Emotion Rendering for Conversational Speech Synthesis with Heterogeneous Graph-Based Context Modeling

Rui Liu[1]*, **Yifan Hu**[1], **Yi Ren**[2], **Xiang Yin**[2], **Haizhou Li**[3,4]

[1]Inner Mongolian University, China
[2]ByteDance
[3]Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[4]National University of Singapore, Singapore
liurui_imu@163.com, hyfwalker@163.com, ren.yi@bytedance.com, yinxiang.stephen@bytedance.com,
haizhouli@cuhk.edu.cn

Abstract

Conversational Speech Synthesis (CSS) aims to accurately express an utterance with the appropriate prosody and emotional inflection within a conversational setting. While recognising the significance of CSS task, the prior studies have not thoroughly investigated the emotional expressiveness problems due to the scarcity of emotional conversational datasets and the difficulty of stateful emotion modeling. In this paper, we propose a novel emotional CSS model, termed ECSS, that includes two main components: 1) to enhance emotion understanding, we introduce a heterogeneous graph-based emotional context modeling mechanism, which takes the multi-source dialogue history as input to model the dialogue context and learn the emotion cues from the context; 2) to achieve emotion rendering, we employ a contrastive learning-based emotion renderer module to infer the accurate emotion style for the target utterance. To address the issue of data scarcity, we meticulously create emotional labels in terms of category and intensity, and annotate additional emotional information on the existing conversational dataset (DailyTalk). Both objective and subjective evaluations suggest that our model outperforms the baseline models in understanding and rendering emotions. These evaluations also underscore the importance of comprehensive emotional annotations. Code and audio samples can be found at: https://github.com/walker-hyf/ECSS.

Introduction

Conversational speech synthesis (CSS) aims to express a target utterance with the proper linguistic and affective prosody in a conversational context (Guo et al. 2020). With the development of human-machine conversations, CSS has become an integral part of intelligent interactive systems (Zhou et al. 2020; Seaborn et al. 2021; McTear 2022) and plays an important role in areas such as virtual assistants and voice agents, etc.

Unlike the speech synthesis technology for single utterance that just predicts the speaking style according to its linguistic content (Wang et al. 2017; Ren et al. 2021; Liu et al. 2021a,b, 2022a, 2024) or attempt to transfer the style information from an additional reference speech (Wang et al. 2018; Li et al. 2022c; Huang et al. 2022), CSS methods usually infer the speaking style of the target utterance according to the dialogue interaction history between two interlocutors. Traditional CSS works attempt to acquire the speaking style information from various aspects, such as inter and intra-speaker (Li et al. 2022a), and multi-modal context (Li et al. 2022a,b; Xue et al. 2023) dependencies modeling, etc. For example, Guo et al. (2020) construct a coarse-grained context encoder at the sentence level and use the recurrent neural network (RNN) for dialogue history encoding. Li et al. (2022a) model inter-speaker and intra-speaker dependencies in a conversation by a dialog graph convolu-

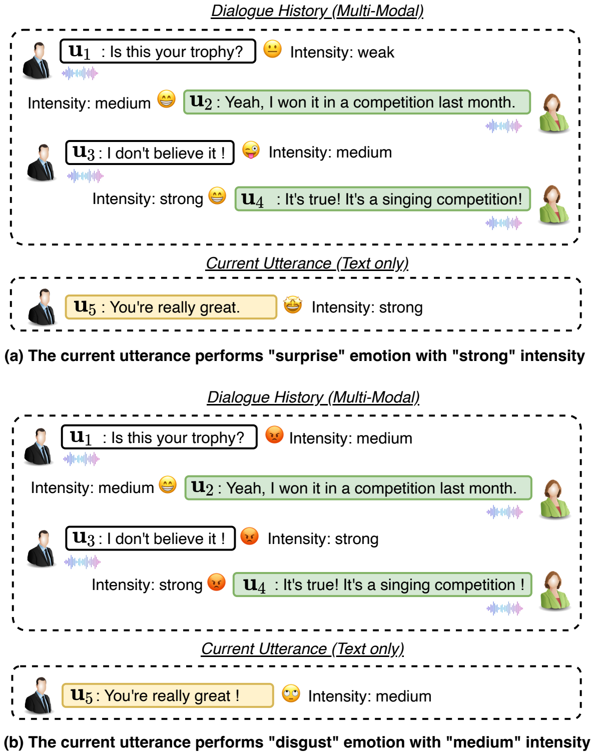
Dialogue History (Multi-Modal)
$\mathbf{u}_1$: Is this your trophy?
Intensity: weak
Intensity: medium
$\mathbf{u}_2$: Yeah, I won it in a competition last month.
$\mathbf{u}_3$: I don't believe it !
Intensity: medium
Intensity: strong
$\mathbf{u}_4$: It's true! It's a singing competition!
Current Utterance (Text only)
$\mathbf{u}_5$: You're really great.
Intensity: strong
(a) The current utterance performs "surprise" emotion with "strong" intensity
Dialogue History (Multi-Modal)
$\mathbf{u}_1$: Is this your trophy?
Intensity: medium
Intensity: medium
$\mathbf{u}_2$: Yeah, I won it in a competition last month.
$\mathbf{u}_3$: I don't believe it !
Intensity: strong
Intensity: strong
$\mathbf{u}_4$: It's true! It's a singing competition !
Current Utterance (Text only)
$\mathbf{u}_5$: You're really great !
Intensity: medium
(b) The current utterance performs "disgust" emotion with "medium" intensity

Figure 1: Graphical depiction of spoken conversation, where different emotion cues in the dialogue history perform a direct impact on the emotion expression for the current utterance.

*Corrposending Author.

tional neural (GCN) network and summarize the output of the graph neural network using the attention mechanism. In more recent studies, Li et al. (2022b) propose a multi-scale relational graph convolutional network (MRGCN) to learn the dependencies in conversations at both global and local scales among the multi-modal information. The above researches contribute to understanding the conversational context and determining appropriate speaking styles in synthesized speeches.

