CONCSS: Contrastive-based Context Comprehension for Dialogue-appropriate Prosody in Conversational Speech Synthesis

Yayue Deng[1], Jinlong Xue[1], Yukang Jia[3], Qifei Li[1], Yichen Han[1], Fengping Wang[1], Yingming Gao[1], Dengfeng Ke[2], Ya Li[1]

[1]Beijing University of Posts and Telecommunications, Beijing, China

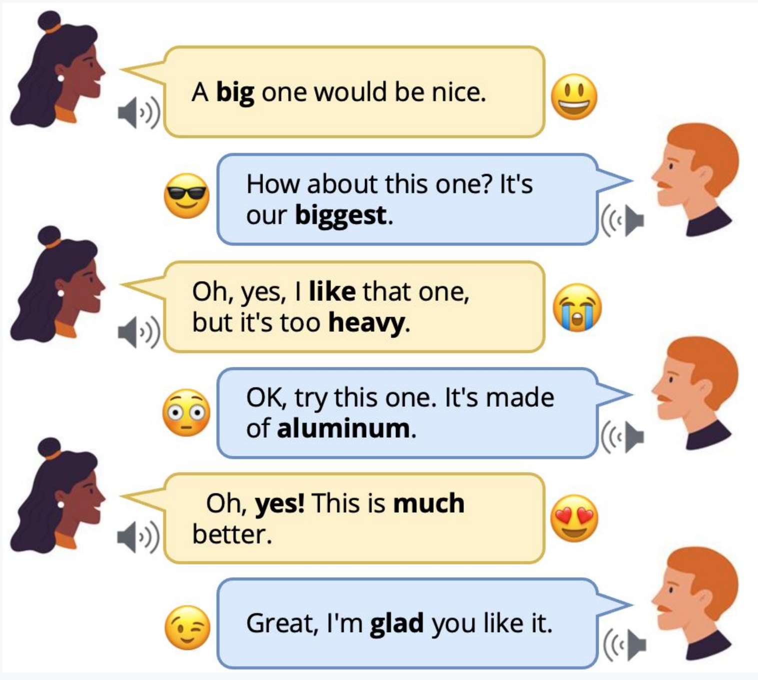
[2]Beijing Language and Culture University, Beijing, China

[3]Perfect World Co.,Ltd, Beijing, China

Conversational Speech Synthesis (CSS)

CSS Task Definition:

Given history dialogue, the CSS task focuses on improving the model's context understanding capability and generating audio with context-appropriate prosody.



Motivation

Limitation of Previous Works:

- Previous CSS approaches mostly rely on jointly training synthesis model and context encoder using the mel-reconstruction loss.
- Without explicit constraints, is this output vector of the context encoder sufficiently indicative of underlying context variations?

Contribution:

- A novel conversational speech synthesis framework CONCSS
- A novel pretext task specific to CSS
- Comprehensively evaluate models on their ability to produce context-sensitive vectors and dialogue-appropriate prosody

