

# Supplementary Materials: Generative Expressive Conversational Speech Synthesis

## A MORE INFORMATION ABOUT THE NCSSD

### A.1 The Detailed Data Source

The collection subset of NCSSD, CL-EN and CL-ZH, are extracted from 79 English and 34 Chinese TV shows respectively. Please refer to the entire TV show list in Figure 1 (The Roman numerals from I to X represent the seasons of a specific TV show).

English TV shows		
1. A Kind of Spark	28. Modern Family Season II	55. American Rust I
2. Back to the Rafters	29. Modern Family Season III	56. After Life II
3. Brothers & Sisters II	30. Modern Family Season VIII	57. Alexa & Katie II
4. Brothers & Sisters III	31. Modern Family Season IX	58. Alexa & Katie III
5. Brothers & Sisters IV	32. Modern Family Season X	59. Alexa & Katie IV
6. Chesapeake Shores I	33. Mom VII	60. All Creatures Great and Small III
7. Chesapeake Shores II	34. Parenthood I	61. All Creatures Great and Small IV
8. Chesapeake Shores III	35. Parenthood II	62. A.P. Bio I
9. Dive Club	36. Parenthood III	63. A.P. Bio II
10. Everybody Loves Raymond I	37. Parenthood V	64. Here and Now
11. Everybody Loves Raymond II	38. Shameless II	65. Life Unexpected I
12. Everybody Loves Raymond III	39. The Baby-Sitters Club II	66. Life Unexpected II
13. Everybody Loves Raymond IV	40. The Fosters I	67. Modern Love I
14. Everybody Loves Raymond V	41. The Fosters II	68. New Amsterdam I
15. Everybody Loves Raymond VI	42. The Fosters III	69. The Durrells IV
16. Everybody Loves Raymond VII	43. The Righteous Gemstones I	70. This Is Us II
17. Everybody Loves Raymond VIII	44. Workin' Moms I	71. This Is Us III
18. Everybody Loves Raymond IX	45. Doogie Kealoha, M.D II	72. This Is Us IV
19. Everybody Hates Chris II	46. Heartland I	73. This Is Us V
20. Everybody Hates Chris III	47. Brothers & Sisters V	74. Everything Now
21. Fuller House I	48. Big Love I	75. My Life with the Walter Boys
22. Fuller House II	49. Big Love II	76. 2 Broke Girls III
23. Fuller House III	50. Big Love III	77. Black-ish IV
24. Fuller House V	51. Big Love IV	78. Suit I
25. Heartland X	52. Big Love V	79. Trying II
26. Manifest I	53. Six Feet Under IV	
27. Modern Family Season I	54. Six Feet Under V	

Chinese TV shows		
1. 小舍得 (A Love For Dilemma)	18. 暖阳之下 (Modern City)	
2. 大江大河 I (Like a Flowing River I)	19. 陪读妈妈 (Always with You)	
3. 大江大河 II (Like a Flowing River II)	20. 千金醒来 (Qian Jin Xing Lai)	
4. 生活家 (My Treasure)	21. 生逢灿烂的日子 (A Splendid Life in Beijing)	
5. 微笑妈妈 (Smile Mom)	22. 熟年 (In Later Years)	
6. 爱情万万岁 (Forever Love)	23. 外婆的新世界 (Grandma's New World)	
7. 安居 (Housing)	24. 我的岳父会武术 (My Kungfu Father-in-law)	
8. 曾少年之小时候 (Once and Forever: The Sun Rises)	25. 我们的日子 (Days of Our Lives)	
9. 飞驰余生 (The Galloping Remaining Years)	26. 我们好好在一起 (Let's be together)	
10. 父母爱情 (Romance of Our Parents)	27. 心居 (Life is A Long Quiet River)	
11. 回来的女儿 (Homesick)	28. 幸福三重奏 (As We Wish)	
12. 加油妈妈 (Mom Wow)	29. 幸福三重奏II (Xing Fu San Chong Zou II)	
13. 结婚的秘密 (Marriage Secret)	30. 装台 (Stage Builder)	
14. 金婚 (Gold Marriage)	31. 纵有疾风起 (In Spite of the Strong Wind)	
15. 辣妈正传 (Hot Mom)	32. 隐秘的角落 (The Bad Kids)	
16. 龙城 (Dragon Town Story)	33. 以家人之名 (Go Ahead)	
17. 漫长的季节 (The Long Season)	34. 生命中的好日子 (Good day in life)	

Figure 1: Detailed list of TV Shows for collection subset of NCSSD.

