ComedicSpeech: Text To Speech For Stand-up Comedies in Low-Resource Scenarios

Yuyue Wang[1,*]*, Huan Xiao*[2,*] *, Yihan Wu*[1]*, Ruihua Song*[1,#]

[1] Gaoling School of Artificial Intelligence, Renmin University of China, China
[2] School of Statistics, Renmin University of China, China

{wangyuyue123, xiaohuan001, yihanwu}@ruc.edu.cn, songruihua_bloon@outlook.com

Abstract

Text to Speech (TTS) models can generate natural and high-quality speech, but it is not expressive enough when synthesizing speech with dramatic expressiveness, such as stand-up comedies. Considering comedians have diverse personal speech styles, including personal prosody, rhythm, and fillers, it requires real-world datasets and strong speech style modeling capabilities, which brings challenges. In this paper, we construct a new dataset and develop ComedicSpeech, a TTS system tailored for the stand-up comedy synthesis in low-resource scenarios. First, we extract prosody representation by the prosody encoder and condition it to the TTS model in a flexible way. Second, we enhance the personal rhythm modeling by a conditional duration predictor. Third, we model the personal fillers by introducing comedian-related special tokens. Experiments show that ComedicSpeech achieves better expressiveness than baselines with only ten-minute training data for each comedian.

Index Terms: Text to Speech, stand-up comedy, spontaneous-style speech

1. Introduction

Many TTS models [1, 2, 3, 4] can synthesize natural and high-quality speech. However, they cannot perform well when synthesizing speech which requires dramatic expressive force, such as stand-up comedies. The stand-up comedy performance has a strong personal speech style, and its three characteristics bring great challenges to this task:

- **Personal Prosody.** Stand-up comedians (also called "comedians" for short in our paper) always use various emotions, high or low intonations, and stress to enhance the expressive and humorous in their own styles, which demands stronger prosody modeling capability of the TTS system.

- **Personal Rhythm.** A personal rhythm related to content including variable pauses and speaking rates is always used by comedians to create humorous effects. These rhythmic variations show the comedian's personal performance style, requiring accurate phoneme duration predicting ability of the TTS system.

- **Personal Fillers** [1] Comedians often use personal fillers (abbreviated as PF) that do not affect the meaning of a sentence, e.g. "well", "you know", "I mean", but enhance the expressiveness of performance. In stand-up comedy scenario, PFs are highly speaker-dependent. Some of them are expressed

---

* The first two authors contribute equally to this work.
# Corresponding author: songruihua_bloon@outlook.com
[1]In this paper, personal filler refers to a mantra, a habitual suffix, or a pause that occurs as personal language habits.

Table 1: *An example of special token ($\langle spc1 \rangle$) replacement. The Personal Filler (PF) which means "you know" is marked in red.*

raw text	我真正开始有一点自信以后，是我开始说脱口秀以后，我觉得我长得挺好的，你知道吧
raw phoneme	uo3 zh e1 n zh e4 ng k ai1 sh ii3 iou3 i4 d ia3 n z ii4 x i4 n i3 h ou4 sh ii4 uo3 k ai1 sh ii3 sh uo1 t uo1 k ou3 x iou4 i3 h ou4 uo3 j ve2 d e5 uo3 zh a3 ng d e2 t i3 ng h ao3 d e5 n i3 zh ii1 d ao4 b a5
phoneme after replacement	uo3 zh e1 n zh e4 ng k ai1 sh ii3 iou3 i4 d ia3 n z ii4 x i4 n i3 h ou4 sh ii4 uo3 k ai1 sh ii3 sh uo1 t uo1 k ou3 x iou4 i3 h ou4 uo3 j ve2 d e5 uo3 zh a3 ng d e2 t i3 ng h ao3 d e5 $\langle spc1 \rangle$

as vague pronunciations with much faster speed rather than word by word, which makes it difficult to synthesize for TTS.

