Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Analyzing the relationship between sound, color, and emotion based on subjective and machine-learning approaches

J. Kurilčik[a], M. Połom[a], M. Jankowski[a], O. Kozłowska[a], A. Łabich[a], E. Skiba[a]
P. Spierewka[a], P. Śliwiński[a], B. Kostek[b]*

[a]*Gdańsk University of Technology, ETI Faculty, Department of Biomedical Engineering, Narutowicza 11/12, 80-232 Gdańsk, Poland*
[b]*Gdańsk University of Technology, ETI Faculty, Audio Acoustics Laboratory, Narutowicza 11/12, 80-232 Gdańsk, Poland*

Abstract

The aim of the research is to analyze the relationship between sound, color, and emotion. For this purpose, a survey application was prepared, enabling the assignment of a color to a given speaker's/singer's voice recordings. Subjective tests were then conducted, enabling the respondents to assign colors to voice/singing samples. In addition, a database of voice/singing recordings of people speaking in a natural way and with expressed emotion was prepared, where discrete colors were assigned in subjective tests. These data were used in a machine-learning approach that consisted in searching for the relationship between sound, color, and emotion. Analyses based on correlational analysis and learning algorithms were performed. It occurred that assigning values of naturally sounding and emotionally charged speech/singing parameters to colors (and their parameters) did not enable finding a correlation between the given voice, emotions, and color features. The machine learning model achieved high accuracy in the relation between the generated colors and the colors corresponding to the emotions in the literature and questionnaire results.

Keywords: Sound; Color; Emotion; Speech-color correlation; Machine learning

* Corresponding author. Tel.: 48-58-3472717.
E-mail address: bokostek@audioakustyka.org

1. Introduction

The phenomenon known as synesthesia, which involves the interconnectedness of visual and auditory perception in the human mind, has been acknowledged for an extended period and has found applications in diverse fields like painting, architecture, environmental design, and pattern design. In psychology, synesthesia is defined as a mental process that elicits an experience from one sense by stimulating another, wherein stimuli to one sense organ provoke responses from other sense organs [1]. It can encompass various senses. Within the realm of audiovisual synesthesia, there exists a phenomenon wherein individuals associate sounds with colors. People psychologically attribute specific colors to sounds, as articulated by the musician Marion, who referred to “audible color” and “visible music” [2]. Numerous experiments employing brain imaging techniques such as fMRI or PET have been conducted to explore this phenomenon, identifying the brain areas responsible for integrating audiovisual stimuli [3, 4, 5].

There are distinct parallels between sound and color, manifesting in the convergence of two artistic domains: music and painting. In the realm of art, colors are often described using sound terms such as harmony and disharmony. Moreover, painting employs musical terminology to characterize form, style, and content [6]. Similarly, in music, form, style, and content can be depicted using colors, as evidenced by composers like Rimsky-Korsakov or Alexander-Skryabin, who assigned specific colors to particular keys [7]. Applications of color music, including color musical instruments or musical fountains, not only impact aesthetics but also offer new avenues in the area of art therapy.

The above short introduction constitutes the motivation behind our study, especially as there are state-of-the-art deep models that may facilitate analyzing color, sound, and emotions automatically and enable to compare the outcomes with subjective tests. Therefore, the paper is organized as follows. It starts with a short literature overview investigating the relationship between sound and color. This is followed by data gathering for experiments. To that end, a survey was created to relate sound, color, and emotions. The respondents’ task was to assign a color or colors that, in their perception, best reflected the impressions associated with each recording. Further on, the results obtained from the survey-based investigation were analyzed.

The second part of the research project focuses on developing predictive models for the automatic recognition of emotional content in sound recordings and then, based on survey results, assigning a specific color to a given sound sample. For that, publicly available collections of sound recordings, such as RAVDESS─Ryerson Audio-Visual Database of Emotional Speech and Song [8], CREMA (Crowd-sourced emotional multimodal actors’ dataset) [9], TESS─Toronto Emotional Speech Set [10], SAVEE─Surrey Audio-Visual Expressed Emotion [11], MIREX─Music Information Retrieval Evaluation eXchange [12] were employed. To explore the relationship between sound, color, and emotions in speech and signing automatically, convolutional neural networks (CNNs) were used. This is followed by a discussion on the algorithmic approach to the sound-color-emotion relationship was carried out on the basis of metrics obtained. Finally, some concluding remarks were delivered.

