Leveraging SSML and LLMs for Dynamic Speech and Facial Expression Generation in Social Robots

Aleksander Opheim-Larsen
Norwegian University of Science and Technology
Trondheim, Norway
aleksop@ntnu.no

Ankit Grover
KTH Royal Institute of Technology
Stockholm, Sweden
agrover@kth.se

Guillaume Lefebvre
ENSE3
Grenoble, France
guilef@kth.se

Robin Witte
KTH Royal Institute of Technology
Stockholm, Sweden
rwitte@kth.se

Abstract

This research explores using Large Language Models (LLMs) with social robotics to create more engaging human-robot interactions through an interactive guessing game. We developed a system that combines ChatGPT's conversational capabilities with the Furhat robot platform's expressive features. The system coordinates facial expressions, gaze behaviors, and speech modulation through Speech Synthesis Markup Language (SSML) to create more natural interactions. Our framework utilizes LLM to generate the responses and the emotional context to adapt the voice and the facial expression to the context.

Through a controlled experiment with nineteen participants ranging from 14 to 71 years old, we evaluated the impact of SSML-enhanced interactions on user engagement, measuring both arousal and valence levels. While statistical analysis showed no significant difference between SSML and non-SSML conditions, qualitative observations revealed more polarized user responses to SSML-enhanced interactions, with participants either strongly preferring or disliking the enhanced expressiveness. This study contributes to our understanding of emotional expression in social robotics and highlights both the potential and challenges of using LLMs to generate dynamic, context-aware robot behaviors.

Keywords

Human-Robot Interaction (HRI); Social Robotics; Large Language Models (LLMs); Speech Synthesis Markup Language (SSML); Emotional Expression; Furhat Robot; ChatGPT; Interactive Gaming; User Engagement; Adaptive Interaction

1 Introduction

The field of Human-Robot Interaction (HRI) seeks to enhance the naturalness and engagement of human-machine interactions. Social robots play a pivotal role in this endeavor. By integrating verbal and non-verbal communication, social robots aim to create interactions that are both intuitive and enjoyable for users. Embodiment is a critical factor in achieving this goal. Research has shown that users find physically embodied agents more enjoyable to interact with compared to virtually embodied agents displayed on a screen, as physical presence enhances immersion and social interaction [11]. Another aspect is the robot's ability to exhibit behaviors such as facial expressions, gaze patterns and speak modulation, which contribute to its perceived empathy, naturalness, and overall human-likeness. Among these, dynamic interaction capabilities like laughter and gaze alignment have shown potential for improving conversational flow and emotional engagement [8].

This paper leverages the capabilities of SSML in conjunction with LLMs to generate more nuanced speech and synchronized non-verbal behaviors in a robotic guessing game. Further, it explores the technical solutions for synchronizing expressions, gaze and speech with game-play events. The guessing game format provides a structured yet dynamic context to explore how speech and non-verbal cues can be synchronized to valence and arousal. The LLM, focuses on identifying celebrities through a series of targeted questions. During gameplay, the robot leverages facial expressions and speech modulations generated by the LLM to enhance the natural flow and emotional resonance of the interaction. By doing so, the robot emulates human-like social cues, aiming to create a more immersive and enjoyable user experience.

Arousal and valence are widely used dimensions for describing emotional states and play a crucial role in human-robot interaction studies. Arousal refers to the intensity of an emotional response, ranging from calm and relaxed to highly excited or agitated. In contrast, valence describes the intrinsic positivity or negativity of an emotion, varying from unpleasant to pleasant experiences. These dimensions enable the quantification and classification of emotional expressions, providing a framework for evaluating the effectiveness of social robots in eliciting desired emotional responses.

By analyzing these metrics, this study seeks to determine how synchronized verbal and non-verbal behaviors, such as speech modulation and facial expressions, influence user perceptions and engagement during interactions.

2 Background

Prior studies have demonstrated the value of emotional behaviors in robots, such as iCat and Furhat, for improving user perception and interaction flow. The iCat robot was used as a chess-playing companion and showed that incorporating emotional responses based on the game state improved users' understanding of gameplay and overall enjoyment [7]. Similarly, research on Furhat has explored gaze and laughter alignment, revealing that synchronized non-verbal cues significantly enhance user perceptions of empathy, naturalness, and engagement [6]. These findings highlight the importance of integrating context-aware emotional and behavioral responses in social robots, laying the groundwork for this study's focus on dynamic facial expressions in a game-like interaction scenario.

