EMO-Music: Emotion Recognition based Music Therapy with Deep Learning on Physiological Signals

Hanzhe Guo
Department of Electronic Engineering，BNRist
Tsinghua University
Beijing, China
guohz20@mails.tsinghua.edu.cn

Jiawen Zhang
Department of Electronic Engineering
Tsinghua University
Beijing, China
jiawen-z20@mails.tsinghua.edu.cn

Yueyao Jiang
Academy of Arts & Design
Tsinghua University
Beijing, China
jyy20@mails.tsinghua.edu.cn

Yifei Qi
Academy of Arts & Design
Tsinghua University
Beijing, China
qyf18@mails.tsinghua.edu.cn

Simeng Chen
School of Environment
Tsinghua University
Beijing, China
chensm21@mails.tsinghua.edu.cn

Zhen Chen
FITC
Tsinghua University
Beijing, China
zhenchen@tsinghua.edu.cn

Weiran Lin
FITC
Tsinghua University
Beijing, China
linwr@tsinghua.edu.cn

Junwei Cao
BNRist
Tsinghua University
Beijing, China
jcao@tsinghua.edu.cn

Shuang Shou Li
FITC
Tsinghua University
Beijing, China
lss@tsinghua.edu.cn

***Abstract*— Music regulates emotions, alleviating negativity and promoting positive mental states. Music therapy is widely used, notably for autistic patients and college students' mental well-being. College students face high stress, poor sleep, and reluctance to seek support, making music therapy particularly suitable. This paper present EMO-Music, which uses a smartwatch to intervene positively by recognizing users' emotions and playing music accordingly. It aims to transform music apps into functional tools for enhanced therapy. User experiments using the SUS scale are conducted to evaluate the feasibility of our system.**

Keywords—Mental Health; Emotion Recognition; Music Therapy; Machine Learning; Deep learning; Transformer.

I. INTRODUCTION

Emotion, a crucial aspect of well-being, influences both physiology and psychology [1]. Models for quantifying emotions include discrete (e.g., Ekman's six fundamental emotions) and dimensional frameworks [1-3]. Emotion recognition relies on cues like facial expressions, vocal intonations, and physiological signals [1]. Yet, integrating these methods into daily life remains challenging due to the complexity of annotating physiological signals [19].

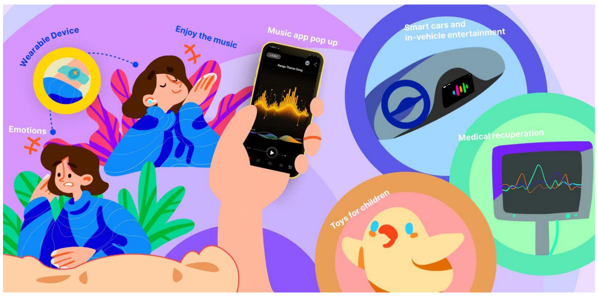

Fig. 1. An illustration ofthe daily use case ofEMO-Music, where a college student wearing a smartwatch is feeling bad. In this situation, EMO-Music is aroused and the app pops up, playing suitable music with permission of the user. Other situations including smart cars, medical recuperation and toys for children are on the way of EMO-Music's future development.

Music therapy, a holistic approach, aids various conditions in diverse settings [4]. It treats schizophrenia, depression, insomnia, neurosis, and pain management, yielding positive outcomes.

For music recommendation, various approaches exist: Gatta et al. used a hypergraph data model [20], Zhang et al. combined collaborative and content filtering [21]. Regarding emotional recognition in music, Chaturvedi, V. et al. reviewed methods but noted a lack of experiments on changing emotions.

Juan [5] studied emotional state detection using smartwatches, utilizing 17 features for Random Forest classification based on specific stimuli. However, this limited approach doesn't cover diverse real-life scenarios and might not be the most effective.

Yande Li [6] used sliding windows to enhance feature extraction for human performance detection from smartwatch data, employing SVM and KNN.

Lin Shu [7] classified emotions in movie-watching scenarios using KNN, RF, and DT, achieving high accuracy but facing similar limitations to Juan's study.