### A.2 The Construction Workflow of NCSSD

As shown in Figure 3, we visualize the construction process of NCSSD. The upper panel displays the automatic pipeline for collection subset, including Video Selection, Dialogue Scene Extraction, Dialogue Segment Extraction, and Dialogue Script Recognition. The lower panel shows the workflow for recording subset, including Dialogue Script Draft Generation, Spoken Dialogue Recording, and Dialogue Script Re-identification.

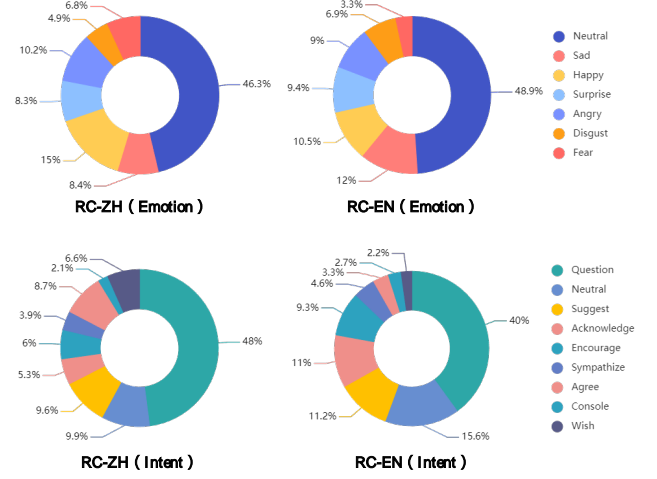


Figure 2: Statistics of emotion and intent labels in the recorded data of NCSSD (RC-EN and RC-ZH).

Table 1: Detailed splits for each subset of NCSSD.

SubSet	Language	Dialogues			
		Train	Valid	Test	All
CL	EN	5626	707	700	7033
	ZH	7120	906	750	8776
RC	EN	957	139	100	1196
	ZH	2006	245	200	2451

