Expressive Prompting: Improving Emotion Intensity and Speaker Consistency in Zero-Shot TTS

Haoyu Wang*, Chunyu Qiang*†, Tianrui Wang*, Cheng Gong‡, Yu Jiang*,
Yuheng Lu*, Chen Zhang†, Longbiao Wang*¶, and Jianwu Dang§
* Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
† Kuaishou Technology Co., Ltd, Beijing, China
‡ Institute of Artificial Intelligence, China Telecom, China
§ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Guangdong, China
¶ Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
Corresponding author: longbiao_wang@tju.edu.cn

***Abstract*—Recent advancements in speech synthesis have enabled large language model (LLM)-based systems to perform zero-shot generation with controllable content, timbre, speaker identity, and emotion through input prompts. As a result, these models heavily rely on prompt design to guide the generation process. However, existing prompt selection methods often fail to ensure that prompts contain sufficiently stable speaker identity cues and appropriate emotional intensity indicators, which are crucial for expressive speech synthesis. To address this challenge, we propose a two-stage prompt selection strategy specifically designed for expressive speech synthesis. In the static stage (before synthesis), we first evaluate prompt candidates using pitch-based prosodic features, perceptual audio quality, and text-emotion coherence scores evaluated by an LLM. We further assess the candidates under a specific TTS model by measuring character error rate, speaker similarity, and emotional similarity between the synthesized and prompt speech. In the dynamic stage (during synthesis), we use a textual similarity model to select the prompt that is most aligned with the current input text. Experimental results demonstrate that our strategy effectively selects prompt to synthesize speech with both high-intensity emotional expression and robust speaker identity, leading to more expressive and stable zero-shot TTS performance. Audio samples and codes will be available at https://whyrrrrun.github.io/ExpPro.github.io/.**

I. Introduction

In recent years, speech synthesis technology has made remarkable progress, with the quality of synthesized speech continuously improving. The GPT model [1] has achieved great success in the field of natural language processing. Inspired by this, language models have gradually been introduced into the field of speech synthesis and have become the mainstream framework paradigm [2]–[5]. The quality of synthesized speech has progressively reached a level comparable to that of human speech. LM-based TTS systems utilize neural audio codecs [5]–[8] to convert speech into discrete tokens, which encapsulate extensive information about the speech. These systems then employ a language model architecture to generate subsequent speech tokens autoregressively. Existing LM-based TTS models implement in-context learning capabilities.

Current LM-based TTS methods [4], [5] employ autoregressive generation to produce subsequent tokens from input prompts and text. These advanced TTS methods achieve zero-shot voice cloning with just a few seconds of prompt speech. However, the quality of the prompts significantly influences the generated speech output, impacting aspects such as timbre, perceptual quality, and emotional expression [9], [10].

Consequently, selecting an appropriate prompt is crucial [11]–[13]. There are two mainstream methods for prompt selection: 1) Random: randomly choosing speech from a specific speaker with a certain emotional speech, or 2) Text-based Methods[14], [15]: selecting prompts based on the similarity between the synthesized text and prompt text. However, these methods are primarily designed for general scenarios and face limitations in emotional speech synthesis. Random selection often fails to provide rich emotional information and expressive capabilities, and focusing solely on the text can yield subpar emotional performances, as there's frequently a weak connection between the text and the desired emotion [2], [16]. Therefore, additional research is required to identify prompts that can enhance emotional expressiveness, speaker similarity, and stability across various LM-based methods in emotional speech synthesis scenarios [17], [18]. This observation motivates our central research question: *Can we improve emotion intensity and speaker consistency in zero-shot TTS through better prompt selection, without any additional training?*

To tackle these challenges, we propose an innovative two-stage prompt selection strategy — ExpPro. In the static selection stage, we evaluate both the inherent emotional quality of the prompt candidates and their specific expressive power within the model. In the dynamic selection stage, we choose the most semantically relevant and contextually appropriate prompts from the candidates after the static selection stage, based on the synthesized text. This strategy aims to systematically screen and rank prompts based on various metrics, ultimately selecting prompts with strong emotional expressiveness, high speaker similarity, and high stability. The specific contributions of this paper are as follows:

