Inferring Speaking Styles from Multi-modal Conversational Context by Multi-scale Relational Graph Convolutional Networks

Jingbei Li*
Tsinghua University
lijb19@mails.tsinghua.edu.cn

Yi Meng*
Tsinghua University
my20@mails.tsinghua.edu.cn

Xixin Wu†
The Chinese University of Hong Kong
wuxx@se.cuhk.edu.hk

Zhiyong Wu†
Tsinghua University
zywu@sz.tsinghua.edu.cn

Jia Jia
Tsinghua University
jjia@tsinghua.edu.cn

Helen Meng
The Chinese University of Hong Kong
hmmeng@se.cuhk.edu.hk

Qiao Tian
ByteDance
tianqiao.wave@bytedance.com

Yuping Wang
ByteDance
wangyuping@bytedance.com

Yuxuan Wang
ByteDance
wangyuxuan.11@bytedance.com

ABSTRACT

To support applications of speech-driven interactive systems in various conversational scenarios, text-to-speech (TTS) synthesis needs to understand the conversational context and determine appropriate speaking styles in its synthesized speeches. These speaking styles are influenced by the dependencies between the multi-modal information in the context at both global scale (i.e. utterance level) and local scale (i.e. word level). However, the dependency modeling and speaking style inference at the local scale are largely missing in state-of-the-art TTS systems, resulting in the synthesis of incorrect or improper speaking styles. In this paper, to learn the dependencies in conversations at both global and local scales and to improve the synthesis of speaking styles, we propose a context modeling method which models the dependencies among the multi-modal information in context with multi-scale relational graph convolutional network (MSRGCN). The learnt multi-modal context information at multiple scales is then utilized to infer the global and local speaking styles of the current utterance for speech synthesis. Experiments demonstrate the effectiveness of the proposed approach, and ablation studies reflect the contributions from modeling multi-modal information and multi-scale dependencies.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Information extraction**; **Neural networks**; • **Human-centered computing** → Human computer interaction (HCI).

KEYWORDS

speech interaction system, conversational speech synthesis, speaking style, context modeling, multi-scale graph convolution network

*Both authors equally contributed to this research. †Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3547831

ACM Reference Format:
Jingbei Li, Yi Meng, Xixin Wu, Zhiyong Wu, Jia Jia, Helen Meng, Qiao Tian, Yuping Wang, and Yuxuan Wang. 2022. 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 (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547831

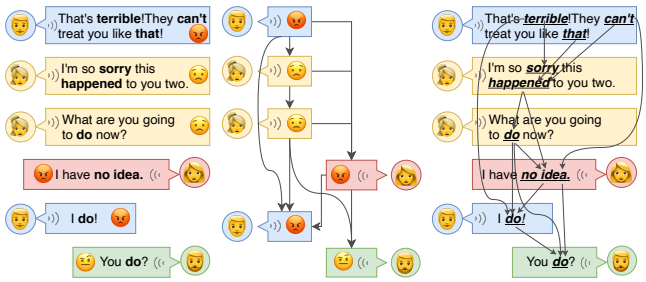
1 INTRODUCTION

With the development of deep learning and speech processing technologies [13, 25–27, 30, 37], speech-driven interactive systems, such as virtual assistants and voice agents are becoming increasingly pervasive in real-world applications [9, 20, 33, 42] which present a diversity of conversational scenarios.

In human-human conversations, people are interacting with different social signals such as humor, empathy, compassion and affect [36] through the contents and the speaking styles of their speeches, where the speaking style(s) of each utterance are dependent on the conversational context encoded in multi-modal forms and in multiple scales. As illustrated by the example[1] in Figure 1, the conversational context is carried multimodally in both acoustic and textual forms across the intentions, attitudes and emotions of the speaker, and also at multiple scales – including the global scale (i.e. utterance level) and the local scale (i.e. word level). Moreover, there is always a major interplay among the dependencies in the conversations [7]. Apart from the temporal dependencies and the inter- and intra-speaker dependencies [7], there are also speaking style dependencies among utterances playing critical roles in the conversations. More specifically, such dependencies are encoded not only at the global scale (such as the utterance level emotions and attitudes), but also at the local scale (such as the word level emphasis and prosody). In human-computer conversations, modeling such dependencies of the speaking style(s) on the conversational context at both global and local scales is crucial for the speech synthesis system to generate speech with the appropriate styles and improve the user experience in speech based interactions.

Recent speech interaction systems have successfully improved the global speaking style in synthesized speech by modeling the dependencies in conversations at the global, utterance level [2, 8, 18].

[1]Video is available at https://thuhcsi.github.io/mm2022-conversational-tts/.



(a) An example conversation

(b) Dependencies at global scale (utterance level)

(c) Dependencies at local scale (word level)

Figure 1: Graphical depiction of a spoken conversation and the illustrations of its inside dependencies at multiple scales.

A conversational context encoder [8] is proposed to sequentially process the textual information in context through a uni-directional gated recurrent unit (GRU) [1] based recurrent neural network (RNN). A conversational TTS system with multi-modal context modeling [18] is further proposed and employs the dialogue graph convolutional network (DialogueGCN) [7] to explicitly model the temporal dependency and inter- and intra-speaker dependencies from the global textual and acoustic information in the context.

However, dependency modeling at the local, word-level is largely missing and the controllability of synthesis style at this scale is also lacking. It should be noted that for synthesis of global speaking styles, the multi-modal information of each utterance is modeled as a fixed-length vector, which cannot be easily used to also model the multi-modal information at the word-level for the local speaking styles. Improper use of local speaking style could seriously affect the user experience of speech-driven interactive systems.

To learn the multi-scale dependencies in conversations and to improve both global and local speaking styles in conversational TTS, we propose a context modeling method for the multi-modal information in context using multi-scale relational graph convolutional network (MSRGCN)[2]. The attention mechanism [35] is used to summarize the context at both global and local scales, according to the multi-scale textual features of the current utterance. The multi-scale context information are then used to infer the speaking styles of the current utterance and synthesize to corresponding speech with FastSpeech 2 [26].

