EmoNews: A Spoken Dialogue System for Expressive News Conversations

Ryuki Matsuura[1*] **Shikhar Bharadwaj**[1*]
Jiarui Liu[1*] **Dhatchi Kunde Govindarajan**[1*]
Carnegie Mellon University
{rmatsuur, sbharad2, jiaruil5, dkundego}@andrew.cmu.edu

Abstract

We develop a task-oriented spoken dialogue system (SDS) that regulates emotional speech based on contextual cues to enable more empathetic news conversations. Despite advancements in emotional text-to-speech (TTS) techniques, task-oriented emotional SDSs remain underexplored due to the compartmentalized nature of SDS and emotional TTS research, as well as the lack of standardized evaluation metrics for social goals. We address these challenges by developing an emotional SDS for news conversations that utilizes a large language model (LLM)-based sentiment analyzer to identify appropriate emotions and PromptTTS to synthesize context-appropriate emotional speech. We also propose subjective evaluation scale for emotional SDSs and judge the emotion regulation performance of the proposed and baseline systems. Experiments showed that our emotional SDS outperformed a baseline system in terms of the emotion regulation and engagement. These results suggest the critical role of speech emotion for more engaging conversations. All our source code is open-sourced.[1]

1 Introduction

In this work, we develop a task-oriented SDS that can regulate emotional TTS based on contextual cues (emotional SDS) to enable more empathetic news conversations. Task-oriented SDSs must balance task- and social-goals to create engaging interactions (Clavel et al., 2022), with emotional speech regulation being crucial among social-goals. For instance, synthesizing "sad" speech for tragic earthquake news can foster user empathy and engagement. Appropriately managing emotional tone can thus enhance both user perception and overall experience (Kurata et al., 2024; Concannon and Tomalin, 2024). To support such needs, the field of affective computing has developed emotional TTS techniques, which generate emotionally expressive speech by adjusting acoustic features like cadence, intensity, and pitch. Recent emotional TTS systems have achieved high-quality oral emotional expressions (Cho and Lee, 2021; Wang et al., 2023; Bott et al., 2024).

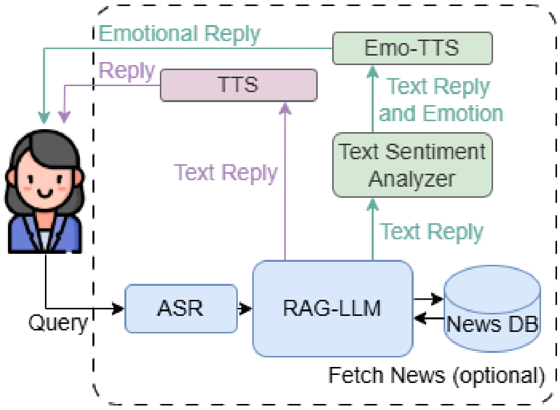

Figure 1: System architecture. Proposed system uses emoTTS and sentiment analyzer.

However, despite these advances, task-oriented emotional SDSs remain underexplored. This is primarily because socio-conversational research has been compartmentalized (Clavel et al., 2022), with SDS and emotional TTS developing separately and lacking an integrated framework. Moreover, evaluating social-goals like emotional speech regulation is difficult (Kurata et al., 2024), as these goals are multidimensional and lack clear definitions, leading to few established evaluation metrics. Thus, the gap between emotional TTS capabilities and their effective integration into SDS highlights

*Equal contributions
[1] https://github.com/dhatchi711/espnet-emotional-news/tree/emo-sds/egs2/emo_news_sds/sds1

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 338–342
Aug 25–27, 2025.

an important area for further research.

We develop a task-oriented emotional SDS and propose its evaluation method. Specifically, we focus on news summarization and Q&A as a target task due to its extensive prior studies. For emotional speech regulation, we adopt a cascade SDS architecture. We employ a PromptTTS (Guo et al., 2022) fine-tuned on the *ESD* dataset (Zhou et al., 2022) as our emotional TTS model. For evaluation, we use an empathy scale originally (Concannon and Tomalin, 2024) and assess the SDS's ability to regulate emotional speech. Additionally, we manually evaluate both the system's emotional speech regulation and task achievement (Walker et al., 1997), comparing it with SDSs that employ non-emotional TTS.

Through this study, we contribute to the studies on socio-conversational system by: (i) providing a method for developing emotional SDS; and (ii) proposesing an evaluation method of emotional SDS.

2 System Design and Method

As depicted in Figure 1, we develop an emotional SDS for news conversations using a cascade architecture, building on a strong baseline (Arora et al., 2025) by adding emotional awareness via sentiment-guided synthesis. Below, we describe the core components of both systems.

The baseline system includes three modules: ASR, LLM, and TTS. The ASR transcribes user speech to text, which is encoded and compared against a News Database for relevant article retrieval. The LLM generates a response based on both the transcript and retrieved news snippets, and the TTS outputs spoken responses in a default tone. We utilize Retrieval Augmented Generation (RAG) in our system. The core ASR and RAG-LLM module are shared: the ASR transcript is passed to a RAG language model that selectively retrieves news to ground its replies, using dynamic in-context prompting for adaptability.

