SMIIP-NV: A Multi-Annotation Non-Verbal Expressive Speech Corpus in Mandarin for LLM-Based Speech Synthesis

Zhuojun Wu
zhuojun.wu@whu.edu.cn
School of Computer Science, Wuhan University
Wuhan, China
Duke Kunshan University
Kunshan, China

Dong Liu
dong.liu@whu.edu.cn
School of Computer Science, Wuhan University
Wuhan, China
Duke Kunshan University
Kunshan, China

Juan Liu*
liujuan@whu.edu.cn
School of Artificial Intelligence, Wuhan University
School of Computer Science, Wuhan University
Wuhan, China

Yechen Wang
yechen@ofspectrum.com
OfSpectrum, Inc.

Linxi Li
linxili@ofspectrum.com
OfSpectrum, Inc.

Liwei Jin
liweijin@ofspectrum.com
OfSpectrum, Inc.

Hui Bu
buhui@aishelldata.com
Beijing AISHELL Technology Co., Ltd., Beijing, China

Pengyuan Zhang
zhangpengyuan@hccl.ioa.ac.cn
Speech and Intelligent Information Processing Laboratory, Institute of Acoustics, Chinese Academy of Sciences
Beijing, China

Ming Li*
ming.li369@dukekunshan.edu.cn
School of Computer Science, Wuhan University
Wuhan, China
Duke Kunshan University
Kunshan, China

Abstract

In natural language communication, emotions are often conveyed through non-verbal sounds (NVs), such as laughter, crying, cough and so on. However, most existing text-to-speech (TTS) corpora lack annotations for these non-verbal sounds, leading to a scarcity of systems capable of generating them. To address this gap, we introduce SMIIP-NV, a non-verbal speech synthesis corpus annotated with both emotions and non-verbal sounds, including laughter, crying, and cough. To the best of our knowledge, SMIIP-NV is the largest publicly available open-source expressive speech corpus that includes non-verbal speech and rich annotations. It comprises 33 hours of speech data, covering five distinct emotions and three types of non-verbal sounds, with detailed transcriptions and precise timestamps for each occurrence of non-verbal sounds. Additionally, the corpus provides annotations for speech segments that contain laughter or crying. To demonstrate the utility of this dataset, we establish a baseline for non-verbal speech synthesis by employing a lightweight large language model (LLM). The SMIIP-NV dataset and static audio demonstrations are publicly available at https://axunyii.github.io/SMIIP-NV. The interactive real-time demonstrations can be accessed at https://huggingface.co/spaces/xunyi/SMIIP-NV_Finetuned_CosyVoice2.

CCS Concepts

• **Computing methodologies** → *Natural language processing*.

Keywords

Speech synthesis, Non-verbal speech corpora, Large language models

1 Introduction

With the continuous advancement of Large Language Models (LLMs) in the field of Natural Language Processing (NLP), their applications in speech synthesis have also made significant progress[1, 4, 11, 15]. Contemporary LLM-based speech synthesis [7, 10, 15, 16, 18, 24] has shifted the focus of speech generation from clarity and intelligibility to naturalness and emotional expressiveness.

Recent research in expressive speech synthesis has primarily focused on emotion modeling and context-aware synthesis. For example, the framework proposed in [2] integrates emotion recognition and contextual reasoning with large-scale language models to improve emotional consistency and speech naturalness. Similarly, [13] enhances speech expressiveness by incorporating annotated textual features, such as sentence pattern, rhetorical device, scene, emotional color, and imitated character[13], enriching onomatopoeic elements and emotional variation in synthesized speech.

Despite these advancements in synthesizing emotional and context-aware speech, the generation of non-verbal sounds (NVs)—such as laughter, crying, and cough—remains an underexplored area[12]. NVs are a crucial component of human communication, not only conveying emotional information but also modulating intonation,

*Corresponding Author: Juan Liu, Ming Li

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758312

reinforcing semantics, and expressing social signals in spoken discourse[3]. However, progress in non-verbal sound synthesis has been hindered by the lack of comprehensive datasets, particularly open-source corpora that contain rich annotations for non-verbal expressions. This limitation constrains the development of systems capable of effectively generating these sounds.

