The Correlation Between Emotion in Text and Speech Segments is Limited: A Cross-Modal Study

David Lindevelt
Leiden University
Leiden, The Netherlands
david.lindevelt@gmail.com

Suzan Verberne
Leiden University
Leiden, The Netherlands
s.verberne@liacs.leidenuniv.nl

Joost Broekens
Leiden University
Leiden, The Netherlands
joost.broekens@gmail.com

Abstract

Although expressive TTS systems aim to capture human-like emotion, little is known about how well emotional signals in text correspond to those in speech. In this short paper, we investigate how emotion (Valence, Arousal, Dominance) in text relates to emotion in speech. We use 8 large language models for identifying emotion in text and two audio models for emotion in speech, across three genres: Podcasts, Audiobooks and TED talks. Findings show that while language models perform well on emotion recognition from situational text, and the audio models perform well on speech, they show a strong correlation for Valence only. Further, the genre of the content significantly impacts the correlation: audiobooks exhibit higher text-audio correlation than TED talks. Finally, we show that more context for LLMs fails to improve this correlation between text and speech emotion prediction. Our results highlight that emotional signals in text do not correspond well to those in speech: emotion prediction from text alone is insufficient for emotional TTS.

1 Introduction

Text-to-speech (TTS) systems have made remarkable progress in recent years, with models achieving near-human-like naturalness in speech synthesis (Li et al., 2023; Lee et al., 2025; Manku et al., 2025). However, one of the remaining challenges in this field is accurate and controllable emotions in synthesized speech. This challenge stems from a fundamental training–inference mismatch. TTS systems learn emotional patterns from text–audio pairs, but at inference time they must rely only on text to generate emotionally appropriate speech. For this to work, the emotional information in the text must be related to the emotion expressed in the corresponding speech audio. We define the *emotion relation* as a correlation between segment-based emotion values from text and speech modalities. We analyze how robust this relation is to changes in context, such as textual genre, and, whether more textual information improves this emotion relation.

The dimensional model of emotions (Russell and Mehrabian, 1977; Bradley and Lang, 1994), which includes three dimensions: Valence (positive – negative), Arousal (active – passive), and Dominance (dominant – submissive), is a commonly used, psychologically grounded representation for emotional states in both text and speech. Recent advances in large language models (LLMs) have shown promising results in assessing dimensional emotions from text (Broekens et al., 2023; Lei et al., 2023; Wu et al., 2025). Our research builds upon these developments. We investigate to what extent the acoustic emotional content relates to the textual emotional content in narrations, such as in audiobooks, podcasts and public lectures.

With this paper, we make three contributions: (1) We develop an analysis pipeline for extracting emotion values from text using state-of-the-art language models and assess their correlation with emotion values derived from corresponding audio narrations. (2) We evaluate the impact of both genre and contextual information (pre-text and post-text) on emotion prediction accuracy, simulating the contextual understanding that human narrators employ. (3) We show a weak relation between emotions identified in speech and text and thereby reveal an important limitation of current text to speech models that aim to express emotions in a natural way based only on text input.

2 Related work

Emotion prediction is an important topic in the fields of affective computing, speech synthesis and natural language processing. Good performance of speech emotion recognition (SER) can enhance human-machine interaction (Mohmad and Delhibabu, 2024).

Findings of the Association for Computational Linguistics: EACL 2026, pages 2611–2621
March 24-29, 2026

Peng et al. (2021) employ an LSTM architecture operating on audio-only cochleagram features for a dimensional SER task. With the success of attention mechanisms researchers focused on transformer architecture to capture feature-rich representation of audio. Wagner et al. (2023) fine-tuned a wav2vec 2.0-based model (Baevski et al., 2020) to predict emotions from speech audio.

Separate from SER, there is extensive research on sentiment analysis and emotion classification from text. Extracting emotions from text has been predominantly investigated with encoder-only (BERT-like) architectures (Siriwardhana et al., 2020; Acheampong et al., 2021). Recently, decoder-only language models have been explored (Broekens et al., 2023) for assessment of emotions in text, showing high correlation with human annotations. Wang et al. (2023) investigated emotional intelligence (EI) of LLMs using psychological tests and showed LLMs can achieve high EI scores.

