IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 November 2023, accepted 5 December 2023, date of publication 13 December 2023,
date of current version 19 December 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3342456*

RESEARCH ARTICLE

# DialoguePCN: Perception and Cognition Network for Emotion Recognition in Conversations

**XIAOLONG WU[1], CHANG FENG[2], MINGXING XU[2], THOMAS FANG ZHENG[2], (Senior Member, IEEE), AND ASKAR HAMDULLA[1]**
[1]School of Information Science and Engineering, Xinjiang University, Urumqi 830000, China
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100080, China

Corresponding author: Askar Hamdulla (askar@xju.edu.cn)

This work was supported in part by the Xinjiang Postgraduate Research Innovation Project of China under Grant XJ2022G023, in part by the Provincial and Ministerial Key Project of the 14th Five-Year Scientific Research Plan of the State Language Commission of China under Grant ZDI145-58, in part by the Natural Science Foundation of Xinjiang Autonomous Region of China under Grant 2023D01C19, and in part by the Natural Science Foundation of China under Grant U2003207.

**ABSTRACT** In the Emotion Recognition in Conversations (ERC) task, extracting emotional cues from the context is an effective strategy for improving model performance. However, current research has two evident limitations: firstly, irrelevant context information severely affects the extraction of emotional features at the utterance level. Secondly, in dialogues, subsequent utterances' retrieval of emotional cues does not benefit from extracted emotional cues from preceding utterances. This paper designs a Dialogue Perception Cognition Network (DialoguePCN) model, which aims to solve the issues above by simulating the perception and cognition phases of emotion in conversations. In the perception phase, DialoguePCN proposes an activation module based on a cosine similarity selection algorithm, providing a dynamic initial emotional state for the predicted utterance. In the cognition phase, the model introduces a new gating mechanism, marking the first attempt to use the extracted utterance emotion representation to reconstruct context information iteratively. This approach reduces the complexity of retrieving emotional cues from the context and solves the inherent cold-start challenge in ERC tasks. Using audio and text features, the accuracy of DialoguePCN reached 68.7% on the IEMOCAP dataset.

**INDEX TERMS** Emotion recognition in conversations, perception-cognition, activation module, dynamic long-term memory update module.

## I. INTRODUCTION

Emotion Emotion Recognition in Conversations (ERC) aims to detect the speaker's emotions within conversations [1], [2], [3], [4], [5]. It is pivotal in intelligent customer service and human-computer interaction systems [6], [7].

For an effective ERC model, it is imperative to possess the capability to simulate human proficiency in retrieving and integrating emotional cues from context [8], [9], as these cues are the primary triggers of emotions. Nevertheless, there are known flaws in using contextual details:

1) Existing studies do not design distinct contexts for utterances at different time steps. For instance, works [10], [11], [12], [13], [14], [15], [16] adopt extracting emotional cues from the context formed by each utterance within the entire dialogue, exacerbating the interference of irrelevant information and complicating the retrieval of emotional cues. Despite [17] choosing the nearest K utterances from the current moment to generate contextual information, it exclusively considers altering the context through positional information, neglecting the potential logical impact of more distantly located speech on the current moment.

2) Current research fails to leverage previously extracted emotional information from the preceding discourse into the feature extraction process of subsequent utterances. In real-world conversations, people can better understand each other's emotions as the conversation progresses [18], [19], [20]. This phenomenon suggests that the recognition results of previous moments can influence people's perception of emotions in subsequent utterances. Therefore, it is crucial

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

to continually update the contextual information to include previous recognition results. Conversely, if the contextual information remains fixed, the difficulty of retrieving emotional cues for subsequent utterances does not decrease due to the preceding recognition results, resulting in the ''cold start'' issue. To the best of our knowledge, this is the first work that has addressed this challenge.

The DialguePCN model proposed in this paper is based on emotion cognition theory to simulate the emotional cognition process in human conversations. Research [21] presents emotional states that may be considered a function of a state of physiological arousal and a cognition appropriate to this state of arousal. Research [22] posits that knowledge is a distal variable that requires an additional evaluation process to generate emotions in the formation of emotions. Building upon this, [23], [24] introduces a general two-stage theory of human reasoning evaluation. During the heuristic process, the heuristic procedure selects ''relevant'' items from task information, while the analytical process operates on the selected items to generate inferences or judgments. Factors deemed to contribute to heuristic selection include perceptual and semantic associations. The analytical process is context-dependent: individuals engage in repeated reasoning from experience. This perspective aligns with the viewpoints presented in the work of DialogueCRN [12]. In our research, the process of selecting task-related information is called the ''perception phase'', which involves choosing information without considering emotional cues. The brain's repeated retrieval and reasoning of information is referred to as the ''cognition phase'', where emotional contextual information is retrieved and integrated.

