

# AI Generated Music Using Speech Emotion Recognition

Roberto Murru<sup>2\*</sup>, Jonas Krug<sup>1\*</sup>, Tom Schmid<sup>1</sup>, Garri Steba<sup>1</sup>, Giorgio Giacinto<sup>2</sup>,  
Alexander von Hoffmann<sup>1</sup>

<sup>1\*</sup>EFI, Technische Hochschule Nürnberg Georg Simon Ohm, Wassertorstraße 10 ,  
Nürnberg, 90489, Bavaria, Deutschland.

<sup>2</sup>DIEE, Università degli Studi di Cagliari, Via Marengo, 2, Cagliari, 09123, Sardinia, Italy.

\*Corresponding author(s). E-mail(s): [roberto.murru@yahoo.it](mailto:roberto.murru@yahoo.it); [jonaskrug@hotmail.de](mailto:jonaskrug@hotmail.de);  
Contributing authors: [tom.schmid@th-nuernberg.de](mailto:tom.schmid@th-nuernberg.de); [garri.steba@th-nuernberg.de](mailto:garri.steba@th-nuernberg.de);  
[giacinto@unica.it](mailto:giacinto@unica.it); [alexander.vonhoffmann@th-nuernberg.de](mailto:alexander.vonhoffmann@th-nuernberg.de);

## Abstract

This study aims to compare two different implementations of speech emotion recognition models. The emphasis is directed towards evaluating their efficacy in capturing and characterizing dialogues portrayed by actors within a film scene to create suitable musical intervals. The goal of the overarching research intends to derive indications to enhance the compositional process of film scores by recognizing the emotion in a particular scene. Based on established deep learning models, the study delves into the exploration of two distinct emotion classification metrics: The Six Emotion Prediction and the Valence/Arousal/Dominance Prediction. To facilitate a comparative analysis, a preliminary study and a following survey is deployed. The preliminary study confirms a significant difference in the generated MIDI data. For this reason, a survey is essential to detect the better fitting algorithm. Participants are tasked to rate the affective suitability of eight generated interval sequences to the corresponding film scenes. The Suitability is verified quantitatively using a bidirectional rating system. Both model assessments are conducted within a uniform sound design, thus ensuring unbiased conditions for evaluation. Upon a thorough examination of our extensive analysis, a preference for method A becomes increasingly evident.

**Keywords:** AI, Deep Learning, Speech Emotion Recognition, Emotion Classification, Affect, Neural Networks, Valence, Arousal

## 1 Introduction

Emotions are an abiding presence in human lives as they get experienced for every entire subjective existence. Still to this day, no unique definition for what an emotion is has been given, but according to the American Psychological Association (APA) an emotion can be described as "*A complex reaction pattern, involving experiential, behavioral, and physiological elements, by which*

*an individual attempts to deal with a personally significant matter or event*"[1].

During the last centuries various studies on emotions have been conducted, fundamental to this study are the works from Mehrabian and Russell [2] and from Ekman [3].

This paper aims to investigate on how AI can be used as a tool that takes emotions from speeches and convert them into music. For that, it will be described the work that has been done

performing Speech Emotion Recognition (SER) on four selected film scenes. The results from SER are used to generate MIDI notes that are in turn sent to an external instrument, in this case a synthesizer.

Two different methods for SER will be described, referred to as A and B. One method, apart from the recognition part, also involves a neural network training part and is based on the emotion classification by Ekman [3]. The other method uses a pre-trained neural network and is focused on the use of the emotion classification by Mehrabian and Russell [2].

The last sections describe the tests carried out to compare the results of the two methods, both in terms of the outcome of the two SER approaches, and in terms of the generated MIDI notes.

## 2 Related Work and novelty

As mentioned in the introduction, two papers inspired the two proposed mechanisms: the paper by Mehrabian and Russell [2], and the paper by Ekman [3]. These papers propose two different emotion classifications, and our goal is to compare the effectiveness of the two methods in providing inputs to the AI module to generate film music.