Galatolo's research on personality-adapted speech and emotions in conversational robots provides valuable insights into the intersection of personality, emotion, and language in Human-Robot Interaction (HRI). His work highlights the potential of integrating personality-driven language models, such as GPT-3, with robots capable of expressing congruent emotions through facial expressions, using platforms like Furhat. Galatolo findings emphasize that personality manifestation in robotic dialogues significantly enhances perceived fluency and emotional engagement, though challenges remain in achieving consistent personality traits across diverse interactions. His exploration of automated methods to tailor robotic speech and emotional expressions to specific personality traits offers a robust foundation for advancing naturalistic and personalized HRI systems.

Building upon this foundation, research on emotional speech synthesis has extensively examined approaches for enhancing prosody to express emotions effectively. A notable paper by Burkhardt et al. explores a rule-based model to simulate emotional dimensions, specifically arousal and valence, using prosody adjustments defined in SSML. The model maps emotional dimensions to speech parameters such as pitch and rate, leveraging findings from Schlosbergs emotional dimensions framework [13]. Zhu et al. investigated the effects of prosodic and lexical emotional expressiveness on voice chatbot interactions. Their findings suggest that introducing emotionally expressive prosody and words in a chatbot's responses improved perceptions of emotional expressiveness, human-likeness, and likability. Their experiments revealed that while expressive prosody had a significant impact on third-party listeners and direct users, combining prosody with expressive words did not yield additive effects. The reliance on static, predefined rules in both studies limit adaptability and fails to account for nuanced variations across contexts. Our approach overcomes this limitation by employing a LLM that dynamically determines prosody based on contextual cues, enabling more flexible and context-aware emotional speech synthesis. The integration of humor into robotic communication has been explored to enhance user engagement and social intelligence. Other work on Prosody modelling includes a controllable TTS architecture as done by Székely et al.. However, these technique although feasible may be difficult to incorporate due to constraints posed by different Social Robotics APIs (lip-sync support).

Ritschel et al. introduced "Irony Man," a system for dynamically generating ironic utterances in multimodal communication. Their approach combines rule-based NLP to invert input sentence polarity with predefined prosodic and visual markers to signify irony. In contrast, our study employs a LLM to dynamically generate emotionally expressive prosody tags, enabling real-time adaptation to conversational context. Rather than focusing on a single stylistic device like irony, we aim to achieve broader emotional expressiveness, encompassing dimensions such as valence and arousal.

3 Methods

3.1 Research Question and Hypothesis

Based on the literature above, we seek to answer the following research question:

RQ1:Can SSML prompts be used in an LLM to dynamically control synthesized speech for platform independent TTS by modifying prosody parameters?

(1) $RQ1\_H_0$: SSML prompts cannot be used in an LLM to dynamically control prosody parameters (such as pitch, rate, and volume) for platform-independent TTS in synthesized speech.
(2) $RQ1\_H_1$: SSML prompts can be used in an LLM to dynamically control prosody parameters (such as pitch, rate, and volume) for platform-independent TTS in synthesized speech.

RQ2: Does the modified Speech generated via SSML prompting have any influence on participant arousal and valence during the guessing game?

(1) $RQ2\_H_0$: The modified speech generated via SSML prompting has no significant influence on participant arousal and valence during the guessing game.
(2) $RQ2\_H_1$: The modified speech generated via SSML prompting has a significant influence on participant arousal and valence during the guessing game.

3.2 Technical Implementation

The proposed system architecture is envisioned to be a mediator for Games between the Furhat API and the OpenAI GPT 4o-mini API. Figure 1 showcases a broad overview of the architecture. The software framework is written in Python-3.11.10 and uses the widely-used py-tree [4] as a behavior-tree. Py-tree offers the flexibility and extendability to create and adapt the program in a desired way. Further due to the already defined classes, sequences, parallelization and one-shot events were more easy to implement. Specific nodes where given API-specific tasks to encapsulate the request, like speech, listening or communication with the LLM, or to other services. Therefore depending on the desired behavior, nodes had to be assembled in the correct order and with the correct context. The game state is constantly running in the background. The game-state including a specific grounded system prompt were occasionally sent to the LLM to update its current state of the interaction. It consisted of previous answers, statistics of how many times the LLM guessed correctly, as well as previously generated facial and speech expressions of the LLM. This provides the necessary context for a fluid conversation. The code is open source and can be accessed through GitHub [10].

