A Appendix and Supplementary Materials

A.1 Zero-shot TTS Model

Our internal zero-shot TTS model is an auto-regressive model, which is similar to BASE TTS [13]. We evaluate our TTS model with some objective metrics. We assess objective metrics including speaker similarity (SIM-O and SIM-R), and robustness (WER) in the following ways: 1) To evaluate speaker similarity, we use the WavLM-TDCNN [5] speaker embedding model. This model measures how closely generated samples match the original prompt (SIM-O) and the reconstructed prompt (SIM-R). 2) For measuring robustness, we calculate the Word Error Rate (WER) using a CTC-based HuBERT model[1] that was initially trained on Librilight and subsequently finetuned on the 960-hour training dataset from LibriSpeech. We compare our models with SOTA auto-regressive TTS models: VALL-E [20], and CLaM-TTS [12], VoiceCraft [16], XTTS-v2[2], and WhisperSpeech[3]. we adapt classifier-free guidance (cfg) [8, 19] for better generation. We use LibriSpeech test-clean for evaluation, which contains 40 distinct speakers. Following [20, 10], we randomly select one sentence for each speaker as the target and a 3-second clip as the prompt from the same speaker's speech.

	Training Data	Sim-O↑	Sim-R↑	WER↓
Ground Truth	-	0.68	-	0.34
VALL-E	LibriLight	-	0.58	5.9
CLaM-TTS	MLS	0.49	0.54	5.11
VoiceCraft	GigaSpeech	0.45	-	6.68
XTTS-v2	-	0.51	-	5.5
WhisperSpeech	LibriLight	0.48	-	4.78
Ours	LibriLight	0.58	0.61	5.56
Ours (w. cfg)	LibriLight	0.60	0.63	4.32
Ours (w. cfg, rerank 5)	LibriLight	0.63	0.66	2.01

A.2 Prompts for Generating Responses

A.2.1 Prompts for Training Set

```
[System]
Let's simulate a conversation between a simulated speaker and you, ChatGPT. I
need your help to finish the following three tasks:
1. Generate your reply which consists of two or more sentences to the
simulated speaker. Your reply should be able to reflect the provided
information.
2. Select an appropriate emotion (happy, sad, fear, angry, surprise, neutral,
disgust) for your reply, which should be conveyed through the language used.
3. Explain how your reply reflects the provided information of the simulated
speaker. Your reason should be less than 3 sentences.

Your output must strictly adhere to JSON format with three keys: "reply",
"reply_emotion" and "reason" corresponding to your answers for the three tasks.

[What the simulated speaker said]
{input_statement}

[Emotion of the simulated speaker]
{emotion}
```

Figure 1: The prompt used to generate responses of utterances for training set related to emotion.

[1] https://huggingface.co/facebook/hubert-large-ls960-ft

[2] https://huggingface.co/coqui/XTTS-v2

[3] https://github.com/collabora/WhisperSpeech

```
[System]
Let's simulate a conversation between a simulated speaker and you, ChatGPT. I
need your help to finish the following two tasks:
1. Generate a reply of two or more sentences to the simulated speaker. Ensure
that your reply mimics the provided information, including adopting a similar
accent style as the simulated speaker.
2. Explain how your reply reflects the provided information of the simulated
speaker. Your reason should be less than 3 sentences.

Your response must be formatted in JSON, with two keys: "reply" for the first
task and "reason" for the second task.

[What the simulated speaker said]
{input_statement}

[Accent of the simulated speaker]
{accent}
```

Figure 2: The prompt used to generate responses of utterances for training related to accent.

```
[System]
Simulate a conversation between a child and ChatGPT. Complete the following
tasks:
1. Generate a reply consisting of at least two sentences, tailored to the
child's age provided in the input.
2. Briefly explain (in less than three sentences) how your reply considers the
child's age.

Your output must strictly adhere to JSON format with three keys: "reply",
"reply_emotion" and "reason" corresponding to your answers for the three tasks.

[What the simulated child said]
{input_statement}

[Child's Age]
{age}
```

Figure 3: The prompt used to generate responses of utterances for training related to age.

```
[System]
You are tasked with simulating a conversation between a simulated speaker and
yourself, ChatGPT. Could you give responses based on a text with a certain
environmental sound? For the scenario, provide a brief description of the
background sound and **five** suitable model response that aligns with the
context.

Your response must be formatted in JSON, each entry should contain the
following keys: "reply", "reason".

[What the simulated speaker said]
{input_statement}

[Background Sound]
{background_sound}
```

Figure 4: The prompt used to generate responses of utterances for training related to background sound.