The first paper introduces a polar representation of emotions on an XYZ-plane, using the following axes: Pleasure (also called Valence) which represent how pleasant or unpleasant one feels about something; Arousal (which represent the energy of the emotion) and Dominance (which represent the dominant or submissive nature of the emotion). This emotional model is called the Pleasure-Arousal-Dominance or Valence-Arousal-Dominance (VAD). This was later developed into a 2-D circumplex model by Russell [4], using Valence and Arousal as the two dimensions.

The second paper on "Facial Expressions and Emotion" [3] is focused on recognising human emotions from facial expressions. In particular, it investigates how emotions are universal across cultures and expressed through distinct facial expression, then human emotions are classified into six basic categories: Anger, Happiness, Surprise, Disgust, Sadness, Fear.

The use of emotion recognition for generating music drones for films, has been investigated in [5]. The International Affective Digitized Sounds (IADS) is used as the main source of emotional

data, and machine learning is employed to generate the music drones.

Finally, matching emotions to notes has been thoroughly investigated by Kaygusuz and Zuluaga in [6], where the connection between western music emotions and certain musical intervals is reported.

## 3 Classification of Emotions with different Metrics

In this study, two different methods for emotion classification are used. Method A is based on the emotion classification in [2], and thus it relies on the VAD model. Method B uses the six emotions model. It's important to note that in both methods semantics has no influence on the emotion recognition part, since they're not part of the analysis process. These models are based on speech recognition models that have been fine-tuned for the purpose of emotion recognition. In fact, the two methods are based on the extraction of features from the human voice such as Mel-Frequency Cepstrum Coefficients (MFCC) or Gammatone Cepstral Coefficients (GTCC), which are then used to train the neural network.

### 3.1 VAD Emotion Prediction

The VAD Classification aims to provide a more diverse representation of emotions than the representation provided by the Six Emotion Prediction (see next subsection). Based on the work of Mehrabian and Russell [2], VAD Prediction classifies emotions into three parameters, namely, Valence, Arousal and Dominance. In this context, Valence gives an estimation of the emotional state of the speech, especially if the feeling is perceived as positive or negative, whereas Arousal is giving an estimation of the intensity of the expression [7]. Although Dominance is a similar Dimension as Arousal and not particularly needed for Emotion Classification, it is helpful to use Dominance to distinguish between certain states of emotions, which are more subtle and harder to classify. Dominance can be used to differentiate between depressive states and tragic states. While both emotions would be located in a similar part of a Valence and Arousal Graph, they are vastly different in expression. A two-dimensional approach of classifying emotions is depicted in figure one.