To address this gap, we introduce SMIIP-NV, a highly expressive speech synthesis corpus designed to support non-verbal sound synthesis research. The SMIIP-NV dataset comprises 33 hours of speech data annotated with five distinct emotional states and three types of non-verbal sounds (laughter, crying, and cough). It provides precise labels and timestamps for non-verbal sounds and accounts for intertwined phenomena, such as speech interspersed with laughter or crying. Furthermore, we establish a baseline for non-verbal speech synthesis using the SMIIP-NV dataset and demonstrate its potential through a lightweight Large Language Model (LLM).

2 Related Work

2.1 Non-verbal Speech Synthesis

Recent advancements in non-verbal speech synthesis have primarily focused on laughter generation. For instance, [6] utilized personal recordings to train a Hidden Markov Model (HMM) for laughter synthesis, blending laughter vowels with neutral or smiling speech to enhance naturalness. Similarly, [14] employed two spontaneous speech corpora within a statistical HMM-based speech synthesis framework to synthesize conversational laughter. This approach incorporated contextual factors, such as phonetic properties and positional aspects, to improve the naturalness of synthesized laughter within speech.

Other studies have explored laughter synthesis through self-supervised learning methods. For example, [22] introduced a Japanese laughter synthesis corpus and employed a pseudo-polyphonic tokens (PPTs)-based approach to achieve high-quality laughter synthesis. This method leveraged HuBERT and K-means clustering for feature extraction. In contrast, [23] applied self-supervised learning to generate Unsupervised Language Units (ULUs) for non-verbal annotation and zero-shot vocalization migration, enabling more flexible synthesis of non-verbal sounds.

Furthermore, [17] proposed EmoCtrl-TTS, a stream-matching-based, zero-shot, emotion-controlled TTS system capable of generating emotive speech containing non-verbal vocalizations, such as laughter and crying. This system utilizes arousal and emotion values, along with laughter embeddings, to generate emotionally expressive speech using large-scale expressive speech data.

2.2 Non-Verbal Speech Dataset

Previous studies on non-verbal speech datasets have demonstrated significant variation in both length and availability. For instance, [6] utilized a dataset comprising 54 minutes of neutral speech, 12 minutes of speech with smiles, and 4 minutes of laughter, all recorded from a single target speaker. Similarly, [14] employed two spontaneous speech corpora: the Utsunomiya University Spoken Dialogue Database (UU Database), which contains 4,840 utterances, including 280 annotated laughs, and the Online Gaming Voice Chat Corpus (OGVC), comprising 9,114 utterances with 1,593 annotated laughs. However, these datasets are not publicly accessible.

In contrast, [22] introduced a 3.5-hour open-source laughter synthesis corpus for Japanese. Expanding on this, [21] developed a Japanese emotional speech corpus that includes both verbal content and non-verbal expressions commonly used in daily conversations, such as laughter and sobbing. This corpus comprises 3.94 hours of speech from four speakers, covering six emotions (anger, disgust, fear, happiness, sadness, and surprise), making it the largest publicly available dataset to date containing non-verbal sounds. Additionally, [23] utilized three private, highly expressive Chinese corpora, totaling 0.24 hours of non-verbal pronunciations. Meanwhile, [17] trained their model on a large-scale real-world sentiment dataset, comprising 27,000 hours of speech derived from 200,000 hours of anonymized English audio. However, due to privacy and security concerns, this dataset has not been publicly released.

These efforts underscore the significant progress in non-verbal sound synthesis, yet much of the existing research has been limited to specific types of non-verbal sounds, such as laughter, and has been constrained by the availability of annotated corpora. To address these challenges, SMIIP-NV provides a rich, open-source dataset that facilitates comprehensive research on the synthesis of various non-verbal sounds, thereby advancing the field of expressive speech synthesis.