### A.3 More Data Statistics

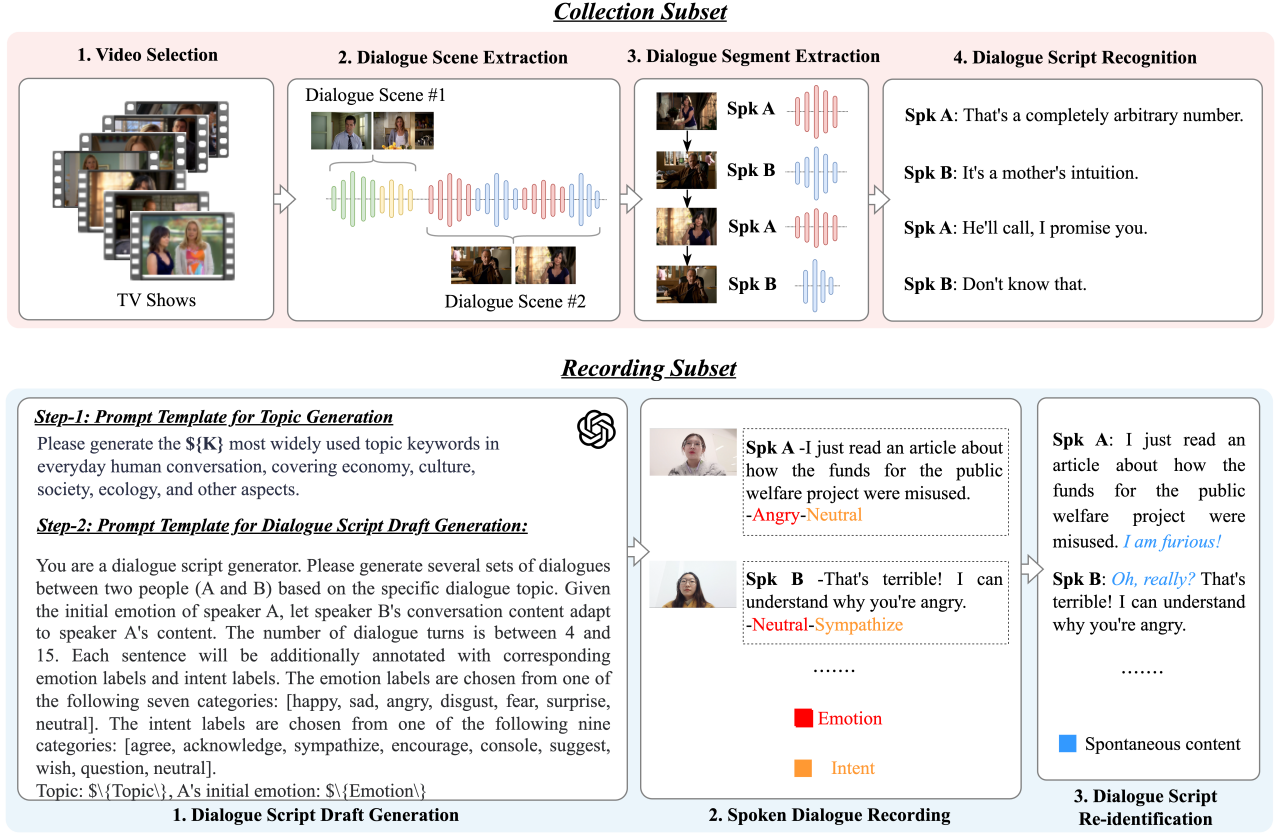
**A.3.1 Emotion and Intent statistics.** During the development of the recording subsets for NCSSD (RC-EN and RC-ZH), we employ ChatGPT to create scripts for two-person dialogues, along with emotional and intent labels for each sentence to help the recording volunteers. As show in Figure 2, we separately statistic the distribution of seven emotion labels and nine intent labels in RC-EN and RC-ZH.

**A.3.2 NCSSD splits.** Table 1 presents the division of Train, Valid, and Test sets for the number of dialogues in each subset of NCSSD.

### A.4 Ethics Statement

This work provides the NCSSD dataset to the research community for free, which is used for conversational speech synthesis research.

For the collection subset (CL-EN and CL-ZH), which primarily derives from Chinese and English TV shows, we employ two Chinese college students to collect 79 TV shows and use our automatic pipeline to extract high-quality dialogues. To ensure equitable compensation, we pay them 50 yuan per hour (\$6.91 USD), which is considered fair locally. Due to copyright issues with TV shows, we



**Figure 3: The construction process of NCSSD. The upper panel shows the automatic pipeline of the collection subset, while the bottom panel shows the workflow of the recording subset.**

will only disclose the names of the TV shows, dialogue scripts, and acoustic features, excluding visual features in the dialogue. Certainly, other researchers can also input videos of other TV shows into our automatic pipeline to extract the dialogue segments.

For the recording subset (RC-EN and RC-ZH), we employ 15 female and 12 male young adults for on-site recording, with a compensation of 800 yuan (\$110.48 USD) per hour. We sign a public usage agreement with each recording personnel, planning to publicly release the transcribed scripts, acoustic features, and facial visual features of the recorded dialogues.

We believe that these high-quality multimodal dialogue resources advance research in conversational speech synthesis.