Most previous works focus on improving the personal speech style by conditioning the source TTS model with the extracted personal representation [5, 6, 7, 8, 9, 10]. They introduce additional encoders, such as reference speech encoder, auxiliary text encoder, etc., to extract style from speech or predict style according to context. However, the personal rhythm, i.e. speaker-dependent duration, is not sufficiently emphasized, but it is an important factor for personal speech style, especially in stand-up comedy scenario. Other works [11, 12, 13, 14] train models with fillers to insert fillers (e.g. "uh", "um") in the text or speech. However, the fillers they choose are fixed and common for all speakers, without taking personal fillers and speaker-dependent pronunciations of fillers into account.

Based on the above analysis, in this paper, we develop ComedicSpeech, a TTS system specially designed for stand-up comedy synthesis. We improve the personal speech style in three aspects. First, we model diverse personal prosody with a well-designed prosody encoder and employ conditional layer normalization to integrate personal prosody into the TTS model. Second, we enhance personal rhythm by learning speaker-dependent phoneme duration distribution with the conditional duration predictor. Third, we apply special tokens to represent personal fillers, imitating personal speaking habits in stand-up comedies. Due to the lack of real-world stand-up comedy datasets, we build a multi-speaker stand-up comedy dataset. Although each speaker has about ten minutes for training, it is a good start for the stand-up comedy synthesis task. Experiment results also indicate that in such a low-resource scenario ComedicSpeech achieves better voice quality, voice similarity, and expressiveness in terms of MOS (Mean Opinion Score), SMOS (Similarity MOS), and Style MOS than baselines. The audio samples are available at `https://xh621.github.io/stand-up-comedy-demo/`.

Table 2: *The statistics of our stand-up comedy speech dataset*

Comedian	Gender	Clips	Total Duration(s)	Words	Words per Second	Average Duration of Pause(ms)	Average Energy	Average Pitch	Phoneme of PF
A	male	120	505	3,357	6.65	206.1	0.24	0.19	"er2"
B	female	140	622	3,824	6.15	188.3	0.11	0.43	"n i3 zh ii1 d ao4 b a5"
C	male	120	626	2,116	3.38	575.5	-0.18	-1.04	None
D	female	120	587	2,891	4.93	142.3	-0.11	0.12	None

2. Real-world Stand-up Comedy Dataset

Some previous works build spontaneous datasets, but they mainly focus on conversations [11, 15, 16]. In this paper, we construct a new real-world stand-up comedy dataset to support research on this new synthesis task by the following steps:

1. Raw Data Acquisition. We collect data from a Mandarin talk show called "Rock & Roast" and select four stand-up comedians (two females and two males) with diverse personal styles as training speakers. We collect about 60-minute speeches for each as raw data.

2. Data Denoising. Stand-up comedy speeches are mixed with noise, e.g., audience laughter and applause, which must be removed in order to train a high quality TTS model. We use Demucs [17, 18], a state-of-the-art music source separation tool to separate the human voice from the background music and other noise in our raw data.

3. Dataset Segmentation. We use an audio processing tool called Audacity [2] to cut the audio. The data is sliced into segments of about 3~8 seconds.

4. Special Token Replacement. To better model the PFs of different comedians, we replace PFs in the data with special tokens and add them into the phoneme set. Table 1 shows an example of replacing "you know" by "⟨spc1⟩". Please note that when two comedians have the same PF such as "you know", we still use two different special tokens to replace them because they are pronounced in different styles.

The analyses and statistics results are shown in Table 2. The statistics support our assumption that comedians have different prosody, rhythm, and fillers, which are part of their personal speech styles. First, the speech speed varies greatly. Among our four comedians, the first two are as fast as 6 words per second, while the other two are slower by 1-2 words per second. Second, the "Average duration of pause" varies widely in comedians. Comedian C takes 575 ms as a pause, which is about two times longer than the other three comedians. The above two points show that the comedians vary widely in rhythm. The "Average Energy" and "Average Pitch" indicate that the comedians vary widely in prosody. In addition, the four comedians use different PFs, as Table 2 shows. Simply inserting fixed fillers like "um" and "uh" as in the previous work is not enough for our scenario. In summary, stand-up comedians have dramatic differences in personal prosody, rhythm, and fillers.