According to the above discussion, this paper introduces the Dialogue Perception-Cognition Network (DialoguePCN), which consists of perception and cognition phases. In the perception phase, we construct contextual information for the speaker's intrinsic emotional inertia and the emotional impact between speakers using a BiLSTM. We introduce a new activation module based on the cosine similarity algorithm to avoid irrelevant contextual interference with utterance emotional representation. This module simulates the heuristic information selection process. In the cognition phase, we utilize the multi-turn reasoning module developed by DiaolongCRN to simulate how the brain processes emotional cues. This module incorporates attention mechanisms and LSTM networks to repeatedly retrieve and integrate emotional cues from the context, resulting in the ultimate emotional representation for recognizing emotions. Additionally, we have designed a dynamic long-term memory update module that utilizes a gating mechanism to resolve the ''cold start'' problem and avoid losing cues during lengthy conversations. This module updates contextual information in real time using the emotional representation of each moment. Extensive experiments were conducted on the IEMOCAP dataset [25] to validate DialoguePCN, achieving an accuracy rate of 68.7%, significantly improving existing works.

The main contributions and innovations of this paper are as follows:

1) We proposed a Dialogue Perception-Cognition Network (DialoguePCN) inspired by emotional cognition theory, which simulates human emotional perception and cognition.

2) We designed an activation module that uses the cosine similarity algorithm to prevent interference from excessive contextual information.

3) We proposed a dynamic long-term memory update module based on a gating mechanism to address the ''cold start'' problem in ERC tasks. This approach can be extended to other sequential prediction problems.

## II. RELATED WORK

### A. EMOTION RECOGNITION IN CONVERSATIONS (ERC)

The challenge in Emotion Recognition in Conversations (ERC) lies in leveraging contextual information to enhance emotion recognition performance. Majumder et al. bc-LSTM [10] preserves the sequential order of utterances and enables consecutive utterances to share information, thus providing contextual information to the utterance-level sentiment classification process. DialogueRNN [13] treats each incoming utterance, considering the speaker's characteristics, which gives finer context. DialogueCRN [12] constructs contextual information by separately considering the emotional inertia of the speakers themselves and the emotional influence between speakers. Reference [13] examined emotional variations both within individual speakers and between different speakers in context. Wang et al. [14] introduced a novel Hierarchical Stacked Graph Convolutional Framework (HSGCF), which employs five interconnected graph convolutional layers hierarchically to establish a more discriminative emotion feature extractor. Song et al. [15] utilized the Conditional Random Field (CRF) to further investigate the probability of emotional transitions for specific speakers during conversations, extracting features from the context related to the specific speaker. To explore the scope of emotion's influence, Shen et al. [26] optimized the utilization of distant and proximate contextual information in conversations using a Directed Acyclic Graph network called DAG-ERC. Wen et al. [16] proposed DIMMN and incorporated temporal convolution networks to assimilate contextual information from individual speakers, aiming to learn Long-term dependent information.

In recent years, several novel efforts have been made that introduce models that integrate knowledge graphs [27] and multi-task learning. COSMIC [28] analyzes emotions by ascertaining speakers' internal and external states based on causal knowledge. Stappen et al. [29] explored an extraction method grounded in lexical knowledge, obtaining emotional understanding from video transcripts. Zhang et al. [30] employed reinforcement learning and domain knowledge for emotion recognition in multimodal conversation videos. Liang et al. [31] designed a graph convolutional network based on dependency trees and emotional common sense to capture emotion dependencies on specific facets.

### B. MEMORY NETWORK AND MULTI-TURN REASONING

In Question Answering (QA) tasks, Memory Networks (MN) have been proposed [32], [33], [34] to infer answers from intricate long-term dialogues. MN operates by utilizing a long-term memory module and an inference module to deduce answers [35], [36]. The long-term memory module acts as a dynamic knowledge base, while the inference module handles the task of answer retrieval. Analogous to the knowledge retrieval in QA tasks, memory networks can be applied to emotion recognition in dialogues, extracting pivotal emotion cues from dialogues. Hazarika et al.'s work CMN [17] first considered interactions between speakers and employed memory networks to more proficiently retrieve emotional cues from context. Subsequently, an Interactive Conversation Memory Network (ICON) was proposed [11], establishing a global memory from the global conversation context constructed at the speaker level. It harnesses attention mechanisms within a recurrent read-write module to retrieve pertinent emotional information from this global memory. Xing et al. [37] introduced the Adaptive Dynamic Memory Network (A-DMN). The model delineates emotions both within individual speakers and inter-speaker relations, employing a situational memory module to incorporate the extracted emotional contextual data. Hu et al. [12] proposed a novel Contextual Reasoning Network(DialogueCRN) that learns contextual information from a cognitive perspective. This model employs a multi-turn inference module to extract and integrate emotional cues.