A.2.2 Prompts for SD-Eval

```
[System]
Let's simulate a conversation between a hypothetical speaker and ChatGPT. I
need you to:
1. Create **five** diverse responses, each consisting of two or more sentences,
in reaction to the speaker's statement. Each response should appropriately
reflect the context and content provided by the speaker.
2. Assign an emotion selected from **joy, sadness, fear, anger, surprise,
neutral and disgust** to each response, with the language of the reply
demonstrating this emotion.

Format each of your responses with XML tags, such as <reply>reply</reply> and
<reply_emotion>reply emotion</reply_emotion>, which are corresponding to the
tasks above.

[Simulated Speaker's Statement]
{input_statement}

[Speaker's Emotion]
{emotion}
```

Figure 5: The prompt used to generate responses of utterances for *test-emo*.

```
[System]
Simulate a conversation between a hypothetical speaker and ChatGPT. Produce
**five** varied responses, each comprising at least two sentences. Ensure that
your reply mimics the provided information, including adopting a similar
accent style as the simulated speaker. Your output should be **strictly**
formatted for each response using XML tags, <reply> and </reply>.

[Simulated Speaker's Statement]
{input_statement}

[Speaker's Accent]
{accent}
```

Figure 6: The prompt used to generate responses of utterances for *test-acc*.

```
[System]
Simulate a conversation between a child and ChatGPT. Produce **five** varied
responses, each comprising at least two sentences, suited to the child's age
specified in the input. Your output should be **strictly** formatted for each
response using XML tags, <reply> and </reply>.

[Simulated Child's Statement]
{input_statement}

[Child's Age]
{age}
```

Figure 7: The prompt used to generate responses of utterances for *test-age*.

```
[System]
You are tasked with simulating a conversation between a simulated speaker and
yourself, ChatGPT. Could you give responses based on a text with a certain
environmental sound? For the scenario, provide a brief description of the
background sound and **five** suitable model response that aligns with the
context.

Your response must be formatted in JSON, each entry should contain the
following keys: "reply", "reason".

[What the simulated speaker said]
{input_statement}

[Background Sound]
{background_sound}
```

Figure 8: The prompt used to generate responses of utterances for *test-env*.

A.3 Prompts for LLM Evaluation

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and emotion displayed
below. Your evaluation should consider whether it contains an appropriate
sentiment with respect to the user's emotion. Please also consider factors
such as the naturalness, coherence, engagingness and groundedness of the
response. Please make sure you read and understand these instructions
carefully. Please be as objective as possible. Begin your evaluation by
providing a short explanation. After providing your explanation, please rate
the response on a scale of 1 to 10 by **strictly** following this format:
"Rating: [[rating]]", for example: "Rating: [[5]]".

[User's Statement]
{statement}

[User's Emotion]
{info}

[AI Assistant's Response]
{response}
```

Figure 9: The prompt for evaluating *test-emo* using LLM.

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and accent displayed below.
Your evaluation should consider whether the AI assistant recognizes the user's
accent correctly so that the response contains appropriate slang with respect
to the user's accent. Please also consider factors such as the naturalness,
coherence, engagingness and groundedness of the response. Please make sure you
read and understand these instructions carefully. Please be as objective as
possible. Begin your evaluation by providing a short explanation. After
providing your explanation, please rate the response on a scale of 1 to 10 by
**strictly** following this format: "Rating: [[rating]]", for example: "Rating:
[[5]]".

[User's Statement]
{statement}

[User's Accent]
{info}

[AI Assistant's Response]
{response}
```

Figure 10: The prompt for evaluating *test-acc* using LLM.

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and age displayed below.
Your evaluation should consider whether it contains an appropriate tone of
voice with respect to the user's age. Please also consider factors such as the
naturalness, coherence, engagingness and groundedness of the response. Please
make sure you read and understand these instructions carefully. Please be as
objective as possible. Begin your evaluation by providing a short explanation.
After providing your explanation, please rate the response on a scale of 1 to
10 by **strictly** following this format: "Rating: [[rating]]", for example:
"Rating: [[5]]".