3. Proposed Method

As shown in Figure 1, the overall architecture of Comedic-Speech is based on FastSpeech 2. We will describe our ideas and technique details in this section.

---
[2]https://www.audacityteam.org/

3.1. Modeling Personal Prosody

Since comedians always show diverse personal prosody in performance, modeling and integrating their personal prosody is very important. Inspired by GST-Tacotron [5], we apply a prosody encoder with GST-like architecture. The prosody encoder in ComidicSpeech extracts prosody representation through a weighted sampling of embedding from the shared prosody space. First, we map reference speech into a vector R, and then take it as the query vector used in Multi-head attention module. Define P with dimension d to be the basis prosody vector in the space shared by all speakers, then we obtain the prosody representation E:

$$Q = RW^Q, K = PW^Q, V = PW^V,$$
$$E = softmax(\frac{QK^T}{\sqrt{d}})V, \tag{1}$$

where W^Q, W^K, W^V are three trainable matrices, and P is initialized randomly. Then, following AdaSpeech 4 [10], we integrate the personal prosody information into the TTS model by applying the conditional layer normalization (CLN) in encoder and decoder (as shown in Figure 1(a)). Specifically, the extracted prosody representation E is used as the input of CLN, determining the scale and bias vector of FFT Block. Consequently, the prosody representation is flexible to change corresponding to different input audio.

3.2. Modeling Personal Rhythm

The rhythm, which can be affected by phoneme duration, is one of the most important attributes of personal style, especially in stand-up comedy scenarios. To better model the personal rhythm and learn speaker-dependent duration distribution, ComedicSpeech has two improvements:

- As shown in Figure 1(b), we build a speaker-dependent duration predictor, which introduces more personal rhythm information to TTS model. Specifically, we use the conditional layer normalization in the duration predictor, generating adaptive scale vector γ and bias vector β in the duration predictor as follows:

$$\gamma = W^\gamma E, \ \beta = W^\beta E, \tag{2}$$

where W^γ and W^β are linear layers for scale and bias respectively.

- We increase penalties for inaccurate duration predictions by directly calculating MSE loss between the target duration and the predicted duration (See Figure 1(b)), instead of calculating loss after converting them into logarithmic domain, like FastSpeech and AdaSpeech series. The duration loss in ComedicSpeech can be calculated as follows:

$$\mathcal{L}_{duration} = \frac{1}{N}\sum_{i=1}^{N}(d_{target} - d_{predicted})^2, \tag{3}$$

(a) Architecture of ComedicSpeech   **(b) Conditional Duration Predictor**
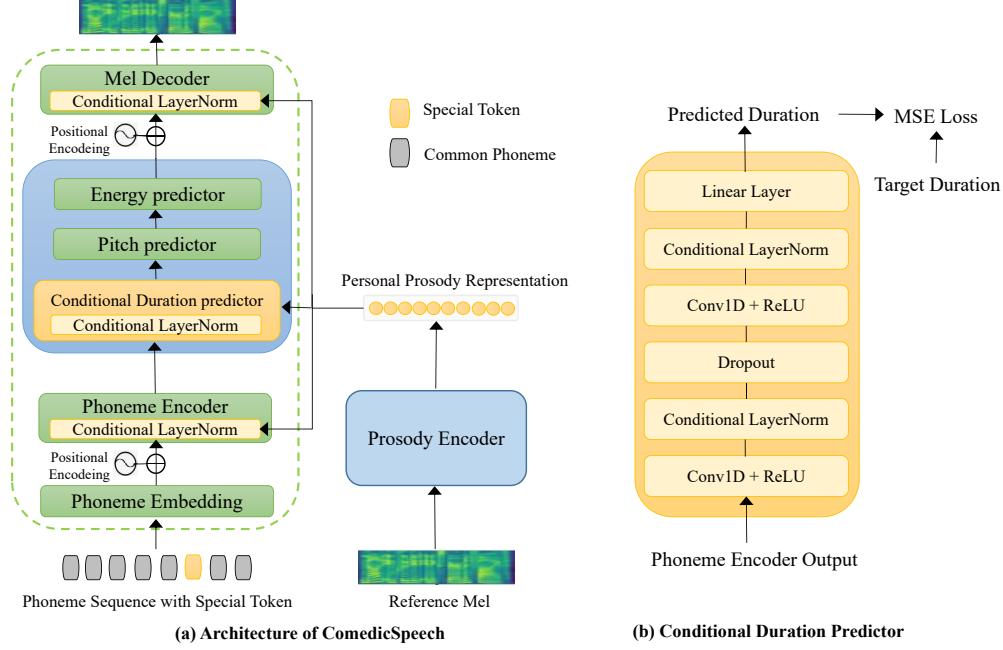
Figure 1: *Model structure of ComdicSpeech, a multi-speaker TTS model based on the backbone of FastSpeech 2.*

