Lubos Marcinek
WASP Software Engineering Course, Module 2025
September 1, 2025

1 Introduction

Human-Robot Interaction (HRI) is a highly challenging area where the theoretical foundations of software engineering for artificial intelligence (AI) systems meet the real-world deployment challenges. My research is focused on designing conversational systems that can perceive and respond to user emotional states, specifically their enjoyment while engaging in open-domain interactions with social robots. The work explores the intersection of affective computing, multimodal machine learning, and adaptive systems, with Software Engineering challenges that are hard to tackle using traditional methods.

The main topic is that effective companion robots, particularly for applications like elder care or education, should dynamically adapt their conversation strategy based on dynamic real-time awareness of user emotional states. This implies moving beyond static, rule-based interaction models to systems with the ability to learn, adapt, and optimize behavior through continuous feedback loops, i.e., a paradigm that directly challenges classical software engineering methodology of deterministic systems.

My research focuses on two connected technical areas that illustrate the complexity of building AI-rich systems. The first area is to develop reliable multimodal enjoyment detection through the combination of speech (prosody and lexical content) and non-verbal cues like facial expressions and body postures (Pereira et al., 2024). This area must deal with the subjectivity inherent in emotional assessment, processing constraints in real-time, and the trust problems that occur when applying machine learning models in robotic environments with limited resources. The second area is the creation of adaptive speech synthesis modules driven by modulating the robot's vocal effort and prosody to generate more conversationally intelligible responses, particularly in noisy real environments where intelligibility is an issue (Marcinek et al., 2025).

A key aspect of this research addresses fundamental architectural decisions in AI system design: making trade-offs between when to use computationally expensive explicit reasoning capability and inexpensive direct prediction for subjective rating tasks. My latest empirical work (Marcinek et al., 2025) evaluates reasoning-capable and non-reasoning LLMs over a variety of granularity levels (turn-level, interaction-level, and overall conversation enjoyment) with the HRI CUES dataset. The findings contradict the general assumption that reasoning will enhance performance on all fronts, yielding instead subtle model-dependent patterns in which reasoning improves turn-level accuracy (0.42 vs 0.36) but has very small influence on overall interaction-level accuracy. These results have immediate practical implications for important system design decisions in real-time AI applications where computational efficiency and user experience must be carefully balanced.

This effort is representative of the bigger challenges to AI engineering: escaping model-centric development and working toward end-to-end system design that incorporates the entire life cycle from development through deployment and maintenance. The journey of taking research prototypes to production-quality HRI systems reveals differences between standard software engineering practices and the requirements of AI-intensive systems, making this work particularly useful in determining how proven software engineering principles should evolve to address the unique requirements of modern AI system development.

2 Lecture Principles

Two concepts selected from Robert's lectures are the following: **Verification vs. Validation** in ML contexts and the **Science-Engineering Loop**.

Verification vs. Validation takes new meaning in ML engineering. Traditional software engineering verifies whether systems are built to specification (building the product right) and validates specifications meet user needs (building the right product). In my research, verification involves checking correct algorithm implementation. However, validation, i.e., ensuring enjoyment predictions correspond to users' true internal states, is more critical. This pushes my research beyond accuracy metrics toward user studies and correlation analysis with self-reported emotional states.

In terms of the **Science-Engineering Loop**, it provides my research framework. Modern ML makes science hybrid (discovery-focused - "why does this work?") and engineering (solution-focused - "how can we make it work reliably?"). This iterative loop of experiment, deploy, refine drives my research project: forming hypotheses about vocal cues, engineering prototype detection models, testing in robots, observing failures and successes. This experiment, deploy, refine cycle drives my research project: generating hypotheses on vocal cues, creating prototype detection models, applying to robots, observing failures and successes. This observation feeds back into hypothesis improvement, moving my work from research prototype to production system.

3 Guest-Lecture Principles

Two of the concepts selected from guest lectures include: Julian Frattini's **Problem-Space vs. Solution-Space** divide and Per Lenberg's **Psychological Safety** idea.

The **Problem-Space versus Solution-Space** principle is the requirement to fully understand a problem before jumping into solutions. It saves us from falling into the trap of proposing solutions like building neural networks for enjoyment detection without asking basic questions first: Why do we even need to detect enjoyment? What does enjoyment in conversation even look like? By interrogating the problem space, we uncover needs that go beyond accuracy. For example, the requirement for real-time performance and tolerance to noisy data, these are essential factors that guide effective solutions.

On the other hand, **Psychological Safety** is a shared perception that teams feel they can take interpersonal risks without fear. This is necessary even in human-robot teams. For users to be willing to speak their minds and for robots to get genuine emotional data, users need to feel psychologically safe. The more users encounter frustrating robot reactions, the more defensive they become, meaning less genuine emotional expression. So, ensuring the user experience is good is not just a system requirement but is actually a fundamental requirement for it to function properly.

4 Data Scientists versus Software Engineers

After reading the CMU Machine Learning in Production chapters (Kästner, 2022), the distinction between software engineers and data scientists is both accurate and vital.