2. Literature overview

Two primary correlation mechanisms emerge when examining the relationship between sound and color: direct timbre-color correlation and indirect correlation through emotions. The former entails direct connections between sound and color elements, with early research explaining this correlation in terms of the physical isomorphism of timbre and color [6]. Investigation of this relationship goes back to ancient Greece and then returned in the 16th century [1]. As already said, artists, but also poets, historians, writers, and researchers were interested in the correspondence between sound and color [1, 13]. Newton suggested a relationship between the seven musical tones and the seven colors of the rainbow, substantiated through calculations involving vibration frequencies and wavelengths [13]. Furthermore, he noted that both color and sound result from the physical effects of external forces, with light reflection generating color and vibration producing sound [2]. Moreover, both color and sound can be digitally represented spectrally and share characteristics of synthesis and decomposition. Sound includes pitch and overtones, as does color, which can be monochromatic or complex.

In a study by Zhou [14], an experimental psychology approach was employed to qualitatively establish a correlation between visual and auditory attributes. Auditory attributes such as pitch, intensity, and sound tension were analyzed, revealing differences in how individuals perceived music in correlation with varying visual perception. While experimental results indicate a certain correlation between visual and auditory perception, specific factors or visual features correlated with distinct auditory features remain unspecified.

Indirect correlation through emotions suggests that both sound and color are linked to emotional experiences. Different sounds evoke various auditory effects, including consonance, disharmony, thickness, delicacy, and lightness. Similarly, different color combinations generate visual sensations analogous to the described auditory effects. A research team affiliated with the University of California, Berkeley, embarked on a series of experiments delving into the correlation between sound and color [15]. In the initial phase, emotions were employed as an intermediary factor [15]. In a study by Palmer and colleagues [16], excerpts from Mozart, Bach, and Brahms were utilized, with variations in length and tempo yielding 18 audio segments. Subsequently, four emotional dimensions were identified. Participants were tasked with selecting the emotion best representing a given audio fragment and then assigning a corresponding color based on that emotion. Analyzing the data from these experiments led to the conclusion that both music and color exhibited associations with specific emotions. In line with prior research, this association mechanism was documented by scientists, confirming the emotional mediation hypothesis, as also affirmed in another article [16]. Palmer et al. extended their study using four different types of Mozart's melodies in a similar experimental approach, confirming the emotional mediation hypothesis based on 64 audio materials.

As the research progressed, adjustments were made to materials, methods, and experimental subjects. In a study detailed in [17], the same research concept was applied, but this time involving individuals without synesthesia. Participants were required to choose the more fitting set of colors for the sounds they heard, alongside rating each sound and color on five emotional dimensions. Through correlation analysis, the authors concluded that even individuals without synesthesia exhibited color associations with musical sounds. Building on previous findings, Griscom [18] focused on cross-modal correspondence in a musical context, exploring whether individuals consistently linked the same colors to musical intervals and chords. The study also investigated the consistent visual association between timbre and music, as well as the relationship between normal cross-modal correspondence and a neurological state termed "sympathetic." Employing experimental methodologies akin to prior studies, researchers found consistent associations between visual and musical features while highlighting the role of semantic and emotional features in driving these cross-modal correspondences.

Liu et al., in their study [19], delved into a thorough quantitative analysis of the sound-color relationship, identifying objective sound parameters and analyzing the connection between sound and color through model creation. An experiment based on subjective perception using audiovisual timbre-color correlation was conducted. A comprehensive set of information regarding sound-color connections was obtained through statistical methods. The analysis of correlations between these data confirmed a certain relationship between the timbre and color dimensions. Subsequently, three algorithms—multiple linear regression, BP neural network, and SVR—were employed to create a timbre-color correlation model, with the precision of these models subsequently verified.

In summary, current research in the area of experimental psychology predominantly centers on analyzing the relationships between visual attributes (such as color, texture, and shape) and auditory attributes (including music, major and minor tones, and rhythm). Emotions serve as a primary focus, indicating a qualitative approach. The outcomes of such research may either substantiate a direct connection between music and color or reveal that the connection between the two occurs through the medium of emotion and perception. There is, however, a limited amount of quantitative research exploring the specific relationship between the objective characteristics of music and color attributes. That is why our research study began with preparing and conducting a survey to collect data regarding the correlation between a voice/singing sample and the color assigned to that sample.

3. Preparation of a database of emotionally charged speech/singing recordings

Creating a dataset with a sufficient number of examples is essential in learning AI (artificial intelligence) models. To achieve satisfactory results, four databases for speech emotion recognition and three for music emotion recognition were combined. Each database consists of audio files in English. For speech, we decided to use the already mentioned RAVDESS, TESS, CREMA-D, and SAVEE databases.

