Wed-O-OS-7-1 INTERSPEECH2022 Presentation

Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History

Yuto Nishimura[1]

Yuki Saito[1]

Shinnosuke Takamichi[1]

Kentaro Tachibana[2]

Hiroshi Saruwatari[1]

[1]The University of Tokyo, Japan, [2]LINE Corp., Japan

Table of contents

1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

Table of contents

1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

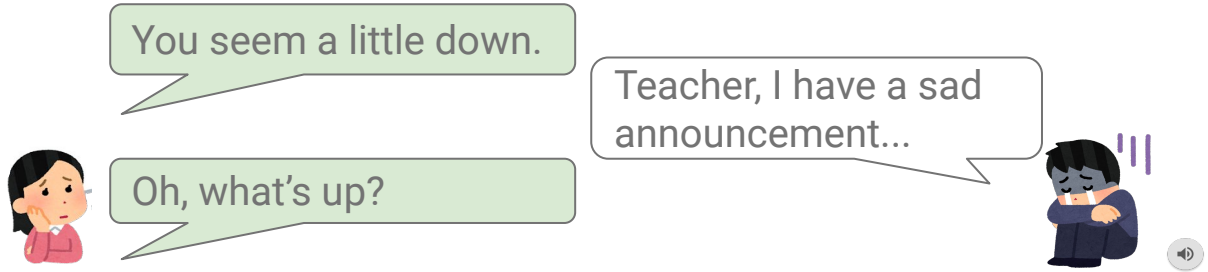
Main topic: dialogue system with empathy

- **Dialogue system:** interact w/ humans by text/speech
  - **Task-oriented**: satisfy user's requrest
    - e.g., Tourist information, restaurant reservation
  - **Non task-oriented**: communicate with user
    - e.g., Chit-Chat
- **Empathy:** active attempt to get inside other person [Davis+18]
  - c.f., Sympathy: synchronize self with other person in emotion

How can we develop dialogue system that can talk to users w/ empathetic speaking style?

Task definition

- **Empathetic Dialogue Speech Synthesis (DSS)** [Saito+22]
  - Reflect main elements of empathy (i.e., **emotion**) on synthetic speech
  - Estimate speech features that contribute to next response, considering **dialogue history** (interaction betw. system & user)



- **Challanging point**
  - Predicting **dialogue context** from linguistic & prosodic features (i.e., modeling **cross-modality** of text & speech)

Overview of our research

- **Conventional DSS method:** using text history only [Guo+20]
  - Learn dialogue context from text embeddings of dialogue history
  - **Limitation:** missing speech modality modeling
- **Proposed DSS method:** using both text & speech history
  - Extract prosody embedding from speech & aggregate two modality
  - Investigate 4 methods for better dialogue context modeling:
    1) pre-trained speech SSL* model, 2) style-guided training,
    3) cross-modal attention, 4) fine-grained embedding modeling
- **Result: more natural DSS than conventional method**

