Music-to-Facial Expressions: Emotion-Based Music Visualization for the Hearing Impaired

Yubo Wang[1*], **Fengzhou Pan**[1*], **Danni Liu**[1*], **Jiaxiong Hu** [2]

[1] Washington University in St. Louis
[2] Tsinghua University
w.yubo@wustl.edu, panfengzhou@wustl.edu, danni.l@wustl.edu, hujx19@tsinghua.org.cn

Abstract

While music is made to convey messages and emotions, auditory music is not equally accessible to everyone. Music visualization is a common approach to augment the listening experiences of the hearing users and to provide music experiences for the hearing-impaired. In this paper, we present a music visualization system that can turn the input of a piece of music into a series of facial expressions representative of the continuously changing sentiments in the music. The resulting facial expressions, recorded as action units, can later animate a static virtual avatar to be emotive synchronously with the music.

Introduction

Music, as a sort of auditory stimulation, has for a long time been a privileged enjoyment for people with normal hearing. Music visualization, the process of graphically interpreting sounds, allows the hearing impaired to appreciate such a form of art. A successful music visualization tool can not only greatly extend the music audience group but create opportunities for novel vision-based music compositions.

Numerous attempts have been approaching music visualization from different aspects, such as frequency spectrum, colors, or 3D particles. Among the various approaches, we pay the most attention to the techniques that grasp meaning of music and convey intelligible information to the hearing-impaired group. Specifically, we focus on emotion-based music visualization, which provides exceptional advantages in resonating with the audience and enriching their music experiences (Chen et al. 2008).

In this paper, we study the history and the current status of music visualization and music emotion prediction, which provides a foundation and inspiration for our work. We have also included a pilot study proving the effectiveness of emotion-based music visualization in aiding music understanding. Then, we present our design methodology for an emotion-based visualization tool that can transfer a piece of music into a video of emotive facial expressions and introduce an implemented pipeline based on our design considerations. Finally, we mention aspects that future work can be done to improve our music visualization system.

[*]These authors contributed equally.

Related Work

Visualization has been proven effective in facilitating the understanding of music and bridging the gap in music accessibility. The technique for music visualization remains an open problem in terms of development and analysis. There are different attempts made to combine visual and auditory media. Prior works ranged from music notation to making use of special graphics. However, visualizing music in an effective, meaningful, and intuitive way is challenging.

Music Visualization

Music visualization can be traced back to Common Music Notation (CMN). Though much music analysis is derived from music notation, the goal of CMN is to assist experienced musicians with music performance (Isaacson 2005). Many users are unfamiliar with CMN due to its complex nature (Smith and Williams 1997).

There are two major approaches to augmenting the learning and listening experience of non-expert users. Some efforts have been made to design tangible devices. Model Human Cochlea is a vibrotactile display designed to help users with limited or no hearing ability access the emotional information in music (Karam et al. 2009). MuSS-Bits wearable sensor-display pairs can provide visual and vibrotactile feedback for deaf people to explore customizable music experience (Petry, Illandara, and Nanayakkara 2016).

The other approach is through the use of a visual interface. Commercial music players are often featured with music visualizations that simulate graphic patterns matching the given music track's tempo, strength, pitch, mood, etc. The iTunes visualizer, for example, provides visualizations based on waveform analysis of the given music. However, prior study suggests that though such visualization is aesthetically pleasing, the selection of visual parameters can be arbitrary, making it less informative and meaningful to the hearing impaired (Fourney and Fels 2009).

