Master's Thesis

Musemo: Express Musical Emotion Based on Neural Network

Wooyeon Kim

Department of Urban and Environmental Engineering
(Convergence of Science and Arts)

Graduate School of UNIST

2020

Musemo: Express Musical Emotion Based on Neural Network

Wooyeon Kim

Department of Urban and Environmental Engineering
(Convergence of Science and Arts)

Graduate School of UNIST

Musemo: Express Musical Emotion Based on Neural Network

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Wooyeon Kim

12. 13. 2019
Approved by

Advisor
JongEun Lee

Musemo: Express Musical Emotion Based on Neural Network

Wooyeon Kim

This certifies that the thesis of Wooyeon Kim is approved.

12. 13. 2019

Advisor: JongEun lee

Jaeweon Cho: Thesis Committee Member #1

Bradley Tatar: Thesis Committee Member #2

Abstract

Music elicits emotional responses, which enable people to empathize with the emotional states induced by music, experience changes in their current feelings, receive comfort, and relieve stress (Juslin & Laukka, 2004). Music emotion recognition (MER) is a field of research that extracts emotions from music through various systems and methods. Interest in this field is increasing as researchers try to use it for psychiatric purposes. In order to extract emotions from music, MER requires music and emotion labels for each music. Many MER studies use emotion labels created by non-music-specific psychologists such as Russell's circumplex model of affects (Russell, 1980) and Ekman's six basic emotions (Ekman, 1999). However, Zentner, Grandjean, and Scherer suggest that emotions commonly used in music are subdivided into specific areas, rather than spread across the entire spectrum of emotions (Zentner, Grandjean, & Scherer, 2008). Thus, existing MER studies have difficulties with the emotion labels that are not widely agreed through musicians and listeners. This study proposes a musical emotion recognition model "Musemo" that follows the Geneva emotion music scale proposed by music psychologists based on a convolution neural network. We evaluate the accuracy of the model by varying the length of music samples used as input of Musemo and achieved RMSE (root mean squared error) performance of up to 14.91%. Also, we examine the correlation among emotion labels by reducing the Musemo's emotion output vector to two dimensions through principal component analysis. Consequently, we can get results that are similar to the study that Vuoskoski and Eerola analyzed for the Geneva emotion music scale (Vuoskoski & Eerola, 2011). We hope that this study could be expanded to inform treatments to comfort those in need of psychological empathy in modern society.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The study began with an attempt to converge science and arts at Science Walden, a research center at the Ulsan National Institute of Science and Technology. Science Walden provides a foundation for using interdisciplinary and convergent approaches to solve modern social problems like personal alienation, intergenerational conflict, and income inequality in our community. Its research does not target specific groups but all people. From this context, this study starts with the question of whether music and science can be applied to society to address the lack of social-psychological empathy. Our final goal is to use music in society to comfort people by creating bonds through mutual emotional understanding and experiences. The first step towards reaching this goal is to understand the underlying emotions in music. This paper describes the process and results of this first step.

Music is an art that reflects human society and culture. It is integrated into our daily lives. In the words of Patrik N. Juslin, a noted music psychologist, “Music is a source of aesthetic pleasure that brings people and culture together, and it may contribute to their health and well-being” (Juslin, 2019). He also demonstrates that music arouses an emotional state in the listener (Juslin, Liljeström, Laukka, Västfjäll, & Lundqvist, 2011; Juslin, Liljeström, Västfjäll, Barradas, & Silva, 2008). Another music psychologist, Petri Laukka, found that listeners use music to change and release their emotions and match their current emotions. They also use music for enjoyment, comfort, and stress relief (Juslin & Laukka, 2004).

Psychologists have developed several models to classify emotions, including Plutchik’s wheel of emotions (Plutchik, 1980), Ekman’s basic emotions (Ekman, 1999), and Russell’s circumplex model of affect (Russell, 1980). However, musical emotions are different from everyday human emotions. Musical emotion has important factors and categories related to musical features in the field of music psychology, including Hevner’s adjectives circle (Hevner, 1936), Juslin’s two-dimensional emotion space in music (Juslin, 2019), and Zentner and Scherer’s Geneva emotional music scale (GEMS) (Zentner et al., 2008). Since there are many different emotion theories, it is crucial to choose a compatible emotion theory that suites the research purposes.

1.2 Motivations

Among the aspects that need to be addressed when interpreting and applying emotions, we focus on two main issues, the first being on the lack of application on music-specific classification systems. According to Juslin, daily emotion and musical emotion are defined differently (Juslin, 2019). However, emotional classifications proposed by non-music specific psychologists are often used in MER studies. Second, this field remains underexplored, so it is difficult to find precedents for the applicable transformation method; for example, the data preprocessing method is not studied compared with the study of the speech emotion recognition field. The speech recognition field has achieved significant growth, and there are many trials and errors involved in speech signal processing. Nevertheless, music signal processing has a lack of various experiments and investigations to achieve proper musical data preprocessing. To attempt solving these problems, we design a MER model called “Musemo” using musical emotion labels following the GEMS system.

1.3 Research Aims

The main goal of this study is to design a neural network model that extracts musical emotions from music files, compare and present various data preprocessing methods, and finally, complete Musemo, a model with less than 15% error. We identify several additional sub-goals. First, to determine the minimum length of a music file that enables learning with an error rate of 20% or less. Second, to use two data preprocessing methods (short-time Fourier transforms (STFTs) and Mel spectrograms) to convert the signals from a music file into two-dimensional representations, and to determine which conversion process is more suitable for neural networks. Third, to map the locations of learned emotion labels in two dimensions to analyze the correlations between musical emotions and to compare these correlations with other studies.

2 Literature Review and Related Work

2.1 Emotion Classification

An emotion is a complex psychological state involving three components: a subjective experience, physiological response, and behavioral or expressive response. Each component is linked to the three categories for emotion theories: psychological, physiological, and cognitive theories. In this paper, we will focus on the psychological theories of emotion.

2.1.1 Psychological Emotion Theories

- Paul Ekman, Six basic emotions (Ekman, 1999)

Ekman studied the relationship between emotions and facial expressions. Because he focused on the universality of emotions, he suggests six basic emotions: happiness, sadness, fear, surprise, anger, and disgust (Figure 1). Charles Darwin argued that facial expressions related to emotions are universal and appear commonly or identically in different cultures (Darwin & Prodger, 1998). Additionally, Carroll Izard performed a similar experiment observing eight different cultures and presented evidence for universality in basic emotions (Izard, 1971). Ekman's six basic emotions can be seen in a universal emotion classification, not limited by cultural differences.

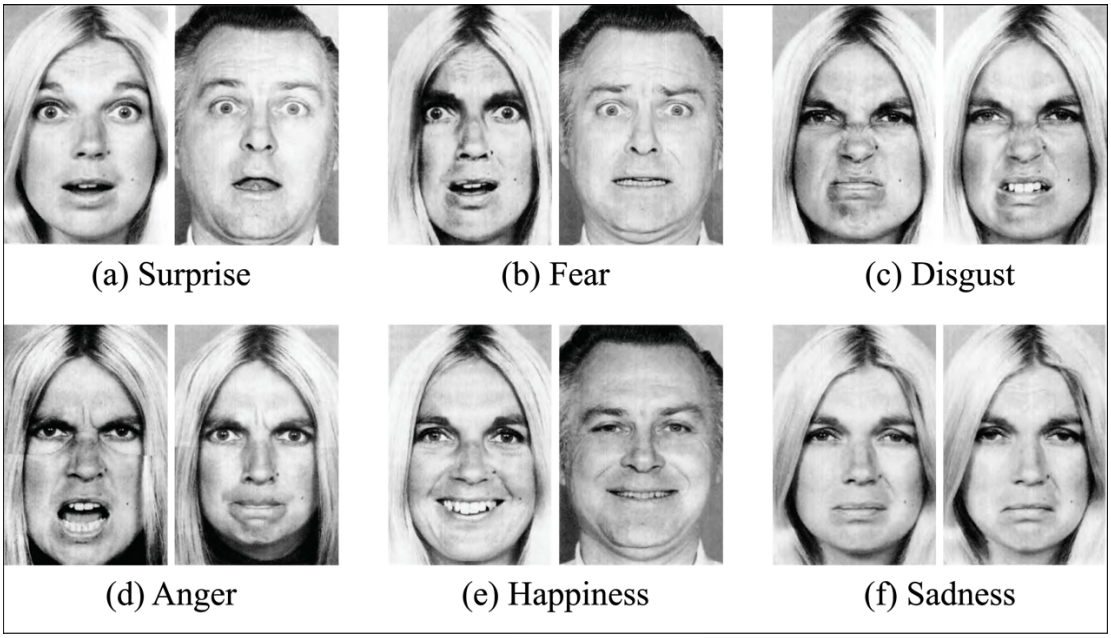


Figure 1. Ekman's six basic emotion, 1999

- Robert Plutchik, Wheel of emotions (Plutchik, 1980)

Robert Plutchik is widely known for his contribution to emotion theories. His main idea is that there are a few primary emotions, and each primary emotion is identifiable in other mammals. His second idea is that emotions are evolutionary adaptations. Plutchik derived this idea from Darwin (Darwin & Prodger, 1998), who inspired his overall psychology studies. Plutchik's other premise is that there is a link between the emotional lives and personalities of people. Based on these premises, Plutchik suggests that mammals have eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Figure 2 shows the wheel of emotions, with the primary emotions located in the second circle. An additional remarkable feature is the blending areas between primary emotions. Love is a blend of joy and trust, implying that the combination of two primary emotions makes another emotion. These combinations lead to love, submission, awe, disapproval, remorse, contempt, aggressiveness, and optimism. Plutchik's psychoevolutionary theory has a significant contribution to the psychology and psychiatry field (Plutchik, 1980).

Figure 2. Plutchik's Wheel of Emotions, 1980

- James Russell, Circumplex model of affects (Russell, 1980)

James Russell's circumplex model of affects is the most widely used in emotion research. Figure 3 shows the circumplex model. Twenty-eight emotional terms are distributed in a circle. The horizontal and vertical axes denote pleasure–displeasure (valence) and activation–deactivation (arousal), respectively. This model, which is also called the valence–arousal model, is mostly used in emotional research. Recently, a circumplex model was developed by Michelle Yik, James A. Russell, and James H. Steiger (Yik, Russell, & Steiger, 2011) (see Figure 4). This model also has a two-dimensional circular space, but space is devoted to 12 points of core affect. The word "core" refers to a form of emotional response that acts as a type of core knowledge about whether an object or event is rewarding or threatening, helpful or harmful, and calling for acceptance or rejection (Barrett, Mesquita, Ochsner, & Gross, 2007). This model is attractive because it visually and spatially represents similarities and differences between neighboring emotions.