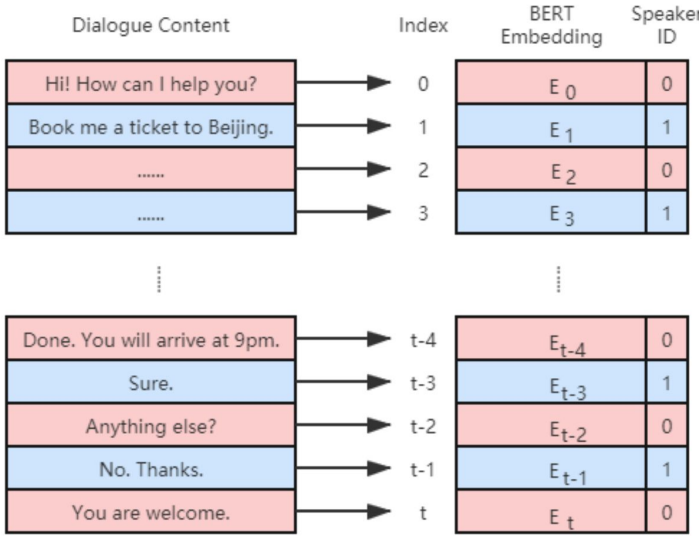
*SSL: Self-Supervised Learning

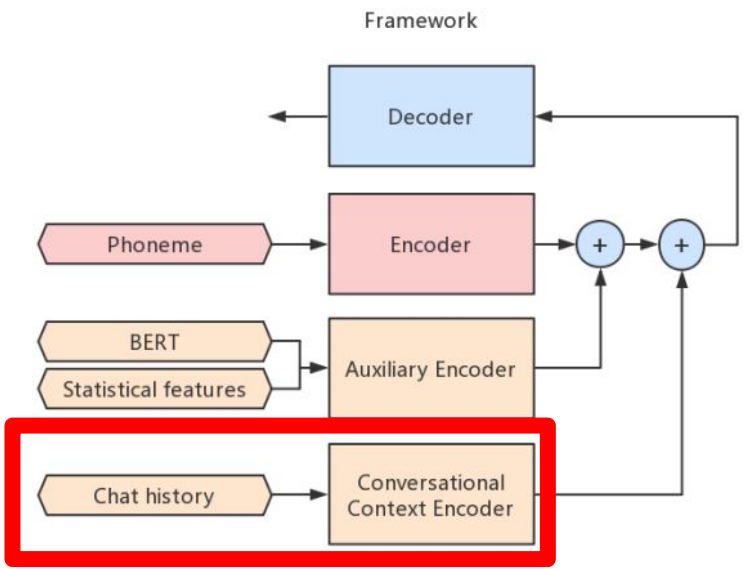
Table of contents

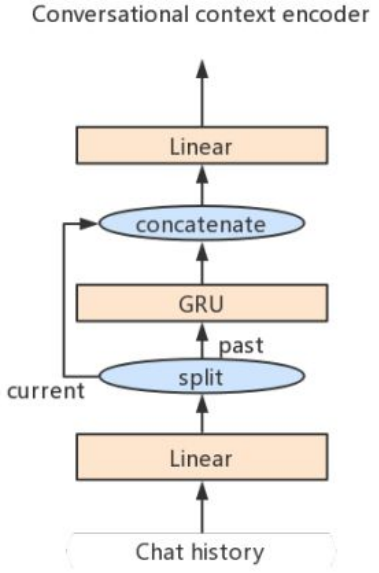
1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

Conventional DSS method [Guo+20]

- **Overview:** E2E TTS w/ **Conversational Context Encoder (CCE)**
  - Step 1: obtain text embeddings using sentence BERT
  - Step 2: extract **context embedding** from chat history w/ CCE