where d_{target} and $d_{prediction}$ represent target duration and predicted duration respectively.

3.3. Modeling Personal Fillers

Special pronunciation of PF conveys a strong personal style. The PF varies from person to person, and even the same PF's pronunciation is different across people. Take comedian B for example, her PF "you know" usually has short duration less than 18 frames and ambiguous pronunciation caused by linking; whereas, when spoken by comedian A, C or D, "you know" lasts longer than 30 frames and is spoken clearly without linking. Therefore, different from AdaSpeech 3 [11] predicting the position of fixed PF (e.g. "uh", "um"), we introduce a special token to replace speaker-dependent PF (e.g. "you know", "I mean"). We find that learning PF as a token can ensure the model to learn speaker-dependent pronunciation of PF, instead of common pronunciation shared by all people. Also, it further reinforces the personal style in the stand-up comedy scenario.

4. Experiments

4.1. Dataset

Due to the lack of stand-up comedy dataset, first we pre-train ComedicSpeech on AISHELL-3 [19], a large-scale and high-fidelity multi-speaker Mandarin speech corpus. Then the model is finetuned on our stand-up comedy dataset. We split our dataset into training, validation, and test sets. All clips in test set are unseen during training. Finally we randomly choose clips synthesized by test set for human evaluation. The way of pre-processing on the speech and text data follows Adaspeech [20].

4.2. Implementation Details

Our model backbone is based on FastSpeech 2 [3], with the configurations follow it unless otherwise stated. Following [5], the prosody encoder contains three modules, a 6-layer con-

Table 3: *The MOS (Mean Opinion Score), Similarity MOS, and Style MOS scores with 95% confidence.*

	MOS (↑)	Similarty MOS (↑)	Style MOS (↑)
GT	4.45 ± 0.07	4.45 ± 0.07	4.54 ± 0.07
GT mel + Voc	4.27 ± 0.07	4.28 ± 0.09	4.35 ± 0.08
FS 2 + spk emb	2.74 ± 0.12	2.47 ± 0.13	2.28 ± 0.13
FS 2 + ref enc	2.96 ± 0.11	2.89 ± 0.11	2.66 ± 0.12
ComedicSpeech	$\mathbf{3.24 \pm 0.10}$	$\mathbf{2.90 \pm 0.11}$	$\mathbf{2.80 \pm 0.12}$

volutional network with stride 2×2 used to down sample reference mel, a GRU with hidden size 128 used to generate a query vector, and a prosody space with 8 prosody embeddings of 256 dimension to represent prosody by a weighted combination.

We first pretrain ComedicSpeech for 300,000 steps on AISHELL-3 dataset [19]. Then we finetune the model on our constructed dataset for 100,000 steps. Both in training and fine-tuning stage, the reference speech fed in the prosody encoder is the ground truth mel-spectrogram extracted from the recording. And the prosody encoder is trained by all the losses in the model (e.g. duration loss, pitch loss) end-to-end. During inference, the reference audio is randomly selected from the training set. In addition, we find that the second half of Fastspeech 2 synthesized speeches often have low voice qualities and electric sounds, so we added weights for the loss of the second half:

