On the Emotion Understanding of Synthesized Speech

Anonymous ACL submission

Abstract

Emotion is a core paralinguistic feature in voice interaction. It is widely believed that emotion understanding models learn fundamental representations that transfer to synthesized speech, making emotion understanding results a plausible reward or evaluation metric for assessing emotional expressiveness in speech synthesis. In this work, we critically examine this assumption by systematically evaluating Speech Emotion Recognition (SER) on synthesized speech across datasets, discriminative and generative SER models, and diverse synthesis models. We find that current SER models can not generalize to synthesized speech, largely because speech token prediction during synthesis induces a representation mismatch between synthesized and human speech. Moreover, generative Speech Language Models (SLMs) tend to infer emotion from textual semantics while ignoring paralinguistic cues. Overall, our findings suggest that existing SER models often exploit non-robust shortcuts rather than capturing fundamental features, and paralinguistic understanding in SLMs remains challenging.

1 Introduction

Speech understanding enables machines to extract meaning and intent from spoken signals, supporting robust interaction and reliable decision-making in real-world settings (Serdyuk et al., 2018; Haghani et al., 2018; Lugosch et al., 2019; Borsos et al., 2023; Wang et al., 2025a). Speech understanding is typically assumed to operate on naturally produced human speech, and the target information extends beyond semantic content to paralinguistic features and speaker identity. For example, the same expression 'Really?' can convey distinct intents when spoken with surprise versus doubt emotions.

Speech understanding is increasingly challenged by the growing prevalence of synthesized speech (Xie et al., 2025; Cui et al., 2025; Hurst et al., 2024). As people can convey the same content via either recorded or synthesized audio, a natural question arises: Does understanding synthesized speech differ from understanding human speech?

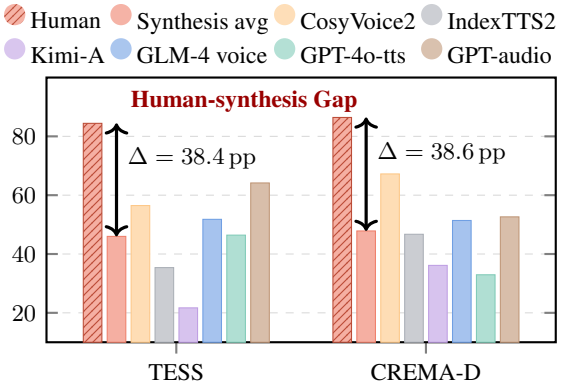

Figure 1: Speech emotion recognition (SER) results on TESS and CREMA-D dataset. Synthetic speech results represent the agreement with human. Results highlight a clear gap between human and synthesized speech.

We investigate this shift through Speech Emotion Recognition (SER), a key component of voice-centric interaction. Prior work reports strong emotion recognition accuracy on human speech using both discriminative SER models and generative Speech Large Language Models (SLMs) (Koh and Dubnov, 2021; Sadok et al., 2023; Ma et al., 2024; Wu et al., 2025a), yet generating speech that reliably expresses a target emotion remains challenging for text-to-speech (TTS) systems and speech-to-speech (S2S) LLMs (Du et al., 2024; Chen et al., 2025b; Zeng et al., 2024; Ding et al., 2025). Recent methods therefore evaluate or optimize synthesis using SER-based signals (Yang et al., 2025; An et al., 2024; Wang et al., 2025b; Ji et al., 2025; Chen et al., 2025a; Zhang et al., 2025). However, they assume that SER models learn emotion representations that transfer to synthesized speech, which has not been rigorously validated. This motivates our central research question: *Can current models reliably understand emotion in synthesized speech?*

To answer it, we systematically evaluate emotion understanding on human and synthesized speech across datasets, SER models (discriminative and generative), and synthesis paradigms (TTS and S2S LLMs). Our results provide four key findings:

- Discriminative SER models generalize poorly under synthesized domain shift, and our findings highlight a clear gap between human and synthesized speech as shown in Fig. 1.
- The dominant source of synthesized domain shift is the speech token prediction process, whereas the subsequent stages, flow matching and vocoder, contribute substantially less.
- Supervised fine-tuning reduces the synthetic-domain gap, but neither standard fine-tuning nor domain-adversarial fine-tuning fundamentally resolves generalization, suggesting that SER models may rely on non-robust shortcuts.
- Speech LLMs tend to infer emotion primarily from textual semantic cues while ignoring paralinguistic signals. Prompt engineering does not change this behavior, indicating a persistent *text dominance* (Wu et al., 2025c).

We released our code to facilitate future research[1].

2 Related work

Speech Emotion Recognition Model The key problem to understand speech emotions is to learn the speech representation of the waveform (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Baevski et al., 2022) and finetune the models in downstream classification tasks (Koh and Dubnov, 2021; Sadok et al., 2023; Ma et al., 2024). We utilized Emotion2vec (Ma et al., 2024), the most powerful and widely used open source SER model training on human speech, to evaluate the recognition performance on synthesized speech.

Text-to-Speech Models Emotional expression is critical in text-to-speech synthesized (Li et al., 2023; Du et al., 2024; Yang et al., 2025; Zhou et al., 2025). However, we observe that the speech synthesized by most models fails to convey distinct emotions perceivable by human listeners. We utilized CosyVoice 2 and IndexTTS 2, the open-source models with the most pronounced emotional expression, to generate TTS synthesized speech.

[1] https://anonymous.4open.science/r/Synthesized_SER

Speech-to-Speech LLMs Emotion matters more in voice interaction between S2S LLMs and human (Hurst et al., 2024; Chu et al., 2024; Zeng et al., 2024; Wu et al., 2025a; Xu et al., 2025a; Ding et al., 2025). We utilized Kimi-Audio and GLM4-Voice, the open-source S2S LLMs with the most pronounced emotional expression, to generate S2S synthesized speech.

