Emotional Voice Game Controller

MARJORIE ANN M. CUERDO* and REBECCA LIETZ*, University of California, Santa Cruz, USA

As technology progresses, novel forms of interaction arise. Video games in particular are highly interactive forms of media that are now going beyond traditional handheld controllers to seeing more immersive forms of player input with more embodied technologies through extended reality (XR). One possible unexpected alternative game controller might be player emotion. We aim to explore the question of whether integrating a player's emotions can enrich and/or better facilitate the game experience by developing a game system which responds to the emotional expression in the player's voice. Using machine learning techniques, our system detects emotion in a player's voice, and treats this as video game input. In other words, we created an emotional voice game controller.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *Sound-based input / output*.

Additional Key Words and Phrases: emotions, neural networks, interaction, audio input, games

ACM Reference Format:
Marjorie Ann M. Cuerdo and Rebecca Lietz. 2022. Emotional Voice Game Controller. In *CHI '22: ACM CHI Conference on Human Factors in Computing Systems, Apr 30–May 06, 2022, New Orleans, LA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Over the last few decades, technological progress has allowed for the development of highly immersive digital experiences for users. One of the most popular and effective forms of immersive technology are video games, which have been established as highly immersive forms of interaction for decades. Audiovisual stimulation in combination with an interactive narrative induces high levels of engagement and identification, which have been shown to aid in learning, emotion regulation, and even pain management [14]. State-of-the-art systems using extended reality (XR) approaches, such as virtual reality (VR) and augmented reality (AR), can put the user in a flow state by engaging multiple senses and providing alternative input methods that open up new possibilities for user engagement. For example, game systems may track user engagement and change the game content when the player gets bored or frustrated [15]. In recent years, a growing number of studies have explored affective gaming, or the consideration of player emotion when interacting with a game. However, most of those studies present games that are either *about* emotions, or that *react to* emotion but do not necessarily prompt the player to reflect on the emotions they are displaying. Our team has taken the concept of affective gaming and developed a game that requires the player to express certain emotions using their voice and responds accordingly, thereby motivating reflection in the player.

*Both authors contributed equally to this research.

2 RELATED WORK

2.1 Role of emotions

Emotions have a large impact on various aspects of our lives. Research has linked emotional states to changes in memory [2], attention [4], and a range of behaviors, such as eating [10] and driving [16]. Based on these findings, it makes sense to view emotions as highly influential in how we experience life. A long history of emotion research has led to the development of numerous emotion models, which generally fall into one of two categories: dimensional and discrete. Dimensional models, such as the Circumplex Model of Affect [1], can describe a wide range of emotions along multiple axes and take into account small variations in a person's mood. While this framework gives a more realistic view of how people actually experience emotion, it can also be helpful to sort emotional states into discrete categories. The most basic model includes six core emotions: anger, sadness, disgust, happiness, surprise, and fear [3]. Discrete emotion models are particularly helpful in projects like ours, where emotion detection is viewed as a classification problem and machine learning techniques are employed to solve it.

2.2 Emotion detection

Since emotions are internal states and inherently difficult to measure and detect, researchers have experimented with numerous ways to identify emotions using external cues, such as facial expressions, speech, physiological data, and behaviors [18].

Behaviors such as fidgeting, mobile phone usage, or interaction on social media, can be indicative of certain emotional states [18]. Although this type of data is relatively easy to obtain, it is often not clear whether the measured behaviors are valid indicators of the emotion they are supposed to represent.

Conversely, facial expressions more explicitly show a person's emotional state. In addition, the only equipment needed to collect facial data is a camera, which is already built into many personal computers. However, facial expressions are easy to fake or hide, which can make it difficult to accurately assess emotions. This method also heavily relies on the availability of suitable data sets in order to train a model reliably.

Physiological responses are least likely to be intentionally modified by a person, and are therefore highly reliable measures for arousal. On the downside, measuring bodily responses usually requires additional sensors and do not account for context, which makes it challenging to label nuanced emotions.