979-8-3315-7206-8/25/$31.00

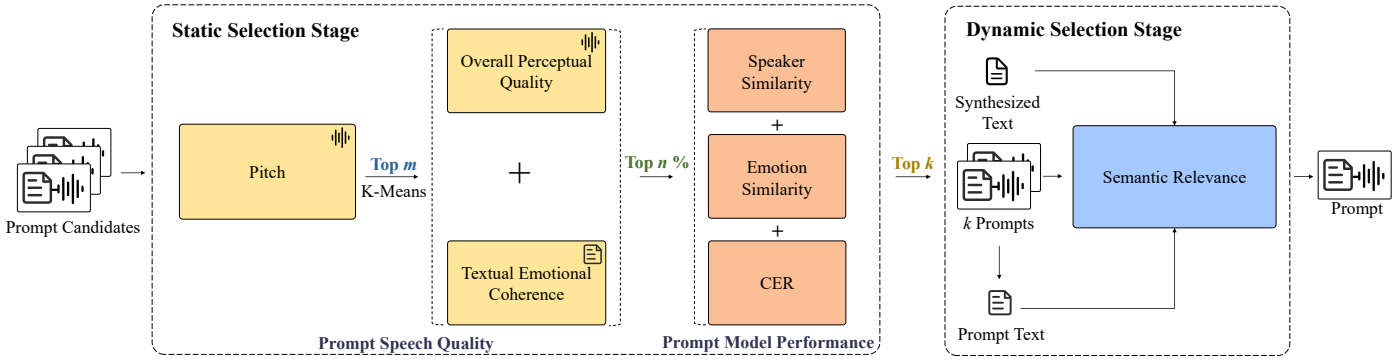

Fig. 1. The overview of ExpPro. It consists of two stages: a static selection stage and a dynamic selection stage. The static selection stage evaluates the intrinsic quality of the prompt and its performance in the specific LM-based model, while the dynamic selection stage chooses the most relevant prompt from k prompts based on the synthesized text.

1) We propose a two-stage emotion prompt selection strategy — ExpPro, which combines static-dynamic selection for LM-based TTS without any additional training.
2) We conduct a multi-perspective analysis about the text and speech of the prompt, taking into account the ability of prompt in specific models as well as the emotional quality of the prompt itself.
3) This is a flexible prompt selection strategy suitable for improving emotional expressiveness in any TTS models that involve the concept of prompt speech.

II. Method

A. Overview

The proposed ExpPro is illustrated in Fig. 1, and it consists of two stages: static selection and dynamic selection. In the static selection stage, we select prompt candidates based on emotional expressiveness, perceptual quality, and textual emotional coherence. The selected candidates are then used for inference with the LM-based TTS methods. The objective metrics are used to evaluate candidates and retain those with high quality, expression, and stability. In the dynamic selection stage, we identify the prompt with the highest semantic relevance to the synthesized text input, choosing from the previously filtered candidates. Finally, this prompt is the one that best reflects the required emotional effect of the synthesized text under the current model.

B. Static Selection

For prompt static selection, we evaluate the quality of the prompt speech across three key dimensions: pitch, overall perceptual quality, and textual emotional coherence derived from a large language model [19]. Additionally, we assess the inference results of the prompt candidates, considering metrics such as character error rate (CER), emotion similarity (ES), and speaker similarity. By integrating these factors, we identify the prompt candidates deemed most suitable for the emotion.

1) Pitch: The pitch, or fundamental frequency, is a preverbal feature that imparts tonal and rhythmic qualities to speech [20]. As a suprasegmental speech feature, pitch conveys information over a longer time scale than segmental features such as spectral envelopes. Features describing overall

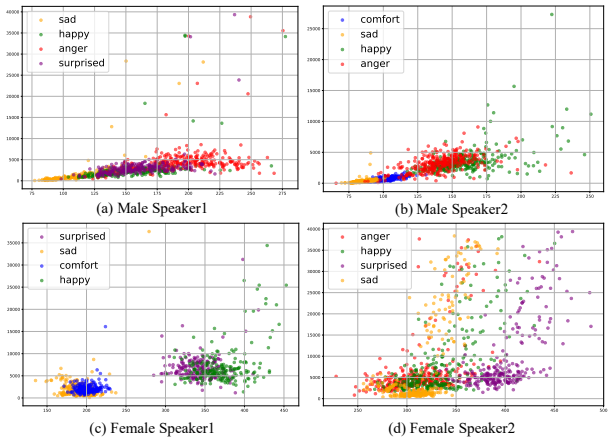

Fig. 2. Mean and variance of emotional speech pitch: red indicates angry, blue indicates comfort, orange indicates sad, green indicates happy, and purple indicates surprised. The x-axis represents the pitch mean, and the y-axis represents the pitch variance.

attributes of the pitch contour, such as mean and variance, are more emotionally resonant than those describing the pitch shape itself, such as slope, curvature, and inflection [21]. a) **Mean**: This refers to the average pitch level over a period of speech. It can indicate the general tone or mood of the speaker. b) **Variance**: This measures the variability in pitch over time. Greater variance might suggest more animated or emotional speech, while less variance could indicate a monotone delivery.

