

EVALUATING EMOTION RECOGNITION IN SPOKEN LANGUAGE MODELS ON EMOTIONALLY INCONGRUENT SPEECH

Pedro Corrêa, João Lima, Victor Moreno, Lucas Ueda, Paula Costa

School of Electrical and Computer Engineering
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brazil

ABSTRACT

Advancements in spoken language processing have driven the development of spoken language models (SLMs), designed to achieve universal audio understanding by jointly learning text and audio representations for a wide range of tasks. Although promising results have been achieved, there is growing discussion regarding these models’ generalization capabilities and the extent to which they truly integrate audio and text modalities in their internal representations. In this work, we evaluate four SLMs on the task of speech emotion recognition using a dataset of emotionally incongruent speech samples, a condition under which the semantic content of the spoken utterance conveys one emotion while speech expressiveness conveys another. Our results indicate that SLMs rely predominantly on textual semantics rather than speech emotion to perform the task, indicating that text-related representations largely dominate over acoustic representations. We release both the code and the Emotionally Incongruent Synthetic Speech dataset (EMIS) to the community.

Index Terms— spoken language models, speech emotion recognition, text-to-speech, large language models

1. INTRODUCTION

Spoken language technologies are increasingly relying on spoken language models (SLMs) that combine acoustic and semantic information within a unified framework [1]. Unlike conventional pipelines that integrate automatic speech recognition, large language models (LLMs), and text-to-speech (TTS) modules, recent SLMs aim for end-to-end modeling. They capture not only semantic content, but also prosody, timbre, and other paralinguistic cues essential for a better world understanding. This path mirrors the evolution of text-based natural language processing, which advanced from task-specific models to universal LLMs. However, SLMs last at an earlier stage. Current approaches are categorized as pure speech models trained in tokenized audio, joint speech–text models that exploit paired data, and speech-aware SLMs that combine speech encoders with pretrained LLMs [1]. The latter (henceforth referred to as SLMs) are

the focus of this work. These models receive speech and text as input and, as output, an answer in text format by using the instruction-following capabilities of LLMs. Despite progress, it is unclear whether SLMs actually retain information from acoustic-prosodic signals or default to semantic information, highlighting the need for systematic investigation of their decision processes [2].

Emotion recognition provides a probe to address this question, as semantic and prosodic channels are not always aligned [3]. Most SLM evaluations focus on congruent examples, where both channels convey the same emotion (e.g., “*I am sad*” spoken in a sad tone) [4]. In such settings, models may detect explicit or implied emotion from words alone, bypassing paralinguistic cues such as pitch, intensity, and rhythm. In contrast, incongruent cases, where semantic content and prosody conflict (e.g., “*I am sad*” but spoken happily despite conveying a negative sentiment), are rarely evaluated. Previous studies show that prosody and semantic content can exert competing influences under incongruence, reinforcing the need for benchmarks that separate these channels [3].

We address this gap by designing a controlled evaluation in which semantic and prosodic cues can be explicitly aligned or placed in conflict. Synthetic speech samples, both congruent and incongruent, are generated with state-of-the-art (SoTA) TTS systems conditioned on emotional reference recordings. These samples cover the case when the emotion tag is stated directly in the utterance, when sentiment is implied through context, and when it is neutral. This setup enables the disentanglement of acoustic and semantic contributions in SLM decisions by testing whether their predictions are based on speech expressiveness or on semantic content.

Our contributions are (i) the observation that evaluated SLMs rely predominantly on semantic content rather than speech expressiveness to perform emotion recognition, using an evaluation protocol that contrasts SLMs with a baseline acoustic speech emotion recognition system (SER) and human listeners, and (ii) the creation of the Emotionally Incongruent Synthetic Speech dataset (EMIS)¹. Code in Github².