3 SMIIP-NV Dataset

In this section, we introduce SMIIP-NV, a mandarin meticulously curated corpus designed for non-verbal speech synthesis. This richly annotated dataset encompasses a diverse range of non-verbal sounds, providing a valuable resource for emotionally expressive speech synthesis across various contexts. All recordings were made using a Neumann U87 microphone in a recording studio with a background noise level of 25 dB.

3.1 Data Recording and Annotation

Given that the production of non-verbal sounds is highly context-dependent, it is crucial to design emotionally rich texts that naturally incorporate these sounds to preserve the authenticity of the recorded data. The data collection process involved several key stages.

In the text organization phase, we developed speech scripts based on everyday conversational scenarios, ensuring that the dialogues elicited highly emotional utterances incorporating non-verbal sounds. These scripts included dialogues designed to evoke strong emotional responses. We crafted expressive sentences for five distinct emotions—happiness, surprise, sadness, neutrality, and anger—and integrated non-verbal sounds into the expression of each emotional state.

Following the design phase, the dataset was recorded by speakers who naturally express non-verbal sounds. Audio recordings were captured with a high-fidelity microphone at a 48 kHz sampling rate, positioned close to the bottom of the speaker's lips to ensure optimal sound quality. Depending on the scene and text content, participants either used scripted prompts or improvised

to add appropriate nonverbal sounds. To maintain contextual accuracy, non-verbal sounds were carefully selected to align with sentence context, and their naturalness was supervised manually throughout the recording process.

Upon completing the recordings, we conducted proofreading and annotation of the recorded data. Initially, we verified and marked the locations of non-verbal sounds within the corresponding text. Subsequently, we manually annotated timestamps for each non-verbal sound to ensure precise alignment with the audio.

In cases where the boundaries of non-verbal sounds were difficult to define—such as when laughter or crying occurred simultaneously with speech—we employed an alternative annotation approach. This method involved labeling the speech segments containing these overlapping phenomena and assigning appropriate start and end timestamps to ensure accurate representation within the dataset.

3.2 Data Processing

We aimed to maintain the duration of each audio utterance within a range of 3 to 23 seconds, reflecting the typical length of everyday speech that includes non-verbal sounds. To ensure consistency and prevent the omission of specific non-verbal sounds, we sought to maintain comparable durations across different types of non-verbal expressions.

During the recording process, we observed that some speakers exhibited greater proficiency in producing certain types of non-verbal sounds, while others were less adept. As a result, the duration of non-verbal sounds varied across speakers. To account for these differences, we preserved speaker-ID information within the dataset, ensuring a more comprehensive and representative resource for non-verbal speech synthesis.

In summary, our dataset includes speaker information, text content, non-verbal vocalization locations, timestamp annotations, and emotion labels for audio containing non-verbal sounds. Given the varied ways in which the data can be processed depending on training objectives, we have not pre-partitioned the original dataset into training and test sets, retaining all labels for flexible partitioning based on specific research tasks.

However, for the baseline task of non-verbal speech synthesis using LLMs, we specifically partitioned the dataset into training and test sets to facilitate model evaluation.

3.3 Train-test Separation

For our proposed non-verbal speech synthesis baseline, we randomly partitioned the dataset into training and test sets using an 85:15 ratio, ensuring that non-verbal labels were retained within transcriptions and sentences. Specifically, the dataset was first partitioned separately within each emotion category, maintaining the 85:15 split for each emotion-specific folder. Subsequently, the training subsets from all emotion categories were merged to form the final training set, while the test subsets were combined to construct the test set. The statistical details of these partitions are presented in Table 1.