However, **emotion understanding** and **emotion rendering** are largely missing in prior CSS research due to the scarcity of emotional conversational datasets and the difficulty of stateful emotion modeling. In human-computer conversations, modeling the emotional expression in the conversational context is crucial for the speech synthesis system to generate speech with the appropriate emotional states and improve the user experience in speech-based interactions. As illustrated by the example in Fig.1, the current utterance in Fig. 1(a) performs the "Surprise" emotion when the emotion flow of dialogue history is "Happy → Happy → Surprise → Happy", while in Fig. 1(b) performs "Disgust" emotion when emotion flow is "Angry → Happy → Angry → Angry". We can conclude that the emotional expression in context can directly affect the speaking style for the target utterance. In addition, humans tend to have multiple emotions with varying intensities, such as weak, medium and strong, while expressing their thoughts and feelings (Firdaus et al. 2020; Im et al. 2022; Liu et al. 2022b; Zhou et al. 2023). Therefore, emotion intensity also has an essential impact on speech expressiveness. In a nutshell, how to fully understand the emotional cues of contextual information, and based on this, adequate emotional rendering in synthesizing conversational speech will be the focus of this paper. Last but not least, the existing emotion-aware multimodal data (Busso et al. 2008; Poria et al. 2018; Saganowski et al. 2022; Dias et al. 2022; Zuo et al. 2023) are mostly targeted at emotion recognition scenarios, in which the speech fidelity of the audio modality is not high enough to meet the data requirements of conversational speech synthesis, resulting in the problem of data scarcity.

To address the above challenges, we propose a novel emotional CSS model, termed **ECSS**, that includes two novel mechanisms: 1) to enhance emotion understanding for the context utterances, *heterogeneous graph-based emotional context modeling* module is proposed to learn the emotion cues of emotional conversational context. Specifically, given the multi-modal context in a conversation, the text, audio, speaker, emotion, and emotion intensity information are treated as the multi-source knowledge and used to build the nodes of the heterogeneous Emotional Conversational Graph called **ECG**. Then, the graph encoding module adopts the Heterogeneous Graph Transformer (HGT) (Hu et al. 2020) as the backbone to learn complex emotional dependencies in the context, and to learn the impact of such complex dependencies on the emotional expression of the current utterance; 2) to achieve the emotion rendering for the current utterance, we employ a contrastive learning-based emotion renderer module to infer the accurate emotion style for the target utterance. Specifically, the emotion renderer takes the emotion-aware graph-enhanced node features from ECG as input and predicts the appropriate emotion, emotion intensity, and prosody features for the current utterance. This information is later aggregated with the content and speaker representations of the current utterance into the acoustic decoder to synthesize the final emotional conversational speech. Note that the new contrastive learning losses are used to enhance the differentiation of emotion and emotional intensity expressions by drawing the same categories of emotion (or intensity) closer together and pushing different categories farther apart. It's worth noting that, to guarantee the successful development of the ECSS model, we designed seven emotion labels (happy, sad, angry, disgust, fear, surprise, neutral), and three emotion intensity labels (weak, medium, strong) for a recent expressive conversational speech synthesis dataset, DailyTalk (Lee, Park, and Kim 2023), and invited professional practitioners to annotate the labels. All annotated data will be open-sourced. The main contributions of this paper include:

- We propose a novel emotional conversational speech synthesis model, termed ECSS. To our best knowledge, this is the first in-depth conversational speech synthesis study that models emotional expressiveness.
- The proposed heterogeneous graph-based emotional context modeling and emotion rendering mechanisms ensure the accurate generation of emotional conversational speech in terms of emotion understanding and expression, respectively.
- Objective and subjective experiments show that the proposed model outperforms all state-of-the-art baselines in terms of emotional expressiveness.

Related Works

The emotion modeling for conversation has been studied in both natural language processing (NLP) (Zhong, Wang, and Miao 2019; Goel et al. 2021) and speech processing fields. However, multi-modal information other than textual information is rarely considered in NLP. In speech processing field, the advanced conversational speech synthesis methods use Graph Neural Networks (GNNs) models (Li et al. 2022a,b) to understand the multi-modal conversational context and infer the appropriate speaking style for the target utterance. For example, multi-modal features of all past utterances in the conversation, including textual information and speaking style information, are modeled by Dialogue Graph Convolutional Network (DialogueGCN) (Ghosal et al. 2019) to produce new representations holding richer context knowledge. However, the GNNs for CSS are designed for homogeneous graphs, in which all nodes and edges belong to the same types (such as utterance nodes from different speakers), making them infeasible to represent the natural heterogeneous structures in conversation. Especially for emotional expressions in conversation, information such as text, audio, speaker, emotion, and emotion intensity can be seen as nodes of the heterogeneous graph.

We note that there are some multi-modal conversational emotion recognition (MMCER) works that adopt heterogeneous graph networks to model the complex dependencies of contexts adequately (Li et al. 2022d; Song et al. 2023). Unlike previous studies, our heterogeneous graph module has some clear differences from these works: 1) we add emotion and emotion intensity nodes into the graph structure to model the dynamic emotion cues in conversation context; 2) we adopt *Heterogeneous Graph Transformer* as the backbone to encode the relations between heterogeneous nodes to learn the high-level feature representation for the constructed graph. Note that our work is the first attempt to model the emotion understanding and rendering in conversations for CSS with heterogeneous graph networks.

Task Definition

A conversation can be defined as a sequence of utterances $(utt_1, utt_2, ..., utt_{\mathcal{J}}, utt_{\mathcal{C}})$, where $\{utt_1, utt_2, ..., utt_{\mathcal{J}}\}$ is the dialogue history till round $\mathcal{J}$ while $utt_{\mathcal{C}}$ means the current utterance to be synthesized. The task of emotional conversational speech synthesis aims to synthesize the audio $a_{\mathcal{C}}$ given the $utt_{\mathcal{C}}$ and the dialogue history $\{utt_1, utt_2, ..., utt_{\mathcal{J}}\}$. For the multi-modal context, each utterance utt_j $(j \in [1, \mathcal{J}])$ in the dialogue history can be represented by five-tuples like <text$_j$, speaker$_j$, audio$_j$, emotion$_j$, emotion intensity$_j$>, in short for $< u_j$, s_j, a_j, e_j, i_j>. Note that the $utt_{\mathcal{C}}$ can be represented by only two-tuples like <text$_{\mathcal{C}}$, speaker$_{\mathcal{C}}$ >, in short for $< u_{\mathcal{C}}$, $s_{\mathcal{C}} >$, since emotion and intensity information need to be generated by ECSS.

