Automated soundtrack generation for fiction books backed by Lövheim’s Cube emotional model

Alexander Kalinin[1][0000-0002-0012-1692] and Anastasia Kolmogorova[2][1111-2222-3333-4444]

[1] Verbalab, Krasnoyarsk, Russia
xyz@verbalab.ru
[2] Siberian Federal University, Krasnoyarsk, Russia
nastiakol@mail.ru

Abstract. One of the main tasks of any work of art is transferring emotion conceived by the author to its recipient. When using several modalities a synergistic effect occurs, making the achievement of the target emotional state more likely. In reading, mostly, visual perception is involved, nevertheless, we can supplement it with an audio modality with the soundtrack’s help via specially selected music that corresponds to the emotional state of a text fragment.

As a base model for representing emotional state we have selected physiologically motivated Lövheim’s cube model which embraces 8 emotional states instead of 2 (positive and negative) usually used in sentiment analysis.

This article describes the concept of selecting special music for the “mood” of a text extract by mapping text emotional labels to tags in LastFM API, fetching music data to play and experimental validation of this approach.

Keywords: Sentiment Analysis, Lövheim’s cube, Soundtrack, Multi-modal art, Emotions

1 Soundtrack for fiction books as element of multimodal art

The main purpose of any piece of art is to communicate ideas and emotions strongly felt by the artist to other people. The more modalities are engaged into the process of transferring artistic message, the more powerful is the aimed emotional impact. That’s why multimodal art is being rapidly developed now with the growth of technology.

The numerous examples of such multimodal art communication are visual performances during musical concerts, interactive installations enabling kinesthetic experience, musical support for paintings exhibitions. All these media enabling visual, audial and kinesthetic perceptive channels have a synergetic effect on the spectator plunging him in a certain atmosphere, in a certain emotional state.

Listening to an appropriate music while reading fiction books can facilitate emotional perception of ideas and feelings expressed by the writer creating a sort of consensuality between author and readers. In this case, reading (visual modality) is supplemented by listening to music (audial modality) and music plays the same role as soundtrack for films where sound creates additional effect to visual scenes. Thus, the main question arises: how to select music to supplement visual models and images, which emerge in readers' mind while reading a story?

The process of creating soundtrack for a text from a very abstract level seems to be a list of the steps below:

1. Split the text into a number of extracts.
2. Define the mood (emotion) of each extract.
3. Find the music, which corresponds to a particular emotion.
4. Map music to text extracts by emotion.

Despite limited number of steps to do and transparency of its logics, to achieve this task of selecting needed music for a given text we have to solve several problems:

1. What set of emotions to choose, i.e. what emotional model is mostly suitable for our task?
2. How to recognize a particular emotion for a given piece of text?
3. What music should we choose for a given emotion?
4. How to align the time of playing music tracks with the speed of reading?

2 Emotional model

A corner-stone problem for the task of mapping texts to music by a certain emotion is to choose an emotional inventory − a set of possible emotional states that user can experience − thus, an emotional model. For it we have to face psychology and neurology querying for available theoretical frames being developed in those fields. Searching for a model viable for computer application building, we were guided by the criterion of the model's capacity to provide numerical or logical input scheme and explicit deterministic approach for emotions differentiation. There are 3 main models to mention as suitable for our needs:

1. Binary model
2. PAD (*Pleasure* , *Arousal* , *Dominance*) model
3. Lövheim Cube

Binary model is a very simplistic approach that assumes that emotion can be either positive or negative. This very model is developed mostly within sentiment analysis [1], an applied field of computational linguistics, and is used to estimate emotional 'color' of a text − whether the text represents positive or negative writer's attitude towards some subject under discussion. Such approach is used for tasks of evaluating different kinds of reviews like movie reviews, internet shop merchandise reviews, restaurant reviews, and also mining users’ opinions on news articles, marketing research etc [2].

As these applied tasks are dealing with the large amount of data and are computationally expensive this simple model is very handy as it doesn’t bring additional complexity and can provide meaningful insights. The mentioned above model also provides continuous interpretation when emotional state can be measured between range of -1 (very negative emotion) and +1 (very positive emotion); 0 value can be interpreted as neutral emotion [3].