## III. METHODOLOGY

This section provides a detailed implementation process of DialoguePCN, with the model's architecture presented in Fig. 1.

### A. PERCEPTION PHASE

The perception phase simulates how humans choose relevant information from conversations and store it in their long-term memory. This phase involves the context construction module and the activation module, both found on the left side of Fig. 1.

#### 1) CONTEXT CONSTRUCTION MODULE

This research aims to fully harness the latent emotional cues inherent in the speaker's emotional inertia and the emotional interactions between speakers. Consequently, the module employs Bi-directional Long Short-Term Memory (BiLSTM) to construct speaker-level contextual information (lower channel) from utterances belonging to Speaker A or B in binary dialogues and to build dialogue-level contextual information (upper channel) from interactive exchanges between both participants [38], [39], [40].

Formally, a dialogue is defined as $U = \{u_1, u_2, \ldots, u_N\}$, where $u_i$ represents the utterance at the time $i$. Utterances belonging to speaker $\lambda$ are defined as $U_\lambda$, where $U_\lambda = \{u_1, u_2, \ldots, u_{len(\lambda)}\}$, $\lambda \in \{a, b\}$. where $len(\lambda)$ represents the total number of utterances by speaker $\lambda$.

In Step 1, the dialogue-level context $c_i^g$ and the speaker-level context $c_{\lambda,i}^s$ are constructed for $u_i$:

$$c_i^g, h_i^g = \overleftrightarrow{LSTM}(u_i, h_{i-1}^g), \tag{1}$$

$$c_{\lambda,i}^s, h_{\lambda,i}^s = \overleftrightarrow{LSTM}(u_{\lambda,i}, h_{\lambda,i-1}^s), \tag{2}$$

where, $h_i$ represents the hidden state of the BiLSTM at time step $i$.

In Step 2, the contextual information from each instance is assembled in temporal order to constitute the global dialogue-level context $C^g = \{c_1^g, c_2^g, \ldots, c_N^g\}$ and the global speaker-level context $C^s = \{c_1^s, c_2^s, \ldots, c_N^s\}$.

In Step 3, to emulate the conversion of contextual information into long-term memory within the brain, the global dialogue-level context $C^g$ and the global speaker-level context $C^s$ are mapped by a fully connected layer to dialogue-level long-term memory $M^g$ and speaker-level long-term memory $M^s$, respectively.

#### 2) ACTIVATION MODULE

The activation module selects relevant information from long-term memory, forming an activated memory that is incorporated into the cognition phase [22], [23]. It has two objectives: Firstly, even when the cognition phase struggles to retrieve emotional cues, the model can still use contextual information to its advantage. Secondly, By activated memory, the model can filter out irrelevant emotional cues retrieved during the cognition phase, resulting in a more accurate final emotional representation. The architecture of the activation module is depicted in Fig. 2.

The specific procedure is as follows: Initially, the correlation between the contextual information $c_i$ of the current utterance and the contextual information $C_o$ of other utterances within the dialogue is computed. Subsequently, a *softmax* function is employed to normalize this correlation, resulting in similarity values denoted as $\alpha_i$:

$$\alpha_i = softmax(CosSim(c_i, C_o)). \tag{3}$$

Subsequently, the similarity value $\alpha_i$ is aggregated with the long-term memory through a weighted summation, resulting in the activated memory $a_i$:

$$a_i = \sum_{i=0}^{N-1} \alpha_i M_o. \tag{4}$$

This module exclusively attends to the influence of the context on the current utterance. Consequently, $M_o$ does not encompass the long-term memory of the current utterance $u_i$. The resulting $a_i$ will be utilized as initializing state information during cognitive reasoning.

### B. COGNITION PHASE

The cognition phase emulates the human process of emotional reasoning and generation. It consists of a multi-turn