3 Preliminaries

To validate whether current SER models reliably understand emotion in synthesized speech, we investigate their performance on three types of speech: real human speech, speech synthesized by TTS, and speech generated by S2S LLMs.

3.1 Task Formulation

Formally, let $\mathcal{M}$ denote a SER model. The emotion recognition process can be formulated as:

$$E = \mathcal{M}(\mathcal{D}_S) \tag{1}$$

where E represents the array of output emotion labels drawn from the discrete set {angry, disgusted, fearful, happy, neutral, sad, surprised, other}. The input $\mathcal{D}_S$ denotes the speech dataset to be categorized, which may originate from natural human recordings, TTS models, or S2S LLMs. We formulate the speech synthesis process for TTS or S2S as:

$$\mathcal{D}_S = \mathcal{S}(\mathcal{D}_T, C) \tag{2}$$

where $\mathcal{D}_T$ is the text dataset and C is the emotion control signal. Despite the unified notation C, the realization of emotional control varies: TTS models typically utilize prompt speech cloning, natural instructions, or emotion embeddings to define emotions, whereas S2S LLMs require multi-turn interactions to achieve significant emotional output. More details are available at Appendix C. Then the reverse process of synthesis is formulated as:

$$\mathcal{D}_T = \mathcal{A}(\mathcal{D}_S) \tag{3}$$

where $\mathcal{A}$ represents Auto Speech Recognition (ASR), transcribing from speech distribution to text distribution using Whisper (Radford et al., 2023).

3.2 SER Results on Human Speech

We investigate the performance of Emotion2vec in two test datasets: TESS (Pichora-Fuller and Dupuis, 2020) and CREMA-D (Cao et al., 2014). In the confusion matrix shown in Fig. 2 (a) and

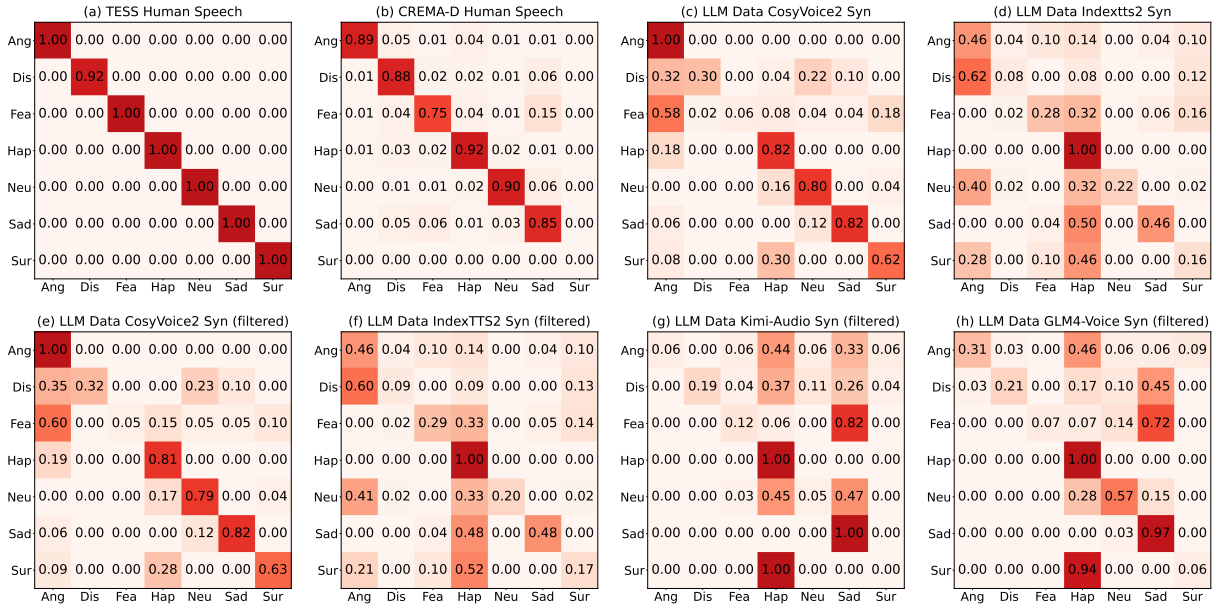

Figure 2: The confusion matrix for speech emotion recognition is shown. The vertical axis represents the ground truth, and the horizontal axis represents the model's predictions. Sub-figures (a) and (b) show SER results on human speech, TESS and CREMA-D. Sub-figures (c) and (d) show SER results for speech synthesized by two TTS models from LLM-generated text. Sub-figures (e) and (f) represent the same results as (c) and (d) after filtering out weak emotional expression. Sub-figures (g) and (h) show the same as (e) and (f), but with synthesis by S2S LLMs.

(b), the vertical axis represents the ground truth, while the horizontal axis represents the model's recognition results. Therefore, the sum of the values in each row equals 1.0. Experimental results demonstrate that Emotion2vec achieves robust performance on human speech, yielding a confusion matrix with a pronounced diagonal structure.

4 Domain Shift to Synthetic Speech

In this section, we apply the discriminative SER model and generative SLMs to synthetic audio, examining the emotion understanding performance and analyzing the potential underlying causes.

4.1 SER Results on Synthesized Speech

Here, we investigate whether the SER model trained on human speech generalizes to synthesized speech. Specifically, we utilized gpt-4o to generate 3,028 text sentences and synthesize into speech using two TTS emotion control methods: prompt speech and natural instruction.

As shown in Fig. 2 (c) and (d), the experimental results demonstrate that Emotion2vec achieves poor performance on TTS synthesized speech, producing a confusion matrix significantly without diagonal structure. Experimental results demonstrate that the SER model trained on human data performs well on human speech, while performs poorly on synthesized speech. So we explore the potential reasons for this performance gap.