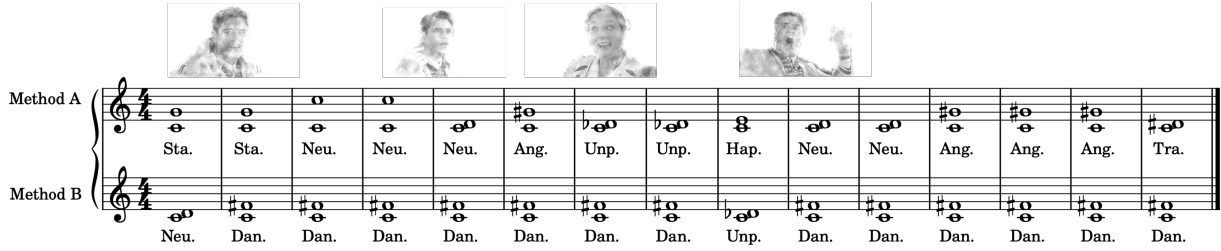


Fig. 1 Scene 1: The Truman Show - METHOD A = WAV2VEC; METHOD B = MATLAB

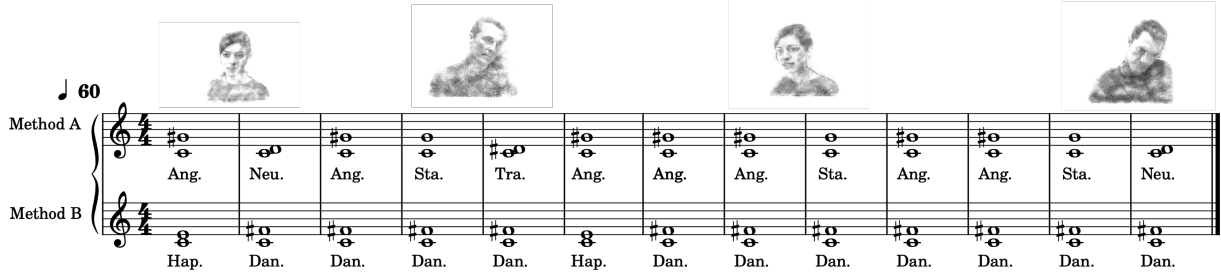


Fig. 2 Scene 2: Interstellar - METHOD A = WAV2VEC; METHOD B = MATLAB

As already established, valence being an indicator of the emotional tone (positive or negative) and arousal being the factor of the Intensity of the emotion, the VA-Diagram can be separated into 4 distinct quadrants. The 0.5 Border of Valence, differentiating into 'positive' or 'negative' perceived emotions, and the 0.5 Border of Arousal, differentiating into high or low perceived arousal. Emotions located around the 0.5 mark along both axes signify states that are perceptually 'Neutral' in nature.

### 3.2 Six Emotion Prediction

The six emotion prediction, in contrast to the VAD prediction, it's more simple and uses the basic emotions defined by Ekman, which are: Fear, Neutral, Sadness, Happiness, Danger and Surprise. Kaygusuz and Zuluaga [6] associate a different emotion for every interval, but define different emotion names compared to the basic emotions by Ekman [3], even though the emotion is the same. In table 1 the connection between the different intervals and the emotion are reported.

Interval	Interval Emotion (Kaygusuz and Zuluaga [6])	Basic Emotion (Ekman [3])
Minor 2nd	Unpleasant	Fear
Major 2nd	Neutral	Neutral
Minor 3rd	Tragedy	Sadness
Major 3rd	Happiness	Happiness
Augmented 4th	Danger	Danger
Major 7th	Aspiration	Surprise

Table 1 The Correlation Between Intervals and Emotions

## 4 Speech Emotion Recognition

The SER process can be summarized into two main parts:

- The audio analysis part
- The AI part

In the audio analysis part, features are extracted from the human voice.

Those features are gathered through the use of tools such as the Mel and Gammatone Cepstrum Coefficients (GFCC and MFCC), and they are used to convert the sound of the human voice into a format that AI can interpret.

This audio analysis is common for both the training phase of a machine learning tool, and the emotion recognition processes: In the first case, the features extracted from the audio database