## B EXPERIMENT SETUP

In ConGPT module, the phoneme encoder utilizes existing method `g2p_en`<sup>1</sup> and `opencpop-strict` dictionary<sup>2</sup> for text discretization. The unit encoder includes a 1D convolutional layer with a stride of 2, and a vector quantization module with a dimension of 256 and a codebook size of 1024. ConGPT features 24 Transformer encoder blocks, each with a Feed Forward block dimension of 2048, 16 attention heads, and a dropout rate of 0.1. During training, the

number of dialogue turns,  $N$ , is set to 3. For the ConVITS module, the Content Encoder includes a text embedding layer with an output dimension of 192 and 6 transformer encoder blocks, also with an output dimension of 192; the Token Encoder consists of three transformer encoder blocks with an output dimension of 192. These blocks have a dropout rate of 0.1 and two attention heads. The Timbre Encoder consists of six 2D convolutional blocks with a kernel size of 3 and stride of 2, one GRU layer, and a linear layer with dimensions scaling from 128 to 512. Apart from the duration predictor, the rest of the configurations are the same as the standard VITS.

During the GPT-Talker training process, we used two NVIDIA A100 GPUs and two A800 GPUs, with a batch size set to 8. The NCSSD was divided into four subsets: CL-EN, CL-ZH, RC-EN, RC-ZH, with the number of dialogue groups in the training set, test set, and validation set divided in an 8:1:1 ratio, see the Appendix A for detailed division methods. It is worth noting that our GPT-Talker was trained with 3 turns ( $N$ ) of dialogue, whereas other baselines had 10 turns. To ensure the fairness of the experiments, during inference, we adjusted the dialogue turns of all baselines to 3 turns.

<sup>1</sup><https://pypi.org/project/g2p-en/>

<sup>2</sup><https://wenet.org.cn/opencpop/>

## C MORE RESULTS AND ANALYSIS

### C.1 Emotion Recognition Results

GPT-Talker is trained on a large emotional speech dataset and can generate natural, emotionally rich speech without explicit emotional labels. To demonstrate the emotional expression capabilities of GPT-Talker synthesized speech, we compare it with the state-of-the-art emotional speech synthesis model ECSS and use a pre-trained SER model<sup>3</sup> to predict the emotion categories of 400 audio samples generated by two models, under the DailyTalk and NCSSD (RC-EN) datasets. The correct emotional labels all come from the emotional category labels of the two datasets.

The confusion matrix depicted in the Figure 4 highlights the performance disparities between them. The pronounced diagonal in the matrix confirms that GPT-Talker surpasses ECSS, proving that the emotional conversational speech produced by GPT-Talker exhibits clear emotional expression.

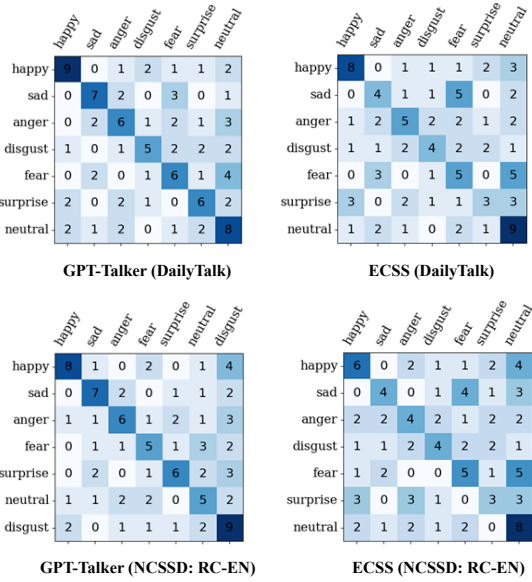


Figure 4: The confusion matrix results of the emotional types of synthetic speech generated by GPT-Talker and ECSS. The X-axis and Y-axis of subfigures represent perceived and true emotion categories.

### C.2 Visualization Analysis (Mel-spectrum and Pitch)

To better understand the quality of GPT-Talker's synthesized speech in terms of naturalness and expressiveness, we analyze the Mel spectrogram and the pitch of its output by comparing it with several baseline models. In example 1 (top of the Figure 5), the corresponding text with synthesized speech from the DailyTalk dataset is "oh, wait a minute. what about a lid for the pan?"; In example 2 (bottom of the Figure 5), the corresponding text with synthesized speech from the NCSSD (EN) dataset is "Great, together we can make a

<sup>3</sup><https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition>

Table 2: Experimental results about various dialogue turns.