Content

1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

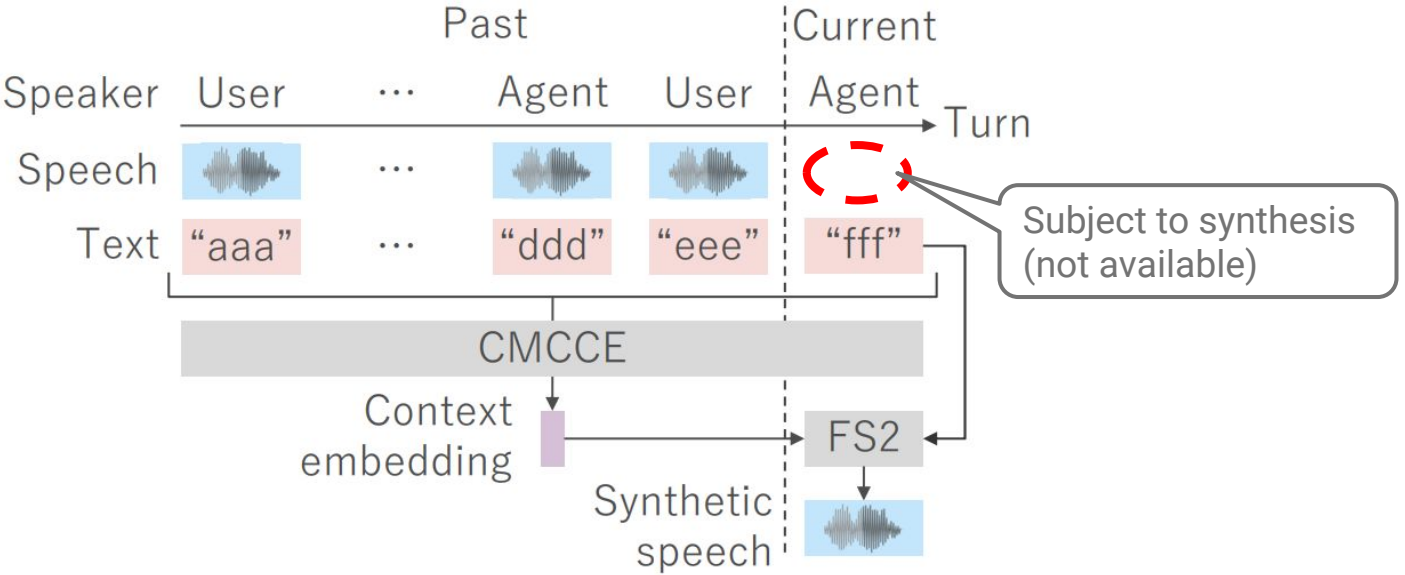
Motivation

- **One-to-many problem in TTS**
  - e.g., “What’s wrong?” w/ various speech prosody


- **Research questions**
  - RQ1: *Can we extract better dialogue context from chat history by considering BOTH text & speech?*
  - RQ2: *How can we learn the cross-modality of text & speech effectively, rather than processing them independently?*

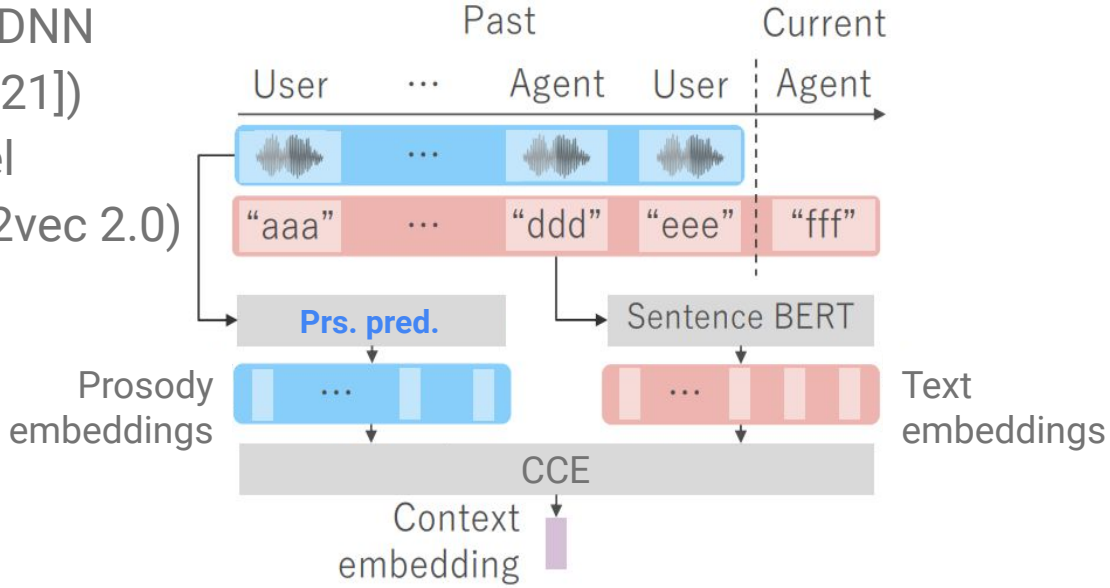
Overview of proposed method

- **Architecture:** FS2-based TTS model w/ **Cross-Modal (CM)CCE**
  - CMCCE: extracting context embedding from text/speech seqs.
  - 4 methods for better context embedding extraction