Lastly, people also express emotion through speech, using different intonations, pitch, and frequency. Similar to facial expressions, speech data can be easily collected using microphones built into a user's personal device. The main disadvantage of this approach is that it, like the facial expression method, relies on the existence of suitable data sets to train a machine learning model. Additionally, recordings from built-in microphones might be noisy and hard for the system to process. Yet, systems that use speech input for emotion recognition enjoy popularity within the scientific and commercial HCI community [13]. Khalil et al. [17] outline the need for speech emotion recognition to employ more non-linear deep learning techniques as opposed to the more traditional techniques utilizing KNN, HMM, and SVM classifiers, which we attempt to realize with our system.

2.3 Emotion in video games

As highly interactive media, games have the ability to evoke a diverse range of specific emotional experiences for players. Game designers manipulate Mechanics-Dynamics-Aesthetics (MDA) [5] to elicit particular types of player experiences according to their vision for their game. Regardless, many methods of game performance evaluation simply rely on achieving a general state of flow [7]. Games are useful for studying emotions, as they can be broken down into components. Players' emotions can be influenced through anything from the game's narrative, aesthetic representation (e.g. colors, imagery, sound, etc.), presented challenges, to even sociocultural context.

In addition to inducing flow, people often express that games should be fun. Lazzaro claimed that there are different types of "fun" that should be differentiated: Easy Fun, Serious Fun, Hard Fun, and People Fun [8]. These types of fun correspond to specific types of human emotions. Easy Fun is more relaxed and focused on feelings of curiosity, wonder, and awe. Serious Fun is a bit more repetitively involved and evokes excitement and zen focus. Hard Fun involves intense fiero – the feeling of triumphing over hardship or experiencing relief after frustration. People Fun occurs with other people, experiencing amusement and admiration.

While categorization can be useful for design, it is becoming more apparent over time that emotions in games are more complex than initially assumed. Bopp et al. [12] found that when reflecting on emotionally-moving game experiences, players most enjoyed and appreciated negatively valenced emotions, counter to what one may initially expect. It can no longer be claimed that a positive player experience is solely depending on experiencing purely positive emotions. Emotions play a substantial role in learning outcomes for serious games [9], so the need to further understand emotions in games goes beyond entertainment purposes.

2.4 Affect-adaptive gaming

Sundstrom [6] describes an affective loop as "an interaction process where:

(1) the user first expresses her emotions through some physical interaction involving the body, for example, through gestures or manipulations of an artifact
(2) the system (or another user through the system) then responds through generating affective expression, using for example, colors, animations, and haptics
(3) this in turn affects the user (both mind and body) making the user respond and step-by-step feel more and more involved with the system"

Keeping those requirements in mind, it is easy to see how games are a natural way for an affective loop to be realized [11]. When people are playing a game, they are required to constantly provide input and receive output (visual, auditory, and/or haptic feedback) from the system. Input to this loop can vary from behavioral data (game metrics/telemetry) to objective data (bodily responses). Games can acknowledge that input by emotionally adapting the experience to the player's assumed emotional state. In turn, emotional adaptation in games can deepen a player's engagement. Emotions can be elicited by manipulating game characters (agents and NPCs) and game content (rules, rewards, levels, narrative, and music) [11].

In our research, we recognize the importance of diverse emotional experiences for the player in order to increase enjoyment and immersion of a game. Drawing from the cited literature, our system demands various emotional responses from the player while rewarding them with a sense of calmness and achievement, leading to a (hopefully) stimulating game experience.

3 METHODS

We were interested in exploring the potential of voice-based player input towards affect-adaptive gaming. Voice-based forms of game interaction are still underexplored despite their affective and accessible qualities.