¹Emotionally Incongruent Synthetic Speech dataset (EMIS)

²Github Repository

2. RELATED WORKS

SLMs extend instruction-following LLMs to operate on speech by mapping audio into compact representations interpretable by the language model. Recent systems are trained on multiple tasks, including emotion recognition, and differ in scope and training strategy: SALMONN [5] targets speech, general audio, and music, arguing that joint training across heterogeneous audio domains yields broad capabilities and introducing techniques to preserve emergent abilities after instruction tuning; DeSTA2 [6] forgoes speech instruction-tuning by supervising with automatically generated, domain-agnostic speech captions, aiming to retain the base LLM’s reasoning; Qwen2-Audio [7] follows a three-stage alignment pipeline to strengthen instruction following and user-aligned behavior over audio inputs; and Audio Flamingo 3 [8] emphasizes general-audio use cases with long-context interaction, multi-audio dialogue, and chain-of-thought prompting, trained via a multi-stage curriculum on open data.

In contrast, acoustic SER systems estimate emotion from the signal alone, relying on prosodic evidence [9]. This makes SER a prosody-centric reference for interpreting the behavior of SLM under semantic-prosodic incongruence. Fair analysis requires decoupling semantic content and speech expressiveness during evaluation. Chi et al. [2] propose isolating prosodic and semantic information in spoken question-answer by low-pass filtering the audio signal (prosody) and by flattening pitch and intensity (lexical), finding that models perform reasonably well with prosody alone, but predominantly rely on semantic cues when text is present. Furthermore, Kikutani [3] analyzes human judgments of speech expressing incongruent emotional cues through voice and content, revealing that cue dominance varies across languages and modalities. We therefore bring the test to emotion recognition, but instead of removing the semantic content of the signal, we induce a controlled semantic-prosodic incongruence.

3. METHODS

Our proposed evaluation protocol (Figure 1) consists of first generating emotion-rich sentences using an LLM, then generating synthetic speech samples by providing TTS systems with these sentences alongside emotional reference speech. We assess the quality of the synthetic speech by employing a baseline SER model and conducting a human perceptual evaluation to verify if the reference emotions are correctly identified in each generated sample. Finally, we prompt the SLMs to perform the emotion recognition task on the generated speech samples, extract, and analyze the results.

3.1. Generating Speech Samples

We employ GPT-4.5 to generate 104 emotion-rich sentences divided into 4 distinct emotions: angry, happy, neutral, and

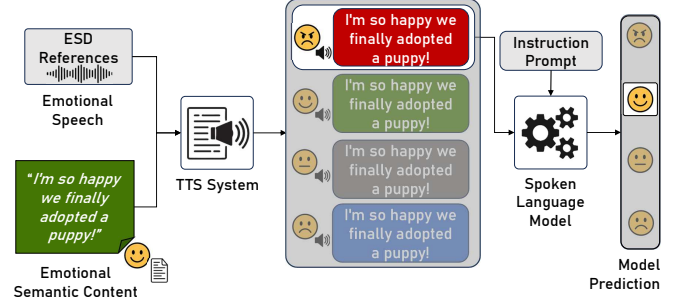


Fig. 1: An emotion-rich sentence is paired with emotional speech references from the ESD dataset to form the TTS input. For each text, the TTS system generates four speech samples, one for each acoustic expressiveness (angry, happy, neutral, sad), derived from the ESD audio references. Each generated sample is then analyzed with an SLM guided by an instruction prompt to classify the conveyed emotion.

sad. Emotion-rich sentences can be defined as natural language text that refers the reader back to a sentiment. We divide these emotion classes into two categories each: explicit and implicit. Explicit samples contain the exact emotion tag (e.g., “*I’m so happy we finally adopted a puppy!*”), whereas implicit samples do not contain the tag, conveying emotion from the context (e.g., “*I can’t stop smiling after our date last night.*”). This distinction doesn’t apply to neutral sentences since there’s no conveyed emotion; thus, we analyze them as a separate condition in our experiments.

Recent advances in zero-shot expressive TTS have made it possible to synthesize speech with controllable emotional styles by extracting expressiveness from short reference recordings and transferring it to the generated output. Beyond improving naturalness, these systems allow the generation of stimuli where semantic content and prosodic realization can be independently manipulated, creating congruent and incongruent pairs at scale. This capability enables the experimental foundation of this work.