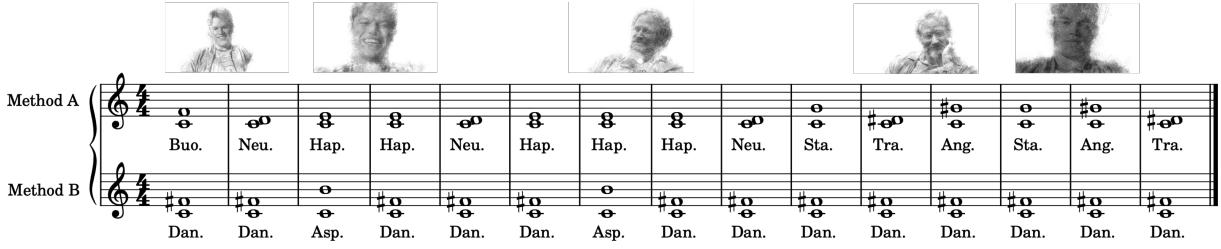


Fig. 3 Scene 3: Good Will Hunting - METHOD A = WAV2VEC; METHOD B = MATLAB

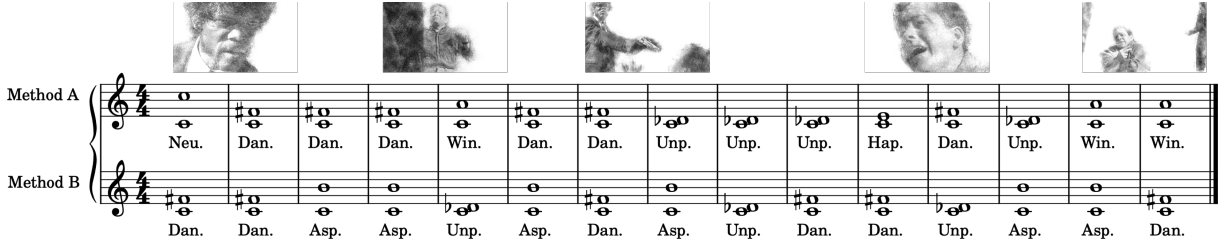


Fig. 4 Scene 4: Pulp Fiction - METHOD A = WAV2VEC; METHOD B = MATLAB

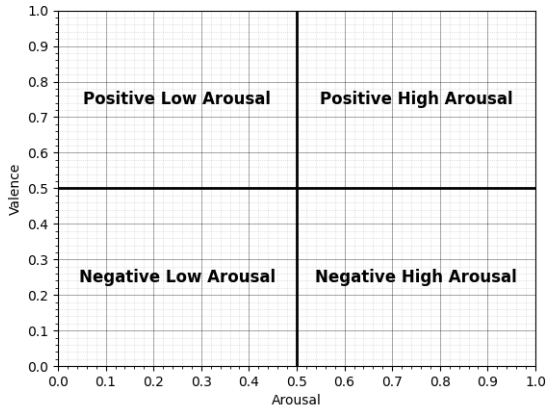


Fig. 5 Valance and Arousal Diagram

files are linked to their respective labeled emotion, in the latter the extracted features are given as an input to the neural network, so that the emotion recognition process can take place.

The AI part acts differently in the two processes. In the learning phase, following the mapping of extracted features to their respective labeled emotional states, AI is tasked with ingesting input data, represented by features, and

employs a predictive function to guess the corresponding emotional outputs. This iterative process is executed multiple times to enhance the predictive accuracy of neural networks, analogous to a sequential trial-and-error learning mechanism observed in human cognition.

Upon the culmination of the learning stage, the neural network is subsequently synthesized, rendering it ready for testing. This synthesis empowers the network to engage in prediction tasks with input data not originally included in its training dataset.

In the emotion recognition phase, the previously synthesized neural network, is fed with the extracted features and after that, a process of analysis between acoustic features and emotional states begins. When new audio samples are fed into the network, AI processes the input features through its learned parameters and outputs a prediction regarding the emotional state conveyed by the speech.

The output of the neural network is typically a likelihood distribution over different emotions. It assigns a likelihood score to each emotion category, and the emotion with the highest likelihood is considered the predicted emotion for that input speech segment.

## 5 Methods

Gathering the musical scores has been carried out with two distinct models, that differ in terms of the network structure and the musical arrangement. A pre-trained network has been used to create compositions based on the VAD Prediction approach, characterized as Method A, while a different Model has been trained from scratch to delve into the musical possibilities of the Six Prediction concept, depicted as Method B. Both will be unfolded in detail in the following chapters.