Music Emotion Prediction

Music has been connected to emotions for a considerable history. With the advances in computer graphics, there is an increasing number of attempts to address communicating emotions. Joyce Horn Fonteles designed a 3D particle visualization system that generates real-time animated particle emitter fountains and measured users' opinions about

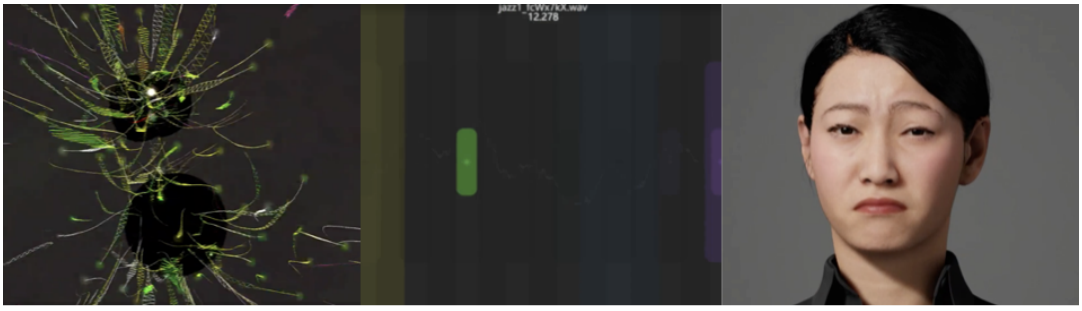

Figure 1: The three music visualization techniques, iTunes, ViTune, and MetaHuman, applied in the pilot study and corresponding results. iTunes is a popular commercial music software that contains a music visualizer. ViTune is designed to convey emotional information in music. Both iTunes and ViTune utilize abstract graphics to visualize some features of the given music. MetaHuman can capture human facial expressions in real time and generate videos of high-fidelity digital human figures expressing the same facial expressions.

the mood expressiveness of the music (Fonteles, Rodrigues, and Basso 2013). Vitune (Deja et al. 2020) evaluated its visualization attributes based on the correlation between the attributes, and the emotions participants felt during certain points in the music.

A number of works have been done for annotating and analyzing emotions from music pieces based on acoustic features. The DEAM dataset (Soleymani, Aljanaki, and Yang 2016), for example, consists of 1802 soundtracks annotated with valence and arousal values, which are essential parameters for analyzing emotions in the dimensional model (Russell 1980). These works give us insights and reference about how to construct our prototype.

Pilot Study

Remarkable achievements have been made in music visualization and music emotion prediction. However, development in these two regions does not necessarily overlap - that is, not all music visualization techniques take emotion into account. Given that emotion is a key factor in music experience that can hardly be ignored, we consider that a good visualizer should be able to convey emotion adequately. In the pilot study, we aimed to compare different visualization methods and tried to find out what kinds of visualization can most convey emotion in music pieces.

We recruited 6 participants for the study. Five of them (5/6) are with normal hearing, and one of them (1/6) is hearing impaired.

We included six music pieces in the study. We selected three different music genres - jazz, R&B, and rock - and for each genre, we picked one music piece with positive emotion and one with negative. We applied three visualization tools to transform each music piece into three corresponding videos - one with a light effect show generated by iTunes visualizer, one with colorful rectangle bars generated by ViTune (Deja et al. 2020), and one with an avatar making continuously changing facial expressions generated by Unreal Engine and MetaHuman. For the first two types of visualization, we generated the videos by directly inputting the music to the visualizers. For the last type, we generated the videos by asking a person to sing and react to the given music, recording his facial expression changes, and generating the avatar version of the reactions in MetaHuman.

The music audio and the visualized music videos were presented in different orders among the six participants to counterbalance the sequence effect. We applied the Valence-Arousal model (Russell 1980) to document the emotional changes of the participants. For each video, while the clip was playing, the participants were asked to rate the arousal and valence levels in real-time on a web page we built. The participants could keep the previous values if they believed the emotion holds and changed them whenever they identify a difference. After watching all three visualization videos, participants with normal hearing were also asked to listen to the music and conduct another round of rating based purely on the audio. The ratings of participants when listening to the music audio are considered to be the baseline indicators of the emotion in the music.