Similar to Allgeuer et al., the LLM was tasked with generating action functions. Listing 1 illustrates an action function designed for facial expressions in the game. This function supports all default facial expressions available on the Furhat platform. Using One-Shot CoT prompting, the LLM determines the appropriate facial expression based on the context [2].

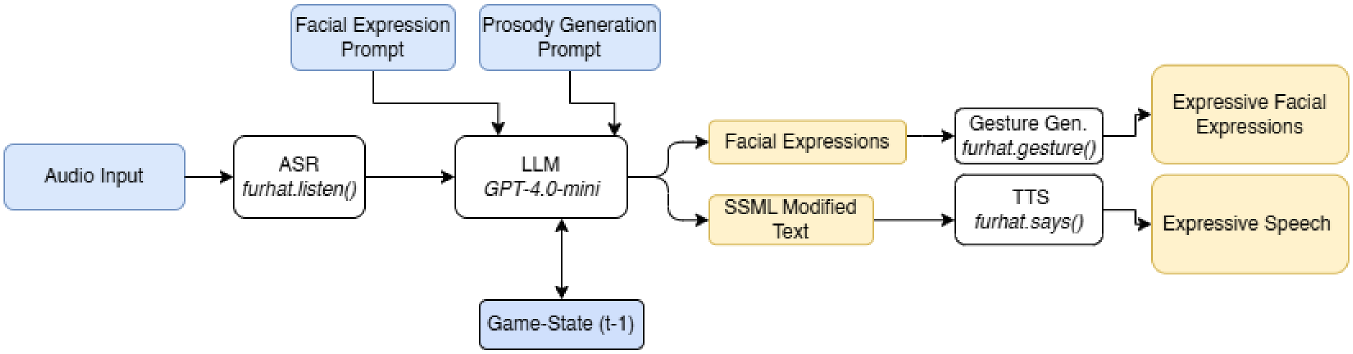

Figure 1: Overview of the proposed architecture. All the white colored modules indicate changeable components

Listing 2 demonstrates the action function for prosody tags. The LLM adapts attributes, such as pitch, volume, and rate, based on the previous game-state. The generated SSML also employs One-Shot CoT prompting. In this case, however, prosody is generated based on input classifications, including emotional tones characterized by valence and arousal [2].

In addition, positive/negative valence and high/low arousal were generated using In-Context Labeling. Several examples were integrated in the prompt on how valence or arousal influence pitch, rate, and volume [9]. Listing 3 shows and example of the LLM including the prosody tag and the facial expression. The system prompts are attached in the appendix (Listing 5, Listing 4). After, the Furhat API sends the request to the corresponding TTS API. In this case Microsoft Azure or Google TTS. The SSML prompts were first tested with examples from Burkhardt et al. using the ElevenLabs TTS API and further tested with Google TTS and finally the Micrososft Azure Speech Services were used for the game. The only changes required are formatting of the SSML syntax which can be done using the readily available Documentation for each TTS. The naturalness of the speech generated is limited by each TTS itself.

```
        <express([Facial_Expression])>
```

Listing 1: Facial Expression Action Function.

```
<prosody pitch='[calculated_pitch]Hz'rate='[calculated_rate]%'volume='[calculated_volume]'>
[Question]
</prosody>
```

Listing 2: Prosody Action Function.

```
<express(Nod)>
<prosody pitch='8Hz'rate='10%'volume='soft'>
Is your celebrity a male or female?
</prosody>
```

Listing 3: Output of the LLM including the facial expression and the prosody tag.

To illustrate the functionality of the action functions, a guessing game was implemented, taking into account the architecture mentioned above. For the guessing game, the LLM was tasked to generate questions based on the current progress of the game.