This paper presents EMO-Music, to enhance emotional detection and utilization in real-life scenarios. It merges emotion recognition with music therapy using deep neural networks to identify real-time emotional changes through physiological signals and recommend tailored music to enhance emotional states. Our contributions are threefold:

1) We conducted a comprehensive survey to substantiate the demand for emotion-based music recommendations, particularly among university students.

2) Leveraging advanced deep learning methodologies, , we applied Bidirectional Encoder Representations from Transformers (BERT) to a novel dataset of physiological signals. Remarkably, our results surpassed those achieved through traditional analytical approaches.

3) In practical application, we developed an application utilizing the Huixin wristwatch platform, which effectively identifies users' emotional states and subsequently offers

tailored music recommendations to positively impact their mood and well-being.

The rest of this article is organized as follows: The first chapter describes the design of music software, and introduces the design blueprint. The second chapter shows the designs of the emotion recognition scheme, which introduces the data set used in this study, the try of the identification model, and the analysis of the results. The third part introduces our preliminary user experiment and puts forward the conclusions and prospects of the current emotion recognition scheme. The last chapter is the conclusion of our whole paper.

II. Music App Design And application

A. *Music App Design*

We surveyed college students for insights into preferences and characteristics. 68 valid data points formed user personas based on music habits and app usage. These personas are crucial for designing our service process.

Our tailored questionnaire covered demographics and music habits. It provided an overview of user behavior patterns. Positive results confirmed strong music demand among college students.

Applying emotion recognition enhances music recommendations, aligning with real-time user needs. Divided into mobile, watch, and backend, our process begins by detecting mood changes on the watch, tailoring song suggestions based on user preferences. Users receive prompts for recommended songs and can play them on either device. A music visualization interface enhances the experience. User interactions influence the algorithm's accuracy, improving future recommendations.

Mobiles prioritize visual experiences with comprehensive functions, while watches focus on real-time data acquisition due to limited hardware. Integrating these elements creates a smoother user experience, utilizing music as an emotional tool. This transforms the app into an emotionally engaging virtual companion, fostering deeper connections and providing increased emotional value. Enhanced user activity and repeated use foster a dynamic user profile model.

Fig.2 shows our EMO-Music system Architecture.

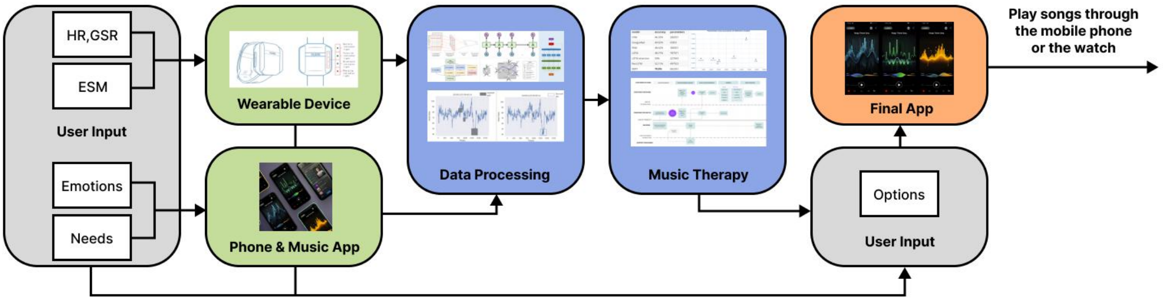

Fig. 2. EMO-Music Architecture. When the user needs music therapy and our app is on, raw data (HR, GSR) is captured from the smartwatch and is used in our deep learning model for emotion detection. Afterward, the emotion result is connected to our music library with certain label and an option ”your emotion is X, will you play music?” appears on the screen. Finally, if the user choose yes, songs will be played through the mobile phone or the watch.

B. *Current Music App Usage*

The usage process of the application can be outlined as follows steps:

1. Activate Bluetooth on your phone and pair it with the wristwatch using the "HUIXIN WRISTWATCH" app.

2. Open the "emo_music" app and navigate to its interface.

3. Click "Listen to the music" to retrieve wristwatch-recorded data for emotion recognition.

4. Confirm the generated emotional label ("Positive," "Calm," or "Negative") to open a relevant emotional state interface.