Figure 2 shows the participant ratings for all six music pieces. Generally, MetaHuman ratings seem to approach baseline the best, which might suggest that participants successfully retrieve the greatest amount of emotion-related information while watching MetaHuman videos. To quantitatively analyze the data, for each visualization of each music piece, a Pearson correlation coefficient between user ratings on the visualization video and baseline ratings was computed. If the correlation is positively high, the visualization is believed to be effective in conveying the emotion of the music and thus can assist music appreciation; if the correlation is close to zero or even negative, the visualization is believed to be unhelpful in supporting music appreciation or, in the worst case, interfere with music emotion understanding. As shown in Figure 3 and Table 1, MetaHuman stands out with a comparably high positive correlation with baseline across almost all music pieces for both arousal and valence, highlighting the advantage of human facial ex-

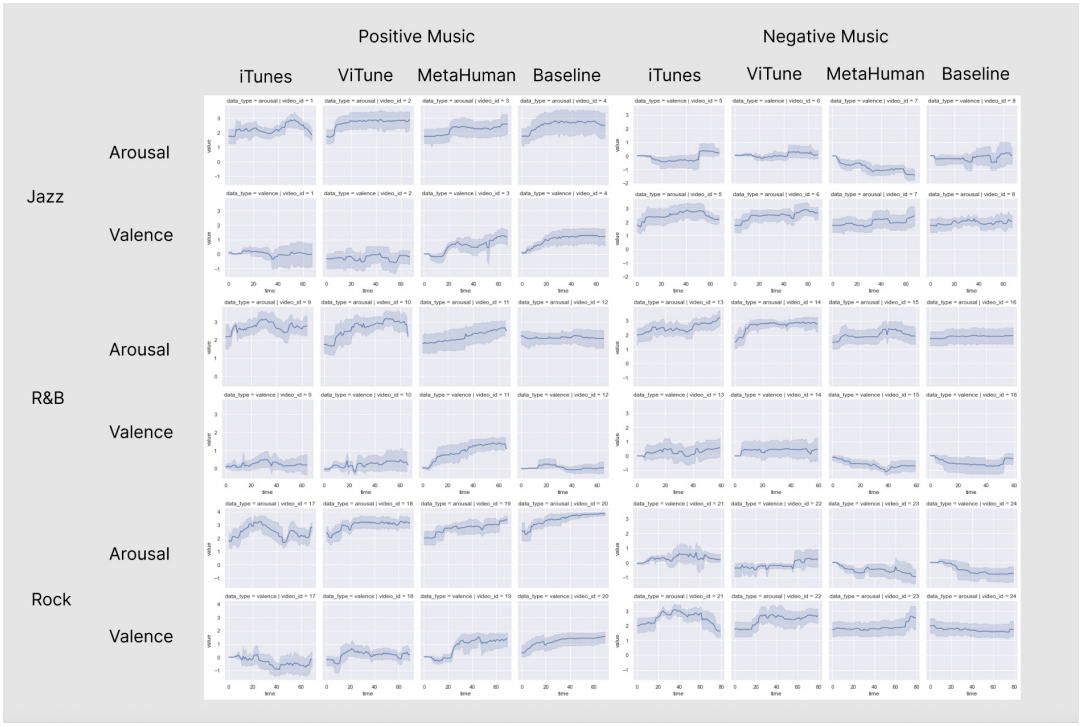

Figure 2: The averaged user ratings of arousal and valence to each piece of presented visualization video and music audio. There are in total six subplots corresponding to six music pieces in the study, with music types of Jazz, R&B, and Rock, and emotions of positive and negative. For each of the six subplots, we show graphs of user ratings across three different visualization videos generated by iTunes, ViTune, and MetaHuman, and the baseline music audio. In the small graph, the horizontal axis represents the change of time in the video or audio from start to end, and the vertical axis represents the ratings. Generally, MetaHuman seems to show the most similar trend of ratings to the baseline.

pressions in facilitating music emotion comprehension. The results strongly encouraged us to rely on the approach of music-to-facial expressions for emotion-based music visualization.