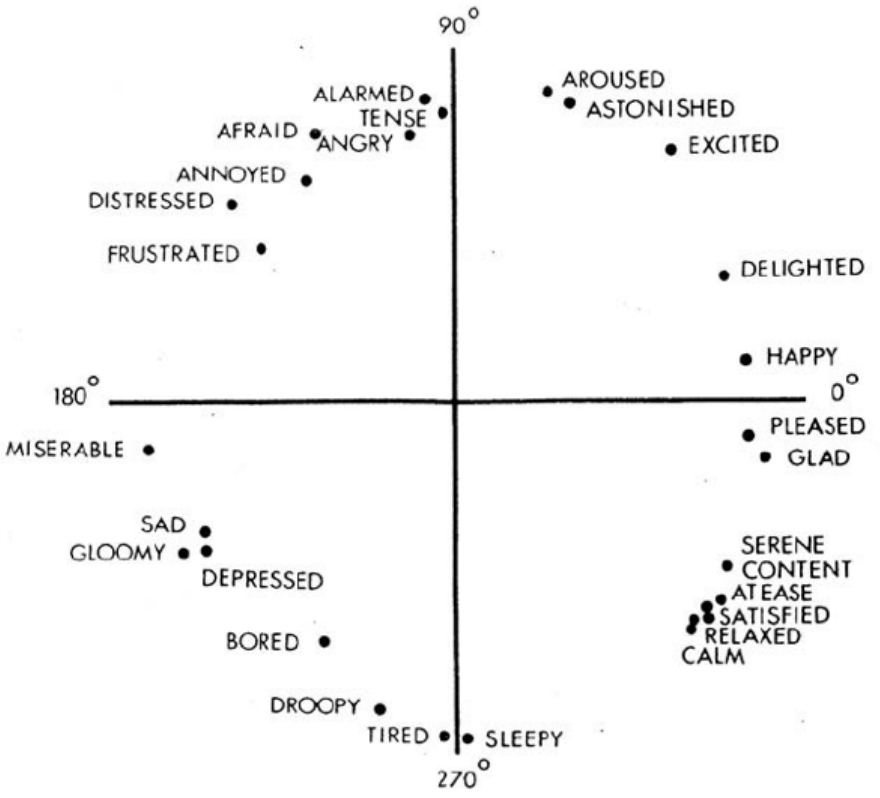


Figure 3 Russell's Circumplex Model of Affect, 1980

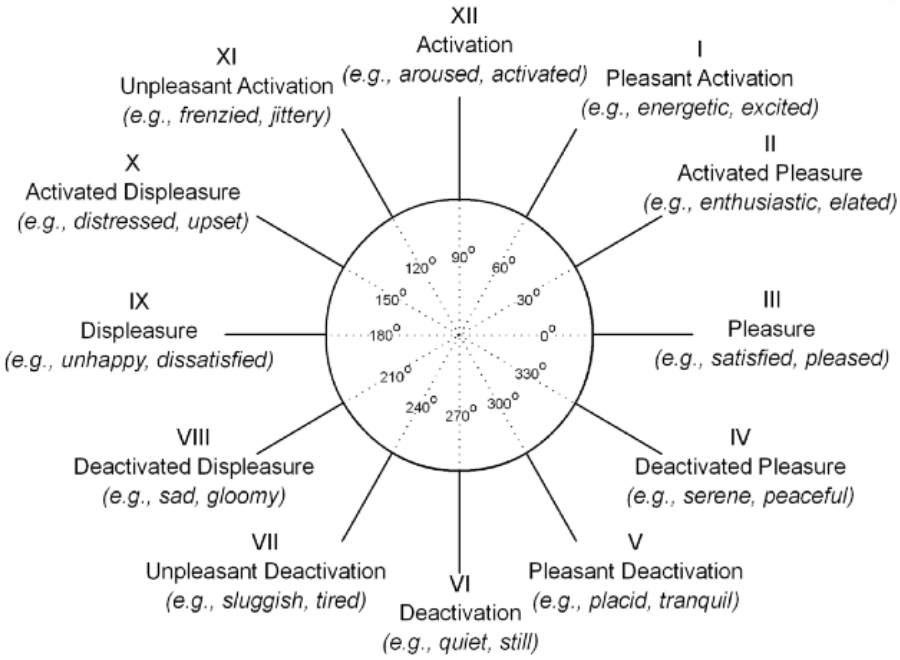


Figure 4. A 12-Point Affect Circumplex, 2011

In this section, we introduced three of the most famous models of psychological emotion classification theory. Psychologists suggest that there are many universal emotions experienced by people around the world, while also believing that emotional experiences can be subjective.

2.1.2 Musical Emotion Theories

- Kate Hevner, Adjective circle (Hevner, 1936)

Hevner argues that discrepancies in the musical emotions perceived by listeners occur due to the different meanings that listeners place on certain words. Her solution was to develop a unique self-reporting scale for musical expression that aims to capture a wide range of emotional word categories. She created an adjective circle (Figure 5) consisting of eight groups containing many emotional terms in a circular composition. The terms in each cluster indicate words with similar or adjacent meanings.

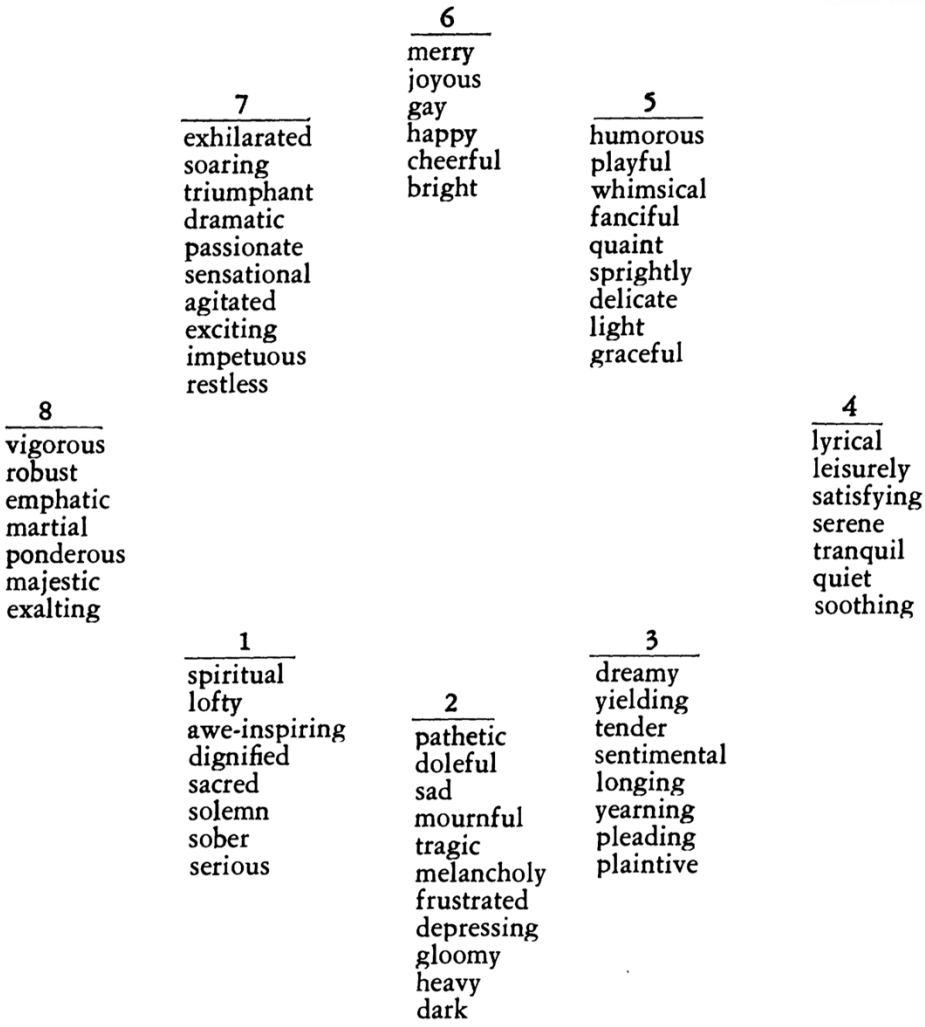


Figure 5. Hevner's Adjective Circles, 1936

- Patrik N. Juslin, Two-dimensional emotion space in music (Juslin & Timmers, 2010)

Juslin suggests five basic emotions related to musical expression. As previously mentioned, the valence–arousal dimension is widely used to classify emotions. Therefore, Juslin maps basic emotions to musical expressions in two dimensions (Figure 6). Additionally, he demonstrates correlations between musical expressions and emotion labels used in psychology (Table 1) (Juslin, 2013). A statistical analysis of the relationship between musical expressions and emotions shows a somewhat surprising correlation (0.76 – 0.98); thus, attempting to group musical expressions and emotions are reasonable.

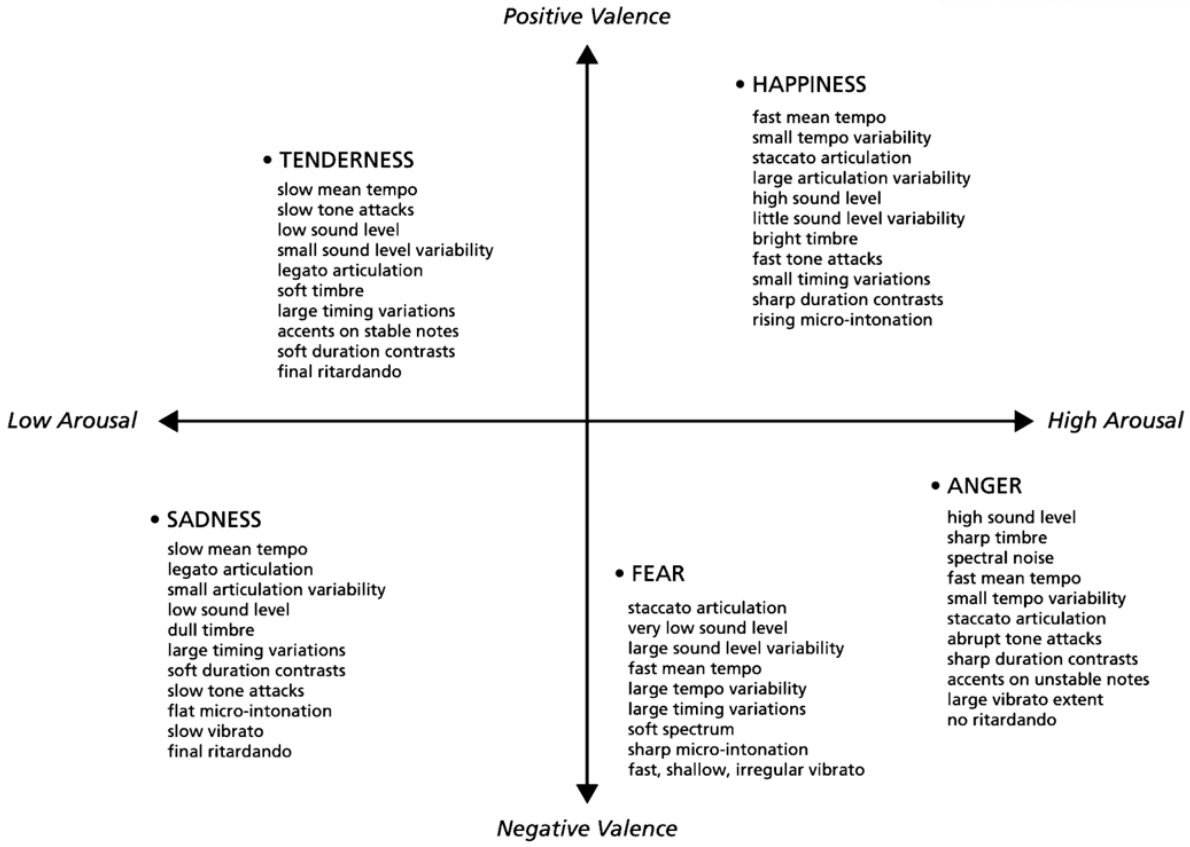


Figure 6. Juslin's Two-dimensional Emotion Space in Music, 2010

Table 1. Correlations between expression marks in musical scores and emotion, Juslin, 2013

Expression mark	Emotion label	Correlation (r)
Dolce	Tenderness	.98
Expressivo	Desire	.85
Furioso	Anger	.92
	Disgust	.79
Grave	Sadness	.88
Scherzando	Happiness	.76
Spiritoso	Surprise	.94
Temoroso	Anxiety	.97
	Fear	.82

- Marcel Zentner, Didier Grandjean, and Klaus R. Scherer, GEMS (Zentner et al., 2008)

Zentner, Grandjean, and Scherer develop a domain-specific musical emotion model, GEMS (Figure 7). GEMS is unique because it was created to describe induced musical emotions and has a level of subdivisions not provided by other models. To create this model, the researchers performed four

consecutive studies. We will discuss these studies in the next section. The first layer of the model includes 40 induced musical emotions, which are grouped into nine first-order musical emotion factors. These nine terms are grouped again into three second-order superfactors: sublimity, vitality, and unease.