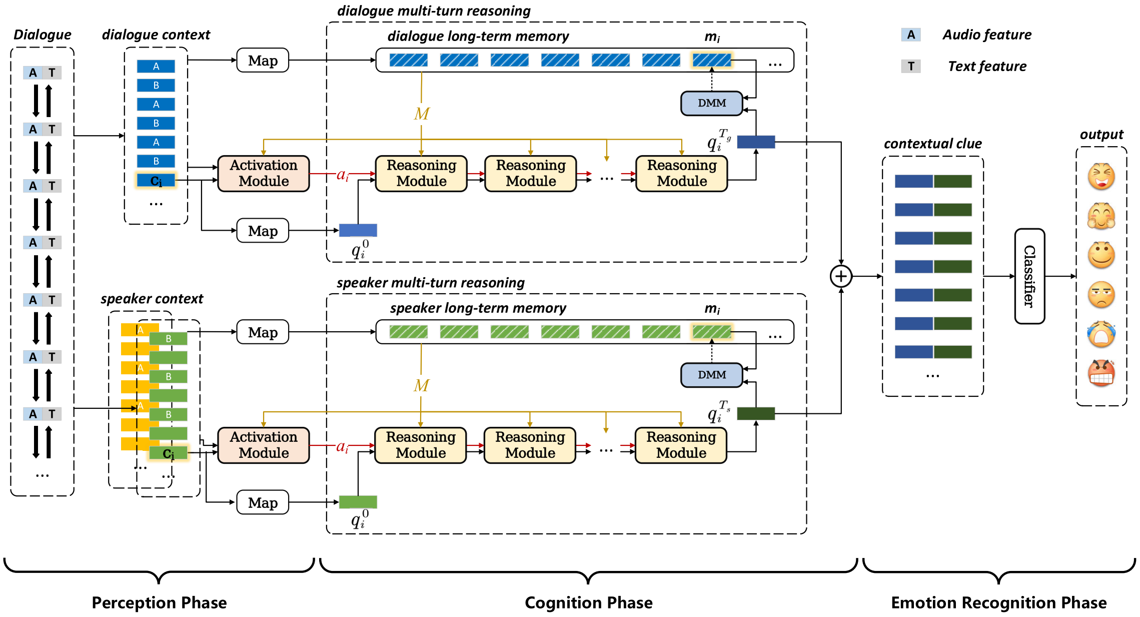

**FIGURE 1.** The architecture of DialoguePCN.

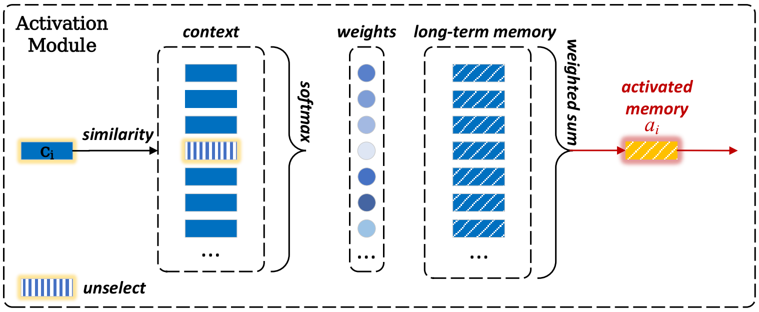

**FIGURE 2.** Activation module.

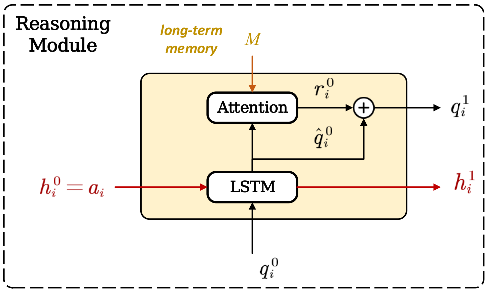

**FIGURE 3.** Reasoning module.

reasoning module and a dynamic long-term memory update module, as illustrated in the middle section of Fig. 1.

### 1) MULTI-TURN REASONING MODULE

The multi-turn reasoning module is designed to retrieve and integrate emotional cues from long-term memory iteratively [24]. The architecture of the reasoning module is illustrated in Fig. 3. The specific process is as follows:

In step 1, the input utterance $c_i$ to be predicted is mapped to $q_i^0$ by a mapping module constructed by fully connected layers(FC), where superscript 0 indicates the first turn reasoning module.

$$q_i^0 = FC(c_i). \tag{5}$$

In step 2, emotional cues from both the context and the activated memory from the perception phase are integrated using an *LSTM*. At the beginning of the first turn of reasoning, emotional cues originate solely from the current utterance and the activated memory. To avoid potential failures in emotional retrieval and minimize irrelevant context interference, the *LSTM's* initial hidden state $h_i^0$ is replaced by the activated memory $a_i$, that is $h_i^0 = a_i$:

$$\hat{q}_i^0, h_i^1 = \overrightarrow{LSTM}(q_i^0, h_i^0), \tag{6}$$

where $\hat{q}_i^0$ encapsulates the representation of both the activation memory and current utterance, subsequently employed to retrieve emotional cues from long-term memory. The resultant $h_i^1$ will be used in the integration process for the next reasoning module.