[User's Statement]
{statement}

[User's Age]
{info}

[AI Assistant's Response]
{response}
```

Figure 11: The prompt for evaluating *test-age* using LLM.

```
[System]
Please act as an impartial judge and evaluate the quality of the response
provided by an AI assistant to the user's statement and background sound
displayed below. Your evaluation should consider whether it considers the
user's background sound and generates an appropriate response. Please also
consider factors such as the naturalness, coherence, engagingness and
groundedness of the response. Please make sure you read and understand these
instructions carefully. Please be as objective as possible. Begin your
evaluation by providing a short explanation. After providing your explanation,
please rate the response on a scale of 1 to 10 by **strictly** following this
format: "Rating: [[rating]]", for example: "Rating: [[5]]".

[User's Statement]
{statement}

[User's Background Sound]
{info}

[AI Assistant's Response]
{response}
```

Figure 12: The prompt for evaluating *test-env* using LLM.

A.4 Prompts for Generating Dialogue Data of *test-env*

```
[System]
You are tasked with simulating a conversation between a simulated
speaker and yourself, ChatGPT. Could you provide examples of how a
speech dialogue model would adapt its responses based on different
environmental sounds? Please include scenarios where the model detects
distinctive background noises (such as in a busy marketplace, at a beach,
children's voice, or during a thunderstorm). For each scenario, provide
a brief description of the background sound and a suitable model
response that aligns with the context. Let's first focus on a special
background noise type: {background_sound}. Your response must be
formatted in JSON, each entry should contain the four keys: "text",
"background", "reply", and "reason", "reason" should focus on why the
reply may change due to the presence of background sound. I will give
you some examples:

{text: "Can you tell me about the effects of adult beverages on health?",
background: "children's voice", reply: ["It sounds like there might be
little ears around! If you're asking for general health information,
it's important to maintain a balanced lifestyle, which includes
moderation in all things."], reason: "The large model can answer the
question about the effects of adult beverages, but because there are
children's voices in the background, it is not suitable for direct
answers, but for discussing more general health topics."}

{text: "What should I wear today?", background: "raining or thundering",
reply: ["Given the rainy and windy weather outside, I recommend wearing
a waterproof jacket and sturdy, water-resistant shoes. Would you like a
reminder to bring an umbrella?"], reason: "Since there are sounds of
rain and thunder in the background, it is recommended to bring an
umbrella."}

{text: "Where's a good place to eat around here?", background: "driving
or traffic", reply: ["Since you're on the road, I'll find something easy
to access. There's a highly-rated grill just off the next exit. Want me
to guide you there?"], reason: "..."}