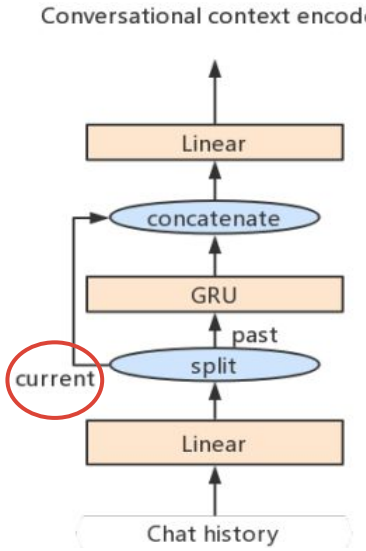
CMCCE w/ prosody predictor

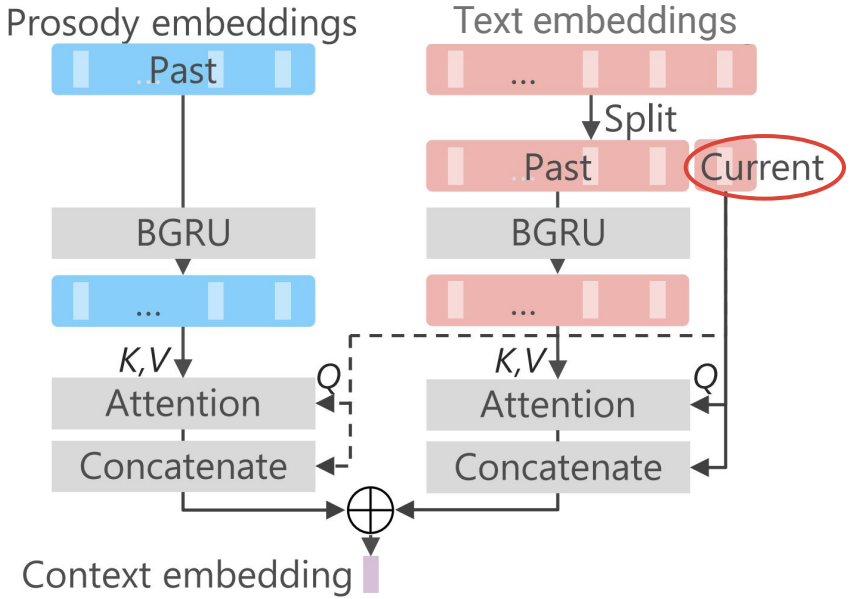
- **Main components**
  - Sentence BERT for text embedding extraction
  - **Prosody predictor** for prosody embedding extraction
    - Trainable DNN (e.g., [Du+21])
    - SSL model (e.g., wav2vec 2.0)