Prior research showed that from text, Valence value predictions are the most feasible, while acoustic features proved to be better for Arousal prediction. To compensate for poor Valence prediction, audio-based prediction models incorporate text and audio modalities together in their architecture (Srinivasan et al., 2022; Atmaja and Akagi, 2020; Triantafyllopoulos et al., 2023).

In our analysis, we make use of decoder-only transformer models for predicting dimensional emotions from text and compare the output with speech-audio-only model predictions. We also examine the impact of text genre and additional textual context in prediction performance.

3 Experimental Methods

We first reproduce and extend the LLM-based emotion prediction capabilities reported in prior work (Broekens et al., 2023). We use a larger set of generative LLMs in the comparison. We compare different model sizes on the same dataset as presented in the paper, the ANET emotional situations dataset with human annotations. We slightly change the prompt for language models used in previous work and applied chain-of-thought prompting strategy with structured output when we predict emotion values using LLMs. The exact prompts used in the prediction process are adapted from prior work and provided in Appendix A.

Further, we reproduce the results of an open-source SOTA speech emotion recognition model on the MSP Podcast test set and compare predicted emotions against human annotated emotions.

Then, we compare text and speech emotions for three data sets representing different genres: podcasts, audio books and TED talks. To do so, we use the same set of language models to predict emotions from text and compare the predicted emotions against the speech-based emotion predictions. For these comparisons, we quantify the relation between speech-based and text-based emotion predictions by computing the Pearson correlation coefficient for each emotion dimension (Arousal, Dominance, and Valence) for each dataset.

Finally, we test whether providing more context, in the form of more text before and after the target sentence, increases the performance of LLMs.

3.1 Models

We select the wav2vec2-based model[1] as SOTA audio model for emotion prediction from speech audio. We also include a second WavLM-based model[2] to strengthen our analysis and evaluate generalizability. We look at how closely their predictions match with human annotations and their agreement across datasets.

We use 8 language models using the ollama[3] framework. These are llama3.3:70b, llama3.1:70b, phi3:14b, gemma2:9b, llama3.1:8b, qwen2:7b, openchat:7b, gemma:7b models. Each inference session follows default settings with the seed value set by experiment run. In total, 5 runs with different seed values are performed. We report the average of five different runs. The bold values in the tables represent statistically significant ($p < 0.01$) values for at least 80% (4/5) of the runs.

3.2 Datasets and pre-processing

We use four datasets for our analysis: 1. **ANET** (Bradley and Lang, 2007): 120 sentences describing emotional situations with dimensional emotion annotations (Pleasure, Arousal and Dominance) by human subjects. 2. **MSP Podcast** (Lotfian and Busso, 2017) version 1.8: 20,539 audio samples from podcasts with human emotion annotations on dimensions Valence, Arousal and Dominance. 3. **Libriheavy** (Kang et al., 2024): 20,539 samples from 'small' subset of dataset. It is an annotated

[1] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

[2] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Multi-Attributes

[3] https://ollama.com

version of Librilight (Kahn et al., 2020) dataset that consists of public domain audiobooks; it includes pre-text information for each segment as extended context. 4. **TEDLIUM** corpus release 2 (Rousseau et al., 2014): 89,986 samples from the train set of English speech recognition corpus from public TED talks.[4] For each sample, we keep preceding and following texts as extended context.

We present details on emotion annotations and emotionally balanced sample selection process for all datasets in Appendix B.

The speech-based audio-only models used in our experiments are trained on the train set of MSP Podcast dataset. We use the combined MSP Podcast test sets (test1 and test2) and selected emotionally balanced subset for our analysis. The MSP Podcast dataset does not include transcriptions. We use the OpenAI Whisper Turbo model (Radford et al., 2023) to transcribe the audio data. Emotion ratings in the MSP Podcast dataset range between 1 (very negative/calm/weak) and 7 (very positive/active/strong), for the dimensions Valence, Arousal, and Dominance. We apply min-max scaling to reduce their range to 0-1, using 1 and 7 as the defined minimum and maximum values.

4 Results

We describe our results along the lines of four research questions.

How good are LLMs at extracting emotions from a given text? Before comparing emotions extracted from text and speech, we first assess the capability of language models to predict dimensional emotions from text. Extending previous work (Broekens et al., 2023), we use 8 language models to predict dimensional emotions for the ANET dataset.