In step 3, emotional cues are retrieved using attention mechanisms, where emotional cues $r_i^0$ are retrieved from dialogue-level long-term memory $M_i^g$ and speaker-level long-term memory $M_i^s$, respectively:

$$e_i^0 = f(M_i^{s/g}, \hat{q}_i^0), \tag{7}$$

$$\alpha_i^0 = \frac{exp(e_i^0)}{\sum_{i=1}^{N} exp(e_i^0)}, \tag{8}$$

$$r_i^0 = \sum_{i=1}^{N} \alpha_i^0 M_i^{s/g}, \tag{9}$$

where, $f$ represents a dot-product function, $M_i$ denotes the global long-term memory corresponding to utterance $u_i$ at time $i$, $\alpha_i$ represents the relevance value between the current utterance and the long-term memory $M^{s/g}$, and $r_i^0$ represents the emotional cues obtained during the first-turn retrieval.

In step 4, emotional cues $r_i^0$ are integrated with the intermediate representation $\hat{q}_i^0$, using a residual connection to obtain the representation $q_i^1$ after the first turn reasoning module.

$$q_i^1 = [\hat{q}_i^0, r_i^0]. \tag{10}$$

To retrieve and integrate deeper emotional cues, this module conducts $T_g$ and $T_s$ turns of reasoning on the speaker- and dialogue-level channels, respectively, resulting in the final dialogue-level emotional representation $q_i^{T_g}$ and speaker-level emotional representation $q_i^{T_s}$.

#### 2) DYNAMIC LONG-TERM MEMORY UPDATE MODULE(DMM)

The Dynamic Long-Term Memory Update Module (DMM) based on gating mechanisms, selectively stores the final emotional representation $q_i^{T_{g/s}}$ at time-step $i$ into the corresponding long-term memory, yielding the latest dynamic long-term memory. By doing so, emotional cues retrieved at time-step $i$ are preserved at a proximity of only one step away from the next predicted utterance. This approach emphasizes the previous time-step's long-term memory during emotional cue retrieval without overemphasizing distant information. DMM not only significantly reduces retrieval complexity and avoids "cold start" issues but also continually retains emotional cues from the previous time step, mitigating the problem of emotional cue forgetting in lengthy dialogues. The DMM is situated in the central position of Fig. 1. The specific procedure is as follows:

In step 1, employing fully connected layers (FC) to ensure that the final emotional representation $q_i^{T_{g/s}}$, has the same dimensionality as its corresponding long-term memory $m_i$.

In step 2, utilize the *tanh* function to compute the similarity weights $w_i$, between the final emotional representation $q_i^{T_{g/s}}$ and its corresponding long-term memory $m_i$.

$$w_i = \tanh(FC(q_i^{T_{g/s}}), m_i). \tag{11}$$

**TABLE 1.** Data distribution of the IEMOCAP dataset.

| Partition | Utterance Count | Dialogue Count |
|---|---|---|
| train+val | 5810 | 120 |
| test | 1623 | 31 |

In step 3, selectively retain the content from $q_i^{T_{g/s}}$ into $m_i$ based on the similarity weights $w_i$, resulting in the latest long-term memory $\hat{m}_i$:

$$\hat{m}_i = w_i q_i^{T_{g/s}} + m_i. \tag{12}$$

In step 4, the latest long-term memory $\hat{m}_i$ is employed to replace the long-term memory $m_i$ at time-step $i$ within $M_i$, generating $M_{i+1}$. $M_{i+1}$ will subsequently be utilized in the multi-turn reasoning process at time-step $i+1$.

### C. EMOTION RECOGNITION PHASE

During the emotion recognition phase, the final emotional representations $q_i^{T_g}$ from the global dialogue channel and $q_i^{T_s}$ from the speaker channel are concatenated and then fed into a classifier. This classifier utilizes a Fully Connected layer (FC) with a *softmax* function for emotion prediction. The cross-entropy loss function is employed to supervise the training of the model:

$$u_i(pred) = softmax(FC([q_i^{T_g}, q_i^{T_s}])). \tag{13}$$

## IV. EXPERIMENTS

### A. DATASET AND SETTINGS

Considering the complexity of emotional cognition, exploring the emotional influence among multiple speakers is challenging as the emotional effect stems not only from a single speaker. Additionally, to validate the model's capacity for context modeling, each dialogue should have sufficient turns. Considering these factors, this research utilizes the widely-used Interactive Emotional Dyadic Motion Capture database IEMOCAP [25]. IEMOCAP consists of 5 sessions that capture dialogues between two participants. We used four sessions for training and one for testing, conducting cross-validation for each session. We identified six emotion categories: happy, sad, neutral, angry, excited, and frustrated. Table 1 illustrates the distribution of samples within IEMOCAP.