Different emotional states are associated with distinct pitch patterns [22]. Both sadness and comfort exhibit relatively low mean and variance in pitch, indicating calmer and lower pitch characteristics, with sadness being slightly more subdued. On the other hand, emotions like happiness and surprise demonstrate higher mean and variance, reflecting more pronounced emotional intensity [23]. Fig. 2 illustrates the mean and variance of pitch across various emotional audio samples.

We select the prompt speech based on the distinct tonal features associated with each emotion category. Initially, we calculated the mean and variance for each emotion type. Subsequently, we apply the K-Means algorithm to cluster 10 groups based on the mean and variance of the prompt candidates for different speakers and emotions. We select m clusters with stronger or weaker means and variances based

TABLE I
THE PROMPT SETTINGS FOR THE CHATGPT IN THE TEXTUAL EMOTIONAL COHERENCE MODULE.

You are tasked with evaluating the degree of match between the input text and the emotional label. I will provide the input text and the corresponding emotional label, and you need to assess the degree of match between them. Please note that your response should be a specific score with four decimal places, without any explanation.

For example:

Text: *The kitten has been gone for many days, and thinking about it still makes me very sad.*
Emotion: *Sad.*
Your output: *0.9357.*

Now, here's my formal question:

Text: [Text], **Emotion**: [Emotion]. **Your output**:

on the various states of different emotional classes.

2) Perceptual and Textual Selecting: We comprehensively consider both perceptual quality and text consistency.

Overall Perceptual Quality: We regard the quality of the prompt speech as a critical factor and utilize DNSMOS [24] for this purpose. DNSMOS is a deep learning-based audio quality assessment tool designed to evaluate the quality of audio signals. It can assess the clarity, naturalness, and overall quality of audio. Leveraging neural network models to simulate human auditory perception, it provides objective scores that are highly correlated with subjective ratings. We measure DNSMOS on all results following pitch selection.

Textual Emotional Coherence: When the text of speech aligns more closely with a particular emotional expression, the sentence can more effectively convey the desired emotion. To assess the relevance of the text to the corresponding emotion in the prompt speech, we use the ChatGPT[1] API. While ChatGPT operates as a black-box system, we use it not as a definitive classifier, but as a semantic comparator to assess the relative coherence between input text and target emotion in a controlled and consistent setting.

Specifically, we fix a benchmarking prompt (shown in Table I) and use ChatGPT under identical conditions for all comparisons. This ensures that all samples are evaluated under the same latent criteria, enabling fair and reproducible ranking across prompt candidates.

Moreover, we verified the stability and reliability of the ChatGPT-based assessment via manual inspection on a sample of cases, which showed strong alignment with human intuition. This setup allows us to benefit from the model's strong contextual understanding while avoiding over-dependence on its opaque internal mechanisms.

Considering the relatively stable distribution of DNSMOS scores within the same dataset, to highlight differences, we directly combine the textual emotional coherence scores with the DNSMOS scores and select the top $n\%$ as the most emotionally expressive data.

3) Selecting with Performance under LM-based TTS Method: The method above focuses on the selection method for evaluating the quality of the prompt speech itself. Additionally, we recognize that even when identical prompt speech is input into different methods, the resulting outputs can vary significantly. This variability primarily depends on factors such as the selection of speech tokens. To address this question, we propose a strategy that considers the specific performance of different models when processing the same prompt speech.

Specifically, we select 20 neutral descriptive sentences for inference based on the prompt candidates from our prompt speech quality selection process. We then evaluate the inference results for all prompt speeches by calculating the CER of the synthesized speech. Furthermore, we use Resemblyzer [25] and WavLM [26] to evaluate speaker similarity and assess the model's capability to generate the same speaker's voice from the given prompt. Finally, we employ the emotion2vec [27] model to assess the cosine similarity between the emotion of synthesized speech and prompt speech, which serves as an indicator of the model's effect in capturing the emotional information of the prompt speech.