### 5.1 Method A

The first Method for classifying emotions, utilizes the already established dimensional approach. A model that is fine-tuned on the dimensions' arousal, valence and dominance was chosen to compare the results of the musical outcome to the Six Emotion Approach. In Speech Emotion Recognition, certain model structures have emerged as particularly effective. A common strategy to a SER-Model involves using an automatic speech recognition (ASR) Model as a Baseline. Subsequently, this baseline model is finetuned to a specific Emotion Recognition Dataset. The model that has been used to gather the results of the Dimensional approach, is the `audeerling/wav2vec2-large-robust-12-ft-emotion-msp-dim` Model, which is published on the Hugging Face Hub and documented within the associated paper [8]. This Model involves an ASR-Baseline that is finetuned with the MSP-Podcast, a Dataset for Emotion Recognition. This model uses an approach of combining `wav2vec2-` and `transformer-`architectures. `Wav2Vec`, originally an algorithm to transcribe raw speech data, is often subject of SER and shows promising results, especially for context-based emotion classification [9]. The use of transformer architectures, lies in their ability to reduce training time compared to common recurrent neural architectures [10]. Before finetuning, the Transformer Layers of this model, have been reduced from 24 to 12 Layers, resulting in a better Performance on the Valence Dimension [8].

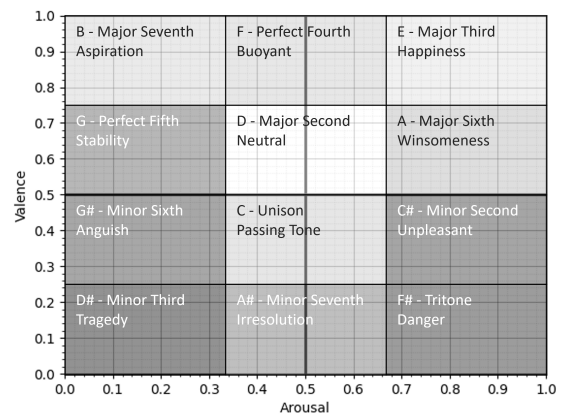
To apply the VAD-Prediction concept in terms of musical arrangement, it is important to view the Valence and Arousal Dimensions as a map where specific rules can be applied. Similarly, to

Note	Semitones	Interval	Feeling
C	0	Unison	Neutral
C#	1	Minor Second	Unpleasant, Dark
D	2	Major Second	Neutral, Passing Tone
D#	3	Minor Third	Tragedy, Sadness
E	4	Minor Third	Joy, Happiness
F	5	Perfect Fourth	Buoyant, Neutral
F#	6	Tritone	Danger, Devilishness
G	7	Perfect Fifth	Stability
G#	8	Minor Sixth	Anguish, Sadness
A	9	Major Sixth	Winsomeness
A#	10	Minor Seventh	Irresolution, Mournfulness
B	11	Major Seventh	Aspiration

**Table 2** Intervals and Their Corresponding Emotion [6]

the Six Emotion Prediction, musical intervals can be mapped to certain parts of the diagram. If we consider everything above the Valence border of 0.5 as mostly positive and everything below 0.5 as mostly negative, it makes sense to locate all major intervals in the upper half of the graph and all minor intervals in the lower half. Using this strategy, a mapping system has been employed, which is based on the work of [Kaygusuz and Zuluaga \[6\]](#). Within the corresponding paper, a diagram shows the correlation of musical intervals and emotions, reported in table 1.

Assuming that the musical interval 'Tritone' is directly linked to an emotion, associated with Danger, it can be located to a part of the graph with low valence and high arousal. Low Valence, because the emotion aligns with the low valence end of the spectrum due to its predominantly negative connotation. High Arousal, because of its connection to intense emotions, places it in the high arousal zone on the graph.



**Fig. 6** VAD-Mapping of intervals

## 5.2 Method B

Method B uses as the main programming environment MATLAB, making use of the Audio Toolbox (for the audio analysis and feature extraction part) and the Deeplearning Toolbox (for the Neural Network part).

The operation of method B can be divided in two main parts:

- Neural Network Synthesis (NNS)
- Speech Emotion Recognition (SER)

### *First Part (NNS)*

The first step is the selection of the emotional audio database. Different databases have been considered and the selection is based mainly on the language and the emotional content of the database itself. The final choice fell to the Surrey Audio-Visual Expressed Emotion (SAVEE) [11], a database in English language with an emotional content in line with the emotions in table 5.