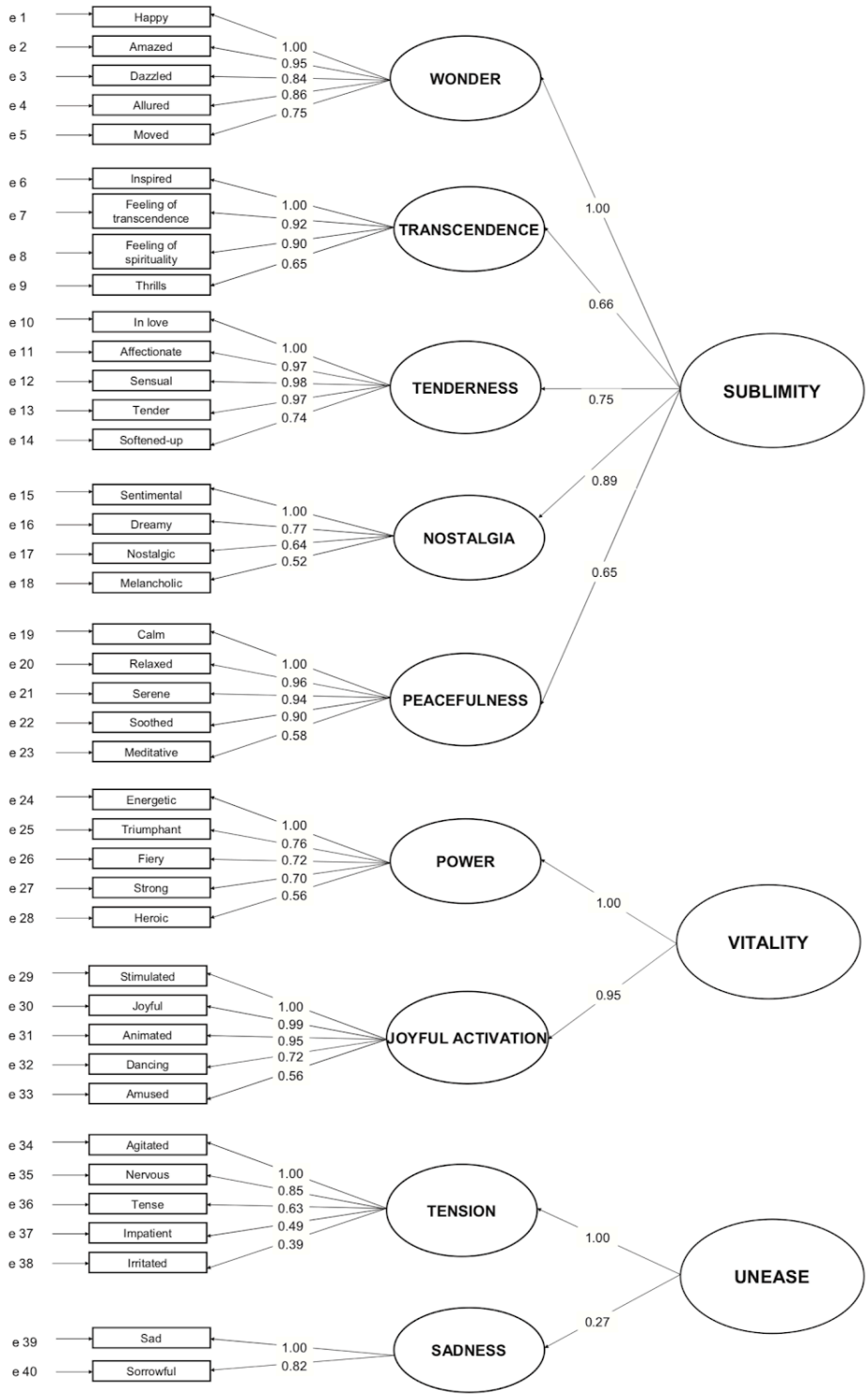


Figure 7. Factor Analysis of Geneva Emotion Music Scale, 2008

2.2 GEMS

In the previous sections, we briefly introduce emotion theories and classifications. In this section, GEMS will be described in greater detail. This model represents a new approach to capturing the essence of music, rather than fitting musical emotions into the categorical or dimensional emotion classifications suggested by psychologists. GEMS includes 40 musical emotion labels that were consistently chosen to describe musically induced emotional states in a relatively wide range of music and listener samples (Figure 7).

First, the researchers collected a list of relevant terms for feelings induced by music and feelings perceived in music. From this initial study, they created a final list of 146 terms, which were used in study 2. In the second study, the researchers investigated which of these 146 terms were relevant to music using factor analysis based on a questionnaire and identified eighty-one emotional terms as being relevant to music.

In the third study, the researchers examined whether the emotional state induced by music could be divided into subunits by which emotions could be classified. They found that a model with nine factors would best be fitted to the data; hence, they set these nine factors as first-order factors in GEMS. At this point, it is essential to understand the true meanings of these factors carefully. For example, it may seem strange to find "*happy*" in the first-order factor "*wonder*". The meaning of the French word "*heureux"* is different from the meaning of the English word happiness. "*Heureux"* represents happiness in the sense of loyalty and accomplishment rather than joy and satisfaction. While some factors in GEMS seem similar to those in general emotion classification models, those similarities may obscure subtle differences in factor meanings. For instance, musical sadness may differ from basic emotional sadness. Because daily emotion of sadness, such as depressed, gloomy, or unhappy, is rarely reported in response to music (Laukka, 2007). The researchers then grouped the nine first-order factors into three second-order factors (sublimity, vitality, and unease) based on their intercorrelations (Table 2). Elating and paradisiac characteristics could be classified as sublimity; joyful activation and power could be classified as vitality; and the two negative factors, tension and sadness, could be grouped into unease.

Finally, in the fourth study, the researchers examined the validity of the domain-specific GEMS model comparing with discrete and dimensional models. They argued that the domain-specific emotion checklist in GEMS tended to improve agreement among listeners in ratings relative to the checklist from the discrete and dimensional models. GEMS was created as a result of these four consecutive studies, and we have referred to the results as being suitable for musical emotion labels.

Table 2. Intercorrelations Among First-Order Musical Emotion Factors, GEMS, 2008

	Wonder	Transcendence	Tenderness	Nostalgia	Peacefulness	Power	Joyful activation	Tension	Sadness
Wonder									
Transcendence	.44								
Tenderness	.40	.42							
Nostalgia	.34	.33	.50						
Peacefulness	.33	.28	.39	.40					
Power	.40	.42	.31	.19	.06				
Joyful activation	.41	.25	.36	.14	.13	.38			
Tension	.04	.16	.12	.07	- .09	.29	.20		
Sadness	.12	.18	.20	.26	.05	.07	.08	.22	

Note. n = 801; all correlations r > .10 are significant at p < .01.

2.3 Music Emotion Recognition (MER)

Music information retrieval (MIR) research involves extracting and inferring important features from music (from external sources such as audio signals, symbolic representations, or web pages), and developing music indexing and search systems using these features (Schedl, Gómez, & Urbano, 2014). MIR subfields include musical feature extraction, similarity analysis, music classification, and applications. MER belongs to the music classification subfield. Its primary purpose is to model the association between music and emotion to facilitate emotion-based music organization, search, and indexing. The critical issue in this field is emotional taxonomy based on the conceptualization of emotions, which is mostly divided into dimensional and categorical approaches. Ekman's six basic emotions and Hevner's adjective circle exemplify the categorical approach, while Russell's valence–arousal model exemplify the dimensional approach. Yi-Hsuan Yang and Homer H.Chen reviewed the overall MER research field (Yang & Chen, 2012). They compared 26 cases by their emotional approach, the number of emotions, number of songs, genre, and the number of subjects per song (Table 3). According to this table, most studies were set to one genre, and the number of songs typically ranged from several hundred to one thousand. The number of subjects evaluating the emotion in each song ranged from 1 to 116.

Table 3. Comparison of selected work on MER, Yang & Chen, 2012

	# emotion	# song	Genre	# subject per song
[Feng et al. 2003]	4	223	pop	N/A
[Li and Ogihara 2003]	13	499	pop	1
[Li and Ogihara 2004]	3	235	jazz	2
[Wang et al. 2004]	6	N/A	classical	20
[Wieczorkowska 2004]	13	303	pop	1
[Leman et al. 2005]	15	60	pop	40
[Tolos et al. 2005]	3	30	pop	10
[Wieczorkowska et al. 2006]	13	875	pop	1
[Yang et al. 2006]	4	195	pop	> 10
[Lu et al. 2006]	4	250	classical	3
[Wu and Jeng 2006]	4	75	pop	60
[Skowronek et al. 2007]	12	1059	pop	6
[Hu et al. 2008]	5	1250	pop	< 8
[Laurier et al. 2008]	4	1000	pop	
[Wu and Jeng 2008]	8	1200	pop	28.2
[Trohidis et al. 2008]	6	593	pop	3
[Lin et al. 2009]	12	1535	pop	from AMG
[Han et al. 2009]	11	165	pop	from AMG
[Hu et al. 2009]	18	4578	pop	from Last.fm
[Korhonen et al. 2006]	2DES	6	classical	35
[MacDorman and Ho 2007]	2DES	100	pop	85
[Yang et al. 2007, 2009]	2DES	60	pop	40
[Yang et al. 2008]	2DES	195	pop	> 10
[Schmidt and Kim 2009]	2DES	120	pop	> 20
[Eerola et al. 2009]	3DES	110	sountrack	116
[Yang and Chen 2011b]	2DES	1240	pop	4.3

2DES represents two dimensional emotion space; 3DES represents three dimensional emotion space;
AMG and Last.fm are music websites to get emotion tags of music

3 Musemo

3.1 Dataset

To create a musical emotion recognition model, we collected annotated music samples. Our research objective is to use a domain-specific musical emotion model rather than using the discrete basic emotions or two-dimensional classifications commonly used in current MER studies. We obtained 400 music files and nine emotion labels following GEMS from the crowdsourcing game Emotify (Aljanaki, Wiering, & Veltkamp, 2014; Aljanaki, Wiering, & Veltkamp, 2016). The research team behind this game produced a public archive annotating musical emotions for classical, electronic, pop, and rock music. Because the categories of emotions in GEMS are in French, changes from the word wonder to amazement, transcendence to solemnity, and peaceful to calmness (Table 4) were made for more natural understanding to the user. The original dataset includes 400 music files, one-minute long each. The dataset also includes nine emotional annotations by participants, their moods before playing the game, ages, genders, and the mother tongues of the participants. Each song was annotated by an average of 20 people, with a standard deviation of 14.05. Participants labeled each emotion as "1" if they experienced that emotion and "0" if they did not. We average these labels to create a new set of data indicating the probability of each emotion is present in the song.

Table 4. GEMS categories with explanations used as in the game, Emotify, 2014

Emotional Category	Explanation	Superfactor
*Amazement	Feeling of wonder and happiness	Sublimity
*Solemnity	Feeling of transcendence, inspiration. Thrills	
Tenderness	Sensuality, affect, feeling of love	
Nostalgia	Dreamy, melancholic, sentimental feelings	
*Calmness	Relaxation, serenity, meditativeness	
Power	Feeling strong, heroic, triumphant, energetic	Vitality
Joyful activation	Feels like dancing, bouncy feeling, animated, amused	
Tension	Nervous, impatient, irritated	Unease
Sadness	Depressed, sorrowful	

3.2 Process

Figure 8 shows the overall structure of the process. By preprocessing the mp3 music files, we create a two-dimensional image of the input data. This input image is used to train a convolutional neural network (CNN) to predict the probability of the nine emotions present in the song.

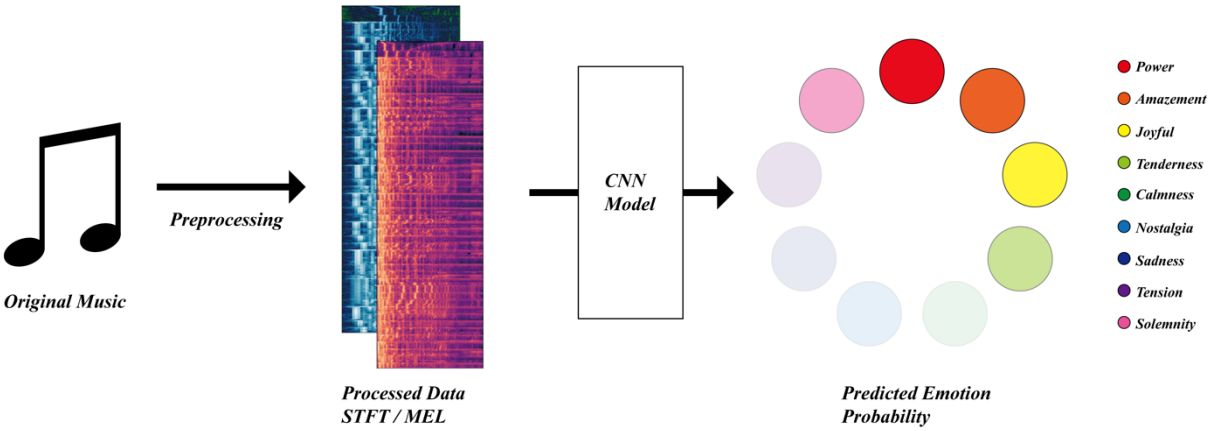


Figure 8. The Overall Process of Musemo

3.2.1 Data Preprocessing

The original dataset contains 400 music files that are one-minute long. The genre of the dataset is Classical, Electronic, Pop, and Rock. We split the music file into intervals of 2, 4, and 8 seconds to compare the performance of models trained by each length of input music. Figure 9 shows how we split the music file into 2, 4, and 8 seconds intervals, and Table 5 presents the number of samples for each duration. Additionally, we transform the music files using STFT and Mel spectrograms to test the performance of models trained by each conversion preprocessing method. STFT divides the audio signal into short intervals, and each piece is transformed using a Fourier transform (FT). The FT decomposes a signal into a sum of periodic functions with various frequencies, but it does not consider time continuity. Therefore, STFT is typically used for audio signal processing to produce decomposed frequencies considering time continuity. Additionally, we transform the music files using Mel spectrograms, which convert the frequency scale to a Mel scale and adjust the frequencies and amplitude ranges so that the frequency differences of sound perceived by the human ear are constant. Table 6 presents the parameters used in the STFT and Mel spectrogram conversion methods.