The three metrics of CER, speaker similarity, and ES form the framework for assessing our model's effect in capturing various speech information. These metrics are further integrated through a weighted sum to provide a comprehensive assessment. In our selection strategy, we prioritize the quality of the prompt candidates themselves. We start with an initial selection of their intrinsic emotional quality before applying the model-specific selection method.

C. Dynamic Selection

The consistency between the prompt text and the synthesized text also affects the results, so we employ a dynamic selection strategy based on the text. The stsb-distilroberta-base[2] analyzes the currently synthesized text alongside the statically selected speeches from the prompt candidates. This allows us to identify the most relevant prompt for the current text, which is then chosen as the final prompt.

III. EXPERIMENTS

A. Data

We utilize a highly emotionally expressive dataset (HE2D), as described in [28], which includes recordings from two male and two female speakers to validate our prompt selection strategy. The dataset encompasses five distinct emotions: comfort, happy, sad, angry, and surprise. Each speaker exhibits four of these emotions, with 200 samples per emotion, amounting to a total of 800 samples per speaker. Additionally, to demonstrate the generalizability of our approach, we also validate our method using the open-source Emotional Speech Database (ESD) [29]. From this dataset, we randomly select recordings from four Chinese speakers (two male and two female), each corresponding to four emotional categories.

[1]https://chatgpt.com/

[2]https://huggingface.co/cross-encoder/stsb-distilroberta-base

TABLE II
COMPARISON OF ZERO-SHOT LM-BASED TTS PERFORMANCE ACROSS VARIOUS PROMPT SELECTION METHODS. SP AND MOS ARE PRESENTED WITH A 95% CONFIDENCE INTERVAL. ESD DOES NOT INCLUDE THE COMFORT EMOTION, THE CORRESPONDING POSITION IS EMPTY.