Although MetaHuman was proved in the pilot study to be an outstanding emotion-based visualizer, its limitation is also apparent - the input of MetaHuman must be human reactions. That is to say, to generate music visualization for each new piece of music, a human actor is needed to stay in front of the camera, listen and react to the music, and get their reaction recorded and transferred to a MetaHuman avatar. The whole process of such music visualization is time-consuming and unintelligent. To overcome the obstacle, we would like to design a visualization system that can automatically learn how to extract emotion-related information from music pieces and then apply the extracted information to perform visualization without the need for human manipulation. Therefore, we seek to explore ways to overcome the obstacle of MetaHuman in the following discussion.

Methodology

There are two main questions involving emotion-based music visualization: how to represent emotion and how to visualize emotion. For the first question, in the following paragraphs, we introduce two different emotion representation approaches - arousal-valence tuple representation and action unit representation. For the second question, we come up with corresponding visualization techniques for each emotion representation approach.

Arousal-Valence Tuple Emotion Representation

A classical way to attribute emotions numerically and continuously is through dimensionalization, commonly rating each type of emotion along the valence (the positivity or negativity of the emotion) dimension and arousal dimension (the intensity of the emotion) and placing the emotion as a point over the arousal-valence coordinate. The arousal-valence decomposition of emotions serves as an effective intermediate step that bridges music and visualization, allowing us to first extract from music crucial emotion arousal-valence labels and then generate images accordingly.

Various methods have been proven to process music for meaningful outcomes successfully. Considering music as audio signals that carry features such as intensity and pitch, (Eyben, Wöllmer, and Schuller 2010) has provided a solution, openSMILE, to emotion recognition by extracting and synthesizing signal information from music. Based on openSMILE, we have implemented a model that takes in a piece of music, segments it with a 0.5-second window,

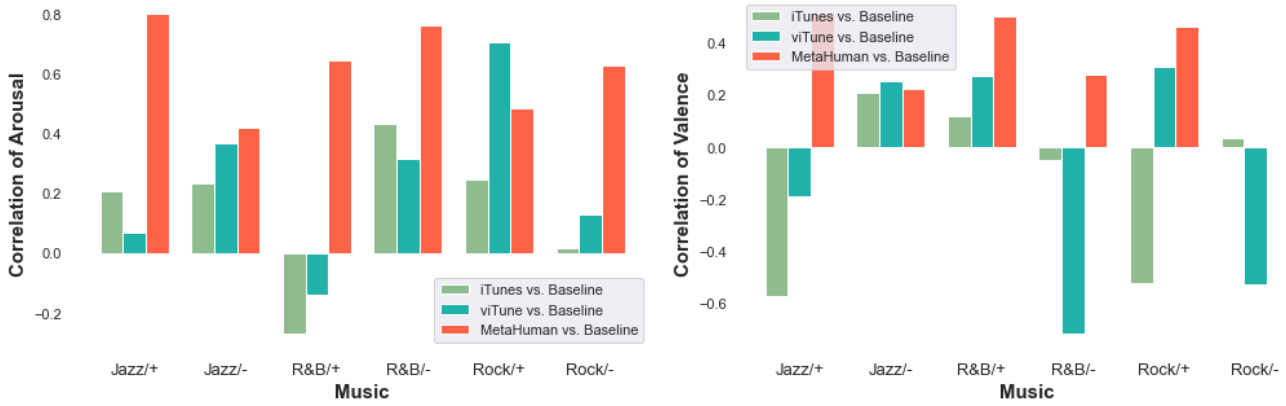

Figure 3: The correlation coefficient values between participant ratings for each visualization and the baseline ratings (participant ratings for the audio) on both arousal (the left graph) and valence (the right graph). The horizontal axis shows all six music pieces in the study across three music types Jazz, R&B, and Rock, and two emotion types, positive and negative. The vertical axis shows the correlation values of baseline ratings with iTunes ratings, ViTune ratings, and MetaHuman ratings respectively. Participant ratings on MetaHuman are generally most positively correlated with baseline ratings.

	iTunes vs. Baseline	ViTune vs. Baseline	MetaHuman vs. Baseline
Arousal	0.14 (SD = 0.242)	0.24 (SD = 0.293)	0.62 (SD = 0.150)
Valence	-0.13 (SD = 0.335)	-0.10 (SD = 0.448)	0.33 (SD = 0.203)

