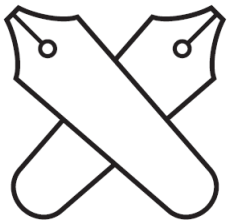
Doctoral Dissertation Academic Year 2015

Emotion Identification System for Musical Tunes based on Acoustic Signal Data



A dissertation for the degree of
Ph.D. in Media and Governance

Graduate School of Media and Governance
Keio University

Tatjana Endrjukaite

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I'd like to thank my main research advisor, Professor Yasushi Kiyoki. I deeply appreciate Professor Kiyoki for his sensible leadership, penetrating advice, and continuous support. He gave me the opportunity to compile my studies into this thesis. Thanks to him, the work that went into this thesis was a delightful quest. He has also provided many opportunities to enrich my academic life, including working on interesting projects, and attending conferences around the world.

I also thank Professor and co-Advisor Hideyuki Tokuda and Professor and co-Advisor Wanglin Yan, Professor and co-Advisor Jin Mitsugi, Professor and co-Advisor Kunihiko Okano for their valuable comments. Their discussions with me on the various aspects of my research improved the quality of this thesis.

I'd like to thank Asako Uraki for the support she has given me not only with reading my papers, watching countless presentations and helping me with my Japanese, but also with everyday issues, and the countless, countless numbers of forms I've had to deal with over my time at Keio SFC.

Shiori Sasaki was also an immense help and support during my entire time at SFC, always willing to help, and giving invaluable advice whenever I needed it.

Thanks to all the participants of the survey for participating in my research, for listening lots of music and answering the questionnaires.

Professor Akiyoshi Hatayama always gave insightful advices, and provided really interesting conversations when I was given a chance to show him a paper or presentation of mine.

I would also like to thank the members of my MDBL laboratory for their support over the years of my membership to that research lab. They have listened to countless presentations, paper overviews and research ideas. They have always provided invaluable support, attention, questions, and friendship over the several years I have been here.

I'd also like to thank Professor Kuniaki Mukai from novel computing, for asking a mix of interesting questions and observations, and for giving good ideas regarding presentations that I have given.

And special thanks to my family who were supporting me all the time and believing in my achievements on every stage of this work.

Tatjana Endrjukaite

Thesis Abstract – Academic Year 2015

Emotion Identification System for Musical Tunes based on Acoustic Signal Data

Music plays an important role in the human's life. It is not only a set of sounds – music evokes emotions subjectively perceived by listeners. The growing amount of audio data wakes up a need for content-based searching. Traditionally, tunes information has been retrieved based on reference information, for example, the title of a tune, the name of an artist, the genre and so on. When users would like to try to find music pieces in a specific mood such standard reference information of the tunes is not sufficiently effective. We need new methods and approaches to realize emotion-based search and tune content analysis.

This thesis proposes a new music-tune analysis approach to realize automatic emotion identification by means of intrinsic musical features. The innovativeness of this research is that it uses new musical features for tune's analysis, which are based on human's perception of the music. Three significant approaches are introduced for calculation of essential features of emotional aspects of music pieces. By means of these approaches the primary musical characteristics can be expressed. These are repeated parts of a tune, thumbnail of a tune, and internal homogeneity pattern. At the heart of the system described in this research, the innovative algorithm to process non-stationary signal data was introduced for audio signals processing in particular. Moreover, this research describes a new way of automatic emotion processing, calculation and presentation on a plane, which has the key idea to quantitatively measure emotions without categorizations.

Finally, this thesis describes the overall architecture of the tunes emotion recognition system, as well as discusses details for specific applications including tunes querying by emotion, querying by tune, and playlist suggestion. We also describe an evaluation section where the proposed approach is tested and then is compared to other researches. Most important distinctions of the proposed approach are that it includes broader range of tunes genres, which is very significant for music emotion recognition system. Emotion description on continuous plane instead of categories also results in more supported adjectives for emotion description which is also a great advantage.

Keywords: music analysis, music emotions, emotion recognition, music similarity, instantaneous frequency spectrum, tune internal homogeneity, music repetitions.

論文要旨- 2015 年度

音響信号データ分析による楽曲感性識別システムの研究

音楽は、人類の歴史において重要な役割を担ってきた。音楽は、単純な音の集合としての現象だけではなく、その表現により聴き手に主観的な感情・感動・印象を与える作用を持っており、これらは人間の音楽に対する記憶想起や比較において極めて重要な特徴の１つである。近年の楽曲データの増加により、膨大な楽曲データを対象とした楽曲検索の必要性が増大してきている。楽曲の感情・感動・印象といった感性的側面を対象とした楽曲検索の実現は重要な研究課題の１つとなっている。楽曲の感性検索の実現において重要な要素は、１）楽曲（コンテンツ）の信号情報から感性的特徴量を自動的に分析する機能を実現すること、および、２）楽曲の感性的特徴量を比較する機能を実現することにある。

本研究は、信号処理による楽曲感性特徴量自動分析方式を提案する。本方式により、音（信号）の持つ感性的特徴量を自動的に定量化することが可能となる。また、本方式を大規模な楽曲データベースに適用することにより、大量の楽曲データを対象とした感性検索の実現が可能となる。本研究の提案方式における主要な機能は、１）楽曲内の信号特徴量の反復自動抽出機能、２）楽曲内における反復の一致性判定機能、３）楽曲間における同種性判定機能である。これらは楽曲の感性的側面を信号処理によって定量化する新しい方式として位置付けられる。

さらに、本論文では、本研究の実現方式として、楽曲感性特徴量自動分析方式のアーキテクチャを示す。また、本研究の応用方式として、楽曲感性検索における感性問い合わせ生成方式、楽曲による問い合わせ方式、および、プレイリスト推薦システムを示す。また、本論文では、本実現方式により構築した実験システムにおける実験の結果と考察を述べ、本研究の有効性を実現可能性を示す。

キーワード: 音楽分析，音楽感性表現，感情認識，楽曲類似性，瞬間フリーケンシースペクトラム，旋律内同種性，反復性。

Chapter 1. Introduction

Huge amount of music pieces already exist and being generated every day all over the world. Never before such a large collection of music pieces has been available and accessed daily via the Internet using personal computers, smartphones and other devices. As the amount of music content continues to increase the way tunes information is created and organized has to evolve. Although standard information such as the name of the artist, genre, and title remains important, these tags have limited applicability in many music-related queries [26]. Nowadays, users expect more semantic metadata to access music, such as similarity, style and mood.

As there are more and more tunes available in personal music libraries it becomes very difficult to find desired ones among them. Reference information of music pieces help to decide if it is what we are searching for or not. However, to find a new tune we will like, we have to listen to each music piece to find it out. This could be very time consuming. We start to realize the importance of creating new kind of metadata that allow users to access music pieces easier.

The growing amount of audio data wakes up a need for content-based searching. When users want to find tunes in a specific mood standard reference information of the pieces is not enough. Nevertheless, many users experience difficulty in formulating a query in words when they want to retrieve emotional aspects of music pieces, since it is difficult to describe such information. For example, users want to retrieve similar music pieces they are listening to, however they don't know the artist and the style of the music, but they want to listen to a tune with a similar emotion to the one they perceived.

It is well established that human beings respond emotionally to music, little is known about precisely what it is in the music that they are responding to. In recent years the design and implementation of tunes playlist suggestion systems is one of the key issues in the field of multimedia research. In the design of such system the important issue is how to define and represent the internal features of music pieces and how to select tunes according to the user's impression.

At present, there are a lot of researches on music features extraction from MIDI files and emotion estimation for MIDI music pieces [11, 12], as presented on Figure 1.1. However, there is still no complete theory for audio music analysis and acoustic signal

processing for emotion recognition. How to deal with emotions in audio signal processing?

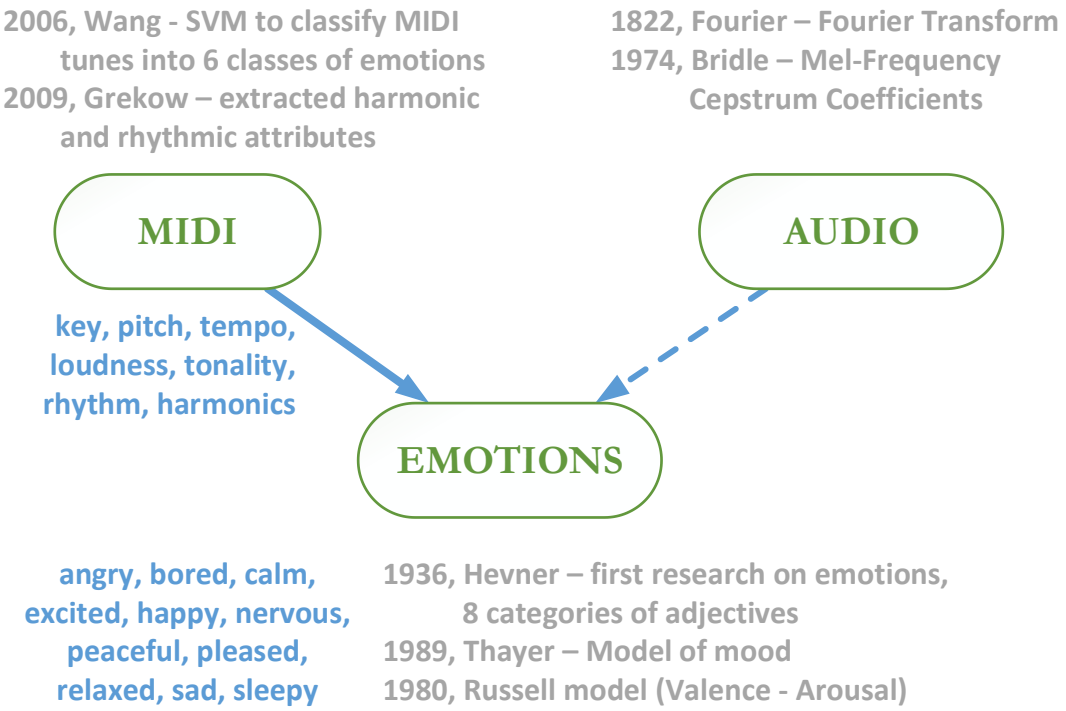


Figure 1.1 Research focus and background.

The reason that many people engage with music, as performers or listeners, is that it has power to evoke or enhance valued emotional states [1, 2]. In the light of this we can say that emotions play a very important role in music. The relation between musical sounds and their influence on the listener's emotion were studied by Hevner through experiments, which substantiated a hypothesis that music inherently carries emotional meaning [3].

1.1. Goal

In PhD research I focus on dependency between intrinsic characteristics of tunes and emotions people experience while listening to those tunes. By analyzing repetitions and tunes internal homogeneity we may discover why a piece of music influences us the way it does. We try to predict what influence an unknown tune would have on us before we listen to it.

The ultimate goal of this research is to construct a system for determining the expected emotional effect of listening to a tune. Such approach could be used in tunes playlist suggestion tasks. For example, the proposed method can be used in a system designed

for selecting appropriate tunes according to the user's wishes to suggest these tunes for listening to the user.

1.2. Definitions

The basic terms used in this thesis are defined as follows.

- ***Tune*** is a complete music piece with or without voice in a form of audio signal data with typical length 5 minutes in average.
- ***Acoustic Signal*** is an acoustic wave digital representation in the form of PCM format with 44100 values per second and with 16 bit per value.
- ***Tune Thumbnail*** is a short fragment of a tune which represents easily recognizable part of the tune with typical length 10–15 seconds.
- ***Emotion*** is a person's perceived state of feeling from listening to a whole tune from the beginning till the end. It is not a mood of a listener, but it is an emotion of a tune.

To process music pieces three significant tunes approaches were introduced: repetitions identification inside a tune, tune's internal homogeneity, and music similarity analysis.

- ***Repetitions*** or ***Repetitive structures*** are fragments that a music piece contains multiple times. These fragments may be slightly different, for example, they might have little difference in tonality and could be performed by different kinds of musical instruments, but are perceived as very similar.
- ***Tune Homogeneity*** is information about internal structure of a tune. Homogeneity for any specific moment within a tune is considered as high when this moment sounds typical to the whole tune. For example, a music piece is played by piano slowly from the beginning till the end, but there are a few short inclusions of irregular musical instrument performances such as drums parts. In that case most of the time tune is typical to itself and can be called homogeneous, but fragments with drums sound untypical and are called inhomogeneous.
- ***Music Similarity*** is a metric that finds similarity between two music tunes. Tunes which have high similarity may have different tonality and may even have different genre but sound very similar to each other.

1.3. Problems and research challenges

There are many research challenges that take place in the field of music analysis and emotions. Following challenges were focused in the current work:

- Music is a non-periodic and non-stationary signal which makes it hard to process music tunes with common signal processing techniques, such as Fourier transform and others. [27, 36, 38, 39].
- People perceive music in a very undetermined way which means that human beings can easily identify similar melody in two fragments that have different tonality, performed by different musical instruments, and may even have different rhythms and speed. That fact makes it very hard to find and estimate repetitions and similarity in music pieces.
- Human brain works in a way to pay more attention to untypical perceptions. This is also observed in listening to music. Untypical parts of a tune are very important fragments that make the impression. However, there is no precise way to determine typical and untypical fragments of a tune.
- Emotions do not have an easy measurement way and an emotion is not a quantitative value that can be easily processed and calculated.

The challenges mentioned above were addressed in this work.

Besides that there are also a number of issues in emotion detection methods [28, 29].

- *Precision or accuracy*

  One of the most obvious criteria for a good emotion or mood detection algorithm is the precision that is achieved. An algorithm gives an output, which depends on the used algorithm. Some give just one emotion others a multi-label output, for example 60% of emotion A and 40% of emotion B. This output is compared with the annotated emotions evoked by a test subject. The percentage of correct answers determines the accuracy.
- *Granularity*

  In a strong relationship with the precision, granularity has a very big influence on the achieved accuracy. When one has to choose between more options, there is a bigger chance that a fault one is chosen. Low granularity, for example 2 emotions can be useful but that is depending on the application. All in all, when there are more adjectives users have more options to describe the emotion precisely. Generally speaking, more adjectives is better.

- *Diversity*

  Some researches only use a limited number of songs or just one or two genres of music. It has its influence on how much the algorithm can be optimized for that particular genre. However, this gives an unfair advantage when comparing the accuracy with other methods, which algorithms uses more kind of music.

- *Cultural background*

  According to [1, 30] emotional expression is best recognized between members from the same ethnical group and that expression can lose their meaning when crossing cultural borders.

1.4. Organization of this Thesis

This thesis is organized as follows.

Chapter 1 gives an introduction to the research area, main definitions and presents problems and challenges.

Chapter 2 describes the related work and their results comparison.

Chapter 3 describes approaches of tune processing and emotion calculation which are used for emotion recognition.

First, we introduce IFS spectrum which is used as a main acoustic signal processing method. Second, we introduce tune structure analysis approach – tune internal homogeneity analysis. Third, we describe music similarity calculation approach which is used for tunes comparison. And finally, the emotion plane is described which is used for emotions representation and processing.

Chapter 4 describes the emotion recognition system design including the overall architecture of the system, some details about tunes comparison and combined emotion calculation, and finally the structure of the system's database.

For the proper operation the system needs to have database filled with data. Two types of information is required: collection of tunes with known emotions, and mapping of emotion adjectives to the emotion values presented in numbers. Both data types are specific to human perception and therefore the data must be collected using polls and questionnaires with some individuals. This is described in Chapter 5.

Chapter 6 describes the evaluation of the tunes emotion recognition approach. First, it is required to specify the accuracy calculation between two emotions which will be used for comparing known and calculated emotions. Second, the validation approach is described

in following chapter. And finally, the comparison with other researches is presented in the end of this chapter.

Chapter 7 describes query by emotion, query by tune, and playlist suggestion applications of the described approach and provides structures of the system for each of these applications.

Chapter 8 discusses future research.

Finally, Chapter 9 summarizes this thesis with concluding remarks.

Chapter 2. Related work

The main motivation of previous researches is to determine audio music characteristics and their mapping to emotions. There are two types of related work presented in this thesis. The first is regarding emotions presentations. And second is dedicated to emotions recognitions in audio data retrieval.

2.1. Emotions in Music

An important question is: which emotions are relevant for detection in music? There has been a lot of research on emotions in psychology such as [3, 13, 31, 32]. K. Hevner was the first one to do scientific research on the topic of emotional expression in music. The relation between musical sounds and their influence on the listener's emotion were studied through experiments, which substantiated a hypothesis that music inherently carries emotional meaning [3]. Figure 2.1 represents Hevner's adjectives grouped into emotion circle.

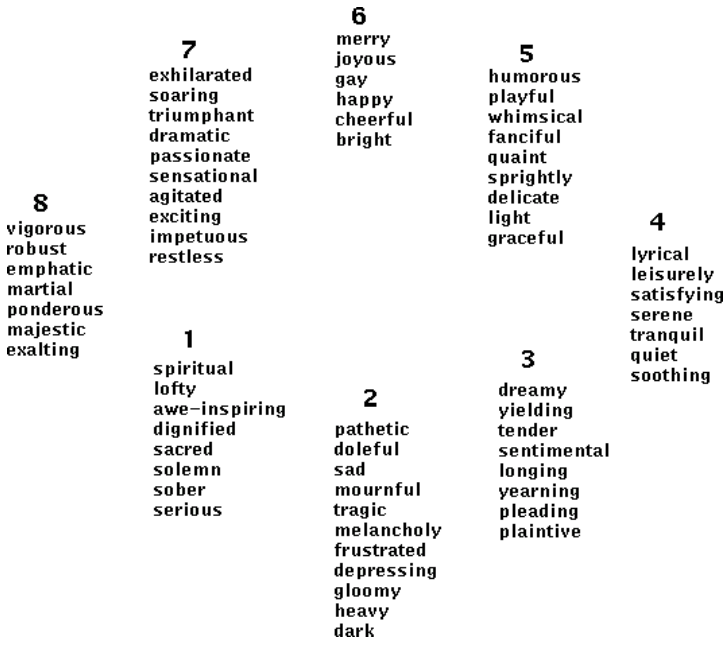


Figure 2.1. Hevner's circle of emotions.

Thayer's model of mood offers a simple model for moods [13]. Along the horizontal axis the amount of stress is measured and along the vertical axis the amount of energy, as shown on Figure 2.2. In music one can think of energy as the volume or the intensity of

sound. Stress can be described as “doing many things” so the difference in tonality and tempo would be a good mapping.

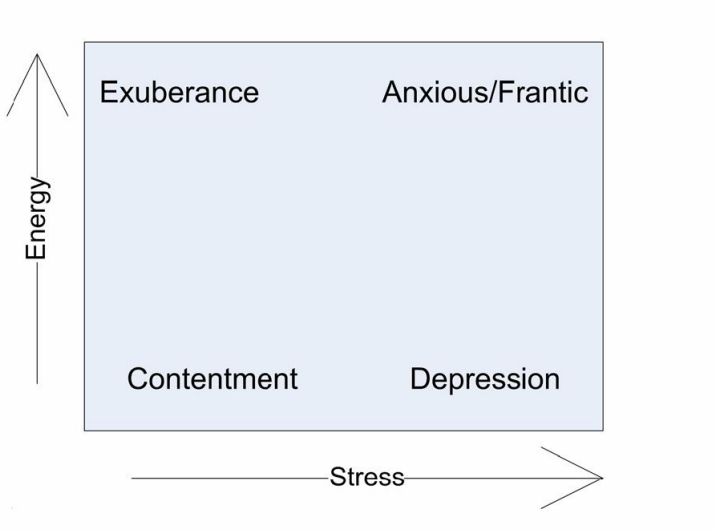


Figure 2.2. Thayers model of mood.

And one more model is the Tellegen-Watson-Clark model of mood [32]. This model contains a lot more emotions and uses the positive/negative affect as one dimension and the pleasantness/unpleasantness versus engagement/disengagement (45 degrees rotated) as the other. This model is shown on Figure 2.3.

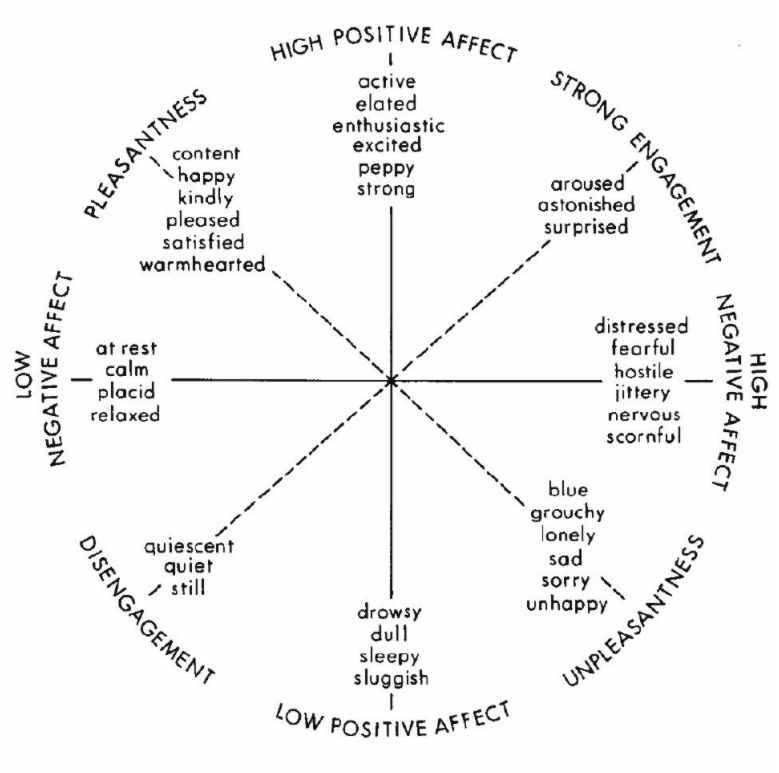


Figure 2.3. The Tellegen-Watson-Clark model of mood.

Summary and Discussion

Which is the best model for emotion presentation? It is depends on the application the model is placed in. The number of different emotions and their correlation has its impact on the precision of the method. For some applications, such as the physiotherapist, aren't a lot of emotions necessary as long as the emotion that needs to be evoked is there and the algorithm has a very high recall percentage on that emotion [29]. On the other hand when a listener is at home, or at some place, and he or she wants to listen to upbeat nice music, but application has just a few, for example three or four emotions, than user will get all happy music. In such a case a higher granularity is required.

2.2. Emotions recognition

Regarding determination of audio music characteristics, researches have made genre recognition analysis [16, 33], music similarity analysis [4, 18], and tried to apply researches on emotions presentation to automatic emotion recognition [21, 22, 23, 24, 25].

Some of the previous research on automatic emotion recognition in audio signal data is listed below. Table 2.1 summarizes their main characteristics.