Methodology

As shown in the pipeline of Fig. 2, the proposed ECSS consists of three components, that are 1) *Multi-source knowledge*; 2) *Heterogeneous Graph-based Emotional Context Encoder* and 3) *Emotional Conversational Speech Synthesizer*. As mentioned before, the multi-modal context, including text, speaker, audio, emotion and intensity, contains natural multi-source information and can therefore be viewed as multi-source knowledge. To enhance emotion understating, the *Heterogeneous Graph-based Emotional Context Encoder* constructs a heterogeneous Emotional Conversational Graph (ECG) by considering each kind of information in the multi-source knowledge as a node, and obtains a graph-enhanced emotional contextual representation for each node after modeling the dependencies among all heterogeneous nodes. To achieve emotion rendering, the *Emotional Conversational Speech Synthesizer* utilizes the graph-enhanced contextual representation in the ECG to make a reasonable prediction of the emotion expression information of the current sentence and further generates the emotional conversational speech.

Heterogeneous Graph-Based Emotional Context Encoder

As shown in the middle panel of Fig. 2, the heterogeneous graph-based emotional context encoder consists of three parts: 1) ECG Construction, constructing the heterogeneous graph with multi-source context; 2) ECG Initialization, initializing different heterogeneous nodes via their own feature representations; 3) ECG Encoding, perceiving emotion cues and generating the emotion-aware feature representation for the heterogeneous nodes.

ECG Construction Unlike previous GNNs based CSS methods, we aim to introduce the multi-source knowledge, which are 5 kinds of nodes including text f_u, audio f_a, speaker f_s, emotion f_e, and intensity f_i and build an emotional conversational graph or ECG $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ denotes the set of nodes, and $\mathcal{E}$ denotes the set of edges representing the relations between two nodes. Note that the speaker, audio and text nodes seek to introduce the basic dialogue attributes, while emotion and intensity nodes can introduce the dynamic emotion traits and bridge the emotion interaction between remote utterances. As shown in the middle part of Fig. 2, different shapes of diagrams mean different kinds of nodes.

Considering the multi-source knowledge, We created 14 different types of edges, as shown by the different colored connecting lines shown in the middle part of Fig. 2. However, due to space limits, not all the edges of the nodes are depicted. In a nutshell, these 14 edges connect 1) the text and each of the other nodes, 2) the audio and speaker nodes, 3) the emotion and speaker, emotion intensity, and audio nodes, and 4) emotion intensity and speaker, audio nodes. Note that all edges include past-to-future and future-to-past connections to model the bidirectional relation.

ECG Initialization To achieve meaningful heterogeneous graph encoding, we need to initialize all nodes with their feature representations. As shown in the middle and bottom panels of Fig. 2, to take the multi-turn dialogue as an example, we employ various encoders to obtain f_{u_j}, f_{s_j}, f_{a_j}, f_{e_j}, f_{i_j} $(j \in [1, 4]])$ for text, speaker, audio, emotion, and intensity nodes.

- **Text Nodes**. We adopt a pre-trained BERT[1] model to extract the linguistic feature: $f_{u_j} = \mathrm{BERT}(u_j)$.
- **Audio Nodes**. We employ the global style token (GST) (Wang et al. 2018) module, which includes a reference encoder and style token layer, as the audio encoder to extract the acoustic features contained in each audio a_j: $f_{a_j} = \mathrm{GST}(a_j)$.
- **Speaker**, **Emotion**, and **Emotion Intensity Nodes**. The speaker, emotion and intensity encoders are used to define three randomly initialized trainable parameter matrices f_{s_j}, f_{e_j}, and f_{i_j} to learn two speaker identity features, seven emotion label (happiness, sadness, anger, disgust, fear, surprise, neutral) features, and three emotion intensity label (weak, medium, strong) features respectively.

Note that the text and speaker nodes $f_{u_{\mathcal{C}}}$, $f_{i_{\mathcal{C}}}$ of current utterance are initialized in the same way as the nodes f_{u_j} and f_{s_j} in the dialogue history.

ECG Encoding After initializing the constructed Graph, the heterogeneous graph encoding module is used to encode

[1] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

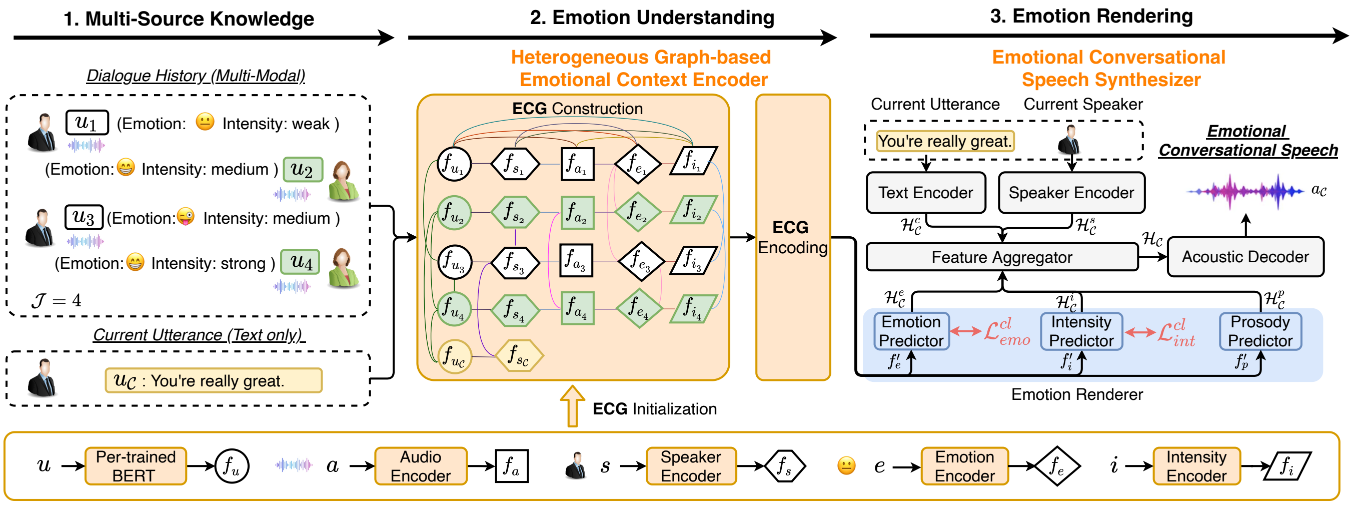
1. Multi-Source Knowledge
2. Emotion Understanding
3. Emotion Rendering
Dialogue History (Multi-Modal)
u_1 (Emotion: 😐 Intensity: weak)
(Emotion: 😁 Intensity: medium) u_2
u_3 (Emotion: 😜 Intensity: medium)
(Emotion: 😁 Intensity: strong) u_4
$\mathcal{J} = 4$
Current Utterance (Text only)
$u_\mathcal{C}$: You're really great.
Heterogeneous Graph-based Emotional Context Encoder
ECG Construction
f_{u_1} f_{s_1} f_{a_1} f_{e_1} f_{i_1}
f_{u_2} f_{s_2} f_{a_2} f_{e_2} f_{i_2}
f_{u_3} f_{s_3} f_{a_3} f_{e_3} f_{i_3}
f_{u_4} f_{s_4} f_{a_4} f_{e_4} f_{i_4}
$f_{u_\mathcal{C}}$ $f_{s_\mathcal{C}}$
ECG Encoding
ECG Initialization
Emotional Conversational Speech Synthesizer
Current Utterance
Current Speaker
You're really great.
Text Encoder
Speaker Encoder
$\mathcal{H}_\mathcal{C}^c$
$\mathcal{H}_\mathcal{C}^s$
Feature Aggregator
$\mathcal{H}_\mathcal{C}$
Acoustic Decoder
Emotional Conversational Speech
$a_\mathcal{C}$
$\mathcal{H}_\mathcal{C}^e$
$\mathcal{H}_\mathcal{C}^i$
$\mathcal{H}_\mathcal{C}^p$
Emotion Predictor
$\mathcal{L}_{emo}^{cl}$
Intensity Predictor
$\mathcal{L}_{int}^{cl}$
Prosody Predictor
f_e'
f_i'
f_p'
Emotion Renderer
u Per-trained BERT f_u
a Audio Encoder f_a
s Speaker Encoder f_s
e Emotion Encoder f_e
i Intensity Encoder f_i