In this experiment, we evaluate both large and small models by comparing their predictions with ground truth text emotions annotated by human subjects. Table 1 reports average agreement over 5 different seeds. Our results show that all models achieve statistically significant and strong correlation with the human labels. Large models exhibit stronger correlations than smaller versions. This experiment confirms that decoder-only language models are able to reliably extract emotion dimensions in a given situational text, particularly when the text explicitly contains affective information, as is the case for the ANET dataset.

[4] https://www.ted.com

Table 1: Correlation between text-based prediction and human ratings – ANET Dataset. Boldface indicates a significance level of $p < 0.01$ in 80% of experiment runs across 5 seeds

	Arousal	Dominance	Valence
llama3.3:70b	**.929** (.003)	**.865** (.004)	**.966** (.002)
llama3.1:70b	**.918** (.008)	**.891** (.010)	**.967** (.009)
phi3:14b	**.822** (.043)	**.828** (.040)	**.870** (.053)
gemma2:9b	**.888** (.008)	**.666** (.026)	**.870** (.037)
llama3.1:8b	**.846** (.015)	**.508** (.048)	**.924** (.013)
qwen2:7b	**.553** (.017)	**.709** (.060)	**.950** (.017)
openchat:7b	**.740** (.032)	**.699** (.054)	**.914** (.046)
gemma:7b	**.803** (.038)	**.392** (.044)	**.901** (.017)

How does text-based emotion prediction relate to speech-based emotions? To quantify the relation between text-based and speech-based dimensional emotion, we use the MSP podcast dataset (Lotfian and Busso, 2017). First, we replicate the in-context performance evaluation of the open-source speech-based model as reported in Wagner et al. (2023). Our replication of their evaluation gave correlations of 0.772 (Arousal), 0.640 (Dominance), and 0.646 (Valence) on the MSP Podcast combined test set, validating the model's performance for speech emotion prediction.

Next, we compare text-based emotion predictions with speech-based emotion annotation and speech-based prediction for podcast transcripts. We find relatively high correlation for Valence (highest 0.630, llama3.1:70b) while performing substantially lower on Arousal (highest 0.384, llama3.3:70b) and Dominance (highest 0.117, gemma2:9b). Valence and Arousal correlations are statistically significant while Dominance shows no statistical significance. This is true for both the correlations between text-based prediction and ground-truth data, and text-based and speech-based predictions (see Table 4 in Appendix D). In other words, predicted text emotions correspond well with speech emotions in terms of Valence, but are less effective at capturing emotional cues related to Arousal and Dominance. This suggests that while both text-based and speech-based predictions are by themselves reliable, there seems to be weak relation and perhaps a genre dependency.

Does text genre impact the relation between emotions predicted from text and speech? To further investigate the effect of genre on the rela-

tion between text-based and speech-based emotion predictions, we repeated the analysis for audiobooks (Libriheavy) and TED talks (TEDLIUM). Since these datasets lack human annotations for text or speech emotions, and the speech-based prediction model worked well for the MSP podcast data, we treat predictions from SOTA audio model as "ground truth" for calculating the emotion relation. For comparison purposes we do the same for the podcast dataset. We find that the correlation for Valence is consistently higher than for Arousal and Dominance. Further, Dominance correlation is consistently low. Also, Valence and Arousal correlations between text-based and speech-based predictions are higher for audiobooks than the other two genres. Tables 5 and 7 in Appendices E and F, respectively, provide detailed results.

Does additional context improve the relation between text and audio emotion predictions? A potential reason for lower correlation on Arousal and Dominance is the lack of textual context. We test whether providing more context improves this relation. We use the same samples from both the audiobooks and TED talks, but we extend their context. For audiobooks, we combine pre-text with the main text, and for TEDLIUM, we include preceding and following sentences. LLMs are then prompted to extract emotions from the main text, considering the full context (see Appendix A).

Tables 2 and 3 show that while extended context may marginally enhance the relation in Arousal for the larger LLM models in the audiobook data, it does not significantly improve and may even slightly weaken the relation for Valence. This indicates that the benefit of additional context may depend on the emotion dimension and the nature of the dataset, and is not conclusive.

Do our findings generalize to another audio model? We find that the two audio models show high agreement with each other and show similar levels of agreement with human annotations, confirming our findings. Agreements measured with Pearson correlation coefficient and presented in Tables 9 (between the two models) and 10 (between models and human annotation) in Appendix G.