Table 1: Utterances Statistics for the Train and Test Set

Set	Emotion	Time (h)	Utterances
train	happy	10.37	5374
	sad	9.30	3992
	neutral	5.94	3790
	angry	2.03	1359
	surprised	0.49	254
test	happy	1.83	949
	sad	1.63	705
	neutral	1.06	669
	angry	0.36	240
	surprised	0.08	45

3.4 Data Statistics

The dataset comprises five distinct emotion categories and features 37 speakers. A speaker may record multiple emotional speech. The statistical distribution of audio data across these emotion categories is presented in Table 2. In total, the dataset contains 33.33 hours of recorded audio. Additionally, the statistics for audio clips containing different non-verbal sounds are detailed in Table 3. The temporal distribution of emotion and nonverbal sounds in SMIIP-NV is displayed in Fig. 1.

Table 2: Emotion Statistics for The Whole Dataset

Emotion	Utterances	Speakers	NV Type
happy	6323	36	laughter
sad	4697	34	crying
neutral	4459	29	cough
angry	1599	11	cough
surprised	299	5	cough

Table 3: Utterance Statistics for Different Non-verbal Sounds

NV Type	Utterances
laughter	6161
laugh while talking	550
crying	4611
cry while talking	367
cough	6348

3.5 Potential Applications

The SMIIP-NV dataset provides a wide range of emotionally expressive non-verbal speech samples, significantly enhancing synthetic audio's ability to realistically and naturally reproduce human paralinguistic cues. Specifically, emotional audio containing non-verbal speech elements can be generated based on a simple baseline model. Further, feature extraction can be performed for different types of non-verbal speech signals, allowing for more

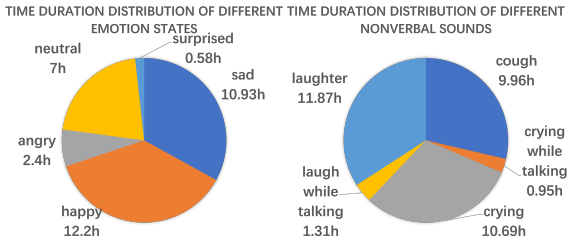

Figure 1: Time Duration Distributions across Different Emotion and Non-verbal Sound Categories

fine-grained control and tuning of the synthesis process. In addition, given that the dataset contains timestamps of nonverbal speech information, this feature enables text-to-speech (TTS) frameworks based on different architectures to be used for the synthesis and precise control of nonverbal speech.

4 Experiments

In this section, we describe the framework of the proposed baseline model and training method. Based on the SMIIP-NV dataset, we construct a baseline model that enables systematic modeling of non-verbal speech synthesis tasks by integrating speech transcription text with annotated data on non-verbal sounds and emotions. To validate the effectiveness of SMIIP-NV, we also fine-tuned it on CosyVoice2[5]. This provides the research community with reproducible and standardized evaluation benchmarks.

4.1 Baseline Setups

4.1.1 Baseline Model. The baseline architecture is adapted from [19], a lightweight non-autoregressive speech synthesis model based on the Transformer architecture. This model is specifically designed to synthesize natural speech while operating within limited computational resources.

For non-verbal sounds, we incorporate them as special tokens within the training process, enabling the model to effectively learn and generate non-verbal components alongside speech synthesis. Additionally, we integrate emotion labels as part of the training pipeline. These emotion labels, which are derived from the SMIIP-NV dataset, allow the model to generate emotionally expressive speech by conditioning the synthesis process on the corresponding emotional states. The overall architecture of our baseline model and the inputs of non-verbal token and emotion labels are shown in Fig. 2.

4.1.2 Training and Inference Procedures. During training, we first expanded the pre-training dataset of the baseline model from 900 hours in [19] to the full 10,000 hours of WenetSpeech4TTS and trained it for 50 epochs, leading to a reduction in Character Error Rate (CER) from 22 to 15.20. This enhancement significantly improves the accuracy of synthesized audio, particularly under limited computational resources.

Following pre-training, we fine-tuned the model using the SMIIP-NV dataset, which includes both non-verbal speech components and emotion labels. The training process was conducted for 500 epochs utilizing 8 NVIDIA A6000 GPUs.

