Prompt-Guided Selective Masking Loss for Context-Aware Emotive Text-to-Speech

Yejin Jeon[1], **Youngjae Kim**[1], **Jihyun Lee**[1], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{jeonyj0612, yj122198, jihyunlee, gblee}@postech.ac.kr

Abstract

Emotional dialogue speech synthesis (EDSS) aims to generate expressive speech by leveraging the dialogue context between interlocutors. This is typically done by concatenating global representations of previous utterances as conditions for text-to-speech (TTS) systems. However, such approaches overlook the importance of integrating localized acoustic cues that convey emotion. To address this, we introduce a novel approach that utilizes a large language model (LLM) to generate holistic emotion tags based on prior dialogue context, while also pinpointing key words in the target utterance that align with the predicted emotional state. Furthermore, we enhance the emotional richness of synthesized speech by incorporating concentrated acoustic features of these key words through a novel selective audio masking loss function. This methodology not only improves emotional expressiveness, but also facilitates automatic emotion speech generation during inference by eliminating the need for manual emotion tag selection. Comprehensive subjective and objective evaluations and analyses demonstrate the effectiveness of the proposed approach.

1 Introduction

Communication between humans is built on the ability to understand, and empathize in alignment to emotional context (Pangaro and Dubberly, 2014). As human-robot interactions continue to grow in prominence through the rise of Intelligent Personal Assistants (IPAs) like Siri and Alexa[1], it is crucial to develop artificial agents that are able to emulate these communicative qualities. This involves generating artificial speech with appropriate empathetic and emotive prosodic patterns that accurately reflect the dialogue context, a task known as emotional dialogue speech synthesis (EDSS; Fig. 1).

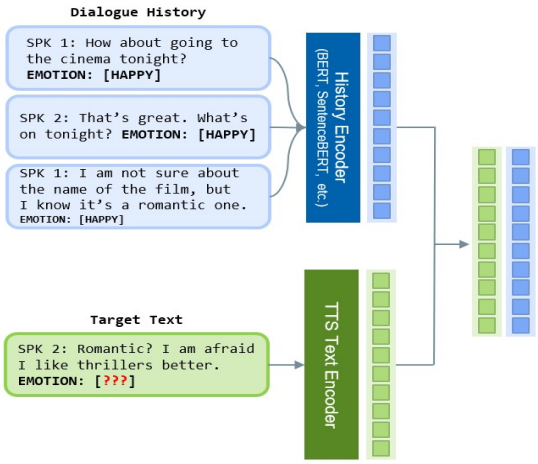

Figure 1: Conceptual illustration of the typical approach utilized for the EDSS task. The preceding dialogue history between two interlocutors serves as the foundation for determining the style in which the subsequent text utterance should be spoken.

EDSS is an interdisciplinary field that merges two main research domains: emotive speech synthesis (ESS) and dialogue speech synthesis (DSS). The first task of ESS focuses on generating speech in a particular emotion style. An intuitive ESS approach involves using discrete emotion labels (Lee et al., 2017; Lorenzo-Trueba et al., 2018; Murata et al., 2024) or a lookup table (Lei et al., 2022) as additional conditions for the text-to-speech (TTS) model. Alternatively, other studies use a reference audio to determine emotive style; a fixed emotion representation is extracted from the reference audio through pretrained speech emotion recognition (SER) models (Tang et al., 2022; Zhou et al., 2021; Tang et al., 2024), attribute disentanglement with auxiliary emotion classification constraints (Zhu et al., 2023; Kang et al., 2023; He et al., 2022), or weighted cluster interpolation (Wu et al., 2019; Um et al., 2020). However, these approaches require manual selection of an emotion label or a reference audio by the user during inference, as the

[1]The number of virtual assistants is forecasted to reach 8.4 billion units by 2024 (Federica Laricchia, 2024).

Findings of the Association for Computational Linguistics: NAACL 2025, pages 638–650

TTS model lacks the reasoning capabilities to autonomously determine the most suitable emotional condition for the target text.

Unlike ESS, the task of DSS incorporates previous conversational utterances between two interlocutors to automatically determine through reasoning capabilities, the overall style in which the subsequent text input should be articulated. Towards this, previous utterances are utilized as additional input to the backbone TTS framework (Liu et al., 2023b). Typically, semantic information from each of the previous utterances are extracted and concatenated together using a pretrained language model like BERT (Guo et al., 2021; Lee et al., 2023). The combined representation then acts as the holistic stylistic condition for the target text input. However, these methods rely solely on textual information, which only includes semantic context. As such, to improve prosodic learning, some studies further utilize acoustic context through cross-modal attention (Cong et al., 2021) or multiple modality-specific encoders (Li et al., 2022; Xue et al., 2023; Li et al., 2024). Nonetheless, these approaches lack the granularity needed to reflect the nuances of human dialogue, where certain words or phrases naturally carry more emotional weight (Patrik N. and Laukka, 2001; Lee and Narayanan, 2005; Gu et al., 2018; Jia1 et al., 2024).

In this paper, we propose a novel EDSS framework that simultaneously addresses both emotion and context-aware dialogue speech synthesis. Moreover, to overcome the limitations seen in prior ESS and DSS research, our approach leverages the sophisticated reasoning and contextual understanding capabilities of large language models (LLMs) to automatically generate the appropriate holistic emotion label for the current text based on the dialogue history. Moreover, to mimic human speech generation and therefore produce emotive speech, we employ LLMs to identify local key words, or rationales within the current text that correspond to the target emotion. In addition, to enhance the learning of paralinguistic attributes, we introduce a novel selective masking loss, which ensures that only the acoustic information for these selected rationales are provided to the backbone TTS model.

In summary, our contributions are as follows. First, unlike traditional methods that require manual annotation of emotion labels, our model automatically generates these labels using LLMs, which reduces user dependency and enhances the model's autonomy. Second, by pinpointing and integrating acoustic information for specific emotion-inducing words and phrases, our method improves the emotive quality of synthesized speech, making it more natural and contextually appropriate. Specifically, emotion salience is enhanced via a novel selective masking loss. The effectiveness of the proposed method against other baselines is validated through subjective and objective validations. In addition, the effectiveness of the proposed method against other baselines is validated through subjective and objective validations.