$$Loss = loss_{firsthalf} + \alpha \times loss_{secondhalf} \quad (4)$$

4.3. Comparing ComedicSpeech with Baselines

We conduct subjective listening tests including MOS (Mean Opinion Score), SMOS (Similarity MOS), and Style MOS to compare ComedicSpeech with baselines. Each clip is listened by 10 native volunteers. All volunteers are required to focus on voice quality, voice similarity, and expressiveness respectively when scoring MOS, SMOS, and Style MOS. We compare the synthesized speech of ComedicSpeech with several
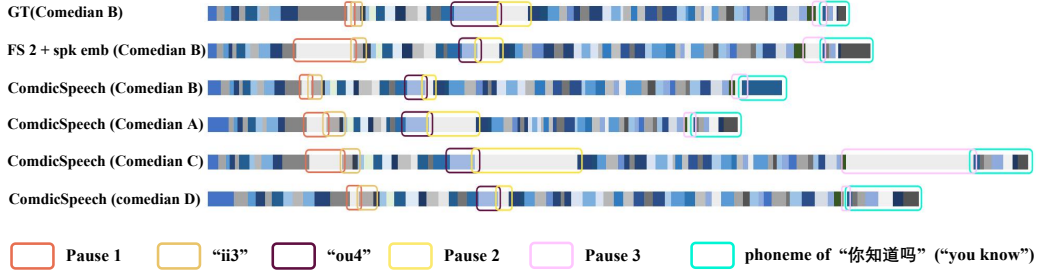
Figure 2: *Phoneme duration of the synthesized speech using the text content in Table 1. A color segment represents a phoneme, the same color means the same phoneme, and the length of the segment represents the duration of the phoneme.*

Table 4: *The CMOS (comparison MOS), Similarity CMOS and Style CMOS scores for ablation study. In this table, CLN stands for conditional layer normalization*

ID	Settings	CMOS	Similarity CMOS	Style CMOS
#1	ComedicSpeech	—	—	—
#2	#1 − duration CLN	− 0.05	− 0.23	− 0.57
#3	#1 + pitch and energy CLN	− 0.65	− 0.4	− 0.86
#4	#1 − special token	− 0.12	− 0.02	− 0.18

baselines: 1) GT, the ground truth recordings; 2) GT mel + Voc, using the ground-truth mel-spectrogram to synthesize waveform with HiFi-GAN vocoder. 3) FS 2 + spk emb, a FastSpeech 2 based multi-speaker TTS model using speaker ID to capture speaker-level acoustic conditions. 4) FS 2 + ref enc, a FastSpeech 2 based multi-speaker TTS model using a prosody encoder to extract speaker-level acoustic conditions. All settings use HiFi-GAN as the vocoder. Results are shown in Table 3. It indicates that ComedicSpeech makes considerable improvements in terms of voice quality, voice similarity, and expressiveness in stand-up comedy scenario. Also, it should be mentioned that there still is a performance gap between ComedicSpeech and recording, because of the difficulty of stand-up comedy synthesis in low-resource scenario.

4.4. Case Study

We use the case from Comedian B in Table 1 to study whether our method can capture the personal style and differences across comedians. We draw lines of phoneme duration output by different models in Figure 2, where the same color corresponds to the same phoneme. First, compared to the ground truth, our ComedicSpeech model for speaker B can better capture relative duration of phonemes, e.g., "Pause 1" and "ii3", and "ou4" and "Pause 2", than the baseline model FS 2 + spk emb [3]. That may result in better MOS, SMOS, and Style MOS scores. There is still room to improve on the absolute duration. Second, when using the ComedicSpeech models of other comedians, the same sentences have diverging speed of speech, pause duration, and PF duration. For example, Comedian C has the longest pause, i.e. "Pause 2", which is consistent with statistics in Table 2. He likes to pause for longer time and speak slowly. In addition, when looking close to the duration of PF, we find that "you know" is a PF for Comedian B but not for other comedians. Thus the other models do not take it as a whole special token. Consequently, they predicate longer du-

ration for "you know", rather than our model for Comedian B, which makes sense.