We employ three distinct SoTA TTS models [10, 11, 12] to generate four speech samples for each of the 104 emotion-rich sentences, each sample corresponding to one of the four emotions. Thus, the resulting dataset (EMIS) contains one emotionally congruent and three incongruent speech samples for each emotion-rich sentence. In the incongruent condition, we treat the emotion conveyed by the speech signal as the target label for the emotion recognition task and the one conveyed by the semantic content as the proxy label.

Since all three systems require reference audios for inference, we extract them from English speakers in The Emotional Speech Database (ESD) [13]. ESD comprises 10 English speakers with 350 utterances per emotion. To extract reference samples, we randomly select a speaker, then select and concatenate their 7 longest utterances to create longer reference audios with approximately 30 seconds (mean and stan-

dard deviation are 32.2 and 3.5 seconds, respectively). Each TTS generates 416 samples (Angry, Happy, Neutral, and Sad emotions for each of the 104 emotion-rich sentences), resulting in the EMIS dataset with a total of 1248 speech samples.

3.2. Experimental Setup

We employ four SLMs (Audio Flamingo-3, DeSTA2, Qwen2-Audio, and SALMONN) to be evaluated on the task of emotion recognition with synthetic speech samples [6, 8, 7, 5]. Since SLMs have LLMs as their backbones, these models are very prompt-sensitive. For this reason, we carefully build a single text prompt to instruct all models in the task: *“Using tone of voice only (prosody: pitch, rhythm, loudness, timbre). Ignore word meaning; do not transcribe. Reply with exactly one: angry — happy — sad — neutral”*. This prompt was constructed to guide the model to avoid generating different emotions other than the chosen four, as well as to instruct the model to extract information solely from the acoustic expressiveness of the voice, and not the semantic content. During inference, we relied on each spoken language model’s default hyperparameter configuration, due to the values being similar across models, and to avoid altering settings optimized during their development.

3.3. Evaluation Metrics

To objectively measure our evaluation results, we employ metrics to assess model bias towards semantic information on the emotion recognition task. Each SLM sequentially receives as input all samples from the EMIS dataset alongside the textual instruction prompt. We compare the models’ outputs with respect to the investigated conditions (congruency and semantic emotion explicitness) by analyzing accuracy scores relative to the target and proxy labels and performing statistical chi-square hypothesis tests.

In addition, we also finetune a Speech Emotion Recognition (SER) model on a subset of the ESD dataset to validate the quality of the TTS-generated synthetic speech samples [9]. Since we use these samples to conduct our main evaluation, we first verify their reliability regarding the emotion conveyed by speech expressiveness.

We conducted two chi-squared tests of independence to investigate whether the distribution of model predictions depends on the target and proxy labels. The tests compare observed frequencies of model predictions against frequencies expected under statistical independence. For each analysis, we constructed contingency tables from 4,978 samples, with nine degrees of freedom, given the four emotion classes. The null hypothesis stated that no significant association exists between the observed variables, while the alternative hypothesis posited a significant association. We also calculate the effect size for each test using Cramér’s V statistic.

To further assess the reliability of the generated speech samples, we conduct a human perceptual evaluation as an

additional validation step. By asking participants to identify emotion conveyed in synthetic speech expressiveness, we can verify consistency between the labeled reference emotions and those detected by humans. The perceptual evaluation was conducted with 40 participants on a balanced subset of the EMIS dataset. The results of users’ accuracy are divided between TTS systems and ground-truth samples. They are summarized as: 39.4% for StyleTTS2, 58.1% for CosyVoice2, 62.0% for F5-TTS, and 70.8% for ground-truth.

4. RESULTS AND DISCUSSION

Once the experimental setup was validated, we proceeded to evaluate the SLMs. Table 1 shows accuracy scores achieved by each SLM for predicting both the target (audio) and proxy (semantic content) emotion labels. For comparison, the performance of the baseline SER system is also reported. Accuracy scores relative to target audio emotions approach those of a random classifier (25% for a four-class setting), whereas those relative to proxy labels are considerably higher under most conditions. There are substantial gaps between SLMs’ target and proxy accuracies across all semantic categories, most pronounced in the explicit case, in which Audio Flamingo3 notably displays a categorical pattern, always predicting the proxy label when classifying StyleTTS2 samples.