2 Related Work

2.1 Style Control in Speech

Current TTS models (Shen et al., 2018; Ren et al., 2019; Kim et al., 2021a) have reached a level of sophistication where they are able to generate artificial speech that sounds remarkably human. Such progress has catalyzed efforts to address more complex tasks, such as modifying audio to simultaneously mimic specific individuals' voices (Choi et al., 2022; Valle et al., 2020; Jeon et al., 2024), or making synthetic speech more emotive. Whereas traditional TTS models typically have a single input (i.e., the target text to convert into audio), such stylistic speech generation requires an additional input for stylistic control, resulting in two inputs.

The additional stylistic input for TTS models can be either an explicit label or an audio sample. In the label-based approach, predefined styles, such as specific timbres using speaker IDs (Chen et al., 2020, 2021) or emotion labels (Kim et al., 2021b), are encoded and concatenated with the embedding of the target text. However, this method is limited by its inability to support zero-shot or few-shot stylistic generation for unseen speakers or emotions, as discrete labels often fall short of capturing the intricate nuances of auditory styles. To address this, some studies incorporate audio as the stylistic input (Wang et al., 2018; Yan et al., 2021), allowing for unsupervised learning of the nuanced acoustic and prosodic characteristics of a speaker's voice or intonation. Nevertheless, both approaches still require the user to manually select the appropriate emotion label or audio sample.

2.2 Prompting in TTS

Given the recent popularity of LLMs, there has been growing interest in leveraging these models within the speech domain. Specifically, these efforts focus on utilizing LLMs for dataset curation,

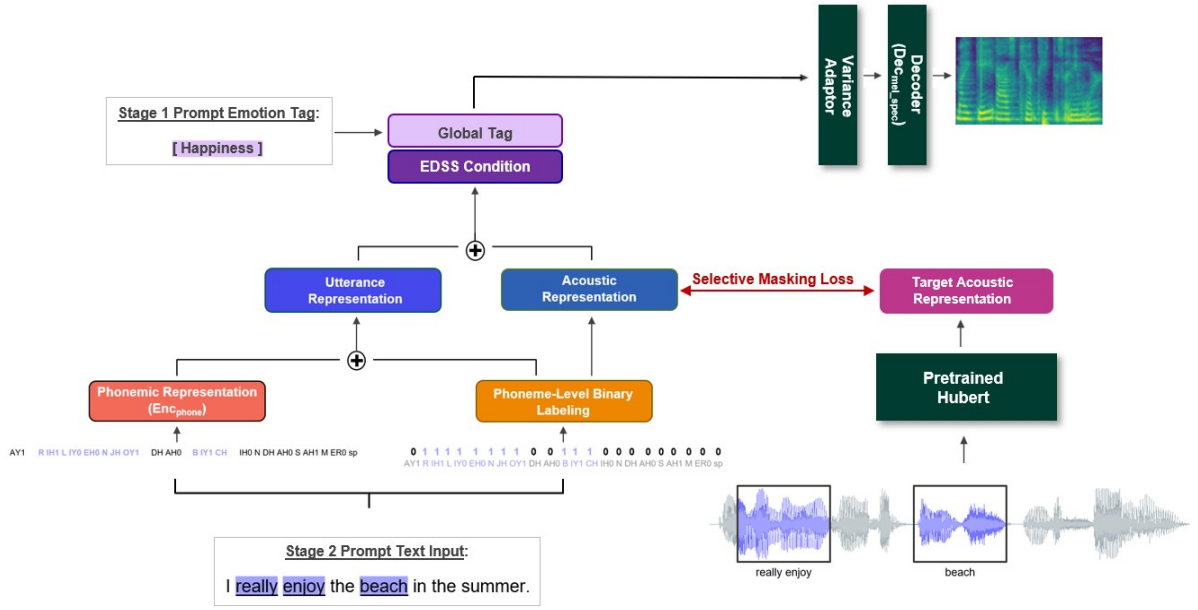

Figure 2: Architecture of the proposed model. Text- and acoustic-based binary rationales are integrated at the beginning with $\text{Enc}_{\text{phone}}$, whereas the holistic tag is incorporated at the end, which produces the final condition to the variance adaptor. Darker sections in the figure represent components that are used without modification.

where individual audio samples are annotated with corresponding emotion or stylistic tags. For example, Saito et al. (2023) employed ChatGPT to generate emotion, intention, and style tags for each utterance within a dialogue history. These tags are subsequently embedded using BERT and concatenated to form a comprehensive context vector, which conditions the backbone TTS model. Similarly, Guo et al. (2023); Liu et al. (2023a) developed approaches where text descriptions were generated for each audio sample and embedded with a BERT (Devlin et al., 2019) model. The output representation was then used as the style conditioning input for the TTS system. Yet, these methods require extensive dataset curation (Guo et al., 2023; Yang et al., 2024; Yoon et al., 2022). To address the challenges associated with dataset limitations, Sigurgeirsson and King (2024) prompted an LLM to directly predict relative fundamental frequencies and energies for each word in a target text utterance on a scale from 0 to 5. Additionally, global pitch, energy, and duration values were predicted on a scale ranging from -5 to 5. In contrast to this numerical prediction of prosodic attributes, our approach adopts a more intuitive method that better aligns with the inherent strengths of LLMs in text processing. Specifically, we focus on utilizing the words and phrases that are pivotal in conveying emotion, thereby enhancing the emotive expressiveness of the generated speech.