By the reason of the model’s simplicity, the algorithms that are based on it are unable to differentiate between more nuanced kinds of positive and negative emotions, as they (emotions) don’t certainly exist in one dimension. For example, anger and fear are both considered as negative emotions, but they are for sure very dissimilar, and experiencing anger is definitely different from experiencing fear. Moreover, fiction books provide much larger palette of emotions than the narrow binary system could embrace. In other words, the binary emotion model seems to be unproductive for our purposes.

PAD [4] uses three-dimensional axis measurable numerically.

The first axis is *Pleasure-Displeasure Scale that* measures how pleasant or unpleasant human feels about something. For instance, both anger and fear are unpleasant emotions, and both score on the displeasure side. However, joy is a pleasant emotion.

The *Arousal-Nonarousal Scale* measures how energized or soporific one feels. It is not the intensity of the emotion – for grief and depression can be low arousal intense feelings. While both anger and rage are unpleasant emotions, rage has a higher intensity or a higher arousal state. On the contrary, boredom, which is also an unpleasant state, has a low arousal value.

The *Dominance-Submissiveness Scale* represents the controlling and dominant versus controlled or submissive emotions one feels. For instance, while both fear and anger are unpleasant emotions, anger is a dominant emotion, whereas fear is a submissive one.

A more abbreviated version of the model uses just 4 values for each dimension, providing only 64 values for possible emotions [5]. For example, anger is a quite unpleasant, quite aroused, and moderately dominant emotion, while boredom is slightly unpleasant, quite unaroused, and mostly non-dominant.

The drawbacks of both models are conditioned by the fact they are phenomenological, i.e. they put an introspection to be the main method for defining to what degree the emotion is pleasant or unpleasant, negative or positive etc. The experiencer discovers and defines the quantitative parameters of emotions, so the estimation based on such approach can be quite subjective.

The last suggested model – Lövheim’s Cube. It takes objective parameters to be the arguments of emotional functions [6]. These parameters are mixtures or proportions of three monoamine neurotransmitters: serotonin, dopamine and noradrenaline. The balance of these monoamines forms a three-dimensional space represented by the following visual model:

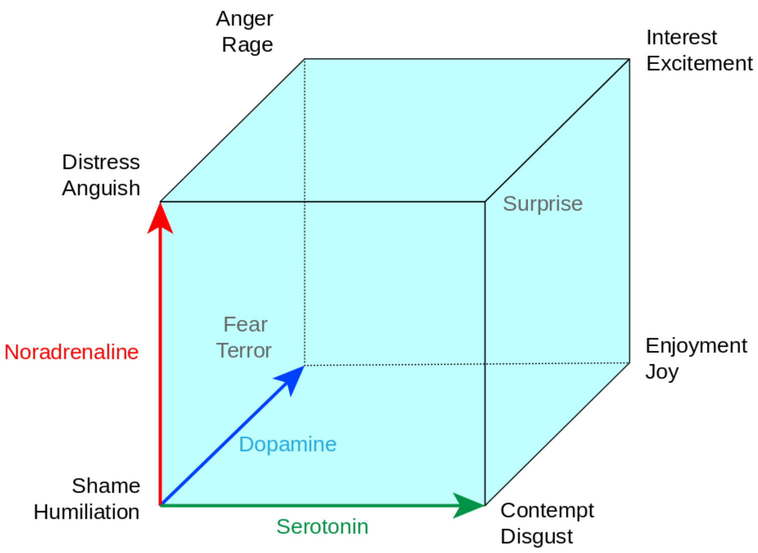

Fig. 1. Visual representation of Lövheim's Cube emotional model

Anger is, for example, according to the model, produced by the combination of low serotonin, high dopamine and high noradrenaline.

The reason for us to choose this model instead of other two ones is following:

1. Unlike previous two models this one is not subjective and introspective and represents emotional state to be funded by neurosomatic process [7].
2. It provide 8 basic emotions as function of monoamine balance and suggests continuous numerical scale estimation for intermediate emotions, so any emotional state can be presented as a vector.

3 Relevant data acquisition

After choosing the emotional inventory to apply for mapping texts to music we should define the mechanism of labeling certain text extract with emotions from Lövheim's cube set. Such objective can be achieved by using automated tools (like applying current sentiment analysis techniques) or manually (labeling text extras for by human readers).