RAVDESS contains recordings of 24 professional actors (12 women, 12 men) speaking two lexically matched statements in a neutral North American accent. Speech and songs contain expressions of peace, joy, sadness, anger, fear, surprise, and disgust. In addition, the song contains emotions of peace, joy, sadness, anger, and fear. Each expression is produced at two levels of emotional intensity (normal and strong), with an additional neutral expression. TESS is a set of 200 target words spoken in the carrier phrase "Say the word" by two actresses (aged 26 and 64). These recordings were made of the set depicting each of the seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). CREMA-D is a dataset of 7,442 original clips from 91 actors. These clips were made by 48 male and 43 female actors between the ages of 20 and 74 coming from various races and ethnicities (African American, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified). The SAVEE database was recorded from four native English male speakers, postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness, and surprise. A neutral category is also added to provide recordings of 7 emotion categories.

The dataset for song emotion recognition is based on the previously mentioned RAVDESS, Multi-modal MIREX-like emotion dataset, and Emotify. Mirex contains 903 audio clips (30 seconds), 764 texts, and 193 midis. To the best of our knowledge, this is the first emotion dataset containing these 3 sources (audio, lyrics, and MIDI). The files were divided into the following categories: Bubbly, Confident, Passionate, Thrilling, Noisy, Sympathetic - good-natured, Cheerful, Fun, Revelry, Sweet, Autumn, Bittersweet, Thoughtful, Literate, Touching, Wistful, Campy, Humorous, Silly, capricious, witty, twisted, aggressive, fiery, intense, tense–restless, visceral, volatile.

Emotify consists of 400 song excerpts (1 minute long) in 4 genres (rock, classical, pop, electronic). The dataset was divided into nine emotional categories: Amazement, Solemnity, Tenderness, Nostalgia, Calmness, Power, Joyful, Tension, and Sadness. The annotations to this database were collected using a game. The annotations produced by the game are spread unevenly among the songs, caused by the design of the experiment and the game. To standardize and combine audio, all files were categorized into six emotions: Anger, Fear, Happiness, Neutral, Sadness, and Disgust (only for speech). Both databases were used in questionnaires to gather responses of randomly encountered people. Participants categorized each given audio track into six emotions. All collected data were then used in machine learning. In addition, data were augmented to increase further the number of examples used for the learning model. The noise was added to the tracks. Also, signals were stretched, and the pitch was changed. This operation tripled the amount of data [20, 21].

4. Building a survey application

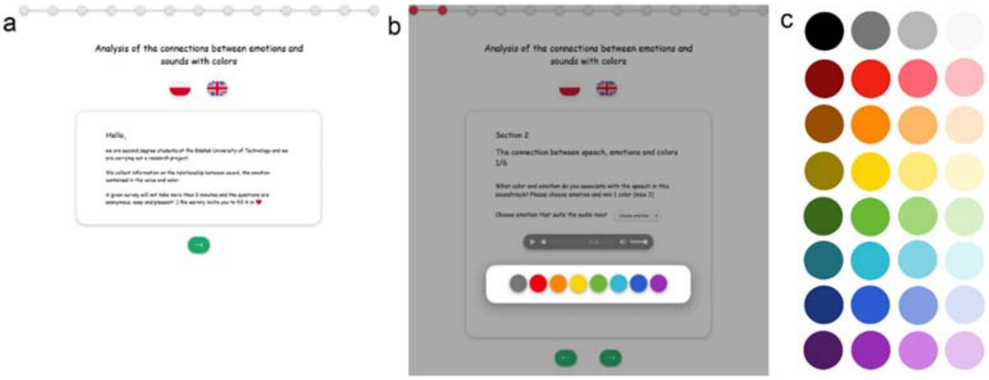
Developing an adequate survey is of great importance because a shared and completed survey is a valuable source of information for research carried out. Namely, these are input data for machine learning. Therefore, in the context of the analysis of the relationship between the sound of the voice and the color assigned to the speaker's/singing voice sample, a survey was created in which the respondents could allocate the most suitable colors to the records based on their feelings.

There were several technology-related requirements necessary to meet while creating a survey, i.e., the ability to add a color palette as a response option, to add an audio track, no limits on the number of questions and answers per survey, or a minimum limitation of at least 1,000 responses. The chosen method to acquire such a tool is building a web application using JavaScript programming language and the React framework. As they provided sufficient tools to build a simple and robust website that the survey could be placed in. To optimize the application development process, pre-made templates were adapted to the study's individual needs [22].

The Python programming language was used with the Django framework when creating the website's back-end. To enable communication between the client and the server, an appropriate API (application programming interface) was created. For this purpose, the Django Rest Framework tool was used. To make the application available online, a Virtual Private Server from the OVHCloud platform was used.

The survey is presented in Fig. 1 and is divided into parts: part 1–an introduction of the purpose of the study, encouraging people to complete the survey (Fig. 1a); part 2–general information about the respondent: gender, age, and possible colorblindness; part 3–the main part of the survey, containing questions about color and emotion associations with shared speech and singing audio tracks (Fig. 1b). The survey provides each respondent with one audio sample from speech and singing category (labeled as disgust, fear, sad, neutral, angry and happy in datasets).