2 RELATED WORK

The dependencies in conversations have been studied in both natural language processing (NLP) and speech processing. In NLP, these dependencies are modeled in many conversation-related researches such as turn-based question answering and response generation [6, 12, 16, 23] and long-term conversational emotion recognition [7, 17, 24, 31]. Previous researches on turn-based conversations only model the dependencies between the question and answer in a single dialogue turn and lacks the modeling on long-term context. In conversational emotion recognition, the DialogueRNN [24] is proposed to use a global GRU and a party GRU to model the dependencies between the global conversational states and the speakers. The global GRU models the dependency at the global scale and provides past conversational states to the party GRU. The party GRU jointly considers the current utterance and these past conversational states to model the dependencies between the speakers and the global states. The outputs of party GRU are further back ported to the global GRU to generate new global state. However, the speaker information of each utterance is not considered in the party GRU, which is crucial for modeling the inter- and intra-speaker dependencies in conversations. The DialogueGCN [7] is then proposed to construct a directed graph to represent the inter- and intra-speaker information flows in conversations and further model the dependencies between or inside different speakers. However, such inter- and intra-speaker flows are only represented at the global scale. The dependencies at local scale, such as the dependency between each pair of words in conversations, are not considered. Besides, the multi-modal information other than the textual information are rarely considered in NLP.

In speech processing, there are studies that jointly consider the multi-modal information in conversations such as multi-modal conversational emotion recognition [10, 11, 21, 22, 28, 32, 40] and conversational speech synthesis [2, 8, 18]. A uni-directional GRU-based conversational context encoder [8] is proposed to embed the global textual information in the past and current utterances into a fixed-length context embedding. Though this approach brings context modeling to conversational TTS, the acoustic information which is also vital to understanding the conversation and generating proper speaking styles are not captured. A context acoustic encoder [2] which encodes the global speaking style of the previous utterance is proposed and used to synthesize speech with spontaneous behaviors for the current utterance. A multimodal DialogueGCN based conversational TTS system [18] is then proposed to further consider the temporal and speaker dependencies to improve the synthesis of the global speaking styles, in which however the local details of speaking styles are ignored.

Compared with the previous works, the contributions of our work include: (1) this is the first attempt to jointly model the dependencies among the utterances and words in conversations at multiple scales, where we advocate modeling of local-scale dependencies between the words within the same and across different utterances to significantly improve the understanding of conversational context; (2) we present a novel framework based on MSRGCN for context modeling in conversations, which outperforms other state-of-the-art context modeling approaches; (3) we propose an effective approach to infer the speaking styles at both global and local scales for conversational speech synthesis from the multi-modal conversational context.

3 DATA OBSERVATION

We first conduct subjective observations to analyze the dependencies in conversations at multiple scales. We use the English conversation corpus (ECC) [18] for observation. We randomly pick 25 conversations in ECC, in which each conversation consists of 5 past utterances and a current utterance. 25 listeners are invited to listen to the selected conversations and answer the following questions

[2]Source code is available at https://github.com/thuhcsi/mm2022-conversational-tts.

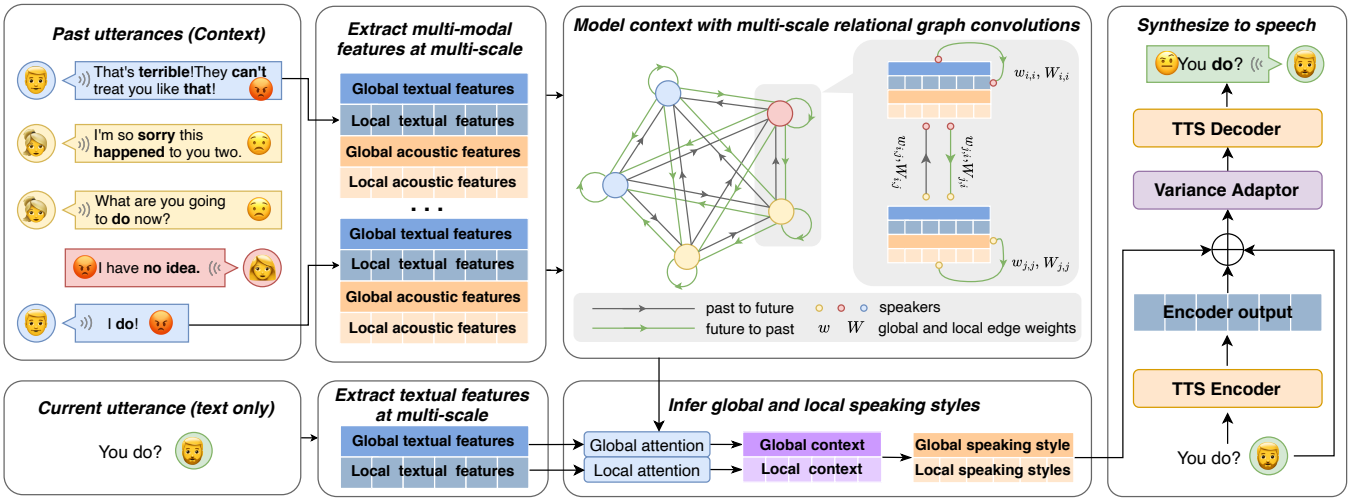


Figure 2: Inferring speaking styles in conversations from multi-modal information at multiple scales with MSRGCN.

for each conversation: (1) Is there a word that has a noticeably different speaking style from those of the other words in the current utterance? If so, which one? (2) Is the global speaking style of the current utterance dependent on a particular past utterance? (3) If there is a word that has a noticeably different speaking style in the current utterance, is this speaking style dependent on one or more particular words in the past utterances? If so, which one(s)?

The first question aims to demonstrate the need to further infer the local speaking style than just using the global speaking style for conversational speech synthesis. The aggregated responses from the listeners (by averaging) indicate that 55.0% of the current utterances contain at least one word with noticeably different speaking style(s) compared to the other words in the utterance. This reflects that local speaking styles at the word level are frequently used and important in conversations.