Model	Dataset	Method	Happy		Sad		Angry		Surprised		Comfort	
			MOS ↑	SP ↑	MOS ↑	SP ↑	MOS ↑	SP ↑	MOS ↑	SP ↑	MOS ↑	SP ↑
CosyVoice	ESD	Random	$3.83_{\pm 0.12}$	$0.683_{\pm 0.023}$	$3.71_{\pm 0.09}$	$0.635_{\pm 0.019}$	$3.84_{\pm 0.13}$	$0.621_{\pm 0.011}$	$3.83_{\pm 0.19}$	$0.667_{\pm 0.015}$	-	-
		MiniLM	$3.79_{\pm 0.11}$	$0.681_{\pm 0.018}$	$3.75_{\pm 0.10}$	$0.647_{\pm 0.015}$	$3.81_{\pm 0.17}$	$0.628_{\pm 0.021}$	$3.88_{\pm 0.18}$	$0.675_{\pm 0.013}$	-	-
		ExpPro	$\mathbf{3.87_{\pm 0.14}}$	$\mathbf{0.689_{\pm 0.012}}$	$\mathbf{3.80_{\pm 0.13}}$	$\mathbf{0.651_{\pm 0.014}}$	$\mathbf{3.86_{\pm 0.12}}$	$\mathbf{0.633_{\pm 0.015}}$	$\mathbf{3.93_{\pm 0.16}}$	$\mathbf{0.681_{\pm 0.014}}$	-	-
	HE2D	Random	$4.13_{\pm 0.11}$	$0.767_{\pm 0.018}$	$4.23_{\pm 0.16}$	$0.789_{\pm 0.023}$	$4.43_{\pm 0.16}$	$0.733_{\pm 0.018}$	$4.11_{\pm 0.11}$	$0.733_{\pm 0.013}$	$4.21_{\pm 0.11}$	$0.833_{\pm 0.013}$
		MiniLM	$4.33_{\pm 0.14}$	$0.767_{\pm 0.013}$	$4.35_{\pm 0.15}$	$0.818_{\pm 0.017}$	$4.46_{\pm 0.12}$	$0.767_{\pm 0.018}$	$4.30_{\pm 0.09}$	$0.744_{\pm 0.015}$	$4.34_{\pm 0.12}$	$0.879_{\pm 0.011}$
		ExpPro	$\mathbf{4.45_{\pm 0.13}}$	$\mathbf{0.811_{\pm 0.017}}$	$\mathbf{4.42_{\pm 0.11}}$	$\mathbf{0.832_{\pm 0.018}}$	$\mathbf{4.50_{\pm 0.12}}$	$\mathbf{0.867_{\pm 0.015}}$	$\mathbf{4.33_{\pm 0.08}}$	$\mathbf{0.767_{\pm 0.017}}$	$\mathbf{4.36_{\pm 0.11}}$	$\mathbf{0.889_{\pm 0.011}}$
GPT-SoVITS	ESD	Random	$3.69_{\pm 0.18}$	$0.641_{\pm 0.021}$	$3.57_{\pm 0.13}$	$0.597_{\pm 0.015}$	$3.59_{\pm 0.12}$	$0.583_{\pm 0.017}$	$3.72_{\pm 0.11}$	$0.641_{\pm 0.013}$	-	-
		MiniLM	$3.71_{\pm 0.15}$	$0.637_{\pm 0.022}$	$3.62_{\pm 0.14}$	$0.628_{\pm 0.014}$	$3.51_{\pm 0.12}$	$0.591_{\pm 0.019}$	$3.77_{\pm 0.13}$	$0.671_{\pm 0.019}$	-	-
		ExpPro	$\mathbf{3.78_{\pm 0.17}}$	$\mathbf{0.659_{\pm 0.017}}$	$\mathbf{3.75_{\pm 0.11}}$	$\mathbf{0.641_{\pm 0.016}}$	$\mathbf{3.65_{\pm 0.11}}$	$\mathbf{0.611_{\pm 0.013}}$	$\mathbf{3.88_{\pm 0.14}}$	$\mathbf{0.675_{\pm 0.016}}$	-	-
	HE2D	Random	$4.02_{\pm 0.15}$	$0.668_{\pm 0.023}$	$3.98_{\pm 0.19}$	$0.727_{\pm 0.024}$	$4.01_{\pm 0.09}$	$0.696_{\pm 0.013}$	$4.12_{\pm 0.11}$	$0.711_{\pm 0.021}$	$4.09_{\pm 0.13}$	$0.789_{\pm 0.016}$
		MiniLM	$4.29_{\pm 0.12}$	$0.709_{\pm 0.020}$	$4.25_{\pm 0.17}$	$0.794_{\pm 0.018}$	$4.27_{\pm 0.15}$	$0.733_{\pm 0.012}$	$4.30_{\pm 0.16}$	$0.756_{\pm 0.024}$	$4.42_{\pm 0.11}$	$0.767_{\pm 0.014}$
		ExpPro	$\mathbf{4.39_{\pm 0.17}}$	$\mathbf{0.733_{\pm 0.019}}$	$\mathbf{4.37_{\pm 0.14}}$	$\mathbf{0.826_{\pm 0.015}}$	$\mathbf{4.44_{\pm 0.12}}$	$\mathbf{0.790_{\pm 0.012}}$	$\mathbf{4.31_{\pm 0.13}}$	$\mathbf{0.778_{\pm 0.016}}$	$\mathbf{4.46_{\pm 0.13}}$	$\mathbf{0.811_{\pm 0.013}}$

B. Compared Methods

To verify the effectiveness of our approach, we compare the following strategies for selecting prompt speech: 1) **Random:** We randomly select from all prompts as the prompt choice. 2) **Text-based Methods:** We achieve the selection by performing semantic similarity analysis between the synthetic text and the prompt text [15], using all-MiniLM-L6-v2[3] (MiniLM) [30] to implement the prompt selection.

C. Test Metrics

For our subjective evaluation, we employ 20 native speakers. For the main method comparison, we provide 5 sentences per emotion, each designed to convey the corresponding emotional state. For other ablation experiments, we select 5 descriptive neutral sentences to verify the effectiveness of our method. We provide participants with detailed evaluation criteria and report both the mean scores and 95% confidence intervals. The test metrics used in the subjective evaluation are as follows:

- **Emotion MOS (MOS):** This metric evaluates the quality and emotional expression of the synthesized speech.
- **Strength Perception (SP):** A subjective strength perception test. The judge is asked to rate the emotional strength on a scale from 0 to 1.