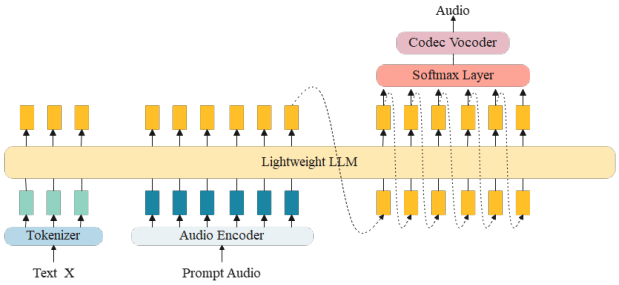

Figure 2: Overall architecture of the proposed baseline model.

During inference, we control the synthesis of speech by incorporating special tokens into the text to designate the placement of non-verbal sounds and emotional states. The special tokens and their corresponding usage are detailed in Table 4.

4.2 Open Source System

We also evaluated the effectiveness and generalization ability of the SMIIP-NV dataset on the open source speech synthesis system CosyVoice2. In our experiments, we used the CosyVoice2-0.5B checkpoint, which is the officially baseline model for CosyVoice2. This checkpoint is the starting point of our fine-tuning process. In the fine-tuning phase, we use only non-verbal tokens, adding them to the sequence of special tokens. The fine-tuning process trained for 200 epochs using a distributed setup of 8 NVIDIA A6000 GPUs. To reduce the influence of the original system, we use different tokens than the original CosyVoice2 non-verbal tokens.

4.3 Evaluation Metrics

In order to better evaluate our proposed baseline model, we performed both subjective and objective evaluations and further analyzed the emotional expressiveness of the synthesized speech through an emotion confusion matrix. For the fine-tuned CosyVoice2, we performed a subjective evaluation to verify the effectiveness of SMIIP-NV for generating non-verbal sounds.

4.3.1 Objective Evaluation. For the objective evaluation, we assessed the performance of the fine-tuned model on the test set from [20]. This external benchmark was augmented with randomized non-verbal labels to evaluate our model's synthesis capability on an external dataset. Using this standard set facilitates comparison of general speech synthesis quality (measured by CER) with the pre-trained baseline.

To assess intelligibility, we measured the Character Error Rate (CER) by comparing the Automatic Speech Recognition (ASR) results of the synthesized speech against the input text. For ASR transcription, we employed the Paraformer-Large model[8].

Additionally, for zero-shot evaluation, we assessed speaker similarity (SS) by measuring the speaker similarity between the generated speech and the speech prompt calculated by Resemblyzer[9].

4.3.2 Subjective Evaluation. For subjective evaluation, we randomly selected three samples from each synthetic non-verbal sentence

and three samples from each non-verbal utterance in the original recorded dataset, in total 18 samples. A total of 15 Chinese native speakers participated as judges.

We employed the Mean Opinion Score (MOS) method to assess the naturalness of the synthesized non-verbal utterances. The MOS ratings were measured on a five-point scale, ranging from 1 to 5, with 0.5-point increments.

4.3.3 Emotion Confusion Matrix Evaluation. To further assess the emotion expressiveness and discriminability of the synthesized non-verbal speech, we constructed an emotion confusion matrix. Specifically, we selected 5 synthesized speech utterances for each emotion, invited 15 native Chinese speakers from the same group to listen to the synthesized samples, and selected the perceived emotion category of each utterance from the five emotion labels.

Then, the confusion matrix was calculated by comparing the expected emotion labels of the synthesized utterances with the perceived emotion categories assigned by the judges. By analyzing the emotion confusion matrix, we can evaluate the effectiveness of emotion expression in synthetic speech.

5 Results

5.1 Baseline Model

In this experiment, we fine-tuned a pre-trained speech synthesis model using a dataset containing non-verbal speech and evaluated its objective performance metrics. The results, presented in Table 5, indicate a Character Error Rate (CER) of 15.46, while the synthesized speech achieved a Speaker Similarity (SS) score of 80.45%.