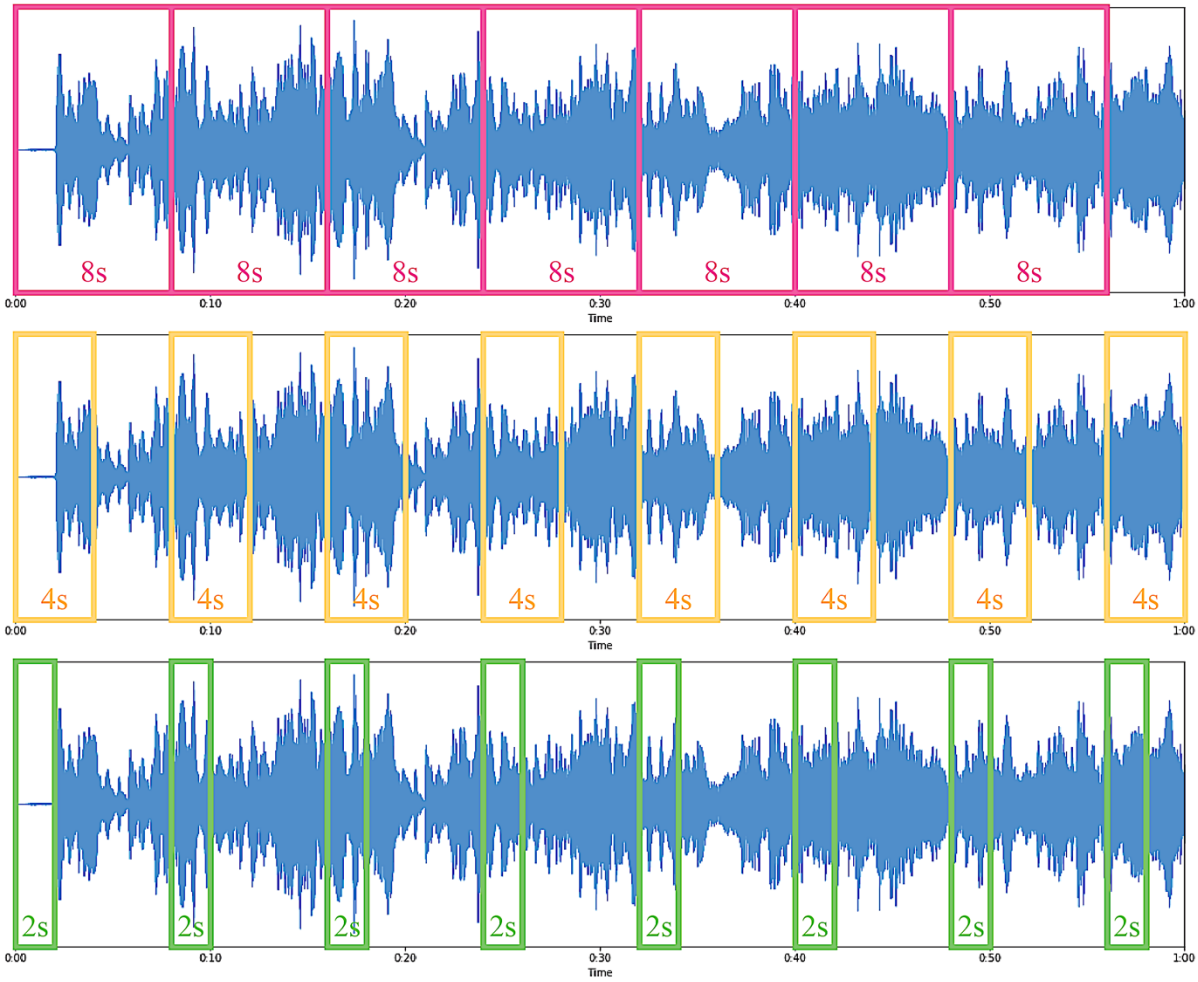


Figure 9. How music samples are split

Table 5. The number of music samples (2seconds, 4 seconds and 8 seconds)

	number of samples
2s	3186
4s	2788
8s	2787

Table 6. Conversion parameters for STFT and Mel Spectrogram

STFT		Mel Spectrogram	
hop_length	512	n_mels	128 bins
n_fft	2048	scale	mel scale
scale	log scale		

3.2.2 Data processing with Musical Understanding

According to Webster and Weir, mode, tempo, and texture are associated with different emotional responses to music (Webster & Weir, 2005). Therefore, a music sample of appropriate length that does not change mode, tempo and texture can be a useful data set to create a MER model. However, specifying an appropriate length of music is difficult, because it depends on the composition of each music and the characteristics of the genre. Therefore, assuming that each song has a common time (4/4) and 120bpm, we chose to compare each performance by making some templates with 2 seconds, 4 seconds, and 8 seconds corresponding to 1 bar, 2 bars, and 4 bars.

3.2.3 CNN

As shown in Figure 8, preprocessed music files are input directly into CNN. CNNs are typically used to analyze data consisting of multi-dimensional arrays (LeCun, Bottou, Bengio, & Haffner, 1998). They are applied in various fields such as image, video, and voice recognition (LeCun, Bengio, & Hinton, 2015). If an input array correlates with the value of nearby data, a CNN may be appropriate to use. Similarly, to use CNN for speech emotion recognition (Lim, Jang, & Lee, 2016), we construct a CNN algorithm using our music samples. Figures 10 and 11 show the architecture of our CNN for STFT and the Mel Spectrogram model. Every model is trained using the same architecture and hyper-parameters but with different input sizes. Table 7 presents the input sizes, and Table 8 presents the hyper-parameters of CNN. Input samples of 2, 4, and 8 seconds have widths of 173, 345, and 690, respectively. In Table 7, for convenience, each input sample width is represented as “8n”, to assign each width to Figures 10 and 11.

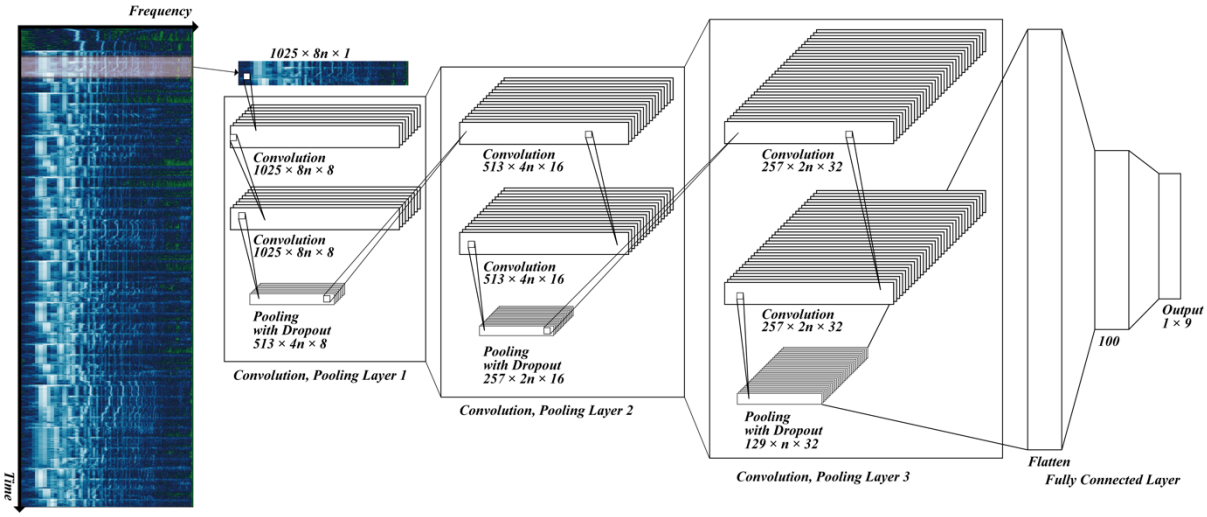


Figure 10. CNN architecture of STFT

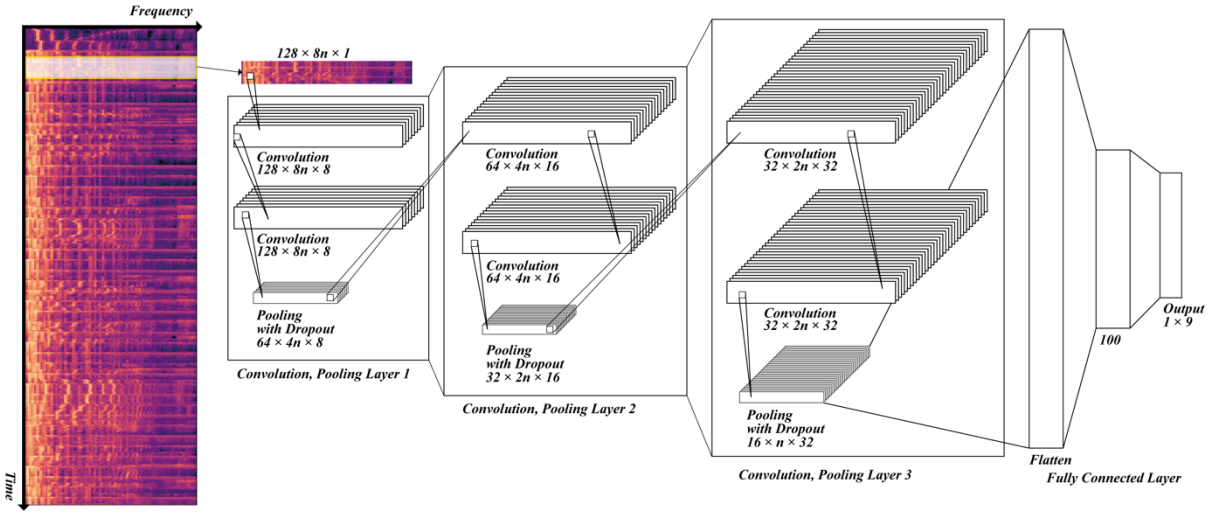


Figure 11. CNN architecture for Mel Spectrogram

Table 7. The input data size

	STFT	Mel Spectrogram
2 seconds	(1025, 8n = 173)	(128, 8n = 173)
4 seconds	(1025, 8n = 345)	(128, 8n = 345)
8 seconds	(1025, 8n = 690)	(128, 8n = 690)

Table 8. Hyperparameters for CNN

Parameter	Value
Convolution filter size	3*3
Activation Function	ReLU
Dropout rate	0.3
Optimizer	Adam Optimizer
Learning Rate	0.0001

3.3 Results and Discussion

3.3.1 Accuracy

We measure accuracy with root mean square error (RMSE). Because we target the probability that each emotion exists in each piece of music, we assess how close the model learns to that probability. Using the RMSE, we calculate the error between the model-predicted probability that each emotion presents, and the probability obtained from the actual survey. Three lengths of music samples (2, 4, and 8 seconds) are provided as inputs and are converted by STFT or Mel spectrogram for a total of six models. The results for each model include 5-fold cross-validation. Table 9 and Table 10 present the RMSE values of the models for the STFT and Mel spectrogram, respectively. RMSEs are expressed as percentages to facilitate comparisons. All six models perform similarly, with error rates of approximately 15% to 16%. While the performance difference is small, the best model is the 4 seconds STFT conversion model (14.91% of RMSE).

Table 9. RMSE for STFT model

STFT	2s	4s	8s
Amazement	10.77%	10.67%	10.99%
Solemnity	14.05%	12.56%	13.41%
Tenderness	14.71%	15.06%	15.10%
Nostalgia	15.20%	14.83%	15.97%
Calmness	17.61%	17.79%	18.17%
Power	14.87%	14.52%	15.28%
Joyful_activation	19.60%	18.99%	19.89%
Tension	15.24%	14.62%	15.69%
Sadness	15.29%	15.17%	15.68%
Total	15.26%	14.91%	15.58%

Table 10. RMSE for Mel Spectrogram model

Mel Spectrogram	2s	4s	8s
Amazement	10.70%	10.59%	11.75%
Solemnity	13.75%	13.36%	13.72%
Tenderness	14.15%	13.85%	15.80%
Nostalgia	14.64%	14.14%	15.88%
Calmness	17.97%	17.28%	19.19%
Power	17.19%	16.32%	15.31%
Joyful_activation	19.44%	19.05%	19.71%
Tension	15.74%	15.78%	15.54%
Sadness	14.18%	14.14%	15.07%
Total	15.31%	14.95%	15.78%

Although the performance differences of the six models above are not significant, we can demonstrate how Musemo better predicts each emotion. Table 11 presents the two emotions predicted with the lowest errors, and the two predicted with the highest. Amazement and solemnity have the lowest errors (11% and 13%, respectively), while calmness and joyful activation has the highest errors (18% and 19%, respectively). Notably, amazement and solemnity have the lowest errors, because according to Jonna and Tuomas (see Figure 12), amazement (wonder) and solemnity (transcendence) has the least consistent ratings in the discrete emotion, dimensional emotion, and GEMS models (Vuoskoski & Eerola, 2011). Musemo exhibits the best performance (with the lowest error) for amazement and solemnity, which are relatively difficult to define on an arbitrary scale.