Figure 2: The overall architecture of ECSS. 1) Multi-source knowledge includes text, speaker, audio, emotion, and emotion intensity information of the multi-turn conversation; 2) Heterogeneous graph-based emotional context encoder aims to model the complex dependencies among the multi-source knowledge with the Emotional Conversational Graph (ECG), thus understanding the emotion cues in context; 3) Emotion rendering for CSS seeks to render the accurate emotion state for synthesized speech, in which *Emotion Renderer* aims to infer the emotion feature according to the heterogeneous ECG encoding.

the emotion cues in dialogue context to obtain the graph-enhanced representations for each node. Inspired by Heterogeneous Graph Transformer (HGT) (Hu et al. 2020), we also adopt a three-stage emotional HGT network, that includes *Heterogeneous Mutual Attention* (HMA), *Heterogeneous Message Passing* (HMP), and *Emotional knowledge Aggregation* (EKA) operations, to model the dependencies in emotional conversations. Assuming that one node in the heterogeneous ECG is the target node, then any node of any type can be viewed as the source node. The goal of HGT is to aggregate information from source nodes to get a contextualized representation for the target node.

Firstly, given a target node $\mathcal{N}_{t\hat{g}t}$ and all its neighbor source nodes $\mathcal{N}_{s\hat{r}c}$, the HMA mechanism maps $\mathcal{N}_{t\hat{g}t}$ into a Query vector, and $\mathcal{N}_{s\hat{r}c}$ into a Key vector, and calculate their dot product as attention, that indicates the importance of each source node for the target node. Then we concatenate h attention heads together to get the attention vector for each node pair $\mathcal{E}_{s\hat{r}c \rightarrow t\hat{g}t} = \{\mathcal{N}_{s\hat{r}c}, \mathcal{N}_{t\hat{g}t}\}$ ($\mathcal{E}_{s\hat{r}c \rightarrow t\hat{g}t}$ means the edge between $\mathcal{N}_{s\hat{r}c}$ and $\mathcal{N}_{t\hat{g}t}$). For each target node $\mathcal{N}_{t\hat{g}t}$, we gather all attention vectors from its neighbors $\mathcal{N}_{s\hat{r}c}$ and conduct softmax to get the final attention score.

Secondly, HMP is computed parallel to pass the dependency information from $\mathcal{N}_{s\hat{r}c}$ to $\mathcal{N}_{t\hat{g}t}$. Specifically, for a pair of nodes $\mathcal{E}_{s\hat{r}c \rightarrow t\hat{g}t} = \{\mathcal{N}_{s\hat{r}c}, \mathcal{N}_{t\hat{g}t}\}$, HMP uses a linear projection to project the feature vector of $\mathcal{N}_{s\hat{r}c}$ into a message vector, that then followed by a matrix $W_{s\hat{r}c \rightarrow t\hat{g}t}$ for incorporating the edge dependency. The final step is to concatenate all h message heads to get the final message vector for each node pair.

At last, the EKA module aims to aggregate the computed attention score and message vector. We use the attention score as the weight to average the corresponding messages from all neighbor source nodes $\mathcal{N}_{s\hat{r}c}$ and get the emotion-augmented vector $f_{\mathcal{N}_{t\hat{g}t}}$ for the target node $\mathcal{N}_{t\hat{g}t}$. It's worth noting that all ECG nodes, including emotion and intensity nodes, can be treated as the source or target nodes to learn its high-level contextual information. Therefore, the EKA operation ultimately incorporates the emotional information from the dialog history into all ECG nodes. The graph-enhanced feature representation of each node incorporates the emotional cues in context.

In this way, the ECG encoding module can obtain the final emotion-aware graph-enhanced feature representation $f_u^{'}$, $f_s^{'}$, $f_a^{'}$, $f_e^{'}$ and $f_i^{'}$ for all text, speaker, audio, emotion and intensity nodes respectively, which can be fed into subsequent models to provide emotional information.

Emotional Conversational Speech Synthesizer

As shown in the right panel of Fig. 2, the emotional conversational speech synthesizer consists of the following four components: *Text encoder, Speaker Encoder, Emotion Renderer and Acoustic Decoder*. The text and speaker encoders seek to encode the content and speaker identity features $\mathcal{H}_\mathcal{C}^c$ and $\mathcal{H}_\mathcal{C}^s$ for the current utterance and speaker. Note that the emotion renderer attempts to predict the current utterance's emotion, intensity, and prosody features $\mathcal{H}_\mathcal{C}^e$, $\mathcal{H}_\mathcal{C}^i$, and $\mathcal{H}_\mathcal{C}^p$ using the graph-enhanced node features. To obtain a robust feature representation of the current utterance, the feature aggregator module set a set of trainable weight parameters for the above five features and then output the final mixup feature $\mathcal{H}_\mathcal{C}$. For the acoustic decoder, we use *FastSpeech2* (Ren et al. 2021) as the backbone, which includes the variance adaptor, mel decoder and vocoder. The variance adapter takes the $\mathcal{H}_\mathcal{C}$ as inputs to predict the duration, energy, and pitch. The Mel Decoder aims to predict the mel-spectrum features. Finally, a well-trained HiFi-GAN (Kong, Kim, and

Bae 2020) is used as the vocoder to generate speech waveform $a_{\mathcal{C}}$ with desired emotion style.