Unfortunately, we couldn't use automated tools based on machine learning approach. For training a model to predict a certain emotion for a given text sample we need data which had been previously labeled by human readers.

The dataset must contain a sentence or a paragraph with mapped emotion label from Lövheim's cube inventory. But in spite of the fact that there lots of data with mapped binary emotions [8] and data containing PDA model metadata[9] we didn't succeed to find any text dataset with Lövheim's emotions set neither for English nor for Russian languages.

Taking into account the lack of needed data and bearing in mind the fact that our aim was only to prove the concept of soundtrack generation for a given "emotion-text" mapping, we decided not to use existing data-driven sentiment extraction pipeline and to follow some listed below steps.

1. Select Russian novel as a source of text data.
2. Split the text into small extracts showing emotional "persistence" and manifesting mainly one emotional "color".
3. Manually label the collection of extracts using evaluation from, at least, four assessors.
4. Finally, for each text extract set the emotion which was selected for a given extract for most times during manual labeling.

As a text data source we selected Russian novel entitled "Pismovnik" ("Epistolary novel"), which consists of many letters written by a man and a woman in love while they were apart. The text from this novel mimics real personal epistolary style and represents a wide palette of emotions.

First, we had split the text into short extracts, which were considered to be emotional invariants, i.e. text where emotion sentiment is stable and doesn't fluctuate. While splitting the text of novel "Pismovnik" we discovered that the optimal length for such invariants was about 300 characters.

After splitting the text into emotional extracts we needed to label it with large number of human assessors. For quick fetching the result set, large data coverage and randomisation of tasks we decided to use crowdsource platform "Toloka" developed by Yandex team. The main reason we selected it was the fact that this platform provided big community of native Russian speakers who could correctly perceive the underlying emotion and label the emotion of a text extract.

We asked human assessors to read a text extract and to select the emotion one from eight possible emotions that mostly suits the sentiment of the text, or if either emotion is suitable or text was very neutral an assessor had to select "neutral" variant. Tasks with text extracts were distributed randomly, and each extract had to be assessed at least by 4 different assessors. After pushing 769 text's extracts we fetched 2893 labeled results provided by 142 native Russian-speaking assessors. Than we processed the novel text once again and labeled each extract with the label that was selected by the most times from 4 different assessments.

4 Mapping music to text extras

To get meta-information relevant to music we decided to use LastFM service. The main reason to choose such resource was the availability of a big set of data concerning various kinds of music: genres, number of times played (scrobbled) and search tags assigned to different tracks by users community. This tags' folksonomy facilitated our search for different music tracks for a set of emotion labels.

For generating a list of music tracks we folded previously labeled text extras into a sequence of emotion items where no item had the same emotion as its neighbor (see. Fig.2).

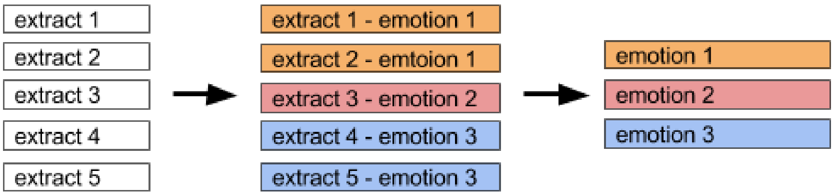

Fig. 2. Scheme for convolution of paragraphs into emotional labels sequence

For each item in this sequence we made a search query to LastFM API to get 100 tracks with search tag matching emotion label. After each tag query fetched music tracks, the list was randomized, so all the songs for each emotion item were shuffled (see Fig. 3). 100 tracks were selected to provide potentially excessive time playing of single emotion track list as we could not definitely know how long it might take to read a text extract.