1. Yang et al. [21]

   This research evaluated a structured emotion rating model for embodiment in software agents to assist human annotators in the music annotation system. The motivation behind this research is that music annotation poses too much pressure on listeners. The goal is to make music annotation (emotion in particular) easier but still provide a human input. As a fundament for emotion detection the Teller-Watson-Clark emotion model is used.

   The researchers want to focus on the more negative emotions in this model because these emotions would be harder distinguishable than positive emotions. Results were given for a single-attribute test to rate emotion intensity (the sum of positive and negative energy in the model), based on 500 songs. About 90% accuracy was achieved using both timbral and rhythmic features. This acoustic based algorithm was used to categorize the songs into two classes of emotions corresponding to the Teller-Watson-Clark emotion model Hostility, Sadness, Guilt and Love, Excitement, Pride respectively.

2. Li et al. [22]

   The problem of emotion detection in music in this research is presented as a multi-label classification problem. Musical pieces can belong to more than one emotion. There are 10 groups of emotions that were used and three added emotions groups. These three extra groups were added according to a test subject who indexed the test songs. The test subject was also asked to group the emotions into groups. The classifiers used in this research are based on support vector machines. The acoustic features that were used are timbral texture features, rhythmic content features (beat and tempo detection) and pitch content features.

   The accuracy is around 50% but it has relatively higher granularity since 4 different genres were used in the research. The half-way point between the precision and the recall, was 46% in micro-averaging and 43% in macro-averaging. In six-supergroup experiment the breakeven point was 50% in micro-averaging and 49% in macro-averaging, so the overall accuracy was improved when the number of categories is reduced.

3. Liu et al. [23]

   Mood detection for a specified part of music is the main subject in this research. The algorithm is based on Thayer's model of mood. Features of intensity, timbre and rhythm are used. Intensity is mapped to energy and both timbre and rhythm are mapped to the stress component.

   Two frameworks are given, a hierarchical and a non-hierarchical framework. In the first framework musical pieces are firstly shifted on intensity into group 1 Depression/Contentment and group 2 Exuberance/Anxious respectively. The second step makes the distinction between the 2-tuples. In the second framework all features have their impact at the same time. This research manages to reach accuracy ranging from 76.6% to 94.5% for the hierarchical framework and 64.7% to 94.2% for the nonhierarchical framework. The accuracy is very high but the algorithm is trained on classical music only with four moods, so the granularity is very low.

4. Carvalho et al. [24]

   This research proposes classification based on how much "happiness" is present in a song. Authors rate a song on a 5 points based scale. They think that in this way the labels will be more appealing to users when used in a real-life application. On the other hand this is a much easier job for their algorithm than when more categories, for example, fear, tranquil, whimsical, would have been used. In advance, the songs are classified by two persons. The accuracy is approximately 82%. Four classes of features were used: Musical surface, Spectral Flatness Measure, Spectral Crest Factor and Mel Frequency Cepstral Coefficients.

5. Li et al [25]

   This research describes a system which does two things based on acoustic features in music. First, a similarity search which gives music which is similar to a given piece and the second, emotion detection in music. The extracted music features used are the Mel-Frequency Cepstral Coefficients, the Musical surface features and the Daubechies Wavelet Filters (DWCH). This method manages to yield a fairly high degree of precision on Jazz musical pieces, at least 70% to a maximum of 83%. Again, there are not so many categories to choose from such as, Cheerful and Depressing, Relaxing and Exciting, and Comforting and Disturbing. So the granularity is somewhat low.

Summary and Discussion

There are various aspects to an emotion detection method. In general methods perform better when fewer emotion categories are used, although a greater number of categories is often desirable. With lower granularity levels emotions like frustration and excitement could fade into each other. Music pieces trimmed to 20 or 30 seconds and are fed to the algorithm. These clips are annotated with a certain emotion and the result of the algorithm will be compared with the annotated value. Then the learning machine will learn by mistakes and finds acoustical features which choose better result for given emotions. In this way the best values can be found and the algorithm gets better precision. However, a segment of 20 or 30 seconds will be far too little to use especially for classical music, because the articulation and tempo can differ greatly in classical music pieces.

Table 2.1. Comparison of previous research.

	Dan Yang, WonSook Lee	Tao Li, Mitsunori Ogihara	Dan Liu, Lie Lu, Hong-Jiang Zhang	Carvalho, Chao	Li, Ogihara
Database size	500	250	250	200	235
Tune length	20 seconds	30 seconds (random)	20 seconds	—	—
Genres	1	4	1	—	1
Method	Support Vector Machine (SVM)	Support Vector Machine (SVM)	Gaussian Mixture Model (GMM)	Support Vector Machine (SVM)	Support Vector Machine (SVM)
Features	timbre, rhythm	timbre, rhythm, pitch	timbre, intensity, rhythm	musical surface, spectral flatness, spectral crest, MFCC	MFCC, musical surface, DWCH
Emotion categories	2 categories	13 groups	2 groups	5 categories	3 categories
Emotion adjectives	—	23 adjectives	4 moods	1 adjective	—
Accuracy	90%	~50%	76 - 95%	82%	70 - 83%

Chapter 3. Approaches

This chapter describes approaches of tune processing and emotion calculation which are used for emotion recognition.

First, we introduce a novel approach of acoustic signal analysis – IFS spectrum. IFS is used as a main acoustic signal processing method because it is intended for processing non-stationary and non-periodic signal, such as music.

Second, we introduce the innovative approach for tune structure analysis – tune internal homogeneity analysis. This approach provides an easy way to see the structure of a tune, as well as find typical and untypical parts, which is extremely important for music analysis from the human perception perspective.

Third, we describe the approach for identifying repetitive structures in tunes which have a strong impact on music perception as well. The important advantage of the proposed approach is in its ability to detect not only identical parts in acoustic signal of a tune, but also those part that are present in different tonality, performed by different musical instruments, but sound similar and perceived by listener the same way.

Next, music similarity calculation approach is described which is used for tunes comparison. This approach uses all of the methods mentioned above to compare tunes for similarity based on aspects of human perception aspects: overall structure, thumbnail, and repetitive parts of a tune.

And finally, we describe a new unique approach for emotions definition, representation and calculation – the emotion plane which is used for describing the listener's perception of music. Unlike other models of emotions, this approach is much more suitable for emotions processing, calculation and aggregation which plays a significant role in music emotional analysis.

3.1. Tune processing

The approach for tunes comparison is extremely important for system successful operation. I based the approach for tunes comparison on the method described in [18], which compares tunes by their descriptors that contain spectrums of most repeated parts or in other words about repetitions inside the tune.

The idea for acoustic signal comparison consists in using instantaneous frequency spectrum. That means that we compare IFS spectrums of signals when we want to compare one tune to another. In the initial work only repetitions were used, but we are proposing to extend the descriptors with tune internal homogeneity information and enrich the list of most representative parts.

Instantaneous frequency spectrum calculation, tune's repetitions detection, and tune internal homogeneity calculation approaches are described in chapters 3.1.1, 3.1.2, 3.1.3 and 3.1.4 accordingly.

3.1.1. Instantaneous Frequency Spectrum

As a part of the related research I was working on time-dependent tune structure analysis for genres recognition where I introduced a new method of signal processing: the Instantaneous Frequency Spectrum (IFS). This IFS method was used as a base approach for genres recognition, and thus demonstrated its effectiveness in tunes description. IFS is used as an important part in the current research.

The Instantaneous Frequency Spectrum approach is based on the Hilbert-Huang Transform approach [10]. As the first step of the approach the signal is decomposed into so-called intrinsic mode functions (IMFs) using Empirical Mode Decomposition (EMD) [8] and then instantaneous frequency data is obtained by means of the Hilbert transform (HT) [34]. Since the decomposition is based on the local characteristics on time scale of the data, it can be applied to non-periodic and non-stationary processes signal data.

The EMD method is a necessary step to reduce any given data into a collection of IMFs, because they have well-behaved Hilbert transforms, from which the instantaneous frequencies can be calculated [8].

To extract IMFs from the signal $X(t)$, all local extrema (minima and maxima) should be found first. Then we should create an upper envelope $e_u(t)$ by local maxima and a lower envelope $e_l(t)$ by local minima.

Envelopes are built by cube-spline interpolation. Using the upper and lower envelopes, the mean $m(t)$ is calculated as in (3.1).

$$m(t) = \frac{e_u(t) + e_l(t)}{2} \qquad (3.1)$$

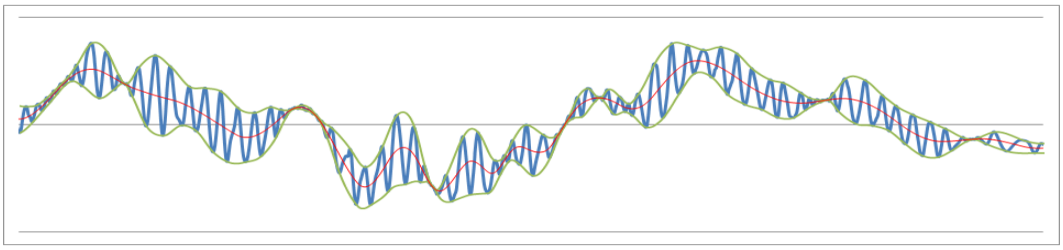


Figure 3.1. Signal (blue), its envelopes (green) and mean (red) by envelopes.

The result is shown in Figure 3.1. The difference between the data and $m(t)$ is the first component $h_1(t)$, which represents *proto IMF*. An IMF is defined as a function that satisfies two requirements:

1) the number of extrema and the number of zero-crossings must either be equal or differ at most by one,
2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

Until $h_1(t)$ does not satisfy the definition of the IMF mentioned above, it should be iteratively refined using the same procedure. Thereby for $h_1(t)$ we get next component $h_2(t)$ and then $h_3(t)$ and so on until stop criteria (3.2) becomes true, where ε is a small number. In this work ε was set to *0.0001*.

$$\frac{\sum_t (h_k(t) - h_{k-1}(t))^2}{\sum_t (h_{k-1}(t))^2} < \varepsilon \qquad (3.2)$$

After repeated refinement up to k times, $h_k(t)$ becomes the first IMF of the signal, called $c_1(t)$. Figure 3.2 shows the first IMF obtained from the data in the Figure 3.1.

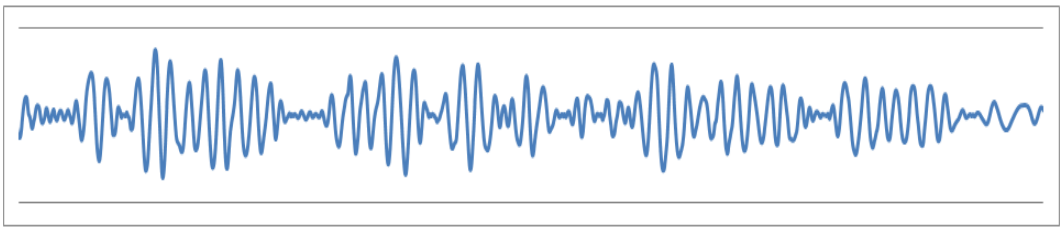


Figure 3.2. First IMF obtained from the signal.

After we obtain the first IMF, we can get the residue $r(t)$ by subtracting $c_1(t)$ from initial data:

$$r(t) = X(t) - c_1(t) \qquad (3.3)$$

The residue of initial signal from the Figure 3.1 is shown in the Figure 3.3.

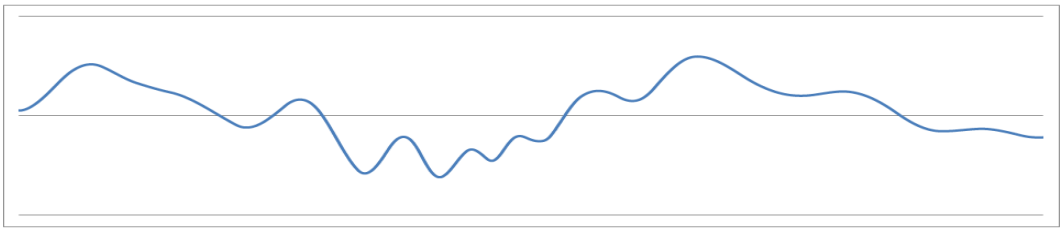


Figure 3.3. Residue after subtracting first IMF c1.

In the next round of the sifting process the residue $r(t)$ is considered as a signal $X(t)$ and the sifting procedure is repeated the same way to obtain $c_2(t)$, then $c_3(t)$, and so on until residue becomes a monotonic function without extrema. When we sum all obtained IMFs with the last residue, we get initial data signal as follows:

$$X(t) = \sum_{i=1}^{n} c_i + r_n \qquad (3.4)$$

The good feature of such decomposition is that each IMF represents an intrinsic component of the real physical effect.

Figure 3.4 shows the original signal and IMFs obtained by means of EMD.

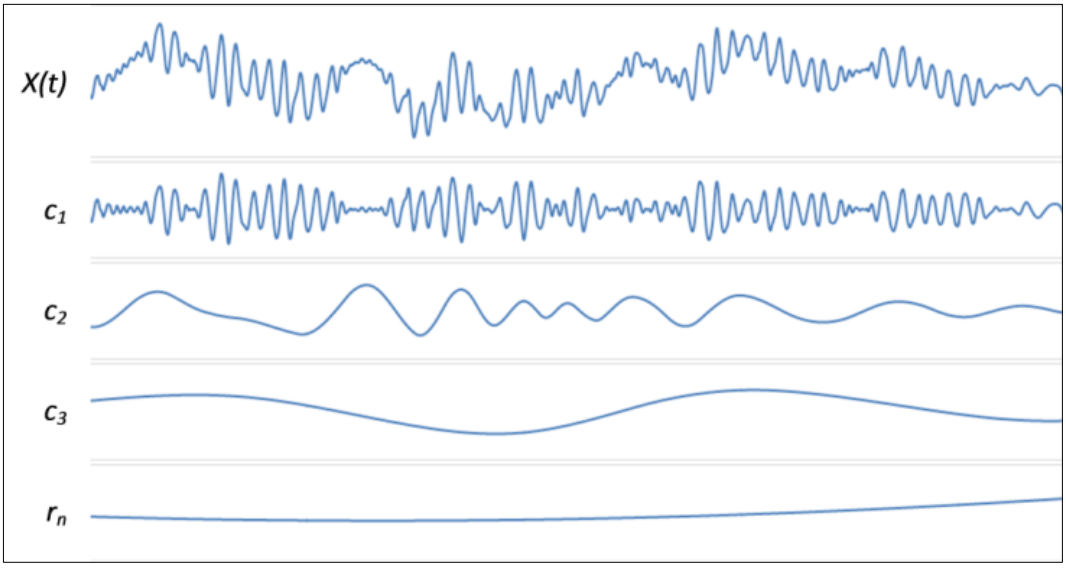


Figure 3.4. The resulting empirical mode decomposition components from the music data: the original data $X(t)$ and the components $c1 - c3$; r_n is a trend.

Listing 3.1 shows the MatLab code which takes the signal as an input and decomposes it into a set of intrinsic function in an array.

Listing 3.1. Empirical Mode Decomposition code using MatLab.

```
function imf = emd(x, maxImfs)

    x = transpose(x(:));
    imf = [];
    while ~ismonotonic(x) && length(imf) < maxImfs
       x1 = x;
       sd = Inf;
       n = 1;
       while (sd > 0.0001) && (n < 30)
          [s1, s2] = GetEnvelopes(x1);
          x2 = x1 - (s1 + s2) / 2;

          sd = sum((x1 - x2) .^ 2) / sum(x1 .^ 2);
          x1 = x2;
          n = n + 1;
       end

       imf{end+1} = x1;
       x = x - x1;
    end
    imf{end+1} = x; % residue
end

function u = ismonotonic(x)
    u1 = length(findpeaks(x)) * length(findpeaks(-x));
    if u1 > 0,
        u = 0;
    else
        u = 1;
    end
end

function u = isimf(x)
    N  = length(x);
    u1 = sum(x(1:N-1) .* x(2:N) < 0);
    u2 = length(findpeaks(x)) + length(findpeaks(-x));
    if abs(u1 - u2) > 1,
        u = 0;
    else
        u = 1;
    end
end

function [lowerEnvelope, upperEnvelope] = GetEnvelopes(data)
    N = length(data);
    [ymax, imax, ymin, imin] = extrema(data);
    upperEnvelope = spline([0 imax N + 1], [0 ymax 0], 1 : N);
    lowerEnvelope = spline([0 imin N + 1], [0 ymin 0], 1 : N);
end

function s = getspline(x)
    N = length(x);
    [pks, locs] = findpeaks(x);
    s = spline([0 locs N + 1], [0 pks 0], 1 : N);
end
```

Hilbert Transform

The Hilbert transform can be interpreted as a phase shifter, which changes the phase of all frequency components of a signal to π/2. To shift a phase, the initial signal is processed with a Fourier transform and then every component of the resultant spectrum is multiplied by imaginary *i* and the spectrum is converted back to signal using the inverse Fourier transform. An example of the original signal and derived signal with shifted phase are shown in Figure 3.5.

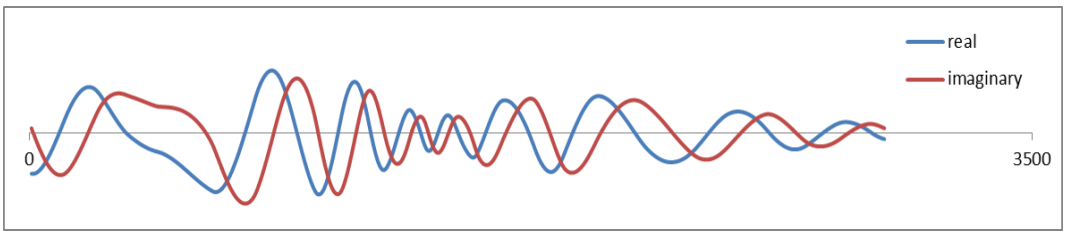


Figure 3.5. Initial Signal (blue) and obtained imaginary signal after HT (red).

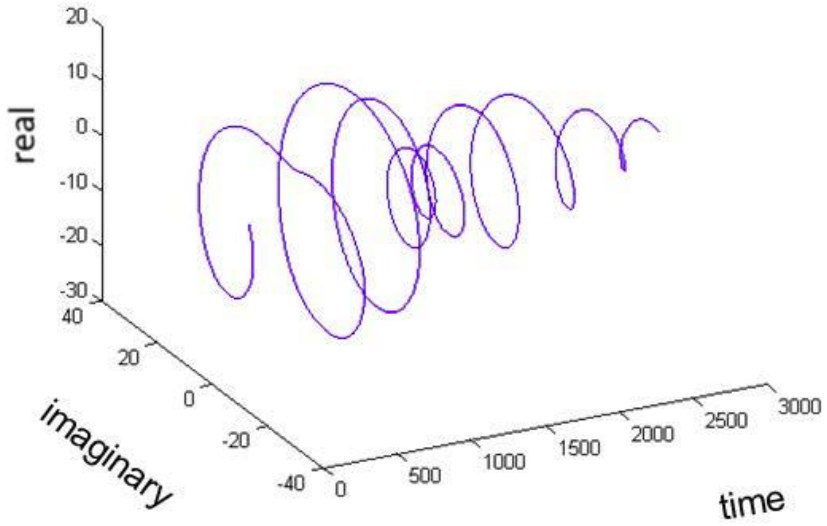


Figure 3.6. Complex analytical signal *H(t)*.

The imaginary signal $\widetilde{\mathrm{X}}(t)$ is orthogonal to original signal $\mathrm{X}(t)$. This feature allows us to develop from $\widetilde{\mathrm{X}}(t)$ and $\mathrm{X}(t)$ a complex analytical signal *H(t)*:

$$H(t) = \mathrm{X}(t) + i\widetilde{\mathrm{X}}(t) \qquad (3.5)$$

H(t) is described as a vector on the complex plane where $\mathrm{X}(t)$ and $\widetilde{\mathrm{X}}(t)$ are projections to real and imaginary axes, respectively.

The advantage of this representation is that we have an opportunity to determine instantaneous parameters of the signal *H(t)*, i.e., the amplitude and frequency, where

the radius of each circle represents the amplitude and the space between circles means the frequency.

Instantaneous amplitude is calculated as complex number length:

$$A(t) = \sqrt{(realH(t))^2 + (imagH(t))^2} \qquad (3.6)$$

Instantaneous frequency is calculated as instantaneous phase derivative of a signal:

$$F(t) = \frac{1}{2\pi}\varphi'(t), \qquad (3.7)$$

where phase φ is calculated as

$$\varphi(t) = \tan^{-1}\frac{realH(t)}{imagH(t)} \qquad (3.8)$$

Instantaneous Frequency Spectrum

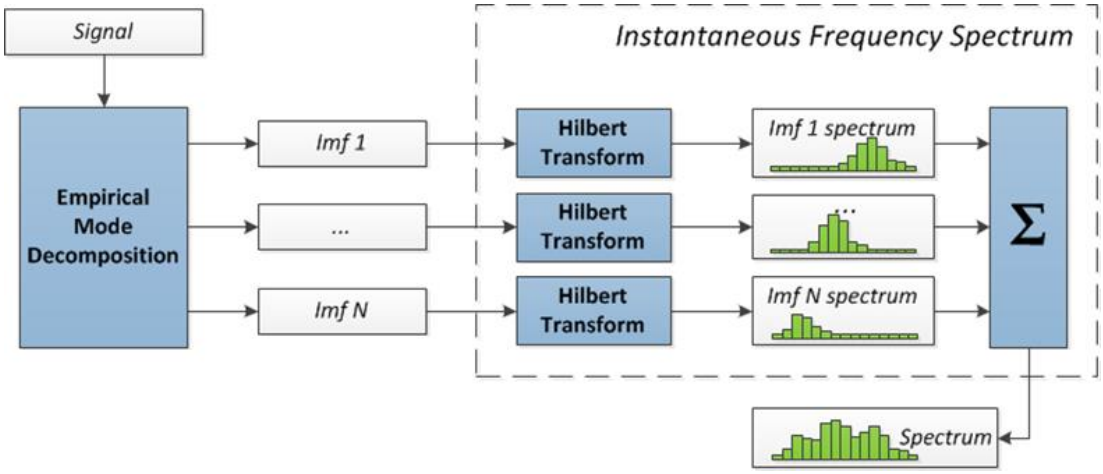


Figure 3.7. Scheme for calculating the IFS.