4.1 Fundamental differences

I strongly agree with the highlighted distinctions. Data science is research-driven, investigatory, and focused on acquiring knowledge by exploring models. Software engineering builds robust, scalable, reliable systems that meet requirements. Data science is concerned with the model and software engineering with the entire system. A data scientist's success may be model correctness in notebooks, but engineers get models to go out to millions of people with low latency, refresh without bringing the system down, and fail gracefully.

4.2 Role evolution

Evolution will combine increased specialization with necessary integration of core competency, creating workers who specialize but also know the basics of other fields. Specialization will deepen, i.e., data science may split into Research Scientist and ML Analyst roles, while software engineering develops MLOps Engineer and AI Infrastructure Engineer positions. But complete segregation is not possible. Models are meaningless without systems, and systems are lost without completely comprehended models. The careers will not merge entirely, but overlap of skills will increase substantially, requiring shared vocabulary and mutual knowledge of AI product life cycles.

5 Paper Analysis

5.1 Paper 1: "A Meta-Summary of Challenges in Building Products with ML Components – Collecting Experiences from 4758+ Practitioners" by Nahar et al. (CAIN 2023)

1. Core ideas and their SE importance: Nahar et al. (2023) covers one of the biggest challenges of learning about software engineering issues that occur with the integration of machine learning in real production systems. The authors performed a large-scale meta-analysis of 50 empirical studies involving views of over 4,758 practitioners. They also created a taxonomy where they categorized more than 500 listed challenges into a number of primary categories: data management, model training and serving, testing and monitoring, team collaboration, and process integration.

Among the paper's most significant findings is that the most frequently mentioned difficulties are overwhelmingly systems engineering and not algorithmic. Practitioners worry most about data management and quality, followed by testing machine learning sub-components' complexities and data scientists and software engineers' coordination. This provides strong proof that systems engineering and not model performance is the main bottleneck in machine learning systems. The paper further presents a detailed examination of AI engineering needs, pointing out that tools and research have to focus on the aspect that production ML systems need significantly varied engineering practices in comparison to traditional software engineering.

2. Relation to my research: The paper examines the difficulties I am currently facing in moving from research prototypes to production HRI systems. The issues surrounding testing are particularly salient as they relate to my enjoyment detection area, as traditional approaches assume deterministic behavior and known ground truths. However, enjoyment requires subjective assessments where it is near impossible to define a ground truth and would vary between each subject. The collaborative challenges regarding teams with differing expertise are highly relevant to my interdisciplinary project as it

includes the integration of knowledge from psychology, speech processing, computer vision, and systems engineering. The data quality emphasis in the paper mirrors the challenges I have faced to obtain clean representative datasets for training enjoyment detection and speech synthesis models that operate in real world HRI scenarios that include noisy audio, variability in lighting conditions, and demographic variability.

3. Integration into a larger AI-intensive project: In a future project for an adaptive elder care companion robot, these findings would basically determine system architecture and development strategies. The study findings would determine an appropriate team structure based on cross functional teams specifically focused on the collaboration needs of ML researchers, systems engineers and eldercare domain experts. Based on the data management findings, the project would take a strategic approach to pipelines that include automation and other quality checks, data set versions and isolation, data lineage, automated data drift detection, and bias monitoring by demographic characteristics. In terms of the testing challenges, the project would evolve thorough approaches not only limited to traditional unit testing by looking at monitoring after continuous deployment, A/B testing approaches for conversational strategies, deciding the need for a simulation environment for exploratory edge cases, and creating user-in-the-loop validation schemes that allow capturing user comments/feedback on subjective interaction quality.

4. Adaptation of my research: This paper would reshape my focus from model-centered to system-centered from the beginning of a project. Rather than just thinking about improving detection accuracy of enjoyment or naturalness of speech synthesis, I would engage more with overlapping engineering issues throughout the project. I would have individual work packages for data (strong data collection and labeling pipelines with quality control), test procedures (e.g., systems for validating subjective model output, mechanisms for resolving disagreement among annotators, creating confidence intervals around the predictions), and collaborative activities (e.g., creating good interfaces around components, shared metrics). Most importantly, the evidence-to-production gap exercise that the paper introduces would support evidence-based user testing throughout development vs viewing testing as an end validation step that suggests rigor - so in other words, to speak to the main finding that this is fundamentally an engineering problem that needs systems thinking, not just more algorithm development.

5.2 Paper 2: "Towards Understanding Model Quantization for Reliable Deep Neural Network Deployment" by Hu et al. (CAIN 2023)

1. Core ideas and their SE importance: Hu et al. (2023) addresses a core issue related to reliability around deep neural networks. More specifically, it resolves questions about how model quantization, the process of reducing the numerical precision of models to deploy on resource-constrained platforms, affects model performance on out-of-distribution data. The authors offer considerable empirical investigation on multiple datasets and models, making an effort to detail reliability effects of quantization that go beyond just accuracy metrics.