The second question aims to verify the dependency between the current utterance and the past utterances at the global scale. The averaged responses to the second question show that 70.7% of the current utterances have global speaking styles dependent on a previous utterance in the context, and this reflects the importance of modeling the synthesis style globally at the utterance level.

The third question is designed to demonstrate the local-scale dependency between the words in the current utterance and those in the past utterances. Averaging the responses shows that for the utterances that have noticeably different speaking styles at word level, 88.5% of them are dependent on previous words in the conversational context. If we further divide these dependent word pairs into different groups according to their word types (namely, nouns or proper nouns, verbs, adjectives, adverbs, and others), we find that 55.21% of these dependent word pairs consist of words from the same group.

4 PROBLEM FORMULATION

A conversation can be defined as a sequence of utterances $u_1, \ldots, u_i, \ldots, u_n, u_{n+1}$, where each utterance u_i consists of $(text_i, speaker_i, speech_i)$. The task of conversational TTS aims to synthesize the $speech_{n+1}$ given the $text_{n+1}$ and $speaker_{n+1}$ of the current utterance and the past utterances $u_1, \ldots, u_n$. Particularly, the speaking style of the synthesized $speech_{n+1}$ should confirm to the conversational context characterized by the utterances $u_1, \ldots, u_n$.

Multi-modal features are extracted at the global and local scales for each past utterance u_i, which will be further used for context modeling. The multi-modal features at the global scale g_i and those at the local scale s_i include the textual and acoustic features at utterance and word levels respectively.

To model the dependencies in conversations at both global and local scales, a context modeling method f_{CM} is required to generate conversational representations at the global scale g'_i and at the local s'_i from the multi-modal features at the multiple scales, g_i and s_i.

Specifically, f_{CM} needs to consider: **(1)** the temporal dependencies between each pair of utterances, including past-to-future and future-to-past; **(2)** the dependencies between different speakers and inside each speaker; **(3)** the dependencies between the multi-modal information in utterances; and **(4)** the above dependencies at both global and local scales.

The learnt conversational representations contain rich information which models the above dependencies at both scales than the multimodal features of each utterance. Such conversational representations could be used in many conversation-related researches, such as conversational emotion analysis and conversational TTS.

5 METHODOLOGY

5.1 Multi-scale multi-modal feature extraction

Multi-modal features are first extracted at both global and local scales for each past utterance in the conversational context. Details of the feature extraction module are shown in Figure 3. The textual features are extracted for each word in the utterance by a pretrained BERT [3, 39] model. The acoustic features include the global and local speaking styles extracted by the global and local speaking style encoders respectively.

Inspired by the reference encoder [34] and the global style token (GST) attention layer [38], the global speaking style encoder consists

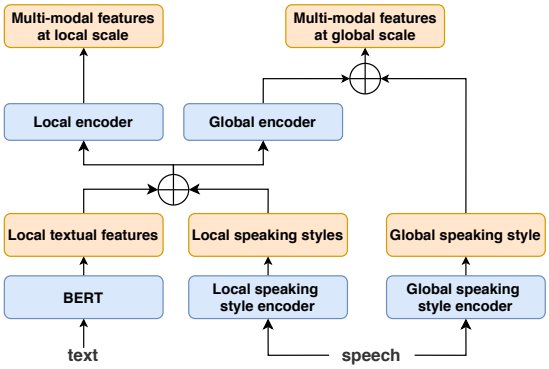


Figure 3: Multi-scale multi-modal feature extraction.

of 6 strided convolutional neural networks (CNNs) composed of 3×3 kernels with 32, 32, 64, 64, 128, 128 filters respectively and 2×2 stride, a 256 dimensional GRU layer and a 128 dimensional style attention layer. The mel-spectrograms of the input speech are first processed by CNNs and GRU. The final state of GRU is further sent to the style attention layer to derive the global speaking style vector GST as the weights for 10 automatically learnt base global speaking style embeddings. The process can be formulated as:

$$q = final(GRU(CNN(speech))) \tag{1}$$

$$GST = softmax\left(q^T GST^{table}\right) \tag{2}$$

where $final$ returns the final state of GRU, q is the query for the style attention layer, GST^{table} contains the 10 automatically learnt base global speaking style embeddings.

The architecture of the local speaking style encoder is same to the global speaking style encoder, except the stride is now 1×2, and the GRU layer now returns the output for each input frame. The outputs of GRU are then summarized for each word in the utterance by multiplying them with the speech-to-text attention weights extracted from a pretrained neural network based forced aligner (NeuFA) [19]. NeuFA employs bidirectional attention mechanism [19] to learn the bidirectional information mapping between a pair of text and speech. The learnt attention weights at the speech-to-text direction could be used to summarize the frame-level information for each word in the utterance, deriving the local speaking style sequence LST. The process of local reference encoder can be formulated as:

$$q' = W_{ASR}^T GRU(CNN(speech)) \tag{3}$$

$$LST = softmax\left(q'^T LST^{table}\right) \tag{4}$$

where W_{ASR} is the attention weights at the speech-to-text direction obtained by NeuFA, q' is the query for the local style attention layer, LST^{table} contains 10 automatically learnt base local speaking style embeddings.

The local textual features and local speaking styles are further concatenated and encoded by a local encoder into the multi-modal features at local scale for each past utterance. The local encoder consists of the pre-net and CBHG networks in Tacotron [37]. The generation of multi-modal features at local scale is:

$$s_i = E_l([BERT_i; LST_i]) \tag{5}$$

where $[;]$ is the concatenating operation, $BERT_i$ are the BERT embeddings and LST_i are the local speaking styles at word level of the i-th utterance, E_l is the local encoder, $s_i \in R^{l_i \times d_s}$ are the multi-modal features at local scale, l_i is the number of words in the i-th utterance and d_s is the dimension of the multi-modal features at local scale.

