The 12th ISCA-SAC Doctoral Consortium (DC2026)
26 September 2026, Sydney, Australia

Natural Language as the Interface to Speaking Style: Describing and Controlling *How* Speech Is Delivered

Shreeram Suresh Chandra [1,2]

[1] Johns Hopkins University, USA
[2] The University of Texas at Dallas, USA
schand30@jh.edu, shreeramsureshchandra@gmail.com

Abstract

The expression and perception of human speaking styles are inherently continuous and unbounded in nature. Yet conventional approaches collapse this continuum into discrete labels or rigid taxonomies that are costly to annotate and too coarse to capture how speech actually varies. Natural language offers a compelling alternative: it is open-ended and compositional enough to describe speaking styles at the resolution at which they occur, and it provides a shared medium for both perceiving and producing them. In my work, I explore natural language as an interface to build paralinguistic-informed speech frameworks that can improve speech understanding and expressive speech generation. In the scope of this paper, we present the research questions that motivate us, our contributions and a brief outline on future work.

Index Terms: computational paralinguistics, speaking style captioning, emotion, speech generation

1. Motivation and research questions

Until recently, computational models of speech paralinguistics represented emotion as a small set of discrete categories (happiness, sadness, neutral, anger etc.), grounded in the categorical theories of Ekman [1] and Plutchik [2]. This paradigm is convenient but fundamentally lossy: human expressivity is continuous, compositional, and high-dimensional, and collapsing it into a handful of labels limits how faithfully we can either understand or generate expressive speech.

To move beyond this, a growing body of work has adopted natural language as the medium for paralinguistic information, given that it provides a human-interpretable, open-ended, and continuous representation capable of describing *how* speech is delivered at the resolution at which it actually varies. My thesis investigates natural language as an *interface* to paralinguistic speech: a single medium through which expressive speech can be aligned with language, described, generated under linguistic control, and rigorously evaluated.

Developing this interface is currently held back by two broad, cross-cutting challenges.

(C1) Data scarcity: The lack of diverse, scalable speech corpora with natural-language paralinguistic annotations

(C2) Evaluation: Standard metrics fail to assess open-ended outputs (e.g., free-form captions, style-conditioned audio), lacking reliable measures for quality and faithfulness.

These challenges span our research: data scarcity is addressed by introducing new datasets (RQ1, RQ3) and developing an annotation-free method (RQ2), while open-ended outputs necessitate novel evaluation paradigms (RQ1, RQ3, RQ4).

Research Questions (RQs) :

RQ1 (Speech-natural language alignment). How can we learn joint speech–language embedding spaces that align emotional speech with natural-language style descriptions while preserving the continuous, ordinal structure of emotion, and use them for speech-emotion retrieval?

RQ2 (Natural language guided speech generation). How can natural language serve as a control medium for expressive speech generation without relying on scarce, explicitly labeled style data?

RQ3 (Speech understanding via Natural Language). Speech large language models (SpeechLLMs) take speech in and produce free-form text out. How can we steer them to caption paralinguistic speaking style on demand, and how do we evaluate generations that are open-ended rather than categorical?

RQ4 (Benchmarking Natural language guided speech generation). As natural-language-guided speech generation systems proliferate, how do we reliably benchmark their controllability and faithfulness, measuring not just whether instructions are followed, but whether outputs are grounded in the signal?

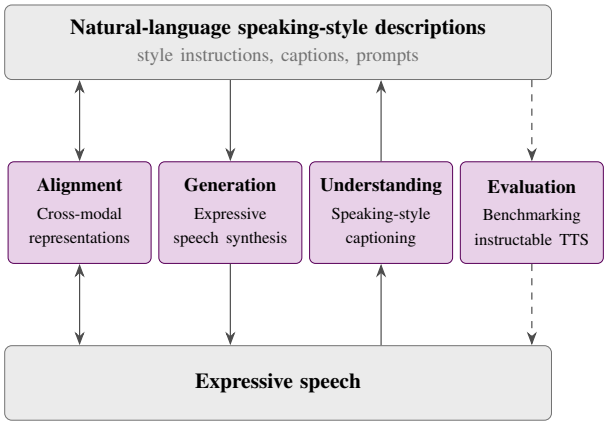

Figure 1: *Natural language as an interface to expressive speech, explored across four dimensions - cross modal alignment, expressive speech synthesis, speaking style captioning, and the evaluation of instruction-guided TTS*

2. Key contributions

My thesis explores the connection between natural-language speaking-style descriptions and expressive speech across several dimensions and tasks as shown in Fig. 1. In each dimension, we pose key questions (KQs) that push the field forward, while the setup remains constant: the models operate over speech and natural-language speaking-style descriptions, and the goal is to tie the two together faithfully.