A script was made, which takes as an input the audio database and lists the content of it. Available datasets include a limited number of samples, and, consequently, to improve the performances of the trained model, the user can perform an augmentation process of the database. that applies random transformations (usually noise, pitch shifting and volume changes) to the original files, thus creating an increased and diverse database. Augmentation, in that way, may be seen as a way to increase the accuracy since it adds information to the database using the same files.

We trained a Bidirectional Layer Short-Term Memory (BiLSTM) network on the augmented dataset.

The next and last step is the network validation, which consists in the LOSO (Leave One Speaker Out)  $k$ -fold validation: the network is trained with  $k-1$  speakers and then validated with the remaining speaker.

### *Second Part (SER)*

After the network is ready, the SER phase can be performed. For that, an app has been developed. It takes as inputs a file, which can be audio or video (either way the script will extract the audio part of it), and the network trained with the script.

The time interval for the analysis of the audio trace, analyzing can be selected, spanned from 1

to 5 seconds: after different trials, the best results were obtained with a time interval of 4 seconds. After the prediction starts, the app analyses the audio by dividing it in sections that have length determined by the selected time interval.

The prediction values are normalized, so the sum of all the predicted values is equal to 1. The emotion highest in value determines the musical interval and the two notes that are related to that interval are outputted in MIDI format from the application through an external MIDI interface.

The remaining emotion values are multiplied by 1.2 which converts them to the 0-127 interval of MIDI commands, making it possible to use them as Control Change (CC) signals in order to modulate other parameters (for example, when using a synthesizer).

## 6 Test Design

The test carried out is two fold. In the first part it is necessary to directly compare the two proposed methods in terms of the produced MIDI-Data. Therefore, the Intervals have to be compared by similarity at each time stamp. If the result shows a significant lack of similarity a subjective test must be used.

### 6.1 Similarity of MIDI-Data

To motivate the need of personal judgement, the similarity of the musical intervals produced is checked in a preliminary study. For this purpose, the Pearson correlation was determined (tab. 3). The linear basic fitting of the data is used to represent the measure of determination and compared with the ideal straight line. This can be additionally expressed by the r square parameter. If the similarity of the interval sequences is too great, a subjective comparison by the test persons is not meaningful.

To transform MIDI-Data into a metric scaling C3 is interpreted as 0 and every interval as their half tone step difference to the root note (C3). Therefore the values range from 0 to 12. The r-value shows the correlation of the intervals. Table 3 shows a low correlation of all four scenes therefore a subjective test is necessary.

Scene	r-value
The Trueman Show	0,03
Interstellar	-0,26
Good Will Hunting	0,00
Pulp Fiction	-0,09

**Table 3** R-Value of the presented movie scenes

## 6.2 Subjective test

In the following, the inductive empirical study to determine the suitability of VAD and labelled emotion categorisation for the generation of musical intervals to support film dialogue is presented. The goal of this test is to assess which method is more suitable in the context of the developed MIDI note generation system discussed in chapter 3. For the suitability test, an online questionnaire with sound samples and the corresponding film scenes is prepared, which has been proposed to 30 participants. The age and gender of the participants is not considered due to the anonymity of the online survey distribution platform. In order to collect valid data, it is essential to define the general conditions for both methods, such as sound design and loudness. In order to obtain quantitative data, a bidirectional rating system was chosen, with a discrete range from -60 to 60. Values towards a negative value classifies method A as more suitable and vice versa. The test design is derived from the ITU/EBU listening tests (ITU-R BS.1284-2 [12]). It is well known that the referenced test procedure is generally used to assess the quality of the audio signal, but this can provide a suitable test platform for general listening tests. However, it is important to define the listening conditions for the subjects. In the survey they must confirm that they are using headphones. As described in chapter 2 the VAD uses 12 different intervals to display the emotion and the labeled only 6. This is a fundamental difference of both systems which needs to be analyzed.