3 Methodology

3.1 Preliminaries

The backbone TTS model is the non-autoregressive FastSpeech2 (Ren et al., 2021). Consider an input utterance $\text{Utt} = [W_1, W_2, W_3, \ldots, W_n]$, consisting of N words. Each word W is further decomposed into its phonemic representations, $W_n = [w_{n1}, w_{n2}, \ldots, w_{np}]$[2]. The phonemic encoder $\text{Enc}_{\text{phone}}$, which is comprised of a feed-forward Transformer block and a 1D convolution layer, transforms the phonemic embeddings into their corresponding hidden phonemic states. The variance adaptor then integrates variable duration, pitch, and energy information into these phonemic states of $\text{Enc}_{\text{phone}}$. The resulting representation is subsequently fed into the mel-spectrogram decoder $\text{Dec}_{\text{mel\_spec}}$, which mirrors the composition of $\text{Enc}_{\text{phone}}$. The final output is a predicted mel-spectrogram. Since the architecture of the variance adaptor and $\text{Dec}_{\text{mel\_spec}}$ remains unmodified in our approach, interested readers should refer to Ren et al. (2021) for specific implementation details. In the following subsections, we explain the proposed EDSS framework in detail. The complete architecture is illustrated in Figure 2.

[2]An example of the input sentence and its corresponding phonemic representation can be seen in the lower left-hand corner of Figure 2.

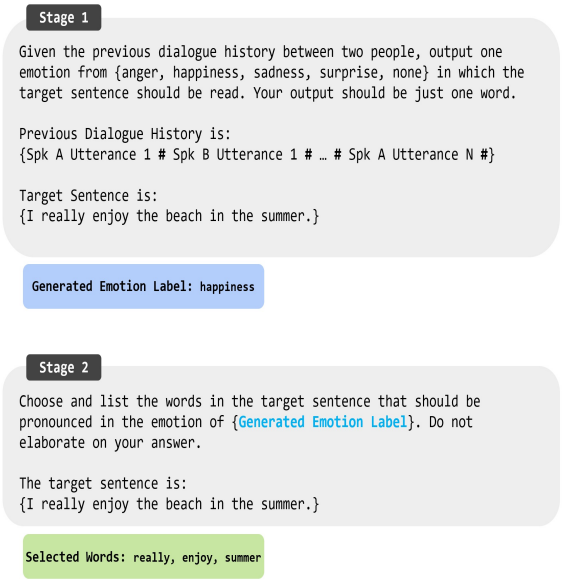

Figure 3: Overview of the dual-stage prompts. In Stage 1, a single emotion tag is predicted for the upcoming target utterance based on the dialogue history. In Stage 2, a list of specific words that align with the predicted emotion tag is generated.

3.2 Dual-Stage Prompting

Given a prior conversation history between two speakers, which is denoted as $\mathrm{H} = [\mathrm{Utt}_1, \mathrm{Utt}_2, \mathrm{Utt}_3, \ldots, \mathrm{Utt}_{T-1}]$, our initial objective is to derive a holistic emotion tag C for which the subsequent target utterance $\mathrm{Utt_T}$ should be articulated. To obtain this global emotion tag C, we utilize the GPT 3.5 API[3] by prompting the model with both the conversation history H and the forthcoming target utterance $\mathrm{Utt_T}$. Each utterance within H is delimited by a special character '#', which clarifies speaker transitions and accommodates utterances that may span multiple sentences. This process results in the generation of a single emotion tag selected from a predefined set of emotions, and is referred to as Stage 1 prompting.

As mentioned in Section §1, while a global emotion tag C is able to guide the overall style delivery about how the target text should be articulated, actual emotive speech often necessitates a more granular approach; human speech typically involves emphasizing certain words or phrases to convey their emotions more effectively than neutral speech (Patrik N. and Laukka, 2001; Lee and Narayanan, 2005; Gu et al., 2018). To address this, we implement a secondary prompting stage. Here, GPT 3.5 is tasked with identifying and highlighting specific words or phrases within non-neutral target utterances $\mathrm{Utt_T}$ that should be emphasized in accordance with the global emotion tag C, which was obtained from the previous prompting stage. This process yields a list of $\mathrm{S_{Words}} = [\mathrm{W}_{i_1}, \mathrm{W}_{i_2}, \ldots, \mathrm{W}_{i_k}]$, where $\{i_1, i_2, \ldots, i_k\}$ are the indices of the words selected from the target utterance $\mathrm{Utt_T}$. The overall prompting procedures are provided in Figure 3.

Before incorporating the context-aware information derived from the previous dual-stage prompting into the target input sentence, a phonemic representation for the target input sentence[4] is first generated. This process involves embedding the list of phonemes using a lookup table, followed by further processing through the phonemic encoder $\mathrm{Enc_{phone}}$. This results in a 256-dimensional phonemic representation e_{phone} of the target utterance $\mathrm{Utt_T}$. This is to retain the basic TTS objective, which is to pronounce the target text in an intelligible manner.

Simultaneously, we integrate phoneme-level binary rationales into the initial list of phonemes. Specifically, from the list of selected words corresponding to the target emotion C that were previously identified through dual-stage prompting, all phonemes making up each word in the selected list of words is assigned a binary label: 1 if the word is in $\mathrm{S_{Words}}$, and 0 otherwise. The phoneme-level rationale sequence is then embedded using a lookup table f into a fixed size of 256 dimensions. The output is concatenated with the phonemic representation of the full utterance. Conceptually, this can be expressed as the following:

$$\mathrm{P_{rationale}}(w_{ni}) = \begin{cases} 1 & \text{if } W_n \in \{W_{i_1}, \ldots, W_{i_k}\} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$e_{\mathrm{rationale}} = f(\mathrm{P_{rationale}}) \quad (2)$$

$$e_{\mathrm{Utt}} = e_{\mathrm{rationale}} \oplus e_{\mathrm{phone}} \quad (3)$$

3.3 Prompt-Guided Alignment Loss

While local rationales have been specifically identified within the text modality, it is imperative to also integrate their corresponding acoustic features to enhance prosodic learning. This integration is essential because text-level information inherently lacks the acoustic cues vital for capturing prosodic nuances. To address this, an acoustic representation e_{acoustic} is initialized from the preceding phoneme-level rationale representation $e_{\mathrm{rationale}}$, and exclusively trained and updated using a novel selective masking loss. This loss is detailed as follows.