4.1.1 Hypothesis 1: TTS Emotion Expression

As described in related work, we observed that the speech synthesized by most models fails to convey distinct emotions perceivable by human listeners. Despite employing TTS and S2S models with superior emotional expressiveness, certain synthesized utterances still fail to express the target emotion.

Consequently, we recruited four annotators to manually filter out synthesized utterances that failed to convey the target emotion. However, as illustrated in Fig. 2 (e) and (f), the confusion matrix still lacks a distinct diagonal structure. Experimental results demonstrate that the SER model trained on human recorded data still generalizes poorly to the synthesized speech, even when the target emotions are clearly expressed.

4.1.2 Hypothesis 2: Lack of Training Data

Secondly, we examine whether the SER model exhibits performance degradation in specific emotion categories. Previous work by Yang et al. reported limited accuracy in Emotion2vec for disgusted, fearful, and surprised categories due to insufficient training data. Consistent with this, Fig. 2 (e) confirm that the model struggles primarily with these three emotions other than neutral. However, as

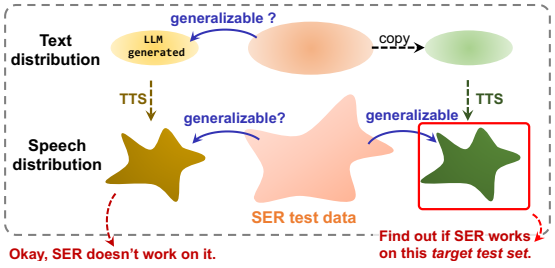

Figure 3: To mitigate the impact of text distribution generated by LLMs, we investigated speech emotion recognition performance on identical text datasets.

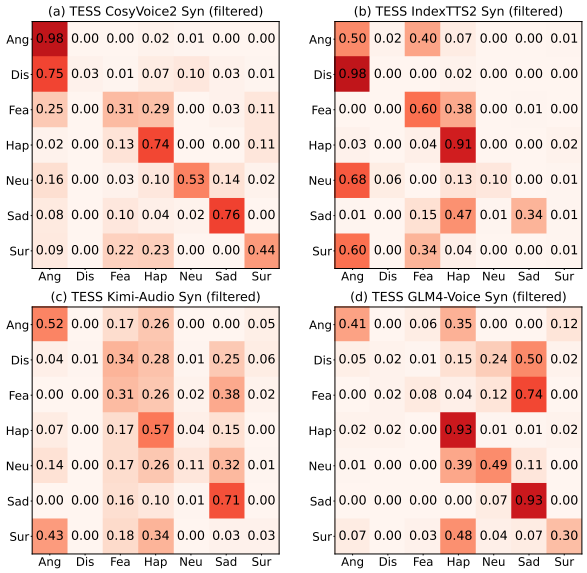

Figure 4: Confusion matrix of speech emotion recognition results for synthetic speech based on TESS text. The vertical axis represents the ground truth, while the horizontal axis represents the model's predictions. Subfigs (a) and (b) show SER results for two TTS models, while (c) and (d) show SER results for two S2S models.

shown in Fig. 2 (f) (g) and (h), further experiments indicate a significant drop in performance across almost all emotion categories for synthesized speech generated by IndexTTS and S2S LLMs, including categories with sufficient training data, such as happy, angry, and sad. To ensure reliability, all synthesized audio employed in this section and hereafter was checked by human annotators to verify the distinctness of the emotional expression.

4.1.3 Hypothesis 3: Text Distribution Gap

Moreover, we investigate to eliminate the concern that Emotion2vec lacks robustness to shifts in textual distribution. It is reasonable to hypothesize that the performance of the SER model is related to the underlying text distribution, where *'text distribution'* is defined as the underlying semantic representation or ground truth ASR text sentences, identical with Eq. 3. As shown in Fig. 3, our initial recognition experiments on LLM-generated text (leftmost portion) yielded poor results. This raises a critical question: is the failure caused by an unfamiliar text distribution, or by the domain gap in the speech synthesis process?

To disentangle these factors, we conducted the ablation experiment depicted on the right side of Fig. 3. We first obtained ASR text transcripts from a test dataset, then synthesized text to generate a new target test set. The process to generate the target test dataset can be formulated as:

$$\mathcal{D}'_S = \mathcal{S}(\mathcal{A}(\mathcal{D}_S), C) \quad (4)$$

where $\mathcal{S}(\cdot)$ represents speech synthesis and $\mathcal{A}(\cdot)$ denotes ASR. This design ensures that any remaining performance drop can be attributed solely to the synthesis artifacts rather than the semantic content.

As shown in Fig. 4, the confusion matrix of two TTS models and two S2S LLMs on synthesized TESS dataset continues to exhibit a lack of diagonal structure. This suggests that performance degradation of the SER model on synthesized speech does not stem from a text distribution gap. Instead, it likely arises from synthesis artifacts introduced by the speech generation process. All synthesized audio employed in this section and below was synthesized from a real human speech transcript.

4.1.4 Hypothesis 4: Distribution Gap between Synthesized and Human Speech

Prior experiments suggest that a SER model trained on human speech does not generalize to understanding tasks on synthesized audio. Even when ASR transcripts are identical, conveying the same semantics and content, there remains a mismatch between the audio distributions produced by natural speech production and by the speech synthesis process.