5. Enjoy emotionally healing music and save preferred songs to a collection via the heart-shaped button.

6. Use the "Flow Mode" button to easily return to the main interface when done listening.

7. Access the favorite music list from the main interface's heart-shaped button for viewing and playback.

8. Easily return to the main emotion recognition interface using the "Listen to Music" button across these interfaces.

III. Data and Models

The dataset, collected by Tsinghua University's Department of Psychology [8], integrates physiological data from Huixin wristwatches and psychological data via the Experience Sampling Method (ESM). It covers heart rate, exercise status, and skin electricity records (DAPPER), involving 100+ participants across 5 days, capturing diverse physiological data.

Participants provided dynamic emotional data through periodic 5-point scale questionnaires, assessing true self and state personality. Sentiment potency quantification served as the primary label for classification.

A. *Data Processing*

Label data processing: Emotion potency ratings (1-5) were reclassified into three categories: 1-2 as negative, 3 as neutral, and 4-5 as positive. Emotion quantification typically relies on discrete or dimensional models [22][23]. Utilizing valence from the dimensional model, emotions were categorized into positive, neutral, and negative.

Sample data processing: Physiological data samples before and after questionnaire completion (1-minute intervals, 120 samples) were used. Missing values were filled using mean imputation, potentially impacting model training. Data normalization (0-1 range) mitigated participant-based physiological variations. ECG signals, reflecting emotional changes, were utilized. Studies indicate a strong correlation between ECG waveform and emotional valence [25]. Wearable ECG devices, like wristwatches, provide reliable and minimally intrusive data, aiding emotion recognition using deep learning techniques

B. Dataset Division and Assessment

Data with significant defects were removed. Data with missing sample point data that accounted for 25% or more of the data were considered to have serious defects and were excluded. A total of 2080 data points met these requirements, with 1750 data points allocated for training and 330 for testing. A batch size of 5 was used, and data shuffling was employed during loading to prevent the trained model from overfitting to specific participants in the training set.

The dataset offers insights into real-life emotional physiology, but its limitations include a small size with 2080 usable data points, potentially restricting complex model training. Additionally, data loss affects 6.07% of samples and renders 10.1% unusable due to missing labels, while the use of questionnaires may introduce subjectivity in emotional analysis, impacting data accuracy

C. Random Forest based Emotion Recognition

Random Forest-based emotion recognition method employed manual feature extraction, generating 26 features, and used sklearn for classification. While achieving 73.3% accuracy in athree-category dataset from literature, it notably dropped to 39.67% in our dataset, possibly due to size constraints and inherent subjectivity in the test scale.

This approach's heavy reliance on manual feature extraction, tailored to a different dataset, led to decreased accuracy. These limitations emphasize the necessity for more adaptable feature extraction techniques in emotion recognition.

D. Deep learning-based Emotion Recognition

We turned to deep learning methods [10] for emotion recognition. We explored various deep learning models, including CNN[11], RNN[13], LSTM[14] and its variants [15][16], and BERT. Recognizing the similarity between the time series classification problem and text classification, we leveraged deep learning models that have shown promise in natural language processing (NLP) for comparison.

Table. 1. Accuracies and parameters of different DL models

Model	Accuracy	parameters
CNN	46.33%	282931
GoogLeNet	49.82%	65831
RNN	48.42%	306051
LSTM	48.77%	787971
LSTM-attention	50%	227843
ResLSTM	52.11%	497923
BERT	70.8%	662961

The experiment results of the models mentioned are listed in Table.1. and Fig. 5. Considering emotion detecting accuracy, BERT model reaches over 70%, far better than other traditional deep learning methods, and the scale of parameters is tolerable.

We adapted BERT for a time sequence task, modifying the word element mapping to physiological data, incorporating token and position embedding to improve signal position information correlation, and using mask recovery for missing data which is shown in Fig. 4. In NLP tasks, position embedding concerns the word position in the sentence, while in our task, the position embedding is based on the relative position point index of our selected physiological data. This method achieved promising results for emotion recognition.