Table 1: The averaged correlation coefficient values between participant ratings for each visualization and the baseline ratings (participant ratings for the audio) on both arousal and valence. The correlation between Metahuman and baseline is the most positive.

and predicts the emotion of each 0.5-second duration in the music. Besides, audio spectrum patterns, even though not a sensible visualization to human beings, have been proven to be significant to the machine and serve as another effective predictor of music emotion (Brotzer, Mosqueda, and Gorro 2019). Combining audio patterns with the extracted features from openSMILE might improve the performance of the whole system in understanding human feelings.

Transforming emotion labels we retrieve from the last step to visualized images is less complicated, but it also deserves careful consideration. Since the very original purpose of inventing the arousal-valence coordinate is to classify diverse types of emotion, it is easy to follow such ideology and represent each arousal-valence tuple with its corresponding emotive facial expression, specifically the six basic emotions - anger, disgust, fear, happiness, sadness, and surprise. For each emotion, we pick five facial images that strongly express such emotion and randomly choose one to show each time the song is playing a snippet with that emotion. However, such oversimplification of the arousal-valence coordinate fails to distinguish nuances or discern minor emotions in music, so we seek to find an advanced visualization method that performs a finer division on the arousal-valence coordinate. Existing study in emotion avatars has been working on placing over thirty emojis, far beyond the six basic emotions, onto the arousal-valence coordinate (Jaeger et al. 2019). By selecting the closest emotion avatar for each point on the coordinate, we are able to assign each music snippet a human-like expression.

Predicting emotions in the form of arousal-valence tuples from music and generating a visualization of the given emotion labels produces some ideal results, but there are mainly two problems. Firstly, the attempt to map an arbitrary pair of values in the continuous 2-dimensional coordinate system of arousal and valence to the range of a harshly limited set of discrete visual representations hugely reduces the variety of visualization output styles, weakening the expressiveness of specific details and the diversification of music particularity. Secondly, the series of emotion labels are fragmented and loosely related to one another, and it is hard for the arousal-valence methodology to connect these emotion labels into sensible video streaming smoothly. Thinking of the process of one facial expression changing to the next facial expression as the movement of one point on the arousal-valence coordinate system to another position, it seems unreasonable to draw out each of the facial expressions along the straight line through these two points to show the transition.

Action Unit (AU) Emotion Representation

In order to deal with the two obstacles in the visualization process based on arousal-valence emotion labels discussed above, a different method is adopted, which makes use of the facial action coding system as a new way to represent emotions. Facial action coding is an anatomical system that divides the whole-face expressions into individual muscle movements (action units or AUs) such as brow raiser and