The created audio files are rendered in 44,1 kHz and 24 bit. To keep the listening performance constant over the duration of the test and to prevent fatigue, the duration of the sequences does not exceed 20 seconds. For the Test four film sequences are prepared:

- The Truman Show
- Interstellar
- Pulp Fiction

- Good Will Hunting

These scenes were chosen because they do not have background music in the original. Thus, the subjects are not biased. The scenes are for each participant presented in the same manner. First they have to listen to Method A and than to Method B. Therefore four depended sample sets are created.

Statistic wise, the gathered data will be analyzed and compared by its mean to classify the overall suitability of the audio files to the movie scenes. Respectively, four paired sample t-test will be carried out.

## 7 Results

The preliminary study shows the differences of the SER systems. In Table 1 it is evident that the generated MIDI-Data for each film significantly differs. For the analysis of the data, however, it must be noted that due to scaling and to the range of values of data, small changes suggest a large deviation in similarity. The results of this survey are displayed within Table 4 and Table 5.

Scene	n	Avg	Std Dev	Std Err Avg
1	32	6.91	44.33	7.84
2	33	-14.03	37.02	6.44
3	33	-12.12	32.94	5.73
4	35	9.46	36.81	6,22

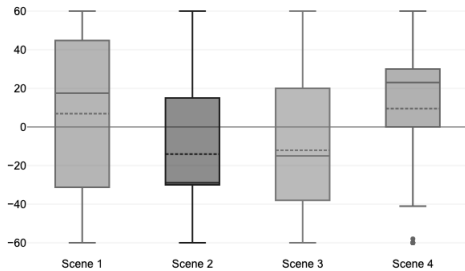
**Table 4** Results of the subject test of four movie scenes

Scene	t	df	p	Avg Diff.
1	0,88	31	0,385	6,91
2	2,18	32	0,037	-14,03
3	2,11	32	0,042	-12,12
4	1,52	34	0,138	9,46

**Table 5** Results of the conducted t-test of four movie scenes

The T-Test Comparison is carried out with the use of Boxplots displayed Figure 7. The statistical results from the survey comparing the effects of Method A and Method B on four different movie scenes reveal variations in the impact of these methods. In Scene 1, the data is relatively spread, with a mean of 6.91 and a median of 17.5, indicating potential differences between the two

methods. Scenes 2 and 3 exhibit negative means (-14.03 and -12.12, respectively), suggesting that Method A might have a more adverse effect on these scenes compared to Method B. Scene 4, on the other hand, shows a positive mean of 9.46, implying that Method B could be more favorable for this particular scene.



**Fig. 7** T-Test Comparison

The statistical results indicate that there is no significant difference between both methods in perception regarding scene 1. The average amounts to 6.91. This value indicates that most subjects don't have a clear Preference for both methods (A or B) in regard to the appropriate depiction of emotions in scene 1. The standard deviation of 44.33 shows, that the reviews of most subjects are varying strongly. This may be an indication, that some subjects prefer one method while other subjects prefer the other. The t-value measures the differences between both averages (6.91) and shows how significant both differences are. In this case the t-value is not significant, because it is above the level of significance of 5 Percent. So, the differences of both methods is not statistically significant. The analysis here shows that the data comes from a population where the average is 0, so our average of 6.91 cannot be considered generally valid. However, it does show a tendency towards method B in the context of this survey.

In Scene 2, the average preference for method A is -14.03, with a standard deviation of 37.02, indicating varying opinions among subjects. The t-value of 2.18 and p-value of 0.037 reveal statistical significance, showing that subjects find method A more appropriate for depicting emotions in Scene 2.

For Scene 3, the average preference leans towards method A at -12.12, with a standard deviation of 32.94, indicating a range of opinions among the subjects. The t-value of 2.11 and a p-value of 0.042 confirm the statistical significance, revealing a distinct preference for method A in conveying emotions in Scene 3.

Moving on to Scene 4, the average preference is 9.46, favoring the music generated using method B. However, the standard deviation of 36.81 suggests substantial variability in subjects' opinions within this scene. Surprisingly, the t-value of 1.52 and a p-value of 0.138 do not demonstrate statistical significance, implying that there is no clear preference between methods for portraying emotions in Scene 4.

