CMCU-CSS: Enhancing Naturalness via Commonsense-based Multi-modal Context Understanding in Conversational Speech Synthesis

Yayue Deng* Jinlong Xue* Fengping Wang
Yingming Gao Ya Li†

Beijing University of Posts and Telecommunications

SCAN ME

Main Objective

With the development of deep learning technology, there is an increasing demand for human-machine interaction in various scenarios, such as virtual robot assistants and virtual home teachers. Human-Computer Interaction (HCI) has become more prevalent in our daily life, and machines need to communicate with us in a spoken style. Therefore, conversational speech synthesis task (CSS) that aims to produce speech suitable for oral communication settings, has received widespread attention.

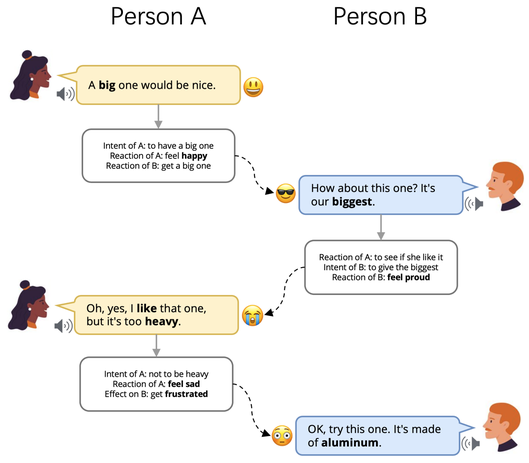

Figure 1. A example dialogue between two interlocutors, with corresponding emotions and implicit states for each utterance.

Task Definition: The objective of conversational speech synthesis (CSS) task is to synthesize context-appropriate audio u_N^a, by giving current text u_N^t with speaker p_N, and a sequence of $N-1$ historical utterances $[(u1, p1), (u2, p2), ..., (u_{N-1}, p_{N-1})]$ in conversation, where each utterance u_i is spoken by party p_i.

Challenges: The conversational speech synthesis system needs to understand the complex context and model the contextual dependencies between inter/intra speakers, which presents a challenge for the CSS task.

Motivation

The prior commonsense knowledge shared by interlocutors is essential in uncovering the implicit variables of the conversation (e.g., intent, emotion). This knowledge acts as a guide for the individuals, assisting them in their cognitive processes related to the conversation's content, including reasoning, emotional convergence, empathy, and other related phenomena. Fig. 1 demonstrates how commonsense knowledge plays an essential role in context understanding. The subtle psychological influence, which is mutual and context-dependent, results in different expressions from the speaker. **Thereby, we investigate the use of commonsense knowledge to model mutual psychological influence between interlocutors in CSS task.**

In summary, the contributions of this work are:

- propose a novel Commonsense-based Multi-modal Context Understanding module (CMCU)
- the first to adopt external commonsense knowledge with multi-modal information into CSS task
- validate our proposed module subjectively and objectively, where the proposed CMCU achieves the best results.

Methodology

As shown in Fig. 2, the whole conversational speech synthesis framework CMCU-CSS that we propose has six parts: 1) CMCU module, 2) text encoder, 3) posterior encoder, 4) flow-based decoder, 5) stochastic duration predictor, and 6) HiFi-GAN generator.

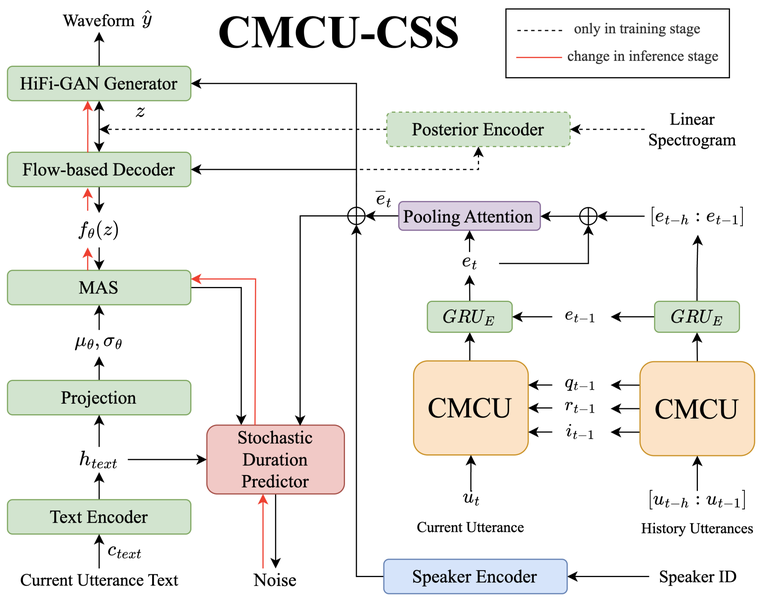

Figure 2. Overview of our proposed conversational speech synthesis system CMCU-CSS that incorporates the Multi-modal Commonsense-based Context Understanding (CMCU) module with speech synthesis framework VITS.

Moreover, Fig. 3 illustrates details of multi-modal features and commonsense knowledge extraction in our framework. For multi-modal feature extraction, we process textual features u_i^t and acoustic features u_i^a for the i-th utterance separately. For the textual modality, we utilize the widely-used large-scale pretrained language model RoBERTa Large [5] and finetune it on the emotion recognition task. For the acoustic modality, we employ the pretrained acoustic model Wave2Vec 2.0 [1] and also finetune it on the downstream emotion classification task to extract acoustic feature vectors w_i that are relevant to emotional expression. For commonsense knowledge extraction, we employ the GPT-based model COMET [2]. Three different states (i.e., intent state, internal state and external state) are established in the CMCU module to model the context dependency between inter/intra speakers on different aspects of conversation.