10.21437/DC.2026-8

The first dimension is **alignment**: how to build joint embedding spaces that align the continuums of dimensional speech emotion and natural-language speaking-style descriptions. The second extends this connection to **generation**, using natural language to control expressive synthesis, with a focus on conversational speech. The third turns to **speech understanding**, examining how large audio–language models perform speaking-style captioning. Finally, we shift to **evaluation**, benchmarking the controllability of instructable text-to-speech (TTS) models.

3. Methods

3.1. Cross-modal alignment

KQ: How can we build a joint embedding space for natural-language speaking-style captions and speech emotion that preserves the continuous, ordinal structure of emotion rather than collapsing it into discrete categories?

In the first dimension, we examine contrastive language–audio pre-training (CLAP) as a means of aligning natural-language speaking-style descriptions with speech emotion. Prior methods naïvely align audio samples with their corresponding text prompts using a symmetric cross-entropy loss [3, 4, 5, 6], treating each pair as an isolated positive. This ignores the relative ordering between emotions and yields a wide audio–text modality gap, which in turn limits fine-grained, inter-emotion understanding. To address this, we adopt the Rank-N-Contrast objective [7] to learn an *ordered* representation space, supervising the geometry with dimensional emotion attributes (valence and arousal) so that embeddings are arranged by emotional similarity rather than merely clustered by pair.

The proposed method, EmotionRankCLAP [8], substantially narrows the modality gap: it reduces the Maximum Mean Discrepancy (MMD) [9] between the audio and text spaces to 0.087 and the Wasserstein distance [10] to 0.065, compared to $0.096/0.180$ for CLAP-SCE and $0.444/0.417$ for SupCon-CLAP (lower is better). The same ordinal structure transfers to downstream speech-emotion retrieval, where EmotionRankCLAP achieves the highest arousal and valence ordinal consistency (Kendall's τ of 0.552 and 0.616), outperforming all CLAP baselines including CLAP-SCE ($0.492/0.505$) and prior models such as CLAP4emo ($0.284/0.533$) and ParaCLAP ($0.283/0.217$).

3.2. Expressive speech generation

KQ: In a natural conversational setting, can we utilize the spoken text itself as context to control the speaking style of a text-to-speech (TTS) system?

In the second dimension, we extend the connection from understanding to generation. Prior emotional TTS methods produce well-articulated emotion only by relying on explicit emotion labels or a stylistically matched reference utterance—neither of which is available in conversational use. We develop TEMOTTS [11], a reference-free framework built on global-style-token codebooks [12] and FastSpeech2 [13], premised on the natural correlation between spoken words and affective prosody: the text itself supplies the emotional control signal.

Among reference-free models, TEMOTTS is both more intelligible and more emotionally faithful. It attains the lowest error rates (CER 4.08%, WER 13.11% versus $8.22/18.57$ for VITS [14] and $6.88/18.34$ for FastSpeech2, is chosen best in 80.95% of Best–Worst Scaling comparisons for emotion-faithfulness (worst in only 1.90%), and achieves the highest MOS (3.75 ± 0.06).

3.3. Speaking style captioning

KQ: Can large audio-language models describe speaking style on demand, generating captions that target a requested stylistic dimension and stay grounded in the speech signal?

Next, we focus on speech paralinguistics understanding. Although large audio-language models excel at tasks like automatic speech recognition and Q&A, paralinguistic captioning lacks a unified task formulation or evaluation paradigm. This ambiguity exists because speaking style is subjective and spans interacting factors like speaker identity, affect, prosody, and contextual delivery. To address this, we reformulate speaking-style captioning (SSC) as an instruction-following audio-language task. By using explicit instructions to specify which aspect of style the model should describe, SSC becomes controllable and tailored to downstream intent, rather than producing a single, generic description.

