Text-to-Speech Based Emotion-Aware Human-Robot Dialogue System

Elnur Alimirzayev and **Burak Can Kaplan** and **Cornelius Weber** and **Stefan Wermter**
University of Hamburg - Knowledge Technology
Bundesstraße 56b, 20146 Hamburg - Germany
www.knowledge-technology.info

Abstract

The rapid progress of large language models (LLMs) has accelerated the rise of interactive AI systems, yet most remain text-only and largely emotion-agnostic. Adding text-to-speech (TTS) enables voice interaction, but current state-of-the-art TTS models mostly offer indirect, coarse control over prosody and frequently fail to realize requested emotions. This mismatch between intended content and delivered tone undermines user trust in high-stakes settings such as customer support and healthcare. Two technical obstacles cause this gap: 1) content and prosody are tightly entangled, so an inappropriate emotion can degrade intelligibility, and 2) models tend to express the emotions that match the given text while neglecting context. Addressing these issues is essential for emotionally coherent and controllable speech in LLM-driven interactions. Our approach proposes a modular human-robot interaction system that integrates a neural model for emotion recognition in conversations (ERC) with TTS to enable emotional awareness. Due to this integration, our TTS can decide on the correct emotion to answer, select appropriate reference audio to adapt its prosody in each interaction, and generate speech of an "emotionally intelligent" robot. Additionally, we set a new baseline for an emotional dialogue system pipeline and automated evaluation of such systems. A demo and code will be made publicly available. A demo and code are publicly available at https://al1ve1t.github.io/eahris/.

1 Introduction

Contemporary human-robot interaction (HRI) systems are mainly characterized by functional properties, such as the tasks they address, their levels of autonomy, and design challenges (Goodrich and Schultz, 2007), while the agenda of "socially interactive robots" proposes social competence as a core requirement. Trust is pivotal since a large meta-analysis shows that human factors (e.g. experience), robot factors (e.g. reliability, transparency), and environmental factors jointly determine perceived trust (Hancock et al., 2011). Emotion-aware, social signal processing equips HRI with methods to perceive and produce signals, including gaze, nods, timing, pitch, and energy, that regulate turn-taking and convey stance. Other aspects of natural dialogue include detecting ends of turns, handling interruptions, coordinating multi-party exchanges. Recent models predict signals for end-of-turn prediction and generation, highlighting open problems in latency and overlap handling (Skantze, 2021). As the first emotional text-to-speech systems have only recently emerged, emotional voice generation still requires integration into dialogue systems.

Therefore, we present a newly developed system with category-controllable, expressive short-latency prosody towards real-world applications that requires limited training data. Our contributions are as follows: 1. a novel system that integrates emotion assessment of the current utterance in context with emotion controlled TTS for dyadic human-robot dialogue, 2. a modular software architecture with easily interchangeable pretrained and fine-tuneable components, 3. an automated evaluation framework coupling emotion accuracy with intelligibility metrics, defining a new benchmark for such integrated systems.

2 Related Work

2.1 HRI Systems and Evaluation

Early works on social robots, such as the affective Kismet (Breazeal, 2000), have led to various emotion-aware applications, such as the facial animation expressing robot Furhat[1], the empathy robot Navel[2], the social companion ElliQ[3], among others.

[1] https://www.furhatrobotics.com/
[2] https://navelrobotics.com
[3] https://elliq.com/