Similarly, the multi-modal features at global scale are generated for each past utterance. A global encoder is employed to summarize the multi-modal features from the local scale to the global scale, which has the same architecture as the local encoder. The difference is that the global encoder only outputs the last step of CBHG. This output is then concatenated with the extracted global speaking style as the multi-modal features at global scale:

$$g_i = [E_g([BERT_i; LST_i]); GST_i] \tag{6}$$

where GST_i is the utterance level global speaking style of the i-th utterance, E_g is the global encoder, $g_i \in R^{d_g}$ are the multi-modal features at global scale and d_g is dimension of the features.

5.2 Context modeling with MSRGCN

To model the dependencies in conversations at both global and local scales, we propose the MSRGCN as the key component for context modeling. MSRGCN is extended from the conventional RGCN [29] with the ability to simultaneously model the dependencies at multiple scales.

The above extracted multi-scale multi-modal features for the past utterances in each conversation are organized as a directed graph, consisting of vertices, edges, edge types and multi-scale edge weights. MSRGCN is then adopted to aggregate the information along the graph, producing new conversational representations holding richer context information regarding temporal, speaker and multi-scale style dependencies.

5.2.1 Vertices. Each past utterance in the conversation is represented as a vertex in the graph. Unlike other graphs used in conventional GCNs [14, 29], the proposed approach defines each vertex to hold a sequence of feature vectors instead of a single feature vector with fixed length. Each vertex is initialized with the corresponding multi-modal features at local scale:

$$\mathcal{V}_i = s_i \tag{7}$$

where $s_i \in R^{l_i \times d_s}$ are the multi-modal features at local scale.

5.2.2 Edges. To model all possible dependencies in conversations, we add every possible edge to the graph, including two directed edges for each two vertices in both directions and a self-loop edge for each vertex in the graph.

5.2.3 Edge types. We also consider the temporal and speaker dependencies like DialogueGCN [7] by defining different edge types for different temporal orders and each possible speaker pair. We define the edge type of each edge as a combination of three components: the temporal order (either *past-to-future* or *future-to-past*), the speaker label of the source vertex, and the speaker label of the target vertex. Particularly, the edge is marked as *future-to-past* when it is a self-loop edge.

5.2.4 Edge weights. The edge weights are the most important part to achieve multi-scale graph convolutions. Instead of being a single number between 0 and 1, the proposed approach defines the weight of each edge to be a combination of a global edge weight and a local edge weight matrix:

$$\alpha_{j,i} = (w_{j,i}, W_{j,i}) \tag{8}$$

where j and i are the indices of the source and target vertices (i.e., the j-th and i-th utterances), $w_{j,i} \in R$ is the global edge weight, $W_{j,i} \in R^{l_j \times l_i}$ is the local edge weight matrix, l_j and l_i are the lengths of corresponding multi-modal features at local scale and also the number of words in the j-th and i-th utterances, and $\alpha_{j,i}$ is the combination of global and local edge weights.

The global edge weights are the attention weights calculated from the global-scale, multi-modal features by a conventional attention mechanism [35]. For each vertex i, the global edge weights for the edges pointing to this vertex are calculated as:

$$w_{1,i}, \ldots, w_{i,i}, \ldots, w_{n,i} = softmax\left(g_i^T [g_1, \ldots, g_n]\right) \tag{9}$$

where n is the length of the context.

The local edge weight matrices are the attention weights calculated from the local-scale, multi-modal features by a bidirectional attention mechanism [19] to learn the bidirectional dependencies between each pair of past utterances. The bidirectional attention mechanism is proposed to capture the bidirectional relations between two sets of key-value pairs. By setting the two sets of keys and values as the multi-modal features at local scale of two vertices i and j, the local edge weights for the edges between these two vertices are calculated as:

$$A_{i,j} = f(s_i) \times f(s_j)^T \tag{10}$$

$$A_{j,i} = f(s_j) \times f(s_i)^T = A_{i,j}^T \tag{11}$$

$$W_{i,j} = softmax(A_{i,j}) \tag{12}$$

$$W_{j,i} = softmax(A_{j,i}) \tag{13}$$

where f is a shared linear projection for both directions, $A_{i,j}$ and $A_{j,i}$ are two internal attention score matrices in the bidirectional attention mechanism, $W_{i,j} \in R^{l_i \times l_j}$ and $W_{j,i} \in R^{l_j \times l_i}$ are the local edge weights. Particularly, for the self-loop edge, we have:

$$A_{i,i} = f(s_i) \times f(s_i)^T \tag{14}$$

$$W_{i,i} = softmax(A_{i,i}) \tag{15}$$

5.2.5 Feature transformation. Inspired by DialogueGCN [7], a two-step multi-scale graph convolution process is employed to model the dependencies in conversations at both global and local scales, which is a stack of two 128-dimensional MSRGCNs.

In the first step, the first MSRGCN jointly considers the dependencies in conversations at multiple scales by virtue of both global and local edge weights in the convolution for each vertex. The local edge weight matrices are used to learn the dependencies at local scale. It should be noticed that the numbers of words in different utterances are not exactly the same, leading to the situation that the lengths of the multi-modal features at local scale for different vertices are also different. The introduction of local edge weight matrices overcomes such challenges of dimension mismatch by mapping the features of the neighbouring vertices into the length of the features of the target vertex. The global edge weights are further utilized to normalize these mapped features under different edge types and model the dependencies in conversations at global scale. The transformation in the first MSRGCN is formulated as:

$$h_i^{(1)} = \sigma\left(\sum_{r \in \mathcal{R}} f_r^{(1)}\left(\sum_{j \in N_i^r} w_{j,i} W_{j,i}^T s_j\right) + w_{i,i} f_0^{(1)}(s_i)\right) \tag{16}$$

where $s_i \in R^{l_i \times d_s}$ is the input feature sequence of vertex i, r is one of $\mathcal{R}$ which are the edge types, N_i^r are the neighbours of vertex i under type r, j is a neighbour vertex in N_i^r, $s_j \in R^{l_j \times d_s}$ is the feature sequence for this neighbour, $w_{i,i} \in R$ is the input global edge weight for the self-loop edge at vertex i, $w_{j,i} \in R$ and $W_{j,i} \in R^{l_j \times l_i}$ are the input global and local edge weights for the edge from j to i, $W_{j,i}^T s_j \in R^{l_i \times d_s}$ is the mapped information of j in shape of s_i, $f_0^{(1)}$ and $f_r^{(1)}$ are the learnable convolution kernels, σ is an activation function, and $h_i^{(1)}$ is the output for vertex i.