CONCSS Framework Overview

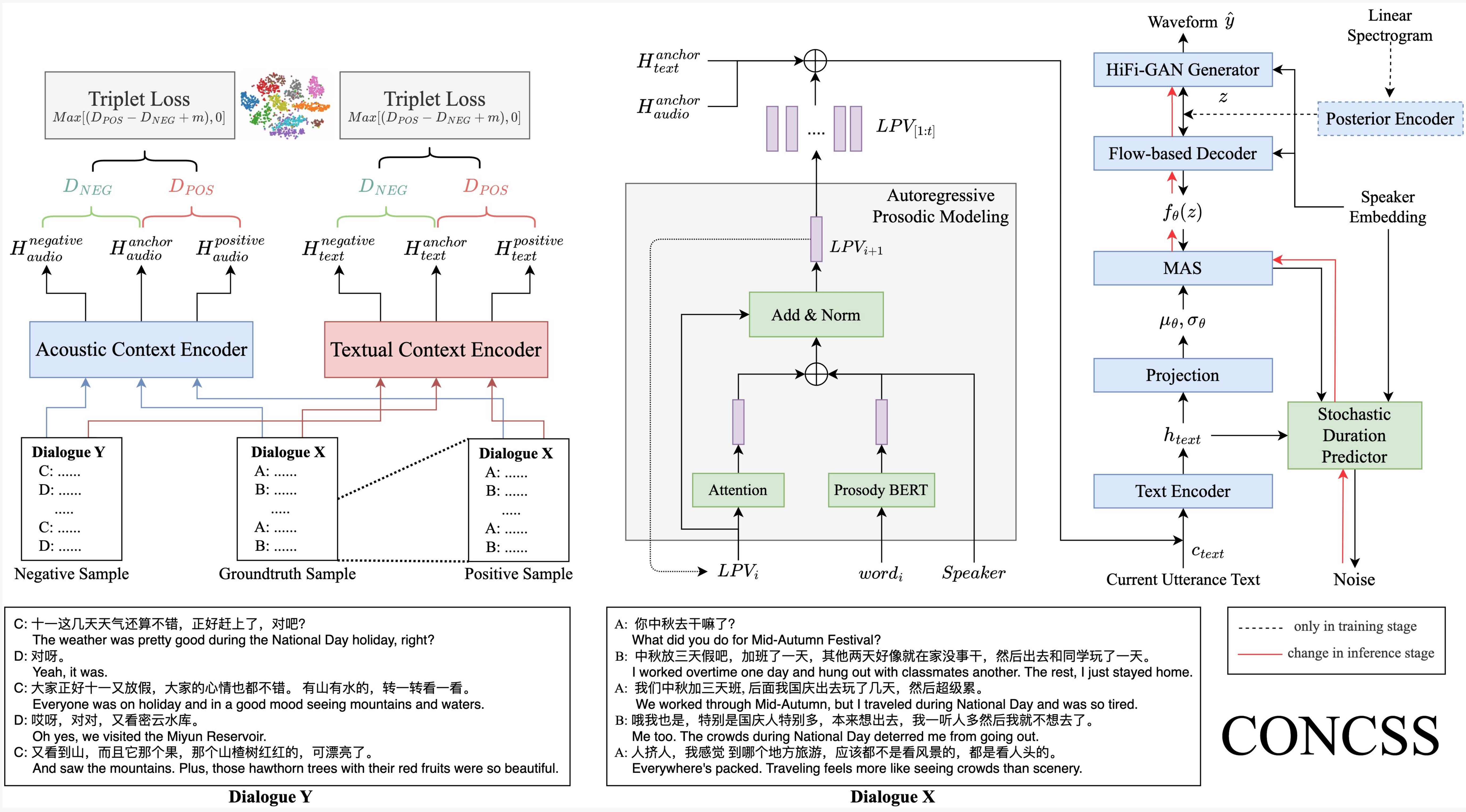


Fig.1. Illustration of our proposed CONtrastive-based Conversational Speech Synthesis (CONCSS).

Method

CONCSS = VITS+Four Enhancements :

- Leveraging an innovative pretext task to create context-dependent pseudo-labels

$$D(h_i + h_i^p) < D(h_i + h_i^n)$$

- Employ triplet loss with a hard negative sampling strategy

$$L(h_i^a, h_i^p, h_i^n) = max\{D(h_i^a - h_i^p) - D(h_i^a - h_i^n) + m,\ 0\}$$

- An Acoustic and Textual Context Encoder

$$\begin{cases} L_{text}^k = L(H_{text}^a, H_{text}^p, H_{text}^n) \\ L_{audio}^k = L(H_{audio}^a, H_{audio}^p, H_{audio}^n) \end{cases}$$

$$L_{contra} = \frac{1}{N}\sum_{k=1}^{N}(L_{text}^k + L_{audio}^k)$$

- Utilize an autoregressive prosodic modeling (APM) module with a pre-trained prosodic language model

Experiments and Conclusion

Table 1. Subjective evaluation (context-appropriate prosody and naturalness) for different models.

Model	GRU-based	M2CTTS	S1	S2	S3	S4
MOS (↑)	3.396 ± 0.107	3.438 ± 0.104	3.528 ± 0.097	3.708 ± 0.108	3.838 ± 0.110	$\mathbf{3.967 \pm 0.120}$

Models	GRU-based vs. M2CTTS	M2CTTS vs. S1	S1 vs. S2	S1 vs. S3	S2 vs. S3	S3 vs. S4	GRU-based vs. S4	M2CTTS vs. S4
CMOS (↑)	0.200	0.388	0.796	0.983	0.492	0.325	1.846	1.788

Table 2. **Objective evaluation metrics primarily focus on the context-sensitive prosody.** The `Real` context type uses the correct context for the current synthesized sentence, whereas the `Fake` type randomly selects from unrelated dialogues.

Method	Set	Type	Mel Loss (↓)	Log F0 RMSE (↓)	MCD (↓)
GRU-based		Real	3.599	0.2949 ± 0.1192	5.3590
		Fake	3.683	0.3001 ± 0.1164	5.3781
M2CTTS		Real	3.579	0.2936 ± 0.1014	5.3236
		Fake	3.596	0.3036 ± 0.1277	5.3882
CONCSS	S1	Real	3.609	0.2911 ± 0.1099	5.3382
		Fake	3.626	0.3203 ± 0.1093	5.4923
	S2	Real	3.556	0.2906 ± 0.1047	5.2883
		Fake	3.638	0.3311 ± 0.1417	5.5157
	S3	Real	3.530	0.2821 ± 0.0960	5.2748
		Fake	3.715	0.3272 ± 0.1455	5.6923
	S4	Real	**3.525**	$\mathbf{0.2803 \pm 0.0961}$	**5.2634**
		Fake	3.649	0.3252 ± 0.1097	5.6041

Table 3. Subjective evaluation between different context types.

Model	MOS (↑)		CMOS (↑)
	Real	Fake	Real vs Fake
GRU-based	3.442 ± 0.111	3.388 ± 0.102	0.325
M2CTTS	3.504 ± 0.100	3.312 ± 0.112	0.445
S1	3.638 ± 0.091	3.250 ± 0.116	0.492
S2	3.796 ± 0.076	3.229 ± 0.101	0.529
S3	$\mathbf{3.958 \pm 0.074}$	3.308 ± 0.100	**0.804**

Acknowledgements

The work was supported by the National Natural Science Foundation of China (NSFC) (No.62271083), the Key Project of the National Language Commission (No.ZDI145-81), the Fundamental Research Funds for the Central Universities (No.2023RC73, No.2023RC13).

References

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in ICML. 2021, vol. 139 of Proceedings of Machine Learning Research, pp. 5530–5540, PMLR

[2] Jinlong Xue, Yayue Deng, Fengping Wang, Ya Li, Yingming Gao, Jianhua Tao, Jianqing Sun, and Jiaen Liang, "M2-ctts: End-to-end multi-scale multi-modal conversational text_x0002_to-speech synthesis," in ICASSP 2023-2023 IEEE Interna_x0002_tional Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2023, pp. 1–5.

[3] Haohan Guo, Shaofei Zhang, Frank K Soong, Lei He, and Lei Xie, "Conversational end-to-end tts for voice agents," in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 403–409.

Contact Information

Yayue Deng
Email: yayue.deng@bupt.edu.cn
Phone: +86 18810956816