Our process to develop an emotional voice game controller involved the following steps:

(1) Pre-processing the audio data
(2) Training the neural network
(3) Creating the API
(4) Developing the game and connecting to API

3.1 Pre-processing the audio data

We used the speech audio files from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This data set contains 1464 audio files from 24 professional voice actors either saying "Kids are talking by the door" or "Dogs are sitting by the door" with any of the following eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.

Due to lack of appropriate hardware, we turned to Google Colaboratory to access GPU power. To be used in a neural network, the voice audio files were converted into spectrogram images using the librosa library. After running the code in this Colab notebook, the spectrogram images were sorted into the respectively labeled folders. We then downloaded the sorted data folder locally and uploaded it to a Google Drive folder for Colab to access later.

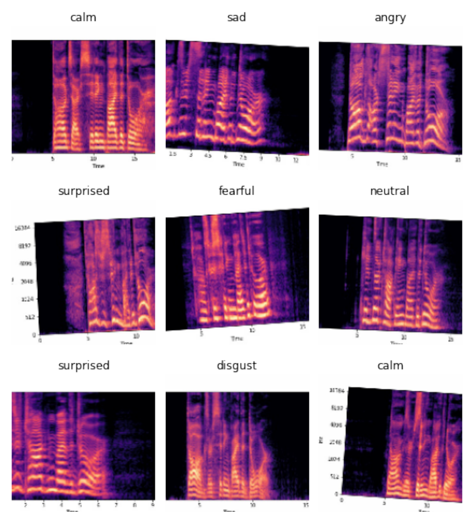

Fig. 1. Examples of spectrograms categorized by emotion.

3.2 Training the neural network

We used the following Colab notebook to perform transfer learning from ResNet-34 to our own neural network. This method required the use of the *fastai* library and ResNet-34, a pretrained Convolutional Neural Network (CNN) model trained on the ImageNet dataset used for image recognition. Our goal was to train a model that moderately detects and classifies the correct emotion in any voice recording regardless of the content of speech.

We used the *fit_one_cycle()* function to train our model with the data obtained from RAVDESS. Each cycle had five epochs each and we modified the learning rate through reading and interpreting the *lr_find()* graph (example depicted in Figure 2). We selected the learning rate range with the steepest decline in loss to use in the following cycle, thereby fine-tuning the overall accuracy of the model.

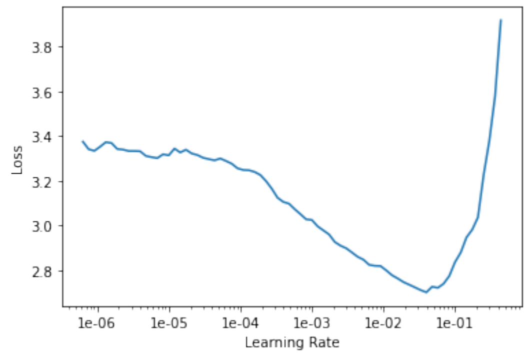

Fig. 2. An example of learning rate accuracy graph.

After a total of 5 cycles, our model achieved an accuracy of 0.809 for correctly classifying the dominant emotion in a speech sample. See the confusion matrix for more details (Figure 3).

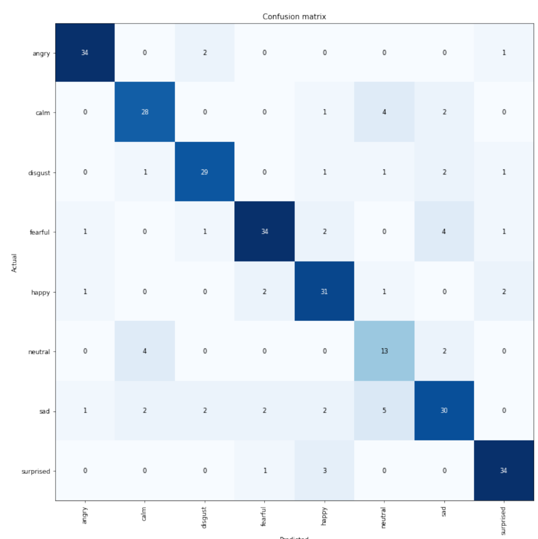

Fig. 3. Confusion matrix for our model.