The object evaluation metrics include speaker similarity, ES, CER. Resemb and WavLM are calculated via cosine similarity between speaker representations of the target and generated speech using Resemblyzer [25] and WavLM-large [26], while ES uses cosine similarity between emotion2vec-large [27] representations. CER compares the synthesized text with Paraformer-zh [31] output.

IV. EXPERIMENTAL RESULTS

In our experiments, we validate our method through multiple approaches. Specifically, we conduct evaluations using two state-of-the-art and widely popular LM-based TTS models: CosyVoice [5] and GPT-SoVITS[4]. First, we compare our method with the baseline approach on prompt speech selection during zero-shot inference for both TTS models. Next, we explore the effects of model parameter selection and systematically validate the contribution of each module.

A. Comparison with Baseline Methods

By applying the prompt selection process during zero-shot inference with these two models, we compare our approach against other baseline methods. Detailed experimental results are presented in Table II. Our proposed method demonstrates significantly better performance across major emotional categories compared to the baselines, with noticeable improvements in both emotional intensity and stability. When comparing the results on the ESD dataset with those on the H2ED dataset, we observe that the stronger the emotional expressiveness of the prompt speech, the more pronounced the improvements achieved by our method. This approach thoroughly considers the quality of the prompts, the performance differences between models, and the correlation between synthesized text and prompt text, resulting in superior experimental outcomes.

TABLE III
THE RESULTS OF EXPPRO IN DIFFERENT LM-BASED TTS WITH THE SAME PROMPT CANDIDATES.

Model	PromptID	CER ↓	Resemb ↑	WavLM ↑	ES↑
CosyVoice	165(Top1)	1.55%	93.66	82.10	98.37
	112(Top2)	2.01%	90.67	81.74	98.45
	119(Top3)	1.86%	91.68	79.17	97.61
	047(Bottom2)	2.94%	84.43	63.67	50.30
	029(Bottom1)	2.63%	89.65	69.63	43.36
GPT-SoVITS	083(Top1)	1.55%	85.82	69.11	93.11
	031(Top2)	2.01%	85.27	61.75	96.17
	064(Top3)	1.70%	87.71	61.53	93.01
	099(Bottom2)	1.75%	79.71	49.39	65.37
	039(Bottom1)	2.24%	77.40	41.89	52.44

B. Evaluation on Different TTS Models

We further demonstrate the importance of the prompt model performance module. The results, presented in Table III, are obtained by ranking the prompts according to the weighted-sum evaluation of CER, speaker similarity, and ES, using the same happy emotion data from female speaker 1 across two different models. Our findings reveal that the Top three most expressive prompts and the Bottom two least expressive

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4]https://github.com/RVC-Boss/GPT-SoVITS

prompts differ significantly between the two models, despite utilizing identical input data. This indicates that the effectiveness of a given prompt is model-dependent, underscoring the necessity of evaluating prompt performance within the context of the specific TTS system.

Furthermore, the results indicate that CosyVoice outperforms GPT-SoVITS in terms of both speaker similarity and emotion similarity. This performance advantage is primarily attributed to CosyVoice's use of an ASR-supervised tokenizer and the integration of additional speaker x-vector inputs. Based on this result, all subsequent experiments are conducted primarily using the CosyVoice model.

TABLE IV
THE RESULT OF DIFFERENT PARAMETER SETTINGS. SP IS PRESENTED WITH A 95% CONFIDENCE INTERVAL.

	m	n	ES ↑	SP ↑
1	5	25	88.85	$0.724_{\pm 0.013}$
2	4	25	89.12	$0.733_{\pm 0.015}$
3	3	25	**89.83**	$0.735_{\pm 0.015}$
4	2	25	89.25	$\mathbf{0.737_{\pm 0.013}}$
5	3	25	89.83	$0.735_{\pm 0.014}$
6	3	20	92.27	$0.745_{\pm 0.016}$
7	3	15	94.01	$0.750_{\pm 0.013}$
8	3	10	**95.59**	$\mathbf{0.767_{\pm 0.011}}$

C. Range of Prompt Selection

We conduct experiments on the range of data selected at each stage, including the selection of m, n, k, and other variables under different conditions. The specific results are shown in Table IV. From the table, we can clearly observe that as the degree of selection of the pitch clusters increases (m decreases), the emotional impact of the prompts also gradually enhances. The results indicating that emotion similarity and strength perception increase as n decreases highlight the effectiveness of using overall perceptual quality and textual emotional coherence to prompt selection. Considering the limited number of prompt speeches, the prominence of the prompter's emotional effect, and the uncertainty of text content during inference, we ultimately choose the parameters $m = 3$, $n = 15$, and $k = 5$.