The IFS calculation method is outlined in the Figure 3.7. White boxes show data and blue boxes show processing steps. As inputs, we use a number of IMFs that represent intrinsic functions of the same signal. As an output, we get a histogram of amplitudes by frequencies. A histogram for every IMF is calculated in the same way. For each IMF, we get instantaneous frequencies and instantaneous amplitudes using the Hilbert transform as described before. These frequencies and amplitudes are used to create a histogram. Formally, this is described as shown in (3.9) and (3.10).

$$b_i = \sum_t A(t), \qquad t: \beta(i-1) \le \log F(t) \le \beta(i), \qquad i = \overline{1,N} \tag{3.9}$$

$$\beta(i) \stackrel{\text{def}}{=} \frac{i}{N} \log F_{max} \tag{3.10}$$

where

b_i is height of i-th bar of the histogram,

$A(t)$ is an instantaneous amplitude at time t,

$F(t)$ is an instantaneous frequency at time t,

$\beta(i)$ is a frequency upper boundary for i-th histogram bar,

N is a number of bars in the histogram,

F_{max} is maximal frequency.

Listing 3.2 shows the MatLab code which calculates Instantaneous Frequency Spectrum for a signal. The code takes the signal data itself, some other settings (such as quantification frequency, size of spectrum, range between minimal and maximal frequencies which is included to the spectrum), and returns the IFS spectrum as a result. The set of spectrums will be included in the result if the sliding window approach is desired.

Listing 3.2. Instantaneous Frequency Spectrum calculation code using MatLab.

```
function result = InstSpectrumTime(data, Fs, winSize,
                  spectrumSize, freqMin, freqMax)

    result = [];
    scale = spectrumSize / log(freqMax);

    % Using EMD signal is split into intrinsic mode functions (imfs)
    imfs = emd(data, 10);
    len = length(data);

    imfCount = length(imfs);
    for n = 1 : imfCount
        currentImf = imfs{n};

        % Hilbert Transform
        imfh = hilbert(currentImf);
        % Instantaneous amplitude
        instAmp{n} = abs(imfh);
        % Instantaneous frequency
        instFrq{n} = max([diff(unwrap(angle(imfh))),0],
                         zeros(1,length(imfh))) * Fs / (2 * pi);
    end
```

```
    for t = 1 : length(instFrq{1})
        ifs = zeros(1, spectrumSize);
        for n = 1 : imfCount
            f = instFrq{n}(t);
            if freqMin < f && f < freqMax
                logF = floor(log(f) * scale);
                if logF > 0
                    ifs(logF) = ifs(logF) + instAmp{n}(t);
                end
            end
        end

        index = floor(t / winSize) + 1;
        if index > length(result)
            result{index} = zeros(1, spectrumSize);
        end

        result{index} = result{index} + ifs;
    end

    maxIndex = floor(length(instFrq{1}) / winSize);
    result = result(1 : maxIndex);
end
```

Processing results

Following is an example of processing result of the tune "Jennifer Lopez – On The Floor", with highlighted fragment from the beginning till 120th second. The tune was processed with sliding window of 1 second, and each frame was processed by IFS. The result is visualized on the Figure 3.8 where horizontal axis represents time, but vertically is presented spectrum frequencies. Low values are shown in blue, higher values are shown through green and yellow till red for the greatest values of the spectrum.

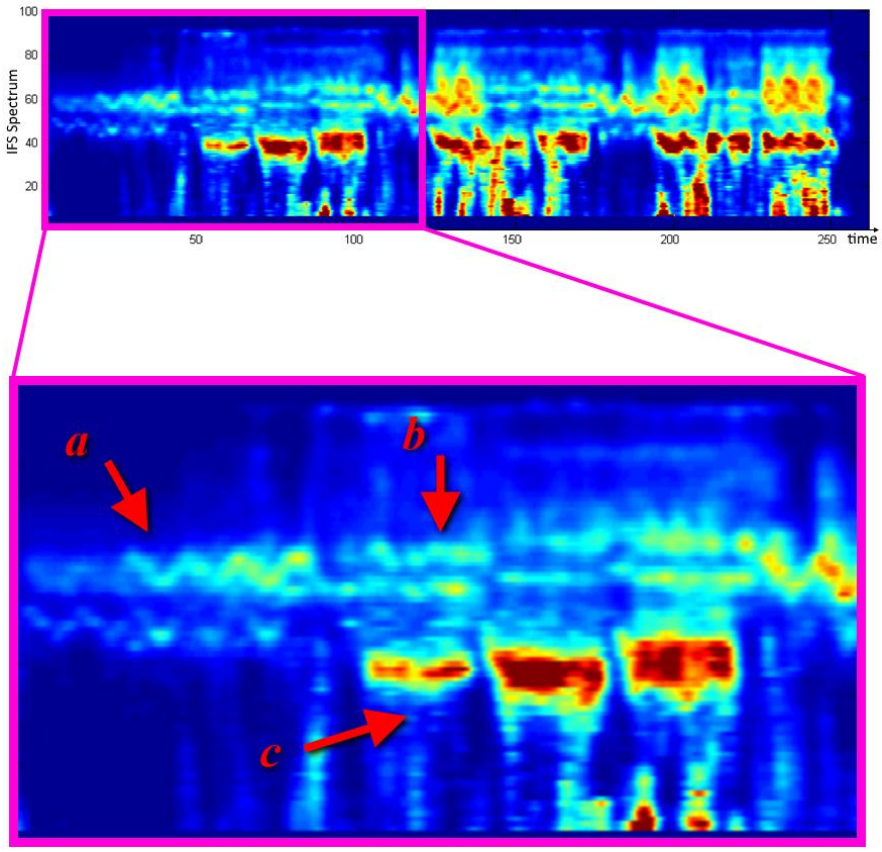


Figure 3.8. Example of full tune processing result (above). Revealed internal structures: a) introduction fade-in, b) two voices in parallel, c) periodic parts of rhythmic music.

The considered example reveals some internal structures of the processed tune. We can see that the tune starts quietly with a narrow spectrum, and then slowly fades in to a louder part. It is shown with letter "a" in the picture. This part of the tune is different and can be called as an introduction. Following part of the tune introduces low frequency components which make the tune rhythmical. It is shown on the picture with letter "c". At the same time we can see that the tune contains two streams of voices in parallel with frequencies nearby. This is shown on the picture with letter "b".

This example shows and proves the value of IFS spectrum and its applicability for good way of describing signal and revealing internal structures of it.

3.1.2. Music Homogeneity Analysis

Another approach I introduced is processing and analyzing the structure of musical tunes. In that method I propose a new innovative technique which summarizes signal of a whole tune and detects regularity within the data. The visualization of the result gives a very good picture of the tune organization, and presence of untypical parts. The

approach is generalized to be used on any kind of sequential data such as non-periodic and non-stationary signals.

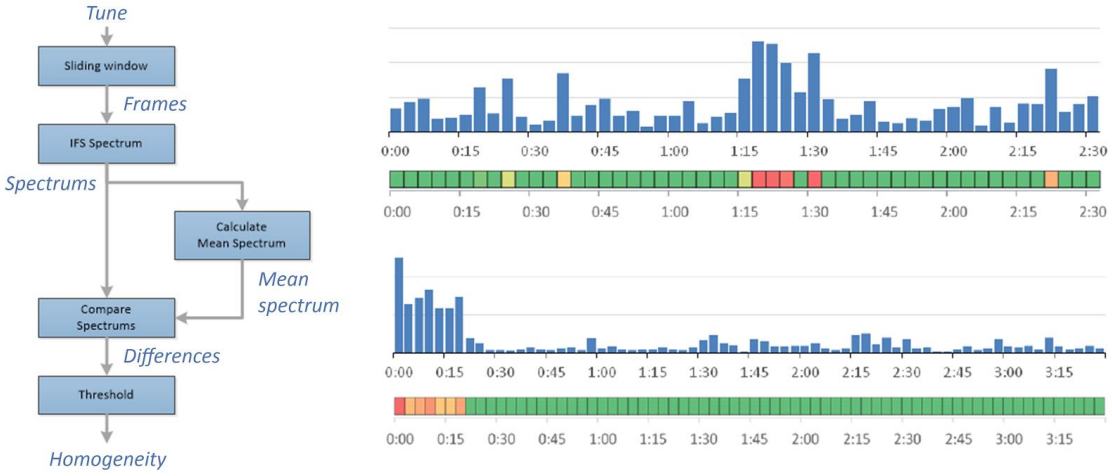


Figure 3.9. Tune homogeneity algorithm steps (left) and two samples' results in time (right).

Tune homogeneity detection can be described as following steps.

Step 1. Split tune into N frames a_i of T seconds size:

$a_i,\ i = \overline{1,N}$.

Step 2. Calculate IFS spectrum s_i for every frame a_i: $s_i = \mathrm{IFS}(a_i)$, $i = \overline{1,N}$.

Step 3. Calculate average spectrum s_a of all spectrums s_i in (3.11).

$$s_a(k) = \frac{1}{N}\sum_{i=1}^{N} s_i(k), \qquad k = \overline{1,M} \qquad (3.11)$$

where k is a spectrum bar number, M is a spectrum bars count.

Step 4. Calculate differences $D = \{ d_i \}$ for every frame spectrum s_i with overall spectrum s_a according to (3.12).

$$D \stackrel{\text{def}}{=} \{d_i = dist(s_i, s_a), \qquad i = \overline{1,N}\} \qquad (3.12)$$

Spectrums difference in step 4 is calculated as follows as shown in (3.13), where x, y are IFS spectrums. $x(k)$, $y(k)$ are k-th frequency range value in IFS spectrum, M is number of frequency ranges in x and y.

$$dist(x,y) = \sum_{k=1}^{M}(x(k) - y(k))^2 \qquad (3.13)$$

After described 4 steps we get differences $D = \{ d_i \}$. D is a first approximation of homogeneity result and it can be further refined.

Listing 3.3. Homogeneity calculation code using MatLab.

```
function result = Homogeneity(sp)
    mean = AverageSpectrum(sp);

    N = length(sp);
    result = zeros(1, N);
    for i = 1 : N
        result(i) = CompareSpectrums(sp{i}(30:100), mean(30:100));
    end
end

function result = AverageSpectrum(sp)
    sum = zeros(1, length(sp{1}));
    N = length(sp);
    for i = 1 : N
       sum = sum + sp{i};
    end
    result = sum / N;
end

function result = CompareSpectrums(ifs1, ifs2)
    result = sum((ifs1 - ifs2).^2);
end
```

Listing 3.3 shows the first approximation of homogeneity calculation using MatLab. The function `Homogeneity(sp)` takes an argument `sp` that contains an array of spectrums. The calculation uses two other functions `AverageSpectrum(sp)` and `CompareSpectrums(ifs1, ifs2)` which calculate the combined average of given spectrums array, and other one compares two IFS spectrums and returns sum of squared differences.

Iterative refinement

First homogeneity approximation D is calculated from average spectrum s_a which contains information not only from homogenous parts, but also from inhomogeneous

parts. Now when we already know which frames are inhomogeneous, we can exclude them from average spectrum s_a to make the result more accurate.

Step 5. Set weight w_i for every frame a_i depending on their homogeneity so that higher distance d_i leads to lower weight w_i as in (3.14).

$$w_i = \frac{\max\limits_{j=\overline{1,N}} d_j - d_i}{\max\limits_{j=\overline{1,N}} d_j - \min\limits_{j=\overline{1,N}} d_j}, \qquad i = \overline{1,N} \qquad (3.14)$$

Step 6. Calculate weighted average spectrum s_{wa} of all spectrums s_i according to (3.15).

$$s_{wa}(k) = \frac{\sum_{i=1}^{N} w_i * s_i(k)}{\sum_{i=1}^{N} w_i}, \qquad k = \overline{1,M} \qquad (3.15)$$

Step 7. Calculate differences $D = \{ d_i \}$ for every frame spectrum s_i with overall weighted average spectrum s_{wa} as in (3.16).

$$D \stackrel{\text{def}}{=} \{ d_i = dist(s_i, s_{wa}), \qquad i = \overline{1,N} \} \qquad (3.16)$$

Step 8. Repeat steps 5, 6, 7 until differences d_i stop changing. Stop criteria can be defined as in (17), where r – iteration number, $d_i^{(r)}$ – difference d_i on iteration r.

$$\frac{\sum_{i=1}^{N} \left(d_i^{(r)} - d_i^{(r-1)} \right)^2}{\sum_{i=1}^{N} \left(d_i^{(r-1)} \right)} < \varepsilon \qquad (3.17)$$

In step 8, ε is a minimal relative change size that can be set to 0.01 meaning 1% change.

Visual presentation. Homogeneity analysis result D can be visualized as a bar chart where horizontal axis represents time, and bars represent homogeneity d_i of the corresponding fragment. An example is shown on Figure 3.10.

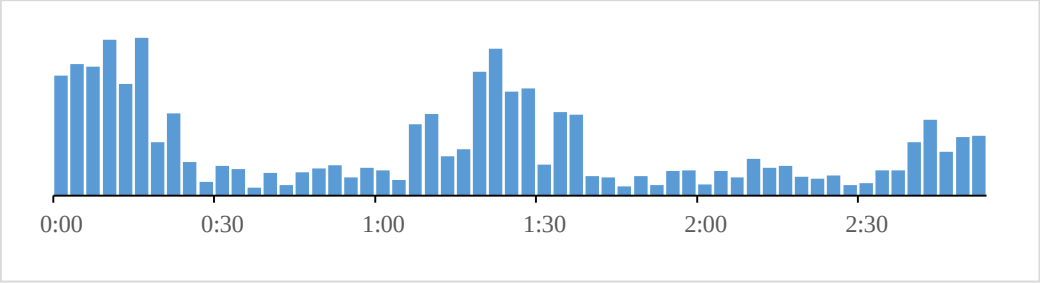


Figure 3.10. Homogeneity result example.

The higher is a bar the greater difference to the tune it represents. On the other hand, shortest bars show most homogeneous parts of the tune.

Tune thumbnail

Tune thumbnail is the most representative part of a tune and it is essential, because by listening to a thumbnail of the tune people can easily recall that tune from their memory and fillings they were experienced during it performance. It senses how our memory is working to memorize some significant parts from the music pieces. So we use tunes thumbnails as a second parameter in the tunes descriptors.

Let's say that we need a thumbnail of length $T_{th} = 30$ seconds. To get most representative 30-seconds part, we use a sliding window on previously calculated differences $D = \{ d_i \}$. For the case when window size $T_{th} = 30$ seconds, and frame size $T = 3$ seconds, sliding window contains $L = T_{th} / T = 10$ items of d_i. Shifting window with step T we calculate sum of items d_i within the window.

Tune fragment, which correspond to the minimal sum value is the most representative fragment for tune thumbnail (3.18).

$$Thmb_{start} = \arg\min_{i} \sum_{j=i}^{i+L} d_j , \qquad i = \overline{1, N-L} \qquad (3.18)$$

where L is a sliding window size, $Thmb_{start}$ is a beginning of the thumbnail.

Tune Homogeneity Pattern

Tune homogeneity pattern is the information about internal structure of a tune, which describes how often tune changes the way of its performance. For example, when internal homogeneity of the music piece changes frequently, listeners may memorize it as highly varying tune, so the emotions they will experience will be exiting and interesting compare to the tunes, which are performed in monotonous way. To make it comparable between tunes we decided to present it as a histogram that shows the frequency of different homogeneity values within tune internal homogeneity result.

In this work we set the homogeneity pattern histogram to consist of 5 bars. Figure 3.11 shows two examples of homogeneity pattern for two different tunes.

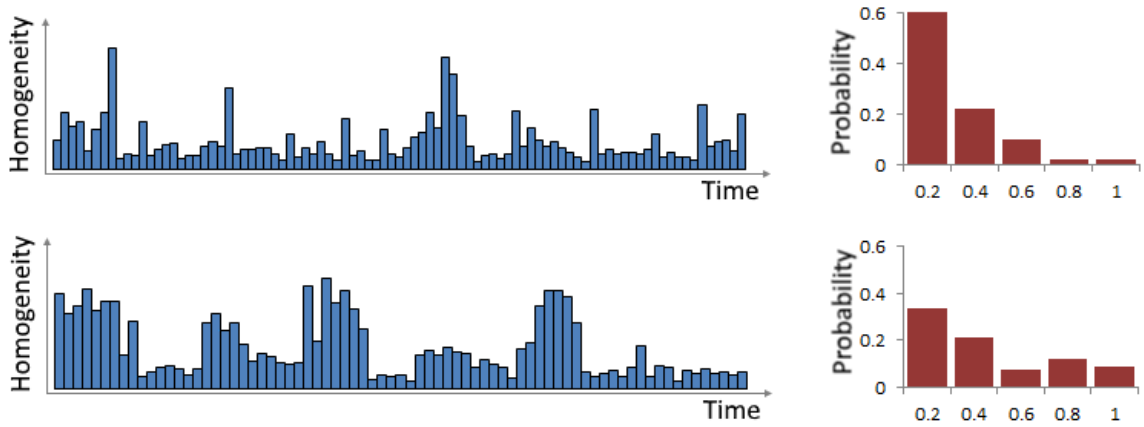


Figure 3.11. Two examples of tunes internal homogeneity (blue) and their homogeneity patterns (red).

For the tune above in Figure 3.11, we can see that most homogeneity bars (blue) are short. That fact can be seen on corresponding homogeneity pattern histogram (red): the probability of low values is very high. The tune below in Figure 3.11 has less homogeneous parts compared to the previous tune. Its homogeneity pattern looks a bit different and has higher probability of greater values (red).

3.1.3. Music Repetitions Detection

The most representative part in a piece has been defined as the most frequently repeated component in it such as repetitions and typical parts during tune's performance. Generally speaking, the more repetitions and similar phases there are in a piece of music, the easier it is for people to have affinity for it [7]. Repetitions have found by performing self-similarity calculations with Mel-Frequency Cepstral Coefficients (MFCCs) [15].

As a part of this research I was working on a method to identify and visualize repetitive structures. The proposed algorithm is able to identify repetitive structures in a tune by using a self-similarity matrix and a sliding two-dimensional window that recognizes and emphasizes the fragments which sound similar. The proposed approach is capable to detect not only identical phrases in a tune, but also those that are slightly different but still sound similar to a listener.

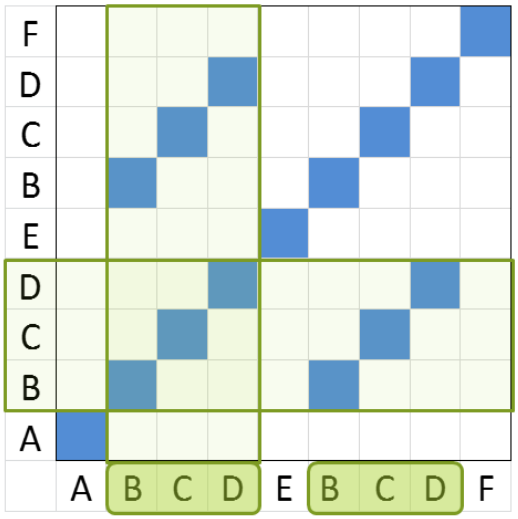


Figure 3.12. The idea of repetitions in similarity matrix: "BCD" fragment is repeated twice and it appears as diagonal line.

The more repetitions and similar phases there are in a piece of music, the easier it is for people to have affinity for it. Since repeated parts are very important, we identify repetitive structures in a tune by using a self-similarity matrix. The most outstanding repetitions of a tune are found to calculate their IFS spectrums for including this information into a tune descriptor. The approach of finding repetitive structures within music pieces was initially introduced in [9].

Finding Repetitions

The tune is divided into frames of size 0.7 seconds and with a shift size 0.1 second. Each of these frames is converted using MFCC and used to calculate distance to each other. A distance equal to 0 means that the fragments are identical; the greater the value is, the less similar the fragments are. From these values we get a self-similarity matrix showing the results of pair-wise comparisons of all frames. This matrix can be represented in the form of a grayscale image where each number is represented by brightness. Black points in the image shows the value of 0. Further processing is based on the data from this matrix.

Listing 3.4. Self-similarity matrix calculation code using MatLab.

```
function result = DistanceMatrix(mfccArray)
    N = length(mfccArray);
    result = zeros(N, N);
    for i = 1 : N
        for j = i + 1 : N
            result(i, j) = CompareSpectrums( mfccArray{i},
                                mfccArray{j} );
            result(j, i) = result(i, j);
        end
    end
end
```

In the self-similarity matrix, similar parts in a tune appear in the form of dark lines. We take a small window and consider only the values within it. The distances at the points marked with letter A should be significantly smaller than those at the points marked with letter B. If this is true, considered area would look like a line, so points marked with letter A are a part of a repetition, see Figure 3.13. To avoid division by zero, a very small constant ε is added to the numerator and denominator (about 1×10^{-5}). Thus, the lower the value we get using this equation, the more desired the line is.

$$f = \frac{A_1 + 2 * A_2 + A_3 + \varepsilon}{B_1 + B_2 + B_3 + B_4 + \varepsilon} \tag{3.19}$$

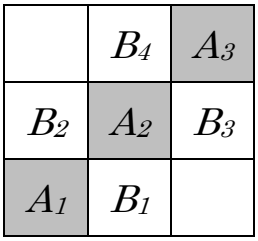
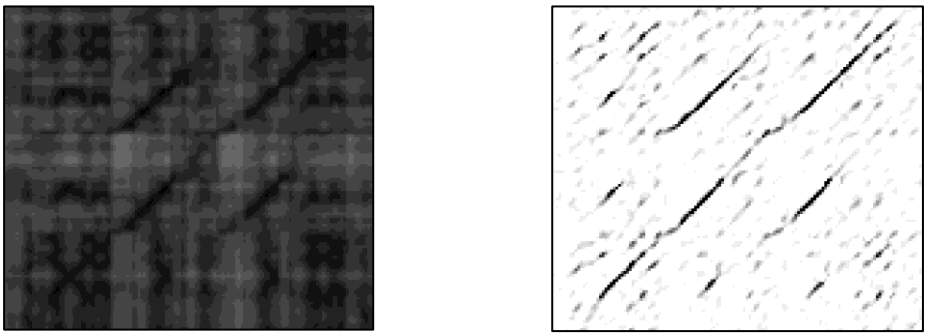


Figure 3.13. Window of size 3 × 3.

According to (3.19), if the values at points marked with the letter A are less than the values at points B, the result is less than 1. Thus, if the result is in the range from 0 to 1, it is a potential line. Values closer to 0 indicate a stronger line, and values closer to 1 show a weakly visible line. We are not interested in values greater than 1, so they can simply be replaced by 1, where 1 means the situation when the sum of A-points is approximately equal to the sum of B-points, meaning that considered area points inside

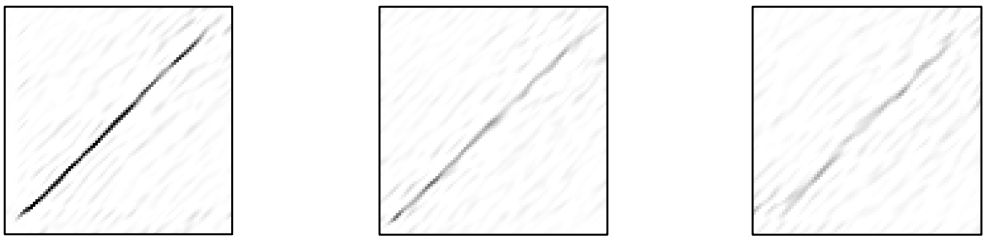
the window does not look like a line. After this process, places with lines are strongly emphasized and it is easy to see them. An example is shown on Figure 3.14 b.



a) part of similarity matrix b) lines detection result

Figure 3.14. Line parts detection results.

