# Machine Learning Driven Musical Improvisation for Mechanomorphic Human-Robot Interaction

Richard Savery
rsavery3@gatech.edu
Georgia Tech Center for Music Technology
Atlanta, GA

## ABSTRACT

As industrial robots and social robots become prevalent in commercial and home settings it is crucial to improve forms of communication with human collaborators and companions. In this work, I describe the use of musical improvisation to generate emotional musical prosody for improved human-robot interaction. This aims to develop a canny approach, where robots perform in a mechanomorphic manner improving collaboration opportunities with humans. I have currently collected a new 12-hour dataset and developed a Conditional Variational Autoencoder to generate new phrases. Generations have then been used to compare the impact of prosody on anthropomorphism, animacy, likeability, perceived intelligence, and trust. Future work will incorporate prosody into groups of robots and humans, using personality to drive emotional decisions and emotion contagion.

## CCS CONCEPTS

• **Hardware → Sound-based input / output**.

## KEYWORDS

robotics, sound, emotion, music, audio, prosody

**ACM Reference Format:**
Richard Savery. 2021. Machine Learning Driven Musical Improvisation for Mechanomorphic Human-Robot Interaction. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3434074.3446351

## 1 INTRODUCTION

As the role of robots and artificial intelligence increases, there is a need to find improved ways to interact with these systems. Many systems focus on achieving human-like features and interactions, building around anthropomorphic design. There is a growing body of work that also suggests the counter approach of mechanomorphic design has a greater potential for future development [16]. Uncanny valley describes that as a robot becomes more human, they become more appealing, until they reach a point where they elicit revulsion. Moore suggests the contrasting "canny" approach, whereby robots are developed that are clearly robots [15] and openly display their capabilities and robotic features.

In my Ph.D. dissertation, I propose that musical improvisation offers a "canny" approach to robot design and communication. Just as the field of human-robot interaction (HRI) requires new thinking and methods beyond those existing in human-computer interaction [4], robotic audio creation can not simply recycle existing approaches. HRI needs audio specifically tailored to robotics, with deep consideration for embodied sound and its implication across use cases and platforms. I believe music can offer methods of interacting that are comparable to human-communication while avoiding the uncanny valley. Music is an inherently dialogue-like medium, between composer, performer and audience [3] while Higgins describes that for most listeners, even music through headphones is "a kind of communication between themselves and other human beings" [11]. In addition to possibilities for direct communication, musical interaction can act as a framework to develop broader principles for human robot interaction. I focus on using emotional musical prosody for communication in both social and industrial robotics.

## 2 RELATED WORK

Verbal language-based interaction is the prominent form of communication used in human-robot interaction [14] covering a wide range of tasks from robot companions [8] to industry [17]. Many robotic interactions do not include language; these non-verbal forms of communication fall into six categories, kinesics, proxemics, haptics, chronemics, presentation, and vocalics [12, 19]. The final category, vocalics, includes ideas such as prosody [7]. The vast majority of these communication techniques require significant technical and financial expense and variation to a system [19].

Robot personality has been shown to improve human-robot interaction with related research on artificial agents and personality traits [2, 13] indicating that an effective approach for collaborating with artificial agents is through conveying emotions using non-verbal communication channels such as prosody. Efforts to generate and manipulate prosody focused on linguistic robotic communication [7], and have been successful in conveying expressions such as approval, prohibition, attention, and comfort [5]. Music is a powerful medium to convey emotions [26], and shares many of the underlying building blocks of prosody such as pitch, timing, loudness, intonation, and timbre [10, 28, 29]. Music generation has been widely addressed as a deep learning task [6], in particular using LSTMs [27, 30]. Music tagged with emotion has also been achieved through LSTMs with logistic regression used to generate music with sentiment [9], or a BALSTM network [31] generating musical phrases corresponding to Russel's valence-arousal emotion space.


*HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA*

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3446351

## 3 CURRENT RESULTS

The first key research area is developing emotion-tagged musical prosody to drive human-robot communication and interaction. I will build a scalable model that allows real-time interactions with emotional musical prosody, that can utilize a newly created dataset. This prosody will then be utilized for multiple benefits to human-robot interaction such as improved trust and likeability. I will then use emotional prosody for both human-to-robot and robot-to-robot communication allowing for consistent team-work between groups of humans and robots. These research areas lead to the primary research question:

> **Can an emotion-tagged musical prosody generative system improve trust, likeability, and perceived intelligence in individual and group human-robot interaction?**

This research question will be split into four sub-questions aiming to address specific parts of each question. I have made significant progress with research questions 1 and 2. The results from these studies have been published, [20–24] and are discussed in the following section. Two of these papers have successfully presented a development of the real-time generative system and have been applied to use cases in social and industrial robotics. The other two published papers have studied the application of prosody to trust and other HRI metrics. Research questions 3 and 4 are in progress and described in the future work section.