OK, now it's your turn! Please provide examples of how a speech dialogue
model would adapt its responses based on different environmental sounds,
let's firstly focus on a special background types: {background_sound}.
In my examples, each "reply" contains only one response. You should
provide five proper responses for each example. Please generate five
examples. Each example contains five responses, and for each example,
only give one reason for general.
```

Figure 13: The prompt for generating dialogue data of *test-env*.

A.5 Potential Negative Societal Impacts

SD-Eval may have potential negative societal impacts for *test-acc* subset, as it currently includes only accents from the United Kingdom, Canada, the United States, Australia, and New Zealand. In the future, we plan to enhance the diversity of the test set by incorporating accents from a broader range of countries.

A.6 License of Public Dataset

Table 1 lists the licenses for all publicly available datasets used to construct SD-Eval. We strictly adhere to the licenses of these datasets. Specifically, we will not directly provide downloads of these datasets or bundle them with our data. Instead, we provide links to the original websites of the datasets. We promise the data we are using/curating does not contain personally identifiable information or offensive content.

Table 1: Licenses of public datasets used to construct SD-Eval.

Dataset Name	License
VCTK [22]	CC-BY-4.0
Common Voice v16.1 [2]	CC-BY-SA-3.0-Unported
RAVDESS [14]	CC-BY-NC-SA
JL Corpus [9]	CC0-public
AudioCaps [11]	MIT
LibriSpeech [15]	CC-BY-4.0
MEAD [21]	Link
MyST [18]	Link

A.7 Statistics of Training Set

Table 2 shows the statistics of training set. For training data related to the environment, we generate one response for each sentence, except for data related to the environment, which has five different responses for each sentence to serve the purpose of data augmentation.

Table 2: Statistics of training set. ChatGPT Version refers to the specific version of ChatGPT used to generate the data.

Type	**# Hours**	**# Utts**	**Constructed From**	**Labels**	**ChatGPT Version**
Emotion	120.60	100.5k	MSP-Podcast [14], IEMOCAP [3], MELD [17], EmoV-DB [1], ESD [23], CREMA-D [4]	Angry, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise,Frustrated Excited, Amused, Sleepiness	GPT-3.5-Turbo
Accent	759.75	508.6k	VCTK [22], UK-Ireland dataset [7]	England, Scottish, Northern Irish, Welsh, Irish, American, Canadian, Australian, Nea Zealand	GPT-4o
Environment	32.06	47.1k	LibriSpeech[15], AudioCaps [11]	Driving, Children's Voice, Sea Beach, Raining or Thundering, Bells Sports Center, Shopping Center, Bus or Subway	GPT-4-Turbo
Age	140.31	73.2k	MyST [18]	Child	GPT-3.5-Turbo
Summary	1,052.72	729.4k	-	-	-

A.8 Datasheets for SD-Eval

A.8.1 Motivation

- **For what purpose was the dataset created?** The dataset is designed as a novel benchmark dataset for multidimensional evaluation of spoken dialogue understanding beyond words. The dataset is to promote the development of more empathetic and intelligent spoken dialogue systems that can generate appropriate responses based on paralinguistic and environmental information.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** N/A. Our submission is double-blinded.
- **Who funded the creation of the dataset?** N/A. Our submission is double-blinded.
- **Any other comments?** None.

A.8.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Our data is comprised of documented forms that contain relevant information.
- **How many instances are there in total (of each type, if appropriate)?** Totally 7,303 instances. See Table 1 in paper.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** We construct the dataset using public datasets from 4 aspects and will add more in the future.
- **What data does each instance consist of?** Each instance includes utt_id, transcript, response, and info (indicating emotion, accent, age, or background sound), the name of the original dataset, and the corresponding audio file (which may need to download from original dataset).
- **Is there a label or target associated with each instance?** Yes. The reference responses generated by ChatGPT are provided.
- **Is any information missing from individual instances?** No.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** No.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes. See Table 1 in paper.
- **Are there any errors, sources of noise, or redundancies in the dataset?** No.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.** See Appendix A.6. We can not ensure the public datasets exist forever and consistent but they should exist for a long time as some of them are published for many years.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** No.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** Yes for age. *test-age* contains data from MyST [18] which is children speech data. Under *test-age*, the ratio of utterances for child and adult is 1:1.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** No.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *test-acc* is related to accent, which are constructed from VCTK [22] and Common Voice [2].
- **Any other comments?** None.

A.8.3 Collection Process

- **How was the data associated with each instance acquired?** See Section 3 of paper.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** See Section 3 of paper.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** See Section 3 of paper.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** No data collection process since the dataset is built by public datasets.
- **Over what timeframe was the data collected?** N/A.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** N/A. The dataset is built by public datasets.

A.8.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes. See Section 3 of paper.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** No. The dataset is built by public datasets.
- **Is the software that was used to preprocess/clean/label the data available?** Part of them is available. And we provide all details in section 3 of the paper.
- **Any other comments?** None.

A.8.5 Uses

- **Has the dataset been used for any tasks already?** Yes.
- **Is there a repository that links to any or all papers or systems that use the dataset?** No.
- **What (other) tasks could the dataset be used for?** Spoken dialogue evaluation for four aspects.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** No.
- **Are there tasks for which the dataset should not be used?** No.
- **Any other comments?** None.

A.8.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** Yes.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** It will be distributed by Github and Huggingface.