Table 11. Average RMSE for each emotion label

	STFT	Mel Spectrogram
Amazement	10.81%	11.02%
Solemnity	13.34%	13.61%
Tenderness	14.96%	14.60%
Nostalgia	15.33%	14.89%
Calmness	17.85%	18.14%
Power	14.89%	16.27%
Joyful_activation	19.49%	19.40%
Tension	15.18%	15.69%
Sadness	15.38%	14.46%

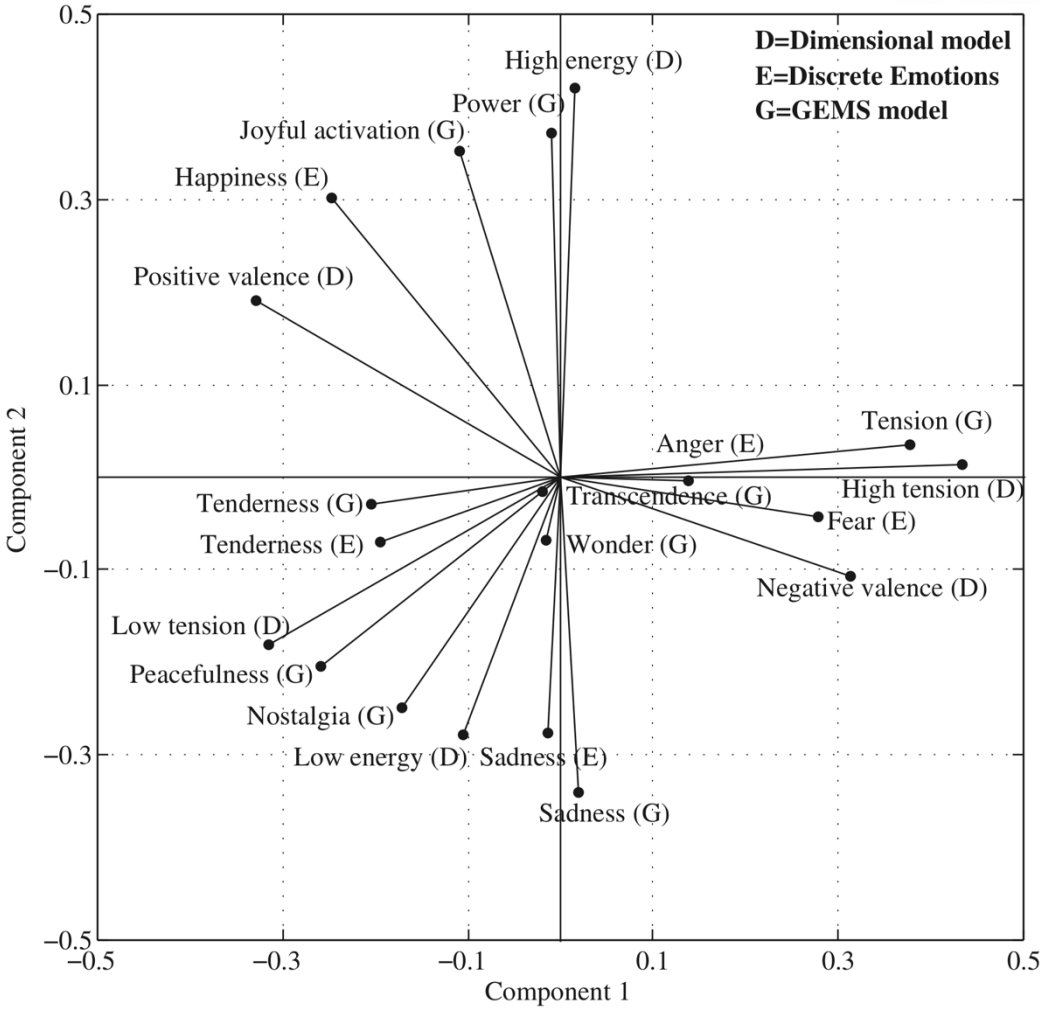


Figure 12. Two-dimensional principal component analysis of three emotion models, 2011

Table 12 shows the top three performances of the MIREX Audio Music Mood Classification contest from 2010 to 2017. Despite various audio features and their combinations have been used, classification accuracy is limited to 70%. Our model is designed to learn the probability of having each emotion, not to classify the emotion which exists or not in the music. Since 15%-16% RMSE is not a high error, if we design our Musemo to learn the emotion classification like other studies, we expect that there will be a comparable result in classification accuracy (Table 12).

Table 12. Top three performances of the MIREX AMMC contests

Contest	Top Three Accuracy		
AMMC 2010	64.17%	63.83%	63.17%
AMMC 2011	69.50%	67.17%	66.67%
AMMC 2012	67.83%	67.67%	67.17%
AMMC 2013	67.83%	67.83%	67.67%
AMMC 2014	66.33%	66.17%	65.50%
AMMC 2015	66.17%	62.50%	59.17%
AMMC 2016	63.33%	62.50%	60.33%
AMMC 2017	69.83%	68.67%	67.83%

3.3.2 Comparison with existing studies

Through the correlation analysis of emotions that Musemo classifies through machine learning, we review whether the correlation analysis among emotion labels suggested by the different study is reasonable. So, we introduced the principal component analysis (PCA). PCA converts samples from high to low dimensional space. PCA linearly transforms data into a new coordinate system so that when the data is mapped onto one axis, the axis with the largest variance is placed as the first principal component and the second largest variance as the second principal component. Figure 13 shows the two-dimensional PCA of Musemo (4s, STFT). As a result of the analysis, the first component (Eigenvalue 0.38) accounts for 55% of the variance, while the second component (Eigenvalue 0.32) accounts for 45% of the variance. Joyful_activation and power appeared to have a significant positive correlation, and they are negatively correlated with tenderness and nostalgia. Amazement is located in the opposite position from sadness. Calmness is located in the opposite position from tension. According to Jonna K. Vuoskoski and Tuomas Eerola, the three emotion models (discrete, dimensional, and GEMS models) described in this paper have very similar scales for emotional traits; GEMS appears to have combined aspects of discrete and dimensional models (Vuoskoski & Eerola, 2011). When comparing Jonna K. Vuoskoski and Tuomas Eerola's PCA results (Figure 12) and Musemo's PCA result (Figure 13), interestingly, the correlation between Joyful_activation and Power appears to be high in both studies. Moreover, the position of Tension is far from Power-Joyful_activation and Tenderness-Nostalgia-Calmness in both studies. As a result, our correlation among the emotion labels is similar to the correlation of emotion labels in Jonna and Tuomas' study.

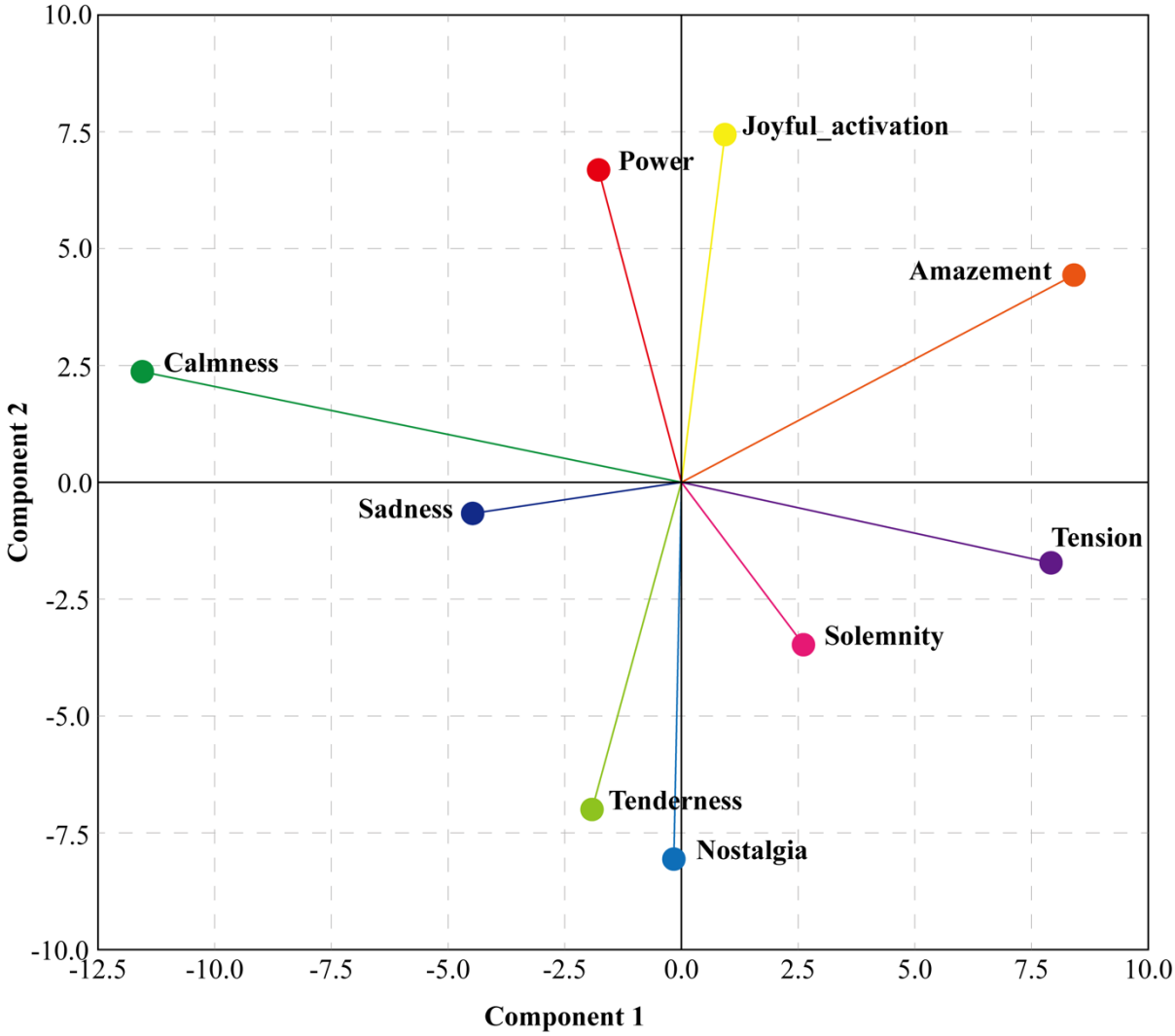


Figure 13. Two-dimensional principal component analysis of Musemo (4s, STFT)

4 Conclusion and Future Work

In this paper, we introduce the first step of research to develop a device that can provide people with psychological comfort and empathy using music. We create a machine learning model that recognizes emotions only with music files of various genres. We focus on comparing the performance of each model created by applying various lengths of music files and two conversion methods. We design the CNN structure and train with input data made of 2, 4, 8 seconds of music samples converted with STFT and Mel Spectrogram. This model targets the probability that each of the nine emotions would exist. As a result, the RMSE of these six models is about 15%-16%, and the best model is a 4 seconds STFT model (14.91%). The principal component analysis show correlations between nine emotion labels, similar to Vuoskoski and Eerola's study (Vuoskoski & Eerola, 2011). In the future, in addition to increasing the accuracy of machine learning models, this research is going to be extended to the study of how Musemo can be applied to people, giving psychological empathy and comfort.

At the end of this paper, we have attached an exhibition report about the Musemo application as an appendix. The title is Musemo Ex.1, and the theme is Express ∞ Empathize. We can visually see how people and Musemo express and empathize with daily or musical emotion based on music, and this allows us to plan the next step of the study by reflecting the empirical opinions of people. Through this, rather than conducting a private research meeting format among researchers, we present a new format of research called "laboratory exhibition". Various interpretations of ideas or data collection and research can be made freely in this exhibition format, and these enable researchers to develop research based on the collected resources from the exhibition participants. Musemo research team experimented model application through this laboratory exhibition called Musemo Ex.1. Please refer to the appendix for details.

5 Reference

Aljanaki, A., Wiering, F., & Veltkamp, R. (2014). Collecting annotations for induced musical emotion via online game with a purpose Emotify. In (Vol. 2014): UU BETA ICS Departement Informatica.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management, 52*(1), 115-128.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol., 58*, 373-403.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*: Oxford University Press, USA.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion, 98*(45-60), 16.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *American journal of Psychology, 48*(2), 246-268.