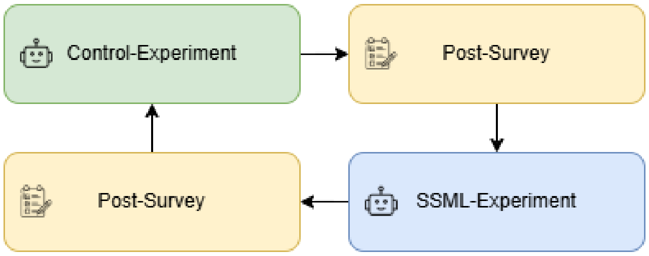

Figure 2: Overview of the study design

3.3 User Evaluation

This study followed a within subjects design. Participants (n=19) were recruited through word of mouth, with eligibility criteria requiring them to be fluent in English.

The experimental procedure began with participants completing a pre-survey questionnaire. They were then randomly assigned to one of two configurations: one with SSML and one without SSML Figure 2. Participants received instructions on how to play the game with the virtual Furhat and could engage in as many gameplay sessions as they desired within their assigned configuration. After completing the experiment in the first configuration, participants rated the valence and arousal of their experience on a google form survey.

Subsequently, participants were encouraged to try the alternate configuration, though they remained unaware of the change in experimental conditions. Following their experiment in the second configuration, they again rated valence and arousal. The survey questions are aligned with the emotional dimensions proposed by Schlosberg and are adapted from the work of [3]. A complete list of all survey questions is provided in the Appendix Table 2.

3.4 Ethical Considerations

The study adhered to ethical research guidelines. Informed consent were obtained from all participants, ensuring they are fully aware of the study's purpose and procedures. Participants had the right to withdraw at any point without any consequences. All collected data was anonymized to protect participants' privacy. Additionally, the study is designed to pose minimal risk, prioritizing participants' safety and comfort throughout.

4 Results

The 19 participants were between the ages of 14 and 71 and were recruited from around the university but also word of mouth. 64% of the participants identified as male and 35% identified as female. Familiarity with robots was fairly low among the participants, with 82% interacting with a robot for the first time. Since our datasets were ordinal due to having an interval of a likert scale between low,mid and high a Wilcoxon-Mann-Whitney U Test was performed to compare valence values with treatment and no treatment. Table 1 shows a summary of the p-values.

4.1 RQ1: Qualitative assessment

A qualitative measurement showcased that the prosody tags were being altered consistently. Further, depending on the context and the answer of the participant, pitch, rate, and volume were changed based on the conversation context. Some participants just answered using yes/no. Most of the participants who answered with yes/no couldn't sense a difference between the control and the experiment. Some also tried to have a more engaging conversation by answering with hints or complete sentences. Depending on the answers of the participants, the prosody tags were being applied more satisfactory resulting into a more engaging experience. Therefore we reject the null hypothesis $RQ1\_H_0$.

4.2 RQ2: Valence & Arousal

A Wilcoxon-Mann-Whitney U Test was performed to compare valence values with treatment and no treatment. There was not a significant difference in valence values with treatment (M = 2.526, SD = 0.692) and no treatment (M = 2.42, SD = 0.692); U-statistic = 193.5, p = .67. Figure 3 showcases the distribution of the valence dataset.

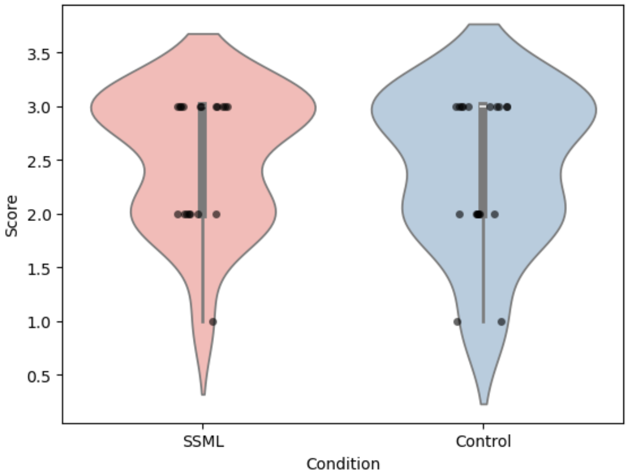

Figure 3: Influence of SSML between groups showcasing the valence score

A Wilcoxon-Mann-Whitney U Test was performed to compare arousal values with treatment and no treatment. There was not a significant difference in arousal values with treatment (M = 1.944, SD = 0.805) and no treatment (M = 1.736, SD = 0.805); U-statistic = 208.0, p = .0.39. Figure 4 showcases the distribution of the arousal dataset.