TABLE V
THE EXPERIMENT OF QUALITY SELECTION. ⊖ REPRESENTS THE REVERSE METHOD OF EXPPRO, PSQ DENOTES PROMPT SPEECH QUALITY. SP IS PRESENTED WITH A 95% CONFIDENCE INTERVAL.

Emotion	Method	CER ↓	Resemb ↑	WavLM ↑	ES ↑	SP ↑
Happy	ExpPro	**2.35%**	**91.37**	73.51	**93.08**	$\mathbf{0.782_{\pm 0.018}}$
	⊖PSQ	2.35%	91.31	**74.75**	90.46	$0.633_{\pm 0.014}$
Sad	ExpPro	**2.23%**	**90.28**	**81.41**	96.31	$\mathbf{0.724_{\pm 0.013}}$
	⊖PSQ	2.32%	89.90	78.03	**97.47**	$0.674_{\pm 0.017}$
Angry	ExpPro	**1.83%**	**92.33**	**76.59**	**96.31**	$\mathbf{0.697_{\pm 0.013}}$
	⊖PSQ	2.23%	86.76	67.26	91.29	$0.579_{\pm 0.016}$
Surprised	ExpPro	**1.67%**	87.41	**77.55**	**93.61**	$\mathbf{0.744_{\pm 0.011}}$
	⊖PSQ	1.95%	**89.25**	77.03	92.15	$0.646_{\pm 0.017}$
Comfort	ExpPro	**1.70%**	91.69	**84.32**	97.56	$\mathbf{0.741_{\pm 0.008}}$
	⊖PSQ	1.70%	**92.11**	82.21	**97.59**	$0.688_{\pm 0.007}$

D. Ablation Study

To better understand the contribution of each component in ExpPro, we conduct ablation studies by isolating and evaluating the impact of prompt speech quality and prompt model performance in two stages. The two stages are analyzed separately as follows:

1) Importance of Prompt Speech Quality: Table V presents the results of our experiments on the prompt speech quality module. We employ an inverse selection strategy compared to ExpPro to finish these experiments. Specifically, we select the m clusters that performed the worst after pitch clustering, along with the Bottom $n\%$ of data based on overall perceptual quality and textual emotional coherence weighted results. Finally, we compare the performance of the Top k prompts candidates through prompt model performance, respectively. The results indicate that ExpPro successfully selects emotionally expressive prompts.

TABLE VI
THE EXPERIMENT OF MODEL PERFORMANCE SELECTION. ⊖ REPRESENTS THE REVERSE METHOD OF EXPPRO, PMP DENOTES PROMPT MODEL PERFORMANCE. SP AND MOS ARE PRESENTED WITH A 95% CONFIDENCE INTERVAL.

Emotion	Method	MOS ↑	SP ↑
Happy	ExpPro	$\mathbf{4.30_{\pm 0.13}}$	$\mathbf{0.787_{\pm 0.017}}$
	⊖PMP	$4.27_{\pm 0.08}$	$0.773_{\pm 0.017}$
Sad	ExpPro	$\mathbf{4.24_{\pm 0.13}}$	$\mathbf{0.817_{\pm 0.012}}$
	⊖PMP	$4.21_{\pm 0.13}$	$0.773_{\pm 0.017}$
Angry	ExpPro	$\mathbf{4.33_{\pm 0.15}}$	$\mathbf{0.700_{\pm 0.015}}$
	⊖PMP	$4.28_{\pm 0.14}$	$0.677_{\pm 0.016}$
Surprised	ExpPro	$\mathbf{4.12_{\pm 0.11}}$	$\mathbf{0.727_{\pm 0.016}}$
	⊖PMP	$4.11_{\pm 0.13}$	$0.723_{\pm 0.015}$
Comfort	ExpPro	$\mathbf{4.27_{\pm 0.15}}$	$\mathbf{0.800_{\pm 0.011}}$
	⊖PMP	$4.22_{\pm 0.15}$	$0.760_{\pm 0.017}$

2) Importance of Prompt Model Performance: Table VI shows the results of the experiments on the prompt model performance module, after the prompt speech quality selected using the positive selection of ExpPro, we compare the performance of the Top k and Bottom k prompts candidates to validate the role of prompt model performance module. The experimental results indicate that the module significantly enhances the user's listening experience.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed ExpPro, a novel two-stage emotion prompt selection strategy that evaluates both the emotional quality of prompts and their generation performance.