To further validate the distribution gap between synthesized and human speech, we fine-tune a pretrained audio representation model, Emotion2vec-base (Ma et al., 2024), using either human or synthesized data under controlled settings. Specifically, for real speech we use three datasets: IEMOCAP, RAVDESS, and TESS. We then generate a synthesized counterpart by first transcribing the real utterances with ASR, performing emotion-conditioned speech synthesis from the transcripts, and filtering out samples with insufficiently perceiv-

	Human			Synthesized	
	IEM.	RAV.	TESS	IEM.	TESS
I. Emotion2vec Train on Human Speech					
Ang	77.23	81.25	97.50	63.27	50.00
Hap	64.14	56.25	100.00	80.30	0.00
Neu	71.98	100.00	100.00	43.24	0.00
Sad	75.68	50.00	100.00	24.59	0.00
Fea	-	93.75	100.00	-	55.00
Dis	-	81.25	100.00	-	0.00
Sur	-	93.75	100.00	-	3.57
WA	71.28	77.88	99.64	49.58	15.31
II. Emotion2vec Train on Synthesized Speech					
Ang	48.51	25.00	0.00	85.71	97.22
Hap	16.67	0.00	0.00	74.24	70.59
Neu	71.98	37.50	100.00	96.22	94.59
Sad	87.16	100.00	2.50	86.89	92.59
Fea	-	0.00	0.00	-	85.00
Dis	-	0.00	2.50	-	96.77
Sur	-	0.00	0.00	-	92.86
WA	55.67	22.12	15.00	89.20	91.84

Table 1: SER performance on real or synthesized speech while trained with human or synthesized speech data.

able emotional expression. Because RAVDESS contains substantial transcript duplication, the resulting synthesized set is relatively small; therefore, all synthesized RAVDESS samples are used for training, without holding out a test split.

As shown in Table 1, models trained on real speech perform well on real audio but degrade sharply on synthesized audio. Conversely, models trained on synthesized audio classify synthesized utterances more accurately yet generalize poorly to real speech. These results confirm a pronounced distribution gap between synthesized and human speech, and highlight the limitation of directly reusing SER models trained on real speech for synthesized audio understanding.

4.2 Commercial Synthesis Models

SER models trained on human audio exhibit significant performance degradation on synthetic audio, even the emotional expression of the synthesized audio is manually verified. Given that open-source models frequently utilize inaccurate labels by Emotion2vec as RL training rewards or evaluation metrics, we question whether this phenomenon persists in leading commercial models and whether this discrepancy contributes to the performance gap between open-source and commercial models.

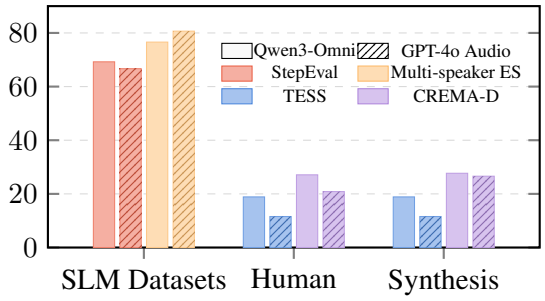

Figure 5: Speech emotion recognition results of SLMs. Solid colors and shaded areas represent the results of Qwen3-omni and GPT-4o Audio respectively.

To this end, we employ the TTS model `GPT-4o mini TTS` and the S2S model `GPT-4o Audio` to synthesize natural, realistic, and emotionally expressive speech, subsequently evaluating the SER models performance. The results show that the gap persists despite stronger affective expressiveness: on CREMA-D, Emotion2vec achieves 32.94% accuracy on gpt-4o-tts outputs and 52.63% on gpt-4o-audio outputs. On the more lexically homogeneous TESS dataset, accuracy improves but remains low at 46.42% and 64.16%, respectively. More detailed results are provided in Appendix D. These findings indicate that even for leading commercial synthesizers, current human-trained SER models do not reliably recognize emotion in synthesized speech.

4.3 SER Results with SLMs

SLMs are central to voice-centric interaction and are believed to have strong speech and emotion understanding capabilities. This section explores the performance of advanced SLMs on both human and synthesized speech datasets. We selected two SLMs including the open-source Qwen3-Omni (Xu et al., 2025b) and the commercial GPT-4o Audio. In addition to TESS and CREMA-D, we also include two datasets used to assess the emotion understanding of SLMs: StepEval-Audio-Paralinguistic (Wu et al., 2025b) and the Multi-Speaker Emotional Speech Dataset[2].

As shown in Fig. 5, both models achieve high SER accuracy on the two SLM datasets, with an average of 73.3%. However, on TESS and CREMA-D, emotion recognition accuracy drops significantly, averaging only 19.6%. The models also perform poorly on synthesized data.

[2] https://magichub.com/datasets/multi-speaker-emotional-speech-dataset/

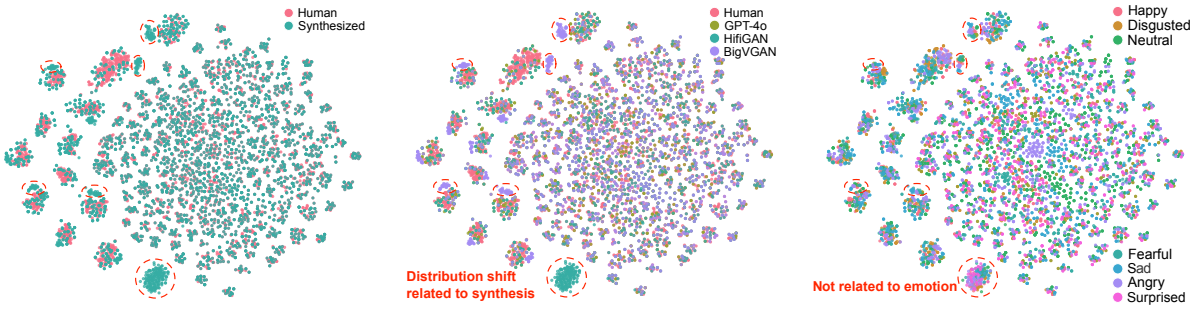

Figure 6: Visualization of the Emotion2vec embedding space, distinguished by synthesis type, vocoder, and emotion.