[3] gpt-3.5-turbo-0125

[4] https://pypi.org/project/g2p-en/

Algorithm 1: Selective Audio Masking

Input : Audio $\mathcal{A}$, Emotion tag $\mathcal{C}$, S_{Words}, Word-level timestamp $\mathcal{T}$

Output : Masked audio $\mathcal{A}'$

Step 1: Initialize $\mathcal{A}'$ as a zero array

$\mathcal{A}' \leftarrow [0] \times \text{len}(\mathcal{A})$;

Step 2: Selective Masking

if $\mathcal{C} = \textit{neutral}$ **then**
  return $\mathcal{A}'$;
else
  foreach w ***in*** S_{Words} **do**
    $(t_{start}, t_{end}) \leftarrow \mathcal{T}[w]$;
    $\mathcal{A}'[t_{start} : t_{end}] \leftarrow \mathcal{A}[t_{start} : t_{end}]$;
  end
  return $\mathcal{A}'$;
end

In order to effectively integrate the acoustic features of specific words identified by GPT 3.5, we apply masking to the time frames[5] that do not correspond to these selected words. Essentially, only the audio segments associated with the words selected by GPT 3.5 during the dual-stage prompting are retained. The masked audio sample is then passed through a pretrained Hubert (Hsu et al., 2021) base model, where the resulting embedding serves as the target for training the acoustic representation e_{acoustic}. To ensure the integrity of the text-based processing, gradient flow is blocked from propagating through the preceding layers. An overview of this selective masking process is provided in Algorithm 1. Moreover, e_{acoustic} is concatenated with the utterance-level embedding e_{Utt} described in subsection §3.2. The resulting multi-modal EDSS output is subsequently integrated with the global emotion tag C, which is embedded through a lookup table. The combined representation is then processed through the variance encoder followed by mel-spectrogram decoder $\text{Dec}_{\text{mel\_spec}}$, which ultimately generates a mel-spectrogram prediction.

The proposed selective audio masking loss is combined into a composite loss (Eq. 4) along with two other components from Ren et al. (2021): the reconstruction loss, which quantifies the difference between the mel-spectrogram of the ground truth audio and the predicted mel-spectrogram output by $\text{Dec}_{\text{mel\_spec}}$, and the losses used for training the modules of the variance adaptor.

$$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{Rec} + \mathcal{L}_{Var} + (1 - \lambda) \cdot \mathcal{L}_{SAM} \quad (4)$$

4 Experimental Settings

4.1 Dataset Preprocessing

We employ the CC-BY-SA 4.0 licensed DailyTalk dataset that has been open-sourced by Lee et al. (2023). The dataset includes seven emotion categories, but for the purposes of our study, we focus on five specific emotions: anger, happiness, neutral (none), sadness, and surprise. The remaining two emotions are excluded due to the limited number of samples available, with each having fewer than 100 examples[6]. The data is partitioned into training, validation, and test sets with a ratio of 8.6:0.7:0.7. Additionally, to enhance the learning of emotive characteristics, we further refine the training and validation subsets by including only those samples where GPT 3.5 accurately predicted the emotion tag during the initial prompting stage. This results in a total of 15,578 utterances.

In addition, following Ren et al. (2021), we utilize the Montreal Forced Aligner[7] (McAuliffe et al., 2017; MFA) to achieve precise alignment between spoken utterances and their corresponding phonemes. Specifically, MFA predicts the starting and ending timestamps for each articulated word within the audio samples. Audio samples are originally sampled at a rate of 22,050 Hz. However, when processing the audio samples through the pretrained Hubert model, they are downsampled to a rate of 16,000 Hz to comply with the model's requirements.

4.2 Baseline Models and Training Setup

As baseline systems, we first employ Lee et al. (2023), as they curate and provide the dataset that is used in this research. This baseline utilizes the FastSpeech2 framework (Ren et al., 2021) and incorporates a conversational context encoder (Guo et al., 2021) that embeds each utterance within the dialogue history using BERT. The output from the context encoder is then concatenated with the original phonemic outputs of the TTS encoder. Furthermore, given that the task of EDSS is both dialogue- and emotion-centric, it is also essential to compare

[5]The exact positions of specific words in the audio are determined during the preprocessing phase. More details are provided in §4.1.

[6]Additional dataset statistics can be found in Appendix C.

[7]https://montreal-forced-aligner.readthedocs.io/en/latest/

Models	MOS (↑)					ABX (↑)					SEA (↑)	WER (↓)
	Ang	Hap	Neu	Sad	Sur	Ang	Hap	Neu	Sad	Sur		
EXT-DT	3.08 (± 0.20)	3.38 (± 0.21)	3.56 (± 0.14)	3.78 (± 0.15)	3.17 (± 0.22)	34%	20%	16%	28%	18%	33.18%	0.104
DailyTalk	3.80 (± 0.22)	3.79 (± 0.19)	3.81 (± 0.19)	3.68 (± 0.23)	3.65 (± 0.20)	28%	26%	26%	32%	30%	29.69%	**0.090**
Proposed	**3.88 (± 0.19)**	**4.04 (± 0.19)**	**4.01 (± 0.16)**	**3.90 (± 0.20)**	**4.11 (± 0.19)**	**38%**	**54%**	**58%**	**40%**	**52%**	**46.16%**	0.109

Table 1: Performance comparison between baseline models and the proposed framework. Emotion expressivity is evaluated across anger, happiness, neutral, sadness, and surprise (left to right). MOS 95% confidence intervals are recorded in parentheses.

our proposed methodology with a purely emotion-based TTS system. For this purpose, we exclusively concatenate a discrete emotion ID without incorporating any dialogue history with the output of the Lee et al. (2023) phoneme encoder. This model is referred to as ExE-DT.

All models were trained using one A6000 GPU for 500,000 iterations with a batch size of 16. Mel-spectrograms were processed using a 1024 window size, 1024 filter length, and 256 hop size. The Adam optimizer (Kingma and Ba, 2015) was used with parameters $\beta_1 = 0.9$ and $\beta_1 = 0.98$, and predicted mel-spectrograms were converted into audio using the HiFi-GAN (Kong et al., 2020) vocoder.