Compared to the pre-training TTS model prior to fine-tuning, the CER under ASR evaluation did not significantly increase, demonstrating that the inclusion of non-verbal data in the fine-tuning process does not adversely affect the model's ability to synthesize standard spoken sentences. Instead, these results show that, while preserving core synthesis quality, the fine-tuned model acquires the additional capacity to generate natural-sounding non-verbal sounds (laughs, cries, coughs) within its outputs.

Table 6 presents the Mean Opinion Scores (MOS) for authentic and synthesized speech across the three non-verbal (NV) categories: cough, laughter, and crying. The results indicate that the MOS scores for synthesized speech are comparable to those of authentic speech, demonstrating that fine-tuning the LLM-based speech synthesis system using SMIIP-NV is effective in synthesizing non-verbal speech.

Notably, the MOS scores for synthesized laughter and crying (both 4.18) are very close to or even slightly higher than those of authentic speech (4.33 and 3.96, respectively). This suggests that the synthesis process effectively preserves the naturalness of these non-verbal sounds, making them perceptually indistinguishable from real speech. Furthermore, due to training with diverse non-verbal contexts, the synthesized outputs may even appear more natural than some real counterparts.

Although the MOS score for synthesized cough (3.99) is slightly lower than that of authentic cough (4.44), this small difference indicates that the synthesized version still maintains a high degree of naturalness.

The emotion perception performance of the synthesized speech was analyzed using a confusion matrix evaluation (Fig.3), showing effective emotion discrimination by all 15 native Chinese speakers. The diagonal dominance confirms the successful encoding of the target emotion. The slight increase in the error rate of the surprise category may be due to its lower amount of data in the training corpus. These results demonstrate the capability of SMIIP-NV for training emotionally expressive speech synthesis models.

5.2 CosyVoice2 Fine-tuning Results

To further evaluate the generalization capability and effectiveness of the SMIIP-NV dataset on a different, large-scale open-source model, we fine-tuned the open-source CosyVoice2 system as described in Section 4.2. Consistent with the evaluation plan for this model, we performed a subjective listening test focusing on the naturalness of the synthesized non-verbal sounds, quantified using Mean Opinion Score (MOS).

For the results of the fine-tuning experiments with CosyVoice2, we re-selected 18 samples for the subjective evaluation, using the

Table 4: Special token usage for non-verbal sounds and emotions.

Token Type	Usage Example
Non-verbal Tokens	
<laughter>	我决定要读取书籍，但又瞬间被可爱猫咪吸引 <laughter>
<laughter/> </laughter>	上周我去参加聚会，<laughter/> 结果发现自己是唯一一个没带瓶酒的 </laughter>
<crying>	曾经的承诺如今变得苍白，<crying> 我只能默默叹息
<crying/> </crying>	<crying/> 我恨自己无法忘记，往事如风 </crying>，痛苦却难以释怀
<cough>	整理居住空间，<cough> 可以让心情变得更加愉悦
Emotion Tokens	
Neutral	<|emo_NEUTRAL|> 在雨天听音乐，<cough> 是一种怡然自得的享受 </|emo_NEUTRAL|>
Sad	<|emo_SAD|> 羞愧和遗憾交织，<crying> 无法再给你一个拥抱 </|emo_SAD|>
Happy	<|emo_HAPPY|> 初次见面，朋友问我最喜欢什么，我说“吃”，他回我“那你真是个美食家”<laughter></|emo_HAPPY|>
Angry	<|emo_ANGRY|> 我心中的怒焰 <cough> 无法熄灭，顾虑毫无意义 </|emo_ANGRY|>
Surprised	<|emo_SURPRISED|> 某些虫子 <cough> 竟然能够在毒雾中生存 </|emo_SURPRISED|>

Table 5: Objective Evaluation Results with Comparison to Fine-Tuned and Non-Fine-Tuned Models. CER stands for Character Error Rate, and SS stands for Speaker Similarity.