Interestingly, this phenomenon suggests that emotion understanding of SLMs relies more on semantic content than on paralinguistic cues. In TESS and CREMA-D, the same text is read with different emotions, requiring emotion inference from audio's paralinguistic features. In contrast, audio in StepEval and Multi-speaker SE can typically be interpreted based on the transcribed text. Furthermore, prompt engineering aimed at directing SLMs to focus more on paralinguistic information does not have an effect (specific prompts are provided in the Appendix B). This highlights the 'text dominance' in SLMs (Wu et al., 2025c) and suggests the need for further enhancement of their ability to process paralinguistic features.

5 Deep Analysis

5.1 Representation Space Bias

To investigate the significant performance drop of Emotion2vec on synthesized audio, we hypothesize that it is related to the model's representation space. The pretraining objective of Emotion2vec focuses on distilling speech representations from the teacher model, data2vec, via self-supervised learning. However, the pretraining data does not include synthesized audio, and data2vec (Baevski et al., 2023) does not model synthetic data well. Consequently, a gap in the representation space may exist between synthesized and real speech, causing fine-tuned classification heads to be more influenced by distributional differences in the synthesized audio rather than emotional features.

Representation space shift of synthetic audio. We begin by visualizing the representation space of Emotion2vec. Specifically, we encode real and synthesized data from the TESS and CREMA-D datasets, then reduce the representation to two dimensions using t-SNE. Three coloring schemes are used to distinguish samples: whether the data are synthesized, the audio's vocoder, and emotion.

Metric(%)	Syn.	Voc.	Mod.	Emo.
Balanced Acc	99.81	96.25	75.01	59.86
Macro-F1	99.81	96.25	76.74	59.19

Table 2: Linear probing results on Emotion2vec representation space. Target probing categories including synthesis / human (Syn.), vocoder name (Voc.), model name (Mol.), and emotion label (Emo.).

Fig. 6 supports our hypothesis. Subfigure 1 shows that the clusters of synthesized data are offset from those of real data, marked by red dashed circles. These offsets are attributed to specific synthesis models, such as CosyVoices 2 with the HiFi-GAN vocoder at the bottom and IndexTTS 2 with the BigVGAN vocoder in other clusters. Subfigure 3 reveals that these outlier clusters encompass all emotions, indicating that the offsets are model-specific rather than emotion-related. Thus, classification models trained in this representation space are likely influenced by differences between synthesis models, hindering the extraction of emotion.

Embeddings capture more synthesis-specific patterns. Probing experiments further corroborate this finding. Using a dataset balanced across emotion labels and synthesis models, we attached linear probes to frozen Emotion2vec embeddings to predict four targets: authenticity, vocoder, synthesis model, and emotion. As shown in Table 2, the probes easily discriminate between real and synthesized audio, vocoders, and even synthesis models, but perform poorly on emotion classification. These results motivate us to reshape the Emotion2vec representation space to disentangle synthesis-related artifacts from emotional features.

5.2 Failure of Generalization

Table 1 reveals a pronounced distribution gap between synthesized and human speech. Meanwhile,

Method	Human Speech					Synthesized Speech			
	In Domain					In Domain		OOD	
	IEMOCAP	RAVDESS	TESS	MELD	CREMA-D	IEMOCAP	TESS	CREMA-D †	CREMA-D ‡
MLP	**71.13**	66.35	**99.64**	**50.15**	**78.43**	**89.20**	88.27	32.77	49.38
DANN_syn	69.22	**70.19**	98.21	47.85	72.10	84.21	84.18	47.30	**50.21**
DANN_vocoder	69.07	69.23	97.14	47.24	71.70	84.49	85.20	**51.35**	46.88
DANN_model	69.37	68.27	98.21	47.97	71.96	81.99	83.16	41.22	43.87
DANN_syn*	67.60	69.23	98.93	47.85	73.16	86.43	**91.33**	42.91	45.01
DANN_vocoder*	70.54	65.38	97.86	47.12	72.10	84.21	88.27	44.93	46.78
DANN_model*	69.81	69.23	98.21	49.00	71.43	85.04	89.80	46.28	47.51

Table 3: Under different fine-tuning strategies, the SER models fail to demonstrate generalization, both to cases where training data includes real CREMA-D and speech synthesis by CosyVoice and Kimi-audio but lacks CREMA-D synthesis with these models (OOD CREMA-D †), and to unseen synthesis models (OOD CREMA-D ‡).

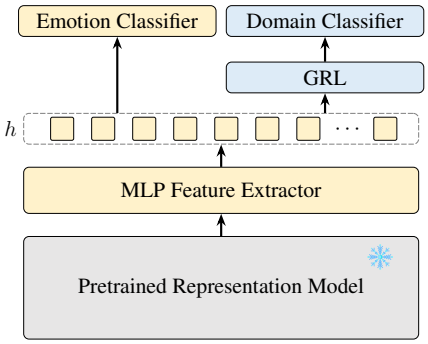

Figure 7: Domain adversarial fine-tuning. An MLP further extracts features and is used for classification in understanding tasks. Meanwhile, a GRL is employed to prevent shortcut learning through artifact features.

labeled synthetic speech is scarce: although fine-tuning uses 40k labeled human speech, only 4k synthetic speech remain after manual validation and filtering of weak emotional expression. A natural approach is to fine-tune on a mixture of both domains, since Table 1 suggests that large-scale human speech pre-training combined with a small amount of labeled synthetic data can yield strong synthesis SER performance, as illustrated in Tabel 1. Moreover, biases in the Emotion2vec representation space motivate disentangling synthesis-related artifacts from emotional cues. Inspired by domain adversarial neural networks (Ganin et al., 2016), we learn latent representations that support the emotion classifier while confusing a domain classifier.