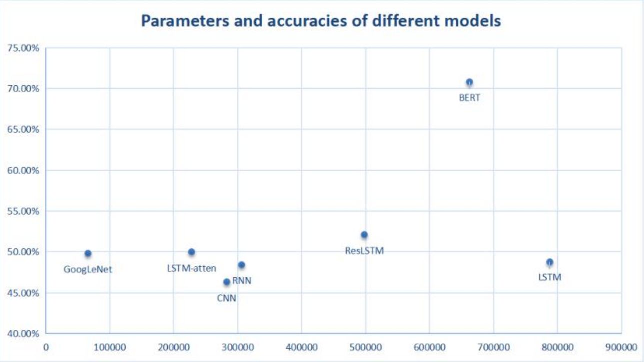

Fig. 3. Accuracies and parameters of different DL models.

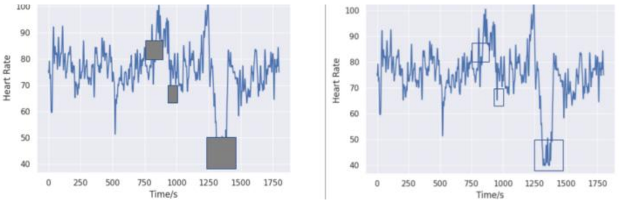

Fig. 4. BERT pretext: mask and recovery.

IV. User Experiment

1) We chose the SUS scale [18] as the prototype for this test because the SUS scale is based on extensive experimental design and is a globally widely used comprehensive feasibility evaluation scale. In this test, we had 36 participants. We aim to assess the usability and learnability of this design through the SUS scale, summarize the test results, and prepare for the next iteration. EMO-Music's experimental usages are shown in Fig. 5. In the test experiment, it includes heart rate data measuring, positive emotion result, negative emotion result and neutral emotion result.

1) SUS Results:

Overall, the SUS scores mainly ranged from 40 to 60, indicating that the usability of this design is at a medium level. Among them, 14 respondents scored above 50 in SUS, suggesting that they rated the product's usability higher than other participants.

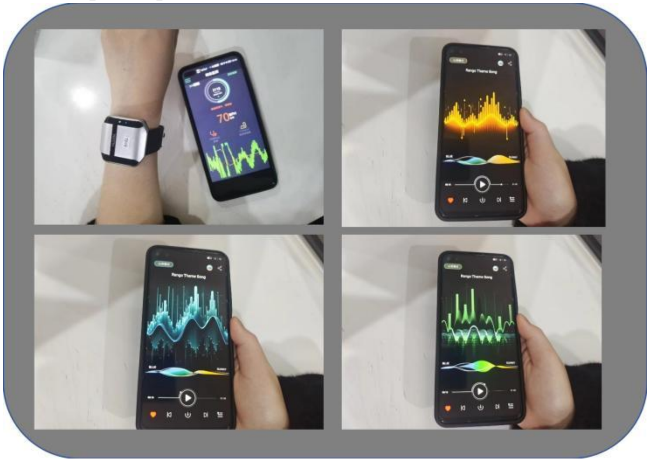

Fig. 5. EMO-Music’s physical usage experiments.

2) Usability Test Results:

In terms of usability, most respondents' scores ranged from 35 to 55. Among them, 8 respondents scored relatively high (above 50) in usability, believing that the product performed relatively well in terms of functional implementation, operational logic, and interactive design.

3) Learnability Test Results:

The scores for learnability are generally low. Since wristwatches are electronic products that people use less frequently, and the process of using hardware and software in tandem can be complex, there might be a learning curve for new users with this design.

In conclusion, we should consider optimizing the learnability aspect more in the future, adding clearer guidance, tutorials, or onboarding processes to help new users familiarize themselves with the product faster. Although the SUS test results were not high, looking at the results from question one, collecting human body data through wristwatch hardware and optimizing software usage in real-time through emotion recognition data still has a high demand.