Note that to achieve emotion rendering, our emotion renderer extracts the encoded node features from ECG and predicts the emotion and emotion intensity feature representations of the current utterance while performing the prosody prediction. More importantly, to achieve accurate emotion category and emotion intensity feature representation prediction, we propose contrastive learning-based emotion and emotion intensity loss functions $\mathcal{L}^{cl}$.

Emotion Renderer As shown in the blue part of Fig. 2, the emotion renderer consists of emotion, intensity and prosody predictors. The prosody predictor adopts a multi-head attention layer to infer the speaking prosody information of the current utterance from the feature representations of text nodes in the dialogue history. The features of audio nodes are not used because we believe that audio and text modalities have already interacted during the ECG encoding, thus the text node already contains audio information. We use the MSE loss for the prosody predictor to constrain its training, where the target is obtained from a GST-based prosody extractor. Next, we will introduce emotion and intensity predictors.

- **Emotion Predictor** uses the encoded features of the emotion nodes in dialog history to infer the emotion representation $\mathcal{H}^e_{\mathcal{C}}$ of the current sentence. It includes two convolutional layers, a bidirectional LSTM layer, and two fully connected layers.

$$\mathcal{H}^e_{\mathcal{C}} = \mathrm{FC}(\mathrm{BiLSTM}(\mathrm{CNN}(f^{'}_{e}))) \tag{1}$$

where $f^{'}_{e}$ represents all emotion-type nodes in the dialogue history after ECG encoding.

- **Intensity Predictor** uses the encoded features of the intensity nodes in dialog history to infer the emotion intensity representation $\mathcal{H}^i_{\mathcal{C}}$ of the current sentence. It consists of two convolutional layers, a bidirectional LSTM layer, two fully connected layers, and a mean pooling layer.

$$\mathcal{H}^i_{\mathcal{C}} = \mathrm{AvgPooling}(\mathrm{FC}_2(\mathrm{BiLSTM}(\mathrm{CNN}_2(f^{'}_{i})))) \tag{2}$$

where $f^{'}_{i}$ is a universal representation of all emotion intensity nodes in the dialog history after ECG encoding.

Contrastive Learning Training Criterion For emotion and intensity predictors of the emotion renderer, inspired by (Khosla et al. 2020), we design the emotion-supervised contrastive learning losses $\mathcal{L}^{cl}$ to motivate the emotion renderer to better distinguish different emotions categories and intensity degrees. Specifically, contrastive learning loss $\mathcal{L}^{cl}_{emo}$ for emotion category and $\mathcal{L}^{cl}_{int}$ for emotion intensity share the same spirits, that is treating all examples with the same emotion category or intensity label in the batch as positive examples while different labels as negative.

For the emotion feature $\mathcal{H}^e_{\mathcal{C}}$, a batch of K emotion representations is denoted as $\mathcal{H}^K = [\mathcal{H}^e_{\mathcal{C}1}, \mathcal{H}^e_{\mathcal{C}2}, ..., \mathcal{H}^e_{\mathcal{C}K}]$, $\mathcal{L}^{cl}_{emo}$ for $\mathcal{H}^e_{\mathcal{C}k}$ as follows,

$$\mathcal{L}^{cl}_{emo} = log\frac{-1}{|\mathcal{P}(k)|}\frac{\sum\limits_{\mathcal{H}^e_{\mathcal{C}q}\in\mathcal{P}(k)} exp(sim(\mathcal{H}^e_{\mathcal{C}k}, \mathcal{H}^e_{\mathcal{C}q})/\tau)}{\sum\limits_{\mathcal{H}^e_{\mathcal{C}d}\in B(k)} exp(sim(\mathcal{H}^e_{\mathcal{C}k}, \mathcal{H}^e_{\mathcal{C}d})/\tau)} \tag{3}$$

where $sim(\cdot,\cdot)$ is a cosine similarity function. τ is a scalar temperature parameter. $B(k) \equiv \mathcal{H}^K \backslash \{\mathcal{H}^e_{\mathcal{C}k}\}$ contains all representations in $\mathcal{H}^K$ except $\mathcal{H}^e_{\mathcal{C}k}$. $\mathcal{P}(k)$ is the set of positive samples that have the same emotion label with $\mathcal{H}^e_{\mathcal{C}k}$ in a batch. Similarity, $\mathcal{L}^{cl}_{int}$ for $\mathcal{H}^i_{\mathcal{C}k}$ as follows,

$$\mathcal{L}^{cl}_{int} = log\frac{-1}{|\mathcal{P}(k)|}\frac{\sum\limits_{\mathcal{H}^i_{\mathcal{C}q}\in\mathcal{P}(k)} exp(sim(\mathcal{H}^i_{\mathcal{C}k}, \mathcal{H}^i_{\mathcal{C}q})/\tau)}{\sum\limits_{\mathcal{H}^i_{\mathcal{C}d}\in B(k)} exp(sim(\mathcal{H}^i_{\mathcal{C}k}, \mathcal{H}^i_{\mathcal{C}d})/\tau)} \tag{4}$$

At last, the total loss function $\mathcal{L}$ is: $\mathcal{L} = \mathcal{L}^{cl}_{emo} + \mathcal{L}^{cl}_{int} + \mathcal{L}^{mse}_{pro} + \mathcal{L}_{fs2}$, where $\mathcal{L}_{fs2}$ refers to the acoustic feature loss of traditional FastSpeech2, including pitch, energy, and duration, as well as the mel spectrum, $\mathcal{L}^{mse}_{pro}$ indicates the MSE loss for prosody predictor.

Experiments and Results

Dataset We validate the ECSS on a recently public dataset for conversational speech synthesis called DailyTalk (Lee, Park, and Kim 2023), that consists of 23,773 audio clips representing 20 hours in total, in which 2,541 conversations were sampled, modified, and recorded. We partition the data into training, validation, and test sets at a ratio of 8:1:1.