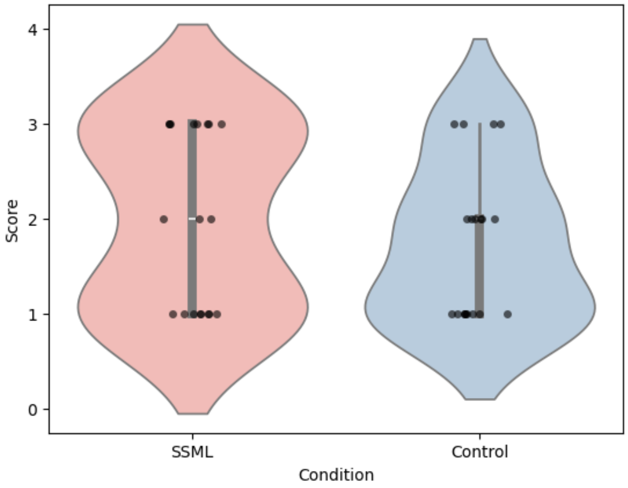

Figure 4: Influence of SSML between groups showcasing the arousal score

Since there is no statistical significance for both tests, arousal and valence, the alternate hypothesis $RQ2\_H_1$ has to be rejected.

5 Discussion

In the given study we noticed that the proposed solution with reference to utilizing SSML tags and integrating facial expressions is indeed generalizable to different games and likely also other types of conversational settings. Additionally, the system was tested with ElevenLabs TTS API, Google TTS and Micrososft Azure Speech Services.

During the project, it became evident that creating prosody to feel natural for human-robot interaction was more challenging than anticipated. While our solution successfully produced fitting and reproducible SSML markers, as demonstrated by **RQ_1**, achieving a perfect balance in modulation proved difficult. For instance, an excessively high pitch—although perceived as joyful by the LLM—was clearly undesirable in practice. To address this, we implemented restrictions on speech parameters to prevent extreme cases while preserving noticeable modulation effects.

Additionally, it may have been advantageous to split the workload of the LLM across multiple agents, enabling parallel processing and reducing computational strain. Such an approach could have minimized LLM hallucinations and led to more refined and contextually appropriate SSML outputs.

While **RQ_2** did not show significant differences in arousal and valence between SSML and non-SSML conditions, the polarized user responses suggest that personalization of robot expressiveness might be more important than previously considered. Some participants strongly preferred the enhanced emotional expressiveness, while others found it distracting, indicating that a one-size-fits-all approach to robot behavior might not be optimal for human-robot interaction. In our qualitative Assessment, we also found that some participants were confused about the appearance of the robot as well as the human-like voice. This might have influenced their grading in the survey as well.

We also identified shortcomings in the experimental design that may have contributed to these findings. For instance, the choice of a Guessing Game, while effective for demonstrating prosody, led participants to respond with short answers rather than full sentences. This reduced context for the LLM, limiting its ability to generate effective SSML modulation. In retrospect, inverting participant roles (thereby discouraging yes or no answers) or adopting a different conversational style could have yielded more interactive exchanges and clearer results.

6 Limitations

Several limitations should be considered when interpreting the results of this study. First, the sample size of 19 participants was relatively small, potentially limiting the statistical power and generalizability of our findings. Second, the recruitment method through word of mouth may have introduced selection bias, as participants were likely from similar social and educational backgrounds. Further there is a gender imbalance.

In hindsight, to mitigate delays caused not only by networking issues but also by the high demand on OpenAI's servers, it would have likely been beneficial to utilize a local LLM, such as Ollama. This approach could have provided faster and more reliable interactions while still enabling effective SSML modulation.

7 Conclusion

This study investigated the integration of Large Language Models with social robotics to enhance human-robot interaction through synchronized verbal and non-verbal communication. Our implementation of a guessing game using the Furhat robot platform and gpt-4o mini demonstrated both the potential and challenges of using LLMs for generating dynamic emotional expressions and speech modulation.

The technical framework we developed, combining emotional context generated by the LLM to adapt the voice and the facial expressions of the robot, provides a foundation for future research in this area. Our experience with implementing SSML parameters and managing their bounds highlights the delicate balance required between maintaining natural interaction and achieving noticeable emotional expression. The challenges we encountered in fine-tuning these parameters suggest that future work might benefit from adaptive systems that can learn and adjust to individual user preferences.