4.5. Ablation Studies

We conduct a series of ablation experiments to verify the effectiveness of each design in ComedicSpeech, including conditional layer normalization in the duration predictor, and the special tokens. We evaluate the CMOS (comparison MOS, focus on voice quality), Similarity CMOS (focus on voice similarity), and Style CMOS (focus on expressiveness) of the synthesized speech when removing each component in ComedicSpeech. The results are shown in Table 4.

We find that removing conditional layer normalization (CLN) from duration predictor impairs voice quality, voice similarity and speech expressiveness. However, adding CLN to all variance adaptor (e.g. pitch predictor, energy predictor) does not bring significant gain but leads to a drop of all three metrics. This is because CLNs in encoder and decoder are sufficient for the modeling of pitch and energy, and more CLNs in variance adaptor cause the model to focus too much on comedian's accent, creating strange tunes. However, the modeling of duration is so important and difficult in this scenario that only two CLNs cannot model it well. So our adding CLNs in only phoneme encoder, duration predictor, and mel decoder makes sence.In addition, directly regarding PF as an ordinary phoneme sequence impairs voice quality, voice similarity, and speech expressiveness, causing a decline in all three metrics. Therefore, our using a special token to replace the whole PF does work.

5. Conclusions

In this paper, we propose ComedicSpeech, a TTS model specially designed for stand-up comedy synthesis in low-resource scenario. We improve the personal speech style in three aspects. First, we introduce the prosody encoder to extract personal prosody of different comedians. Then we learn a more accurate duration distribution to improve personal rhythm by duration predictor. Next, we model the personal fillers by introducing a special token to ensure that model learns speaker-dependent pronunciation. Due to the lack of dataset, we build a multi-speaker dataset from a real-world stand-up comedy scenario. Experimental results show that our proposed ideas improve voice quality, voice similarity, and expressiveness for different comedians over strong baselines. In future, we will improve the modeling ability of the diverse personal style and speech expressiveness in stand-up comedy scenario. We will also try to combine TTS model with pretrain language model, to insert personal fillers for comedians automatically.

---

[3]Baseline FS 2 + spk emb performs better than FS 2 + ref enc in terms of duration loss, so we choose it to do the case study.

6. References

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 4006–4010. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1452.html

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 4779–4783. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3165–3174. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/f63f65b503e22cb970527f23c9ad7db1-Abstract.html

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[5] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5167–5176. [Online]. Available: http://proceedings.mlr.press/v80/wang18h.html

[6] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end TTS for voice agents," in *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*. IEEE, 2021, pp. 403–409. [Online]. Available: https://doi.org/10.1109/SLT48900.2021.9383460

[7] Y. Zhou, C. Song, X. Li, L. Zhang, Z. Wu, Y. Bian, D. Su, and H. Meng, "Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2573–2577. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10054

[8] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 7748–7759. [Online]. Available: http://proceedings.mlr.press/v139/min21b.html

[9] Y. Wu, X. Wang, S. Zhang, L. He, R. Song, and J. Nie, "Self-supervised context-aware style representation for expressive speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 5503–5507. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-686

[10] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2568–2572. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-901

[11] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W. Zhang, and T. Liu, "Adaspeech 3: Adaptive text to speech for spontaneous style," *CoRR*, vol. abs/2107.02530, 2021. [Online]. Available: https://arxiv.org/abs/2107.02530

[12] J. Cong, S. Yang, N. Hu, G. Li, L. Xie, and D. Su, "Controllable context-aware conversational speech synthesis," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 4658–4662. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-412

[13] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[14] J. Gustafson, J. Beskow, and E. Székely, "Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," *Proc. 11th ISCA SSW*, pp. 48–53, 2021.

[15] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," *CoRR*, vol. abs/2207.01063, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2207.01063

[16] A. O. Adigwe and E. Klabbers, "Strategies for developing a conversational speech dataset for text-to-speech synthesis," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2318–2322. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10802

[17] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 23*, 2023.

[18] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[19] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," *CoRR*, vol. abs/2010.11567, 2020. [Online]. Available: https://arxiv.org/abs/2010.11567

[20] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=Drynvt7gg4L