In the second step, since the dependencies of edge types have already been considered in the previous step, the convolution in the second MSRGCN is processed with all the edge types set as a same universal edge type. Also, since the number of neighbours is now equal for each vertex under the same edge type, the global edge weights used to normalize the information of different edge types are also omitted. The transformation in the second MSRGCN can then be formulated as:

$$h_i^{(2)} = \sigma\left(f^{(2)}\left(\sum_{j \in N_i} W_{j,i}^T h_j^{(1)}\right) + f_0^{(2)}\left(h_i^{(1)}\right)\right) \tag{17}$$

where $f_0^{(2)}$ and $f^{(2)}$ are the learnable convolution kernels for the second MSRGCN, and $h_i^{(2)}$ is the representation output for vertex i.

After the processing of MSRGCNs, we concatenate the final representation of each vertex with its initial multi-modal features at local scale as the conversational representation at local scale:

$$s_i' = [s_i; h_i^2] \tag{18}$$

A post global encoder is further adopted to summarize the conversational representation at local scale to global scale for this utterance. The architecture of the post global encoder is the same as the global encoder in feature extraction. It also uses the last step of CBHG as the output. The output of the post global encoder is then concatenated with the multi-modal features at global scale as the conversational representation for each utterance at global scale:

$$g_i' = [g_i; E_g^p(s_i')] \tag{19}$$

where E_g^p is the post global encoder.

5.3 Multi-scale speaking style inference

As shown in Figure 4, to infer the global and local speaking styles for the current utterance, the conversational representations learned above are summarized at both global and local scales. The textual features of the current utterance are extract at global and local scales with Equation (5) and (6), except that the corresponding speaking style features are set to zero:

$$s_{n+1} = E_l([BERT_{n+1}; 0]) \tag{20}$$

$$g_{n+1} = [E_g([BERT_{n+1}; 0]); 0] \tag{21}$$

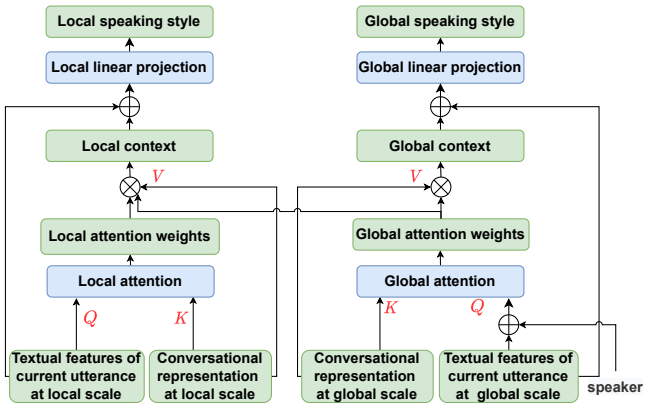


Figure 4: Inferring global and local speaking styles for the current utterance from multi-scale conversational representations.

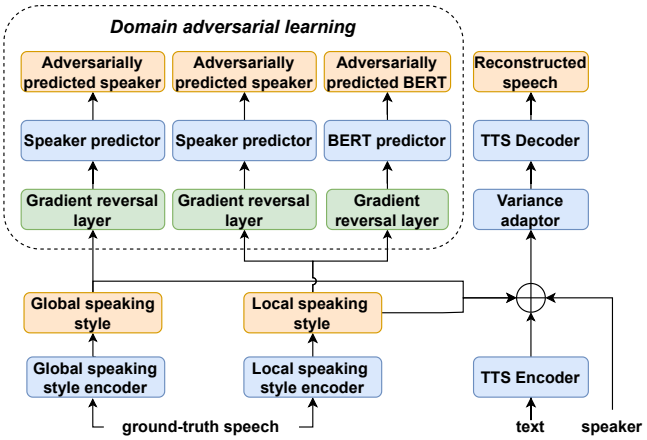


Figure 5: Pretraining the global and local speaking style encoders and FastSpeech 2.

where g_{n+1} and s_{n+1} are the textual features of the current utterance at global and local scales respectively.

Similar to the global and local edge weights used in MSRGCN, we calculate global and local attention weights to summarize the context at different scales. The global attention weights are calculated by querying on the concatenation of the textual features at global scale and the speaker information of the current utterance:

$$w_{1,n+1}, \ldots, w_{n,n+1} = softmax\left(\begin{bmatrix} g_{n+1}^T \\ speaker_{n+1}^T \end{bmatrix} [g'_1, \ldots, g'_n]\right) \quad (22)$$

where $w_{i,n+1}$ is the global attention weight for each past utterance. The local attention weights are also calculated by a conventional attention mechanism since only the past-to-current direction is considered for each past utterance:

$$A_{i,n+1} = s'_i \times f_a(s_{n+1}) \quad (23)$$

$$W_{i,n+1} = softmax(A_{i,n+1}) \quad (24)$$

where $A_{i,n+1}$ is the internal attention matrix, f_a is a linear projection, and $W_{i,n+1}$ is the local attention weight matrix.

The representations in context are then summarized at multiple scales with the global and local attention weights of past utterances:

$$c_g, c_l = \sum_{i=1}^{n} w_{i,n+1} g'_i, \sum_{i=1}^{n} w_{i,n+1} W_{i,n+1}^T s'_i \quad (25)$$

where $c_g \in R^{d'_g}$ is the summarized global context, $c_l \in R^{l_{n+1} \times d'_s}$ is the summarized local context, d'_g is the dimension of g'_i, l_{n+1} is the number of words in the current utterance, d'_s is the dimension of s'_i.