We introduce *StyleInstructCaps*, the first standardized benchmark for controllable SSC. It evaluates captions across five axes: (1) **metadata groundedness**, (2) **hallucination**, (3) **instruction-following ability**, (4) **speaker-style consistency**, and (5) **generalization to unseen speech and prompts**.

To support this benchmark, we built a multi-style corpus expanding on ParaSpeechCaps (PSC) [15]. This includes 2,900 hours of speech where each utterance is paired with six complementary caption types: speaker identity, contextual delivery, emotion, pragmatics, listener perception, and a holistic summary. These parallel annotations allow us to test whether models can isolate individual stylistic dimensions, integrate multiple cues when required, and generalize across diverse conditions.

Evaluating nine SOTA audio-language models reveals that task-aligned, instruction-diverse training is critical for SSC. Beyond generic large-scale pretraining, this targeted training improves metadata groundedness, reduces hallucination, strengthens speaker-style consistency, and enables robust generalization to new speakers, datasets, and prompting conditions.

4. Current and future research directions

4.1. Evaluating instruction-guided TTS

KQ: As natural-language prompts become the dominant control interface for expressive TTS, how do we reliably evaluate whether generated speech actually realizes the emotion described in the prompt?

In the final dimension, we turn from building natural language-guided systems to *measuring* them. Natural-language prompts offer a continuous, free-form control surface that is well suited to emotional speech, but their open-endedness makes evaluation hard: there is no fixed label to check against, and subjective listening tests are costly and slow. I plan to develop *InstructEmotionEval*, a benchmark that quantifies cross-modal emotion control in instruction-guided TTS along four complementary axes. The benchmark spans four axes: **cross-modal alignment** (does generated speech match the prompted emotion, via CLAP-based retrieval and cosine similarity), **emotion monotonicity** (do ordered prompts yield monotonically ordered acoustic responses), **disentanglement** (does changing one emotion axis leak into others), and **emotion coverage** (how much of the valence–arousal–dominance space is reachable via prompts). Together, these probe whether language is not merely an expressive control interface, but also a *faithful* and *complete* one, thereby extending the theme of groundedness in this thesis.

5. Generative AI Use Disclosure

Generative AI tools were employed solely for language polishing of text written by the authors. These tools were not used to generate scientific content, results, experimental designs, analyses, or conclusions. All authors are responsible for the full content of this paper and consent to its submission.

6. References

[1] P. Ekman, "Are there basic emotions?" 1992.

[2] R. Plutchik, *The emotions*. Bloomsbury Publishing PLC, 1991.

[3] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, "Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 021–10 025.

[4] H. Dhamyal, B. Elizalde, S. Deshmukh, H. Wang, B. Raj, and R. Singh, "Prompting audios using acoustic properties for emotion representation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 936–11 940.

[5] W.-C. Lin, S. Ghaffarzadegan, L. Bondi, A. Kumar, S. Das, and H.-H. Wu, "Clap4emo: Chatgpt-assisted speech emotion retrieval with natural language supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 791–11 795.

[6] X. Jing, A. Triantafyllopoulos, and B. Schuller, "Paraclap–towards a general language-audio model for computational paralinguistic tasks," *arXiv preprint arXiv:2406.07203*, 2024.

[7] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi, "Rank-n-contrast: Learning continuous representations for regression," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 882–17 903, 2023.

[8] S. S. Chandra, L. Goncalves, J. Lu, C. Busso, and B. Sisman, "EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion via Rank-N-Contrast," in *Interspeech 2025*, 2025, pp. 3000–3004.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[10] G. Peyré and M. Cuturi, *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[11] S. S. Chandra, Z. Du, and B. Sisman, "Exploring speech style spaces with language models: Emotional tts without emotion labels," in *Proc. odyssey 2024*, 2024, pp. 194–200.

[12] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[13] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*.

[14] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 5530–5540.

[15] A. Diwan, Z. Zheng, D. Harwath, and E. Choi, "Scaling rich style-prompted text-to-speech datasets," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 3639–3659.