After line detection with a 3 by 3 window, we can better see the lines we are about to detect. These lines vary in brightness (See Figure 3.15). Darker lines indicate a strong similarity between corresponding musical fragments. Lighter gray lines indicate less accurate repetition, which may be because of differences in tone, in performance, or because of different musical instruments.



a) very similar b) medium c) not very similar

Figure 3.15. Different similarity level comparison.

When there is at least a light line of repetition that can be seen in an image, the listener should be able notice similarity between these fragments.

Finally, to detect lines we iterate through all closed sets of dark points and search for coordinates of the bottom left point and the top right point within every set. These are the boundaries of lines – the result of the line search in the image. When a line with coordinates (x_1, y_1) - (x_2, y_2) is found, this means that the fragment $(x_1 - x_2)$ is similar to the fragment $(y_1 - y_2)$ within this piece of music. After selecting the most significant lines from all found lines, we get a picture such as shown in Figure 3.16.

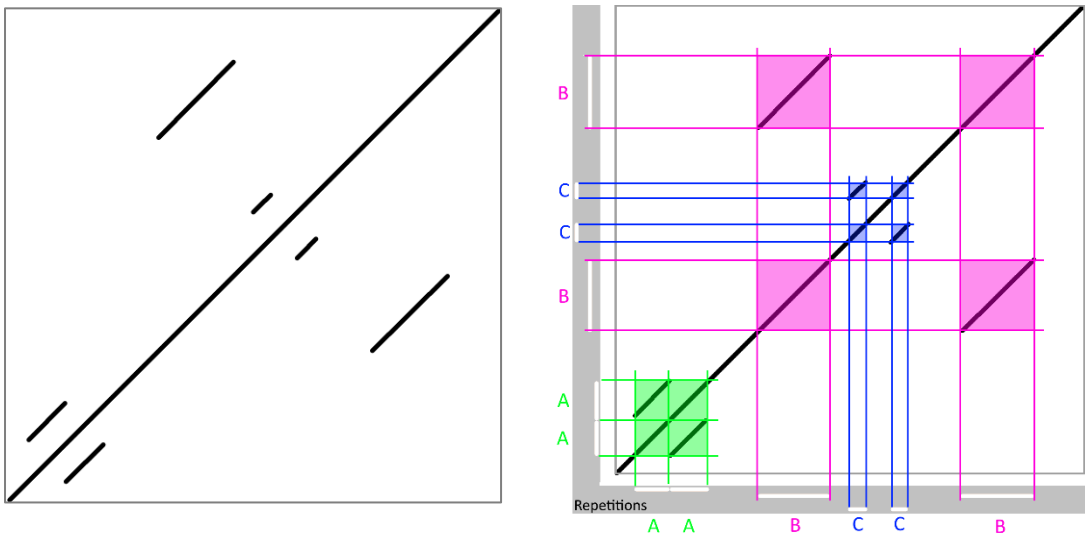


Figure 3.16. Example of most significant repetitions found.

Figure 3.17. Final result of identified repetitive structures.

3.1.4. Music Similarity Analysis

A new approach for tunes similarity calculation based on repetitions was proposed as part of this research. Information about repetitions in tunes is important since repetitions make very significant impression to a listener. I am providing a way to describe tunes in descriptors which contain frequency information about repetitions of tunes. Frequency information is retrieved by means of a signal processing approach IFS. Further tunes comparison takes this repetitions frequencies information from one tune descriptor and compares to the second tune's descriptor. Significant result is that similarity between compared different tunes can be obtained.

To compare two tunes we have to compare every repetition of first tune to every repetition of second tune. Those pairwise repetitions comparison results have to be considered to calculate the tunes difference. Schematically it is displayed on Figure 3.18.

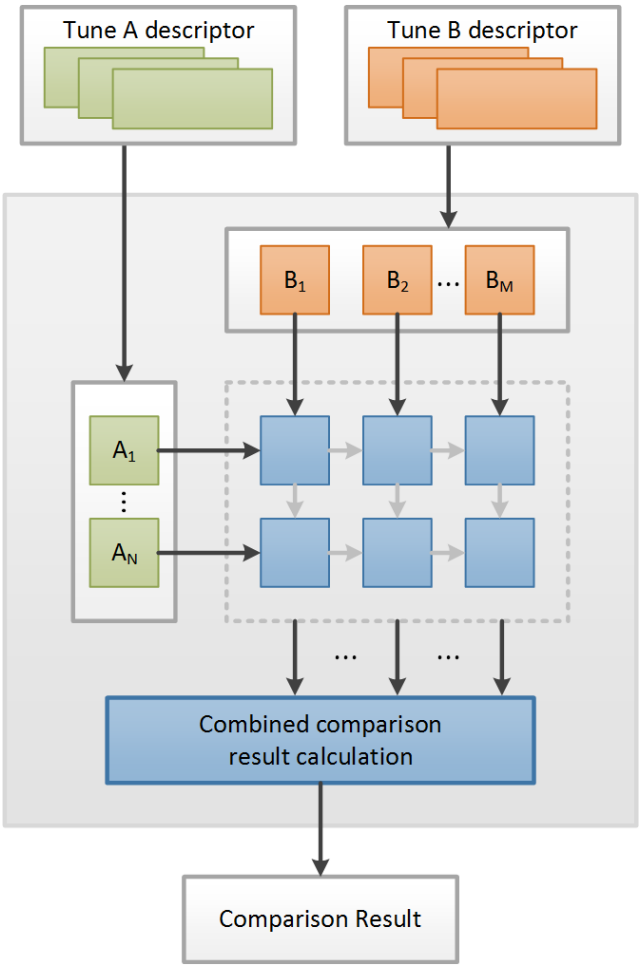


Figure 3.18. Tunes comparison method

There are different approaches to combine multiple results from repetitions comparison into one single number for tunes difference. We give three approaches how it can be done. Hereafter we reference them as AVG, MIN and AVG_k.

AVG approach requires taking average of all repetition comparison as shown in (3.20).

$$AVG(A,B) = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M} comp(A_i, B_j) \qquad (3.20)$$

where $comp(A_i,B_j)$ is a comparison result for i-th repetition from tune A and j-th repetition from tune B.

MIN approach requires taking the smallest value, which means that tunes become as different as their most similar repetitions difference is. Equation (3.21) shows the way how to calculate it.

$$MIN(A,B) = \min_{\substack{i=1..N \\ j=1..M}} comp(A_i, B_j) \qquad (3.21)$$

AVG$_k$ approach requires taking average of k smallest values from all repetition pairs comparisons. k should be calculated from tunes repetitions count as in (3.22).

$$AVG_k(A,B) = \frac{1}{k}\sum_{p=1}^{k} ord\{x, x = comp(A_i,B_j), i = \overline{1,N}, j = \overline{1,M}\}_p, \qquad k = int\left(\sqrt{N * M}\right) \qquad (3.22)$$

where k is the number of elements we take into account; $ord\{x\}_p$ is a p-th element of set of x elements ordered in ascending order. In other words, we take k minimal elements from all items.

Two different tunes comparison

Here we consider example of two tunes comparison: Tune A and Tune B. Both have three repetitions. Let's assume that we compared all repetitions of those tunes. Figure 3.19 shows the results of those comparisons.

		Tune A		
		Rep 1	Rep 2	Rep 3
Tune B	Rep 1	3.0	1.0	2.0
	Rep 2	2.5	2.0	3.5
	Rep 3	2.0	1.0	1.5

Figure 3.19. Two different tunes comparison example

Table 3.1. Methods results comparison for different tunes

Type	Result	Comment
AVG	2.06	All information is taken into account
MIN	1.00	Too few information is considered
AVG$_k$	1.17	Part of information is considered

In Table 3.1 three methods results are represented. For different tunes comparison AVG method looks as the best because it uses all of the repetitions data to give the overall tunes similarity result. MIN approach for the same data gives the minimal value, which means that only one repetitions pair comparison is used, even when there are lots of repetitions in tunes being compared, it is too few to give a comprehansive result. AVG$_k$ approach uses more than one pair of repetitions comparison results, but not all of them.

Tune comparison to itself

Here we consider another example when tune A is compared to itself. Figure 3.20 shows the results of those comparisons. In Table 3.2 three methods results comparison are represented.

		Tune A		
		Rep 1	Rep 2	Rep 3
Tune A	Rep 1	0.0	3.0	2.0
	Rep 2	3.0	0.0	4.0
	Rep 3	2.0	4.0	0.0

Figure 3.20. Tune comparison to itself example

Table 3.2. Methods results comparison for one tune

Type	Result	Comment
AVG	2.00	INACCEPTABLE
MIN	0.00	OK
AVG_k	0.00	OK

When tune is compared to itself, AVG method gives a non-zero result which is inacceptable, because tune is always maximally similar to itself, so the difference must be zero. MIN approach always will give zero value because values on main diagonal always equal to zero. AVG_k approach also always gives zero because k elements that are being taken into account are exactly all zeros from main diagonal.

As a conclusion we can say that AVG_k approach is the best among described three approaches because it behaves as a trade-off, trying to consider as much repetitions comparisons as possible, and not breaking the requirement of zero difference for tune comparison to itself.

Experiments

We selected seven famous classical music pieces for our experiments. Table 3.3 shows the selected tunes and their short names. We use the short names hereafter in this paper. Seven pieces of music were processed with the method described in this paper.

Table 3.3. Target music pieces used in the work

Name	Description
V	Four Seasons: Summer 3rd movement by Vivaldi
B	Air on G string Orchestral Suite No3 in D major by Bach
C	Etude Op. 25 No. 11 by Chopin
S	Gnossienne 4 by Satie
M	Meditation for Thais by Massenet
AH	Bugler's Holiday by Anderson
AP	Plink Plank Plunk by Anderson

Table 3.4 shows the repetitions in the musics scores we used to calculate tunes descriptors for terget music pieces and their performances time in CD. For example, for V tune the number of bars and the performance times of each part are as follows: from the 10th to 17th bar and from the 101th to 108th bar are the same. Performed from 12.5 to 22.0 seconds and from 122.0 to 132.0 seconds accordingly.

Table 3.4. Repetitions in target music pieces

Tune	Nr.	Repetition (bar)	Performance (seconds)
V	1	10 – 17; 101 – 108	12.5 – 22.0; 122.0 – 132.0
B	1	1 – 6; 7 – 12	0 – 43.2; 47.2 – 90.1
	2	13 –24; 25 –36	94.6 – 185.0; 186.5 – 278.3
C	1	5 – 7.5; 13 – 15.5; 69 – 71.5; 77 – 79.5; 23 – 25.5; 31 – 33.5	23.9 – 30.5; 38.6 – 46.5; 151.6 – 159.1; 167.8 – 176.0; 58.5 – 66.5; 75.0 – 83.0
	2	5 – 15.5; 69 – 79.5	23.9 – 46.5; 151.6 – 176.0
S	1	11 – 12; 24 – 25	58.1 – 68.6; 133.4 – 146.7
	2	13; 15	69.2 – 74.5; 80.6 – 85.7
	3	26 – 27; 28 – 29	147.4 – 159.2; 160.0 – 172.7
	4	19 – 20;	102.5 – 114.4;

Tune	Nr.	Repetition (bar)	Performance (seconds)
		31 – 32	180.1 – 195.4
	5	14; 17	74.5 – 79.7; 91.3 – 96.8
	6	18–20; 30–32	97.1 – 113.2; 173.6 – 195.1
M	1	3 – 10; 40 – 47	11.3 – 47.1; 174.5 – 211.1
	2	15 – 20.5; 52 – 57.5	63.9 – 96.7; 229.9 – 258.8
	3	3 – 4; 11 – 12; 40 – 41	11.3 – 19.0; 47.8 – 56.4; 175.1 – 182.1
AH	1	9 – 22.5; 35 – 38.5	5.7 – 16.1; 16.4 – 28
	2	59 – 74; 141 – 156	43.0 – 54.6; 104.0 – 116
	3	75 – 88; 157 – 170	54.8 – 64.8; 116.3 – 126.3
	4	97 – 104; 113 – 120	70.1 – 75.7; 83.0 – 87.9
AP	1	4.5 – 19; 20.5 – 19; 37.5 – 52; 37.5 – 52; 75.5 – 90; 107.5 – 122; 123.5 – 122	3.1 – 15.4; 15.9 – 27.2; 40.1 – 50.7; 62.4 – 72.9; 99.6 – 110.5; 122.0 – 133.4; 133.5 – 143.1
	2	21 – 36; 53 – 36; 91 – 106	27.4 – 38.9; 51.0 – 62.0; 110.9 – 121.7
	3	55 – 70; 55 – 70	74.0 – 85.0; 85.2 – 96.4

Figure 3.21 represents the color scale, where red color means dissimilarity and green color means high similarity. Figure 3.22 shows repetition to repetition comparison results in color.

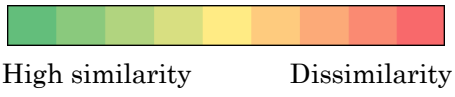


Figure 3.21. Color scale

As expected, the main diagonal is green because every repetition is maximally similar to itself. By solid lines we separate tunes. Dotted lines separate repetitions within tune.

In Figure 3.22 we can see repetitions comparison between tunes, as well as within tunes. We may notice that some music pieces have dissimilar repetitions within. For example, it can be seen inside (6x6) cells rectangle for S tune comparison to itself: there are red and yellow cells. For other tunes comparisons to themselves we mainly see green cells, that means that repetititons within them are very similar.

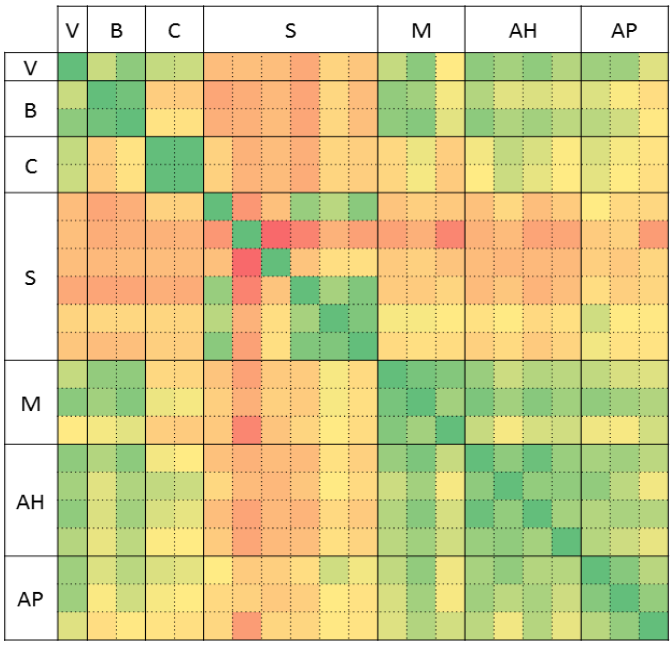


Figure 3.22. Tunes descriptors comparison

According to comparison between two different tunes, we can see various results. Some rectangles mostly consist of green cells meaning that tunes are alike, some consist mostly of yellow and red cells meaning tunes dissimilarity.

Tunes Comparison Results

To get overall result for tunes comarison we consider all cells within every rectangle and calculate a single difference by means of AVG_k method. Those tune to tune similarity results are shown in Table 3.5 and it's colored presentation is shown in Figure 3.23.

Table 3.5. Comparison results

	V	B	C	S	M	AH	AP
V	0.0000	0.0013	0.0029	0.0082	0.0012	0.0013	0.0018
B	0.0013	0.0000	0.0057	0.0080	0.0012	0.0016	0.0029
C	0.0029	0.0057	0.0000	0.0073	0.0042	0.0030	0.0037

	V	B	C	S	M	AH	AP
S	0.0082	0.0080	0.0073	0.0000	0.0049	0.0058	0.0042
M	0.0012	0.0012	0.0042	0.0049	0.0000	0.0012	0.0019
AH	0.0013	0.0016	0.0030	0.0058	0.0012	0.0000	0.0018
AP	0.0018	0.0029	0.0037	0.0042	0.0019	0.0018	0.0000

Main diagonal is green meaning that every tune is similar to itself. Tune S is very different from all other tunes, as we can see by red cells on row and line that correspond to that tune. S tune is very slow and contains many silent fragments and pauses. This fact makes that tune very different from all other tunes.

The same as S, tune C is also somehow different to most of the tunes. This etude is performed by piano and the tempo is very fast. Such piano recording sounds not very similar to orchestral performances as in B and M. But it is somehow similar to V and AH, as they sound similar because of the speed. Tunes comparison values are 0.0029 and 0.0030 correspondigly, and displayed as yellow cells.

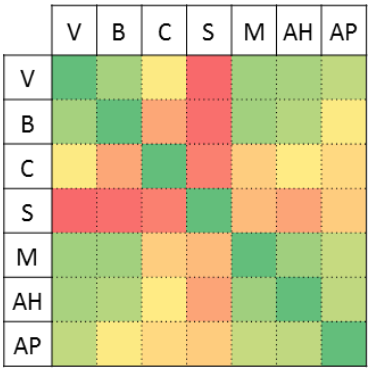


Figure 3.23. Target music analysis similarity result

Music pieces V, B, M, AH, and AP are very similar to each other as we can see from comparison values that are between 0.0012 and 0.0019, displayed as green cells. These tunes sound similar because they are played by orchestra and by quartet, they have similar energetic mood. The only exception is B to AP pair (0.0029) that sounds just somehow similar because of performance mood and small speed difference.

The proposed method for tunes comparison is adaptive to the data, and it is its advantage. The approach shows which tunes are similar and dissimilar among given tunes. In the described example there are only 7 tunes where for example Vivaldi is much more similar to Bach than to Satie.

3.2. Emotions processing

As a base for emotions processing approach eight classes of emotions were used which were proposed by Kate Hevner in [3]. The Hevner's list of emotions is presented in Table 3.6.

Table 3.6. Hevner's categories.

Category	Description
C1	spiritual, lofty, awe-inspiring, dignified, sacred, solemn, sober, serious
C2	pathetic, doleful, sad, mournful, tragic, melancholy, frustrated, depressing, gloomy, heavy, dark
C3	dreamy, yielding, tender, sentimental, longing, yearning, pleading, plaintive
C4	lyrical, leisurely, satisfying, serene, tranquil, quiet, soothing
C5	humorous, playful, whimsical, fanciful, quaint, sprightly, delicate, light, graceful
C6	merry, joyous, gay, happy, cheerful, bright
C7	exhilarated, soaring, triumphant, dramatic, passionate, sensational, agitated, exciting, impetuous, restless
C8	vigorous, robust, emphatic, martial, ponderous, majestic, exalting

To make it simpler to choose, measure and process emotions I transformed the Hevner's adjectives categories into a plane with two axes: activeness and happiness. The choice of axes Activeness and Happiness is coming from the Hevner's categories. Four of the categories can be commonly titled as Happy emotions (C4), Sad emotion (C2), Active emotions (C8), and Passive emotions (C4), while other four categories are somewhere in between. Similar tendency can also be found in subsequent related researches of emotions in music.

When a listener has to provide an emotion, he or she has to put a point on the plane or provide two estimates for two independent axes. Compared to the Hevner's 8 classes the plane-approach has an advantage because (a) it does not require selection of appropriate adjectives, (b) has higher granularity, and (c) provides much better emotions aggregation and processing ability. Emotions presentation in a plane takes into account relation between Hevner's emotion classes the same as in Hevner's ring, but it is beneficial because it can differentiate emotions described by adjectives allocated within one Hevner's category. The emotions plane is presented in the Figure 3.24. The plane has horizontal axis that represents activeness with values range from *-10* to *10*, and

vertical axis that represents happiness with values from *-10* to *10*. Descriptive keywords are shown near all sides of the square region for a better feeling of the emotion space.

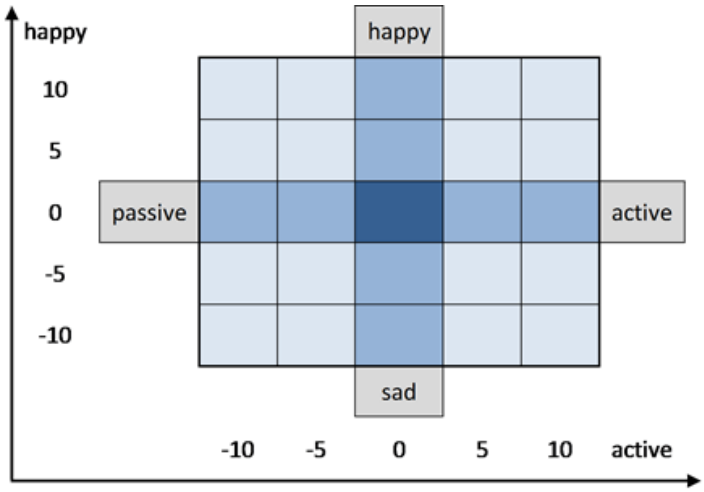
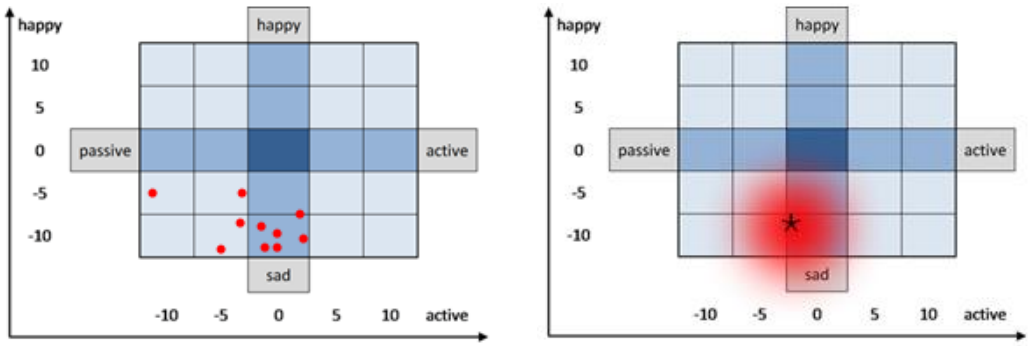


Figure 3.24. Emotions plane for representing emotions.