The singing section does not have a ‘disgust’ labeled audio as it is not available in the used datasets. Part 4 contains consent to data processing. The answers are sent to the database after clicking the "Consent and submit survey" button.



Analysis of the connections between emotions and sounds with colors

Hello,

we are second degree students of the Gdańsk University of Technology and we are carrying out a research project.

We collect information on the relationship between sound, the emotion contained in the voice and color.

A given survey will not take more than 3 minutes and the questions are anonymous, easy and pleasant :) We warmly invite you to fill it in ❤

Analysis of the connections between emotions and sounds with colors

Section 2

The connection between speech, emotions and colors 1/6

What color and emotion do you associate with the speech in this soundtrack? Please choose emotion and min 1 color (max 3)

Choose emotion that suits the audio most: choose emotion

Fig. 1. (a) Title page of the survey; (b) Color selection; (c) 32 possible colors to choose from.

As mentioned earlier, the main point of the survey is to ask the respondents to assign an emotion and a color to a presented recording of speech and singing. A person can choose one emotion from the following: disgust, fear, sadness, neutral, anger, and happiness. The idea is to check if the chosen emotion will be the same as labeled in datasets. A maximum of three colors can be selected to prevent situations when a person cannot choose one color because of changing feelings. The palette consists of seven basics plus monochromatic colors, each of them available in four different shades. In total, this gives 32 possible colors to choose from. The idea of using 7 basic colors (also known as ROYGBIV - the acronym describing the sequence of rainbow colors) is based on the color circle division proposed by Sir Isaac Newton. The number of colors of the proposed palette is big enough to provide a satisfactory range of choices to the respondent and small enough to provide clear results. The color selection process is simple: click on one of the gray circles, display the color palette (shown in Fig.1b), and expand one of the main 8 colors to get an additional four shares (presented in Fig.1c)

5. Survey results

Based on the created survey, subjective tests were carried out, in which respondents assigned firstly emotions and then colors to the presented audio of speech and singing. Responses were received from 141 persons; 81 of them were women, 59 were men, and one person did not state their gender. Eight people described themselves as having problems with color recognition.

Some people expressed their opinions on the structure of the survey. The main observation was the occurrence of only one positive emotion (“happiness”) among the rest of the emotions considered as more negative. However, the emotions included in the study were selected based on the scientific articles and existing datasets. Respondents found some recordings difficult to evaluate because the emotions were less clearly outlined and required greater focus. The survey was perceived as easy to complete; participants went through all its stages efficiently, and the instructions themselves were specific and understandable.

The results obtained for the correlation between speech/singing and emotions are shown in Fig. 2. For the correlation between speech and colors (see Fig. 2a), it is noticeable that the chosen emotion “neutral” appears often (for audio with labels “happy,” “sad,” and “fear”). Emotion neutral may be considered a conservative option, especially for less emotionally charged recordings. Nevertheless, many respondents correctly assigned emotions to the audio labels. For two out of six audio labels (“angry” and “neutral”), more than half of the respondents chose the same emotion as assigned in the datasets. The emotion that was recognized the best is “anger,” which is perceived as very intense. The most difficult to distinguish is “disgust,” and only 9% of people managed to "guess" it.

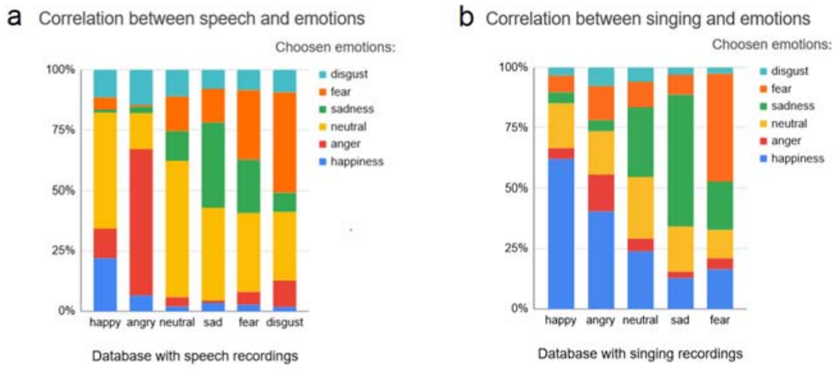


Fig. 2. (a) Correlation between speech and emotions; (b) Correlation between singing and emotions.

Fig. 2b presents the results for the correlation between singing and emotions. This part of the survey did not include the audio labeled as “disgust” because no recordings were included in most datasets. For three out of five audio labels (“happy,” “sad,” and “fear”), the highest percentage of correct answers was obtained. The best-recognized audio emotion is “happy,” and 62% of respondents correctly described it as happiness.

