Emotional Musical Prosody: Validated Vocal Dataset for Human Robot Interaction

Richard Savery, Lisa Zahray, Gil Weinberg *

Georgia Tech Center for Music Technology
rsavery3@gatech.edu

Abstract. Human collaboration with robotics is dependant on the development of a relationship between human and robot, without which performance and utilization can decrease. Emotion and personality conveyance has been shown to enhance robotic collaborations, with improved human-robot relationships and increased trust. One under-explored way for an artificial agent to convey emotions is through non-linguistic musical prosody. In this work we present a new 4.2 hour dataset of improvised emotional vocal phrases based on the Geneva Emotion Wheel. This dataset has been validated through extensive listening tests and shows promising preliminary results for use in generative systems.

Keywords: robotics, emotion, prosody, music

1 Introduction

As the use of robotics and artificial agents continue to expand, there is a growing need for better forms of communication between human and computer. While speech based mediums are common in home assistants and robots, we contend that for many human-agent collaborations, semantic meaning is not always required. One alternate to speech communication is non-linguistic emotional musical prosody where a robot communicates through musical phrases. Using musical phrases can avoid the challenges of uncanny valley, allowing robot interaction to have it's own human-inspired form of communication (Savery, Rose, & Weinberg, 2019b). In past work we have successfully shown the capabilities of prosody in robotic systems through simulations of a potential generative system. This includes increased trust in social robots (Savery, Rose, & Weinberg, 2019a) and industrial robots (Savery, Zahray, & Weinberg, 2020). In this paper we present the collection and validation of a emotional prosody dataset that can be used to improve future human robot interactions.

2 Custom Dataset

Our custom dataset with 4.2 hours of audio was created with the singer Mary Carter. After a pilot study collecting data, we decided to use the Geneva Emotion

* This material is based upon work supported by the National Science Foundation under Grant No. 1925178

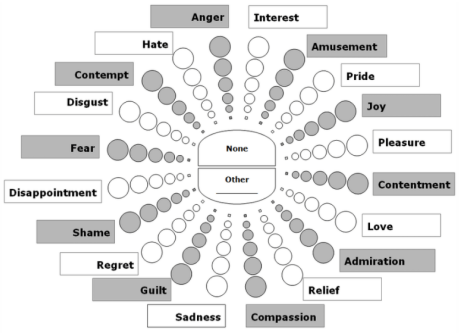

Fig. 1. Geneva Emotion Wheel

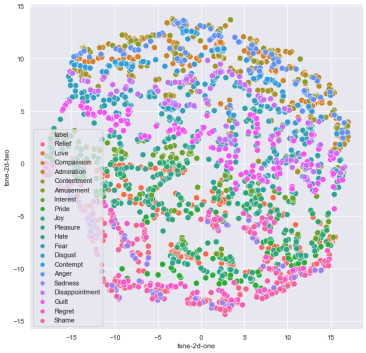

Fig. 2. Vanilla VAE Latent Space

Wheel (GEW) (Sacharin, Schlegel, & Scherer, 2012), which is a circular model, containing 20 emotions with position corresponding to the circumplex model (see Figure 1).

One of the primary advantages of the GEW is that it includes 20 different emotions, but these emotions can also be reduced to four separate classes which align with a quadrant from the circumplex model. GEW also includes most of the Eckman's basic emotions - fear, anger, disgust, sadness, happiness - only leaving out surprise. The ability to use different models of emotion allows for significant future use cases of the dataset (Savery & Weinberg, 2020).

This dataset only has one musician, and therefore only captures one person's perspective on musical emotion. While the dataset can make no claim to represent all emotion and does not create a generalized emotion model, we believe using one person has advantages. By having only one vocalist generative systems have the possibility of recreating one person's emotional style, instead of incorrectly aggregating multiple peoples to remove distinctive individual and stylistic features.

2.1 Process and Data

Carter was paid $500 and recorded the samples over a week long period at her home studio, using a template we created in Logic while maintaining the same microphone positioning. For the samples we requested phrases to be between 1 and 20 seconds, and to spend about 15 minutes on each emotion, allowing jumping between any order of the emotions. We allowed deletion of a phrase if she felt after singing it did not fit the emotion. The final recorded dataset includes 2441 phrases equalling 4.22 hours of data with an average of 122 for each emotion. Samples from the dataset can be heard online.[1]

[1] www.richardsavery.com/prosodycvae

2.2 Dataset Validation

To validate the dataset, we performed a study with 45 participants from Prolific and Mechanical Turk, paying each $3. Each question in the survey asked the participant to listen to a phrase and select a location on the wheel corresponding to the emotion and intensity they believed the phrase was trying to convey. Phrases fell under two categories of "best" and "all", with each participant seeing 60 total phrases selected at random. The "best" category consisted of 5 phrases for each emotion that were hand-selected as best representing that emotion, ensuring an even distribution of phrase lengths in each emotion set. The "all" category consisted of a phrase sampled from all phrases in the dataset for that emotion, with a new phrase randomly selected for each participant. Rose plots of the validation results that combine the "best" and "all" categories can be seen in the appendix, separated into each Geneva Wheel quadrant.