To provide an emotion person should put a point on the plane that mostly corresponds to their emotions from listening to a tune. The way people express emotions may slightly vary from person to person. This is caused by the fact that every individual is unique and he or she feels emotions subjectively. However, there are emotion adjectives which are very well commonly understood. Besides that, people very well understand the connection between these adjectives giving the appearance of one adjectives relative to others. We place the whole emotion space into a two-dimensional plane, and we assume that the space is linear. As we can see from the experiments results in Chapter 5.4, most of given adjectives were placed on the plane by survey participants in a very consistent way, so the deviations of values by each individual from the common values are not significant. In other words, the experiment showed that the emotion plane is understood and treated consistently by all users.



a) survey respondents answers b) processed emotion result

Figure 3.25. Survey and result for the music piece “A. Winehouse - Back to black”.

In Figure 3.25 is presented example of survey results for determining tune's emotion. Red points in Figure 3.25a represent emotions that respondents feel from listening to the given tune. Figure 3.25b represents an aggregated result for performed tune that we obtained as a mean value, which is presented as an asterisk with value (*-1.8; -7.2*). Variance which is represented as red area is equal to a value of *3.48*, and equal to standard deviation value equal to *1.86*.

Summary and Discussion

Some models of mood were proposed in related researches to represent emotions. Some of them have very low granularity, such as Thayer's model of mood which actually has only 4 possible emotions. Others tend to describe emotion state more precisely, by giving more degrees of freedom and more adjectives to describe them. For example, Tellegen-Watson-Clark model suggests to use 4 non-orthogonal axes to unite 8 groups of adjectives into one model of mood. On the other hand, the proposed emotion plane approach is not limited to a given set of specified adjectives as in other models. Instead, it provides continuous 2-dimensional space which is able to describe not only a finite number of adjectives but also emotions which are hard to explain in words. If the space would be multidimensional with 3 or more dimensions, it would make the model much more complex. Whereas two dimensional space could be imagined, presented and easily understood by humans, and therefore it is simpler to operate with. Simplicity for understanding is a very important factor because the plane is intended to be used by users of the system to specify their emotions.

Chapter 4. Emotion Recognition System Implementation

This chapter describes the design and the overall architecture of the system for automatic emotion identification by means of intrinsic musical features. Its innovativeness is that it uses new musical features for tune's analysis, which are based on human's perception of the music. Therefore the results are more relevant to listener expectations. Three significant approaches are used for calculation of essential features of emotional aspects of music pieces. These are repeated parts of a tune, thumbnail of a tune, and internal homogeneity pattern. At the heart of the described system is the innovative algorithm to process non-periodic and non-stationary signal data for audio signals processing in particular. Moreover, the described system instead of emotions categorizations uses a new way of automatic emotion processing, calculation and presentation on a plane, which has the key idea to quantitatively measure emotions that is very significant for emotions calculation.

4.1. System design

The system for music processing to detect tune's emotion is outlined in Figure 4.1, where white boxes show data and blue boxes show processing steps. Acoustic data is used as an input. Output is a detected tune's emotion. Tunes database store information about every tune and its corresponding emotion that was determined statistically. Tune itself is presented in the database by the tune descriptor, which contains three significant physical parameters of the tunes. Emotion for a tune is stored as a point on the emotional plane. This makes it possible to aggregate emotions from multiple tunes into a one single estimation.

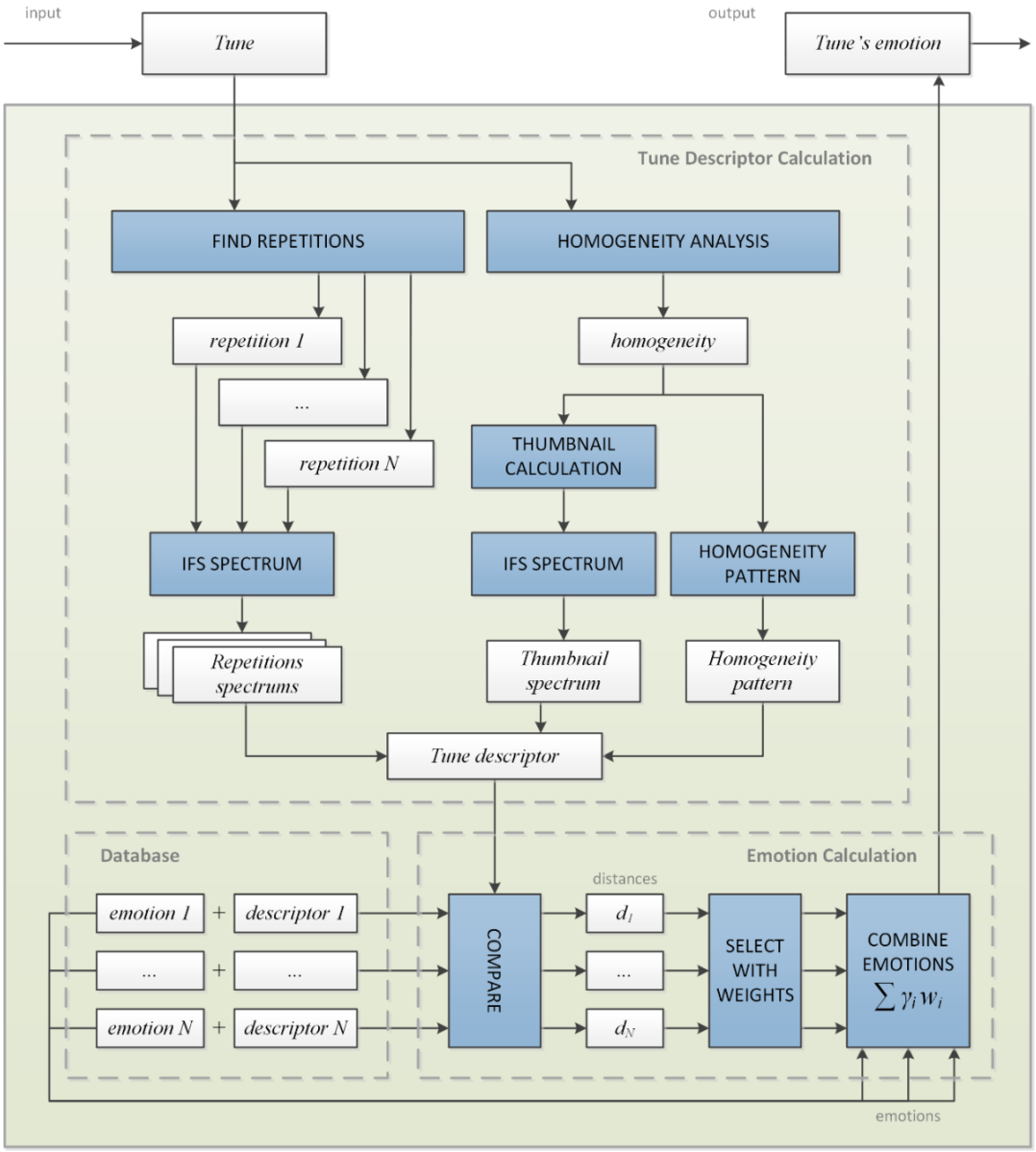


Figure 4.1. System architecture of tune emotion detection system.

Tune emotional effect estimation

Tune processing within the system has two main parts: finding matching tunes in the database and corresponding emotions combining. In the first part we search for multiple matches instead of a one single best match for better result precision since more related records are aggregated to calculate the result. Such approach is helpful for reducing the effect of outstanding values. In the second part we wisely combine emotions from all matches to get the result emotion for the test tune.

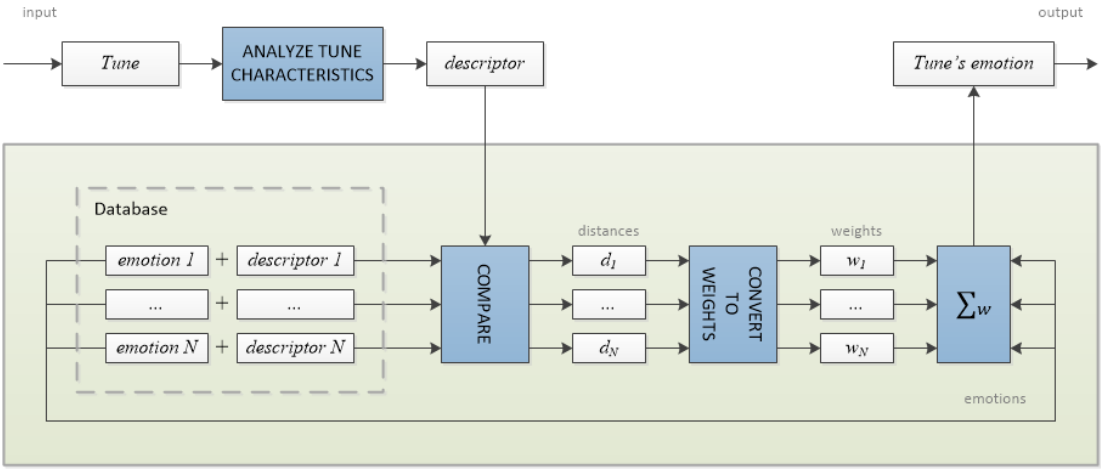


Figure 4.2. System workflow to detect tune's emotion.

Tunes comparison

Tunes matching within the database is a challenging task by itself. The approach we used in this paper is based on the one described in [2]. Every tune is processed to get the descriptor which is like a fingerprint that precisely identifies the characteristics of the tune but at the same time is small and easy to process, compare and store. As described before tune descriptor contains a set of fields: repetitions spectrums, thumbnail spectrum and a homogeneity pattern. Every field is compared separately and has its own weight. Table 4.1 describes descriptor fields, comparing approaches.

Table 4.1. Descriptor fields comparison approaches

	Field	**Comparison approach**
1.	Repetitions spectrums	AVG_k comparison using Music Similarity approach
2.	Thumbnail spectrum	Euclidian distance
3.	Homogeneity pattern	Euclidian distance

Comparison result is a tunes difference that is calculated as a sum of weighted differences from every field comparison as shown in (4.1).

$$d_j = \sum_i w_i * diff(TT.field_i, T_j.field_i) \qquad (4.1)$$

where d_j is a difference between test tune TT and j-th tune from the database, TT is a test tune, T_j is a j-th tune from the database, $TT.field_i$ is a i-th field in the TT tune's descriptors field, w_i is a weight of i-th field.

Matching tunes selection and emotions combining

When a test tune is provided to the database to find matching records, its descriptor is compared to all records in the database. N most matching tunes are selected, i.e. N tunes with the smallest difference values compared to the test tune. Further, matching tunes emotions are combined with respect to the match distance in the way where tunes with smaller distance have higher weight. The emotion is calculated as shown in (4.2).

$$\gamma^* = \frac{\sum_{i=1}^{N} w_i * \gamma_i}{\sum_{i=1}^{N} w_i}, \quad w_j = \left(\frac{\max_i d_i - d_j}{\max_i d_i - \min_i d_i} \right)^2 \tag{4.2}$$

where γ^* is a combined emotion,
γ_i is an i-th record's emotion,
w_i is an i-th record's weight.

4.2. Database

The database stores tunes, descriptors and corresponding emotions. Every tune record in the database contains information about artist and title of the music piece and the tune content. Every tune record has corresponding descriptor information which consists of three parts:

- information about repetitions
- tune thumbnail
- homogeneity pattern

All of these parts are stored in separate tables in the database. Since tune may have many repetitions or may not have repetitions at all, every tune record may have zero to many corresponding records in repetitions table. Every tune has a thumbnail and homogeneity pattern therefore there will always be one thumbnail record and one homogeneity pattern record for each tune.

The information about homogeneity pattern is defined in 5 numerical values which are stored inside the table attributes p1, p2, p3, p4, p5. The tune thumbnail is described as a spectrum, as well as every repetition in tunes are described as signal spectrum. These spectrums data are extracted to a separate table for easy access and comparison between spectrums. The spectrum table stores IFS spectrum information consisting of 100 spectrum bars.

Besides the tune descriptor, database contains emotion information for every tune. It is described in two numerical attributes: activeness axis value and happiness axis value. This data is stored in separate table as well which has a relation to tunes table. Its attributes are following:

- activeness
- happiness
- standard deviation

Standard deviation information represents the statistical parameter which shows how varying were experts estimates for current tune emotion. This information is not directly used by the system when calculating tunes emotion, but it is helpful when analyzing recognition results. For example, when the system uses tunes with relatively big standard deviation, the result may also be imprecise. This is also the case when tunes are used for approach validation.

A separate part of the database is responsible for emotion adjectives mapping. Three tables contain a list of adjectives in different languages and their emotion numerical representations in terms of happiness and activeness. The system is capable of working with emotion adjectives in different languages. What is important, every language has its own set of adjectives and their own mapping to emotions on the emotion plane. Table *Adjective* contains the actual adjective text presentation and a reference to language record in *Language* table. Emotion mapping for adjectives is stored in a separate table *AdjectiveEmotion*. This table has a reference to adjective record and the numerical values for activeness axis value and happiness axis value. Figure 4.3 shows the overall database tables structure.

The same as for *TuneEmotions* table, *AdjectiveEmotion* table also has standard deviation value attribute. The standard deviation of adjectives is taken into account by the system to achieve better accuracy. Those adjectives which have small values of standard deviation were equally mapped by experts to the emotion plane, and therefore such adjectives are commonly well-understood. On the other hand, adjectives with big standard deviation are understood less equal and have some ambiguity. Adjectives with high values of standard deviation should not be taken into account when higher precision of the output is desired, and those adjectives could be filtered out by a threshold. The threshold for standard deviation is a variable configuration of the system that can be adjusted during operation. When it is set to small values such as *2.0* or less, the list of adjectives becomes very short, but those adjectives are very basic like sad,

serious, happy, depressing, and calm. When the threshold is chosen to be equal to *4.0* or higher, many more adjectives become available to describe tunes emotions, however their meaning is less precise. The default threshold in this work was chosen equal to *3.0* which is a compromise between precision and variety of adjectives.

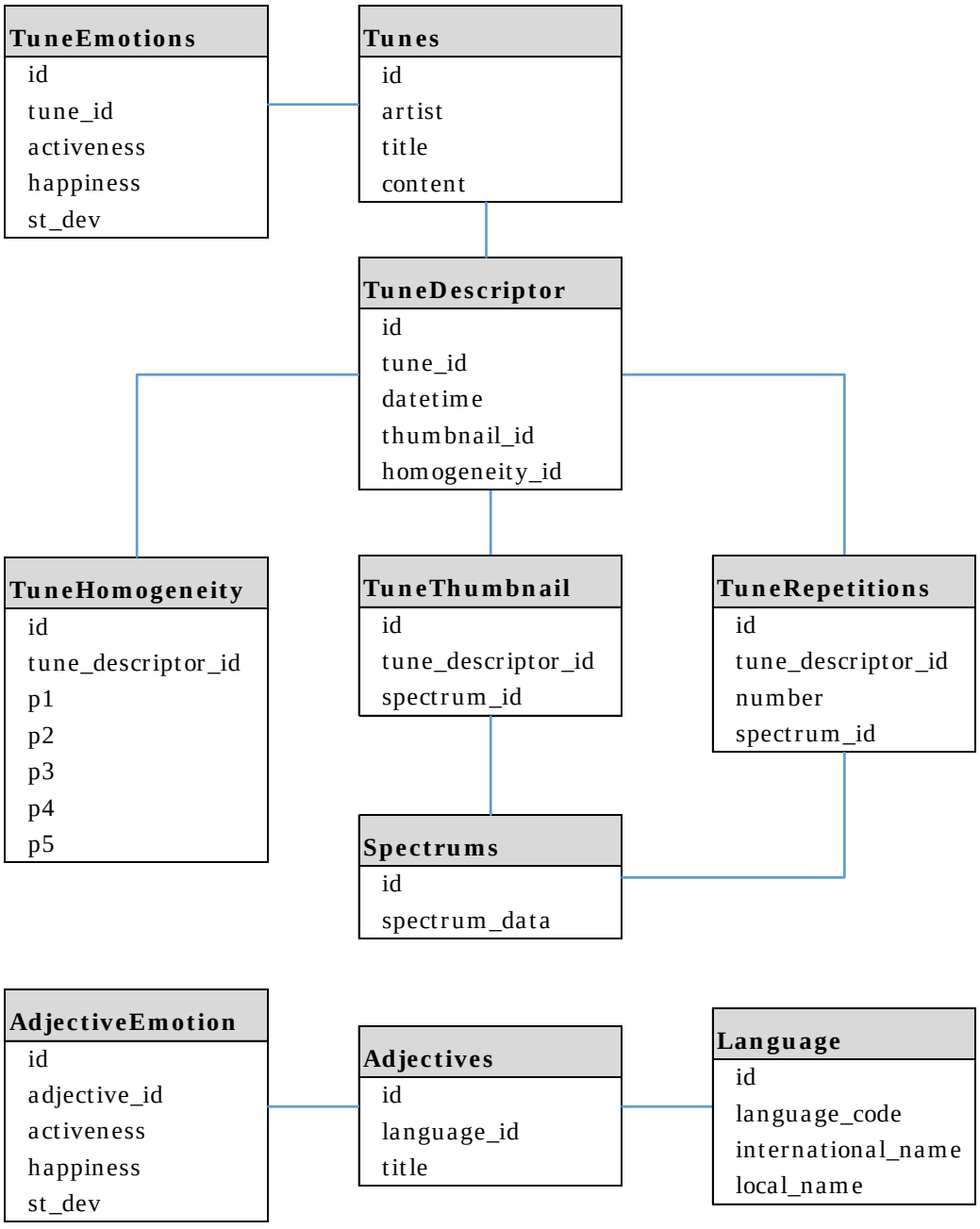


Figure 4.3. Database tables structure.

Chapter 5. Data collection

For the proper operation the system needs to have representative data. Two types of information is required: collection of tunes with known emotions, and mapping of emotion adjectives to the emotion values presented in numbers. Both data types are specific to human perception and therefore the data must be collected using surveys and questionnaires with individuals. This chapter describes the process of data collection and surveys results.

5.1. Tunes database

Tunes database contained tunes from various musical genres as listed in Table 5.1. The full list of used tunes is presented in Table 5.2.

Table 5.1. List of tunes genres used in the system.

Number	Genre name
1	Classical XVII - XX
2	Classical organ
3	Classical orchestra
4	Classical string quartet
5	Classical quartet
6	Classical piano
7	Rock
8	Hard rock
9	Metal
10	Heavy metal
11	Pop rock
12	Funk rock
13	Soft rock
14	Pop
15	Dance
16	Latino
17	Hip-hop
18	Dance-pop
19	RnB
20	Soul
21	Electro-pop

Table 5.2. List of tunes used in the system.

Nr.	Artist	Title
1	Johann Sebastian Bach	Toccata and fugue in D minor
2	Johann Sebastian Bach	Air on the G string
3	Johann Sebastian Bach	Brandenburg Concerto No 1 - Adagio
4	Johann Sebastian Bach	Brandenburg Concerto No 3 - Allegro
5	Johann Sebastian Bach	Brandenburg Concerto No 4 - Andante
6	Johann Sebastian Bach	Concerto D minor BWV 1059 - Movt 2
7	Ludwig van Beethoven	Symphony No 5 - Allegro con brio
8	Ludwig van Beethoven	Symphony No 6 - Allegro
9	Ludwig van Beethoven	Symphony No 7 - Allegretto
10	Ludwig van Beethoven	Symphony No 7 - Presto
11	Ludwig van Beethoven	Symphony No 9 "Ode to Joy"
12	Ludwig van Beethoven	Moonlight Sonata (Sonata al chiaro di luna)
13	Ludwig van Beethoven	Fur Elise
14	Ludwig van Beethoven	Rondo a Capriccio in G Major
15	Ludwig van Beethoven	Minuet in G Major
16	Ludwig van Beethoven	Piano Concerto No 5 Emperor Movt 2
17	Ludwig van Beethoven	Piano Sonata Op. 13 movt. 2 Adagio
18	Wolfgang Amadeus Mozart	Eine Kleine Nachtmusik Movt 1
19	Wolfgang Amadeus Mozart	Symphony No 35 "Haffner Symphony" Movt 1
20	Wolfgang Amadeus Mozart	Symphony No 40 Movt 1 "Molto Allegro"
21	Wolfgang Amadeus Mozart	Magic Flute "Overture"
22	Wolfgang Amadeus Mozart	Marriage of Figaro "Overture"
23	Wolfgang Amadeus Mozart	String Quartet No 23 K 590 Movt 4
24	Wolfgang Amadeus Mozart	Violin Sonata K378 "Rondeau"
25	Wolfgang Amadeus Mozart	Piano Concerto No 21 "Andante"
26	Wolfgang Amadeus Mozart	Turkish March Piano Sonata No 11
27	Wolfgang Amadeus Mozart	Violin Concerto No 3 K 216 Movt 1
28	Wolfgang Amadeus Mozart	Flute Concerto No 2 K 314 Movt 2
29	Wolfgang Amadeus Mozart	Horn Concerto No 3 K 447 Movt 2
30	Frederic Chopin	Etude Opus 25
31	Frederic Chopin	Nocturne Opus 15
32	Frederic Chopin	Nocturne B Flat minor, op 9 No 1
33	Frederic Chopin	Nocturne C minor, op 48 No 1
34	Frederic Chopin	Nocturne C sharp minor, op 27 No 1
35	Frederic Chopin	Nocturne D Flat Major, op 27 No 2
36	Frederic Chopin	Piano Prelude No 4 Opus 28
37	Frederic Chopin	Piano Prelude No 7 Opus 28
38	Frederic Chopin	Polonaise No 1 C sharp minor op 26 No 1
39	Frederic Chopin	Prelude No 18 F minor op 28
40	Frederic Chopin	Prelude No 8 F sharp minor op 28
41	Frederic Chopin	Waltz A Flat Major, op 69 No 1
42	Antonio Vivaldi	SPRING Concerto No.1 in E, Op.8 - Allegro
43	Antonio Vivaldi	SPRING Concerto No.1 in E, Op.8 - Largo
44	Antonio Vivaldi	SPRING Concerto No.1 in E, Op.8 - Allegro Pastorale
45	Antonio Vivaldi	SUMMER Concerto No.2 in g, Op.8 - Allegro
46	Antonio Vivaldi	SUMMER Concerto No.2 in g, Op.8 - Adagio e piano
47	Antonio Vivaldi	SUMMER Concerto No.2 in g, Op.8 - Presto
48	Antonio Vivaldi	AUTUMN Concerto No.3 in F, Op.8 - Allegro
49	Antonio Vivaldi	AUTUMN Concerto No.3 in F, Op.8 - Adagio molto
50	Antonio Vivaldi	AUTUMN Concerto No.3 in F, Op.8 - Allegro
51	Antonio Vivaldi	WINTER Concerto No.4 in f, Op.8 - Allegro non molto
52	Antonio Vivaldi	WINTER Concerto No.4 in f, Op.8 - Largo
53	Antonio Vivaldi	WINTER Concerto No.4 in f, Op.8 - Allegro
54	Antonio Vivaldi	Symphony No.1 In C RV 116 - Allegro
55	Antonio Vivaldi	Symphony No.1 In C RV 116 - Andante
56	Antonio Vivaldi	Symphony No.1 In C RV 116 - Presto
57	Antonio Vivaldi	Symphony N. 2 IN G RV 146 - Andante
58	Antonio Vivaldi	Storm at sea, concerto per violino e archi
59	Franz Schubert	Symphony No. 5 Movt 1
60	Franz Schubert	Ellens Gesang 3, Op. 52/6, D 839,Ave Maria
61	Franz Schubert	Impromptu In G Flat, D 899