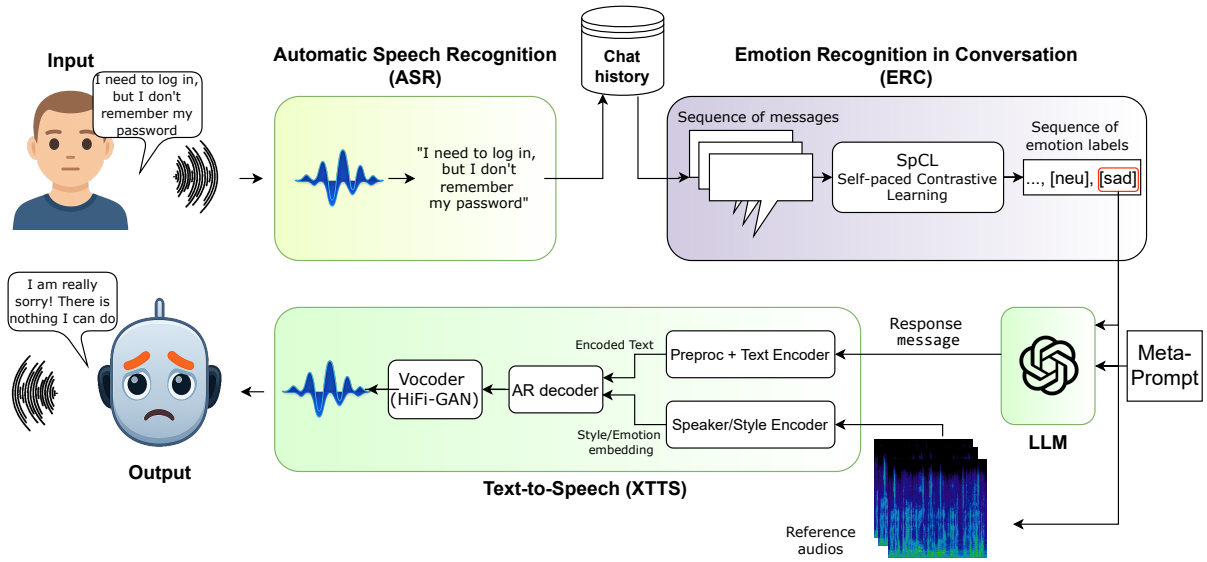

Figure 1: Software architecture involving ASR, ERC, LLM and TTS. Boxes represent replaceable modules.

They cover appraisal-based affect, projected-face control, multimodal emotion dialogue, in order to enable empathic elder care, proactive companionship, and educational platforms. They typically use an LLM service via internet or use Jetson-class hardware. Yet, despite advances in language understanding and nonverbal perception, prosodic requirements remain underspecified and many solutions depend on heavyweight sensing. Evaluation practices such as Wizard-of-Oz (Riek, 2012) and Godspeed (Bartneck, 2023) remain central and come with reporting guidelines that support replicability and ethical conduct. Common research corpora of emotional conversations include IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2018). Moreover, HRI increasingly integrates LLM-based conversation with perception and control. For instance, grounding language in robot capabilities (Ichter et al., 2023), fusing continuous sensor modalities with language (Driess et al., 2023), and training vision-language-action models (Zitkovich et al., 2023), thus broadening generalization and task coverage.

2.2 TTS Models

Current TTS models tend to mix the characteristics of autoregressive and diffusion-based approaches. XTTS (Casanova et al., 2024) couples a VQ-VAE codebook with a GPT-2 decoder and a style encoder to deliver multilingual zero-shot voice cloning with strong intelligibility and reference style transfer. Diffusion approaches push emotional control further: EmoDiff (Guo et al., 2023) introduces soft-label guidance for continuous intensity, while EmoMix (Tang et al., 2023) mixes emotions by embedding derived from speech emotion recognition (SER), enabling both mixture and intensity control. Zhou et al. (2023) implement mixed emotions via a relative-attribute scheme in an emotional seq2seq TTS. i-ETTS (Liu et al., 2021) uses an SER-style critic to reinforce clearer emotion. A hybrid approach, EmoVoice (Yang et al., 2025), jointly emits phoneme and audio tokens and leverages LLM for emotional appropriateness. Together, these studies map a path to emotion-capable TTS that balances intelligibility with precise, context-aware prosody. Several recent TTS models implement emotion-aware HRI, like Tortoise-TTS (Betker, 2023), F5-TTS (Chen et al., 2025), OpenVoice (Qin et al., 2024) and XTTS-v2 (Casanova et al., 2024). XTTS-v2 (VQ-VAE + GPT-2 AR with HiFi-GAN) performs zero-shot voice cloning with explicit emotion/style transfer by cloning, supports cross-language synthesis with output at 24kHz frequency, and was trained on around 27k hours across currently up to 17 languages.