Next conclusions are derived based on the part of the survey focusing on the correlation between speech/singing and colors. The chosen colors are presented in Fig. 3 in columns on a percentage scale. It is noticeable that both speech (Fig. 3a) and singing (Fig. 3b) results are similar to each other. For the audio labeled with “happy,” bright shades of yellow, orange, and green were chosen. The “angry” label recordings were mostly associated with more aggressive colors - intense red, orange, and purple. For the “neutral,” users selected various light-colored hues. The “sad” label audio corresponded to shades of light and dark blue. Finally, there was no dominant color for the “fear” and “disgust” labels, but all received colors were intense.

The developed survey regarding the correlation between speech, singing, emotions, and colors helped to obtain general, valuable observations based on respondents' subjective assessments. The obtained results constitute part of the input data for the machine learning performed, so they are an important element of the study.

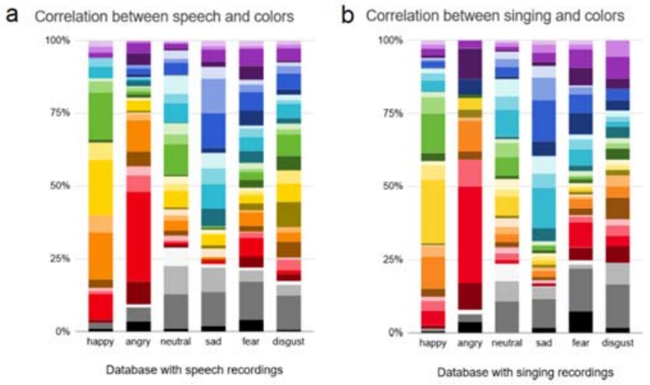


Fig. 3. (a) Correlation between speech and colors; (b) Correlation between singing and colors.

6. Correlation between speech/singing signal parameters and the assigned colors in subjective tests

The next step in preprocessing was audio signal feature extraction. From each track, the following features were extracted: Zero Crossing Rate, Chroma stft, MFCC, RMS (root mean square) value, and Mel-Spectrogram. To find a correlation between each parameter and the chosen color, the value of color was represented in two formats: RGB and HSV. The first one is the common way of coding a value of color, but the second one refers to how the human vision organ works. The results show that there is no direct correlation between parameters and each component. To find the most significant parameters to be used in training, an algorithm of principal component analysis was used. This led to the conclusion that Mel-Spectrograms and RMS values retain the most information from all features.

7. CNN-based analysis

A part of the research project focused on developing predictive models for automatic recognition of emotional characteristics in audio recordings related to speech and separately for music/singing. The models employed were deep convolutional neural networks (CNNs) for both speech and singing. The use of a feature-based CNN model is supported by the fact that the input data in the realized approach is a large number of complex coefficients. There are also different tools that could be used: LSTM (Long Short-Term Memory) and wavelet transform. Other researchers reached accuracy rates of 70.34% [23], 61% [24] using LSTM, 81.12% [25], and 88.67% [26] using wavelet transform. Researchers who used CNN reached 80.3% [27] and 80.89% [28]. Based on survey results, specific color assignments were made for each audio sample with regard to class assignment scores. Model training involved iterative adjustments of weights and parameters to achieve optimal predictive capabilities and avoid overfitting. The training process was based on the earlier audio recording datasets and survey data presented earlier. Training, validation, and testing were conducted on different sets of records. Finally, the models were evaluated using metrics.

The initial data preprocessing began with the separate loading of records from each dataset in a standardized manner to facilitate their later combination into a larger set. The file path and emotion label corresponding to each recording were extracted in the record loading. The file path served as the basis for data augmentation and feature extraction used in model training. Data augmentation aimed to expand the original dataset by adding additional records with random Gaussian noise and pitch modification. Feature extraction from recordings involved loading a specified length of the audio signal: 2.5 seconds for speech and 8 seconds for singing/music. Additionally, these signals were divided into time windows, for which coefficients such as Zero Crossing Rate, Root Mean Square Energy, and Mel-Frequency Cepstral Coefficients were computed. These three types of features were eventually combined into a single resulting array, used as the input for the feature space along with labels for the models.

Both speech and music/singing utilized the same architecture of a deep convolutional neural network. The networks consisted of four convolutional layers (CL) with 256, 256, 128, and 64 filters, a kernel size of 5, a stride of 1, a ReLU (rectified linear unit) activation function, and zero-padding (see Fig. 4). A batch normalization layer and a max-pooling layer with a size of 5 and a stride of 2 were added after each CL. The output from the 4th CL was an input to a fully connected (dense) layer with 32 neurons, followed by another batch normalization layer. The model's output was a dense layer with 6 neurons (number of classes) and a SoftMax activation function used for the final multi-class classification. Additionally, dropout regularization with dropout probabilities of 0.2 and 0.25 was applied at the output of the 3rd CL and the output of the fully connected layer, respectively.