Nr.	Artist	Title
62	Franz Schubert	German Dance No. 1 In C, D 90
63	Franz Schubert	String Quintet In C Major D. 956 - II. Adagio
64	Franz Schubert	Symphony No. 9 In C Major Great D. 944 - III. Scherzo, Allegro vivace
65	Franz Schubert	"Serenade" Standchen
66	Franz Schubert	Moment Musical No. 3 In F Minor, Op. 94, D
67	Franz Schubert	Impromptus, Op. 90, D 899 - #4 In A Flat
68	Franz Schubert	Symphony No. 3 In D, D 200 - Allegretto
69	Pyotr Tchaikovsky	Piano Concerto No. 1 In B Flat minor, Op. 23 - 1. Allegro Non Troppo E Molto Maestoso
70	Pyotr Tchaikovsky	Piano Concerto No. 2 in G Major, Op. 44: III. Allegro con Fuoco
71	Pyotr Tchaikovsky	Swan Lake, Op. 20 - Dance Of The Swans
72	Pyotr Tchaikovsky	The Nutcracker, Op. 71 - Waltz Of The Snowflakes
73	Pyotr Tchaikovsky	The Nutcracker, Op. 71 - Waltz Of The Flowers
74	Pyotr Tchaikovsky	The Nutcracker, Op. 71: XIIId. Trepak (Russian Dance)
75	Pyotr Tchaikovsky	Les Saisons, Op. 37B - June: Barcarolle
76	Pyotr Tchaikovsky	Sleeping Beauty, Op. 66 - Waltz
77	Pyotr Tchaikovsky	Symphony #6 In B Minor, Op. 74, "Pathétique"
78	Pyotr Tchaikovsky	Eugene Onegin, Op. 24 - Polonaise
79	Pyotr Tchaikovsky	Suite #4 In G, Op. 61, "Mozartiana" - 3. Preghiera
80	Pyotr Tchaikovsky	Romance In F Minor, Op. 5
81	Pyotr Tchaikovsky	1812 Overture, Op. 49
82	Dmitri Shostakovich	Symphony No. 5 in D minor, Op. 47: Moderato
83	Dmitri Shostakovich	Symphony No. 5 in D minor, Op. 47: Allegro non troppo
84	Dmitri Shostakovich	Symphony No. 7 in C Major (Leningrad), Op. 60: Memories, Moderato (poco allegretto)
85	Dmitri Shostakovich	Symphony No. 8 in C minor, Op. 65 (Stalingrad): Allegro non troppo
86	Dmitri Shostakovich	Symphony No. 10 in E minor, Op. 93: Andante
87	Dmitri Shostakovich	Sonata for piano No. 2 in B minor, Op. 61: Allegretto
88	Dmitri Shostakovich	String Quartet No. 8 in C minor, Op. 110: Largo
89	Dmitri Shostakovich	String Quartet No. 11 in F minor, Op. 122: Introduction (andantino)
90	Dmitri Shostakovich	String Quartet No. 11 in F minor, Op. 122: Recitativo (adagio)
91	Edvard Grieg	In The Hall of the Mountain King
92	Edvard Grieg	Morning Mood
93	Edvard Grieg	Piano Concerto in A minor: I. Allegro molto moderato
94	Edvard Grieg	Peer Gynt Suite No. 1: Anitra's Dance
95	Edvard Grieg	Peer Gynt Suite No. 1: Aase's Death
96	Nikolai Rimsky-Korsakov	Flight of the bumblebee
97	Nikolai Rimsky-Korsakov	Scheherazade: the sea and Sinbad's ship
98	Nikolai Rimsky-Korsakov	Scheherazade: the story of the kalander prince
99	Nikolai Rimsky-Korsakov	Capriccio Espagnol: Alborada - Vivo e strepitoso
100	Nikolai Rimsky-Korsakov	Variazioni: Andante con moto
101	Jennifer Lopez ft. Pitbull	On the floor
102	Jennifer Lopez	Let's get loud
103	Jennifer Lopez ft. Pitbull	Dance again
104	Jennifer Lopez	Papi
105	Katy Perry	Firework
106	Katy Perry	I kissed a girl
107	Katy Perry	Wide awake
108	Pussy cat dolls	Don't cha
109	Pussy cat dolls	Buttons
110	Rihanna	Umbrella
111	Rihanna ft. Justin Timberlake	Rehab
112	Rihanna	Diamonds
113	Maroon 5	Animals
114	Maroon 5	Sugar
115	Maroon 5 ft. Christina Aguilera	Moves like jagger
116	Maroon 5	One more night
117	Christina Aguilera	Hurt
118	Christina Aguilera	Fighter
119	Christina Aguilera	Genie in a bottle

Nr.	Artist	Title
120	Britney Spears	My Prerogative
121	Britney Spears	Baby one more time
122	Britney Spears	Toxic
123	Garbage	World is not enough
124	R.E.M.	Loosing my religion
125	Shakira	Whenever wherever
126	Shakira	Hips don't lie
127	Shakira	Waka waka
128	Celine Dion	My heart will go on
129	Celine Dion	I'm alive
130	Tony Braxton	Un-break my heart
131	Tony Braxton	Yesterday
132	Bryan Adams	I do it for you
133	Bryan Adams	Please forgive me
134	Ricky Martin	Livin la vida loca
135	Ricky Martin	María
136	Robbie Williams	Supreme
137	Robbie Williams	Rock dj
138	Sting	Shape of my heart
139	Sting	Desert rose
140	Spice Girls	Viva forever
141	Spice Girls	Wannabe
142	Spice Girls	Say you'll be there
143	Backstreet Boys	I want it that way
144	Backstreet Boys	Show me the meaning of being lonely
145	P!nk	Family portrait
146	P!nk	Try
147	P!nk ft. Nate Ruess	Just give me
148	Enrique Iglesias	Hero
149	Enrique Iglesias	Not in love
150	Enrique Iglesias	Rhythm divine
151	Mariah Carey	All I want for Christmas is you
152	Mariah Carey	My all
153	Mariah Carey	Hero
154	Kylie Minogue	Can't get you out of my head
155	Sia	Chandelier
156	Sia	Elastic heart
157	Sia	Big girls cry
158	Elton John	Sorry seems to be the hardest word
159	Elton John	Circle of love
160	Bruno Mars	Grenade
161	Bruno Mars	Locked out of heaven
162	Bruno Mars	Just the way you are
163	Bruno Mars	When I was your man
164	Taylor Swift	Blank space
165	Taylor Swift	Style
166	Taylor Swift	Wildest dreams
167	John Legend	You & I
168	John Legend	All of me
169	John Legend	Tonight
170	Adele	Hello
171	Adele	Rolling in the deep
172	Adele	Someone like you
173	Hozier	Take me to church
174	Hozier	Someone new
175	Gotye	Eyes wide open
176	Gotye ft. Kimbra	Somebody that I used to know
177	Madonna	American pie
178	Madonna	Die another day
179	Madonna	Music
180	Miley Cyrus	We can't stop
181	Miley Cyrus	Wrecking ball
182	Miley Cyrus	Can't be tamed
183	Lady Gaga	Bad romance

Nr.	Artist	Title
184	Lady Gaga	Poker face
185	Lady Gaga	You & I
186	Justin Bieber	Sorry
187	Justin Bieber	What do you mean?
188	Justin Bieber	Baby
189	Fun.	Some nights
190	Fun.	Carry on
191	Fun. ft. Janelle Monáe	We are young
192	Owl City	Fireflies
193	Owl City	Vannila twilight
194	Timbaland ft. Katy Perry	If we ever meet again
195	Timbaland ft. Nelly Furtado, Justin Timberlake	Give it to me
196	Justin Timberlake	Mirrors
197	Justin Timberlake	Cry me a river
198	Justin Timberlake	What goes around
199	Leona Lewis	Bleeding love
200	Leona Lewis	Better in time
201	Queen	Bohemian rhapsody
202	Queen	The show must go on
203	Queen	I want to break free
204	Queen	We are the champions
205	Linkin Park	Numb
206	Linkin Park	In the end
207	Linkin Park ft. Busta Rhymes	We made it
208	Bon Jovi	It's my life
209	Bon Jovi	Always
210	Bon Jovi	Bed of roses
211	Alice Cooper	Poison
212	Metallica	Nothing else matters
213	Metallica	Enter sandman
214	Metallica	Fade to black
215	Europe	Final countdown
216	Europe	Open your heart
217	Europe	Carrie
218	Bonnie Tyler	Total eclipse of the heart
219	Bonnie Tyler	it's a heartache
220	Bonnie Tyler	Holding out for a hero
221	Pink Floyd	Another brick in the wall
222	Pink Floyd	Wish you were here
223	Led Zeppelin	Stairway to heaven
224	Guns N' Roses	November rain
225	Guns N' Roses	Paradise city
226	Guns N' Roses	Sweet child o' mine
227	Heart	All I wanna do is make love to you
228	Heart	Alone
229	Scorpions	Maybe I maybe you
230	Scorpions	Still loving you
231	Scorpions	Wind of change
232	Scorpions	Send me an angel
233	Scorpions	You and I
234	Scorpions	When you came into my life
235	The hooters	Johnny B
236	Aerosmith	I don't want to miss a thing
237	Aerosmith	Cryin'
238	Aerosmith	Dream on
239	Aerosmith	Crazy
240	Gary Moore	Empty rooms
241	Gary Moore	One day
242	Bon Jovi	This ain't a love song
243	The Pretenders	I'll stand by you
244	No doubt	Don't speak
245	Joan Osborne	One of us
246	Skid row	18 and life

Nr.	Artist	Title
247	Skid row	I remember you
248	Eagles	Hotel California
249	Eurythmics	Sweet dreams
250	Red Hot Chili Peppers	Californication
251	Red Hot Chili Peppers	Otherside
252	The police	Every breath you take
253	Nirvana	Smells like teen spirit
254	Rammstein	Du hast
255	Rammstein	Mein herz brennt
256	Rammstein	Sonne
257	System Of A Down	Chop Suey!
258	System Of A Down	Aerials
259	System Of A Down	B.Y.O.B.
260	Marilyn Manson	Tainted Love
261	Marilyn Manson	Rock is dead
262	Marilyn Manson	Personal Jesus
263	Scooter	Fire
264	Scooter	How much is the fish
265	Shakespeare Sisters	Stay
266	Shakespeare Sisters	Hello
267	The Cranberries	Zombie
268	The Cranberries	Promises
269	Megadeth	A tout le monde
270	Ozzy Osbourne	Dreamer
271	Ozzy Osbourne	No more tears
272	Ozzy Osbourne ft. Lita Ford	If I close my eyes forever
273	Joan Jett	I hate myself for loving you
274	Joan Jett	I love Rock N Roll
275	Joan Jett	Dirty deeds done dirt cheap
276	AC/DC	Thunderstruck
277	AC/DC	You shook me all night long
278	AC/DC	Back in black
279	Phil Collins	In the air tonight
280	Phil Collins	Another day in paradise
281	Cyndi Lauper	Time after time
282	Richard Marx	I will be right here waiting for you
283	Richard Marx	Hazard
284	Richard Marx	Endless summer nights
285	Deep Purple	Smoke on the water
286	Deep Purple	Soldier of fortune
287	Deep Purple	Perfect strangers
288	Bob Dylan	Knockin on heaven's door
289	Rolling Stones	Wild horses
290	Rolling Stones	Angie
291	Rolling Stones	Beast of burden
292	Judas Priest	Painkiller
293	Vince Neil	Tattoos and tequila
294	Vince Neil	Set me free
295	Evanescence	Bring me to life
296	Evanescence	My immortal
297	Evanescence	Call me when you're sober
298	Evanescence	Going under
299	Nickelback	How you remind me
300	Nickelback	Photograph

5.2. Tunes emotions survey

Tunes from the tunes database were provided to the audience to evaluate emotions. The questionnaire was conducted using an application with interface as shown on Figure 5.1. The application used for survey contained:

- list of tunes which participants listened
- audio player for current tune
- emotion plane for providing the answer as emotion

To provide the emotion for the listened tune participants can click anywhere on emotion plane to set the answer. Anytime during the test participants could adjust and change their answers by navigating between given tunes and modifying answers by moving the point on emotion plane.

The tunes in questionnaires were given full from the beginning till the end as they were composed by authors and presented to the public. Poll participants were asked to listen full tunes and provide the overall emotion for each tune.

For this questionnaire 20 persons were selected. All of them are from Europe. Half of them are men, other half are women. The ages vary from 21 years old till 55 years old.

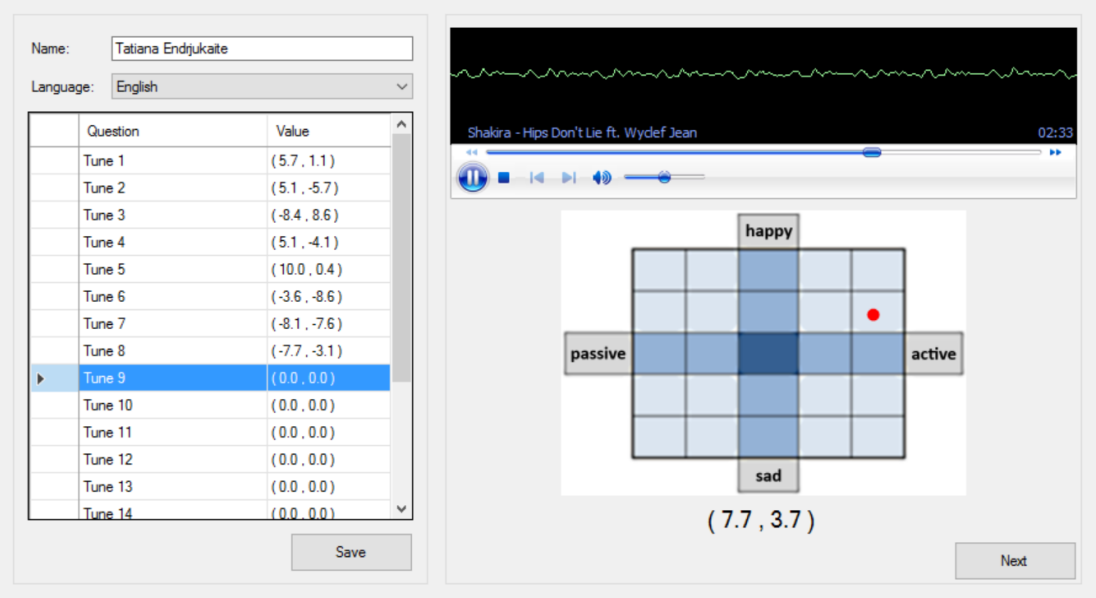


Figure 5.1. Tunes emotions evaluation questionnaire application interface.

The total length of questionnaire for every participant took approximately 25 hours which included brief explanation of the rules and them main part when participants listened to 300 tunes. Since tunes listening was taking so long time, the questionnaire

was divided into 10 separate parts for every person. Therefore, every participant was taking the survey 10 days with approximately 2 hours per day. Every survey part contained mixed tunes sequence from all genres.

5.3. Emotion adjectives

Hevner's adjectives were taken as a base list of adjectives for emotion representation in words. The adjectives list is shown in below Table 5.3 in the first column. For using them in the emotion recognition system all of them should have corresponding location on emotion plane. This cannot be determined automatically since it depends on perception by humans, therefore it can only be through a survey. The survey and its results are described in detail in the next chapter.

5.4. Emotion adjectives survey

This chapter describes the experiment which was conducted for collecting the emotion plane values of adjectives which people usually use to describe their emotions. Figure 5.2 the interface of the application which was used to conduct the survey and to collect participants' answers.

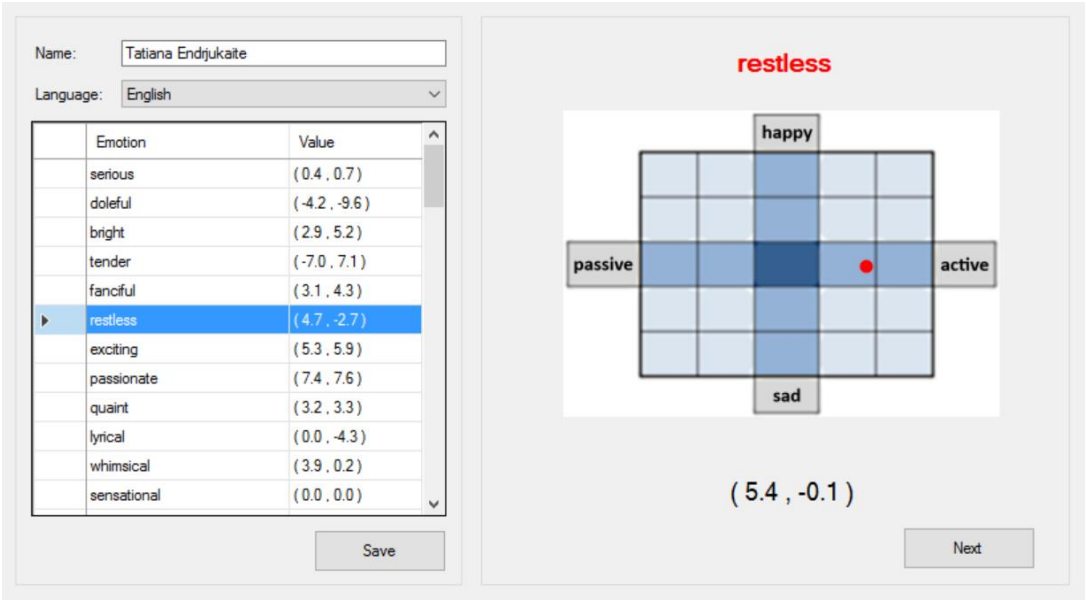


Figure 5.2. Adjectives emotions evaluation survey application.

The initial survey shown that sometimes it is hard for people to project their emotions onto the given plane. As a result, the answers from participants were varying and sometimes inconsistent. In the second survey participants used additional sheet of paper

with brief mapping of some basic emotion adjectives placed on the plane as shown on the Figure 5.3.

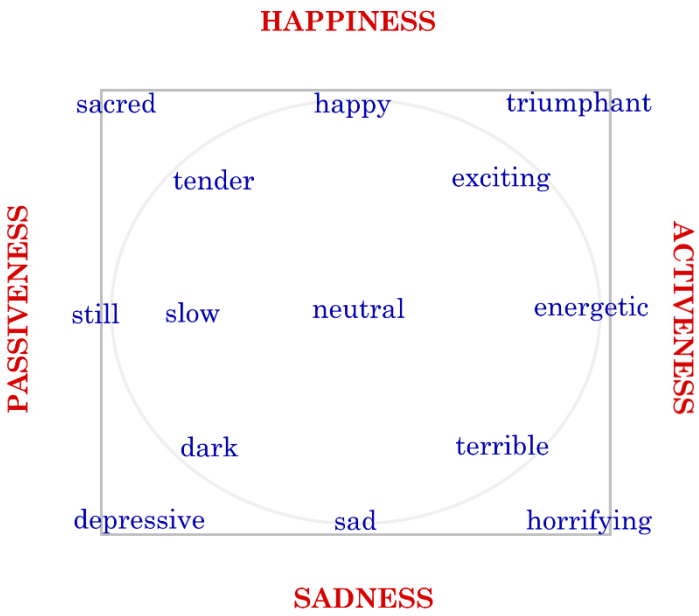


Figure 5.3. Basic emotions mapping provided for survey participants.

Emotions collected during both surveys were summarized and results were processed with analysis of variance. Answers for every adjective were grouped together. Mean and variance was calculated among answers for every adjective. The results of first survey show that average variance among adjectives is *33.17*, average standard deviation is *5.3*. The result of second survey show that average variance was reduced till *11.17*, average standard deviation became *2.6*. Distribution of standard deviations for emotion adjectives for both surveys are shown on the Figure 5.4 and Figure 5.5 for comparison.

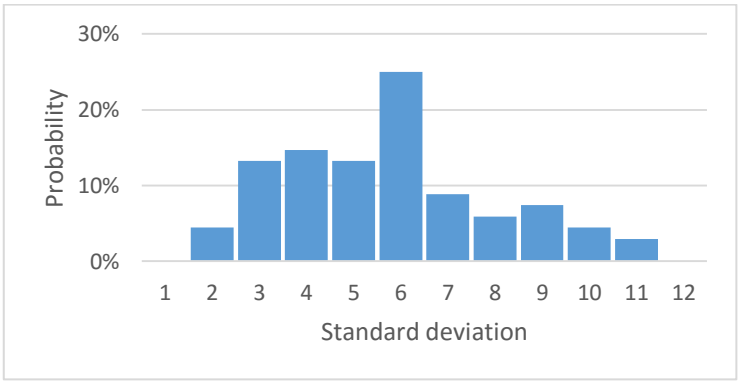


Figure 5.4. Distribution of standard deviation for first survey without additional emotions map.

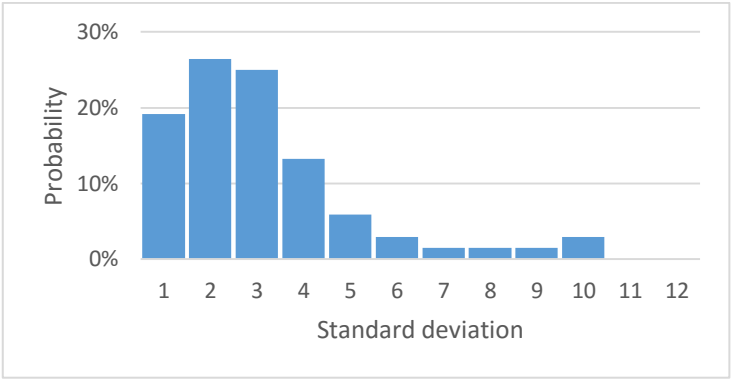


Figure 5.5. Distribution of standard deviation for second survey with additional emotions map.