Our proposed system enhances the baseline with a Sentiment Analyzer that infers the emotional tone (`neutral`, `happy`, `sad`, `angry`, or `surprised`) from the LLM's text response. The emotion tag is fed to PromptTTS, an emotional TTS module that conditions speech synthesis on both text and emotion, producing expressive and empathetic responses. Compared to the emotionally neutral baseline, our system delivers more human-like, engaging interactions through sentiment understanding and emotional prosody.

3 Experiments

To evaluate the extent to which the proposed method can control proper speech emotion, we compared it with a baseline system using human subjective judgments.

3.1 Datasets

For emotional TTS fine-tuning, we used the English portion of the ESD dataset (Zhou et al., 2022), splitting 17,500 utterances into training, validation, and evaluation subsets across five emotions. For sentiment analyzer fine-tuning, we used GoodNewsEveryone (Bostan et al., 2020) and GoEmotions (Demszky et al., 2020), mapping their emotion tags to five target categories following Koufakou et al. (2024). As the news database for retrieval, we used Free News[2], filtering for English articles and embedding news titles with Chroma[3] and Sentence Transformers (Reimers and Gurevych, 2019).

3.2 System Setups

Proposed System We used Whisper Large for ASR, LLaMA 3.2 1B for the language model, and a sentence transformer for retrieving the top 1 relevant news. For emotional TTS, we fine-tuned PromptTTS (pretrained on LJSpeech). Our preliminary analysis showed that its quality was comparable to FastSpeech (Ren et al., 2019) and VITS (Kim et al., 2021) in terms of UTMOS, DNSMOS, PLCMOS, and WER, and qualitative analysis confirmed clear emotional variation. For the sentiment analyzer, we fine-tuned a distilled RoBERTa model (batch size 8, learning rate 0.00001, 4 epochs) after finding that prompt-based LLM approaches tended to over-predict sadness and surprise, achieving better performance than Koufakou et al. (2024).

Baseline System The baseline system shared the same modules as the proposed system, except the sentiment analyzer and a VITS

[2]https://github.com/Webhose/free-news-datasets
[3]https://www.trychroma.com/

Metric	Item
RAG Evaluation	The system was helpful to understand the retrieved news.
Task Achievement 1 (Usefulness)	The news that the system retrieved matched the information you wanted to know.
Task Achievement 2 (Response Consistency)	The system consistently responded according to the retrieved news.
Speech Emotion Appropriateness	The system seemed to vary its emotional state of speech to demonstrate expressiveness and modify its responses to accommodate the mood of the context.
Engagement	Did you have favorable feelings toward the one you were talking to? Did you feel a sense of familiarity with the one you were talking to? Did you feel that the system you were talking to understood the mood of contexts?

Table 1: Emotional SDS Evaluation Questionnaire.

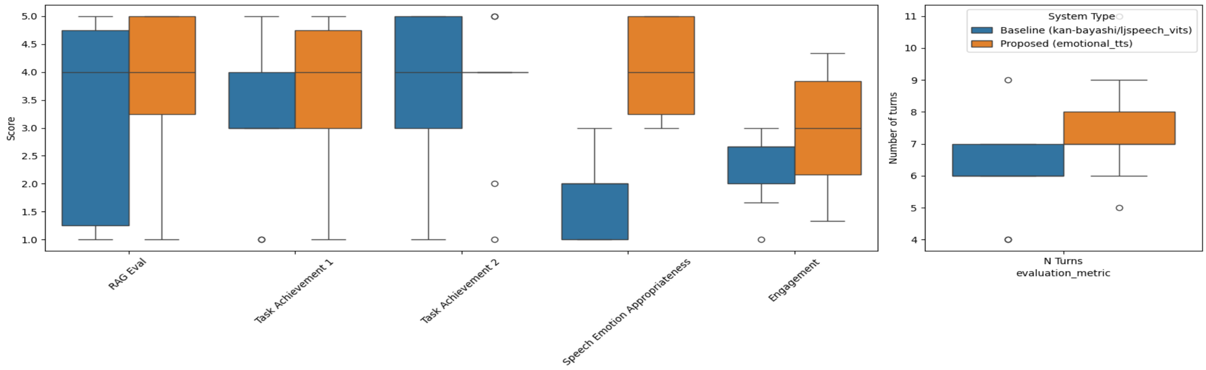

Figure 2: Comparison of Evaluation Metrics by System Type.

Metric	**U**	**p-value**	**Cohen's d**
RAG Evaluation	42.5	0.580	0.301
Task Achievement 1	42.0	0.558	0.206
Task achievement 2	51.0	0.968	0.070
Speech Emotion Appropriateness	1.5	< 0.001	3.070
Engagement	27.5	0.090	0.824
N Turn	26.5	0.073	0.831

Table 2: Statistical Comparison Between Baseline and Proposed Systems

model pre-trained on LJSpeech instead of emotional TTS.