2.3 Dataset to Midi

We converted each phrase's audio into a midi representation to use as training data. We first ran the monophonic pitch detection algorithm CREPE (Kim, Salamon, Li, & Bello, 2018) on each phrase, which outputs a frequency and a confidence value for a pitch being present every 0.01 seconds. As the phrases included breaths and silence, it was necessary to filter out pitches detected with low confidence. We applied a threshold followed by a median filter to the confidence values, and then forced each detected pitch region to be at least 0.04 seconds long. We then converted the frequencies to midi pitches. We found the most common pitch deviation for each phrase using a histogram of deviations, shifting the midi pitches by this deviation to tune each phrase.

3 Conclusion

We have so far analyzed the dataset by training a Vanilla VAE (without emotion labels) on the pitch alone. Figure 2 shows the latent space reduced from 100 dimensions through t-distributed stochastic neighbor embedding (t-SNE) (Hinton & Roweis, 2003). This demonstrates that the latent space is able to separate by emotion, without any knowledge of the emotional labels. We then used a conditional VAE (conditioned on emotion labels) to generate phrases with playback through a separate audio sampler. These phrases were validated following the format of our dataset evaluation, and the results are shown in the appendix. In the future we will run extensive studies with generated prosody applied to human-robot interactions. This will take place between varying group sizes from one human and robot, to groups of humans and robots with different embedded personalities. We expect for emotional musical prosody to enable many future collaborations between human and robot.

References

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems* (pp. 857–864).

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). Crepe: A convolutional representation for pitch estimation. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 161–165).

Sacharin, V., Schlegel, K., & Scherer, K. (2012). Geneva emotion wheel rating study (report). geneva, switzerland: University of geneva. *Swiss Center for Affective Sciences*.

Savery, R., Rose, R., & Weinberg, G. (2019a). Establishing human-robot trust through music-driven robotic emotion prosody and gesture. In *2019 28th ieee international conference on robot and human interactive communication (ro-man)* (pp. 1–7).

Savery, R., Rose, R., & Weinberg, G. (2019b). Finding Shimi's voice: fostering human-robot communication with music and a NVIDIA Jetson TX2. *Proceedings of the 17th Linux Audio Conference*, 5.

Savery, R., & Weinberg, G. (2020). A survey of robotics and emotion: Classifications and models of emotional interaction. In *Proceedings of the 29th international conference on robot and human interactive communication.*

Savery, R., Zahray, L., & Weinberg, G. (2020). Emotional musical prosody for the enhancement of trust in robotic arm communication. In *29th ieee international conference on robot & human interactive communication.*

4 Appendix

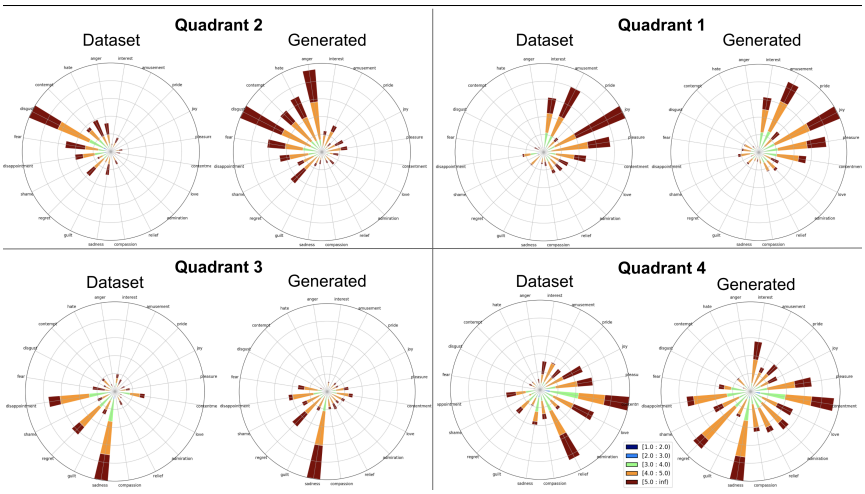

Fig. 3. Rose plots of dataset validation for each emotion quadrant