowski and Josep Sola, for their support. Without their devotion and active help, this work would not have been feasible. We also thank Robbie Tingey, Chris Liu and Kaz Mugura at Design Center London for their contribution to the experiments.

References

[1] K. Hevner, *Expression in music: a discussion of experimental studies and theories*, Psychological review **42** (1935), pp. 186-204

[2] R. Tato, R. Santos, R. Kompe, *Emotional space improves emotion recognition*, Proc. ICSLP 2002, Denver, Colorado, September 2002

[3] A. Wang, *An industrial-strength audio search algorithm*, Proc. of the Fourth International Conference on Music Information Retrieval, pp. 7 ff, Baltimore, MD, October 2003

[4] J.J. Aucouturier, F. Pachet, *Finding songs that sound the same*, Proc. of the IEEE Benelux workshop on model based processing and coding of audio, University of Leuven, Belgium, November 2002

[5] F. Pachet, A. Laburthe, A. Zils, J.J. Aucouturier, *Popular Music Access: The Sony Music Browser*, Journal of the American Society for Information Science, 2003

[6] D. Bainbridge, S. Cunningham, J. Downie, *How People Describe Their Music Information Needs: A Grounded Theory Analysis Of Music Queries*, Proc. of the Fourth International Conference on Music Information Retrieval, pp. 221 f, Baltimore, MD, October 2003

[7] D. Liu, L. Lu and H.J. Zhang, *Automatic mood detection from acoustic music data*, Proc. of the Fourth International Conference on Music Information Retrieval, pp. 81 ff, Baltimore, MD, October 2003

[8] F. Pachet, D. Cazaly, *A Taxonomy of Musical Genres*, in Proc. RIAO 2000, Paris, France, April 2000

[9] H. Gish, M.H. Siu, R. Rohlicek, *Segregation of Speakers for Speech Recognition and Speaker Identification*, in Proc. ICASSP-91, S. 873 ff

[10] S. Pauws, *Effects of song familiarity, singing training and recent song exposure on the singing of melodies*, Proc. of the Fourth International Conference on Music Information Retrieval, pp. 57 ff, Baltimore, MD, October 2003

[11] T. Kageyama, K. Mochizuki, Y. Takashima, *Melody retrieval with humming*, Proc. ICMC 1993, pp 349-351

[12] R. Dannenberg, W. Birmingham, G. Tzanetakis, C. Meek, N. Hu, B. Pardo, *The MUSART testbed for query by humming evaluation*, Proc. of the Fourth International Conference on Music Information Retrieval, pp. 41 ff, Baltimore, MD, October 2003

[13] T. Sonada, M. Goto, Y. Muraoka, *A www-based melody retrieval system*, Proc. ICMC 1998, pp. 349-352

[14] J. Foote, M. Cooper, U. Nam, *Audio retrieval by rhythmic similarity*, Proc. of the Third International Conference on Music Information Retrieval (ISMIR), Paris, France, October 2002

[15] M. Cooper, J. Foote, *Automatic music summarization via similarity analysis*, Proc. of the Third International Conference on Music Information Retrieval (ISMIR), Paris, France, October 2002

[16] R. E. Thayer, *The biopsychology of mood and arousal*, Oxford University Press, ISBN 019506827-0, several prints exist