- **When will the dataset be distributed?** When we are ready, or when the paper is accepted.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** Yes. The license is CC-BY-NC 4.0.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances** Yes, please check Appendix A.6.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.
- **Any other comments?** None.

A.8.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?** The authors will be supporting/hosting/maintaining the dataset.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The authors can be contacted by email. And issues are also accepted on Github/Huggingface.
- **Is there an erratum?** N/A.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes. We will continue to maintain the dataset after releasing.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** N/A.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** The decision depends on the reason for the update. We will specify the reason and indicate whether the original version will be retained.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Yes, we allow others to contribute to our dataset. We will specify the way for contribution after releasing dataset.
- **Any other comments?** None.

A.9 URL to Website / Platform Where the Dataset/Benchmark can be Viewed and Downloaded by the Reviewers

We upload four json files called *test-emo*, *test-acc*, *test-age*, and *test-env* under folder json_data, which contains all data except the audio. We will not directly provide downloads of these datasets or bundle them with our data. Instead, we provide links to the original websites of the datasets as shown in Table 1. In addition, we offer our synthesized data, including portions of the *test-age* and *test-env* datasets. The link to download the synthesised data is `https://drive.google.com/file/d/1ILnV2fS7kRoPofmG2GZcD5Y2qM-tjqU7/view?usp=sharing`. For each utterance, it contains the following keys:

- utt_id: Utterance id.
- wav_path: Wav file path to find the audio inside the dataset.
- transcript: Ground truth transcript for each audio.
- info: What emotion, accent, age or background sound for this utterance.
- dataset_name: Which dataset this utterance is from.
- target: The responses generated by ChatGPT.

A.10 The Machine Learning Reproducibility Checklist

For all models and algorithms presented, check if you include:

- **A clear description of the mathematical setting, algorithm, and/or model.** ✓
- **An analysis of the complexity (time, space, sample size) of any algorithm.** N/A
- **A link to a downloadable source code, with specification of all dependencies, including external libraries.** × We upload source code in supplementary materials, which is under the folder "xtuner-speech-main". You can try to reproduce and check the hyper-parameters by using the config files under "xtuner-speech-main/xtuner/configs/llama_speech".

For any theoretical claim, check if you include.

- **A statement of the result.** ✓

- **A clear explanation of any assumptions.** ✓
- **A complete proof of the claim.** N/A.

For all figures and tables that present empirical results, check if you include:

- **A complete description of the data collection process, including sample size.** ✓
- **A link to a downloadable version of the dataset or simulation environment.** × We provide dataset in supplementary materials.
- **An explanation of any data that were excluded, description of any pre-processing step.** ✓
- **An explanation of how samples were allocated for training/validation /testing.** ✓
- **The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.** × We do not provide the method to select the best hyper-parameter configuration but we provide almost all details to conduct experiments and most of the hyper-parameters are default which are used in the Xtuner [6].
- **The exact number of evaluation runs.** ✓
- **A description of how experiments were run.** ✓
- **A clear definition of the specific measure or statistics used to report results.** ✓
- **Clearly defined error bars.** ×
- **A description of results with central tendency (e.g. mean) & yariation (e.g. stddev)** ×
- **A description of the computing infrastructure used.** ✓

A.11 Author Statement

We bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

A.12 Hosting, Licensing, and Maintenance Plan

The license of this dataset will be CC-BY-NC 4.0. We will host this dataset on Github and Huggingface. The dataset will be distributed when we are ready, or when the paper is accepted. We will ensure continuous availability and easy access for the research community. We will continuously update the dataset to correct potential errors or to add more data. We welcome feedback and inquiries from the research community. We ensure the dataset will be available for a long time with a clear maintenance plan.

A.13 Structured Metadata to a Dataset's Meta-Data Page Using Web Standards

Please check "metadata.json" in supplementary material.

References

[1] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[4] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

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

[6] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, 2023.

[7] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6532–6541, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.804`.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[9] Jesin James, Li Tian, and Catherine Inez Watson. An Open Source Emotional Speech Corpus for Human Robot Interaction Applications. In *Proc. Interspeech 2018*, pages 2768–2772, 2018. doi: 10.21437/Interspeech.2018-1349.

[10] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

[11] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. 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 119–132, 2019.

[12] Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. *arXiv preprint arXiv:2404.02781*, 2024.

[13] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*, 2024.

[14] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4): 471–483, 2019. doi: 10.1109/TAFFC.2017.2736999.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[16] Puyuan Peng, Po-Yao Huang, Daniel Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024.

[17] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL `https://aclanthology.org/P19-1050`.

[18] Sameer Pradhan, Ronald A. Cole, and Wayne H. Ward. My science tutor (MyST)–a large corpus of children's conversational speech. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1052.

[19] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

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

[21] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

[22] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2019.

[23] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2021.11.006. URL https://www.sciencedirect.com/science/article/pii/S0167639321001308.