3 Methodology and Evaluation

3.1 System Description

Our main contribution is a modular HRI architecture for human-robot dialogue with context-adapted emotional speech expressed by the robotic agent NICO. It integrates automatic speech recognition (ASR), emotion recognition in conversation (ERC), and emotional TTS into an LLM-controlled

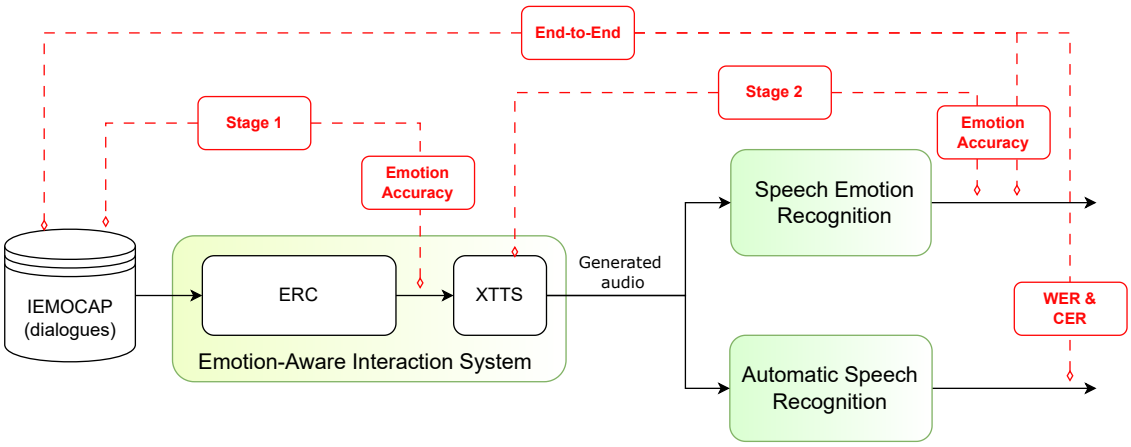

Figure 2: Automated Evaluation Framework for emotion-aware HRI system. Red color shows the stages with evaluated values.

system, as visualized in Figure 1. ASR transcribes user speech, while the ERC model predicts the next-turn emotion. For ERC we use a lightweight model, SpCL (Song et al., 2022), which is pretrained on IEMOCAP. For TTS we use XTTS-v2 (Casanova et al., 2024), which produces emotion and style from a reference voice via cloning. We prompt the LLM to produce a response in that target emotion style, then synthesize it with XTTS using a reference utterance for the predicted emotion drawn from a per-emotion repository (same speaker across references). The architecture's modularity allows the replacement of any component with minimal effort.

For our HRI constraints, predictable, reference-conditioned emotional transfer, and operational simplicity outweigh diffusion's parallelism or OpenVoice's separate style controls. XTTS-v2 provides consistent affect without label taxonomies or prompt heuristics and fits our modular software requirements. While originally around 6 second long reference audios are suggested (Casanova et al., 2024), we found that 10-15 seconds long clean, single-speaker clips yield more robust emotion/style transfer. We therefore maintain a per-emotion repository of reference utterances.