*RQ 1: How can we develop a scalable, real-time model for emotion-tagged prosody generation utilizing a newly created data-set of improvisations?*

To develop a prosody generator I created a new dataset consisting of the material from three vocalists, each one improvising using the Geneva Emotion Wheel (GEW) [18] for 4 hours. GEW includes 20 discrete emotion categories, five for each quadrant of the circumplex model of emotion. The generative system was designed foremost to operate in real-time and allow rapid generation and dialogue exchange between human and robot. For this reason, the system combines symbolic deep learning through a Conditional Convolution Variational Auto-encoder, with an emotion-tagged audio sampler. This aimed to combine the best generation options available with a system that can be implemented on a variety of platforms. I evaluated this system primarily through listening tests, with users rating their ability to recognize emotions and qualitative questions and showing success rates matching those achieved by the dataset.

*RQ 2: How can non-verbal prosody through musical phrases accurately increase the level of anthropomorphism, animacy, likeability, perceived intelligence, and trust?*

To analyze these HRI metrics and prosody I then compared their implementation in a robotic industrial arm and Shimi, a social robot. I conducted three separate studies to answer these questions. The first study focused on Shimi interacting with emotional phrases that match a users' choice emotion and then measuring users' trust, and had 28 in-person participants. I was able to show a significant result ($p<0.001$) for the prosody system over gestures, or speech-to-text systems. The second study involved participants interacting with a robotic arm assisting in a pattern recognition task and responding emotionally to the users' choices, and had 92 online participants. A post-interaction survey was conducted using the Godspeed metrics [1] for anthropomorphism, animacy, likeability, perceived intelligence, and the Schaefer2016 survey [25] for trust. Figure 2 shows a comparison for the Godspeed metrics between the robot arm with no audio, baseline audio, and the prosody system. These results indicated that prosody has a significant difference for anthropomorphism and likability, but not for perceived safety. The final study compared the arm and social robot between groups completing the same task. Figure 2 shows the variation in results between a social robot (Seekerbot), a Robotic Arm, a and mobile robot (Stretch). While further research is required, these results indicated that audio's performance is not consistent across platforms and does require refinement based on the robot and its desired use case.

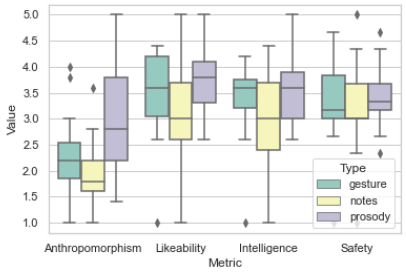


**Figure 1: Godspeed metrics for Robotic Arm with different audio**

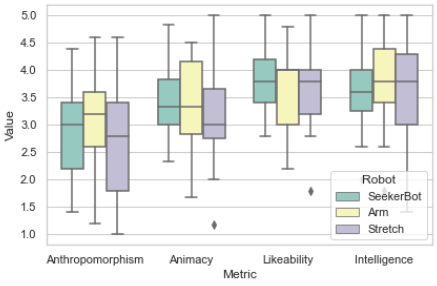


**Figure 2: Godspeed metrics comparing platforms with prosody**

## 4 FUTURE WORK

Research question 3 will explore how personality can be used to choose emotional reactions for robots. This will emphasize both the robots taking on personality traits and identifying personality traits from human collaborators. The robotic responses will be dictated by common strategies used for personality types with high and low neuroticism and extraversion. These interactions will be expanded to large groups of humans and robots for research question 4, allowing an exploration of the range of possibilities for musical prosody while developing new knowledge about humans' preference for robotic emotional response.

## REFERENCES


[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[2] Joseph Bates. 1994. The Role of Emotion in Believable Agents. *Commun. ACM* 37, 7 (July 1994), 122–125. https://doi.org/10.1145/176789.176803

[3] Bruce Ellis Benson. 2003. *The improvisation of musical dialogue: A phenomenology of music.* Cambridge University Press.

[4] Cindy L Bethel and Robin R Murphy. 2010. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics* 2, 4 (2010), 347–359.