Criteria	Pretrain Model(With NVs)	Fine-Tuned Model(With NVs)	Pretrain Model(Without NVs)	Fine-Tuned Model(Without NVs)
CER	N/A	15.46	15.20	15.37
SS	N/A	81.20	80.45	80.68

Table 6: The Subjective MOS Score Results of Our Baseline Model Under Different Non-verbal Utterance Types

NV Type	Authentic Speech MOS	Synthesized Speech MOS
cough	4.44 ±0.04	3.99 ±0.07
laughter	4.33 ±0.05	4.18 ±0.06
crying	3.96 ±0.06	4.18 ±0.07

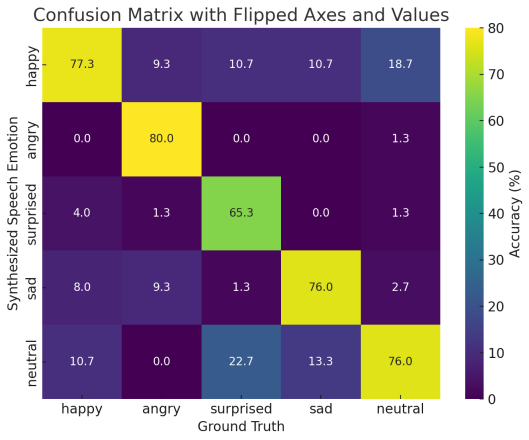

Figure 3: Emotion Confusion Matrix in Subjective Evaluation for Synthesized Non-verbal Speech

same group of judges. Table 7 presents the MOS results obtained from the subjective evaluation of the fine-tuned CosyVoice2 model's ability to synthesize cough, laughter, and crying. For comparison, we also include the MOS scores for the authentic speech.

Table 7: The Subjective MOS Score Results of Our Finetuned Cosyvoice2 Model Under Different Non-verbal Utterance Types

NV Type	Authentic Speech MOS	Synthesized Speech MOS
cough	4.12 ±0.07	4.28 ±0.05
laughter	4.22 ±0.06	4.15 ±0.06
crying	4.31 ±0.05	4.13 ±0.05

The results presented in Table 7 clearly demonstrate that fine-tuning CosyVoice2 with the SMIIP-NV dataset significantly enhances its capability to synthesize natural-sounding non-verbal utterances. The MOS scores for synthesized cough, laughter, and crying are highly competitive with, and in one case surpass, the scores for authentic speech. Specifically, synthesized cough achieved a MOS of 4.28, which is notably higher than the authentic cough score of 4.12. For laughter and crying, the synthesized scores (4.15 and 4.13 respectively) are very close to their authentic counterparts (4.22 and 4.31), demonstrating a high degree of perceptual naturalness. All synthesized non-verbal types obtained MOS scores exceeding 4.0, signifying excellent perceptual quality and acceptability.

The results of these subjective evaluations demonstrate that the SMIIP-NV dataset can effectively transfer non-verbal synthesis functions to different advanced TTS architectures such as CosyVoice2, provided that provisions for special tokens are made in the tokenizer. The fine-tuned models successfully learned to generate these complex non-verbal sounds in a highly natural way, which can be used to develop more expressive and human-like speech synthesis systems for various models.

6 Conclusion

This paper introduces SMIIP-NV, a novel and highly expressive audio dataset focused on non-verbal representations. As the currently largest open-source dataset of its kind, SMIIP-NV provides a valuable resource for research in expressive speech synthesis. We also propose a baseline system that achieves strong performance in generating non-verbal expressions, corroborated by a high Mean Opinion Score (MOS). Furthermore, SMIIP-NV's effectiveness and versatility are demonstrated through its integration with other open-source systems, enhancing the expressiveness and naturalness of generated speech.

7 Dataset and Code Availability

The SMIIP-NV dataset, accompanied by static audio demonstrations, is publicly available at https://axunyii.github.io/SMIIP-NV and licensed for non-commercial academic research and use. To promote transparency and facilitate further research, we provide the open-source fine-tuning code for the CosyVoice2 model on the SMIIP-NV dataset, along with the resulting fine-tuned model weights, at https://huggingface.co/xunyi/SMIIP-NV_finetune_CosyVoice2. Furthermore, an interactive, real-time demonstration of the fine-tuned CosyVoice2 model is hosted at https://huggingface.co/spaces/xunyi/SMIIP-NV_Finetuned_CosyVoice2.