Table 2: Correlation between *contextual* text and speech emotion prediction on Libriheavy

	Arousal	Dominance	Valence
llama3.3:70b	**.416** (.016)	.175 (.054)	**.700** (.006)
llama3.1:70b	**.427** (.021)	.015 (.025)	**.697** (.016)
phi3:14b	**.392** (.080)	.108 (.064)	**.591** (.066)
gemma2:9b	**.450** (.015)	**.297** (.026)	**.641** (.023)
llama3.1:8b	**.473** (.064)	.119 (.117)	**.642** (.040)
qwen2:7b	**.267** (.067)	.064 (.044)	**.679** (.016)
openchat:7b	**.320** (.099)	.006 (.037)	**.591** (.089)
gemma:7b	.230 (.052)	.144 (.100)	**.549** (.053)

Table 3: Correlation between *contextual* text and speech emotion prediction on TEDLIUM

	Arousal	Dominance	Valence
llama3.3:70b	.210 (.011)	.214 (.028)	**.627** (.007)
llama3.1:70b	.223 (.030)	.174 (.022)	**.624** (.017)
phi3:14b	.244 (.058)	.110 (.072)	**.538** (.077)
gemma2:9b	.135 (.029)	.116 (.058)	**.586** (.021)
llama3.1:8b	.208 (.041)	.217 (.037)	**.590** (.019)
qwen2:7b	**.340** (.041)	**.292** (.057)	**.647** (.011)
openchat:7b	**.237** (.058)	.181 (.054)	**.546** (.106)
gemma:7b	.155 (.043)	.048 (.063)	**.504** (.012)

5 Conclusions

Language models excel on the emotion recognition task from situational text, but their predictions correlate poorly with speech-based emotions, particularly on the Arousal and Dominance in all tested genres. This shows that text alone is insufficient for emotional TTS. Second, genre impacts emotion relation, with audiobooks showing better text-speech correlation than TED talks and podcasts. Third, contrary to our expectation, increasing context did not improve the relation between text and speech emotions. Our results confirm the view of Gunes and Schuller (2013) that Activation (Arousal) is easily captured through acoustic features, Valence (positivity) is more reliably extracted from linguistic features, highlighting that emotional signals are not always shared across modalities.

Our findings motivate the development of text-based emotion predictors that match the emotion expressed in speech, but also show a need for improved TTS conditioning strategies that accurately control affective expression. Detailed understanding of the relation of textual and acoustic emotional content is essential for next generations of TTS. This will help natural-sounding speech synthesis but also enable better control over the subtle emotional nuances present in the source text, much like a skilled human narrator would have.

Limitations

For Libriheavy and TEDLIUM, we rely on an open-source audio-only model trained on podcasts, potentially introducing bias. Additionally, we use English-only data to conduct the analysis, future work should include human annotations, multilingual data and consider alternative metrics to account for imbalanced emotion distributions, particularly for extreme values that may skew Pearson correlation results. Further, binned emotion distributions for MSP Podcast, Libriheavy audiobooks, and TEDLIUM (see Appendix C) reveal that audiobooks show more neutral emotions, while TEDLIUM emotions in higher bins similarly to MSP Podcast; all show limited coverage of edge emotions. This may influence the correlations.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Bagus Tris Atmaja and Masato Akagi. 2020. Improving valence prediction in dimensional speech emotion recognition using linguistic information. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 166–171. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

M.M. Bradley and P.J. Lang. 2007. Affective norms for english text (anet): Affective ratings of text and instruction manual. *Techical Report. D-1, University of Florida, Gainesville, FL.*

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE.

Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136.

Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, and 1 others. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2025. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36:19594–19621.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alex Smola. 2025. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *arXiv preprint arXiv:2505.23009*.

GH Mohmad and Radhakrishnan Delhibabu. 2024. Speech databases, speech features and classifiers in speech emotion recognition: A review. *IEEE Access*.

Zhichao Peng, Jianwu Dang, Masashi Unoki, and Masato Akagi. 2021. Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech. *Neural Networks*, 140:261–273.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Anthony Rousseau, Paul Deléglise, Yannick Esteve, and 1 others. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*, pages 3935–3939.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*.

Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. 2022. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6442–6446. IEEE.