[5] Cynthia Breazeal and Lijin Aryananda. 2002. Recognition of affective communicative intent in robot-directed speech. *Autonomous robots* 12, 1 (2002), 83–104.

[6] Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. 2017. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620* (2017).

[7] Joe Crumpton and Cindy L Bethel. 2016. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics* 8, 2 (2016), 271–285.

[8] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nehaniv, A. Sisbot, R. Alami, and T. Siméon. 2006. How May I Serve You? A Robot Companion Approaching a Seated Person in a Helping Context. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, Utah, USA) *(HRI '06)*. Association for Computing Machinery, New York, NY, USA, 172–179. https://doi.org/10.1145/1121241.1121272

[9] Lucas Ferreira and Jim Whitehead. 2019. Learning to Generate Music With Sentiment.. In *ISMIR*. 384–390.

[10] Maija Hausen, Ritva Torppa, Viljami R Salmela, Martti Vainio, and Teppo Särkämö. 2013. Music and speech prosody: a common rhythm. *Frontiers in psychology* 4 (2013), 566.

[11] Kathleen Marie Higgins. 2011. The music of our lives. (2011).

[12] Richard Jones. 2013. *Communication in the real world: An introduction to communication studies.* The Saylor Foundation.

[13] Michael Mateas. 1999. Artificial Intelligence Today. Springer-Verlag, Berlin, Heidelberg, Chapter An Oz-centric Review of Interactive Drama and Believable Agents, 297–328. http://dl.acm.org/citation.cfm?id=1805750.1805762

[14] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35.

[15] Roger K Moore. 2017. Appropriate voices for artefacts: some key insights. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots.*

[16] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots.* Springer, 281–291.

[17] J Norberto Pires and Amin S Azar. 2018. Advances in robotics for additive/hybrid manufacturing: robot control, speech interface and path planning. *Industrial Robot: An International Journal* (2018).

[18] Vera Sacharin, Katja Schlegel, and KR Scherer. 2012. Geneva Emotion Wheel rating study (Report). Geneva, Switzerland: University of Geneva. *Swiss Center for Affective Sciences* (2012).

[19] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608.

[20] Richard Savery, Ryan Rose, and Gil Weinberg. 2019. 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)*. IEEE, 1–7.

[21] Richard Savery, Ryan Rose, and Gil Weinberg. 2019. Finding Shimi's voice: fostering human-robot communication with music and a NVIDIA Jetson TX2. *Proceedings of the 17th Linux Audio Conference* (2019), 5.

[22] Richard Savery and Gil Weinberg. 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.*

[23] Richard Savery, Lisa Zahray, and Gil Weinberg. 2020. Emotional Musical Prosody for the Enhancement of Trust in Robotic Arm Communication. In *Trust, Acceptance and Social Cues in Human-Robot Interaction, RO-MAN 2020.*

[24] Richard Savery, Lisa Zahray, and Gil Weinberg. 2020. ProsodyCVAE: A Conditional ConvolutionalVariational Autoencoder for Real-timeEmotional Music Prosody Generation. In *2020 Joint Conference on AI Music Creativity.* CSMC + MUME.

[25] Kristin E. Schaefer. 2016. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI".* Springer US, Boston, MA, 191–218. https://doi.org/10.1007/978-1-4899-7668-0_10

[26] John Sloboda. 1999. Music: Where cognition and emotion meet. In *Conference Proceedings: Opening the Umbrella; an Encompassing View of Music Education; Australian Society for Music Education, XII National Conference, University of Sydney, NSW, Australia, 09-13 July 1999.* Australian Society for Music Education, 175.

[27] Bob L Sturm, Joao Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. 2016. Music transcription modelling and composition using deep learning. *arXiv preprint arXiv:1604.08723* (2016).

[28] William Forde Thompson, E Glenn Schellenberg, and Gabriela Husain. 2004. Decoding speech prosody: Do music lessons help? *Emotion* 4, 1 (2004), 46.

[29] Ann Wennerstrom. 2001. *The music of everyday speech: Prosody and discourse analysis.* Oxford University Press.

[30] Jian Wu, Changran Hu, Yulong Wang, Xiaolin Hu, and Jun Zhu. 2019. A hierarchical recurrent neural network for symbolic melody generation. *IEEE Transactions on Cybernetics* 50, 6 (2019), 2749–2757.

[31] K. Zhao, S. Li, J. Cai, H. Wang, and J. Wang. 2019. An Emotional Symbolic Music Generation System based on LSTM Networks. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).* 2039–2043.