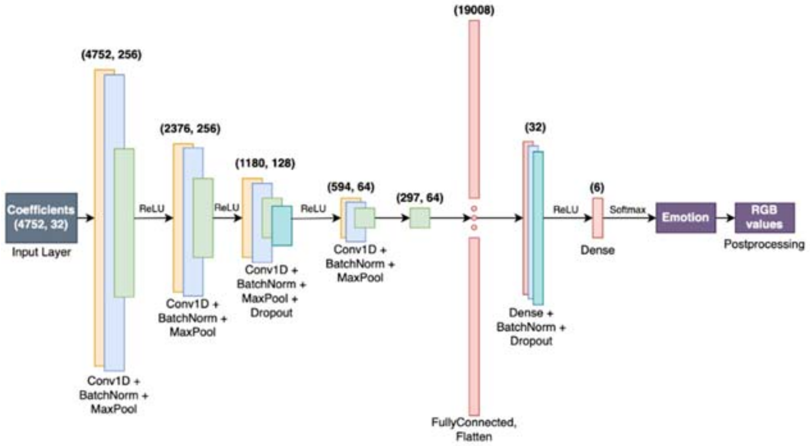


Fig. 4 Deep convolutional neural network architecture used in the study for speech/singing.

The Adam optimizer was employed with an initial learning rate of 0.0005. Categorical cross-entropy as the loss function and accuracy were selected as the measures of training performance, and its value evolved during training epochs. A learning rate reduction technique was applied based on the accuracy metric for the validation set. The learning rate was halved whenever the accuracy on the validation set did not improve for the next three epochs. Training for the models lasted for 100 epochs for singing/music and 50 epochs for speech. The final prediction results for specific recordings were used to select colors from the RGB space based on the investigated color-emotion correlation.

Table 1 shows the color for each emotion acquired for the survey, and Table 2 presents a sample final result of color assignment and classification with regard to the average-weighted color of a probability output. Both were done separately for speech and singing/music in the survey-related part of the project. The algorithm considered specific RGB values for each emotion, adjusted according to the model's confidence level. The proposed post-processing algorithm can use any set of colors as an input. It was decided to use color-emotion data from the survey to connect both parts of the research. Personalized color palettes were obtained for each prediction, allowing for a visual representation of the emotion-color correlation, similar to the survey results.

Table 1. Average color values obtained as a result of the speech/music survey.

emotion		RGB for speech		RGB for colors
Happy	🟨	(196; 155; 63)	🟨	(191; 164; 76)
Angry	🟥	(188; 74; 59)	🟥	(185; 57; 47)
Neutral	🟩	(165; 175; 141)	🟩	(169; 178; 149)
Sad	🟦	(132; 154; 178)	🟦	(119; 146; 172)
Fear	🟫	(136; 118; 119)	🟪	(127; 100; 118)
Disgust	🟫	(159; 151; 106)	🟫	(152; 117; 117)

Table 2. Model predictions for first 10 labels for speech/singing using weighted average probabilities of emotion attribution.

	Speech model			Singing model		
Index	Actual label	Predicted Labels	Color	Actual label	Predicted Labels	Color
1	Happy	Happy	rgb(195; 155; 64)	Fear	Neutral	rgb(169; 177; 149)
2	Angry	Neutral	rgb(167; 165; 134)	Happy	Happy	rgb(191; 164; 76)
3	Fear	Sad	rgb(132;153;177)	Neutral	Neutral	rgb(169; 178; 149)
4	Neutral	Sad	rgb(140; 159; 169)	Fear	Neutral	rgb(155; 152; 139)
5	Happy	Happy	rgb(196; 155; 63)	Happy	Happy	rgb(191; 163; 76)
6	Neutral	Neutral	rgb(165;175; 141)	Sad	Happy	rgb(187; 163; 81)
7	Neutral	Neutral	rgb(164; 173; 143)	Fear	Fear	rgb(127; 100; 118)
8	Angry	Angry	rgb(188; 74; 59)	Neutral	Neutral	rgb(169; 177; 149)
9	Angry	Angry	rgb(183; 82; 67)	Happy	Happy	rgb(189; 123; 65)
10	Fear	Fear	rgb(136; 118; 119)	Angry	Neutral	rgb(183; 71; 59)

8. Machine learning results

Input data from the databases employed were augmented. This process consisted of noisemaking and stretching the samples, shifting them in the time domain, as well as changing the tone values. Next, their characteristic features were extracted from the samples, and it was ensured that there were no incorrect values in the data. Further, the classes were encoded using the OneHotEncoder class, the division into learning, verification, and test sets was done, and the data were standardized. The next steps encompassed the design of the neural network model.