To obtain the multi-source knowledge, we invited a professional practitioner to perform fine-grained labeling of DailyTalk data for emotion category and intensity labels while listening to speech and understanding the semantics of utterance. The final distribution of the data is as follows, the number of each of the 7 emotion category labels (happy, sad, angry, disgust, fear, surprise, neutral) are 3871, 722, 226, 186, 74, 497 and 18,197, and the number of each of the 3 emotion intensity labels (weak, medium, strong) are 19,973, 3,646 and 154.

Experimental Setup In the heterogeneous graph-based emotion context encoder, the dimension of the text node representation f_{u_j} is set to 512, and the dimensions of the remaining type node representations f_{e_j},f_{i_j},f_{s_j}, and f_{a_j} are all set to 256. For multi-head attention-based methods, we set the head number as 8. For the emotion predictor of emotion renderer, the convolutional layer has a convolutional kernel of 3, the LSTM input dimension is 384, the hidden state size is 256, and the forward and backward outputs from the last time steps of the LSTM are spliced using concat before going into the linear layer. For the intensity predictor of emotion renderer, the mean pooling layer convolution kernel size is 2. For the prosody predictor of emotion renderer, the dimensions of the $Query$, Key and $Value$ and output feature $\mathcal{H}^p_{\mathcal{C}}$ are 512, 384, 384 and 256. The acoustic decoder is configured with reference to FastSpeech2 (Ren et al. 2021).

Systems	**N-DMOS (↑)**	**E-DMOS (↑)**	**MAE-M (↓)**	**MAE-P (↓)**	**MAE-E (↓)**	**MAE-D (↓)**
No emotional context modeling	3.232 ± 0.030	3.100 ± 0.026	0.681	0.506	0.346	0.300
GRU-based	3.314 ± 0.022	3.288 ± 0.011	0.675	0.506	0.352	0.296
Homogeneous Graph-based	3.384 ± 0.027	3.493 ± 0.031	0.662	0.456	**0.204**	**0.150**
ECSS (Proposed)	**3.506 ± 0.022**	**3.619 ± 0.028**	**0.654**	**0.455**	**0.215***	**0.152***
w/o emotion	3.424 ± 0.018	3.496 ± 0.026	0.658	0.467	0.224	**0.150**
w/o intensity	3.487 ± 0.035	3.523 ± 0.024	0.660	**0.453**	0.210	0.157
w/o speaker	3.401 ± 0.019	3.511 ± 0.027	0.666	0.456	0.231	0.152
w/o audio	3.391 ± 0.023	3.505 ± 0.022	0.656	0.457	**0.198**	0.154
w/o $\mathcal{L}^{cl}$	3.388 ± 0.025	3.312 ± 0.031	0.665	0.459	0.222	0.156

Table 1: Subjective (with 95% confidence interval) and objective results with different CSS systems. (* means the metric value achieved the suboptimal result among all systems.)

we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$. For text input, we adopt Grapheme-to-Phoneme (G2P) toolkit[2] to convert all text into its phoneme sequence. All speech samples are re-sampled to 22.05 kHz. The mel-spectrum features are extracted with a window length of 25ms and a shift of 10ms. The model is trained on a Tesla V100 GPU with a batch size of 16 and 600k steps. During ECSS training, the context or dialogue history length is set to 10. More detailed experimental settings are accessed in the *Appendix* section.

Evaluation Metrics For the subjective evaluation metrics, we organized a dialogue-level Mean Opinion Score (DMOS) (Streijl, Winkler, and Hands 2016) listening test with 30 master students whose second language is English and provided specialized training on the rules to all listeners. Given the dialogue history, they were asked to rate the naturalness DMOS (N-DMOS) and emotional DMOS (E-DMOS) of the synthesized speech of the current utterance, ranging from 1 to 5. Each listener was access to 50 audio samples. Note that the N-DMOS focuses on the speaking prosody, while the E-DMOS focuses on the richness of the emotional expression of the current utterance and whether it matches the emotional expression of the context.

For the objective evaluation metrics, we calculate the Mean Absolute Error (MAE) between the predicted and real acoustic features to assess the emotional expressiveness of the synthesized audio. Specifically, we assess the acoustic feature in terms of mel-spectrum, pitch, energy, and duration with MAE-M, MAE-P, MAE-E and MAE-D. In addition, we conduct a visualization study with the help of a third-party Speech Emotion Recognition (SER) model, that was used to identify the emotion categories of the synthesized emotional conversational speech. We plot the confusion matrix to validate the ECSS.

Comparative Study To demonstrate the effectiveness of our ECSS, we employ three advanced approaches which also employ FastSpeech2 as the TTS backbone as the baseline systems.

- **No emotional context modeling**. The first baseline approach is a vanilla FastSpeech2 (Ren et al. 2021) with no context modeling, which is also representative of state-of-the-art non-conversational TTS systems.
- **GRU-based context modeling**. This method involves only text modality and uses an RNN-based unidirectional GRU network to model contextual dependencies in a dialogue sequentially (Guo et al. 2020).
- **Homogeneous Graph-based emotion context modeling**. In the homogeneous graph-based approach (Li et al. 2022b), each past utterance in the conversation is represented as a node in the graph. Each node is initialized with the corresponding multi-modal features. To achieve emotion rendering using this method, we can extract the graph-enhanced feature of all nodes in dialogue history to predict the emotion and intensity information.

Main Results As shown in the first five rows of Table 1, the ECSS achieves state-of-the-art performance on average, which obtains the optimum results in MAE-M (0.654) and MAE-P (0.455), and suboptimal results in MAE-E (0.215) and MAE-D (0.152). However, objective experiments may not fully reflect human feelings. By observing the subjective results, the proposed ECSS model outperforms all baselines with an N-DMOS score of 3.506 and an E-DMOS score of 3.619, which reflects the superiority of our ECSS. ECSS contributes to adequate emotion understanding based on heterogeneous graph context modeling, in which the emotion renderer fully mines the emotion cues to infer the emotion expression state of the current sentence, thus achieving satisfactory emotion rendering effects for conversational speech synthesis.

Ablation Results To evaluate the individual effects of various heterogeneous nodes, including emotion, intensity, speaker and audio, in ECG and the contrastive learning loss $\mathcal{L}^{cl}$ for emotion renderer, we remove these components to build various systems and conduct a series of ablation experiments, and the subjective and objective results are shown in the rows 6 through 10 of Table 1.