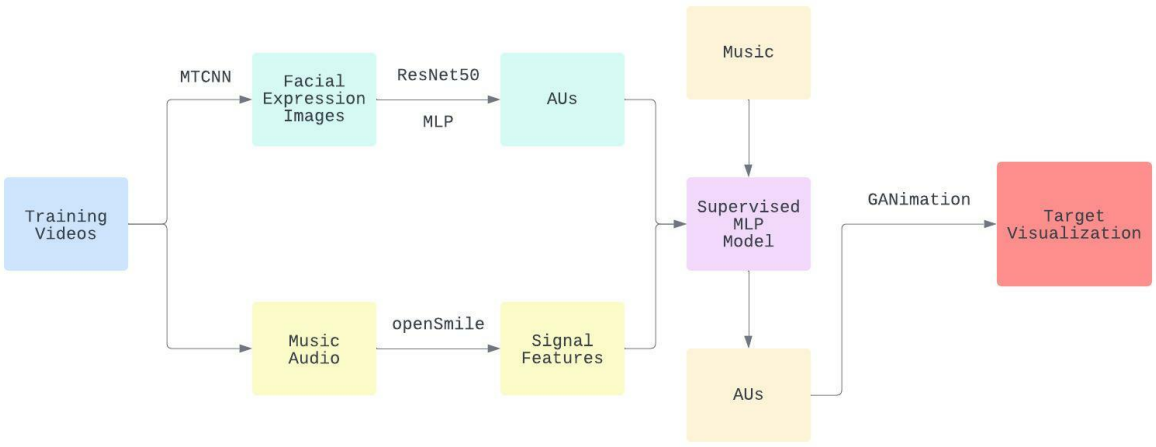

Figure 4: The overall flowchart of the pipeline. The pipeline starts by breaking down training videos into facial expression images and music audio and then extracts features respectively from the two. The resulting AUs and signal features from the last step are used to train an MLP model for music-to-AU mapping prediction. Finally, GANimation transforms the AUS generated from music input into the MLP model into visualization.

lip stretcher (Ekman and Friesen 1978). Later, this system is applied by GANimation to generate various expressions by manipulating the degrees of activation of a subset of AUs on a human face (Pumarola et al. 2018).

The new approach of using AU to represent emotions overcomes the two major shortcomings of the previous arousal-valence representation: restrictiveness and disconnectedness. Regarding restrictiveness, the set of over 40 AU variables and over 7000 possible AU combinations enables the generation of almost all possible human facial expressions, which can greatly enrich the content of music visualization; regarding disconnectedness, unlike arousal-valence tuples that work on representing individual static facial expressions but fail on representing the shifts between expressions, manipulating numeric values of AUs can easily create the effect fading changes of faces to capture the in-between transitions. Besides, another significant advantage of emotions being labeled in AUs is that the results we receive from the music emotion prediction phase can be directly transformed into visual works, enabling us to omit an unnecessary intermediate step of turning non-visualizable emotion labels into visualizable representations.

Based on the new approach, we manage to generate visualization by learning human facial expressions from music videos, which we assume should present certain emotive features of the corresponding music. We have selected a dataset with YouTube videos involving human singers or listeners with clear facial expressions who perform or respond to pieces of music. For each video, screenshot images of facial expressions are sliced out, and each image is linked to its timestamp in its original music. Then, an AU detector is applied to extract features from facial expressions in the images and produce a series of AU values in a time sequence. These AUs, together with the pure music soundtracks separated from the original videos, are trained in a random forest model aimed at predicting AUs from any given music. Once we input a new piece of music into the model and receive the resulting AUs, GANimation is to control the activation magnitudes of multiple AUs in a given human face picture with mild facial expressions and generates serialized emotive facial expression frames that are later turned into a video that matches the original music emotions. Lastly, we conduct video frame interpolation that synthesizes additional frames in between any two neighboring frames to smooth the produced video.

Design and Implementation

We introduce a pipeline to generate a video based on a piece of audio and a face image. The pipeline includes three parts: learning AU representations from labeled facial images, learning the mapping from audio to AU representations, and generating facial expressions based on AU representations.

Learning AU Representations from Facial Images

Based on the existing study (Deng, Chen, and Shi 2020), we were able to learn a mapping from video frames to AU representations. To solve the issue of missing labels, the researchers use teacher-student networks. The teacher and student networks share the same structure and size, but the teacher network is exposed to an incomplete set of labels, and the student network is exposed to an imperfect complete set of labels. For our task, we only used part of their work which only includes the regression branch for predicting AUs. The network adopts a ResNet50 backbone and an MLP for regression for AUs.