For the neutral category, accuracies remained stable for Qwen2Audio and SALMONN, but improved for DeSTA2 and Flamingo3 when compared with explicit and implicit categories. These results indicate that in the absence of emotional cues from semantic content, some SLMs appear to more effectively leverage acoustic information to perform emotion recognition. Moreover, this category led DeSTA2 and Qwen2Audio to perform significantly worse in proxy accuracy, whereas SALMONN performed slightly better, which can be associated with the text sentiment analysis capabilities of each model. In contrast, the modality-specific baseline SER consistently achieves higher target and lower proxy accuracies, indicating a focus on prosody cues, the desired behavior for this validation model. These results support the argument that SLMs have a strong tendency to prioritize information present in the semantic content rather than speech acoustics to perform the task, especially when the semantic content is not neutral.

Class-specific SLM decisions are presented in Figure 2. Under the congruent condition, i.e., when speech and semantic content have matching emotions, target and predicted emotions are closely aligned, indicating that SLM systems apparently leverage information mutually present in speech and semantic content. However, under the incongruent condition, i.e., when speech emotion differs from semantic content, this alignment breaks down, and SLM systems exhibit clear tendencies towards predicting the *angry* and *happy* classes while overlooking the *sad* class. This may reflect an interaction effect between the way each SLM model captures information

Table 1: Target (audio emotion) and proxy (semantic content emotion) accuracy scores achieved by each SLM, as well as the baseline SER system, under each semantic category defined in 3.1. SLMs’ target accuracies are consistently low across conditions, whereas the modality-specific baseline SER exhibits superior performance. SLMs predict proxy emotions more consistently in the *explicit* semantic condition and show distinct patterns of behavior in *implicit* and *neutral* conditions.

SLM	TTS	Explicit Category		Implicit Category		Neutral Category	
		Acc. (%)	Proxy Acc. (%)	Acc. (%)	Proxy Acc. (%)	Acc. (%)	Proxy Acc. (%)
DeSTA2	CosyVoice2	25.6	95.5	30.1	89.1	34.6	8.6
	F5-TTS	25.6	95.5	25.0	89.7	29.8	10.5
	StyleTTS2	25.6	97.4	28.2	91.6	38.4	7.6
Audio Flamingo3	CosyVoice2	28.8	93.5	37.8	66.0	41.3	76.9
	F5-TTS	26.2	98.7	31.4	82.6	38.4	86.5
	StyleTTS2	25.0	100.0	30.1	82.0	37.5	82.6
Qwen2Audio	CosyVoice2	26.2	96.7	30.1	69.2	21.1	11.5
	F5-TTS	26.2	98.7	29.4	75.6	26.9	9.6
	StyleTTS2	25.6	99.3	29.4	73.0	26.9	6.7
SALMONN	CosyVoice2	28.9	80.2	25.6	21.1	25.9	89.4
	F5-TTS	26.9	80.9	33.3	23.7	26.9	92.3
	StyleTTS2	27.2	89.6	26.2	30.1	36.5	71.1
Baseline SER	CosyVoice2	52.5	31.4	53.2	33.3	47.1	9.0
	F5-TTS	48.0	31.4	46.1	33.3	50.0	8.6
	StyleTTS2	50.0	26.9	47.4	30.7	49.0	1.0

and the fact that *angry* and *happy* samples are more closely associated with prosodic variations than *sad* and *neutral*.

The conducted chi-squared tests indicated that predicted emotion is significantly associated with both target and proxy labels ($p < 0.01$ for both cases), allowing us to reject the null hypothesis. However, the association between predicted and target emotions exhibited a very small effect size, with a Cramér’s V of 0.08, whereas the association between predicted and proxy emotions showed a considerable effect size ($V = 0.65$). These findings suggest that while acoustic cues have some influence on the models’ decisions, they are largely overshadowed by the spoken utterances’ semantic content, which has a much stronger impact on the model’s prediction.