4.3 Evaluation Protocol

To thoroughly evaluate the proposed EDSS model, we employ a combination of subjective and objective metrics. Subjective evaluations are conducted using two primary methods: Mean Opinion Scores (MOS) and ABX testing. For MOS, participants rate how accurately the synthesized audio conveys the target emotion using a 5-point Likert scale, where higher scores indicate a better alignment with the intended emotion. In ABX testing, participants are presented with three synthetic audio samples in parallel, each generated by one of the two baseline models and the proposed EDSS model. Participants are then asked to identify which one of the three samples most effectively conveys the specified emotion tag. These subjective evaluations are conducted via the Amazon Mechanical Turk platform[8], with 25 participants (Appendix D).

In addition to subjective assessments, objective metrics are employed to quantify the model's performance. The Speech Emotion Accuracy (SEA) metric utilizes a pretrained speech emotion recognition model (Ullah et al., 2023) to evaluate how accurately the synthesized audio aligns with the intended emotion tag. In addition, to assess the intelligibility of the synthesized speech, we use a pretrained Whisper large model (Radford et al., 2023) for automatic speech recognition (ASR). The transcriptions generated by the ASR system are then analyzed using the jiwer[9] library to calculate the Word Error Rate (WER).

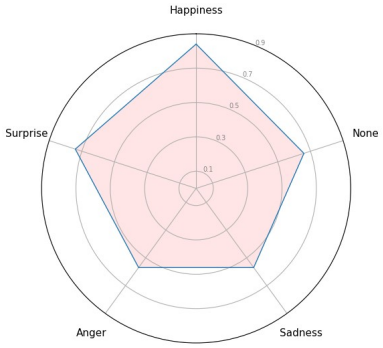

Figure 4: Radar graph demonstrating Stage 1 prompting accuracies across all emotion categories.

5 Results and Analyses

5.1 Comparisons with Baselines

In Table 1, we conduct subjective and objective evaluations to compare the performance of the proposed methodology against baseline models. Subjective results reveal that the DailyTalk model generally achieves higher MOS scores across all emotions compared to the emotion-label-only model (ExE-DT). A similar trend is observed in the ABX results, though the margin in emotive expressivity between the two baseline models is relatively modest. The SEA metric, however, presents a slight advantage for the ExE-DT model over DailyTalk. While there are variations in performance between the two baseline models, the differences are not pronounced, with each model occasionally outperforming the other even within the same metrics.

[8]https://requester.mturk.com/

[9]https://pypi.org/project/jiwer/

Emotion	Avg. # of Words		Overlap
	LLM	Human	
Happiness	2.531	3.191	0.740
Sad	2.250	3.917	0.673
Surprise	1.896	2.708	0.726
Anger	2.497	3.644	0.649
Total	2.497	3.197	0.735

Table 2: Congruence between LLM and human annotations. The average number of words selected by GPT 3.5 across all five emotion categories, and the extent of overlap (i.e., number of words jointly identified by both ChatPGT and human annotation) illustrates the model's alignment with human judgements. Degree of overlap is quantified on a scale from 0 to 1, with higher values indicating greater agreement in word selection.

This similarity in performance can be attributed to the fact that both models rely solely on global information, whether in the form of emotion labels or contextual information.

This notion is further substantiated by the results of the proposed model, which demonstrates significant improvements in both MOS and ABX metrics, alongside a substantial increase in the objective SEA metric (with a Δ12.98 and Δ16.47 over the ExE-DT and DailyTalk baselines, respectively). These results suggest that integrating local information in additional to holistic features plays a crucial role in enhancing emotive expressivity. In addition, given that the primary goal of speech synthesis is to produce intelligible speech, we evaluate all models based on WER. We find that the proposed model demonstrates comparable performance to the baseline models while achieving superior EDSS results.

5.2 Prompting Assessment

Holistic Comprehension Before assessing the influence of prompt-guided labelling of localized features on emotive synthesis, we first evaluate the accuracy with which GPT 3.5 assigns the correct emotion tag to target text utterances, given the prior dialogue context (i.e., Stage 1 Prompting). This involves comparing the predicted emotion tags with the ground truth tags. Our analysis reveals that the average accuracy for correctly identifying the holistic emotion tag across all five categories is 69.3%. In particular, as illustrated in Figure 4, the categories of happiness (84%), surprise (74%), and neutral (66%) exhibit the highest accuracies, while anger and sadness have lower accuracies, at 57%.

	Settings	Singular	Boolean	SEA
Labelling	TRN	✓		37.99%
	INF	✓		
	TRN		✓	44.77%
	INF		✓	

Table 3: Comparison of various labelling settings pertaining to the employment of the proposed selective masking loss and AE. TRN and INF refer to training and inference, respectively.

Interestingly, the emotion categories with the highest annotation accuracy correspond with the highest MOS and ABX for our proposed model.

Affective Rationales In order to evaluate GPT 3.5's ability to identify emotionally salient words within target sentences, we compared the model's selections with those made by a native English speaker. As can be seen in Table 2, GPT 3.5 selected an average of approximately 2.5 words across all emotion categories, which is closely aligned with the average number chosen by the human annotator. Additionally, the overlap between the selections of GPT 3.5 and the human annotator yielded a congruence score of around 0.74. This indicates that GPT 3.5 is able to select words that are particularly relevant to the holistic emotion label assigned to the target sentence[10].

We further investigate phonemic labelling in various configurations, including singular and boolean labelling during training and inference. In the singular setting, all phonemic states are labelled as 0, regardless of whether a word in the target utterance is selected by GPT 3.5 during Stage 2 Prompting. In contrast, the boolean setting uses 0 and 1, where the phonemic sequence of a selected word is marked with 1 and otherwise with 0, as per the proposed methodology. Table 3 shows that using the singular setting for both training and inference results in the lowest performance, as it provides no information about which words align with the target emotion. Conversely, employing boolean labelling during both training and inference yields a performance improvement of 6.78%. This indicates that the model effectively learns to distinguish between neutral and non-neutral words.