In the Introduction, it is mentioned that emotional cognition is a complex process. Emotion cognition is not solely derived from tone and intonation but also from the information conveyed by text. Therefore, this study uses acoustic and linguistic features as input to ensure the diversity of cognitive sources. We conducted experiments using the same data and features to compare our baseline models fairly. We utilized BERT-base [41] and Wav2vec2.0-base [42] as feature extractors, producing 768-dimensional text and audio features as input, and replicated the published code. Our parameter settings included using BiLSTM with two layers and 200 hidden units for context generation in both the dialogue-level and speaker-level. The map layer was fully connected with 100 neurons, while the batch size was 32. We set the dropout rate to 0.2, the learning rate for the Adam optimizer to 0.0001, and utilized L2 weight decay of 0.0002 as the optimizer. The remaining model parameters were set following the paper [10], [12], and [13]. We set the number of reasoning modules for dialogue and speaker levels as a hyperparameter, which we will evaluate and analyze in ablation experiments.

Considering the imbalance in data distribution across different emotional categories, this study employs accuracy and weighted F1-score as evaluation metrics.

### B. BASELINES FOR COMPARISON

To ensure fairness, this experiment replicated existing Emotion Recognition in Conversation (ERC) models for comparison.

DialogueCRN [12] is a model that uses multi-turn reasoning modules to extract speaker and global emotional information from context separately.

DialogueRNN [13] is a dynamic model that recognizes emotions by extracting global context and speaker states over time.

CMN [17] is a memory network that captures the context of different speakers in a conversation.

ICON [11] is an LSTM-based model that preserves the order of utterances to generate contextual features for emotion classification at the utterance level.

bc-LSTM [10] uses LSTM to create contextual features, preserving the order of utterances and providing contextual information for emotion classification at the utterance level.

### C. RESULTS AND DISCUSSIONS

Table 2 shows that the performance of DialoguePCN is significantly better than the baseline model. Compared with DialogueCRN, DialoguePCN's 5-fold accuracy increased by 3.14%, and the weighted F1 score increased by 3.96%. In conjunction with the experimental results given in Table 2, this paper provides the following explanation:

The bc-LSTM model ignores the impact of the context range on the current discourse and lacks a mechanism to separately model for the emotions of different speakers, thus yielding the worst results. CMN and DialogueRNN adopt speaker-level modeling but do not account for dynamic contextual information. ICON and DialogueCRN consider both the global conversational context and speaker-level context and perform multiple emotional retrievals, but their context is still fixed. In summary, baseline methods lack the utilization of already extracted emotional cues, leading to the incapacity of contextual information to adjust based on the recognition result of the previous moment. This results in significant interference from emotion-irrelevant information and an increased difficulty extracting emotional cues. The strength of DialoguePCN lies in the activation module(AM) that provides a unique context initialization state for each sentence. Even when the reasoning module fails to acquire effective emotion cues, the model can still draw on the context initialization state information to capture context information. When the reasoning module retrieves invalid emotion cues, the context initialization state information can diminish the proportion of irrelevant emotion cues in the final emotional representation. Additionally, DialoguePCN uses the emotion information retrieved each time for speech emotion recognition and employs a dynamic long-term memory update module(DMM) to store emotional cues in the context information. Doing so facilitates the retrieval of all previous emotional cues during the cognition phase of the discourse at the next moment, effectively mitigating the cold start problem and overcoming the issue of long-distance emotional cue forgetting.

**TABLE 2. Experimental results on the IEMOCAP dataset.**

| Methods | Modality | Acc(%) | Weight-F1(%) |
|---|---|---|---|
| CMN [17] | A+T+V | 61.90 | 61.40 |
| ICON [11] | A+T+V | 64.00 | 63.50 |
| bc-LSTM | A+T | 59.03 | 59.03 |
| DialogueRNN | A+T | 60.69 | 61.36 |
| DialogueCRN | A+T | 65.56 | 65.13 |
| DialoguePCN | A+T | **68.70** | **69.09** |

T: Text A: Audio V: Video

Fig. 4 displays the confusion matrix of the baseline models on the test set to assess DialoguePCN's recognition performance across various emotions. DialoguePCN excels in predicting instances of the "Happy" category, although some samples are misclassified as "Excited." The distinction between happy and excited can be challenging in real-life scenarios. We believe the fundamental difference between these two emotions lies in the intensity of the emotion. However, due to the diversity in human cognitive levels, using a clear intensity boundary to differentiate between the categories of "Happy" and "Excited" is difficult. The "Sad" category poses a challenge as DialoguePCN wrongly identifies 56 samples as "Frustrated" due to their similar acoustic characteristics of low intensity, pitch, and tone. In addition, we have also observed two significant phenomena: (1)In a dialogue with a negative emotional atmosphere, utterances expressing "sad" and "Frustrated" often alternate. (2) In the IEMOCAP dataset, each sample is annotated by three annotators, with the final label determined by majority voting. We noticed a recurring trend where a single annotator often labels "sad" samples as "Frustrated" and vice versa. The phenomena above indicate that the features of "sad" and "Frustrated" are easily confused. However, DialogueRNN and DialogueCRN misclassify several samples as "Neutral" or "Happy," which is unacceptable. For the "Angry" category, DialoguePCN outperforms all other models. However, like the other models, it also misclassifies many samples as "Frustrated." This