We can find that removing different types of nodes in the heterogeneous ECG brought about a decrease in objective metrics performance in the vast majority of metrics, and the subjective DMOS scores also showed a drop. For example, after removing the emotion node, the MAE-M, MAE-P and MAE-E decreased by 0.004, 0.012 and 0.009 respectively, while N-DMOS and E-DMOS decreased by 0.082 and 0.123. This suggests that our heterogeneous graph nodes can learn the complex emotional dependencies in the dialog

[2] https://www.github.com/kyubyong/g2p

Length	MAE-M	MAE-P	MAE-E	MAE-D
2	0.667	0.468	0.233	0.160
3	0.664	0.461	0.229	0.157
6	0.662	0.456	0.223	0.155
9	0.658	**0.453**	0.219	0.153
10	**0.654**	**0.455***	**0.215**	**0.152**
13	0.660	0.456	0.220	0.154
14	0.661	0.457	0.221	0.153

Table 2: Objective results of various context lengths. Lower values mean better performance.

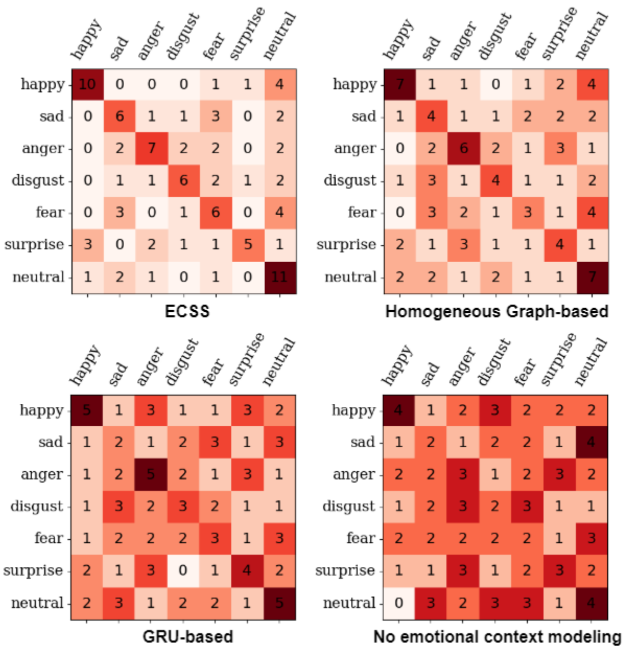

Figure 3: The confusion matrix results about the emotion category rendering with the help of speech emotion recognition. The X-axis and Y-axis of subfigures represent perceived and true emotion categories.

history and achieve full emotional understanding. In addition, to validate the $\mathcal{L}^{cl}$, we replace it with the cross-entropy loss. As shown in the last row of Table 1, all subjective and objective values are reduced after removing the $\mathcal{L}^{cl}$. This demonstrates that the contrastive learning strategy allows the emotion renderer to distinguish between different emotion categories and intensities better.

Context Length Analysis We also explore the effectiveness of emotional context modeling with different context lengths. Specifically, considering the average number 9.3 of dialogue turns in the DailyTalk, we set the utterance length of dialogue history ranging from 2 to 14 to compare the objective performance. As shown in Table 2, from a general view, all values decrease when enlarging the context length from 2 to 10, and increase from 10 to 14. This shows that either insufficient or redundant context information will interfere the understanding of emotion cues in context.

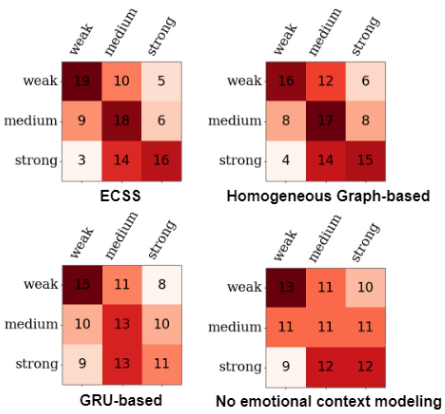

Figure 4: The confusion matrix results about the emotion intensity rendering. The X-axis and Y-axis of subfigures represent perceived and true intensity categories.

Visualization Study To demonstrate the emotional expressiveness of synthesized speech more visually, we employ a pre-trained SER model[3] to identify the emotion category of 400 audio samples synthesized from the ECSS and all baselines, respectively. As shown in Fig. 3, we plot the confusion matrices to show the gap between the different systems. It can be seen that our ECSS outperforms all baselines by presenting a clear diagonal line in the confusion matrix. It proves that the emotional conversational speech synthesized by ECSS performs a clear emotional expression. In addition, we further conduct a visualization study to validate the emotion intensity rendering. Specifically, five listeners were invited and asked to rate the emotion intensity labels of 400 audios. As shown in Fig. 4, the confusion matrix suggests that the emotional conversational speech synthesized by ECSS performs a clear emotional intensity expression. The above results also provide further evidence that our ECSS draws on the contextual modeling capabilities of heterogeneous graphs and the effective constraints of contrastive learning.

Conclusion

To improve the emotion understanding and rendering in CSS systems, we present a novel ECSS model whereby a heterogeneous Emotional Conversational Graph (ECG) armed with multi-source knowledge in context is used for emotional context modeling, and the emotion renderer with contrastive learning constraint to achieve accurate emotional style inference. Experimental results demonstrate the superiority of ECSS over state-of-the-art CSS systems. The contribution of heterogeneous nodes in ECG and emotion renderer are further demonstrated in ablation studies. To the best of our knowledge, ECSS is the first in-depth conversational speech synthesis study that models emotional expressiveness. We hope that our work will serve as a basis for future emotional CSS studies.

[3] https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition

Appendix. Detailed Experimental Settings

We list the detailed model setup of ECSS. The detailed experimental settings of the *Heterogeneous Graph-based Emotional Context Encoder* and *Emotional Conversational Speech Synthesizer* are shown in Table 3.