Andreas Triantafyllopoulos, Uwe Reichel, Shuo Liu, Stephan Huber, Florian Eyben, and Björn W Schuller. 2023. Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Frontiers in Computer Science*, 5:1072479.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2025. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2202–2218.

A LLM Prompts

Here is the prompt structure used in extracting emotions from a given text using LLMs.:

```
[{
   "role": "system",
   "content": "Valence, Arousal and Dominance are three affective dimensions that you can use to
        identify the sentiment in sentences. Assume that these dimensions can take values between 0
        and 1, with 0 being low, and 1 being high. Remember that dominance assesses the extent to
        which the main person in the situation experiences the amount of control it can assert over
        the situation. Assess according to these dimensions the sentiment in the sentences I will give
         you after. Be precise, and output the values (up until two digits after the decimal point) in
         a JSON format please. Use 'Valence', 'Arousal' and 'Dominance' as keys in JSON format. Just
        acknowledge you got it."
},
{
   "role": "assistant",
   "content": "I'll assess the sentiment in the given sentence according to the Valence, Arousal, and
         Dominance dimensions and output the values in a JSON format. Please go ahead and provide the
        sentence. I'll output the values in a JSON format once all the assessments are complete."
},
{
   "role": "user",
   "content": "[SENTENCE]"
}]
```

Prompt structure for contextual emotion extraction:

```
[{
    "role": "system",
    "content": "Valence, Arousal and Dominance are three affective dimensions that you can use to
         identify the sentiment in sentences. Assume that these dimensions can take values between 0
         and 1, with 0 being low, and 1 being high. Remember that dominance assesses the extent to
         which the main person in the situation experiences the amount of control it can assert over
         the situation. You will be provided with three parts of text; 1. Pre-text: The text that
         comes before the main sentence. 2. Main sentence: The sentence to be evaluated. 3. Post-text:
          The text that comes after the main sentence. Consider the entire context (pre-text, main
         sentence, and post-text) when evaluating the emotional dimensions, but focus your assessment
         primarily on the main sentence. The surrounding context should inform your understanding and
         interpretation of the main sentence's emotional content. Assess according to these dimensions
          the sentiment in the main sentence. Be precise, and output the values (up until two digits
         after the decimal point) in a JSON format please. Use 'Valence', 'Arousal' and 'Dominance' as
          keys in JSON format. Just acknowledge you got it."
},
{
    "role": "assistant",
    "content": "I'll assess the sentiment in the given sentence according to the Valence, Arousal,
         and Dominance dimensions and output the values in a JSON format. I will focus on assesing
         main sentence while using surrounding text as contextual cue for my assessment. Please go
         ahead and provide the previous text, main sentence and post text. I'll output the values in a
          JSON format once all the assessments are complete."
},
{
    "role": "user",
    "content": "Pre-Text: [PRE-TEXT}"; Main Sentence: [SENTENCE]; Post-Text: [POST-TEXT]"
}]
```

B Details on data processing

To calculate correlations, we sample all except the ANET datasets and force the sample to a) cover the VAD space as best as possible and b) keep the number of segments close to ANET dataset. We use the speech-based model to predict emotion values from audiobooks and TED talks and then sampled based on predicted speech emotions. We sampled a subset of the MSP Podcast dataset based on ground truth emotion values. For the sample selection process we aim to cover emotional space as equally as possible to not bias the statistical correlation analyses. We divide the 0-1 emotion value range into 5 bins. For each bin we randomly select maximum 3 samples. If the bin does not have enough available samples to select we select the remaining as whole. We reduce the sampling criteria for MSP-Podcast as maximum

2 samples to keep sample sizes closer to each other across datasets. MSP Podcast, audiobooks, and TEDLIUM datasets yield 162, 115, 148 samples respectively.

Libriheavy and TEDLIUM datasets lack human annotated emotion labels. After validating the performance of SOTA speech-only Wav2Vec2 model, we use it as proxy and its prediction for emotion annotations. For a fair comparison we also use its prediction of MSP Podcast dataset in our analysis.