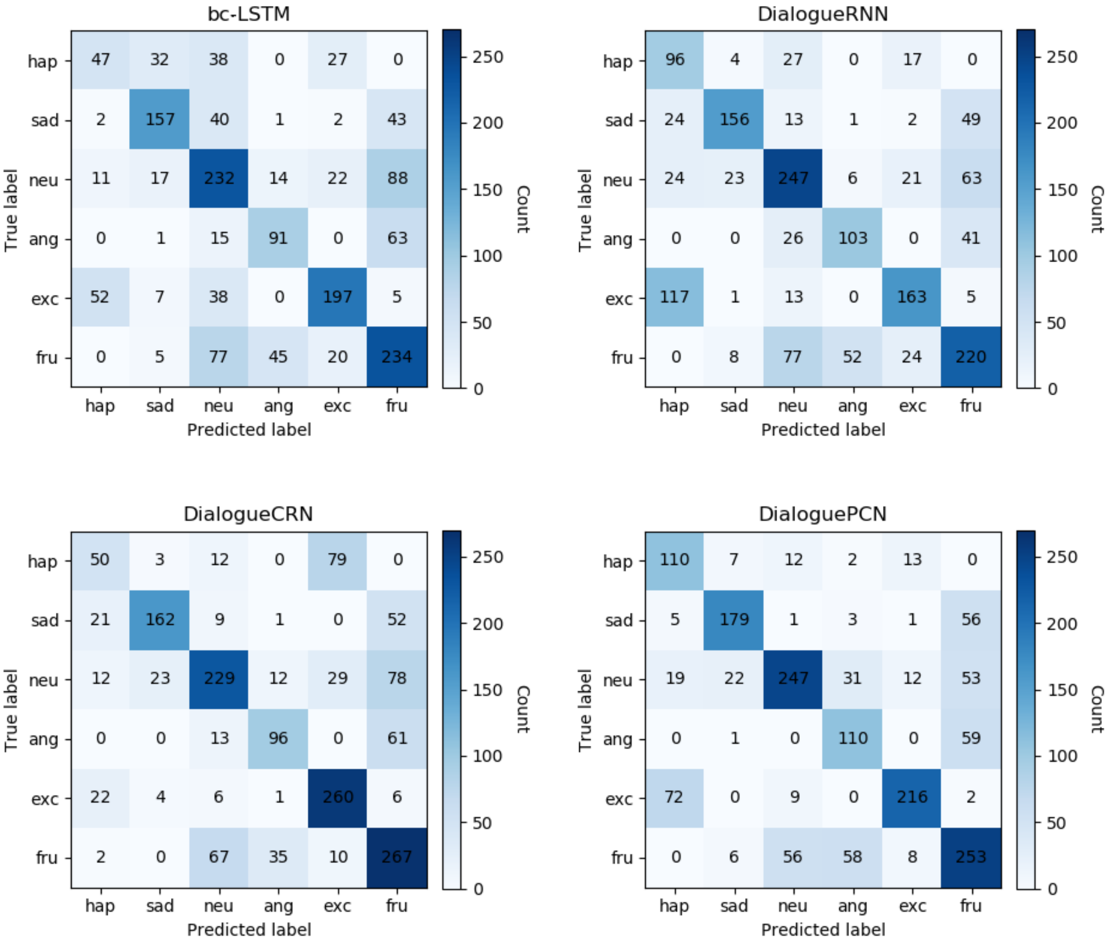

**FIGURE 4.** Confusion matrix for all models.

**TABLE 3.** Emotion category results for all models. The best results(%) are marked in bold.

| Models | | bc-LSTM | RNN | CRN | PCN |
|---|---|---|---|---|---|
| Happy | Acc. | 32.64 | 66.67 | 34.72 | **78.47** |
| | We-F1. | 36.72 | 47.41 | 39.72 | **62.95** |
| Sad | Acc. | 64.08 | 63.67 | 66.12 | **76.73** |
| | We-F1. | 67.67 | 71.40 | 74.14 | **77.69** |
| Neutral | Acc. | 60.42 | **64.32** | 59.64 | 64.06 |
| | We-F1. | 56.31 | 62.77 | 63.61 | **69.49** |
| Angry | Acc. | 53.53 | 60.59 | 56.47 | **62.35** |
| | We-F1. | 56.70 | **62.05** | 60.95 | 59.72 |
| Excited | Acc. | 65.89 | 54.52 | **86.96** | 71.57 |
| | We-F1. | 69.49 | 61.98 | 76.71 | **77.96** |
| Frustrated | Acc. | 61.42 | 57.75 | **70.08** | 65.09 |
| | We-F1. | 57.49 | 57.97 | **63.12** | 62.71 |
| Avg. | Acc. | 59.03 | 60.69 | 65.56 | **68.70** |
| | We-F1. | 59.03 | 61.36 | 65.13 | **69.09** |

RNN: DialogueRNN CRN: DialogueCRN PCN: DialoguePCN

misclassification results from the textual similarity between utterances conveying "Frustrated" and "Anger"-related emotions. For the "Excited" category, DialoguePCN mainly misidentifies instances as "Happy." For the "Frustrated" category, DialoguePCN mostly misclassifies instances as "Neutral" or "Angry." While some frustrated samples share similar acoustic features with neutral samples, the co-occurrence of these two emotions within the same conversation may contribute to mutual interference.

Table 3 presents the accuracy and weighted F1 scores for each emotion category. It is worth noting that both bc-LSTM and DialogueRNN have lower accuracy rates for the "Happy" category, with rates of 32.64% and 34.72%, respectively. This phenomenon can be explained by the infrequent "Happy" instances in the IEMOCAP dataset. Another contributing factor is the simplicity of the bc-LSTM and DialogueRNN model architectures, which struggle to effectively retrieve and integrate emotional cues for the "Happy" category from the limited data available. However, except for the "Frustrated" and "Angry" categories, DialoguePCN achieves the highest weighted F1 scores across all other emotion categories.