ECSS Model Setup		
Heterogeneous ECG	Text Node Representation	512
	Emotion,Intensity, Speaker and Audio Node Representation	256
	HGTConv Hidden Channels	384
	HGTConv Head Number	2
	HGTConv Layer	1
Emotion Predictor of Emotion Renderer	ConvBlock2D Kernel	3
	ConvBlock2D Layer	2
	LSTM Input Dimension	384
	LSTM Hidden State Size	256
	Emotion Predictor Output Dimension	7
Intensity Predictor of Emotion Renderer	ConvBlock2D Kernel	3
	ConvBlock2D Layer	2
	LSTM Input Dimension	384
	LSTM Hidden State Size	256
	AvgPool2d Kernel	2
	Output Dimension	3
Prosody Predictor of Emotion Renderer	Query Dimension	512
	Key and Value Dimension	384
	Prosody Predictor Head Number	2
	Output Dimension	256
Prosody Extractor	Reference Encoder Mel Channels	80
	Reference Encoder GRU Size	128
	Style Token Embedding Size	256
	Style Token Layer GSTs	7
Text Encoder	Word Vector Dimension	256
	Position Encoding Hidden	256
	Transformer FFTBlock Layer	4
	Transformer FFTBlock Head	2
	Transformer FFTBlock Dropout	0.2
Speaker Encoder	Speaker Embedding Size	256
Feature Aggregator	Feature Aggregator Output Dimension	256
Acoustic Decoder	Word Vector Dimension	256
	Position Encoding Hidden	256
	Transformer FFTBlock Layer	6
	Transformer FFTBlock Head	2
	Transformer FFTBlock Dropout	0.2
	Postnet Embedding Dimension	512
	Postnet Kernel	5

Table 3: Detailed Model Setup of ECSS.

Acknowledgments

The research by Rui Liu was funded by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62206136), and Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B1212010004). The research by Yifan Hu was funded by the Research and Innovation Projects for Graduate Students in Inner Mongolia Autonomous Region (No. 11200-121028). The work by Haizhou Li was supported by the National Natural Science Foundation of China (Grant No. 62271432), and Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002).

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Dias, W.; Andalo, F.; Padilha, R.; Bertocco, G.; Almeida, W.; Costa, P.; and Rocha, A. 2022. Cross-dataset emotion recognition from facial expressions through convolutional neural networks. *Journal of Visual Communication and Image Representation*, 82: 103395.

Firdaus, M.; Chauhan, H.; Ekbal, A.; and Bhattacharyya, P. 2020. MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4441–4453. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164. Hong Kong, China: Association for Computational Linguistics.

Goel, R.; Susan, S.; Vashisht, S.; and Dhanda, A. 2021. Emotion-aware transformer encoder for empathetic dialogue generation. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 1–6. IEEE.

Guo, H.; Zhang, S.; Soong, F. K.; He, L.; and Xie, L. 2020. Conversational End-to-End TTS for Voice Agent. *CoRR*, abs/2005.10438.

Hu, Z.; Dong, Y.; Wang, K.; and Sun, Y. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, 2704–2710.

Huang, R.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2022. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in Neural Information Processing Systems*, 35: 10970–10983.

Im, C.-B.; Lee, S.-H.; Kim, S.-B.; and Lee, S.-W. 2022. EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6317–6321.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *CoRR*, abs/2004.11362.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33: 17022–17033.

Lee, K.; Park, K.; and Kim, D. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Li, J.; Meng, Y.; Li, C.; Wu, Z.; Meng, H.; Weng, C.; and Su, D. 2022a. Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7917–7921. IEEE.

Li, J.; Meng, Y.; Wu, X.; Wu, Z.; Jia, J.; Meng, H.; Tian, Q.; Wang, Y.; and Wang, Y. 2022b. Inferring speaking styles from multi-modal conversational context by multi-scale relational graph convolutional networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5811–5820.

Li, T.; Wang, X.; Xie, Q.; Wang, Z.; and Xie, L. 2022c. Cross-Speaker Emotion Disentangling and Transfer for End-to-End Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 1448–1460.

Li, Z.; Tu, G.; Liang, X.; and Xu, R. 2022d. Developing Relationships: A Heterogeneous Graph Network with Learnable Edge Representation for Emotion Identification in Conversations. In *CAAI International Conference on Artificial Intelligence*, 310–322. Springer.

Liu, R.; Hu, Y.; Zuo, H.; Luo, Z.; Wang, L.; and Gao, G. 2024. Text-to-Speech for Low-Resource Agglutinative Language With Morphology-Aware Language Model Pre-Training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 1075–1087.

Liu, R.; Sisman, B.; Bao, F.; Yang, J.; Gao, G.; and Li, H. 2021a. Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 274–285.

Liu, R.; Sisman, B.; Gao, G.; and Li, H. 2021b. Expressive TTS Training With Frame and Style Reconstruction Loss. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1806–1818.

Liu, R.; Sisman, B.; Gao, G.; and Li, H. 2022a. Decoding Knowledge Transfer for Neural Text-to-Speech Training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 1789–1802.

Liu, R.; Sisman, B.; Schuller, B.; Gao, G.; and Li, H. 2022b. Accurate emotion strength assessment for seen and unseen speech based on data-driven deep learning. *arXiv preprint arXiv:2206.07229*.

McTear, M. 2022. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *CoRR*, abs/1810.02508.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Saganowski, S.; Komoszyńska, J.; Behnke, M.; Perz, B.; Kunc, D.; Klich, B.; Kaczmarek, Ł. D.; and Kazienko, P. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data*, 9(1): 158.

Seaborn, K.; Miyake, N. P.; Pennefather, P.; and Otake-Matsuura, M. 2021. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4): 1–43.

Song, R.; Giunchiglia, F.; Shi, L.; Shen, Q.; and Xu, H. 2023. SUNET: Speaker-utterance interaction Graph Neural Network for Emotion Recognition in Conversations. *Engineering Applications of Artificial Intelligence*, 123: 106315.

Streijl, R. C.; Winkler, S.; and Hands, D. S. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2): 213–227.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. *Proc. Interspeech 2017*, 4006–4010.

Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.-S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; and Saurous, R. A. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, 5180–5189. PMLR.

Xue, J.; Deng, Y.; Wang, F.; Li, Y.; Gao, Y.; Tao, J.; Sun, J.; and Liang, J. 2023. M 2-CTTS: End-to-End Multi-Scale Multi-Modal Conversational Text-to-Speech Synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhong, P.; Wang, D.; and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *CoRR*, abs/1909.10681.

Zhou, K.; Sisman, B.; Rana, R.; Schuller, B. W.; and Li, H. 2023. Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Transactions on Affective Computing*, 14(1): 31–48.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1): 53–93.

Zuo, H.; Liu, R.; Zhao, J.; Gao, G.; and Li, H. 2023. Exploiting Modality-Invariant Feature for Robust Multimodal Emotion Recognition with Missing Modalities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.