C Coverage of Emotions in Audio Datasets

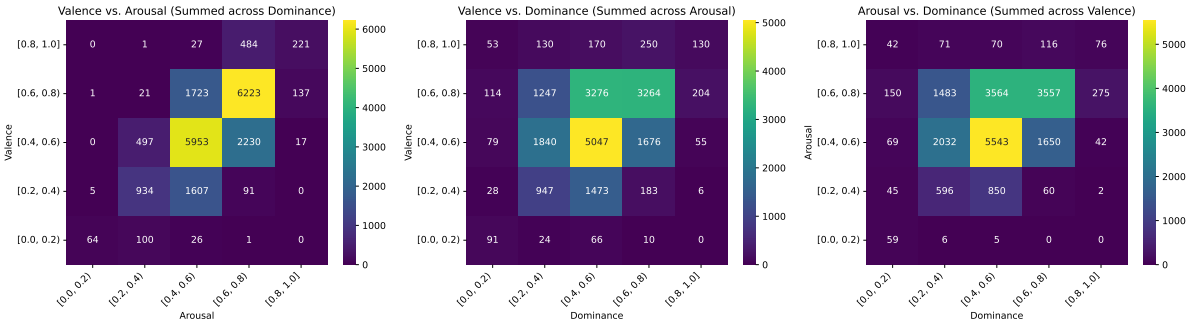

Figure 1: MSP Podcast dataset emotion coverage (Ground Truth Human Annotations).

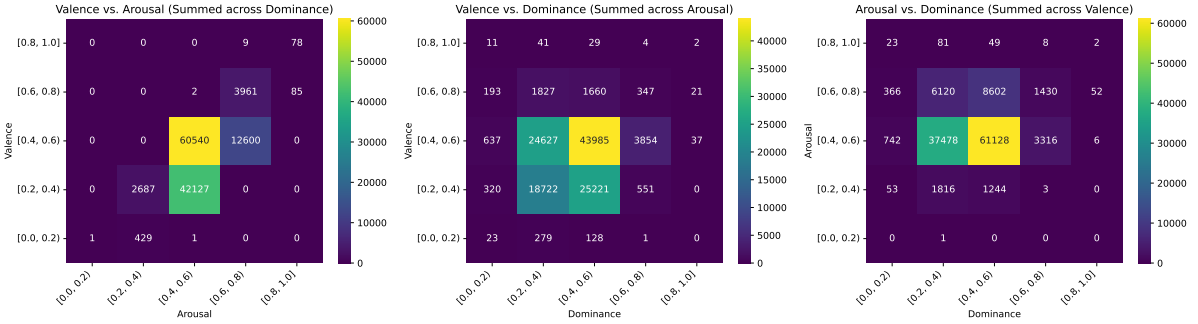

Figure 2: Libriheavy dataset emotion coverage (Audio Model Predictions).

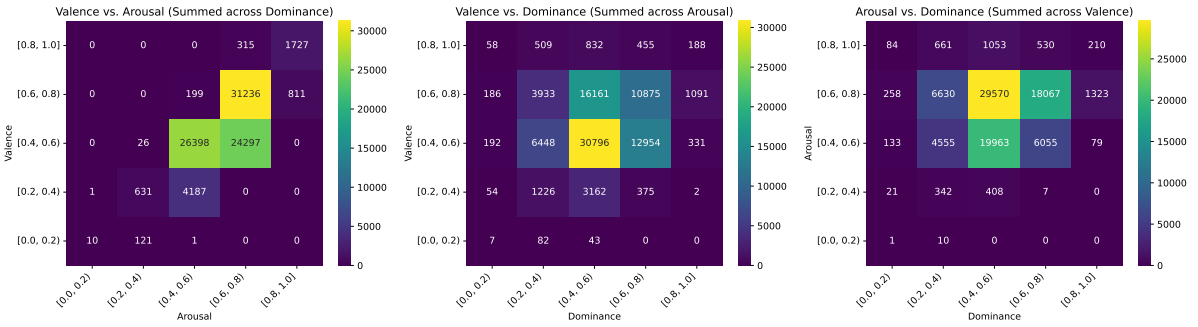

Figure 3: TEDLIUM dataset emotion coverage (Audio Model Predictions).