V. Conclusion

In this paper, we preset EMO-Music, a novel music interaction software to meet the daily psychological therapy needs of college students. Drawing upon data from young college students, a specialized model was trained and integrated into the designed interaction system. In addition to conceptual design and model application, we have also conducted a small-scale preliminary user test, receiving relatively positive feedback. Our primary contribution lies in introducing a novel concept: utilizing physiological data for emotion recognition and subsequently providing services.

References

[1] Nie Dan, Wang Xiao Wei, Duan Ruonan & Lu Baoliang. (2012). Summary of EEG-based emotion recognition studies. Chinese Journal of Biomedical Engineering (04), 595-606.

[2] Zhang Guanhua, Yu Minjing, Chen Guo, Han Yiheng, Zhang Dan, Zhao Zhen & Liu Yongjin. (2019). Review of EEG features for emotion recognition. Chinese Science: Information Science (09), 1097-1118.

[3] Zhao Guozhen, Song Jinjing, GeYan, Liu Yongjin, YaoLin, WenTao. Advances in Emotion Recognition Based on Physiological Big Data[J]. Journal of Computer Research and Development, 2016, 53(1): 80-92.

[4] Shen Jing. (2003). Review of music therapy and its related psychological research. Psychological Science (01), 171-172. doi:10.16719/j.cnki.1671-6981.2003.01.059.

[5] Quiroz JC, Geangu E, Yong MH Emotion Recognition Using Smart Watch Sensor Data: Mixed-Design Study

[6] Li Y, Yu L, Liao J, Su G, Ammarah H, Liu L, Wang S. A single smartwatch-based segmentation approach in human activity recognition. Pervasive and Mobile Computing 2022;83:101600

[7] Shu, L., Yu, Y., Chen, W., Hua, H., Li, Q., Jin, J., & Xu, X. (2020). Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet. Sensors, 20(3), 718. https://doi.org/10.3390/s20030718

[8] Shui, X., Zhang, M., Li, Z. et al. A dataset of daily ambulatory psychological and physiological recording for emotion research. Sci Data 8, 161 (2021). https://doi.org/10.1038/s41597-021-00945-4

[9] Scikit-Learn,https://scikit-learn.org/.

[10] Goodfellow I., Bengio Y., Courville A.， Deep Learning, MIT Press, 2016. (http://www.deeplearningbook.org/)

[11] Chen, Y. (2015). Convolutional neural network for sentence classification (Master's thesis, University of Waterloo).

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[13] Jordan, Michael I. (May 1986). Serial order: a parallel distributed processing approach. Tech. rep. ICS 8604. San Diego, California: Institute for Cognitive Science, University of California.

[14] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

[15] Vaswani, A., Shazeer, N., Parmar,N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)

[17] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[18] James R. Lewis (2018) The System Usability Scale: Past, Present, and Future, International Journal of Human-Computer Interaction, 34:7, 577-590, DOI: 10.1080/10447318.2018.1455307.

[19] Chaturvedi, V., Kaur, A.B., Varshney, V. et al. Music mood and human emotion recognition based on physiological signals: a systematic review. Multimedia Systems 28, 21–44 (2022). https://doi.org/10.1007/s00530-021-00786-6.

[20] V. L. Gatta, V. Moscato, M. Pennone, M. Postiglione and G. Sperlí, "Music Recommendation via Hypergraph Embedding," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 10, pp. 7887-7899, Oct. 2023, doi: 10.1109/TNNLS.2022.3146968.

[21] R. Zhang, S. Tu and Z. Sun, "A hybrid music recommendation method based on music genes and collaborative filtering," 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Falerna, Italy, 2022, pp. 1-6, doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927924.

[22] Broek E L. Ubiquitous emotion-aware computing. Personal and Ubiquitous Computing, 2013, 17: 53-67

[23] J. O. N. A. T. H. A. N. POSNER, J. A. M. E. S. A. RUSSELL, and B. R. A. D. L. E. Y. S. PETERSON, “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology,” Development and Psychopathology, vol. 17, no. 3, pp. 715–734, 2005.

[24] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," in IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 374-393, 1 July-Sept. 2019.

[25] F. Agrafioti, D. Hatzinakos and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 102-115, Jan.-March 2012.