Izard, C. E. (1971). The face of emotion.

Juslin, P. N. (2013). What does music express? Basic emotions and beyond. *Frontiers in psychology, 4*, 596.

Juslin, P. N. (2019). *Musical Emotions Explained: Unlocking the Secrets of Musical Affect*: Oxford University Press, USA.

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of new music research, 33*(3), 217-238.

Juslin, P. N., Liljeström, S., Laukka, P., Västfjäll, D., & Lundqvist, L.-O. (2011). Emotional reactions to music in a nationally representative sample of Swedish adults: Prevalence and causal influences. *Musicae scientiae, 15*(2), 174-207.

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: listener, music, and situation. *Emotion, 8*(5), 668.

Juslin, P. N., & Timmers, R. (2010). Expression and communication of emotion in music performance. *Handbook of music and emotion: Theory, research, applications*, 453-489.

Laukka, P. (2007). Uses of music and psychological well-being among the elderly. *Journal of happiness studies, 8*(2), 215.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lim, W., Jang, D., & Lee, T. (2016). *Speech emotion recognition using convolutional and recurrent*

neural networks. Paper presented at the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33): Elsevier.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39*(6), 1161.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval, 8*(2-3), 127-261.

Vuoskoski, J. K., & Eerola, T. (2011). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae scientiae, 15*(2), 159-173.

Webster, G. D., & Weir, C. G. (2005). Emotional responses to music: Interactive effects of mode, texture, and tempo. *Motivation and Emotion, 29*(1), 19-39.

Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(3), 40.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion, 11*(4), 705.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion, 8*(4), 494.

6 Appendix

6.1 Musemo Ex.1

Express ∞ Empathize

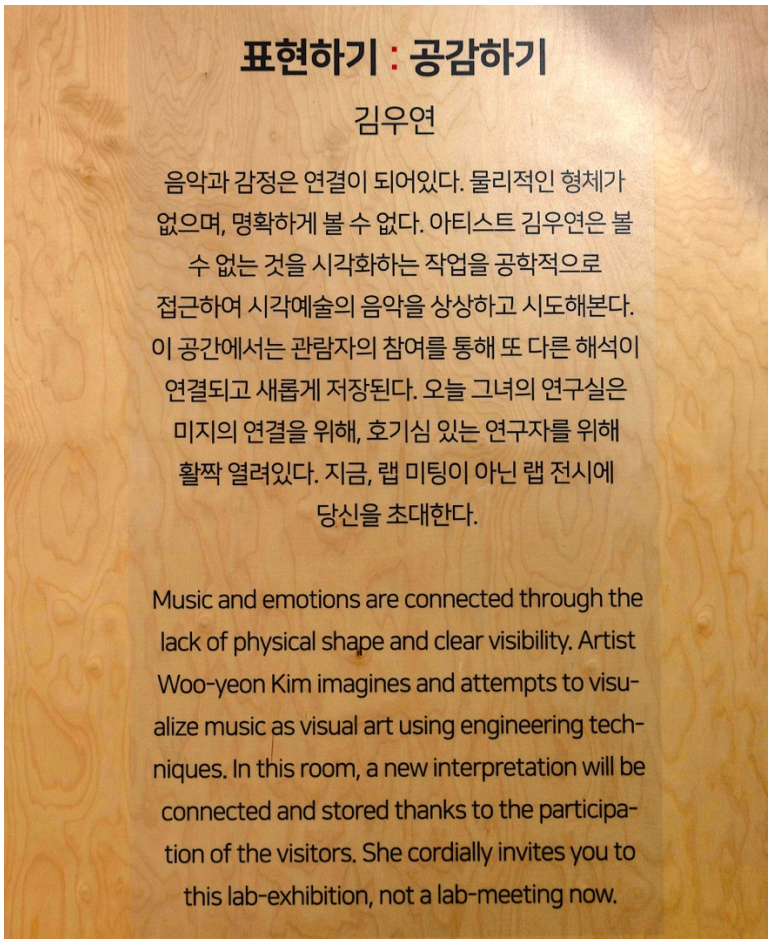
Science Walden has a convergence methodology of science and arts and uses various perspectives to solve social problems. As one of these efforts, the emotional recognition model Musemo was linked to an art exhibition to collect the resources needed for research naturally. Like other art exhibitions, the exhibition has been organized so that anyone can participate without limiting the conditions and settings of sample groups, such as the age, sex, and occupation of participants.

There is an invisible link between music and emotion. The music itself may have musical emotions, and it may affect changes in one's emotions and psychological state. However, music and emotion have no physical form and cannot be seen. This fact leaves many barriers to engineering research. However, data visualization enables the expression of invisible things and enables another interpretation that has not been thought of before. The exhibition offers a variety of data visualization methods to allow participants to understand sensitively, such as seeing, listening to, and feeling emotions and music. Musemo is a machine learning model that recognizes nine musical emotions. The exhibition proposes a laboratory exhibition, a new type of audience participation research, using this model. Entering this space will bring the participants a handbook with directions, and they can experience the process of research indirectly as if they were attending a laboratory meeting.

This work consists of five steps, each representing the process of Express or Empathize. Express and Empathy could be seen as one of the structures of communication and that communication could help solve problems in modern society. This time, Musemo Ex.1 separated the steps to identify a small communication link with four materials: people, music, AI, and emotion, but the Musemo Ex would be completed if the future expanded to create a space where expression and empathy can be freely made among these objects.

Starts

People begin to participate in the exhibition as they get a handbook at the same time.



Picture 1. Entrance

Picture 2. Handbook, Express ∞ Empathize

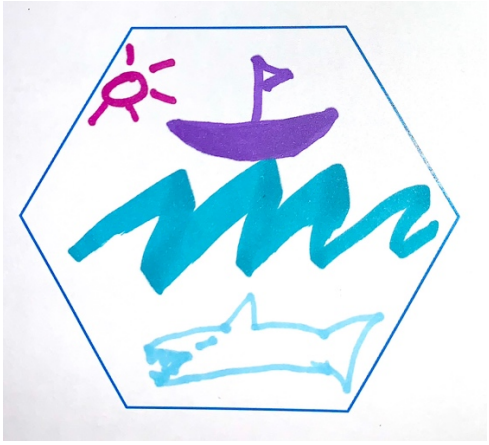
Step 1

Express your emotion using colored markers (Express)

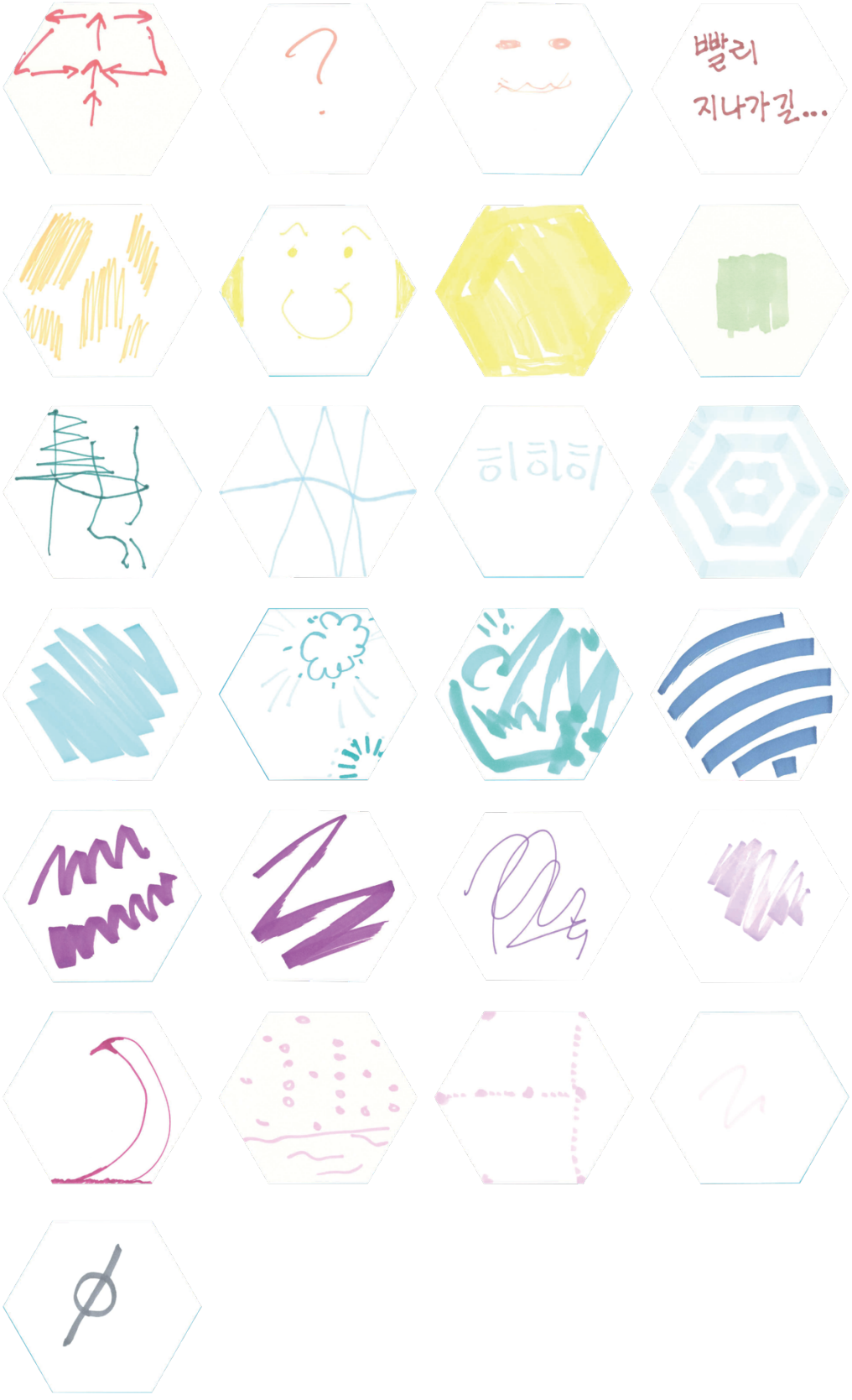
Participants freely expressed their current state of emotion visually using paper and color markers.

Picture 3. Colored Markers

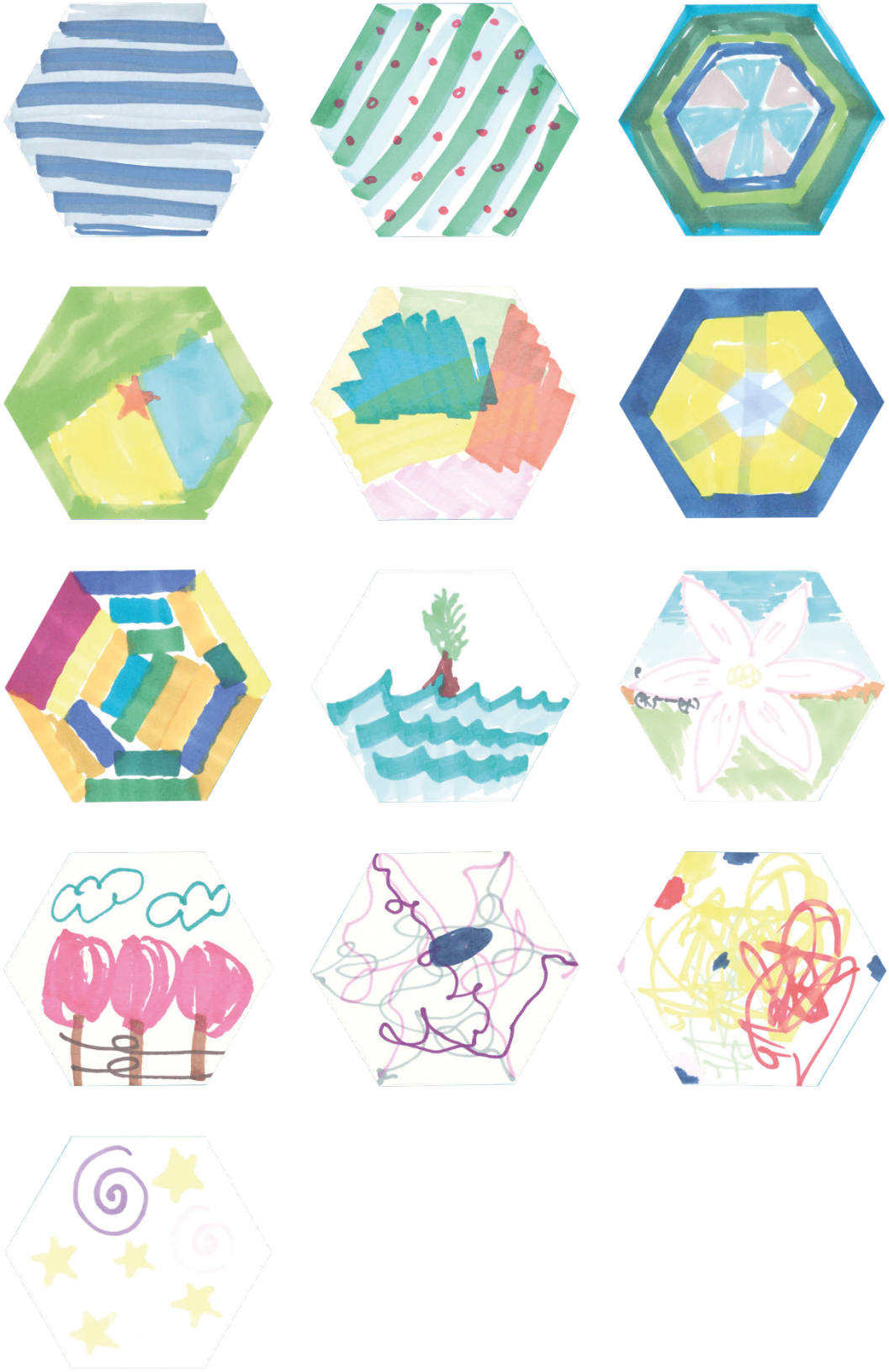
Picture 4. Participants visualizing their emotion