The confusion matrix in Fig. 5a, as well as the metrics in Table 3 (Speech), indicate a balanced distribution of classification errors. One can distinguish better overall recognition of anger and slightly worse precision score values for neutrality and disgust. This aligns with general knowledge, according to which anger is the most expressive emotion. Neutrality, on the other hand, is an emotion that can be confused with others more easily, which explains the fact that precision is at a lower level than recall.

In Fig. 5b and Table 3 (Singing Model), one can observe that the network model for singing performs slightly worse than the one for speech. What is noteworthy is the large number of errors visible on the confusion matrix in Fig. 5b when distinguishing between happiness and anger. This may be due to the fact that both of these emotions are stronger than the others, which translates into the parameters used in learning. The metrics show more instability between precision and sensitivity compared to the previous model, but this may be due to a poor data set.

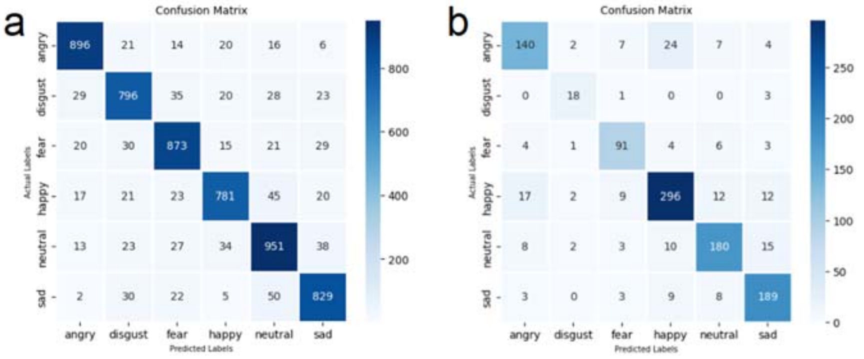


Fig. 5. (a) Confusion matrix for speech model; (b) Confusion matrix for singing model.

Table 3. Evaluation metrics for speech/singing model.

	Speech model				Singing model			
Emotion	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Angry	0.92	0.92	0.92	973	0.81	0.76	0.79	184
Disgust	0.86	0.85	0.86	931	0.72	0.82	0.77	22
Fear	0.88	0.88	0.88	988	0.80	0.83	0.82	109
Happy	0.89	0.86	0.88	907	0.86	0.85	0.86	348
Neutral	0.86	0.88	0.87	1086	0.85	0.83	0.84	218
Sad	0.88	0.88	0.88	938	0.84	0.89	0.86	212
Accuracy			0.88	5823			0.84	1093
Macro avg	0.88	0.88	0.88	5823	0.81	0.83	0.82	1093
Weighted avg	0.88	0.88	0.88	5823	0.84	0.84	0.84	1093

9. Conclusion

As part of the study, a detailed survey was conducted with Europeans (mostly Poles). The results of the survey may vary across different cultural backgrounds; the feeling that a red-colored thing may evoke could differ between Europe, Middle Asia, and East Asia. The survey provided data on the correlation between emotions, sound (speech and singing), and color. Thanks to the analysis of this data, it was possible to supplement the data set with information crucial for the success of the research. The classifier achieved high accuracy, and the generated colors are related to the colors corresponding to the emotions in the literature. This is only a basic network, but given more effort and examples, it is clearly possible to develop an even more sophisticated and robust neural network used to connect emotions, sounds, and colors.

References

[1] "A historical perspective on the relationship between sound and color, from Newton to the XXI century" Lecture in Berlin, 10 Nov. 2012, International Guitar Academy. http://www.marcodebiasi.info/en/a-historical-perspective-on-the-relationship-between-sound-and-colour/ (Retrieved Feb. '2024).

[2] Safran, Avinoam B. and Nicolae Sanda (2015) "Color synesthesia. Insight into perception, emotion, and consciousness." *Curr Opin Neurol*. **28** (**1**): 36-44. doi: 10.1097/WCO.0000000000000169. PMID: 25545055; PMCID: PMC4286234

[3] Bushara, Khalafalla O. Jordan Grafman, and Mark Hallett (2001) "Neural correlates of auditory-visual stimulus onset asynchrony detection." *J. Neuroscience*, **21** (**1**): 300–304. doi: 10.1523/JNEUROSCI.21-01-00300.2001.