The summarized global and local contexts are then respectively concatenated with the global and local textual features of the current utterance to predict the global and local speaking styles:

$$GST'_{n+1} = softmax\left(f_g\left([g_{n+1}; c_g]\right)\right) \quad (26)$$

$$LST'_{n+1} = softmax\left(f_l\left([s_{n+1}; c_l]\right)\right) \quad (27)$$

where f_g and f_l are two linear projections, GST'_{n+1} and LST'_{n+1} are the predicted global and local speaking styles for the current utterance.

5.4 Speech synthesis with predicted multi-scale speaking styles

To synthesize speech with proper speaking styles at both global and local scales, a FastSpeech 2 [26] based acoustic model is adopted as the TTS backbone, as shown in Figure 2. The speaker embedding of the current utterance and the predicted global and local speaking styles are upsampled to phoneme level and concatenated with the encoder outputs. The concatenated results are then passed to the variance adaptor to infer the pitch, duration and energy, and are further converted to mel-spectrogram by the decoder. A well-trained HiFi-GAN [15] is used as the vocoder to generate speech from the predicted mel-sprectrogram with desired speaking styles confirming to the conversational context.

5.5 Training strategy

To ensure the extracted global and local speaking styles are compatible with the FastSpeech 2 TTS backbone, the global and local speaking style encoders in Section 5.1 and the FastSpeech 2 are first jointly pretrained in a same framework, as shown in Figure 5. The global and local speaking styles extracted from each ground-truth speech are used to reconstruct the same speech. To disentangle the text and speaker information from the extracted global and local speaking styles, we also employ the gradient reversal layer (GRL) [4] and domain adversarial learning [5, 41] in pretraining:

$$speaker'_{GST} = f_{speaker}^{GST}\left(GRL(GST)\right) \quad (28)$$

$$speaker'_{LST} = f_{speaker}^{LST}\left(GRL(LST)\right) \quad (29)$$

$$BERT' = f_{BERT}\left(GRL(LST)\right) \quad (30)$$

where $f_{speaker}^{GST}$, $f_{speaker}^{LST}$ and f_{BERT} are linear projections serving as the adversarial speaker predictors and text predictor, GRL reverses the gradients of these adversarial predictors, $speaker'_{GST}$, $speaker'_{LST}$ are the adversarially predicted speaker embeddings, $BERT'$ is the adversarially predicted textual embedding of this utterance. The loss for pretraining is the sum of reconstruction loss and adversarial losses, where the former one is the mean squared error

(MSE) between the predicted and ground-truth mel-spectrograms, and the latter ones are the MSEs between the predicted and ground-truth speaker and BERT embeddings.

After pretraining, the global and local encoders are frozen and used to extract the speaking styles as mentioned in Section 5.1. And the FastSpeech 2 is also frozen and used to synthesize speech as mentioned in Section 5.4.

To train the proposed context modeling method and infer the speaking styles for the current utterance, the MSRGCNs and the following networks mentioned in Section 5.2 and 5.3 are trained with the loss function defined as the MSEs between the predicted and ground-truth global and local speaking styles:

$$loss = MSE(GST_{n+1}, GST'_{n+1}) + MSE(LST_{n+1}, LST'_{n+1}) \tag{31}$$

6 EXPERIMENTS

6.1 Baselines

To demonstrate the effectiveness of inferring speaking styles from the multi-modal features with MSRGCN, we employ 3 approaches with different context modeling methods as the baselines, which also employ FastSpeech 2 as the TTS backbone.

6.1.1 No context modeling. The first baseline approach is a vanilla FastSpeech 2 [26] with no context modeling, which is also a representative of state-of-the-art non-conversational TTS systems.

6.1.2 GRU-based context modeling. We employ the GRU-based context modeling method [8] as the second baseline approach, in which the context information is simply summarized at global scale by a uni-directional GRU:

$$c_g = final\left(GRU\left(SBERT_1, \ldots, SBERT_n\right)\right) \tag{32}$$

where $SBERT_1, \ldots, SBERT_n$ are the sentence level BERT embeddings of the past utterances. Equation (26) is employed to predict the global speaking style of the current utterance. Particularly, the global speaking styles in this baseline are pretrained without local speaking style control, which are different from the global speaking styles in our proposed approach. As a result, the global speaking styles in this baseline may embed part of the local speaking styles.

6.1.3 DialogueGCN-based context modeling. We further employ the DialogueGCN-based context modeling [18] as the most competitive baseline for comparison. In this approach, the multi-modal features, edge weights and attention weights are only obtained at global scale with Equations (6), (9) and (22). And the feature transformation is processed with DialogueGCN:

$$h_i^{(1)} = \sigma\left(\sum_{r\in\mathcal{R}} f_r^{(1)}\left(\sum_{j\in N_i^r} w_{j,i} g_j\right) + w_{i,i} f_0^{(1)}\left(g_i\right)\right) \tag{33}$$

$$h_i^{(2)} = \sigma\left(\sum_{j\in N_i} f^{(2)}\left(h_j^{(1)}\right) + f_0^{(2)}\left(h_i^{(1)}\right)\right) \tag{34}$$

Comparing with Equations (16) and (17), in DialogueGCN, the first RGCN only takes the global features as inputs and neglects the edge weights at local scale. Also the edge types and global edge weights are omitted in the second GCN. The outputs of DialogueGCN are concatenated with the global features as the new global multi-modal features for each past utterance:

$$g'_i = \left[g_i; h_i^{(2)}\right] \tag{35}$$

The global textual feature of the current utterance is extracted as:

$$g_{n+1} = E_g(BERT_{n+1}) \tag{36}$$

and used to summarize the new global features with Equation (25) and projected to the global speaking styles for the current utterance with Equation (26). Same to the previous GRU-based context modeling approach, the global speaking styles in this approach are pretrained without local speaking style control.