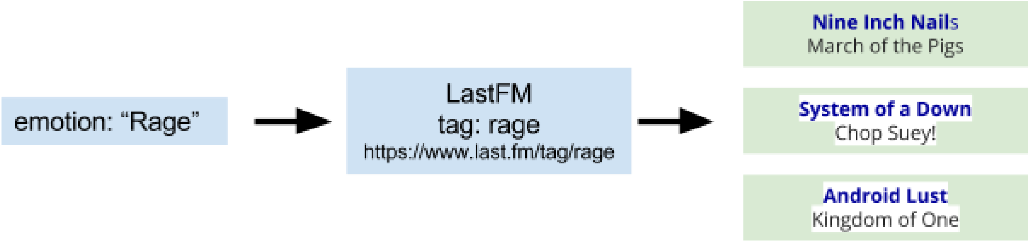

Fig. 3. Pipeline for fetching track list for emotion label

5 Evaluation of concept

It is to notice that in current project we didn't plan to solve the problem of aligning reading speed with switching track lists as reading speed is a personal constantly changing parameter, which is hard to track. Of course, such task is very crucial and definitely will arouse in case of commercial and production release of this model, but for now, we discuss a prototype with a "proof of concept" aim. The achievement of this goal presupposes the evaluation of how well a music track with certain emotion – "mood" – can complement reading a text extract with the same mood. To answer the question we have proceeded with the estimation process.

Each of 10 volunteers was given at least 30 text extracts. Each text extract was shown on the screen. During one extract demonstration no other extracts labeled with different emotion were shown, and music from track list matching this emotion labeled was played in headphones. After reading each extracts each volunteer was asked to rank the selected music track from 1 to 10, where 10 meant "music perfectly complements the text", and 1 meant "this music is completely inappropriate for this text". After volunteer assessed one emotion mapping one was shown a text extract with another emotion and played music with a matching tag. Overall, 328 text extracts were evaluated. Results are presented in table 1.

Table 1. Results for evaluation "text-music" mappings

Emotion	Mean	Standart deviation
Anger/rage	7.3	1.43
Passion/excitement	6.9	1.21
Contempt/disgust	3.1	0.31
Enjoyment/joy	2.3	0.42
Shame/humiliation	2.1	0.35
Distress/anguish	7.9	0.51
Surprise/startle	2.8	0.39
Fear/terror	6.7	0.26
Neutral	5.2	1.17

As we can see in Table 1 this concept works well for mapping distress, passion, rage and fear emotions, for other ones and neutral texts the estimations are not satisfactory.

6 Conclusion and further work

As our experiment has shown an approach for generating soundtracks for fiction books by mapping text to music via emotion model that uses Lövheim's Cube model can do well for a number of emotions, but not for all, and that shows that current version of our approach can't be taken as somewhat applicable. However, good results for 4 emotions encourage us to look for workarounds and further work that can involve following directions:

- Automated text labeling using sentiment analysis techniques and machine learning.
- Using extended and modified music search queries (for example look for such tags as “anger, wrath, aggression, war, battle” etc. for “rage” text label;
- Adding filters for fetched music (for example strip out tracks with lyrics as voices and song narratives can interrupt reader);
- Solving problems for automated tracking speed of reading and switching tracks intelligently.

References

1. Su, Fangzhong; Markert, Katja (2008). "From Words to Senses: a Case Study in Subjectivity Recognition" (PDF). *Proceedings of Coling 2008, Manchester, UK*.
2. Hu, Minqing; Liu, Bing (2004). "Mining and Summarizing Customer Reviews". *Proceedings of KDD 2004*.
3. Kim, S. M.; Hovy, E. H. (2006). "Identifying and Analyzing Judgment Opinions." (PDF). *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY*.
4. Mehrabian, Albert (1980). *Basic dimensions for a general psychological theory*. pp. 39–53.
5. Lance, Brent; et al. (2008). "Relation between Gaze Behavior and Attribution of Emotion". In Prendinger, Helmut. *Intelligent virtual agents: 8th international conference*. IVA. pp. 1–9.
6. Lövheim, H (2012). "A new three-dimensional model for emotions and monoamine neurotransmitters". *Med Hypotheses*. **78**: 341–348.
7. Talanov, M; Toschev, A. "Computational Emotional Thinking and Virtual Neurotransmitters.". *International Journal of Synthetic Emotions*. **5** (1): 1–8.
8. Affective Text: data annotated for emotions and polarity. Dataset. http://web.eecs.umich.edu/~mihalcea/downloads.html#affective
9. EmoBank. 10k sentences annotated with Valence, Arousal and Dominance values. Dataset. https://github.com/JULIELab/EmoBank