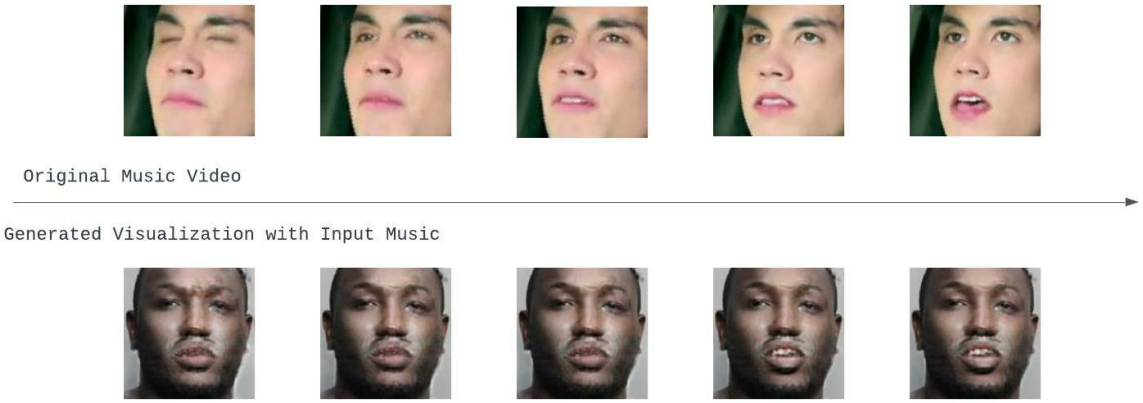

Figure 5: The comparison between facial expressions from a music video and the produced visualization given the same music as input. The above series of pictures are sliced out from a video of a YouTube singer. The series above are generated out of a randomly selected base picture with a human face.

Learning Mapping from Audio to AU Representations

To learn the mapping from audio to AUs, we first used openSMILE to extract the signal level features from audio. We adopted the 0.5-second window to generate statistical features from the audio. By selecting suitable step lengths between two windows, we generated the AUs with the same frequency and frame rate in the final video. In this step, we used an MLP with skip layers that directly learned the mapping from audio features to AUs of the current frame in a supervised manner.

Data. We selected a few YouTube videos involving human singers or listeners with clear facial expressions performing or reacting to the music.

Training. We first used the pre-trained model from (Deng, Chen, and Shi 2020) and GANimation to reduce the difficulty of training. Then we used MTCNN as a face detector that extracts faces from the video. After that, we adopted the model from "Multitask Emotion Recognition with Incomplete Labels" to generate a sequence of AUs for further use and then openSMILE to generate a sequence of audio features with the same sample rate as the sequence of AUs. After having the sequence of audio features and AUs, we trained the MLP to learn the mapping between them.

Inferencing. The input of inferencing was a piece of audio and a face image. We first used openSMILE to generate a sequence of audio features and the MLP to estimate the sequence of AUs for the audio. Then, we used GANimation to generate a sequence of video frames based on the given image and the sequence of AUs.

Generating Facial Expressions Based on AU Representations

Based on GANimation, we were able to generate an altered image with a new facial expression based on a set of AUs and the given image. GANimation is a generative adversarial network that is divided into three parts: a generator, a critic to evaluate the quality of the generated image, and an expression estimator to penalize differences between the desired AU expression and its fulfillment. In our pipeline, we only used the generator to generate the image based on AUs.

Results

Based on our introduced pipeline, we generated a primitive version of emotion-based music visualization. In Figure 5, we present a demonstration snapshot of our visualization of the song "Someone Like You", with the above row showing the original training music video of a human performer singing the song with his face zoomed in and the below row showing the corresponding facial expressions on a randomly-picked face template generated from the input audio file of the song. We can see that there are some facial expression features, such as the frown, are captured by the generated images, suggesting that a sense of melancholy that matches the general tone of the song is also captured. However, due to the incompleteness of successfully-learned features, the generated faces are relatively ambiguous and are not sufficiently strong indicators of emotion. Further refinement can be done for this system to produce satisfactory facial expressions along with the song that highlights the key emotion.

Conclusion and Future Works

This paper introduces a workable emotion-based music visualization pipeline for transforming music into facial expressions. The generated video of facial expressions serves to improve the music experiences of the hearing impaired.