As we can see from the Figure 5.5, in the second survey where participants used supplementary map, 71% of all adjectives (first three bars in the histogram) were placed to the plane with standard deviation less than value of *3.0* which gives 48 adjectives. The value of *3.0* is a good compromise between good precision and variety of adjectives.

Table 5.3 shows survey results with mean activeness and happiness value for every adjective, as well as variance and standard deviation for every adjective. Adjectives are ordered ascendingly by variance, so adjectives which appear at the top of the table have meaning which is easier to understand and therefore were described more consistent by survey participants. On the other hand, adjectives at the bottom of the table have big variance which means that participants had difficulties in placing them on the emotion plane and the results were inconsistent. Further in the system only adjectives with standard deviation less than *3.0* were used as similarly understandable ones for precise description of emotions.

Table 5.3. Adjectives values on emotion plane based on survey results.

Adjective	Activeness	Happiness	Variance	St. dev
sad	-0.10	-9.92	0.017	0.130
awe-inspiring	9.94	-9.90	0.050	0.224
triumphant	9.76	10.00	0.063	0.251
serious	0.08	0.14	0.098	0.313
happy	-0.18	9.98	0.162	0.402
lyrical	0.08	-4.36	0.293	0.541
exhilarated	6.22	-0.10	0.312	0.559
exciting	6.08	6.20	0.327	0.572
sacred	-9.66	9.72	0.498	0.706
depressing	-9.64	-9.60	0.620	0.787
yearning	-0.04	-6.06	0.733	0.856
light	-4.62	4.48	0.962	0.981
dark	-6.44	-6.54	0.983	0.991
passionate	8.18	8.24	1.122	1.059

Adjective	Activeness	Happiness	Variance	St. dev
calm	-8.86	0.10	1.188	1.090
spiritual	-9.18	9.36	1.322	1.150
sober	-0.94	0.02	1.658	1.288
quiet	-6.82	-0.24	1.892	1.375
vigorous	9.00	0.06	1.955	1.398
serene	-7.92	0.20	1.987	1.410
tender	-5.96	5.78	2.047	1.431
bright	2.50	5.00	2.065	1.437
playful	3.88	4.82	2.087	1.445
leisurely	-8.88	0.04	2.397	1.548
joyous	0.06	7.90	2.470	1.572
merry	0.70	8.38	2.702	1.644
tranquil	-7.42	-0.04	2.972	1.724
soothing	-4.70	1.18	3.292	1.814
doleful	-1.36	-9.34	3.348	1.830
longing	6.64	5.58	3.392	1.842
yielding	-4.00	1.56	3.703	1.924
ponderous	-6.12	-5.04	4.457	2.111
quaint	2.32	3.70	4.702	2.168
restless	4.80	-3.12	4.762	2.182
gay	0.38	7.56	5.173	2.274
whimsical	5.36	1.28	5.353	2.314
melancholy	0.96	-8.08	5.652	2.377
sprightly	5.50	4.36	5.885	2.426
mournful	-6.74	-9.54	5.948	2.439
dramatic	0.00	-5.86	6.003	2.450
cheerful	3.46	5.80	6.125	2.475
dreamy	-5.20	5.72	6.175	2.485
delicate	-6.16	2.46	6.383	2.526
robust	3.12	1.00	7.502	2.739
heavy	-6.34	-6.04	7.673	2.770
soaring	-5.90	5.74	8.460	2.909
martial	7.36	-6.06	8.758	2.959
sensational	6.72	6.60	8.887	2.981
gloomy	-0.22	-5.92	9.097	3.016
plaintive	-2.92	-8.38	9.172	3.029
solemn	7.72	7.74	9.303	3.050
sentimental	-2.14	-1.14	9.448	3.074
impetuous	5.62	2.50	9.590	3.097
fanciful	3.18	4.04	11.152	3.339
pleading	1.32	-6.32	12.117	3.481
humorous	2.08	3.24	13.072	3.616
emphatic	7.14	2.46	14.308	3.783
agitated	6.54	-1.58	16.367	4.046
pathetic	-1.04	-6.96	17.408	4.172
graceful	-1.28	4.36	20.412	4.518
tragic	-4.78	-7.96	22.222	4.714
frustrated	-3.60	-5.62	32.275	5.681
lofty	6.14	2.80	35.985	5.999
satisfying	-3.70	5.36	42.555	6.523
dignified	3.24	7.68	59.743	7.729
exalting	-2.06	7.94	77.908	8.827

Adjective	Activeness	Happiness	Variance	St. dev
inspiring	-3.06	7.94	88.868	9.427
majestic	0.60	8.68	94.120	9.702

5.5. Emotions map

The emotions adjectives survey result can be shown on the emotions plane as a map. Figure 5.6 shows the result by using only those adjectives that had standard deviation less than *3* in the survey that gives 48 adjectives. Locations of all these adjectives look very consistent and logical in the sense of distances. Synonyms and adjectives with similar meanings appear nearby whereas antonyms and adjectives with very different meanings are located with big distance between each other.

Border values on the map are very similar to those provided on the given base emotions map which was given to participants. That is expected since that was the basic condition for getting consistent view to the plane for all participants. Distribution of all other adjectives is more interesting and is a valuable knowledge that was collected from the survey.

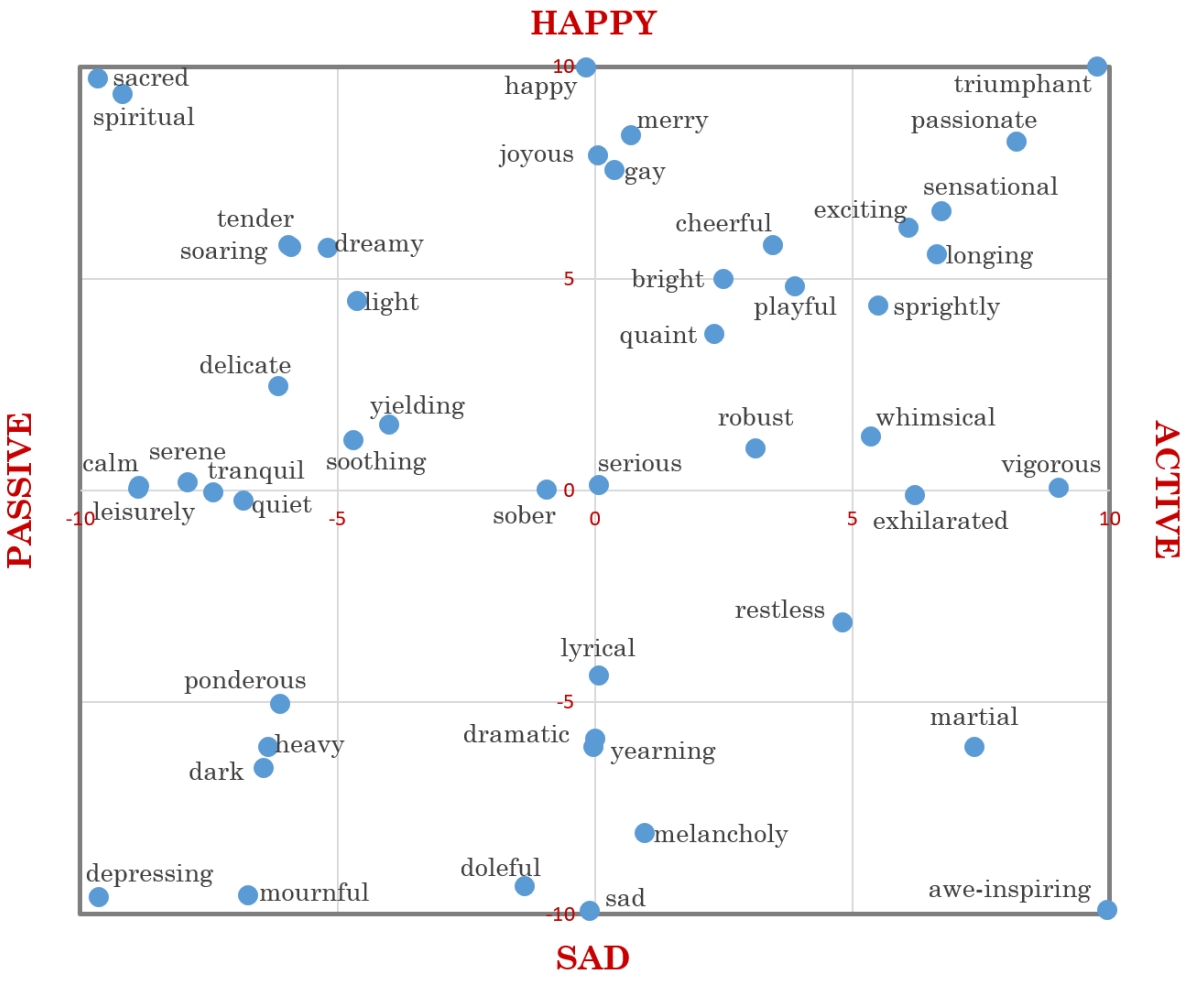


Figure 5.6. Emotion adjectives map based on survey data.

Chapter 6. Evaluation of the method

This chapter describes the evaluation of the tunes emotion recognition approach. First, we specify the accuracy calculation which is used for comparing known and calculated emotions. Second, the validation approach is described which was used for testing the approach. The comparison with other researches shows that it includes broader range of tunes genres which is very significant for music emotion recognition system. Another very important distinction is that emotions are described on continuous plane instead of categories which resulted in much more supported adjectives for emotion description.

6.1. Accuracy calculation

The accuracy calculation is based on emotions, i.e. the difference between calculated emotion and known emotion. This difference gives the error of the output of the system and it is calculated as a distance on emotion plane.

The values of emotions on happiness and activeness axes ranges between *-10* and *10*, so every axis may have maximum possible error equal to *20*. This distance was chosen as a limit boundary and gets 0% accuracy. Points on the emotion plane may have even longer distance for example when these points are located in bottom-right and top-left corners, but anyway all distance longer than *20* are getting 0% accuracy. On the other hand, when difference between calculated and known emotion is *0*, its accuracy is 100%.

More formally, accuracy is calculated from distance on emotional plane as shown in (6.1).

$$A = \begin{cases} (1 - 0.05 * d) * 100\% & , \quad d \leq 20 \\ 0\% & , \quad d > 20 \end{cases} \quad , \quad d = \sqrt{\sum_{i=1}^{2} (\hat{\gamma}_i - \gamma_i)^2} \tag{6.1}$$

where A is accuracy, γ is a known emotion, $\hat{\gamma}$ is calculated emotion estimation.

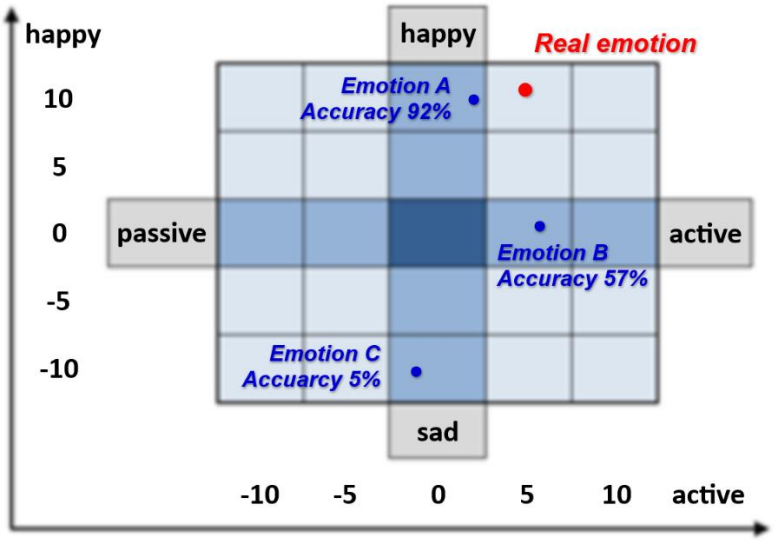


Figure 6.1. Accuracy calculation example.

Table 6.1. Values for accuracy calculation example.

Emotion	Happiness	Activeness	Distance	Accuracy
Real	4.0	9.0	0.0	100%
A	2.4	8.7	1.6	92%
B	5.5	0.5	8.6	57%
C	-2.0	-9.0	19.0	5%

Table 6.1 and Figure 6.1 show an example of accuracy calculation. Red dot represents a given emotion, blue dots show calculated emotions by the system. The emotion A shown as a blue dot is located near the target with Euclidean distance *1.6* which gives it accuracy 92%. The emotion B shown as a blue dot has the distance equal to *8.6* to the target and therefore its accuracy is 57%. The emotion C has the distance *19.0* to the target so gets only 5% accuracy as a result.

6.2. Validation of the system

To evaluate the method of tunes emotion recognition, the database was divided into two parts. 80% of the tunes database was left as required data for system operation, 20% of tunes were extracted as a known data for system validation and testing purposes. Tunes were used as an input for the system, but given emotions were compared to the calculated output of the system. Figure 6.2 shows the validation approach of emotion recognition.

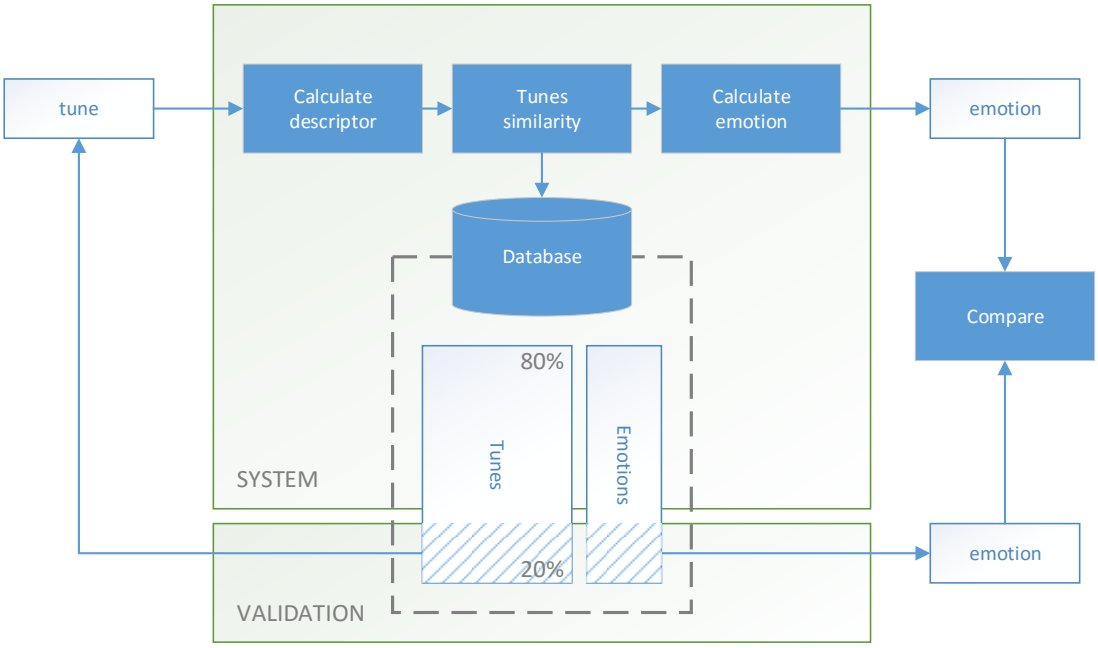


Figure 6.2. Tunes emotion recognition method evaluation with 20% of database.

Since the full database size is 300 tunes, 20% of it gave 60 tunes for validation. These tunes were randomly selected to cover all genres: classical, variations of rock, and modern styles of popular music. All these selected 60 tunes were removed from the system's database and only used as an input for the system, but known emotions were used as a ground truth data. The histogram of acquired results is shown on Figure 6.3.

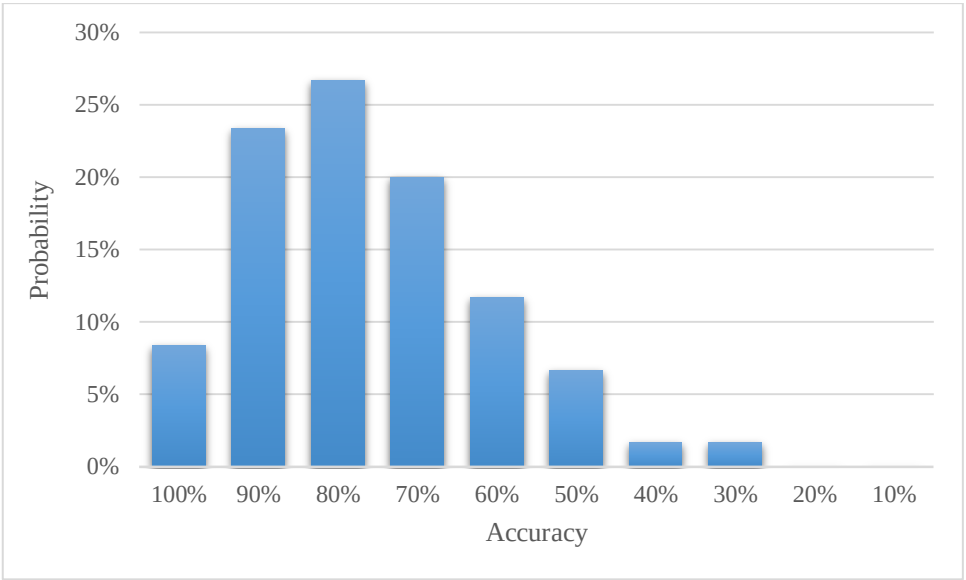


Figure 6.3. The histogram of accuracy distribution of test tunes.

The average accuracy of the system is 76.2%. The first two bars of histogram which represent accuracy above 80% show very good *match* of the calculated result. Next two bars have accuracy between 60% and 80%. These tunes were identified to a *very similar* emotion. The accuracy between 40% and 60% are not too low because the calculated emotion appears *adjacent* to the ground truth. Results of 20% – 40% and 0% – 20% can be considered *wrong* and completely *opposite*. The overall picture of validation results is shown in Figure 6.4.

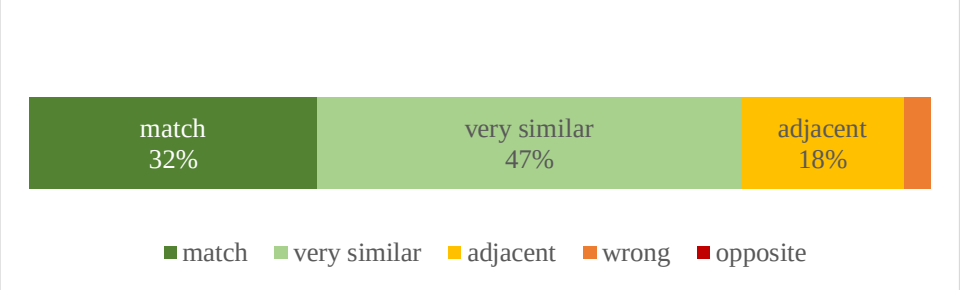


Figure 6.4. The overall picture of validation results.

The *match* and *very similar* results make up 79% total which shows very good validation results. Besides that, next 18% of test tunes were not calculated with a wrong emotion but with an adjacent emotion that is somewhat related to the correct emotion.

6.3. Comparison with other researches

Comparison with other related researches on audio signal tunes classification into emotions is shown in Table 6.2. The most important differences of current work to other researches are emotion description on a continuous plane instead of categories, and the calculation method which is based on pattern matching instead of machine learning.

Related researches take into account only a short fragment of a tune while the current thesis proposes the whole tune analysis. Unfortunately, no researches exist about whole tune processing for emotion recognition.

Table 6.2. Comparison with other researches.

	Dan Yang, Won Sook Lee	Tao Li, Mitsunori Ogihara	Dan Liu, Lie Lu, Hong-Jiang Zhang	Carvalho, Chao	Li, Ogihara	Current work
Database size	500	250	250	200	235	300
Tune length	20 seconds	30 seconds (random)	20 seconds	—	—	whole tune, 5 minutes in average
Genres	1	4	1	—	1	21
Method	Support Vector Machine (SVM)	Support Vector Machine (SVM)	Gaussian Mixture Model (GMM)	Support Vector Machine (SVM)	Support Vector Machine (SVM)	Pattern match
Features	timbre, rhythm	timbre, rhythm, pitch	timbre, intensity, rhythm	musical surface, spectral flatness, spectral crest, MFCC	MFCC, musical surface, DWCH	IFS, MFCC, repetitions, thumbnail, homogeneity pattern
Emotion categories	2 categories	13 groups	2 groups	5 categories	3 categories	—
Emotion adjectives	—	23 adjectives	4 moods	1 adjective	—	48 adjectives
Accuracy	90%	~50%	76 - 95%	82%	70 - 83%	76.2%

Chapter 7. Tunes Emotion Estimation for Music Service

This chapter describes three main applications of the described approach: query by emotion, query by tune, and playlist suggestion. This chapter also provides structures of the system for each of these applications.

7.1. Querying by Emotion

In this use case user provides an emotion and the system gives a list of tunes that match the provided emotion the most. The emotion can either be provided as an adjective, as a list of adjectives, or as a point on emotion plane. As a result the system provides a list of tunes from its database which are most relevant to the requested emotion.

The function of querying by emotion in the system is achieved in 2 steps. First, emotion adjective is converted to emotion. Second, the database is queried to fetch tunes with lowest distances on emotion plane. Finally, tunes are presented to the user sorted by relevance.

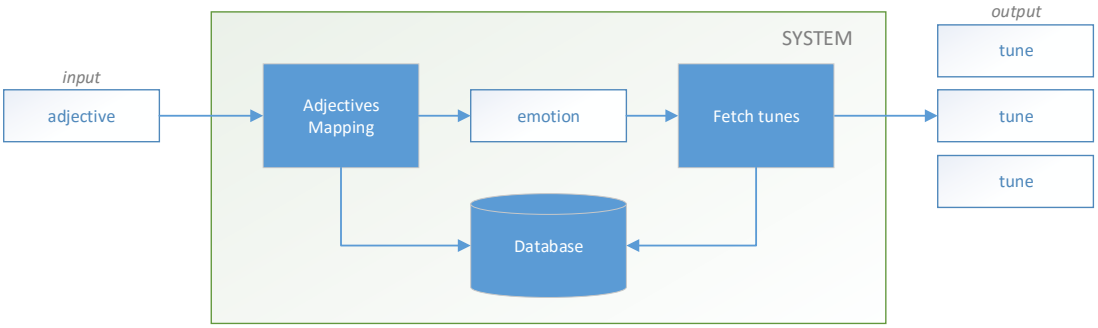


Figure 7.1. Querying by emotion using adjective.

Figure 7.1 shows the query by emotion feature operation structure. The main work is performed by adjectives mapping part which translates user-friendly text which describes user input emotion into numerical value which is suitable for further processing. Next the database performs the remaining part to select corresponding tunes for given emotion.

