

Supplementary Materials of Paper: SpeechCraft: A Fine-grained Expressive Speech Dataset with Natural Language Description

Anonymous Authors

A DEMO WEBPAGE

In addition to the illustration in the main paper, a larger quantity of detailed dataset samples and experimental results are demonstrated on the web page: <https://speechcraft2024.github.io/speechcraft2024/> in the form of speech-language pairs. As an alternative, we pack the page repository (check *Page/index.html*) for local browsers in the supplementary materials. We highly recommend that the reviewers take a listen. Please open the demo webpage in *Chrome* for an enhanced experience.

B DETAILS IN SPEECH ANNOTATION

B.1 Example of Metadata

In Section 3.2 of the paper, we conduct data preprocessing to establish standard metadata before the whole annotation framework and further utilize the raw metadata to extract prior information of the audio clip such as the topic. An example of preprocessed metadata is shown as follows.

```
{
  'path': './giga_00000616/
    ↳ POD0000013293_S0000160.wav',
  'subdatasets': 'POD',
  'sampling_rate': 16000,
  'title': 'Law_Report_A_decade_since_911_
    ↳ _54_new_anti-terror_laws',
  'url': 'https://abcmedia.akamaized.net/rn/
    ↳ podcast/2011/09/lrt_20110906_0845.
    ↳ mp3',
  'sid': 'POD0000013293_S0000160',
  'speaker': 'N/A',
  'begin_time': 0,
  'end_time': 10.220000000000027,
  'text_raw': "And_that_can_actually_fuel_
    ↳ the_possibilities_of_extremism_and_
    ↳ recruitment_by_terrorists_and_
    ↳ that's_why_the_laws_need_to_be_
    ↳ balanced_out_with_other_programs_
    ↳ which_enhance_social_cohesion_",
  'category': 'News_and_Politics'
}
```

B.2 Label Categories

As described in Section 3.3, our proposed annotation system characterized speech in terms of various style properties including pitch, energy, speed, age, gender, emotion description, and word emphasis. The subset labels of each attribute are listed below.

Gender: Male, Female

Age: Child, Teenager, Youth adult, Middle-aged, Elderly

Pitch: low, normal, high

Speed: slow, normal, fast

Volume: low, normal, high

Emotion (English): Fearful, Happy, Disgusted, Sad, Surprised, Angry, Neutral

C SPEECHCRAFT DATA SOURCES

As illustrated in Section 4.1, we implemented the annotation system across large-scale bilingual speech datasets to conduct speech descriptions, resulting in the SpeechCraft dataset. The detailed information of the four open-source speech datasets is introduced as follows.

AISHELL-3 [8] is a high-fidelity Mandarin speech corpus containing roughly 85 hours of recordings spoken by 218 speakers and a total of 88,035 utterances.

Zhvoice¹ corpus consists of eight open-source subdatasets, processed through noise reduction and quality enhancement, with approximately 3200 speakers, and 900 hours of audio, totally 1,032,940 utterances.

LibriTTS-R [4] is a restored English TTS Corpus with 585 hours of speech from 2,456 speakers. It is derived by applying speech restoration to the LibriTTS [9] corpus with sound quality improved.

GigaSpeech [1] is an evolving, multi-domain ASR Corpus collected from three different sources of data, including audiobook, podcast, and YouTube videos. It provides a wide range of choices on the dataset scale. The GigaSpeech-m is an officially recommended subset as a 1000-hour dataset for research experiments.

D DETAILS IN EMPHASIS STUDIES

D.1 Emphasis Regeneration

As introduced in Section 4.2, we employed FastSpeech 2 [7] as the backbone to regenerate the AISHELL-3 and LibriTTS-R datasets with keyword emphasized in each piece.

D.1.1 FastSpeech 2 Backbone Model. FastSpeech 2 achieved modulation of phoneme-level characteristics with the key component of Variance Adaptor. It consists of three primary Predictors for energy, pitch and duration respectively. Adjusting the output of Predictors by scale, we can obtain loud volume, high pitch, and elongated sounds on the designed phoneme. The different combination of scaling factors determines various acoustic effects.

D.1.2 Keyword Extraction. We used the TextRank [5] algorithm for Chinese content and Gensim [6] for English to conduct keyword extraction. Words such as particles and proper nouns are overlooked as they are seldom stressed in conversational speech.

¹<https://github.com/fighting41love/zhvoice>

D.2 Emphasis Detection

We introduced a word-level emphasis detection model in Section 3.3. The detection model works by conducting forced alignment for each waveform to get the separate audio slice for a minimum word unit. As the emphasis is a relative concept that becomes significant only when compared to the surrounding words, we concatenate features of predecessor and successor to form the final representation for each audio unit in the neural network training.

D.3 Emphasis Evaluation on Real-Life Dataset

In Section 4.3, we demonstrated the accuracy of emphasis detection on the testset of the regenerated AISHELL-3-stressed and LibriTTS-R-stressed data. To further evaluate the effectiveness of our detection approach in modeling real-life stress patterns, we utilized an internal dataset with human annotation on word emphasis over 10,000 audio utterances read by professional voice actors to test the model's generalization ability on real-world data. The emphasis detection models achieve 66.90% on word-level accuracy and 41.63% on sentence-level accuracy on the human-annotated dataset, which showcase the model's promising ability to generalize on natural word emphasis. However, the limited accuracy may stem from the inherent complexity of real-world emphasizing effects, which are more nuanced and varied than the mixed feature adjustments in our data construction.

E MODEL CONFIGURATION IN TTS EXPERIMENTS

We adopted the Salle [3] model to conduct Expressive Speech Synthesis in Section 5.1, which facilitate speech synthesis task by codec language modelling with RVQ [2]. RVQ is a method of high fidelity neural audio compression which features multi-layer discrete quantizers that can be reconstructed to high-quality waveforms by the pre-trained neural audio codec model.

Salle employed a hybrid approach combining an autoregressive style conditional codec model (AR) and a non-autoregressive TTS codec model (NAR). The AR model is employed to generate the first layer of RVQ, which encapsulates the fundamental speaking information. Conversely, the NAR model is reserved for the subsequent layers of quantizers that capture fine acoustic details. Text style prompt embedding is concatenated ahead of the phoneme text embedding, serving as the condition. The strategic embedding integration significantly enhances the expressive capacity and accuracy of the speech synthesis.

REFERENCES

- [1] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909* (2021).
- [2] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High Fidelity Neural Audio Compression. <http://arxiv.org/abs/2210.13438> arXiv:2210.13438 [cs, eess, stat].
- [3] Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. TextrolSpeech: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 10301–10305. <https://doi.org/10.1109/ICASSP48485.2024.10445879>
- [4] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*

- (2023).
- [5] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.
- [6] Radim Řehřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. (2010).
- [7] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* (2020).
- [8] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567* (2020).
- [9] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882* (2019).