5.3 Prompt-Guided Alignment Loss

To incorporate the acoustic features associated with the emotion salient words selected during Stage 2

[10]Case studies can be found in Appendix A.

	Settings	SEA	Labelling Relative
+ Loss	Without AE	39.22%	∇5.55
	With AE*	36.57%	∇8.20
	With AE	46.16%	∆1.39

Table 4: Comparison of different AE settings, with localized labelling and loss as the default configurations. Each of the three settings is relatively compared against the boolean training and inference labelling-only settings presented in Table 3.

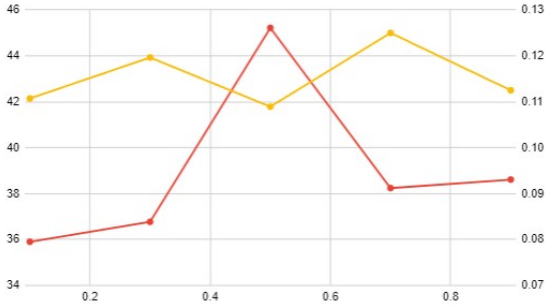

Figure 5: The x-axis denotes the value of λ, and the left and right y-axes denotes SEA (red) and WER (yellow) metric scales, respectively.

Prompting, we have introduced a separate acoustic embedding (AE) that is trained with selective masking loss. In order to analyze the effectiveness of this approach, we compared various configurations with and without AE. The results in Table 4 show that omitting AE led to a substantial 5.55-point drop compared to the highest performance observed in the labelling ablation studies. Furthermore, when compared to the proposed model, which leverages AE to capture acoustic information independently from phoneme-level binary labelling, the absence of AE resulted in poorer performance (6.94% SEA decrease). This suggests that forcing the phoneme-level binary labelling representation to simultaneously learn textual and acoustic features impedes effective learning. Therefore, it is crucial to employ a separate AE to learn acoustic features in a modality-independent manner.

We also experimented with utilizing full audio representations as the target for learning during masking loss. In the proposed methodology, segments corresponding to words not selected by GPT 3.5 were masked out. To evaluate whether this approach enhances acoustic learning regarding emotive characteristics, we conducted an experiment where neutral words were not masked (With AE* in Table 4). This adjustment resulted in a notable performance decline, with an 8.20-point decrease compared to the labelling-only settings and a 9.59-point drop relative to the proposed model. These results indicate that GPT 3.5 is effective in identifying words aligned with the target emotion. Moreover, it can be inferred that the inclusion of neutral information, which also encompasses intonational and other paralinguistic features, can disrupt the learning of emotion-specific features.

Finally, we conducted empirical tests to determine the optimal value of λ, which balances the mel-reconstruction loss and the proposed selective audio masking loss. As illustrated in Figure 5, setting both the mel-reconstruction and selective audio masking loss parameters to 0.5 resulted in the highest SEA accuracies and the lowest WER. This balance likely arises because assigning more weight to the mel-reconstruction loss could cause the neutral information from the ground truth mel-spectrogram to interfere with emotion expressivity. Conversely, placing greater emphasis on the selective masking loss, which relies on a limited amount of audio information, may result in insufficient data to effectively support audio pronunciation learning.

6 Conclusion

This paper investigates the application of LLM prompting to autonomously generate emotion tags and identify specific lexical cues within target utterances that contribute to more expressive speech. This approach obviates the need for additional dataset curation. We further enhance expressivity through a novel selective masking loss function. Our findings underscore the necessity of integrating both global and local information, and highlight the substantial advantages of separately learning text and acoustic features for emotion synthesis. Notably, despite GPT 3.5 being a text-based LLM, it is able to identify words that correspond with assigned emotion tags to some extent. The efficacy of the proposed model is rigorously assessed using a combination of subjective and objective metrics.

7 Limitations

In our pursuit towards automatic and effective emotion generation in dialogue settings, we have leveraged GPT 3.5 for its strong reasoning abilities and contextual understanding. However, our approach assumes that GPT 3.5 can inherently determine the correct holistic emotion tag, as well as which words should serve as cues that align with the pre-

dicted emotion. Specifically, we have assumed that GPT 3.5 can effectively correlate text-based words with acoustic features, a task that humans naturally perform when reading text aloud.

While our ablation studies and both subjective and objective metrics indicate that GPT 3.5 can align text and acoustic information to some extent, a more comprehensive analysis is warranted to better understand the extent of GPT 3.5's speech-related knowledge. Future work will focus on conducting a deeper investigation into this capability, not only within GPT 3.5 but also across other large language models.

Moreover, the considerable imbalance within the dataset, both between the neutral and emotional categories and among the different emotional categories themselves, presents significant challenges in training models to generate speech with pronounced emotional salience and to achieve high emotion classification accuracy for synthesized audio. Despite this, we have demonstrated notable improvements in the emotional expressiveness of the synthesized speech. Nevertheless, addressing the dataset imbalance problem is a notable point of exploration as future work.

8 Ethical Considerations

The field of EDSS, like TTS in general, essentially seeks to closely replicate human speech. While these advancements offer numerous benefits, such as improving accessibility in healthcare, enabling personal AI assistants, and enhancing media experiences, they also pose ethical risks. One significant concern is the potential for misuse, where synthesized speech could deceive listeners into believing it is produced by a human. To address this, it is essential to ensure transparency by clearly disclosing when speech is artificially generated. Another possible solution is to develop techniques akin to watermarking for speech, which allows for the identification of synthesized content and preventing deception in an automatic manner.

9 Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2025-RS-2024-00437866, 47.5%), by Smart HealthCare Program (www.kipot.or.kr) funded by the Korean National Police Agency (KNPA, Korea) [Project Name: Development of an Intelligent Big Data Integrated Platform for Police Officers' Personalized Healthcare / Project Number: 220222M01, 47.5%)], and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH), 5%).

References

Murtaza Bulut, Sungbok Lee, and Shrikanth Narayanan. 2005. Analysis of emotional speech prosody in terms of part of speech tags. In *Interspeech*.