Dialogue Turns (N)	DailyTalk			
	SSIM (↑)	DTWD (↓)	N-DMOS (↑)	E-DMOS (↑)
2	0.877	52.536	3.613+0.027	3.636+0.012
3	<b>0.902</b>	<b>43.014</b>	<b>3.901+0.014</b>	<b>3.944+0.023</b>
4	0.713	60.382	3.324+0.018	3.452+0.021
Dialogue Turns (N)	NCSSD (EN)			
	SSIM (↑)	DTWD (↓)	N-DMOS (↑)	E-DMOS (↑)
2	0.871	53.123	3.602+0.011	3.620+0.026
3	<b>0.896</b>	<b>42.872</b>	<b>3.912+0.025</b>	<b>3.934+0.016</b>
4	0.732	61.621	3.331+0.029	3.328+0.027
Dialogue Turns (N)	NCSSD (ZH)			
	SSIM (↑)	DTWD (↓)	N-DMOS (↑)	E-DMOS (↑)
2	0.883	51.764	3.636+0.022	3.712+0.025
3	<b>0.901</b>	<b>43.192</b>	<b>3.905+0.015</b>	<b>3.956+0.014</b>
4	0.728	60.726	3.418+0.024	3.392+0.035

difference and ensure a stronger future for a regime." For example 2, due to limited space for the synthesized speech which is relatively long, we only visualize the first 4 seconds.

Firstly, the green box in example 1 clearly shows that GPT-Talker exhibits more defined envelope details and superior synthesis quality, as also evidenced in example 2. Secondly, the white polyline in both examples objectively demonstrates that the speech pitch synthesized by GPT-Talker generates richer prosody. Additionally, the red box indicates that the pause frequencies in some of the speech synthesized by GPT-Talker are closer to those of real speech. In a nutshell, our GPT-Talker outperforms all baselines and performs significant advantages in the naturalness and expressiveness of synthesized conversational speech.

### C.3 Dialogue Turns Analysis

To further investigate the influence of dialogue turns (the length of dialogue context) on conversational speech synthesis, we adjust the number of dialogue turns in the input model during inference for comparative analysis. We use models trained on three datasets for validation: DailyTalk, NCSSD (EN) which contains the CL-EN and RC-EN, and NCSSD (ZH) which contains the CL-ZH and RC-ZH. Notably, our model is initially trained with  $N$  set at 3, which includes two utterances of dialogue history and one current utterance for synthesis. In this section, we synthesize 50 examples for each length of turns and ask 30 evaluators to conduct subjective evaluations. As evidenced in Table 2, when the dialogue history is reduced to just one sentence ( $N=2$ ), both objective and subjective experimental measures show a decline, suggesting that a more extensive dialogue history enriches the semantic and prosodic style information of the current sentence. However, increasing  $N$  to 4, beyond the training configuration of the model, allows for speech synthesis but with noticeably diminished naturalness and emotional expressiveness. This reduction is likely due to the model's constrained ability to handle excessively long sequences. This will be a key issue for our future research.

### C.4 Context Serialization Analysis

In the Con-GPT module of GPT-Talker, the input section utilizes various methods for concatenating dialogue contexts. The first method, which we call *ABAB - Format*, alternates splicing utterances:  $\{U_1^t$ ,