Domain adversarial fine-tuning. As shown in Fig. 7, the input text is first processed by a pre-trained Emotion2vec, and we keep the pretrained parameters frozen to prevent catastrophic forgetting. The output embeddings are then fed into a trainable MLP Feature Extractor. This module projects the fixed embeddings into a task-specific latent representation, denoted as h, which serves as the shared feature space for subsequent tasks. Then an emotion classifier is optimized to predict the correct emotion label y by minimizing the cross-entropy loss $\mathcal{L}_{emo}$. To achieve disentanglement, we train an adversarial classifier so that the learned representation is invariant to the domain label d. In practice, we implement this using a Gradient Reversal Layer (GRL, Ganin and Lempitsky, 2015), which negates the gradients flowing from the domain classifier to the MLP during backpropagation. The overall training objective is to learn a representation h that is discriminative for emotion recognition but invariant to domain shifts.

In experimental settings, the human speech data comprise IEMOCAP, RAVDESS, TESS, MELD, and CREMA-D. For synthesized speech, we transcribe IEMOCAP and TESS into text and synthesize audio using CosyVoice2 and Kimi-Audio. All data are split into training and test sets. The test set additionally includes two out-of-domain (OOD) partitions, CREMA-D† and CREMA-D‡, to assess generalization. Notably, training does not include synthesized CREMA-D, but it does include human CREMA-D and synthesized audio from other datasets produced by CosyVoice2 and Kimi-Audio; thus, CREMA-D† evaluates the *compositional generalization*. CREMA-D‡ evaluates generalization to unseen synthesis models, including IndexTTS2, GLM4-voice, gpt-4o-tts, and gpt-4o-audio.

Generalization limits of fine-tuned SER. As shown in Table 3, simply training an MLP on mixed-domain data yields high in-domain accuracy, averaging 73.14% on human speech and 88.73% on synthesized speech, yet generalization remains poor. Performance drops sharply on OOD CREMA-D† and CREMA-D‡. Although we adopt

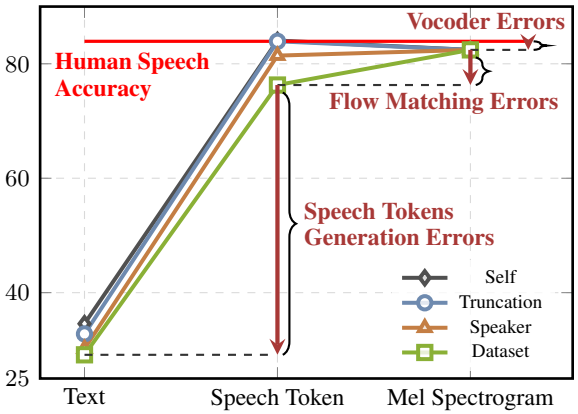

Figure 8: The impact of three synthesis stages. The horizontal axis represents emotion recognition accuracy and deep red arrows indicate the error at each stage.

domain-adversarial training with a GRL to prevent the latent representation h from encoding synthesis attributes, and the rising domain loss $\mathcal{L}_{\text{domain}}$ suggests that this objective is being enforced. However, contrary to our expectation, this does not improve OOD generalization for SER. Balancing positive and negative samples within each mini-batch also fails to help. These results suggest that SER models still exploit in-domain shortcuts rather than learning fundamental emotion features.

5.3 Token Prediction Dominates the Bias

Furthermore, we investigate the contribution of each stage in the speech synthesis process to the distributional bias. Recent speech synthesis (Xie et al., 2025; Cui et al., 2025) typically consists of three stages: (1) an autoregressive (AR) language model generating discrete speech tokens; (2) a flow matching process synthesizing mel-spectrograms conditioned on the speech tokens and a reference audio; and (3) a vocoder converting mel-spectrograms into waveforms. To investigate the impact of each stage on SER accuracy, we conducted an ablation study using CosyVoice 2 on the TESS dataset. We isolated the error contribution of each stage as follows:

- Vocoder Error: We reconstructed audio from ground-truth mel-spectrograms. The performance drop compared to real speech represents the vocoder-induced error.

- Flow Matching Error: We synthesized mel-spectrograms using ground-truth speech tokens derived by the tokenizer as conditions. The performance gap compared to reconstruction from mel-spectrograms reflects the error introduced by flow matching process.

- Speech Tokens Generation Errors: We generated speech tokens from ASR-transcribed text, and synthesis speech with the following two stages. The discrepancy with the ground-truth token setup isolates the error introduced by the auto regressive language model.

Synthesis gap arising from speech token error. To simulate the transition from ideal to realistic synthesis scenarios, we employed four types of reference speech: (1) the real speech itself, (2) truncated real speech with half length, (3) cross-speaker within TESS, and (4) out-of-domain speaker in CREMA-D. As shown in Fig. 8, real speech achieves an SER accuracy of 83.93%, while Mel-spectrogram reconstruction reaches 82.43%, indicating that the vocoder has a negligible impact on emotion recognition. When using ground-truth tokens with the original audio as a reference, accuracy reaches 84.07%, slightly surpassing real speech. Averaged across the four reference settings, flow matching introduces slightly more error than the vocoder but remains a minor factor.

In contrast, driving the synthesis with text inputs utilizing the AR language model causes a drastic performance drop to approximately 30%. Even using the target audio itself as a reference yields only 34.57%, improving to just $\sim$ 50% after manually filtering for emotional expressiveness. While Mel-based reconstruction remains unaffected by reference selection, accuracy for text-based generation declines progressively from ideal to realistic reference conditions. Consequently, our results conclusively show that the discrepancy between AR-generated tokens and ground-truth tokenizer outputs is the dominant factor degrading emotion recognition performance in synthesized speech.