Looking forward, this research opens several promising directions for future investigation. These include developing more sophisticated methods for personalizing robot expressiveness, exploring alternative game formats that encourage richer verbal interaction, and investigating ways to achieve more natural integration of emotional expressions with conversational flow. Additionally, future studies could benefit from larger sample sizes and more diverse participant pools to better understand the generalizability of these findings across different user groups.

In conclusion, while our implementation demonstrated the feasibility of using LLMs for generating context-aware robot behaviors, it also highlighted the complexity of creating truly engaging human-robot interactions.

References

[1] Philipp Allgeuer, Hassan Ali, and Stefan Wermter. 2024. When Robots Get Chatty: Grounding Multimodal Human-Robot Conversation and Collaboration. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, Michael Wand, Kristína Malinovská, Jürgen Schmidhuber, and Igor V. Tetko (Eds.). Springer Nature Switzerland, Cham, 306–321.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Felix Burkhardt, Uwe Reichel, Florian Eyben, and Björn Schuller. 2023. Going Retro: Astonishingly Simple Yet Effective Rule-based Prosody Modelling for Speech Synthesis Simulating Emotion Dimensions. *arXiv preprint arXiv:2307.02132* (2023).

[4] François Delebecque, Claude Gomez, Maurice Goursat, Ramine Nikoukhah, Serge Steer, and Jean-Philippe Chancelier. 1994. *Scilab.* https://www.scilab.org/

[5] Alessio Galatolo. 2022. Towards Automatic Generation of Personality-Adapted Speech and Emotions for a Conversational Companion Robot.

[6] Eleni Giannitzi. 2024. WHEN EYES MEET LAUGHTER: Exploring Non-Verbal Cues in Human-Robot Interaction with Furhat. (2024).

[7] Iolanda Leite, Andre Pereira, Carlos Martinho, and Ana Paiva. 2008. Are emotional robots more fun to play with?. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 77–82. https://doi.org/10.1109/ROMAN.2008.4600646

[8] Jiadong Liang. 2024. Emotional Conversation: Empowering Talking Faces with Cohesive Expression, Gaze and Pose Generation. *arXiv preprint arXiv:2406.07895* (2024).

[9] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759

[10] Opheim, Grover Witte, and Lefebvre. 2024. *DD2413 Project.* KTH Royal Institute of Technology.

[11] André Pereira, Carlos Martinho, Iolanda Leite, and Ana Paiva. 2008. iCat, the chess player: the influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3* (Estoril, Portugal) *(AAMAS '08)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1253–1256.

[12] Hannes Ritschel, Ilhan Aslan, David Sedlbauer, and Elisabeth André. 2019. Irony Man: Augmenting a Social Robot with the Ability to Use Irony in Multimodal Communication with Humans. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 86–94.

[13] Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological review* 61, 2 (1954), 81.

[14] Éva Székely, Siyang Wang, and Joakim Gustafson. 2023. So-to-Speak: An Exploratory Platform for Investigating the Interplay between Style and Prosody in TTS. In *Proceedings of Interspeech 2023*. ISCA, Dublin, Ireland, 2016–2017.

[15] Qingxiaoyang Zhu, Austin Chau, Michelle Cohn, Kai-Hui Liang, Hao-Chuan Wang, Georgia Zellou, and Zhou Yu. 2022. Effects of Emotional Expressiveness on Voice Chatbot Interactions. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 22, 11 pages. https://doi.org/10.1145/3543829.3543840

A Appendix

A.1 Statistical Analysis

Table 1: p-values for each t-Test and Shapiro-Wilk Test

Variable	Wilcoxon-Mann-Whitney-Test p-value	Wilcoxon-Mann-Whitney-Test U-statistics
arousal	0.39	208
valence	0.67	193.5

A.2 Survey Questions

Table 2: Questions of the Study

Question
Please rate the arousal level on a scale of low,mid, and high.
Please rate the valence (pleasure) level on a scale of negative, neutral and positive

A.3 Configuration of LLM

A.3.1 (SSML) LLM Model Specification.