6.2 Training setups

We employ English conversation corpus (ECC) [18] as the dataset for training the proposed MSRGCN based context modeling method for conversational TTS, from which the first 61 videos are used for training, the remaining 5 videos are used for evaluation. The conversations in these videos are converted into 20,996 and 1,407 conversation chunks for the training and test sets with a chunk size of 6, in which the first 5 utterances are used as the context for the last utterance. The model is trained for 10 epochs with a batch size of 32 and a learning rate of 10^{-4}.

We follow the training setups of FastSpeech 2 [26] to train the pretraining framework mentioned in Section 5.5. The model is trained for 500,000 iterations with a batch size of 16. The inputs for the global and local speaking style encoders are mel-spectrograms extracted with a window length of 25ms and a shift of 10ms.

6.3 Evaluations

We adopt the MSE between the predicted and ground-truth mel-spectrograms as the metric for the objective evaluation. The predicted mel-spectrogram is resized to match the length of the ground-truth mel-spectrogram with the nearest-neighbour interpolation. Moreover, we further calculate the MSEs between the high (last 10 dimensions) and low (first 10 dimensions) frequency bands of the predicted and ground-truth mel-spectrograms.

For the subjective evaluation, 20 conversation chunks[3] are further randomly selected and evaluated by 25 listeners. The listeners are asked to rate on how the speaking styles of synthesized speeches match their conversation context on a scale from 1 to 5 with 1 point interval, from which subjective mean opinion scores (MOS) are calculated. Meanwhile, the listeners are asked to choose a preferred speech generated by the proposed or the baseline approaches, from which preference rates are calculated.

6.3.1 Comparing with state-of-the-art baselines. The results of the objective and subjective evaluations are shown in Table 1. By applying the dependency modeling at global scale, the DialogueGCN-based approach mainly optimizes the mel-spectrogram at the low frequency bands than the no context modeling approach, with MOS increased by 0.152 and preference rate exceeded by 4.93%. This demonstrates the need of considering the global level dependencies in conversations. Moreover, compared with the DialogueGCN-based approach, the proposed MSRGCN-based approach further

[3]Some samples are available at https://thuhcsi.github.io/mm2022-conversational-tts/.

Table 1: Subjective and objective evaluations for different approaches. *The second to fourth columns are the MSEs between the full, high (last 10 dimensions) and low (first 10 dimensions) frequency bands of predicted and ground-truth mel-spectrograms.

Context modeling method	MSE* (Mel)	MSE* (High)	MSE* (Low)	MOS ± 95% confidence interval	Preference rate
No context modeling [26]	2.681	1.585	2.691	3.386 ± 0.062	19.54%
GRU-based [8]	3.016	2.083	3.134	2.433 ± 0.060	2.82%
DialogueGCN-based [18]	2.576	1.563	**2.550**	3.538 ± 0.056	24.47%
MSRGCN-based (Proposed)	**2.547**	**1.466**	2.556	**3.807 ± 0.063**	**53.17%**

Table 2: Objective evaluations on different chunk lengths.

Chunk length	MSE (GST)	MSE (LST)	MSE* (Mel)
3	3.28×10^{-3}	3.81×10^{-3}	2.568
4	3.31×10^{-3}	3.74×10^{-3}	2.566
5	3.15×10^{-3}	3.70×10^{-3}	2.549
6	$\mathbf{3.10 \times 10^{-3}}$	$\mathbf{3.64 \times 10^{-3}}$	**2.547**
11	3.18×10^{-3}	3.70×10^{-3}	2.556
16	3.34×10^{-3}	3.67×10^{-3}	2.573
21	3.33×10^{-3}	3.73×10^{-3}	2.574

Table 3: Ablation studies on modalities and scales.

Approach	MSE (GST)	MSE (LST)	MSE* (Mel)
Proposed	$\mathbf{3.10 \times 10^{-3}}$	$\mathbf{3.64 \times 10^{-3}}$	**2.547**
w/o textual modal	3.15×10^{-3}	3.72×10^{-3}	2.562
w/o acoustic modal	3.88×10^{-3}	4.20×10^{-3}	2.682
w/o local scale	6.78×10^{-3}	5.11×10^{-3}	2.776
w/o global scale	3.34×10^{-3}	3.75×10^{-3}	2.587

optimizes the mel-spectrogram at the high frequency bands while the MSEs at the low frequency bands are almost same. And the MOS and preference rate of the proposed approach are further improved by 0.269 and 28.70% respectively, which demonstrate the superiority of dependency modeling at multiple scales, and especially indicate the importance of modeling the local dependencies.

The GRU-based approach achieves even worse results than the no context modeling method, with MOS greatly decreased by 0.953 and the preference rate of being just 2.82%. It should be noted that the GRU-based approach only models the temporal dependencies in conversations. This demonstrates the importance of other dependencies in conversations for context modeling, and shows that insufficient context modeling will seriously affect the synthesis of speaking styles in conversational TTS systems.

6.3.2 Experiments on context lengths. We also explore the effectiveness of context modeling with different context lengths. Since two past utterances are the minimal requirement for batch normalization in CBHG, we train the proposed MSRGCN-based context modeling method with chunk lengths ranging from 3 to 21. We adopt the MSEs between the predicted and ground-truth global and local speaking styles as additional metrics. Results are shown in Table 2. The MSEs decrease when enlarging the chunk length from 3 to 6, and increase when enlarging the chunk length from 6 to 21. This shows that either insufficient or redundant context information will interfere the understanding of context.

6.4 Ablation studies

We then explore the effectiveness of multi-modal information and multi-scale dependency modeling. The textual and acoustic modalities are respectively removed from the proposed approach by setting the corresponding inputs as zeros. According to the results shown in the second and third lines of Table 3, the absence of information on each modal will lower the effectiveness of context modeling.

The dependency modeling at global and local scales are also respectively removed from the proposed approach. At the global scale, the dependency modeling is removed by setting the input global speaking styles to zero and the global edge and attention weights to one. The dependency modeling at the local scale is removed by directly using the DialogueGCN based context modeling methods to predict the global and local speaking styles in the proposed approach. As the results shown in the fourth and fifth lines of Table 3, missing the dependency modeling on each scale will lead to worse speaking style inference. In particular, removing the dependency modeling on the local scale causes significant losses on the performance than the global scale, showing the importance of dependency modeling at local scale for speaking style inference.