6 Conclusion

This paper highlights a clear gap between human and synthesized speech, leading to poor emotion recognition performance in discriminative Speech Emotion Recognition (SER) models on synthesized audio. This generalization issue arises from the speech token prediction stage in synthesis, which induces a representation mismatch. Additionally, generative speech LLMs rely on textual semantics to recognize emotion, often neglecting paralinguistic features. Overall, our findings suggest that existing SER models exploit non-robust shortcuts rather than capturing intrinsic emotional features.

Limitation

While our work offers insights into the generalization gap between human and synthesized speech in emotion recognition, several limitations remain. Firstly, to ensure the validity of our evaluation, we relied on manual annotation to filter out synthesized utterances that failed to convey the target emotion. This process introduces potential human subjectivity and restricts the scale of our synthetic testing data. Secondly, although we attempted to bridge the domain gap using Domain Adversarial Neural Networks, our results indicate that this approach yields limited generalization to unseen synthesis models. Future work should focus more on pretraining better representation models.

Ethical Considerations

Potential Risks. Our work involves the analysis of synthesized speech with emotional paralinguistic features. We acknowledge that advancements in emotional TTS and S2S models carry potential risks, including the creation of deepfakes for fraud, impersonation, or manipulation. However, the primary goal of this research is to critically evaluate the limitations of current Speech Emotion Recognition (SER) models in detecting and understanding these synthetic artifacts. By highlighting the "human-synthesis gap" and the vulnerabilities of SER models to token-prediction errors, our work contributes to the development of more robust detection systems and safer AI interactions.

Human Annotation and Fair Compensation. To ensure the validity of our synthesized datasets, we recruited four human annotators to manually verify the emotional expressiveness of the generated audio clips.

- **Recruitment and Demographics:** The annotators were recruited from graduate students from our research group. They are proficient in English with normal hearing capabilities.
- **Task and Instructions:** Annotators were instructed to listen to synthesized audio samples and determine if the target emotion (e.g., angry, happy, fearful) was clearly perceptible. Samples deemed ambiguous were discarded.
- **Compensation:** All participants were compensated for their time. The payment rate was set at ¥22.5 per hour, which exceeds the local minimum wage in China.
- **Consent and Data Privacy:** We obtained informed consent from all annotators. They were informed that the task involved listening to emotional speech (which may include negative emotions like anger or fear) and that they could withdraw from the study at any time without penalty. No personally identifiable information about the annotators was collected or stored in the final dataset.

Data Usage. The real human speech data used in this study comes from publicly available datasets. We strictly adhered to the usage licenses and terms of these datasets.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR.

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.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025a. Emova: Empowering

language models to see, hear and speak with vivid emotions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5455–5466.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. 2025. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6981, Vienna, Austria. Association for Computational Linguistics.

Eunjeong Koh and Shlomo Dubnov. 2021. comparison and analysis of deep audio embeddings for music emotion recognition. In *CEUR Workshop Proceedings*, volume 2897, pages 15–22. CEUR-WS.

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.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.

M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS).

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.

Samir Sadok, Simon Leglaive, and Renaud Séguier. 2023. A vector quantized masked autoencoder for speech emotion recognition. In *2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW)*, pages 1–5. IEEE.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Dingdong Wang, Junan Li, Mingyu Cui, Dongchao Yang, Xueyuan Chen, and Helen M. Meng. 2025a. Speech discrete tokens or continuous features? a comparative analysis for spoken language understanding in SpeechLLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24913–24924, Suzhou, China. Association for Computational Linguistics.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025b. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025a. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025b. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025c. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. 2025. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *ICLR*. OpenReview.net.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Appendix

A Evaluation Details

In this paper, we report the accuracy of speech emotion recognition models on diverse datasets, including both human and synthesized speech. Since the inference process of Emotion2vec is not affected by random seeds, we run inference only once and report the average accuracy of SER models over each dataset. All experiments were conducted with 8 NVIDIA GeForce RTX 3090 GPUs.

We employ 4 manual annotators to evaluate all synthesized audio samples in data filtering process. The detailed instructions are shown below:

- **Task Overview**

  For each example, you will be given:

  - an audio clip S (about 10 seconds), and
  - a target emotion label $L \in \{$`surprised`, `angry`, `sad`, `disgusted`, `fearful`, `happy`, `neutral`$\}$.

  Your goal is to judge whether the audio S **saliently expresses** the target emotion L. Firstly, listen to the audio. Then decide YES/NO.

  - Select **YES** if the dominant emotion is **consistent with** L and the expression is **clear and strong enough**.
  - Otherwise, select **NO**.

- **Output Format**

  For each example, selects **YES** or **NO**

- **Reference Cues for Each Emotion**

  - `angry`: high energy, tense/pressed voice, strong stress, possible shouting.
  - `disgusted`: clear aversion/contempt, "ew"-like quality, scoffing tone.

- `fearful`: nervous/tense, unstable or trembling voice, rapid breathing, high-/unstable pitch, panic-like urgency.
- `happy`: bright and lively tone, upward intonation, relaxed energy, possible laughter or smiling voice.
- `neutral`: steady, controlled delivery with minimal affect; no strong emotional coloration.
- `sad`: low energy, slower pace, downward intonation, heavy/flat tone, possible sighing quality.
- `surprised`: sudden pitch rise, short burst/exclamation, clear "unexpected" reaction with abrupt prosodic change.

B Speech LLM Prompt for SER

The following shows the prompt used for inference in Speech Large Language Models (SLMs). If you wish to obtain the results reported in the paper, please remove the first two rules from the Rules section. To verify whether prompt engineering can alter the tendency of SLMs to overly rely on semantic content when judging emotions, we used the complete prompt below.