## 8 Discussion

Based on our comprehensive analysis, an inclination towards favoring method A becomes evident. It is crucial to consider the factors that may have influenced participants' preferences for method A (12 intervals) over method B (6 intervals) in our study on speech emotion recognition and music generation. While our results revealed that for two of the four film scenes, participants favored method A, and for the remaining two, they preferred method B, it is essential to acknowledge that the observed differences were not statistically significant, with p-values exceeding the conventional threshold of 5 Percent. This lack of statistical significance suggests that factors other than the number of intervals might be at play in determining participants' choices.

One plausible explanation for these findings is that the perceived preference for method A might not necessarily be a result of the music's better alignment with the emotional content of the speech data. Instead, it is reasonable to assume that participants may have favored method A because the music generated with 12 intervals was potentially less monotonous compared to the music generated with only 6 intervals. In other words, the greater variation in pitch and melody offered by method A could have made the music more engaging and enjoyable for the participants, even if it did not necessarily better convey the intended emotions from the speech data. This suggests that future studies should consider the influence of musical complexity and diversity on

participants' preferences, as these factors may play a significant role in the subjective evaluation of emotionally expressive music.

## 9 Conclusion

Since the focus of this paper is on the coherence between emotion and interval, no extensive research has been done on sound design and rhythm. Intervals are a fundamental dimension of the production of affective music, but they represent only a subset of the broad spectrum of musical expression. Emotions can be evoked by a wide variety of principles of music, which could not be dealt with extensively in the present work. In particular, rhythm, chords, and melodies are other aspects that can have a significant influence on the emotional impact of music. Future research could therefore deal with these aspects more intensively and thus deepen the understanding of the many ways in which emotions can be generated by musical design elements.

## Declarations

On behalf of our Team we would like to express our gratitude to 'Felix Dennerlein' for his contributions to the paper. Mr. Dennerlein's transformation of images into illustrations added a distinctive dimension to our paper and his creative input and versatility in undertaking various tasks beyond his core responsibility greatly enriched our project.

## References

- [1] American Psychology Association. Apa dictionary. URL <https://dictionary.apa.org/emotions>.
- [2] Albert Mehrabian and James A. Russell. An approach to environmental psychology. 1974. URL <https://api.semanticscholar.org/CorpusID:143333487>.
- [3] Paul Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38, 1992. doi: 10.1111/j.1467-9280.1992.tb00253.x. URL <https://doi.org/10.1111/j.1467-9280.1992.tb00253.x>.
- [4] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.
- [5] Stuart Cunningham, Harrison Ridley, Jonathan Weinel, and Richard Picking. Supervised machine learning for audio emotion recognition. *Personal and Ubiquitous Computing*, 25(4):637–650, Aug 2021. ISSN 1617-4917. doi: 10.1007/s00779-020-01389-0. URL <https://doi.org/10.1007/s00779-020-01389-0>.
- [6] Cengiz Kaygusuz and Julian Zuluaga. Impact of intervals on the emotional effect in western music, 2018.
- [7] Kompetenzzentrum Ohm-Ux and Nürnberg Georg Simon Ohm. Methoden zur Gestaltung von UX-Sounds.
- [8] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap, March 2022. URL <http://arxiv.org/abs/2203.07378>. arXiv:2203.07378 [cs, eess].
- [9] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, April 2021. URL <http://arxiv.org/abs/2104.03502>. arXiv:2104.03502 [cs, eess].
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL <http://arxiv.org/abs/1706.03762>. arXiv:1706.03762 [cs].
- [11] Sana ul haq and Philip Jackson. Multi-modal emotion recognition. *Machine Audition: Principles, Algorithms and Systems*, 01 2010. doi: 10.4018/978-1-61520-919-4.ch017.
- [12] BS.1284 : General methods for the subjective assessment of sound quality. URL <https://www.itu.int/rec/R-REC-BS.1284/en>.