D Results on the MSP Podcast dataset

Table 4: Correlation between text-based emotion prediction and speech-based emotion on MSP Podcast

	Wav2Vec2			WavLM			Human Annotations		
Model	A	D	V	A	D	V	A	D	V
llama3.3:70b	**.384** (.012)	.083 (.015)	**.618** (.014)	**.351** (.017)	.007 (.013)	**.589** (.014)	**.298** (.010)	.041 (.011)	**.604** (.002)
llama3.1:70b	**.344** (.026)	.092 (.042)	**.630** (.022)	**.313** (.030)	.018 (.051)	**.610** (.021)	**.265** (.035)	.088 (.040)	**.584** (.037)
phi3:14b	**.295** (.032)	-.041 (.049)	**.506** (.088)	**.269** (.023)	-.075 (.035)	**.471** (.091)	**.223** (.036)	-.011 (.037)	**.478** (.057)
gemma2:9b	**.381** (.027)	.117 (.028)	**.483** (.024)	**.360** (.021)	.092 (.025)	**.460** (.028)	**.281** (.034)	.074 (.023)	**.467** (.022)
llama3.1:8b	**.323** (.050)	.075 (.069)	**.545** (.026)	**.298** (.059)	.066 (.065)	**.514** (.016)	**.252** (.038)	.018 (.035)	**.545** (.025)
qwen2:7b	.210 (.079)	.097 (.060)	**.589** (.027)	.156 (.074)	.063 (.076)	**.541** (.023)	.161 (.063)	.025 (.058)	**.509** (.029)
openchat:7b	**.296** (.066)	.014 (.041)	**.511** (.090)	**.252** (.069)	-.032 (.052)	**.479** (.100)	.197 (.075)	.037 (.038)	**.495** (.107)
gemma:7b	**.358** (.060)	-.039 (.069)	**.417** (.017)	**.369** (.062)	-.061 (.056)	**.402** (.009)	**.306** (.031)	-.036 (.066)	**.419** (.034)

Note: Values show mean correlation (standard deviation) across 5 seeds.
Bold values indicate statistical significance ($p < 0.01$) in $>= 80\%$ of runs.

E Results on Libriheavy Dataset

Table 5: Correlation between text-based and speech-based emotion prediction on Libriheavy dataset

	Wav2Vec2			WavLM		
Model	A	D	V	A	D	V
llama3.3:70b	**.328** (.011)	.172 (.021)	**.783** (.010)	**.399** (.010)	.094 (.022)	**.664** (.011)
llama3.1:70b	**.374** (.039)	.072 (.044)	**.793** (.018)	**.422** (.031)	.005 (.040)	**.701** (.024)
phi3:14b	**.350** (.048)	.104 (.063)	**.671** (.091)	**.353** (.041)	.040 (.063)	**.615** (.093)
gemma2:9b	**.437** (.010)	.198 (.030)	**.713** (.013)	**.494** (.011)	.147 (.028)	**.636** (.027)
llama3.1:8b	**.418** (.050)	.201 (.108)	**.704** (.019)	**.484** (.042)	.180 (.093)	**.628** (.029)
qwen2:7b	**.345** (.023)	.195 (.051)	**.728** (.019)	**.303** (.021)	.102 (.039)	**.626** (.032)
openchat:7b	**.332** (.065)	.057 (.083)	**.720** (.101)	**.331** (.071)	-.032 (.088)	**.634** (.102)
gemma:7b	.250 (.054)	.031 (.075)	**.619** (.007)	**.254** (.038)	-.003 (.058)	**.577** (.006)

Note: Values show mean correlation (standard deviation) across 5 seeds.
Bold values indicate statistical significance ($p < 0.01$) in $>= 80\%$ of runs.

Table 6: Correlation between *contextual* text-based emotion prediction and speech-based emotion prediction on Libriheavy

	Wav2Vec2			WavLM		
Model	A	D	V	A	D	V
llama3.3:70b	**.416** (.016)	.175 (.054)	**.700** (.006)	**.468** (.018)	.060 (.047)	**.618** (.007)
llama3.1:70b	**.427** (.021)	.015 (.025)	**.697** (.016)	**.484** (.017)	-.070 (.029)	**.626** (.023)
phi3:14b	**.392** (.080)	.108 (.064)	**.591** (.066)	**.392** (.076)	.023 (.060)	**.519** (.064)
gemma2:9b	**.450** (.015)	**.297** (.026)	**.641** (.023)	**.511** (.015)	**.266** (.027)	**.599** (.021)
llama3.1:8b	**.473** (.064)	.119 (.117)	**.642** (.040)	**.502** (.056)	.067 (.100)	**.576** (.027)
qwen2:7b	**.267** (.067)	.064 (.044)	**.679** (.016)	.251 (.053)	-.004 (.039)	**.563** (.014)
openchat:7b	**.320** (.099)	.006 (.037)	**.591** (.089)	**.303** (.088)	-.076 (.031)	**.496** (.097)
gemma:7b	.230 (.052)	.144 (.100)	**.549** (.053)	.195 (.050)	.102 (.090)	**.458** (.063)

Note: Values show mean correlation (standard deviation) across 5 seeds.
Bold values indicate statistical significance ($p < 0.01$) in $>= 80\%$ of runs.