Mingjian Chen, Xu Tan, Bohan Li, Yangqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2021. AdaSpeech: Adaptive Text to Speech For Custom Voice. In *ICLR*.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. 2020. MultiSpeech: Multi-Speaker Text to Speech with Transformer. In *Interspeech*.

Byoung Jin Choi, Myeonghun Jeong, Joun Yeop Lee, and Nam Soo Kim. 2022. Snac: Speaker-normalized affine coupling layer in flow-based architecture for zero-shot multi-speaker text-to-speech. *IEEE Signal Processing Letters*, 29:2502–2506.

Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. 2021. Controllable Context-aware Conversational Speech Synthesis. In *Interspeech*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federica Laricchia. 2024. Number of digital voice assistants in use worldwide 2019-2024. https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie. 2021. Conversational End-to-End TTS for Voice Agents. In *IEEE Spoken Language Technology Workshop (SLT)*.

Zhifang Guo, Yuchong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. PromptTTS: Controllable Text-to-Speech with Text Descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jiaxu He, Cheng Gong, Longbiao Wang, Di Jin, Xiaobao Wang, Junhai Xu, and Jianwu Dang. 2022. Improve emotional speech synthesis quality by learning explicit and implicit representations with semi-supervised training. *Interspeech*.

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 Trans. Audio, Speech and Lang. Proc.*, 29.

Yejin Jeon, Yunsu Kim, and Gary Geunbae Lee. 2024. Enhancing Zero-Shot Multi-Speaker TTS with Negated Speaker Representations. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*.

Bonian Jia1, Huiyao Chen, Yueheng Sun, Meishan Zhang, and Min Zhang. 2024. LLM-Driven Multimodal Opinion Expression Identification. In *Interspeech*.

Minki Kang, Wooseok Han, Sung Ju Hwang, and Eunho Yang. 2023. ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models. *Interspeech*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021a. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *ICLR*.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021b. Expressive Text-to-Speech using Style Tag. In *Interspeech*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High FIdelity Speech Synthesis. In *NeurIPS*.

Chul Min Lee and Shrikanth S. Narayanan. 2005. Toward Detecting Emotions in Spoken Dialogs. In *IEEE Transactions on Speech and Audio Processing*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. DailyTalk: Controllable Context-Aware Conversational Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. 2017. Emotional End-to-End Neural Speech Synthesizer. *Neural Information Processing Systems (NIPS)*.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*

Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. 2022. Enhancing Speaking Styles in Conversational Text-to-Speech Synthesis with Graph-Based Multi-Modal Context Modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Xiaojiang Peng, and Alexander G. Hauptmann. 2024. MM-TTS: A Unified Framework for Multimodal, Prompt-Induced Emotional Text-to-Speech Synthesis. In *arXiv*.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Want, Zhifei Li, and Lei Xie. 2023a. PromptStyle: Controllable Style Transfer for Text-to-Speech with Natural Language Descriptions. In *Interspeech*.

Yuchen Liu, Haoyu Zhang, Shichao Liu, Xiang Yin, Zejun Ma, and Qin Jin. 2023b. Emotionally Situated Text-to-Speech Synthesis in User-Agent Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai. 2018. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. *Interspeech*.

Masato Murata, Koichi Muyazaki, and Tomoki Koriyama. 2024. An Attribute Interpolation Method in Speech Synthesis by Model Merging. *Interspeech*.

Paul Pangaro and Hugh Dubberly. 2014. What is Conversation? How Can We Design for Effective Conversation? *Driving Desired Futures: Turning Design Thinking into Real Innovation*, pages 144–159.

Juslin Patrik N. and Petri Laukka. 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. In *Emotion*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of Machine Learning Research*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Tan Liu. 2021. Fastspeech 2: Fast and High-Quality End-to-End Text to Speech. In *ICLR*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *NeurIPS*.

Yuki Saito, Shinnosuke Takamichi, Eiji Limori, Kentaro Tachibana, and Hiroshi Saruwataru. 2023. ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-derived Context Word Embeddings. In *Interspeech*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Atli Thor Sigurgeirsson and Simon King. 2024. Controllable Speaking Styles Using a Large Language Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Haobin Tang, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. 2024. ED-TTS: Multi-Scale Emotion Modeling Using Cross-Domain Emotion Diarization for Emotional Speech Synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2022. EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, page 853–864.

Rizwan Ullah, Muhammad Asif, Wahab Ali Shah, Fakhar Anjam, Ibrar Ullah, Tahir Khurshaid, Lunchakorn Wuttisittikulkij, Shashi Shah, Syed Mansoor Ali, and Mohammad Alibakhshikenari. 2023. Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. *Sensors*, 23(13).

Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. 2020. Emotional Speech Synthesis with Rich and Granularized Control. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *ICML*.

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, and Lirong Dai. 2019. End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*.

Jinlong Xue, Yayue Deng, Fengping Wang, Ya Li, Yingming Gao, Jianhua Tao, Jianquing Sun, and Jiaen Liang. 2023. M^2-CTTS: End-to-End Multi-Scale Multi-Modal Conversational Text-to-Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yuzi Yan, Xu Tan, Bohan Li, Tao Qin, Sheng Zhao, Yuan Shen, and Tie-Yan Liu. 2021. Adaspeech2: Adaptive Text to Speech With Untranscribed Data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. InstructTTS: Modelling Expressive TTS in Discrete Latent Space With Natural Language Style Prompt. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hyun-Wook Yoon, Ohsung Kwon, Hoyeon Lee, Ryuichi Yamamoto, Eunwoo Song, Jae-Min Kim, and Min-Jae Hwang. 2022. Language Model-Based Emotion Prediction Methods for Emotional Speech Synthesis Systems. In *Interspeech*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and Unseen Emotional Style Transfer for voice Conversion with A New Emotional Speech Dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, and Lei Xie. 2023. Multi-Speaker Expressive Speech Synthesis via Multiple Factors Decoupling. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

A Holistic Tag and Affective Rationale Selection

We evaluated various Stage 1 prompting methods, including chain-of-thought (CoT) and reasoning (see Fig. 6) to identify the most accurate approach for generating emotion tags. As shown in Table 5, the proposed prompting method demonstrated the highest accuracy when compared to the ground truth labels. Notably, the largest discrepancy was observed when predicting neutral tags, suggesting that the reasoning prompts may lead the GPT 3.5 model to misclassify neutral utterances by over-interpreting them as different emotion classes. The entire cost of utilizing the GPT 3.5 API was approximately 10 dollars.