3.3 Metrics

We create a seven-item questionnaire in Table 1, using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The first item assesses RAG performance on relevance and coherence, while the second and third address task achievement (Walker et al., 1997): system helpfulness in understanding retrieved news and consistency of responses. The fourth item measures speech emotion appropriateness, adapted from empathy scales for dialogue systems (Concannon and Tomalin, 2024). The last three items assess user engagement, based on Kurata et al.'s questionnaire (Kurata et al., 2024). We also recorded the number of SDS turns as an additional engagement indicator (Aoyama et al., 2024).

3.4 Evaluation Procedure

We collect 20 conversation samples by conducting 10 dialogues with each system. To avoid bias, emotion tags predicted by the sentiment analyzer were hidden from the SDS interface. We test differences in mean scores using Mann-Whitney U tests ($\alpha = .05$) due to the small sample size, and calculate Cohen's d for effect sizes (Cohen, 2013). We assess the internal consistency of the three engagement items using Cronbach's alpha, which was .860, indicating substantial reliability; thus, we averaged them into a single engagement score.

3.5 Results and Discussion

Figure 2 shows the boxplots of human-judgment scores. The proposed system significantly outperformed the baseline in speech emotion appropriateness with a large effect size ($d = 3.070$; 4.100 vs. 1.700), confirming its ability to control emotions according to context. Although engagement scores and the number of turns showed no significant differences, both had large effect sizes (d = 0.824, 0.831), suggesting that emotional control may promote more engaging conversations (Concannon and Tomalin, 2024). However, the mean engagement score remained moderate (around 3), possibly due to abrupt, discrete emotional shifts without considering prior conversational context. Finally, no significant differences were observed in RAG performance or task achievement, and both systems scored around 3, indicating room for improvement in task-goal fulfillment.

4 Conclusion

We presented an emotional SDS that enhances empathetic interactions in task-oriented news conversations by combining a sentiment analyzer with PromptTTS for dynamic emotional speech generation. Our system integrates sentiment-driven emotional control within a prompt-based architecture, improving emotion appropriateness and engagement without compromising task performance. Its modular design enables easy adaptation to other domains, supporting broader development of emotionally aware conversational agents.

References

Takumi Aoyama, Joseph S. Yamazaki, Sachiko Nakamura, Alyssa Vuogan, Hyejin An, Claudia J. Kim, and Ali H. Al-Hoorie. 2024. Conceptualization and Operationalization in L2 Task Engagement Research: Taking Stock and Moving Forward. *Language Teaching*, 57(4):597–601.

Siddhant Arora, Yifan Peng, Jiatong Shi, Jinchuan Tian, William Chen, Shikhar Bharadwaj, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Shuichiro Shimizu, Vaibhav Srivastav, and Shinji Watanabe. 2025. Espnet-sds: Unified toolkit and demo for spoken dialogue systems. In *NAACL Demo*.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Thomas Bott, Florian Lux, and Ngoc Thang Vu. 2024. Controlling Emotion in Text-to-Speech with Natural Language Prompts. *arXiv preprint*. ArXiv:2406.06406 [cs] version: 1.

Sungjae Cho and Soo-Young Lee. 2021. Multi-speaker emotional text-to-speech synthesizer. In *Proceedings of Interspeech 2021*, pages 2337–2338, Brno, Czech Republic. ISCA.

Chloé Clavel, Matthieu Labeau, and Justine Cassell. 2022. Socio-Conversational Systems: Three Challenges at the Crossroads of Fields. *Frontiers in Robotics and AI*, 9. Publisher: Frontiers.

Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Routledge, New York.

Shauna Concannon and Marcus Tomalin. 2024. Measuring Perceived Empathy in Dialogue Systems. *AI & SOCIETY*, 39(5):2233–2247.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2022. PromptTTS: Controllable Text-to-Speech with Text Descriptions. *arXiv preprint*. ArXiv:2211.12171 [eess].

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *arXiv preprint*. ArXiv:2106.06103 [cs].

Anna Koufakou, Elijah Nieves, and John Peller. 2024. Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 196–206, Miami, Florida, USA. Association for Computational Linguistics.

Fuma Kurata, Mao Saeki, Masaki Eguchi, Shungo Suzuki, Hiroaki Takatsu, and Yoichi Matsuyama. 2024. Development and Validation of Engagement and Rapport Scales for Evaluating User Experience in Multimodal Dialogue Systems. In *The Proceedings of The 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. *arXiv preprint*. ArXiv:1905.09263 [cs].

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280.

Shijun Wang, Jón Guðnason, and Damian Borth. 2023. Fine-grained emotional control of text-to-speech: Learning to rank inter- and intra-class emotion intensities. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional Voice Conversion: Theory, Databases and ESD. *arXiv preprint*. ArXiv:2105.14762 [cs].