3.2 Training Configuration

The emotion selection model (SpCL) is trained on IEMOCAP to predict the emotion of the next utterance, which does not necessarily mirror the input emotion. This makes it different from the classic emotion recognition models and therefore more suitable for our setup. SpCL uses SimCSE as a backbone (Song et al., 2022). Optimization uses Adam optimizer and a scheduled learning rate configuration with the start value $lr = 1e-3$ and a pretraining learning rate value $ptmlr = 1e-5$. We used a batch size of 32, a maximum token length of 256, and a dropout value of 0.1. We are using a pretrained XTTS-v2, which has been trained on audio at a 22,050 Hz sample rate using optimizer AdamW with betas (0.9, 0.96), epsilon value 1e-8 and weigh decay 1e-2. The weight decay is applied only to weight parameters. The learning rate is 5e-6 with a dynamic schedule and decay factor $\gamma = 0.5$. We used OpenAI's GPT-4o[4] as an LLM component. The full conversation history is sent to the GPT-4o API with default parameters, together with a meta-prompt, containing a description of NICO's personality and an instruction for the model to generate NICO's next response.

3.3 Evaluation Framework

To evaluate our emotion-aware HRI system and quantify the performance of its ERC, TTS and SER subsystems, we built an automated framework (see Figure 2) centered on two questions: 1) does the synthesized speech adhere to the requested emotion, and 2) does emotion control preserve intelligibility (CER/WER)? As ground truth, we use a curated, dialogue-structured derivative of IEMOCAP with 31 dialogues from session 5, which is held out from training, comprising 1596 scripted and improvised utterances. We use IEMOCAP as the basis for evaluation because it provides dyadic, dialogue-structured emotional speech with utterance-level emotion annotations, transcripts, and speaker information. We restrict the evaluation to the four commonly used IEMOCAP emotion classes: angry,

[4]https://openai.com/index/hello-gpt-4o/

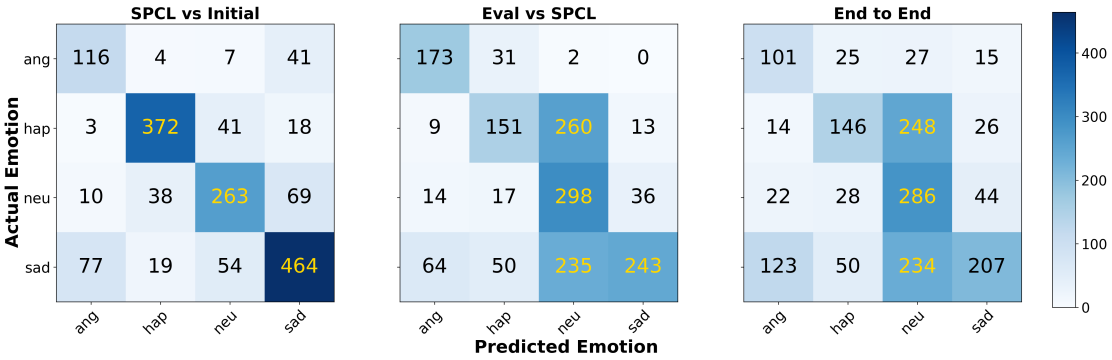

Figure 3: Confusion matrices for evaluation stages. Stage 1 (SpCL vs Initial): evaluation between IEMOCAP ground truth and SpCL prediction. Stage 2 (Eval vs SpCL): between SpCL output and SER evaluation. End-to-End: between IEMOCAP ground truth and SER evaluation.

happy, neutral, and sad. These categories provide the most reliable class support and are widely used in prior ERC/SER work, whereas rarer or more ambiguous labels such as fear, surprise, disgust, frustration, or other would lead to sparse and less stable evaluation. Nevertheless, using IEMOCAP allowed us to ensure the same audio quality for all audio samples, which was useful because TTS pipelines are usually very sensitive to the quality of the reference audio.