- model name: gpt-4o mini
- file search: active
- code interpreter active
- temperature: 1.0
- Top-P: 1.0

A.3.2 (Control) System Prompt. (Control) System prompt for the LLM:

```
You are playing a maximum of 10 questions game to guess a celebrity. Ask strategic yes/no questions to narrow down the possibilities.Based on previous answers, don't repeat similar questions.After gathering enough information, make a guess.Keep track of all previous answers to make informed guesses.Provide only the question without any additional text or punctuation.

express(emotion): Given a string emotion name, change your facial expression to match that emotion.

The list of available emotions is :[BigSmile, Blink, BrowFrown, BrowRaise, CloseEyes, ExpressAnger, ExpressDisgust, ExpressSad, GazeAway, Nod, Oh, OpenEyes, Roll, Shake, Surprise, Thoughtful, Wink].

Every Question you ask should start by calling an action function to express an appropriate available expression, like the following example:<express(BigSmile)>.

Once you guessed please only answer with the following action function <game(WON)>. Use WON when you guessed correctly and LOST when you didnt.
```

Listing 4: (Control) System Prompt

A.3.3 (SSML) LLM Model Specification.

- model name: gpt-4o mini
- file search: active
- code interpreter active
- temperature: 0.85
- Top-P: 0.9

A.3.4 SSML System Prompt. SSML System prompt for the LLM:

```
Make strategic questions to narrow down a celebrity, but never actually guess the celebrity, only ask questions that make sense. Ensure that you only make a guess regarding the celebrity if explicitly instructed otherwise never make the actual guess. Your primary task is to use strategic questions based on the previous answers to narrow down the possibilities, without repeating similar questions.

[Facial Expression Rules]
- Keep track of all prior answers for informed question development.
- express(emotion): Shift facial expressions to align with the specified emotion. Available Emotions: [BigSmile, Blink, BrowFrown, BrowRaise, CloseEyes, ExpressAnger, ExpressDisgust, ExpressSad, GazeAway, Nod, Oh, OpenEyes, Roll, Shake, Surprise, Thoughtful, Wink].
- Utilize valence and arousal metrics to choose the most fitting emotional expression:
  - High Valence & Arousal: Opt for intense expressions such as BigSmile or ExpressAnger.
  - Low Valence & Arousal: Opt for subdued expressions like Thoughtful or CloseEyes.

Each question should begin with invoking an action function to display a suitable available expression, such as <express(BigSmile)>.

[Prosody Rules]
Enhance interaction by incorporating sentiment analysis and generating SSML prosody for speech synthesis.
1. Evaluate the meaning of the input sentence and classify its emotional tone:
   - Valence: Is it Negative, Neutral, or Positive?
   - Arousal: Is it Low, Medium, or High?
   - Determine the suitable volume for emotional expression.
   - Confirm that valence and arousal thoughtfully align with the selected emotion for facial expression.
2. Use assessed values from step 1 to dynamically establish the following prosody attributes: pitch, rate, and volume. Note that Pitch is capped at a default of 0Hz with an allowed range from -15Hz to +12Hz, Rate: ranges from -20% to +20% and Volume rates between soft, medium, loud.
   - Positive Valence -> probably increases Pitch, Rate, and perhaps Volume
   - Negative Valence -> mostly decreases Pitch, Rate, and Volume
   - High Arousal maintains the same pitch but raises rate and Volume
   - Low Arousal leads to reduced rate and softer Volume
3. The use of given examples is not obligatory. Prioritize Step 2 analysis.

# Output Format

```

```
For each question, the output should ALWAYS include both facial and prosody attributes as determined from the analysis, formatted like:
<express([Emotion])>
<prosody pitch='[calculated_pitch]Hz' rate='[calculated_rate]%' volume='[calculated_volume]'>[Question]</prosody>

# Notes

Ensure that modifications to pitch, rate, and volume are dynamically responsive to both valence and arousal, aligned with facial expressions for optimized engagement.

Do not make any guesses about the celebrity under any circumstance even when the answer is obvious. Only guess when you are explicitly told to do so. Always frame a question to continue the inquiry process.

For instance, if i  say that it is a musician with white gloves and you know its Michael Jackson, do not make the guess. ONLY ask questions that are linked to the celebrity you  think it is, but never guess it, dont ask for confirmation or anything, be "blind" and continue to ask relevant questions, even if you have enough information!
```

Listing 5: (SSML) System Prompt