The key findings have notable implications for production: quantized models produce significantly more mismatches than full-precision models when run against distribution-shifted data; quantization-aware training produces more robust models than using post-training quantization; standard uncertainty metrics are insufficient to fully capture behaviors influenced by quantization. From a software engineering perspective, this work is addressing essentially quality assurance issues for software systems that

guarantee system dependability when compromising to make it suitable for deployment purposes. This work exposes that compression methods commonly used to deploy realistic AI systems introduce subtle but significant issues of reliability, many of which are missed by standard evaluation metrics, and exemplifies typical SE trade-offs between attributes of software systems (e.g., latency/resource efficiency vs. reliability/correctness).

2. Relation to my research: This paper is very applicable because quantization is a requirement for using my speech synthesis and enjoyment detection models to deploy companion robots in real-time. Elder care robots will be deployed on constrained hardware and must demonstrate responsiveness, meaning they must be quantized. These results regarding quantization effects on out-of-distribution data are especially troubling for my particular use case: the elder care context has data distributions that are very different from training data due to different accents, speech patterns, ambient noise environments, face shapes, and lighting differences. If quantization decreases the model's robustness to these normal variations, then this would create a serious effect on the accuracy of enjoyment detection in the real world. Also, the limitations of traditional measures for uncertainty in predicting quantization failure would require entirely new validation methods for quantized model variants because current enjoyment detection validation is largely reliant on confidence levels and prediction uncertainty.

3. Integration into a larger AI-intensive project: As part of the adaptive elder care companion robot project, these insights would become part of a larger model reliability assurance framework, operated over the stages of the system lifecycle. The framework would be actively under development, and would undertake recommended quantization-aware training approaches for both enjoyment detection and speech synthesis models, training from scratch with quantization limitations, rather than applying quantization after that training is complete. Reliability monitoring would involve custom uncertainty metric results - building systems independent of the usual confidence scores, by including margin-based metrics identified as demonstrating improved error prediction due to quantization-induced effects. There would be automated testing pipelines which would evaluate the performance of quantized models on artificially generated out-of-distribution data that simulated conditions being met in many elder care environments that could arguably include variation of ambient noise, illumination, speech profiles, and population characteristics either unrepresented or underrepresented in the training data.

4. Adaptation of my research: This paper would expand my research agenda to include additional measures of deployment reliability beyond model accuracy. The research methods would consider quantization at the development stage rather than optimizing the model later when deploying it. In this work I would develop quantization-resistant architectures that would also include designs that are more resistant to precision loss, and explore other neural network architectures, activation functions, or training objectives that would support performance when quantized. This project would construct expert rating datasets that simulate the distribution shifts that occur in eldercare sites, beyond the types of benchmarks that are typically considered in research, towards individual instances of applied evaluation. The primary outputs will encourage development of new reliability measures, predicting failures of quantized models in real world conditions; the design of test suites that systematically explore the full input variation space; and how to specify those conditions in which quantization influences behaviors clearly. Most importantly, it will shift the deployment mentality to be more about considerations of safety comes first, preferential for repeated, known safe behavior across the range of deployment conditions over mainly improved performance, so that reliability will be prioritized as a first-class engineering process.

6 Research Ethics & Synthesis Reflection

Search and screening: I searched articles from CAIN proceedings that involved the words machine learning engineering, production systems, and reliability. Several papers were screened, of which I selected papers presenting systematic solutions. The Nahar meta-summary paper provided comprehensive empirical information, while the quantization paper focused on specific deployment reliability issues.

Pitfalls and mitigations: I avoided narrow-scope papers as I focused on papers providing descriptions of frameworks and empirical work. I checked that papers covered system-level issues and not end-to-end ML research focused on real-world deployment.

Ethical considerations: All analysis represents my understanding of the SE principles as they relate to HRI issues. I have appropriately given credit to concepts and not copied material.

References

Hu, Q., Guo, Y., Cordy, M., Xie, X., Ma, W., Papadakis, M., and Le Traon, Y. (2023). Towards understanding model quantization for reliable deep neural network deployment. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (CAIN)*, pages 165–176.

Kästner, C. (2022). *Machine Learning in Production: From Models to Products*. Carnegie Mellon University. Available at: `https://mlip-cmu.github.io/`.

Marcinek, L., Irfan, B., Skantze, G., Pereira, A., and Gustafson, J. (2025). Role of reasoning in llm enjoyment detection: Evaluation across conversational levels for human-robot interaction. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Marcinek, L., Beskow, J., and Gustafson, J. (2025). Towards adaptable and intelligible speech synthesis in noisy environments. In *Proceedings of Interspeech 2025*.

Nahar, N., Zhang, H., Lewis, G., Zhou, S., and Kästner, C. (2023). A meta-summary of challenges in building products with ml components – collecting experiences from 4758+ practitioners. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (CAIN)*, pages 153–164.

Pereira, A., Marcinek, L., Miniota, J., Thunberg, S., Lagerstedt, E., Gustafson, J., Skantze, G., and Irfan, B. (2024). Multimodal user enjoyment detection in human-robot conversation: The power of large language models. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI)*, pages 469–478. `https://doi.org/10.1145/3678957.3685729`.