4 Results and Analyses

Using the evaluation framework described above, we assessed the HRI pipeline both end-to-end and at intermediate checkpoints to gain finer insight into subsystem performance. We conducted a three-stage assessment: **Stage 1**: accuracy of SpCL emotion predictions compared to the IEMOCAP ground truth; **Stage 2**: accuracy of the XTTS-generated audio emotion, as estimated with SER, compared with SpCL's predicted label; **End-to-End**: the final SER evaluation compared to the IEMOCAP ground truth. ASR provides an estimate of the WER and CER of the generation, as compared to the labels in IEMOCAP. Note that the Stage 2 and end-to-end system performance estimates are underestimated by using an imperfect SER (provided by the Speechbrain toolkit), which has a reported accuracy of 78.7%[5]. Furthermore, the evaluation metrics are sensitive to the choice of reference audio. We obtained an average emotion classification accuracy of 76% for Stage 1, 54% for Stage 2, and 46% for the end-to-end evaluation pipeline. Figure 3 presents the four-class confusion matrices for each of the stages. As expected, Stage 2 and end-to-end performance are lower, due to the imperfect SER that imposes a ceiling of 78.7%. Another compounding factor is that the test set has a majority of *happy* and *sad* emotions, whereas the model's detections of *happy* and *sad* are reduced in favor of the *neutral* emotion, which appears the most in train set (IEMOCAP's sessions 1 to 4). Additionally, we obtained our system performance in terms of intelligibility. Our end-to-end system has a 14% word error rate (WER), and a 13% character error rate (CER). These results reflect a tension between reference-voice clarity and expressive strength. Stronger expressiveness raised CER and WER, whereas weaker expressiveness induced a bias to the *neutral* emotion. We measured the SpCL latency which, depending on chat history length, is 0,61 (±0,8) seconds per every new message (test system: Macbook-M1Pro-16GB). The LLM and TTS produce further latencies, depending on the chosen models and setups.

5 Conclusion and Future Work

We presented a novel modular emotion-aware text-to-speech model integrated via an LLM into an HRI system. We set up an evaluation framework involving a pipeline of emotion classification (ERC), generation (XTTS), and further classification for evaluation (SER). Our newly established baseline provides a challenge for future integrated systems of this kind. The system's modularity allows easy integration into LLM-driven conversational systems, as well as easy switching of components to incorporate future improved models. In future work, the system could be trained end-to-end (while freezing SER and ASR), to prevent overfitting due to limited data while reusing IEMOCAP. Extension

[5]https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

to multimodal input and multi-speaker scenarios is a further development goal.

Acknowledgments

Burak Can Kaplan gratefully acknowledges a Study Abroad Graduate Scholarship by the Ministry of National Education of Türkiye.

References

Christoph Bartneck. 2023. *Godspeed Questionnaire Series: Translations and Usage*, pages 1–35. Springer.

James Betker. 2023. Better speech synthesis through scaling. arXiv:2305.07243.

Cynthia Breazeal. 2000. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. Ph.D. thesis, Department of Electrical Engineering and Computer Science. MIT.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: A massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech 2024*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 6255–6271. Association for Computational Linguistics.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, and Aakanksha Chowdhery, et al. 2023. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. JMLR.org.

Michael A. Goodrich and Alan C. Schultz. 2007. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2023. EmoDiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Peter Hancock, Deborah Billings, Kristin Schaefer, Jessie Chen, Ewart de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53:517–27.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, and Karol Hausman, et al. 2023. Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pages 287–318. PMLR.

Rui Liu, Berrak Sisman, and Haizhou Li. 2021. Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. In *INTERSPEECH*, pages 4648–4652.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *CoRR*, abs/1810.02508.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2024. OpenVoice: Versatile instant voice cloning. arXiv:2312.01479.

Laurel D. Riek. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction Steering Committee*, page 119–136.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. EmoMix: Emotion mixing via diffusion models for emotional speech synthesis. In *Proc. INTERSPEECH*, pages 12–16.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. EmoVoice: LLM-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757. Association for Computing Machinery.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W. Schuller, and Haizhou Li. 2023. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, and Ted Xiao, et al. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229, pages 2165–2183. PMLR.