F Results on TEDLIUM Dataset

Table 7: Correlation between text-based emotion prediction and speech-based emotion prediction on TEDLIUM dataset

	Wav2Vec2			WavLM		
Model	A	D	V	A	D	V
llama3.3:70b	**.281** (.005)	**.274** (.035)	**.685** (.013)	**.306** (.008)	.190 (.029)	**.698** (.016)
llama3.1:70b	**.301** (.024)	.218 (.020)	**.645** (.029)	**.317** (.037)	.128 (.025)	**.638** (.044)
phi3:14b	**.243** (.027)	.025 (.029)	**.599** (.090)	**.271** (.042)	-.028 (.048)	**.626** (.086)
gemma2:9b	**.263** (.018)	**.270** (.058)	**.560** (.020)	**.289** (.021)	**.250** (.059)	**.556** (.005)
llama3.1:8b	.179 (.047)	.138 (.056)	**.616** (.040)	.210 (.062)	.123 (.051)	**.612** (.029)
qwen2:7b	**.270** (.052)	**.292** (.057)	**.662** (.042)	**.246** (.055)	.247 (.052)	**.672** (.036)
openchat:7b	.203 (.118)	.128 (.077)	**.609** (.109)	.170 (.108)	.060 (.086)	**.615** (.103)
gemma:7b	**.281** (.022)	-.052 (.041)	**.472** (.009)	**.279** (.032)	-.110 (.044)	**.528** (.007)

Note: Values show mean correlation (standard deviation) across 5 seeds.
Bold values indicate statistical significance ($p < 0.01$) in $>= 80\%$ of runs.

Table 8: Correlation between *contextual* text-based emotion prediction and speech-based emotion prediction on TEDLIUM

	Wav2Vec2			WavLM		
Model	A	D	V	A	D	V
llama3.3:70b	.210 (.011)	.214 (.028)	**.627** (.007)	**.241** (.023)	.125 (.024)	**.654** (.005)
llama3.1:70b	.223 (.030)	.174 (.022)	**.624** (.017)	**.233** (.027)	.078 (.027)	**.664** (.016)
phi3:14b	.244 (.058)	.110 (.072)	**.538** (.077)	.244 (.055)	.045 (.068)	**.561** (.081)
gemma2:9b	.135 (.029)	.116 (.058)	**.586** (.021)	.164 (.029)	.046 (.063)	**.578** (.024)
llama3.1:8b	.208 (.041)	.217 (.037)	**.590** (.019)	**.242** (.038)	.220 (.043)	**.629** (.037)
qwen2:7b	**.340** (.041)	**.292** (.057)	**.647** (.011)	**.263** (.054)	.216 (.045)	**.653** (.011)
openchat:7b	**.237** (.058)	.181 (.054)	**.546** (.106)	.214 (.043)	.104 (.062)	**.569** (.107)
gemma:7b	.155 (.043)	.048 (.063)	**.504** (.012)	.151 (.033)	.007 (.061)	**.508** (.009)

Note: Values show mean correlation (standard deviation) across 5 seeds.
Bold values indicate statistical significance ($p < 0.01$) in $>= 80\%$ of runs.

G Audio Model Comparison

Table 9: Correlation between emotion predictions of the two audio models Wav2Vec2 and WavLM across datasets

	Arousal	Dominance	Valence
MSP Podcast	**.928**	**.915**	**.823**
Libriheavy	**.934**	**.919**	**.869**
TEDLIUM	**.935**	**.931**	**.868**

Table 10: Correlation between audio models on MSP Podcast dataset human annotations

	Arousal	Dominance	Valence
Wav2Vec2	**.772**	**.640**	**.646**
WavLM	**.765**	**.677**	**.577**