3.3 Creating the API

We exported the trained model as .pkl file and uploaded it to archive.org for easy public access. The model can be downloaded and used from this link. We then created a Flask API that can be deployed locally (Colab notebook here). The API accepts a .wav audio file, transforms it into a spectrogram, and uses the pretrained model to classify the emotional content of the recording. Only the string value of the top class is returned to the requester. While there is much potential to improve the usability of this tool, the current functionality was sufficient for the game project we had in mind.

3.4 Developing the game and connecting to API

The largest hurdle in our project was to figure out the more technical aspects in regards to machine learning and the backend flow of our model to a game. Therefore, we decided that our game should be simple but succinct in demonstrating the capability of emotion detection in voice. We used Unity to create a simple 2D side-scrolling game where the player controls a character who goes on an emotionally difficult journey towards realizing they need therapy (a screenshot displayed in Figure 3).

We were inspired to pursue this theme because of the various stages involved in the journey to acceptance, such as denial, anger, depression, bargaining, and acceptance. Each one of those stages could be narratively related to one of the emotions present in our emotion detection model (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). To invoke the player to emotionally react in a more natural way, we embedded emotion detection into the game's mechanics. To "win" the game, the player must get past their mental blocks (represented by actual physical blocks in the game) by emotionally speaking in a way that's appropriate for the current situation presented in the game. For example, if the character just got unexpectedly fired, the player would have to attempt to sound surprised ("How could you fire me?!"). If the player was able to vocally emote in the way the game wants, the player progresses.

Regarding the backend of this emotional voice game controller, the player has to tap "Record", speak into their device, then tap "Send". A .wav file is then created which is sent to our Emotion Prediction API. Our model then sends the detected dominant emotion label back to the game in Unity, which is displayed to the player and, if correct, unlocks new areas of the game. The player can attempt this as many times as needed until they reach the end of the game.

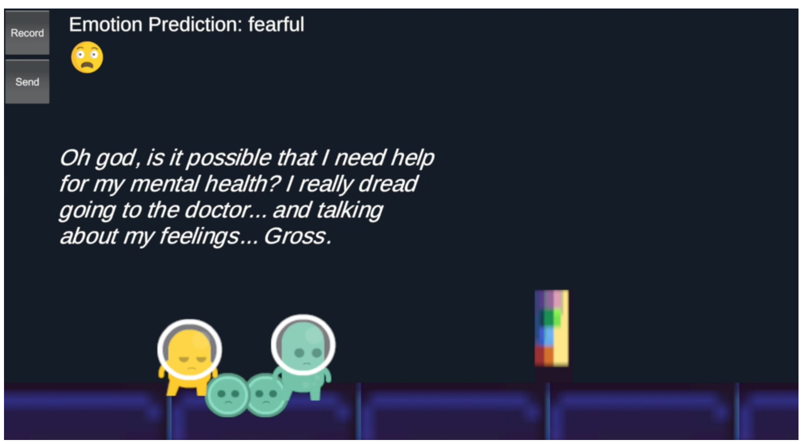

Fig. 4. A screenshot of the game interface.

3.5 Resources

All of our code (Colab notebooks and Unity game) can be found in this Github repository.

4 RESULTS

After training our model, we were able to achieve about 80% accuracy for emotion detection in voice. However, perceived in-game accuracy was much lower. Depending on the player's natural level of expressiveness, it can be hard to vocalize a certain emotion in a way that the model recognizes. For example, one author's "sad" recordings would often be classified as "calm" by the system. In addition to individual difference, this misclassification might also be due to noisy recordings and low sound quality.