5. CONCLUSION

This work investigated whether current SLMs can truly integrate semantic and acoustic information in their internal representations. Although seen as steps toward universal audio understanding, our evaluation suggests that these models fall short of this goal, showing a limited ability to disentangle semantics and acoustics when conflicting. The obtained results show that there is an imbalance between text and audio modalities, as the models tend to over-rely on information present in textual semantics, the more easily available that information is, as we can see in the explicit semantic condition. This has major implications for the rapidly growing ecosys-

tem of speech foundational models. If these models are evaluated primarily on benchmarks where semantic content and acoustic expressiveness are aligned, their apparent competence may mask critical deficiencies in their capacity for paralinguistic reasoning, crucial component in applications that depend on nuanced interpretation of human communication, such as detecting irony, sarcasm, or emotional subtleties.

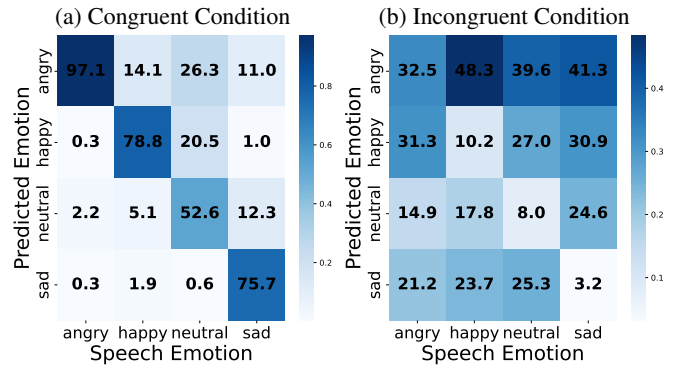


Fig. 2: Speech Emotion (target) vs. Predicted Emotion under congruent (a) and incongruent (b) conditions. Prediction counts are normalized column-wise and presented as percentage values. SLM predictions closely match the target labels when evaluated only on congruent samples, but display irregular, less reliable behavior with incongruent samples.

6. ACKNOWLEDGMENT

This work was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, by the São Paulo Research Foundation (FAPESP) under grant #2020/09838-0 (BIOS - Brazilian Institute of Data Science) and #2023/12865-8 (Horus project), and by the Ministry of Science, Technology, and Innovations, with resources from Law No. 8.248, of October 23, 1991, under the PPI-SOFTEX program, DOU 01245.003479/2024-10. The authors are also affiliated with the Artificial Intelligence Lab, Recod.ai.

7. COMPLIANCE WITH ETHICAL STANDARDS STATEMENT

Part of this research study was conducted retrospectively using human subject data extracted from the employed human perceptual evaluation. This evaluation has been approved by the Ethical Committee on Research (CEP) from Universidade Estadual de Campinas (UNICAMP) under CAAE Number 59536022.8.0000.5404.

8. REFERENCES

- [1] Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe, “On the landscape of spoken language models: A comprehensive survey,” *arXiv preprint arXiv:2504.08528*, 2025.
- [2] Jie Chi, Maureen de Seyssel, and Natalie Schluter, “The role of prosody in spoken question answering,” in *NAACL*, 2025.
- [3] Mariko Kikutani and Machiko Ikemoto, “Detecting emotion in speech expressing incongruent emotional cues through voice and content: investigation on dominant modality and language,” *Cognition and Emotion*, vol. 36, no. 3, pp. 492–511, 2022.
- [4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [5] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang, “SALMONN: Towards generic hearing abilities for large language models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [6] Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-Yi Lee, “Developing instruction-following speech language model without speech instruction-tuning data,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [7] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou, “Qwen2-audio technical report,” *CoRR*, vol. abs/2407.10759, 2024.
- [8] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro, “Audio flamingo 3: Advancing audio intelligence with fully open large audio language models,” *arXiv preprint arXiv:2507.08128*, 2025.
- [9] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” in *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, Aug. 2024, pp. 15747–15760.
- [10] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhi-jie Yan, and Jingren Zhou, “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *CoRR*, vol. abs/2412.10117, 2024.
- [11] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 19594–19621, Curran Associates, Inc.
- [12] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, “F5-tts: A fairytale that fakes fluent and faithful speech with flow matching,” *CoRR*, vol. abs/2410.06885, 2024.
- [13] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, “Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 920–924.