For our next step, we hope to work on revising our visualization system to produce more accurate and advanced emotion-based music visualization. One of the biggest concerns of our current system is the lack of effective metrics for evaluating the performance of visualization. Without metrics, it is hard to determine the direction of model tuning. We think of calculating RMSE between the AUs extracted from

the music video and the AUs generated by the system before being sent into GANimation to be a way of accurate measurement. However, one consideration is that the importance of AUs might vary from each other, but in the calculation of RMSE, all AUs are considered to be equally contributive to the generation of each emotion. For instance, a 0.1 error in AU2 (outer brow raiser) can be less destructive compared to a 0.1 error in AU12 (lip corner puller) because, generally speaking, the lip corner direction carries more straightforward information to speak for emotion. Besides, the same value of difference at different range levels might also carry different significance - a 0.1 error from 0.5 to 0.4 of AU12 (lip corner puller) can turn a smiling mouth into an apathetic one, while a 0.1 error from 0.9 to 0.8 might only weaken the intensity of happiness but do not reverse the emotion. Therefore, we look for more explanatory metrics being applied to our system to assist our model training and system working.

We would also like to conduct formal user studies in the future to quantitatively and qualitatively validate our visualization efficacy. The pilot study results demonstrate the potential of using human facial expressions to augment and convey emotions in music pieces. While we validate our idea before implementing the visualization, the cost is we need to gain more knowledge about if our visualization is different from the MetaHuman visualization we used in the pilot study and exactly how accurate and effective our visualizations are. A more comprehensive effectiveness study and analysis between our prototype and the state-of-the-art music visualizations can be conducted to understand how our prototype is against them and how our model can be improved. We hope our approach can facilitate future work in the research field of music visualization, in particular, visualizing emotional information in music

Acknowledgements

We wish to thank Team Aparecium for helping us design and set up the pilot study and all of our participants for taking the time to complete this study.

References

Brotzer, J. M.; Mosqueda, E. R.; and Gorro, K. 2019. Predicting emotion in music through audio pattern analysis. *IOP Conference Series: Materials Science and Engineering*, 482: 012021.

Chen, C.-H.; Weng, M.-F.; Jeng, S.-K.; and Chuang, Y.-Y. 2008. Emotion-based music visualization using photos. In *International Conference on Multimedia Modeling*, 358–368. Springer.

Deja, J. A.; Dela Torre, A.; Lee, H. J.; Ciriaco IV, J. F.; and Eroles, C. M. 2020. Vitune: A visualizer tool to allow the deaf and hard of hearing to see music with their eyes. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Deng, D.; Chen, Z.; and Shi, B. E. 2020. Multitask emotion recognition with incomplete labels. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *PsycTESTS Dataset*.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the international conference on Multimedia - MM '10*.

Fonteles, J. H.; Rodrigues, M. A. F.; and Basso, V. E. D. 2013. Creating and evaluating a particle system for music visualization. *Journal of Visual Languages & Computing*, 24(6): 472–482.

Fourney, D. W.; and Fels, D. I. 2009. Creating access to music through visualization. In *2009 ieee toronto international conference science and technology for humanity (tic-sth)*, 939–944. IEEE.

Isaacson, E. J. 2005. What You See Is What You Get: on Visualizing Music. In *ISMIR*, 389–395.

Jaeger, S. R.; Roigard, C. M.; Jin, D.; Vidal, L.; and Ares, G. 2019. Valence, arousal and sentiment meanings of 33 facial emoji: Insights for the use of emoji in consumer research. *Food Research International*, 119: 895–907.

Karam, M.; Nespoli, G.; Russo, F.; and Fels, D. I. 2009. Modelling perceptual elements of music in a vibrotactile display for deaf users: A field study. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, 249–254. IEEE.

Petry, B.; Illandara, T.; and Nanayakkara, S. 2016. MuSSbits: sensor-display blocks for deaf people to explore musical sounds. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, 72–80.

Pumarola, A.; Agudo, A.; Martinez, A.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. GANimation: Anatomically-aware Facial Animation from a Single Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

Smith, S. M.; and Williams, G. N. 1997. A visualization of music. In *Proceedings. Visualization'97 (Cat. No. 97CB36155)*, 499–503. IEEE.

Soleymani, M.; Aljanaki, A.; and Yang, Y. 2016. DEAM: MediaEval database for emotional analysis in Music. https://cvml.unige.ch/databases/DEAM/. Accessed: 2022-09-02.