Despite this shortcoming, we were pleasantly surprised at the seamlessness of recording the player's voice and obtaining the emotion prediction from our model within the game. Though the prediction wasn't instantaneous, it took only about 1 second to receive a classification after creating a recording and the game flow did not get interrupted. We are still hoping to improve the backend connection process to minimize the time it takes to create and classify a recording.

5 CONCLUSION

Through our project, we were able to demonstrate that voice-based input for affective game interaction and adaptivity is possible. Our system was able to use a CNN model in a game to predict the player's emotion in speech within a few seconds with 80% accuracy. We aim to expand on our work by improving our model's accuracy, the backend flow from the model to game engine, and the accompanying game's synchronization of player emotions to its mechanics. We hope our future work makes contributions to affect-adaptive gaming and its applications towards both entertainment and learning domains.

REFERENCES

[1] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[2] Gordon H. Bower. "Mood and memory". In: *American Psychologist* 36.2 (1981), p. 129. ISSN: 0003-066X.

[3] Paul Ekman. "An argument for basic emotions". In: *Cognition and emotion* 6.3-4 (1992), pp. 169–200. ISSN: 0269-9931. URL: http://www.tandfonline.com/doi/abs/10.1080/02699939208411068.

[4] Karen Gasper and Gerald L. Clore. "Attending to the Big Picture: Mood and Global versus Local Processing of Visual Information". In: *Psychological Science* 13.1 (2002), pp. 34–40. ISSN: 09567976.

[5] Robin Hunicke, Marc LeBlanc, and Robert Zubek. "MDA: A formal approach to game design and game research". In: *Proceedings of the AAAI Workshop on Challenges in Game AI*. Vol. 4. 1. San Jose, CA. 2004, p. 1722.

[6] Petra Sundström. "Exploring the affective loop". PhD thesis. 2005.

[7] Jenova Chen. "Flow in games (and everything else)". In: *Communications of the ACM* 50.4 (2007), pp. 31–34.

[8] Nicole Lazzaro. "Why we play: affect and the fun of games". In: *Human-computer interaction: Designing for diverse users and domains* 155 (2009), pp. 679–700.

[9] Andreas Alexiou, Michaéla Schippers, Ilan Oshri, et al. "Positive psychology and digital games: the role of emotions and psychological flow in serious games development". In: *Psychology* 3.12 (2012), p. 1243.

[10] L.D Stafford and R Collins. "The interactive effects of mood and memory on food intake". In: *Appetite* 71 (2013), pp. 486–486. ISSN: 0195-6663.

[11] Georgios N Yannakakis and Ana Paiva. "Emotion in games". In: *Handbook on affective computing* 2014 (2014), pp. 459–471.

[12] Julia Ayumi Bopp, Elisa D Mekler, and Klaus Opwis. "Negative emotion, positive experience? Emotionally moving moments in digital games". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 2996–3006.

[13] Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. "Emotion detection: a technology review". In: *Proceedings of the XVIII international conference on human computer interaction*. 2017, pp. 1–8.

[14] Mark D Griffiths, Daria J Kuss, and Angelica B Ortiz de Gortari. "Videogames as therapy: an updated selective review of the medical and psychological literature". In: *International Journal of Privacy and Health Information Management (IJPHIM)* 5.2 (2017), pp. 71–96.

[15] Carlos Alberto Lara et al. "Induction of emotional states in educational video games through a fuzzy control system". In: *IEEE Transactions on Affective Computing* (2018).

[16] Klara Steinhauser et al. "Effects of emotions on driving behavior". In: *Transportation research part F: traffic psychology and behaviour* 59 (2018), pp. 150–163.

[17] Ruhul Amin Khalil et al. "Speech emotion recognition using deep learning techniques: A review". In: *IEEE Access* 7 (2019), pp. 117327–117345.

[18] Rebecca Lietz et al. "Survey of mood detection through various input modes". In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 2019, pp. 28–31.