> You are an audio emotion classification model.
>
> Your task: Given the following audio input, classify the speaker's emotion.
>
> Emotion categories (choose exactly ONE):
> - angry
> - disgust
> - fearful
> - happy
> - neutral
> - sad
> - surprised
>
> Rules:
> 1. Completely ignore the textual content of the speech. The transcript may be misleading or emotion-neutral.
> 2. Judge emotion only from acoustic characteristics, not semantics.
> 3. Do NOT default to "neutral" when uncertain. Instead, choose the closest non-neutral category unless the audio is clearly flat, monotone, and emotionless.
> 4. Output only the emotion label (all lowercase, no punctuation).
> 5. Do not output explanations unless explicitly requested.
>
> Output format:
> <emotion_label>

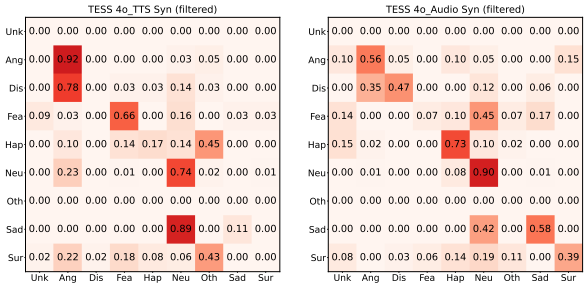

Figure 9: Confusion matrix of speech emotion recognition results. All samples are synthesized by GPT-4o TTS and GPT-4o Audio.

C Speech Synthesis in S2S LLMs

We employ Kimi-Audio and GLM-4-Voice to synthesize speech with salient emotional features for testing the classification ability on S2S LLMs synthetic data. Unlike TTS models which explicitly control emotion through prompt speech and natural language instructions, S2S LLMs generate speech in a conversational manner. During the interaction process, however, we find that the emotional expression in the first-turn response is insufficient. Hence, we introduce additional prompts in subsequent turns, such as 'Repeat the previous answer and speak with a clearly angry tone'. Through this iterative prompting strategy, we are able to obtain audio samples with more pronounced emotional features in the second or third-turn response. These audio samples are manually filtered and subsequently used as synthetic data for testing the emotion classification models.

D Emotion2vec Performance on Commercial Synthesized Models

As shown in Fig. 9, we present detailed confusion matrix for speech emotion recognition on the TESS dataset synthesized by two commercial systems: GPT-4o-TTS and GPT-4o-Audio. Despite the strong synthesis ability of commercial models and the salient emotional features of filtered audio samples, the classification performance of Emotion2vec remains unreliable.

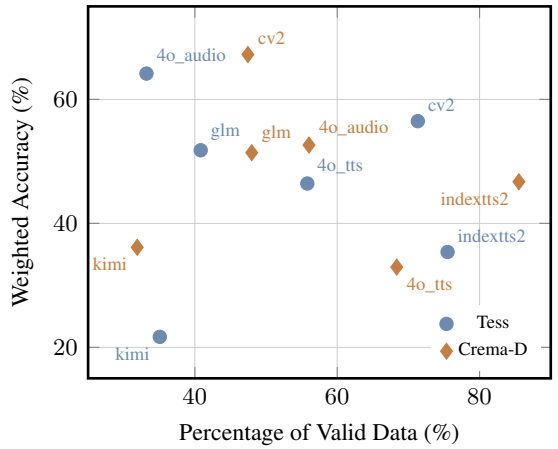

Figure 10: The horizontal axis represents the proportion of data where humans perceive significant emotional expression, while the vertical axis indicates the Emotion2vec SER accuracy.

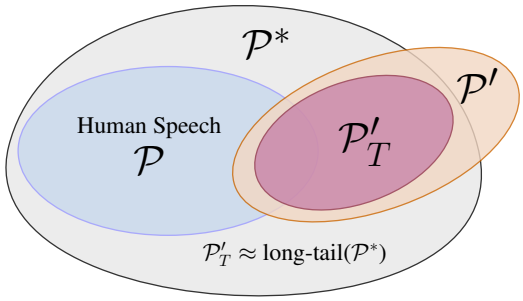

Figure 11: The significance of synthetic data.

E Emotion Recognition Relevance

We visualized the relationship between the significance of emotion expression in synthetic models and the emotion recognition accuracy of Emotion2vec. As shown in Fig. 10, the x-axis represents the proportion of synthetic audio in which human annotators believe the emotion is significantly expressed. The y-axis represents the emotion recognition accuracy of Emotion2vec. The visualization in Figure 10 shows that there is no strong positive correlation between the two.

F The Significance of Synthetic Data in Ideal Scenarios

We investigate the unique value of synthetic data in speech understanding tasks. For instance, can synthetic audio, by supplementing long-tail distribution data, enable neural networks to uncover more intrinsic features relevant to the task?

As shown in Fig. 11, consider a scenario where authentic speech-emotion data follows a common distribution $\mathcal{P}$. Conversely, certain synthetic audio-emotion data, influenced by the generative process, may form a less common distribution $\mathcal{P}'$. Within this, a subset $\mathcal{P}'_T$ represents synthetic data where emotions remain discernible to humans. We posit that $\mathcal{P}'_T$ effectively constitutes the long-tail portion of the true, complete distribution $\mathcal{P}^*$. Consequently, we investigate whether leveraging the more comprehensive distribution formed by combining $\mathcal{P}$ and $\mathcal{P}'_T$ allows for the discovery of more intrinsic features for emotion recognition, thereby transcending the limitations of relying solely on the common distribution $\mathcal{P}$ Furthermore, we explore the potential of leveraging scaling laws to enhance the model's representation of the true distribution $\mathcal{P}^*$ through the use of massive pseudo-labeled synthetic data. Recent work has demonstrated that in language modeling tasks (Zeng et al., 2024), utilizing large-scale synthetic data facilitates the learning of more robust speech-text alignment relationships. Inspired by this, we aim to investigate the potential of scaling pseudo-labeled synthetic data specifically within the context of speech understanding tasks.