Cross-modal attention

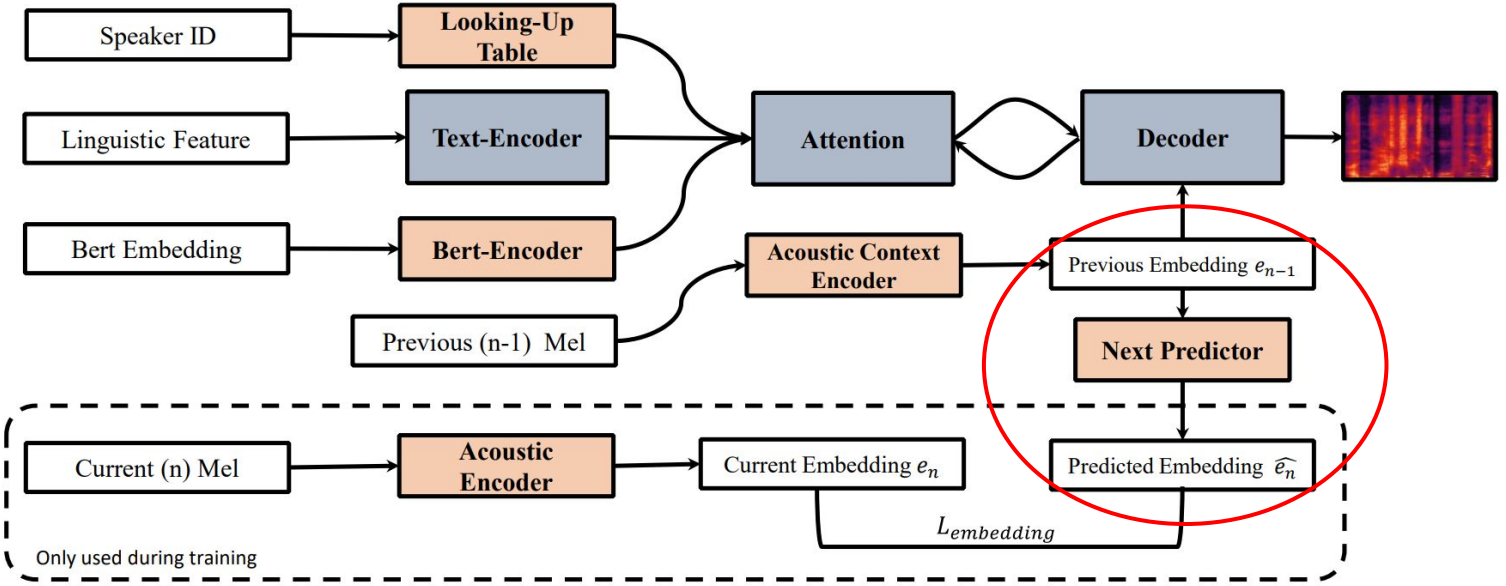
- **How to compress past information of dialogue history**
  - Guo et al.'s [Guo+20]: bi-directional Gated Recurrent Unit (GRU)
  - Ours: attention using embedding of current text as query





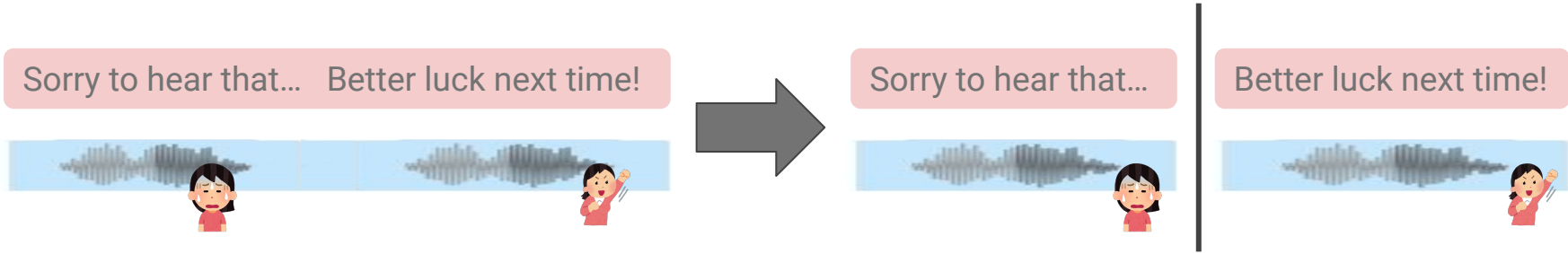
Style-guided context embedding learning

- **Core idea:** Cong et al's method [Cong+21]
  - Associating context embedding with current prosody embedding



Fine-grained context embedding modeling

- **Unit of embedding modeling**
  - Guo et al.'s [Guo+20]: utterance-wise
    - Cannot model change of prosodic variation within one utterance
  - Ours: **sentence**-wise
    - Divide current utterance into sentences by punctuation symbols
    - Extract text/prosody embedding for each sentence
    - Predict sentence-wise context embedding from extracted embeddings using CMCCE