[4] Busse, Laura, Kenneth C. Roberts, Roy E. Crist, D. H. Weissman, and Marty G. Woldorff (2005) "The spread of attention across modalities and space in a multisensory object." *Proc. Nat. Acad. Sci.* USA, **102** (**51**): 18751–18756. doi: 10.1073/pnas.0507704102

[5] Eimer, Martin and Erich Schröger (1998) "ERP effects of intermodal attention and cross-modal links in spatial attention." *Psychophysiology* **35** (**3**): 313–327. doi: 10.1017/s004857729897086x

[6] Lee, Christine (2018) "What We See is What We Desire to See for Color and Instruments: Color as an Inspiration for Musical Composition." Ph.D. thesis, University of California, USA. https://escholarship.org/uc/item/0b12k48x (Retrieved February '2024).

[7] Spence, Charles, Nicola Di Stefano (2022) "Coloured hearing, colour music, colour organs, and the search for perceptually meaningful correspondences between colour and sound." *Iperception* **13** (**3**) doi: 10.1177/20416695221092802

[8] Livingstone, Steven R., and Frank A. Russo (2018) "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." *PLoS ONE*, **13** (**5**): e0196391. doi: 10.1371/journal.pone.0196391

[9] Kaggle. Crema-d. https://www.kaggle.com/datasets/ejlok1/cremad (Retrieved February '2024).

[10] Pichora-Fuller, M. Kathleen, and Kate Dupuis (2020) "Toronto emotional speech set (TESS)" (Retrieved February '2024).

[11] Surrey audiovisual expressed emotion (Savee). https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee (Retrieved Feb. '2024).

[12] Kaggle. MIREX emotion dataset. https://www.kaggle.com/datasets/imsparsh/multimodal-mirex-emotion-dataset (Retrieved Feb. '2024).

[13] Caivano, Jose Luis (2015) "Color and sound: Physical and psychophysical relations." *Color Res. Appl*., **19** (**2**): 126–133. doi: 10.1111/J.1520-6378.1994.TB00072.X

[14] Zhou, HaiHong "The World of Music and Its Expression." Central Conservatory of Music Press, Beijing, China, 2008.

[15] Schloss, Karen B., Patrick Lawler, and Stephen E. Palmer (2008) "The color of music." *J. Vision*, **8** (**6**): 580. doi: 10.1167/8.6.580

[16] Palmer, Stephen E., Thomas Langlois, Tawny Tsang, Karen B. Schloss, and Daniel J. Levitin (2011) "Color, music, and emotion." *J. Vision*, **11** (**11**): 391. doi: 10.1167/11.11.391

[17] Griscom, William S., and Stephen E. Palmer (2012) "The color of musical sounds: Color associates of harmony and timbre in non-synesthetes." *J. Vision*, **12** (**9**): 74. doi: 10.1167/12.9.74

[18] Griscom, William S. (2015) "Visualizing sound: Cross-modal mapping between music and color." M.Sc. thesis, Gradworks, Regina, Canada.

[19] Liu, Jingyu, Anni Zhao, Shuang Wang, Yiyang Li, and Hui Ren (2021) "Research on the correlation between the timbre attributes of musical sound and visual color." *IEEE ACCESS* **9**: 97855–97877. doi: 10.1109/ACCESS.2021.3095197

[20] Aljanaki, Anna, Frans Wiering, and Veltkamp Remco C. (2016) "Studying emotion induced by music through a crowdsourcing game." *Information Processing & Management* **52** (**1**): 115-128. doi: 10.1016/j.ipm.2015.03.00.

[21] Panda, Renato, Ricardo Malheiro, Bruno Rocha, Antonio Oliveira, and Rui P. Paiva (2013) "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis." In: Proc. of 10th CMMR'2013, Marseille, France.

[22] CodePen. Survey template. https://codepen.io/webbarks/pen/QWjwWNV?editors=1111, 2023 (Retrieved Feb. '2024).

[23] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," 2019 IEEE International Conf. on Signals and Systems (ICSigSys), Bandung, Indonesia, 2019, 40-44, doi: 10.1109/ICSIGSYS.2019.8811080

[24] H. Li, X. Zhang, and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, 795-799, doi: 10.1109/ICSIP52628.2021.9689043

[25] Huang, Y., Wu, A., Zhang, G., Li, Y. (2014). Speech Emotion Recognition Based on Coiflet Wavelet Packet Cepstral Coefficients. In: Li, S., Liu, C., Wang, Y. (eds) Pattern Recognition. CCPR 2014. Communications in Computer and Information Science, 484. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-45643-9_46

[26] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier," in IEEE Access, 8, 96994-97006, 2020, doi: 10.1109/ACCESS.2020.2991811

[27] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, 3020-3024, doi: 10.1109/ICASSP39728.2021.9414286

[28] S. Han, F. Leng and Z. Jin, "Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, 803-807, doi: 10.1109/CISCE52179.2021.9445906