Acknowledgments

This research is funded by DKU foundation project "Emerging AI Technologies for Natural Language Processing". Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

References

[1] James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243* (2023).

[2] Jun-Hyeok Cha, Seung-Bin Kim, Hyung-Seok Oh, and Seong-Whan Lee. 2025. JELLY: Joint Emotion Recognition and Context Reasoning with LLMs for Conversational Speech Synthesis. *arXiv preprint arXiv:2501.04904* (2025).

[3] Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. 2019. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest* 20, 1 (2019), 69–90.

[4] Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, Zhikang Niu, Shuai Wang, Hui Zhang, Xie Chen, and Kai Yu. 2025. Vall-t: Decoder-only generative transducer for robust and decoding-controllable text-to-speech. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[5] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117* (2024).

[6] Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, and Thierry Dutoit. 2015. Speech-laughs: an HMM-based approach for amused speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4939–4943.

[7] Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[8] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317* (2022).

[9] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[10] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun MA, and Zhou Zhao. 2024. Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=mvMI3N4AvD

[11] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics* 11 (2023), 1703–1718.

[12] Alkis Koudounas, Moreno La Quatra, Sabato Marco Siniscalchi, and Elena Baralis. 2025. voc2vec: A foundation model for non-verbal vocalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[13] Sen Liu, Yiwei Guo, Xie Chen, and Kai Yu. 2024. StoryTTS: A Highly Expressive Text-to-Speech Dataset with Rich Textual Expressiveness Annotations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11521–11525.

[14] Tomohiro Nagata and Hiroki Mori. 2018. Defining laughter context for laughter synthesis with spontaneous speech corpus. *IEEE Transactions on Affective Computing* 11, 3 (2018), 553–559.

[15] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[16] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710* (2025).

[17] Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, et al. 2024. Laugh Now Cry Later: Controlling Time-Varying Emotional States of Flow-Matching-Based Zero-Shot Text-To-Speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 690–697.

[18] Zhichao Wu, Yueteng Kang, Songjun Cao, Long Ma, Qiulin Li, and Qun Yang. 2025. MPE-TTS: Customized Emotion Zero-Shot Text-To-Speech Using Multi-Modal Prompt. *arXiv preprint arXiv:2505.18453* (2025).

[19] Zhuojun Wu, Dong Liu, and Ming Li. 2024. Lightweight Language Model for Speech Synthesis: Attempts and Analysis. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 501–505.

[20] Kangxiang Xia, Dake Guo, Jixun Yao, Liumeng Xue, Hanzhao Li, Shuai Wang, Zhao Guo, Lei Xie, Qingqing Zhang, Lei Luo, et al. 2024. The ISCSLP 2024 Conversational Voice Clone (CoVoC) Challenge: Tasks, Results and Findings. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 506–510.

[21] Detai Xin, Junfeng Jiang, Shinnosuke Takamichi, Yuki Saito, Akiko Aizawa, and Hiroshi Saruwatari. 2024. JVNV: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access* 12 (2024), 19752–19764.

[22] Detai Xin, Shinnosuke Takamichi, Ai Morimatsu, and Hiroshi Saruwatari. 2023. Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus. *arXiv preprint arXiv:2305.12442* (2023).

[23] Haitong Zhang, Xinyuan Yu, and Yue Lin. 2023. NSV-TTS: Non-Speech Vocalization Modeling And Transfer In Emotional Text-To-Speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[24] Kun Zhou, You Zhang, Shengkui Zhao, Hao Wang, Zexu Pan, Dianwen Ng, Chong Zhang, Chongjia Ni, Yukun Ma, Trung Hieu Nguyen, et al. 2024. Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions. *arXiv preprint arXiv:2409.16681* (2024).