Content

1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

Experimental conditions

Corpus	STUDIES [Saito+22] (downsampled to 22,050 Hz)
Data splitting	{ Training, Validation, Test } = { 2,209, 221, 211 }
TTS model (w/o teacher forcing)	Text2Mel: FastSpeech 2 (FS2) [Ren+21] Vocoder: HiFi-GAN [Kong+20]
Dialogue history length	10 (same setting as [Guo+20])
Compared methods	**Baseline**: FS2 + CCE (Guo et al's method [Guo+20]) **Proposed**: FS2 + CMCCE ● SSL: pretrained SSL model as prosody extractor ● Attn: attention for cross-modal aggregation ● SG: style-guided embedding learning ● FG: fine-grained context modeling
Subjective evaluation	**Stage 1: Pairwise comparison (AB/XAB tests)** **Stage 2: MOS test**

Results of preference AB/XAB tests

- **w/o SSL**
  - Significant improvement by:
    - **+SG**
    - **+SG**+Attn and **+SG**+FG

    → **SG was effective in training for CMCCE w/ prosody predictor.**

Baseline	Naturalness	Similarity	Proposed (w/o SSL) SG	Attn	FG
	0.45 vs. 0.55	**0.54 vs. 0.46**			
	0.44 vs. 0.56	0.53 vs. 0.47	✓		
	0.50 vs. 0.50	0.54 vs. 0.46		✓	
	0.48 vs. 0.52	**0.54 vs. 0.46**			✓
	0.43 vs. 0.57	0.47 vs. 0.53	✓	✓	
	0.45 vs. 0.55	**0.45 vs. 0.55**	✓		✓

- **w/ SSL**
  - Significant improvement by:
    - **+Attn**
    - +SG+**Attn**

    → **Attn aggregated SSL-derived prosody & text embeddings.**
  - Fewer cases w/ improvement, compared with w/o SSL results

Baseline	Naturalness	Similarity	Proposed (w/ SSL) SG	Attn	FG
	0.50 vs. 0.50	**0.61 vs. 0.39**			
	0.53 vs. 0.47	0.46 vs. 0.54	✓		
	0.51 vs. 0.49	**0.44 vs. 0.56**		✓	
	0.52 vs. 0.48	0.50 vs. 0.50			✓
	0.43 vs. 0.57	0.50 vs. 0.50	✓	✓	
	0.46 vs. 0.54	**0.54 vs. 0.46**		✓	✓

Results of MOS test

- **Compared methods: Baseline vs. Proposed (w/o SSL)**
  - +SG+FG (best combination)
  - +SG, +FG (ablation)
  - +SG+Attn+FG (bonus)

Method			Naturalness MOS
Proposed (w/o SSL)			
SG	Attn	FG	
			3.59±0.10
✓			3.62±0.10
		✓	3.59±0.10
✓		✓	3.66±0.10
✓	✓	✓	3.55±0.10
Baseline			3.55±0.10

- **Summary of results**
  - **+SG+SG achieved the highest MOS.**
    - No significant difference betw. Baseline & Proposed…
  - +SG+Attn+FG did not improve the naturalness.
    - Richer model → more difficult training?

Speech samples (available online)

100 listeners (24 answers per listener)

Content

1. Introduction
2. Conventional method
3. Proposed method
4. Experimental evaluation
5. Conclusion

Conclusion

- **Purpose: development of more natural voice agent**
  - Control speaking style according to user's emotion w/ **empathy**
- **This talk: modeling dialogue context from text/speech history**
  - Extract prosody embedding from speech & aggregate two modality
  - Investigate 4 methods for better dialogue context modeling:
    1) pre-trained SSL* model, 2) style-guided training,
    3) cross-modal attention, 4) fine-grained embedding modeling
- **Result: more natural DSS than conventional method**
- **Future work: (semi-)supervised learning using emotion label**