7.2. Querying by Tune

In this use case user uploads a tune to the system and the system returns a list of tunes from its database which are most relevant to the requested emotion – have the same or similar emotion.

The function of querying by tune in the system is achieved in 3 steps. First, tune is processed to build its descriptor. Second, the database is queried to fetch tunes which are similar to the description of provided tune. Third, corresponding emotions are combined to calculate the resulting emotion. At the fourth step the database is queried again to fetch tunes with relatively low distance by emotion. Finally, tunes are presented to the user sorted by relevance.

Figure 7.2 shows the system structure for querying by tune. This scenario involves tune descriptor calculation part with emotion estimation using database, and querying tunes from database by emotion.

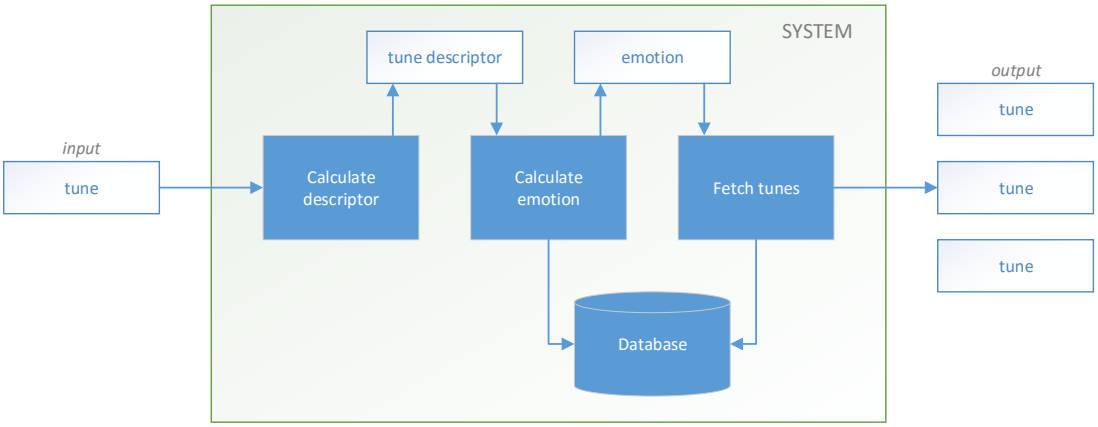


Figure 7.2. Querying by tune.

7.3. Playlist suggestion

In this use case user just selects and listens tunes one by one in the system. In the background system collects statistics of tunes that user listened recently, and automatically gives suggestions. The suggestions are given with two options. In one option tunes are suggested which match to the average emotion by last 5 tunes listened by the user. Other option suggests tunes taking into account emotion change tendency and trying to predict which emotion should follow next by means of 2-dimensional linear approximation using last 5 tunes played by the user.

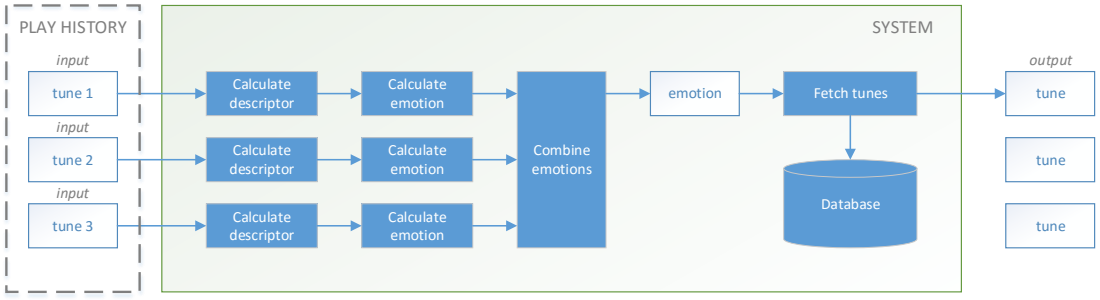


Figure 7.3. Playlist suggestion.

An example of tune suggestion by means of emotion prediction from last five tunes is shown in the Figure 7.4. Blue points connected with a blue line show the played tunes emotions by activeness and happiness, thin red lines show linear approximations for the those values, and red points show the activeness and happiness prediction for the next suggestion. The overall suggestion result is displayed in Figure 7.5.

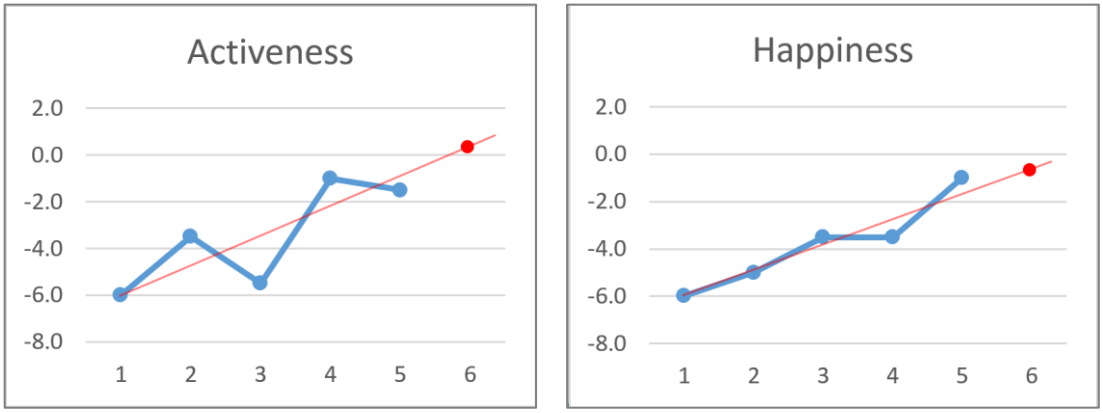


Figure 7.4. Emotion trend prediction example by means of linear approximation.

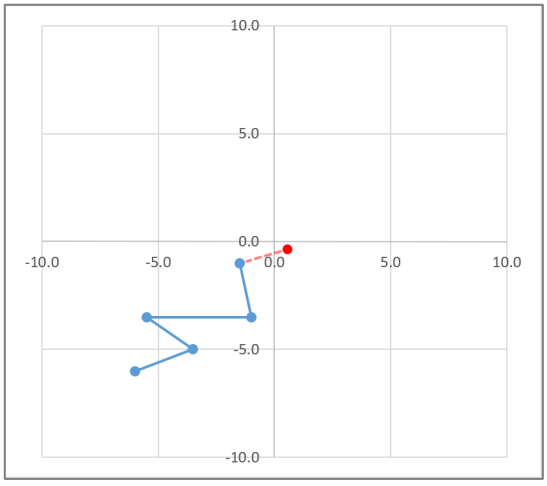


Figure 7.5. Tune suggestion by emotion trend prediction.

Nowadays a number of music recommendation systems exist which give playlist suggestions. Such system are based on other people's preferences using played tunes statistics. They show pretty good results among existing and well known tunes. However, such systems cannot be used where statistical data is not available. For example, the system would never suggest to listen to a new song of an artist. Users have to choose the new song directly many times before the system collects required statistics for the tune. Moreover, private music collections would never have such statistics to give suggestions. Therefore, music recommendation systems that are based on other people's preferences have very limited area of usage and their results are tightly related to popularity. So, a more objective methods are needed for recommendation systems which performs content-based analysis of the tune.

Chapter 8. Future applications

8.1. Influence on Humans Mood

Extend the system with a feature to influence the listener's mood. For example, when listener feels depressive we suggest tunes playlist which moves his mood towards feeling happy. Or vice versa, when listener is overexcited we suggest playlist which will calm user down. Figure 8.1 shows the sequence of tunes which are suggested to the user to slowly shift his emotion from heavy and slightly depressive towards playful and bright. Such approach may have some problems and user may just reject the suggested tunes at all, for example if the suggested sequence of tunes moves too fast to the new mood while the listener is going slower. Another cause of rejection could be that some of suggested tunes user does not know and doesn't want to listen.

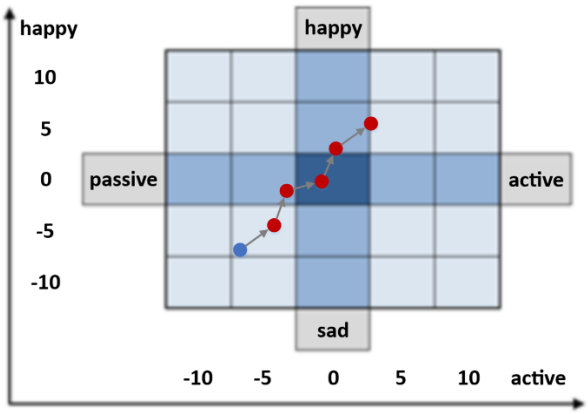


Figure 8.1. Sequence of tunes for direct influence on human emotion.

A more flexible way should be used for affecting the mood of the listener. On each step when next tune should be selected, the system gives some choices to the user to choose which tune he wishes to listen. These suggestions will try to move to the destination emotion point, but at the same time they give flexibility to the listener to provide some feedback and prevent from moving too fast or play undesired tunes.

Figure 8.2 (a – e) shows the sequence of five steps where on each step the system gives 6 suggested tunes with emotions around the last played tune and following the tendency to move towards the goal emotion somewhere around cheerful and exciting. The actual played tunes sequence which is shown as blue points is completely not direct and has a lot of turns, but the overall movement trend is correct. Every person is unique and

therefore there could be no common way which works well for everyone. Listeners are choosing what they want to listen themselves, but the system softly and unobtrusively guides them through suggestions.

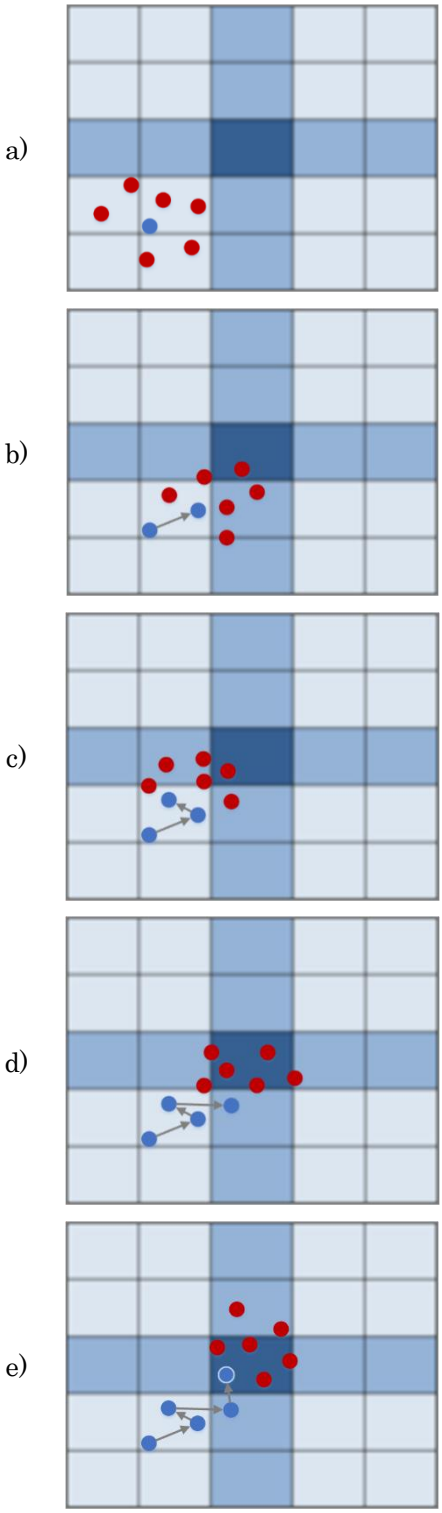


Figure 8.2. Sequence of tunes suggestions (red) to influence on human emotion, and actual selected and played tunes (blue).

Chapter 9. Concluding remarks

The reason that many people engage with music is that it has power to evoke or enhance valued emotional states. In the light of this we can say that emotions play a very important role in music.

The growing amount of audio data wakes up a need for content-based searching. Traditionally, tunes information has been retrieved based on reference information, for example, the title of a tune, the name of an artist, the genre and so on. When users would like to try to find music pieces in a specific mood such standard reference information of the tunes is not sufficiently effective. We need new methods and approaches to realize emotion-based search and tune content analysis.

In this thesis we introduce a number of innovative approaches for acoustic signal processing, emotion calculation, and music-tunes emotion recognition techniques.

First, we introduce a novel approach of acoustic signal analysis – IFS spectrum. IFS is used as a main acoustic signal processing method because it is intended for processing non-stationary and non-periodic signal, such as music.

Second, we introduce the innovative approach for tune structure analysis – tune internal homogeneity analysis. This approach provides an easy way to see the structure of a tune, as well as find typical and untypical parts, which is extremely important for music analysis from the human perception perspective.

Third, we describe the approach for identifying repetitive structures in tunes which have a strong impact on music perception as well. The important advantage of the proposed approach is in its ability to detect not only identical parts in acoustic signal of a tune, but also those part that are present in different tonality, performed by different musical instruments, but sound similar and perceived by listener the same way.

Fourth, music similarity calculation approach is described, which is used for tunes comparison. This approach uses all of the methods mentioned above to compare tunes for similarity based on aspects of human perception aspects: overall structure, thumbnail, and repetitive parts of a tune.

Fifth, we describe a new unique approach for emotions definition, representation and calculation – the emotion plane which is used for describing the listener's perception of music. Unlike other models of emotions, this approach is much more suitable for emotions processing, calculation and aggregation which plays a significant role in music emotional analysis.

Further in this thesis we describes the design and the overall architecture of the system for automatic emotion identification by means of intrinsic musical features. Its innovativeness is that it uses new musical features for tune's analysis, which are based on human's perception of the music. Therefore the results are more relevant to listener expectations. These significant features are repeated parts of a tune, thumbnail of a tune, and internal homogeneity pattern. Moreover, the described system instead of emotions categorizations uses a new way of automatic emotion processing, calculation and presentation on a plane, which has the key idea to quantitatively measure emotions that is very significant for emotions calculation.

This thesis also describes the process of data collection which is required for the system proper operation. Two types of information was collected: a set of tunes with known emotions, and mapping of emotion adjectives to the emotion plane. Both of these kinds of data are specific to human perception, and therefore the data was collected using surveys with individuals and thoroughly processed after that.

Finally, the comparison with other researches shows that the approach presented in this thesis includes broader range of tunes genres which is very significant for music emotion recognition system. Another very important distinction is that emotions are described on continuous plane instead of categories which resulted in much more supported adjectives for emotion description. And the most important advantage of the described approach is that it uses a new set of music features which are based on music perception by human brain, and therefore provides more relevant results which are easily understandable by human.

Publications list

Journal Papers

1. Endrjukaite, T. and Kosugi, N. (2012), "Music Visualization Technique of Repetitive Structure Representation to Support Intuitive Estimation of Music Affinity and Lightness," Journal of Mobile Multimedia Volume 8 Issue 1, pp. 49-71, 2012.
2. Endrjukaite, T. and Kiyoki,Y.(2013), "Music Similarity Analysis through Repetitions and Instantaneous Frequency Spectrum," International Journal of Signal Processing Systems, Vol. 1, No. 2, pp. 170-176, 2013.
3. Endrjukaite, T. and Kiyoki, Y. (2014), "Emotion Identification System for Musical Tunes based on Characteristics of Acoustic Signal Data," Journal of Information Modeling and Knowledge Bases XXVI, pp. 88-107, 2014.

Conference papers

4. Endrjukaite, T. and Kosugi, N. (2012), "Time-dependent Genre Recognition by means of Instantaneous Frequency Spectrum based on Hilbert-Huang Transform". In Proc. of 15th International Conference on Information Integration and Web-based Applications & Services, pp. 64-71, Bali, Indonesia, December 3-5, 2012.
5. Endrjukaite, T. and Kiyoki,Y (2013), "Music homogeneity analysis through instantaneous frequencies". In Proc. of the International Conference on Advances in Mobile Computing & Multimedia, ACM International Conference Proceeding Series pp.405-411, Vienna, Austria, December 2-4, 2013.
6. Endrjukaite, T., Okano, K., Yamamoto, H., and Hatayama A. (2015), "Renewable Energy Introduction in Latvia with Best Energy Mixture Model based on a Concept of a Combined Power System of Biomass and Wind". In Proc. of Conference on Energy, Economy, and Environment, pp. 527-532, Tokyo, Japan, January 27-28, 2015.
7. Endrjukaite, T. and Kiyoki, Y. (2015), "Environmental Meaning Analysis Approach Case Study Hawaii, USA". In Proc. of the 17th International Electronics Symposium, co-sponsored IEEE conference, pp.151-156, Surabaya, Indonesia, September 29-30, 2015.

Bibliography

1. J.Sloboda. Exploring the musical mind: cognition, emotion, ability, function, Oxford University Press, 2005.
2. P.Juslin and J.Sloboda. Music and emotion: Theory and research. Oxford University Press, 2001.
3. K.Hevner. Experimental Studies of the Elements of Expression in Music. American Journal of Psychology, Volume 48, pp. 246-268, 1936.
4. J.Foote. Visualizing music and audio using self-similarity. In Proceedings of the 7th ACM International Conference on Multimedia, pp. 77-80, 1999.
5. D.Levitin. This Is Your Brain on Music: The Science of a Human Obsession. Dutton Adult Press, 2006.
6. P.Lindsay, N. Donald. Human Information Processing: Introduction to Psychology. Academic Press Inc, 2nd edition, 1977.
7. T.Endrjukaite and N.Kosugi, Music visualization technique of repetitive structure representation to support intuitive estimation of music affinity and lightness, Journal of Mobile Multimedia, Vol. 8(1), pp. 49-71, 2012.
8. N.Huang et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Article in Proc. R. Soc. Lond. A, vol. 454 no. 1971, pages 903-995, 1998
9. T.Taniguchi. Ongaku to Kanjoh (Music and Emotion). Kitaohji-Shobo, 1998.
10. N.Huang and S.Shen. Hilbert–Huang Transform and Its Applications. Interdisciplinary Mathematical Sciences; volume 5. World Scientific Publishing Co. Pte. Ltd., 2005.
11. J.Grekow and Z.W.Ras. Detecting Emotions in Classical Music from MIDI Files. Springer Lecture Notes in Computer Science, Vol. 5722, pp. 261-270, 2009.
12. Y.Lin, X.Chen, D.Yang. Exploration of Music Emotion Recognition based on MIDI. In Proceedings of 14th International Society for Music Information Retrieval, 2013.
13. R.Thayer. The Biopsychology of Mood and Arousal. New York: Oxford University Press, 1989.
14. K.Trochidis, et al. Psychophysiological Measures of Emotional Response to Romantic Orchestral Music and their Musical and Acoustic Correlates. In

Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval, pp. 44-57, 2012.

15. B.Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In Proceedings of the 1st International Symposium on Music Information Retrieval, 2000.
16. G.Tzanetskis and P.Cook. Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing, Vol. 10(5), pp. 293-302, 2002.
17. T.Endrjukaite and N.Kosugi, Time-Dependent Genre Recognition by means of Instantaneous Frequency Spectrum. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, 2012.
18. T.Endrjukaite and Y.Kiyoki, Music Similarity Analysis through Repetitions and Instantaneous Frequency Spectrum. In Proceedings of the 5th International Conference of Signal Processing Systems, 2013.
19. T.Endrjukaite and Y.Kiyoki. Music Homogeneity Analysis through Instantaneous Frequencies, In Proceedings of the 11th International Conference on Advances in Mobile Computing & Multimedia, 2013.
20. J.Russell. A circumplex Model of Affect. Journal on Personality Social Psychology, Vol. 39, pp. 1161-1178, 1980.
21. D.Yang and W.Lee. Disambiguating Music Emotion using Software Agents. In Proceedings of International Conference on Music Information Retrieval, 2003.
22. T.Li and M.Ogihara. Detecting Emotion in Music. In Proceedings of International Conference on Music Information Retrieval, 2003.
23. D.Liu, L.Lu, H.Zhang. Automatic Mood Detection and Tracking of Music Audio Signals. In IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14(1), pp. 5-18, 2006.
24. V.Carvalho and C.Chao. Sentiment Retrieval in Popular Music based on Sequential Learning. In Proceedings of ACM SIGIR, 2005.
25. T.Li and M.Ogihara. Content-based Music Similarity Search and Emotion Detection. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 5, pp. 705-708, 2004.
26. D.Huron. Perceptual and Cognitive Applications in Music Information Retrieval. In Proceedings of International Symposium on Music Information Retrieval (ISMIR), 2000.
27. M.Heideman, D.Johnson, C.Burrus. Gauss and the History of the Fast Fourier Transform. IEEE ASSP Magazine vol, 1(4), pp. 14-21, 1984.

28. X, Hu. Music and Mood: Where Theory and Reality Meet. In Proceedings of iConference, 2010.
29. B.Laar. Emotion Detection in Music, a Survey. In Proceedings of the 4th Twente Student Conference on IT, 2006.
30. H.Elfenbeinand and N.Ambady. On the Universality and Cultural Specificity of Emotion Recognition: a meta-analysis. Psychological Bulletin Vol. 128, pp. 203-235, 2002.
31. R.Lazarus. Progress on a Cognitive-motivational-relational Theory of Emotion. Journal American Psychologist, Vol. 46(8), pp. 819-834. 1992.
32. A.Tellegen, D.Watson, L.Clark. On the Dimensional and Hierarchical Structure of Affect. Journal of Psychological Science Vol. 10(4), pp. 297-303, 1999.
33. G.Tzanetskis, G.Essl, P.Cook. Automatic Musical Genre Classification of Audio Signals. In Proceedings of International Symposium on Music Information Retrieval, 2001.
34. F.King. Hilbert Transforms: Encyclopedia of Mathematics and its Applications, Vol.1, Cambridge University Press, 2009.
35. T.Endrjukaite and Y.Kiyoki. Emotion Identification System for Musical Tunes based on Characteristics of Acoustic Signal Data. Journal of Information Modeling and Knowledge Bases XXVI, Vol. 272, IOS Press, 2014.
36. C.K.Chui. An Introduction to Wavelets: Wavelet Analysis and Its Applications. Academic Press 1 edition, 1992.
37. J.Smith. Mathematics of the Discrete Fourier Transform (DFT): with Audio Applications. W3K Publishing, 2nd edition, 2007.
38. E.Chu. Discrete and Continuous Fourier Transforms: Analysis, Applications and Fast Algorithms. Chapman and Hall, 1 edition, 2008.
39. H.G.Stark. Wavelets and Signal Processing: An Application-Based Introduction. Springerr, 1st edition, 2005.
40. J.Robinson. Deeper than Reason: Emotion and its Role in Literature, Music, and Art. Clarendon Press, 2007.