Regarding Stage 2 rationale selection, we compared the words and phrases selected by GPT 3.5

COT

```
Given the previous dialogue history between two people, after
providing your explanation, output one emotion from {anger,
happiness, sadness, surprise, none} in which the target
sentence should be read. Your output should be just one word.

Previous Dialogue History is:
{Spk A Utterance 1 # Spk B Utterance 1 # … # Spk A Utterance N
#}

Target Sentence is:
{I really enjoy the beach in the summer.}
```

Reasoning

```
Given the previous dialogue history between two people, in
what emotion should the target sentence be spoken? Output one
emotion from {anger, happiness, sadness, surprise, none}, and
then provide the reason for that emotion.

Previous Dialogue History is:
{Spk A Utterance 1 # Spk B Utterance 1 # … # Spk A Utterance N
#}

Target Sentence is:
{I really enjoy the beach in the summer.}
```

Figure 6: Prompt variations used for Stage 1 prompting.

	Prompt Accuracy
CoT	30.77%
Reasoning	44.30%
Proposed	69.29%

Table 5: Average accuracies of emotion tags generated via different prompting methods.

with human annotations. Parallel GPT 3.5 and human annotations are demonstrated in Figure 7.

B Generalizability

To assess the broader applicability of the proposed method beyond GPT 3.5, we conducted additional experiments using the Llama 8b model[11] with a temperature setting of 0.5. Our initial analysis focused on the number of selected words and the degree of overlap with human annotations. As shown in Table 6, the model selected an average of 2.9 words, which closely matches the 3.2 words selected by human annotators. The SEA score, at 41.11%, was slightly lower than the results achieved by the proposed model using GPT 3.5. This may be due to the differences in word selection patterns between GPT 3.5 and Llama.

To examine the reason for the performance difference between LLMs, we performed a part-of-speech (POS) analysis using the Stanford POS tagger tool with the `english-bidirectional-distsim model`[12].

[11] casperhansen/llama-3-8b-instruct-awq

[12] Explanations and examples pertaining to each POS category can be found in https://web.stanford.edu/~jurafsky/slp3/old_oct19/8.pdf.

Happiness
LLM: This balcony is perfect for barbecuing.
Human: This balcony is perfect for barbecuing.

Anger
LLM: They sound like the neighbors from hell!
Human: They sound like the neighbors from hell!

Surprise
LLM: Oh, really? what do you do there exactly?
Human: Oh, really? what do you do there exactly?

Sadness
LLM: Oh, that's too bad. should I call a doctor ?
Human: Oh, that's too bad. should I call a doctor ?

Figure 7: Case studies examining the selected rationales generated by GPT 3.5 and those annotated by humans across various emotive categories.

As shown in Figure 8, GPT 3.5 selected a greater proportion of adjectives (`JJ`) and nouns (`NN`) compared to Llama. Given that adjectives and nouns convey more information that are strongly linked to emotive expressivity in speech prosody (Bulut et al., 2005), it can be inferred that GPT 3.5 is more adept at identifying emotionally salient words than Llama. This, in turn, contributes to its superior performance in both emotion salience in the generated speech, and the resulting classification accuracy.

Emotion	Avg. # of Words		Overlap
	LLM	Human	
Happiness	2.849	3.191	0.634
Sad	3.500	3.917	0.671
Surprise	2.688	2.708	0.697
Anger	3.222	3.644	0.666
Total	2.866	3.197	0.639

Table 6: Congruence between Llama and human annotations.

C Dataset Statistics

Out of a total of 15,578 utterances, approximately 20.4% of the utterances belong to an emotion category (i.e., happiness, sadness, surprise, or anger). Among the utterances with an emotion label, happiness accounts for 85.8%, followed by surprise (7.13%), sadness (4.48%), and anger (2.55%).

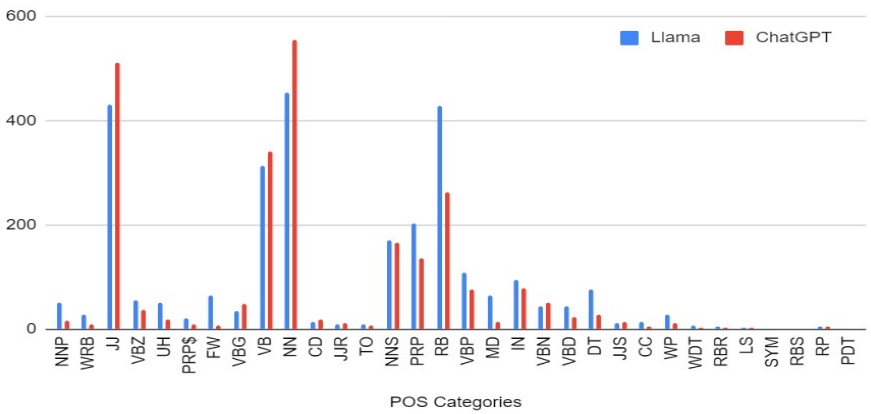

Figure 8: A comparative analysis of Llama and GPT 3.5 word frequencies is presented side-by-side for each part-of-speech (POS) category. The most commonly identified POS categories are adjectives (`JJ`), nouns (`NN`), and adverbs (`RB`), in order.

D Evaluation Metrics

Participants for the MOS and ABX evaluations were recruited through the Amazon Mechanical Turk platform. They were informed prior to the assessment that their responses would be used exclusively for research purposes, with no personal information collected or utilized. Participants were compensated based on the standard hourly rates of the authors' nationality, and the average completion time for the survey was 1 hour and 27 minutes per participant.