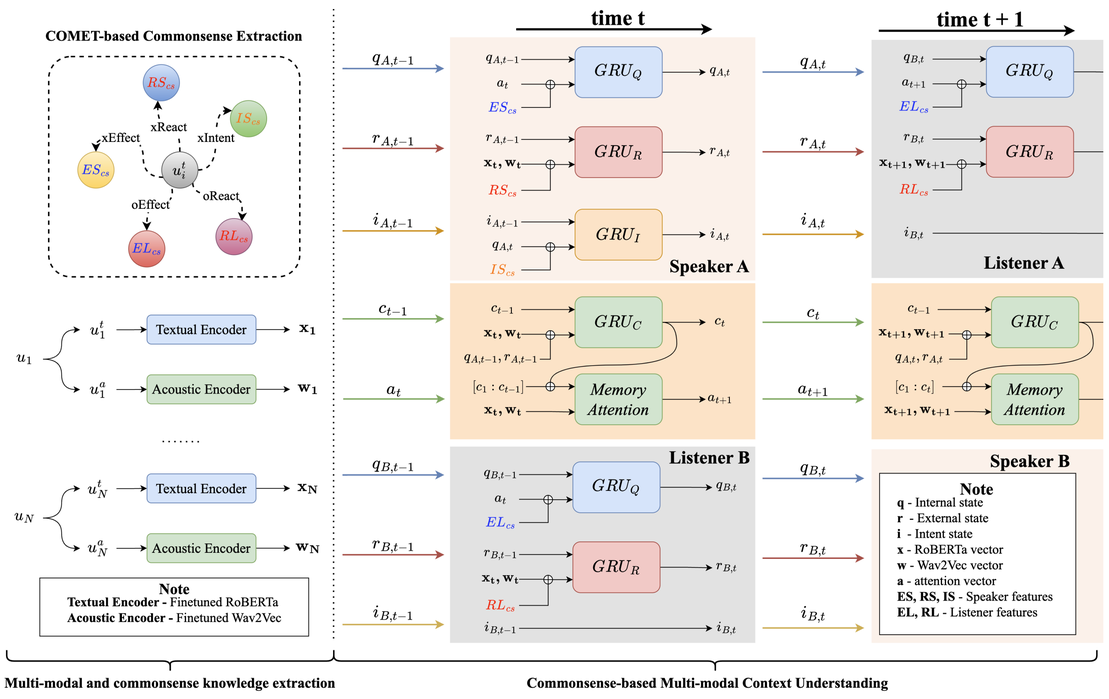

Figure 3. Illustration of our proposed Commonsense-based Multi-modal Context Understanding (CMCU).

Experimental Evaluations

To verify the effectiveness of our proposed method in conversational speech synthesis, we compare three approaches with different historical context modeling, and they all employ VITS as synthesis backbone.

- **No Context Modeling:** vanilla VITS which has no additional conditional input except phoneme sequences and speaker information, denoted as "M_1"
- **GRU-based [4]:** The model is denoted as "M_2", where textual context embedding is compassed by a uni-directional gated recurrent unit (GRU) to model context dependency.
- **Commonsense-based Context Modeling:** This method uses context modeling similar to CMCU, except it only adopts textual feature, denoted as "M_3".
- **CMCU-based Context Modeling (Proposed):** Our proposed method denoted as "M_4".

Table 1. The results of the CMOS tests in DailyTalk and IEMOCAP datasets. CMOS describes the preference degree between A and B (denote as A vs. B). The Preference(%), including Left, Neutral and Right, is calculated according to the CMOS score.

Comparisons	DailyTalk				IEMOCAP			
	CMOS	Left	Neutral	Right	CMOS	Left	Neutral	Right
M_1 vs. M_2	0.037	15.15	65.90	18.93	0.078	21.42	49.28	29.28
M_2 vs. M_3	0.129	12.97	62.59	24.42	0.289	12.15	52.14	35.71
M_3 vs. M_4	0.196	9.84	61.36	28.78	0.382	12.14	40.01	47.85
M_2 vs. M_4	0.274	9.16	58.77	32.06	0.607	14.28	26.42	59.28

To evaluate the capacity of context understanding and dependency modeling, we validate our proposed CMCU module in the task of emotion recognition in conversation (ERC) task. Our proposed approach is validated on various datasets, including IEMOCAP, MELD, where CMCU achieves the best results.

Table 2. The results of emotion recognition in conversation.

Models	IEMOCAP	MELD
	W-Avg F1	W-Avg F1 (7-cls)
GRU-based [4]	62.57	57.03
Graph-based [3]	64.18	58.10
Commonsense-based-w/o-listener	65.33	64.28
Commonsense-based-w-listener	66.37	65.17
CMCU-based-w/o-listener (Proposed)	67.32	65.10
CMCU-based-w-listener (Proposed)	67.42	65.62

CONCLUSION

In this work, we **present a novel conversational speech synthesis system (CMCU-CSS)** for enhancing context understanding and generating natural speech with context-appropriate emotion. Specifically, we **propose a Commonsense-based Multi-modal Context Understanding (CMCU) module** to capture context dependency between inter/intra speakers which is further utilized to infer emotion vectors. Furthermore, we **conduct an ERC task** to verify the effectiveness of CMCU-based context modeling method. Meanwhile, the results of subjective and objective evaluations show that CMCU-CSS can generate more natural and appropriate speech compared with other CSS systems, as **the prior knowledge from commonsense vectors and multi-modal information strengthens the model's understanding of the context, resulting in more appropriate emotion inference**.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL (1)*, pages 4762–4779. Association for Computational Linguistics, 2019.

[3] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP/IJCNLP (1)*, pages 154–164. Association for Computational Linguistics, 2019.

[4] 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)*, pages 403–409. IEEE, 2021.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.