Picture 5. An example of emotion visualization



Picture 6. Collected daily emotion of participants with a single color



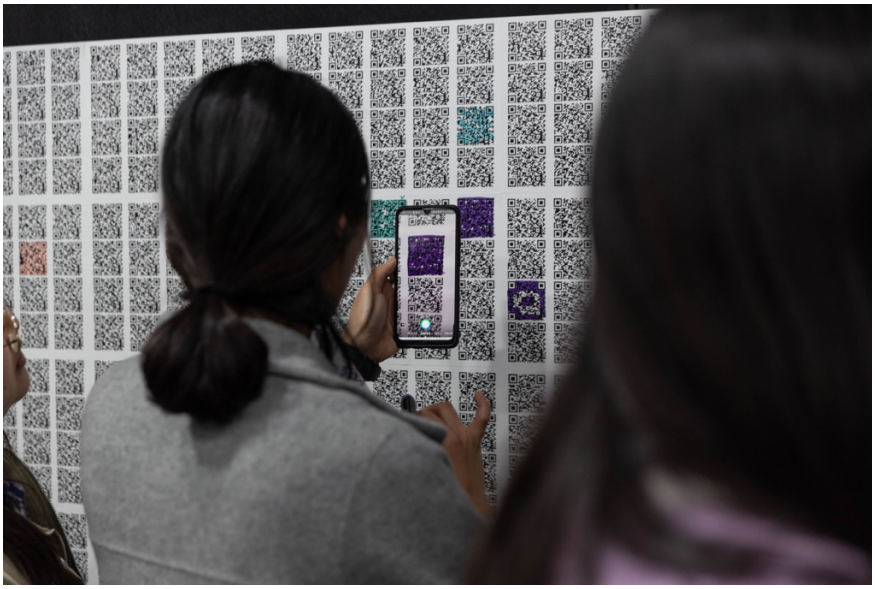
Picture 7. Collected daily emotion of participants with multiple colors

Step 2

Listen to the music and feel the musical emotions (Empathize)

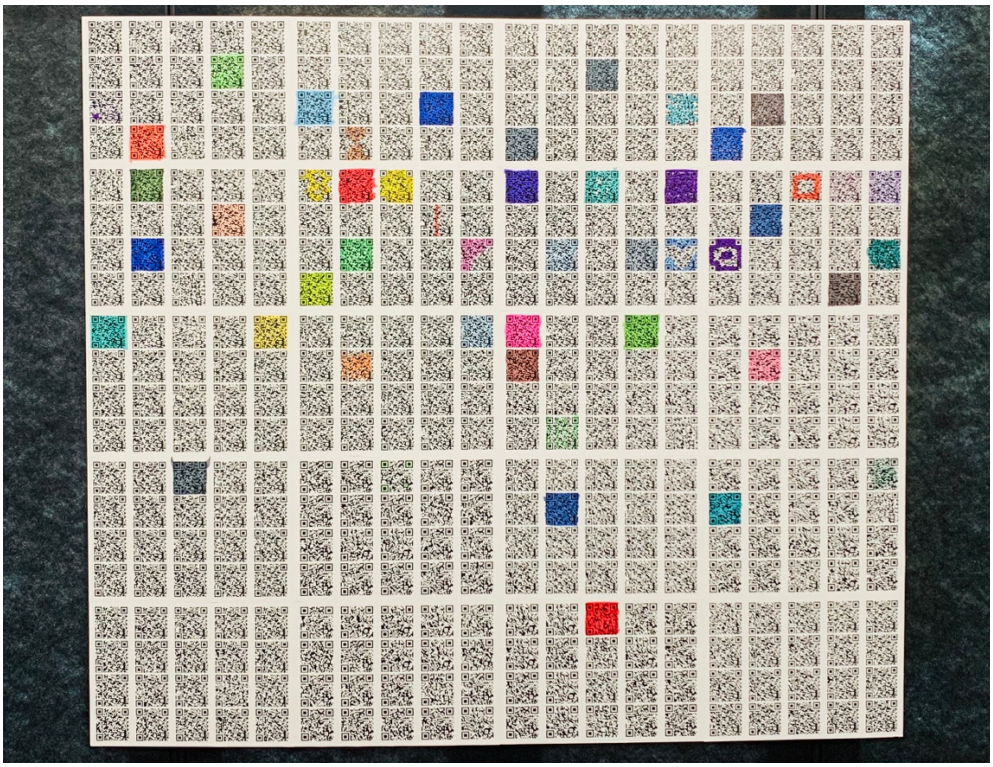
Participants conduct a survey in which they can play and listen to music by scanning a QR code containing a music sample and check the musical emotion felt in the music sample. There are 400 music samples used in the QR code, and the genre consists of classical, electronic, pop, and rock music. Participants choose one QR code and paint the code with their emotion color, and they conducted the survey using the painted code.

Picture 8. Choosing one QR code and painting the code with markers



Picture 9. Scanning the QR code

Picture 10. Playing the music sample in the QR code



Picture 12. Painted QR codes

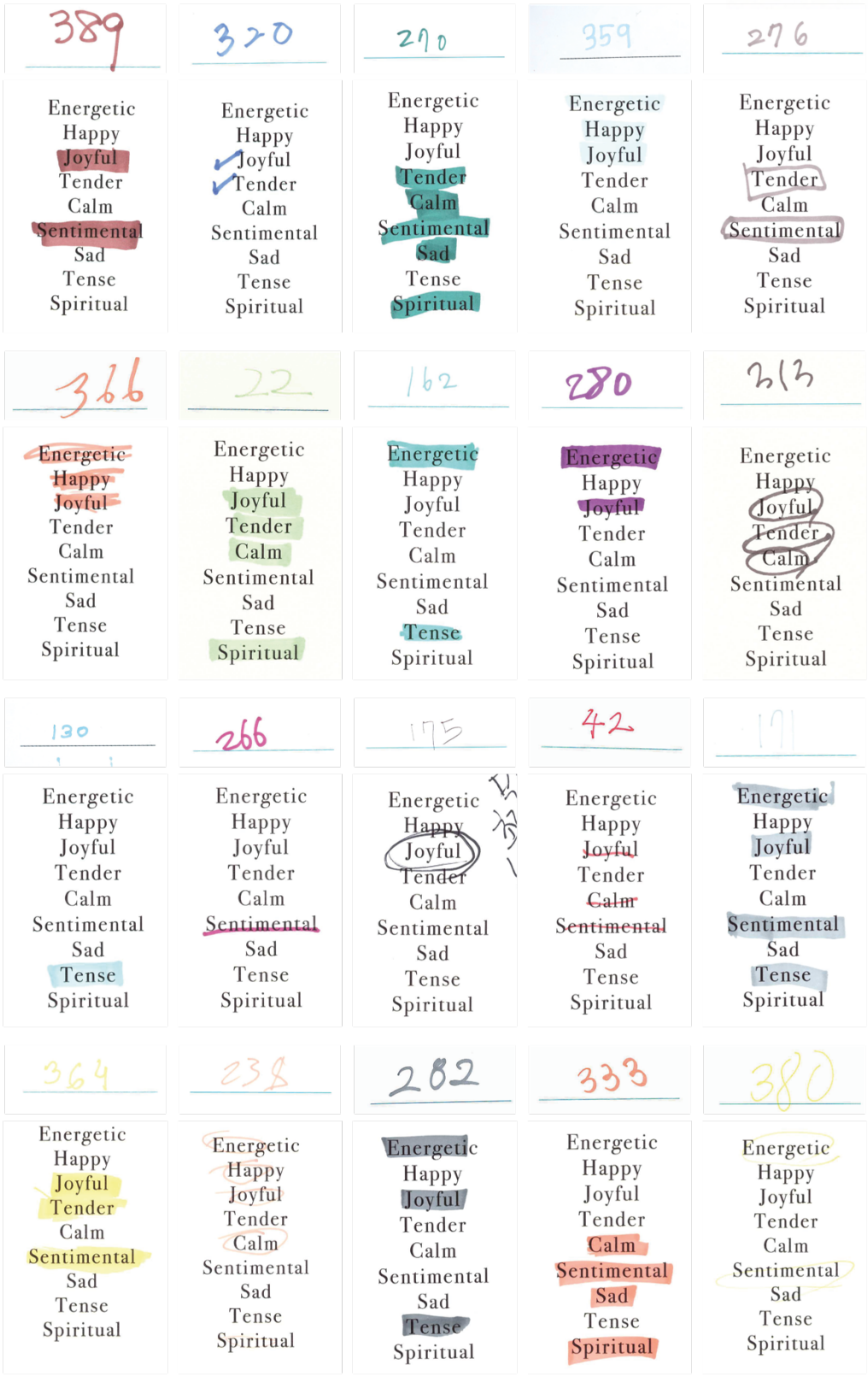
Step 3

Compare the musical emotion that other people have felt with musical emotion that you found in the music (Empathize)

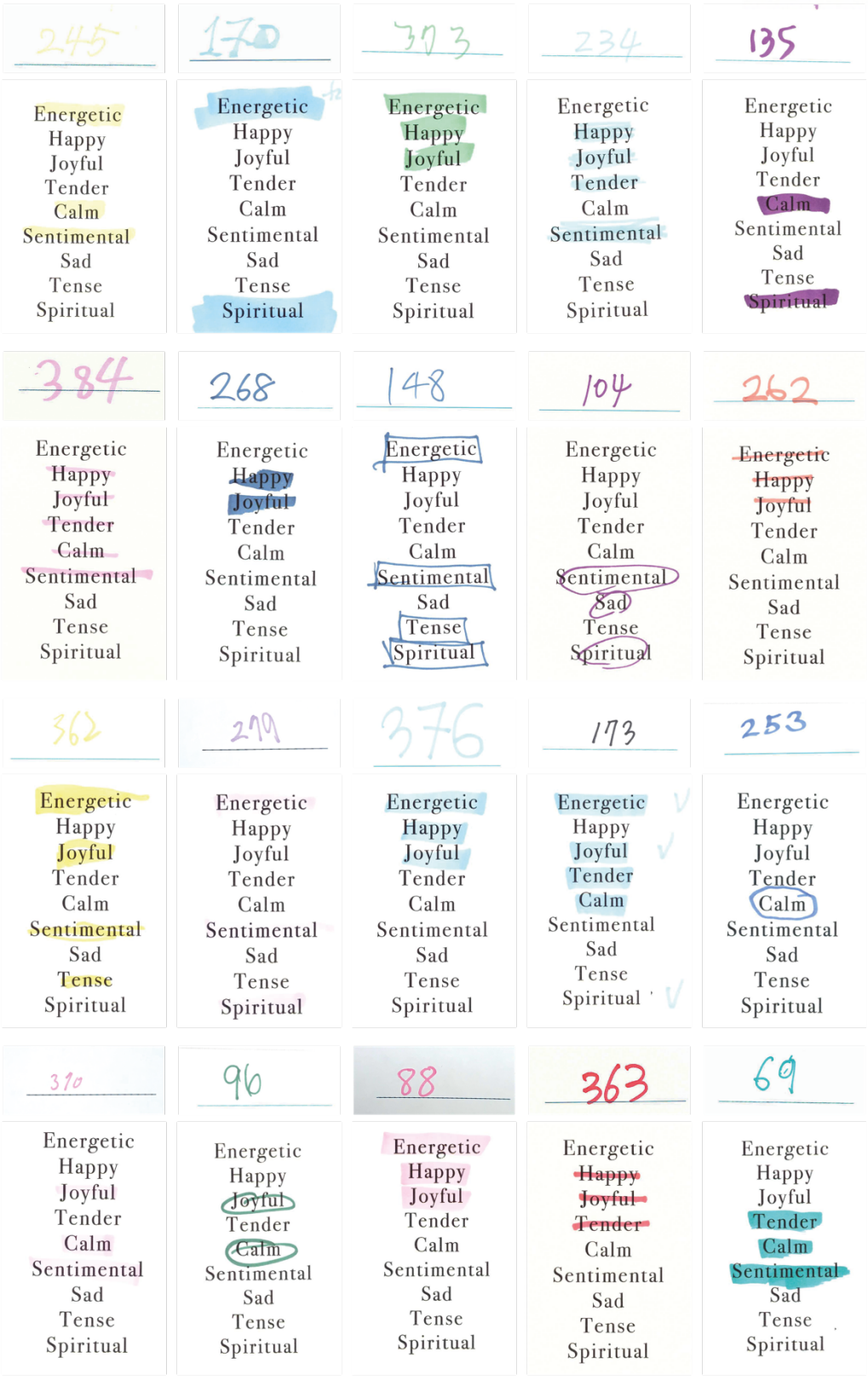
Participants conduct the survey, recalling their musical emotions after listening to the music sample. Among Energetic Happy Joyful Tender Calm Sentence Spiritual, they can select all the emotions felt, and compare their responses to the existing survey's statistical data. Using response results, the artist collects musical emotion label data for the music files.