ExpPro also performs dynamic prompt selection based on the input text to select the most relevant prompt among the emotional prompt candidates. The experiments show that, compared to the baseline methods, the speech generated using the prompt selection strategy proposed in this paper demonstrates advantages in emotional expressiveness, perceptual quality, and content accuracy. In the future, we will further explore prompt selection strategies across other dimensions and try to apply them to various tasks such as text-to-audio.

VI. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant U23B2053 and Grant 62176182.

References

[1] A. Radford, "Improving language understanding by generative pre-training," 2018.

[2] P. Anastassiou, J. Chen, J. Chen, *et al.*, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[3] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[4] T. Wang, L. Zhou, Z. Zhang, *et al.*, "VioLA: Conditional language models for speech recognition, synthesis, and translation," *IEEE Trans. ASLP.*, 2024.

[5] Z. Du, Q. Chen, S. Zhang, *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. ASLP.*, vol. 29, pp. 3451–3460, 2021.

[7] C. Qiang, H. Li, Y. Tian, *et al.*, "Learning speech representation from contrastive token-acoustic pretraining," in *Proc. ICASSP*, IEEE, 2024, pp. 10 196–10 200.

[8] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Proc. NIPS*, vol. 35, pp. 22 199–22 213, 2022.

[10] H. W. Chung, L. Hou, S. Longpre, *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[11] N. Nashid, M. Sintaha, and A. Mesbah, "Retrieval-based prompt selection for code-related few-shot learning," in *Proc. ICSE*, IEEE, 2023, pp. 2450–2462.

[12] Z. Wang, Z. Zhang, C.-Y. Lee, *et al.*, "Learning to prompt for continual learning," in *Proc. IEEE CVPR*, 2022, pp. 139–149.

[13] K. Shum, S. Diao, and T. Zhang, "Automatic prompt augmentation and selection with chain-of-thought from labeled data," *arXiv preprint arXiv:2302.12822*, 2023.

[14] J. Lou, Y. Lu, D. Dai, *et al.*, "Universal information extraction as unified semantic matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 13 318–13 326.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. EMNLP*, Association for Computational Linguistics, Nov. 2019.

[16] T. Bott, F. Lux, and N. T. Vu, "Controlling emotion in text-to-speech with natural language prompts," *arXiv preprint arXiv:2406.06406*, 2024.

[17] X. Li, Z.-Q. Cheng, J.-Y. He, X. Peng, and A. G. Hauptmann, "Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis," *arXiv preprint arXiv:2404.18398*, 2024.

[18] C. Gong, X. Wang, E. Cooper, *et al.*, "Zmm-tts: Zero-shot multilingual and multispeaker speech synthesis conditioned on self-supervised discrete speech representations," *IEEE Trans. ASLP.*, pp. 1–16, 2024.

[19] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *Journal of voice*, vol. 25, no. 1, e25–e34, 2011.

[21] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. ASLP.*, vol. 17, no. 4, pp. 582–596, 2009.

[22] D. Gharavian, M. Sheikhan, and M. Janipour, "Pitch in emotional speech and emotional speech recognition using pitch frequency," *Majlesi Journal of Electrical Engineering*, vol. 4, no. 1, p. 19, 2010.

[23] R. W. Frick, "Communicating emotion: The role of prosodic features.," *Psychological bulletin*, vol. 97, no. 3, p. 412, 1985.

[24] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, IEEE, 2021, pp. 6493–6497.

[25] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*, IEEE, 2018, pp. 4879–4883.

[26] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[27] Z. Ma, Z. Zheng, J. Ye, *et al.*, "Emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[28] C. Qiang, P. Yang, H. Che, Y. Zhang, X. Wang, and Z. Wang, "Improving prosody for cross-speaker style transfer by semi-supervised style extractor and hierarchical modeling in speech synthesis," in *Proc. ICASSP*, IEEE, 2023, pp. 1–5.

[29] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, IEEE, 2021, pp. 920–924.

[30] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Proc. NIPS*, vol. 33, pp. 5776–5788, 2020.

[31] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.