7 CONCLUSION

To improve the synthesis of speaking styles in speech-driven interactive systems, we present a novel approach whereby MSRGCN is used for context modeling in conversations to achieve better speaking style inference. The dependencies among the multi-modal information in conversational context at both global and local scales are captured by MSRGCN as the multi-modal context information at multiple scales, which are further used to infer the global and local speaking styles of the current utterance and synthesized to speech. Experimental results demonstrate the superiority of MSRGCN based approach over state-of-the-art conversational TTS systems with only context modeling at the global scale. The contribution of modeling multi-modal information and multi-scale dependencies are further demonstrated in ablation studies.

ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan (2021QY1500), National Natural Science Foundation of China (NSFC) (62076144), the joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK40415) and Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004).

REFERENCES

[1] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

[2] Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. 2021. Controllable Context-aware Conversational Speech Synthesis. *arXiv preprint arXiv:2106.10828* (2021).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[6] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating Multiple Diverse Responses for Short-Text Conversation. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 6383–6390. https://doi.org/10.1609/aaai.v33i01.33016383 Number: 01.

[7] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

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

[9] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

[10] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7037–7041. https://doi.org/10.1109/ICASSP43922.2022.9747397 ISSN: 2379-190X.

[11] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7360–7370. https://doi.org/10.18653/v1/2020.emnlp-main.597

[12] Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2020. Conversational Question Answering over Passages by Leveraging Word Proximity Networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2129–2132. https://doi.org/10.1145/3397271.3401399

[13] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 5530–5540. https://proceedings.mlr.press/v139/kim21f.html ISSN: 2640-3498.

[14] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)* (2017).

[15] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 17022–17033.

[16] Chia-Chih Kuo, Shang-Bao Luo, and Kuan-Yu Chen. 2020. An Audio-enriched BERT-based Framework for Spoken Multiple-choice Question Answering. *arXiv:2005.12142 [cs]* (May 2020). http://arxiv.org/abs/2005.12142 arXiv: 2005.12142.

[17] Siddique Latif, Junaid Qadir, and Muhammad Bilal. 2019. Unsupervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 732–737. https://doi.org/10.1109/ACII.2019.8925513 ISSN: 2156-8111.

[18] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su. 2022. Enhancing Speaking Styles in Conversational Text-to-Speech Synthesis with Graph-based Multi-modal Context Modeling. https://doi.org/10.48550/ARXIV.2106.06233

[19] Jingbei Li, Yi Meng, Zhiyong Wu, Helen Meng, Qiao Tian, Yuping Wang, and Yuxuan Wang. 2022. NeuFA: Neural Network Based End-to-End Forced Alignment with Bidirectional Attention Mechanism. https://doi.org/10.48550/ARXIV.2203.16838

[20] Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring user emotive state changes in realistic human-computer conversational dialogs. In *Proceedings of the 26th ACM international conference on Multimedia*. 136–144.

[21] Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring User Emotive State Changes in Realistic Human-Computer Conversational Dialogs. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, New York, NY, USA, 136–144. https://doi.org/10.1145/3240508.3240575 event-place: Seoul, Republic of Korea.

[22] Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. 2020. Conversational Emotion Recognition Using Self-Attention Mechanisms and Graph Neural Networks. *Proc. Interspeech 2020* (2020), 2347–2351.

[23] Hongyin Luo, Shang-Wen Li, and James Glass. 2020. Prototypical Q Networks for Automatic Conversational Diagnosis and Few-Shot New Disease Adaption. *arXiv:2005.11153 [cs]* (May 2020). http://arxiv.org/abs/2005.11153 arXiv: 2005.11153.

[24] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6818–6825.

[25] Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7209–7213. https://doi.org/10.1109/ICASSP40776.2020.9054484 ISSN: 2379-190X.

[26] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

[27] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: fast, robust and controllable text to speech. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Number 285. Curran Associates Inc., Red Hook, NY, USA, 3171–3180.

[28] Yu-Ping Ruan, Shu-Kai Zheng, Taihao Li, Fen Wang, and Guanxiong Pei. 2022. Hierarchical and Multi-View Dependency Modelling Network for Conversational Emotion Recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7032–7036. https://doi.org/10.1109/ICASSP43922.2022.9747123 ISSN: 2379-190X.

[29] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*. Springer, 593–607.

[30] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv:1712.05884 [cs]* (Dec. 2017). http://arxiv.org/abs/1712.05884 arXiv: 1712.05884.

[31] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 15 (May 2021), 13789–13797. https://ojs.aaai.org/index.php/AAAI/article/view/17625 Number: 15.

[32] Xiaohan Shi, Sixia Li, and Jianwu Dang. 2020. Dimensional Emotion Prediction based on Interactive Context in Conversation. *Proc. Interspeech 2020* (2020), 4193–4197.

[33] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.

[34] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*. PMLR, 4693–4702.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[36] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F. Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, Dirk Heylen, Rene Kaiser, Maria Koutsombogera, Alexandros Potamianos, Steve Renals, Giuseppe Riccardi, and Albert Ali Salah. 2015. Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. *Cognitive Computation* 7, 4 (Aug. 2015), 397–413. https://doi.org/10.1007/s12559-015-9326-z

[37] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. *Proc. Interspeech 2017* (2017), 4006–4010.

[38] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*. PMLR, 5180–5189.

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[40] Yunhe Xie, Chengjie Sun, and Zhenzhou Ji. 2022. A Commonsense Knowledge Enhanced Network with Retrospective Loss for Emotion Recognition in Spoken Dialog. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7027–7031. https://doi.org/10.1109/ICASSP43922.2022.9746909 ISSN: 2379-190X.

[41] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. *arXiv:1907.04448 [cs, eess]* (July 2019). http://arxiv.org/abs/1907.04448 arXiv: 1907.04448.

[42] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.