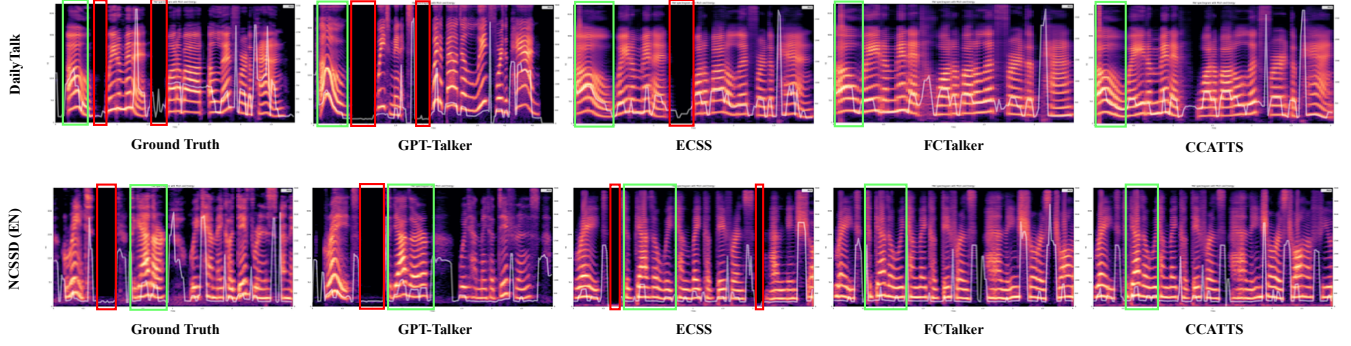


Figure 5: Visualization of Mel-spectrum and Pitch results. The green boxes highlight the spectrum details, while the red boxes highlight the prosody break.

Table 3: The results of different contextual serialization methods comparison.

Dataset	Method	SSIM (↑)	DTWD (↓)	N-DMOS (↑)	E-DMOS (↑)
NCSSD (EN)	ABAB-Format	<b>0.904</b>	<b>42.076</b>	<b>3.910 ± 0.019</b>	<b>3.922 ± 0.022</b>
	AABB-Format	0.873	44.924	3.812 ± 0.015	3.861 ± 0.023
NCSSD (ZH)	ABAB-Format	<b>0.908</b>	<b>43.002</b>	<b>3.906 ± 0.02</b>	<b>3.987 ± 0.021</b>
	AABB-Format	0.891	45.016	3.856 ± 0.033	3.877 ± 0.017

$U_1^a, U_2^t, U_2^a, \dots, U_N^t$ , is our chosen approach. The second method, called *AABB-Format*, employs sequential splicing:  $\{U_1^t, U_2^t, \dots, U_N^t, U_1^a, U_2^a, \dots, U_{N-1}^a\}$ . In this section, we use the NCSSD Chinese and English data to train the model. After that, we use each context connection method to synthesize 50 examples, respectively, and ask 30 evaluators to conduct subjective evaluation. According to the experimental findings presented in Table 3, the *ABAB-Format* method proves more effective for dialogue speech synthesis tasks. This is because this method more closely aligns with the sequence of information exchange in human-machine interactions. Additionally, this method’s integration of each utterance’s text and audio may help the model more accurately determine the placement of statements, thus fostering a complex interdependency at the sentence level within the dialogue context.

### C.5 Zero-shot Timbre Rendering Performance Analysis

Our model boasts a unique ability to perform zero-shot timbre rendering for the synthesized conversational speech for agent. To showcase this remarkable capability, we feed the model unseen speakers from external datasets, resulting in the successful synthesis of their speech.

Specifically, we randomly sample 20 sets of dialogue speech from 5 unseen speakers in IEMOCAP and M<sup>3</sup>ED, respectively, and replace one speaker in the original dialogue in DailyTalk and NCSSD (EN and ZH) to synthesize the speech of unseen speakers. We conduct experiments using the speaker similarity evaluation method in 5.3. As the results show in Table 4, the SSIM values are all above 0.82, which is lower than the speaker similarity values within the collection listed in Table 4 of main text, but it also shows that

Table 4: Results of zero-shot timbre rendering experiment.

Evaluation metric	Datasets		
	DailyTalk	NCSSD (EN)	NCSSD (ZH)
SSIM (↑)	0.838	0.834	0.822

GPT-Talker can basically synthesize the speech of unseen speakers. Of course, a single objective indicator cannot fully reflect the real situation, and we will also place samples on the demo page for readers to reference.

## D LIMITATIONS AND FUTURE DIRECTIONS

During dataset construction, the recorded and collected dialogues were accompanied by videos, incorporating visual modality information. We haven’t explored this aspect in our research yet, so fully utilizing visual modality data to enhance the synthesis of expressive speech will be our subsequent focus.