Picture 11. Marking the musical emotions that participants felt in the music

Picture 13. Emotion book that contains statistical emotion data of each music sample



Picture 14-1. Sample number and musical emotion label from participants

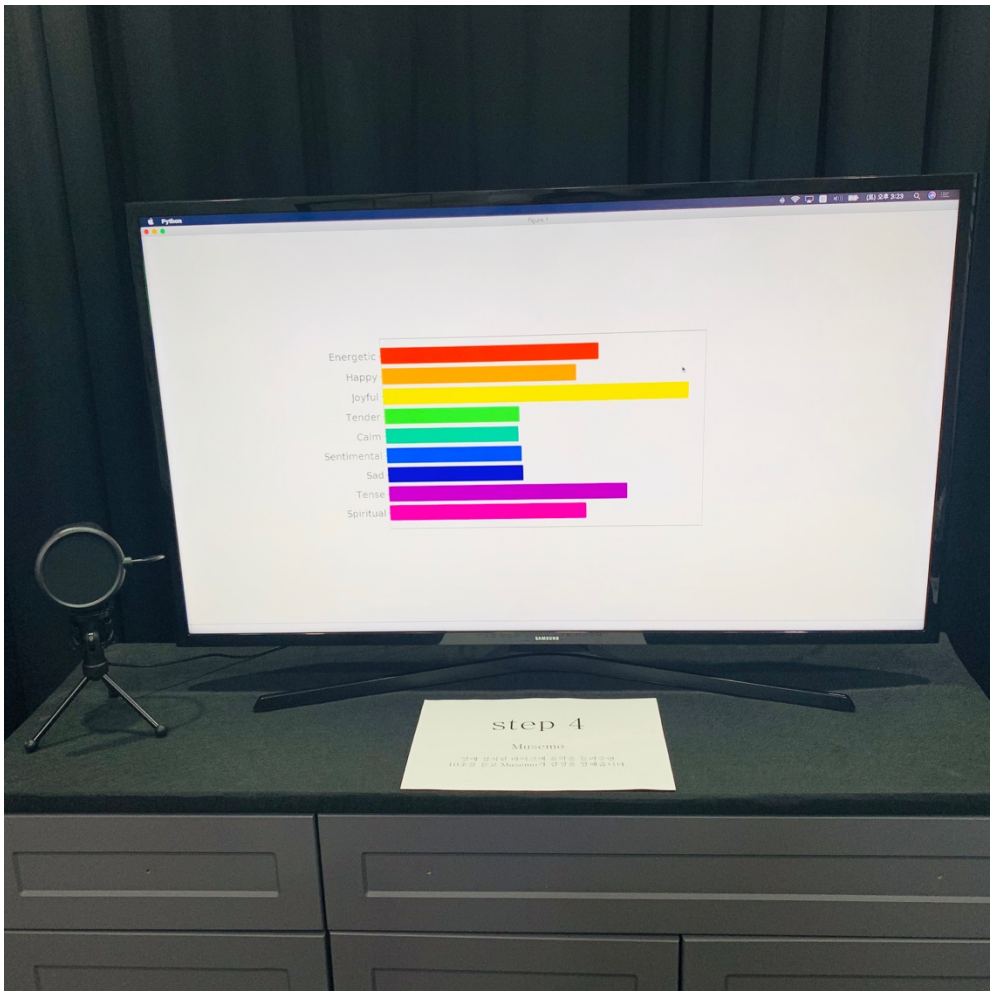


Picture 14-2. Sample number and musical emotion label from participants

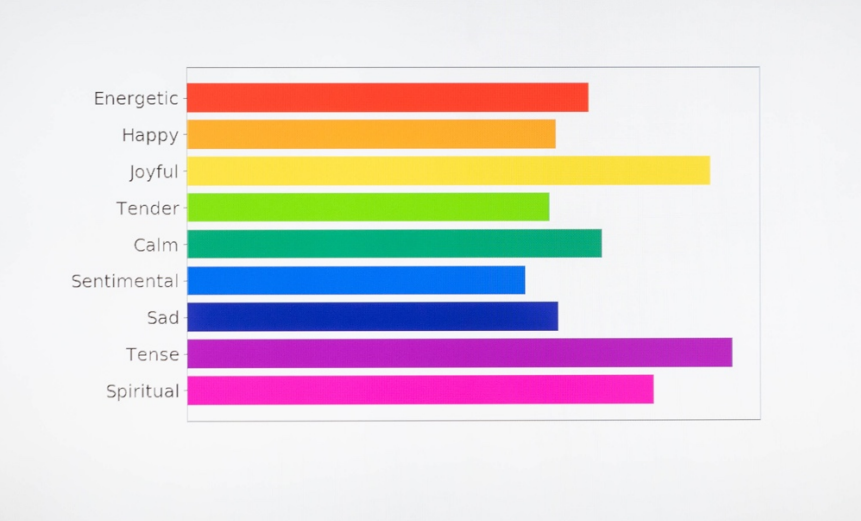
Step 4

Play music that you recommend to Musemo. Musemo will express musical emotions (Express and Empathize)

Participants recommend songs that can empathize with their emotions. They play the songs to the Musemo using their smartphones. Musemo prints the musical emotions it feels after it listens to a song. Participants write information about song recommendation in their handbooks and the most dominant emotion which Musemo feels.



Picture 15. Installation of the microphone and the Musemo



Picture 16. Nine emotions used to learn by Musemo

Picture 17. A student playing a song to Musemo

Picture 18. Students playing a song to Musemo

Table A. Recommended music by participants and Musemo's response

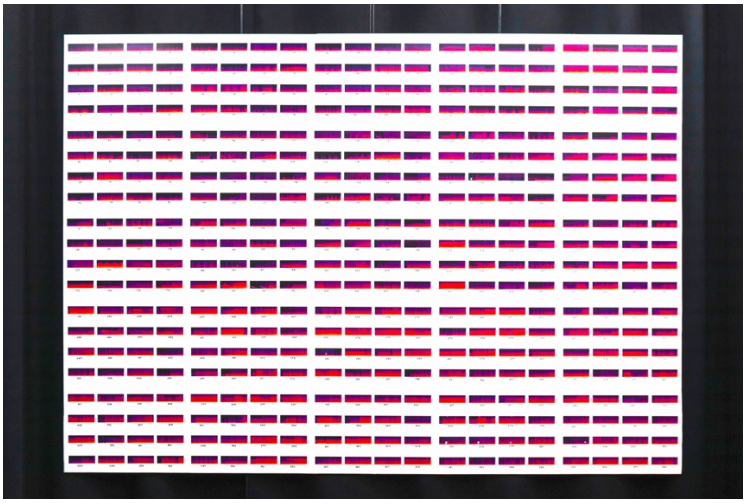
Artists / Composer	Title	Musical Emotion
Aaron Carter	I want candy	Energetic
Standing Egg	오래된 노래	Energetic
Cosmo's Midnight	Lovelight	Happy
형돈이와 대준이	MUMBLE	Joyful
Taylor Swift	22	Joyful
Andy Williams	Can't take my eyes off you	Joyful
볼빨간 사춘기	여행	Joyful
Crush	잊어버리지마	Joyful
a-ha	Take on Me	Joyful
Queen	Bohemian Rhapsody	Joyful
CHEEZE	어떻게 생각해	Joyful
BTS	Mikrokosmos	Joyful
Loreen	Euphoria	Joyful
자전거 탄 풍경	너에게 난 나에게 넌	Joyful
The Breeze	너무나 눈부신	Joyful
Reamonn	Tonight	Joyful
Giriboy	인체의 신비	Joyful
Madeon	Be Fine	Joyful
Franz Schubert	Fantasie in f minor d.940	Calm
Enrique Iglesias	Sombody's me	Calm
The The	내게다시	Calm
Before You Exit	Clouds	Calm
Beatles	Let it be	Calm
Chris Stapleton	Tennessee Whiskey	Calm
Woodkid	Brooklyn	Calm
Vivaldi	Cello sonata in e minor	Calm
김동률	Contact	Calm
IU	마음	Calm
IU	Love poem	Calm
이적	걱정말아요 그대	Sentimental
Billie Eilish	Bad guy	Sentimental
Idealism	Another perspective	Spiritual
Yiruma	Dream	Tense

Step 5

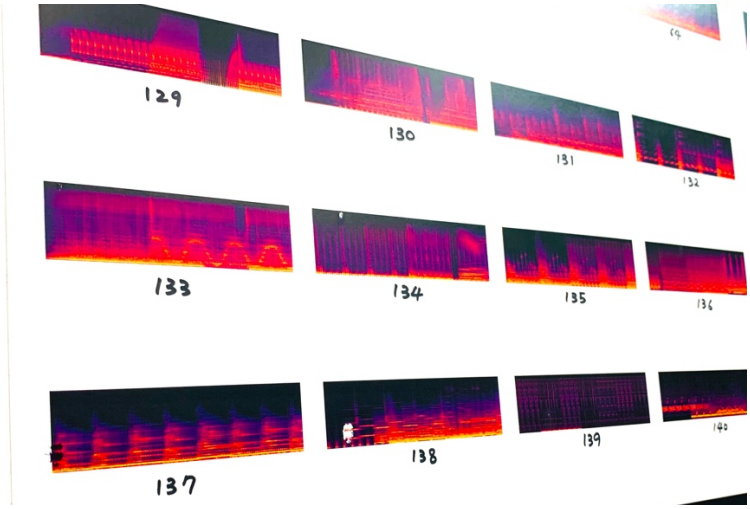
Play with every visible and audible resource or data

The participants who complete all the tests get access to 400 entire music samples. They utilize all the resources in the exhibition space. While listening to music, they find musical characteristics that appear on images and analyze patterns while comparing music samples with the images used for model making. In order to find out the emotions that the Musemo best detects, they play various genres of music and share their own opinions with other participants on musical emotions. Feedback and comments related to this study are also freely available to the artist.

Picture 19. QR code to access the entire music samples



Picture 20. 400 images used to create the model



Picture 21. Examples of images, close-up shot

Picture 22. Participants comparing music file with the converted images

Picture 23. Artist talks

Complete

The participant returns the completed handbook to the artist and exits.

Picture 23. Returned handbook

Conclusion

Musemo Ex.1 is a process-driven research project and art exhibition that combines research subjects, processes and analysis with the form of exhibitions and works with various audiences. Also, because it is a topic of ongoing research, researchers gain the various perspectives and insights needed for the next step of the research, and the synergy of the researchers in that various people participated in the research activities. Researchers are able to analyze and review their research more actively and share concerns with people. Visualizing invisible concepts is the most crucial part of facilitating the involvement of others. It is able to break down language barriers and increase efficiency in understanding and analyzing various data consistently. In order for a lower version of Musemo to be upgraded to a higher version, data from a variety of people gathered based on the understanding of this research is needed. Musemo Ex.1 set up the concept of emotion and music, model making process, and Muemo verification in one space. Finally, researchers propose this type of “laboratory exhibition” as an attempt in the research of convergence of science and arts.