## D. ABLATION STUDY

### 1) MODULE PERFORMANCE TEST

Previously, Hu et al [12]. confirmed the importance of perception and cognition phases through ablation experiments.

**TABLE 4.** Experimental results of modules performance.

| AM | DMM | Acc(%) | Weight-F1(%) |
|---|---|---|---|
| × | × | 65.56 | 65.13 |
| √ | × | 67.34 | 67.03 |
| × | √ | 68.05 | 68.24 |
| √ | √ | **68.70** | **69.09** |

AM: Activation Module DMM: Dynamic long-term Memory Update Module

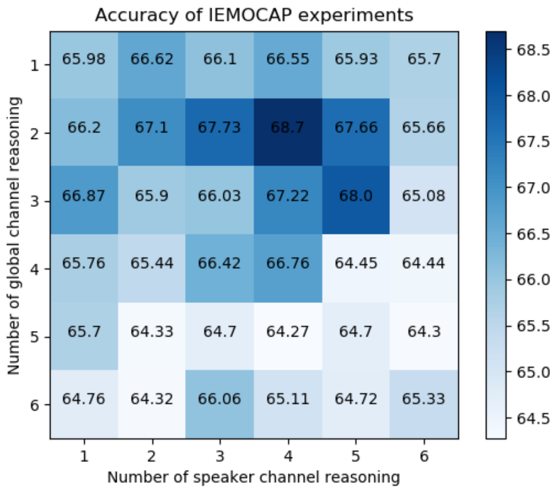

**FIGURE 5.** Test results for the number of reasoning modules.

In this section, we will analyze how the Activation Module (AM) and Dynamic Long-Term Memory Update Module (DMM) affect the performance of DialoguePCN. The results are presented in Table 4.

The first row shows the DialogueCRN model without the proposed AM and DMM. The reproduced accuracy and weighted F1 score are 65.56% and 65.13%, respectively. After adding only the Activation Module to the DialogueCRN model in the second row, we found that the accuracy improved by 1.78%, and the weighted F1 improved by 1.90%. Our interpretation suggests that the Activation Module has an advantage over DialogueCRN because it allows context information to be used when the multi-turn reasoning module does not effectively retrieve emotional cues. Furthermore, the activated memory can reduce the weight of irrelevant emotional cues, enhancing the model's ability to resist interference. In the third row, the model only incorporates the Dynamic Long-Term Memory Update Module (DMM), resulting in an accuracy improvement of 2.49% and a weighted F1 improvement of 3.11%. The DMM continuously adds previously retrieved emotional cues from the last time step utterance to the current long-term memory, making it easier to retrieve emotional cues from the current context. It also solves the issue of forgetting distant emotional information in the conversation. This strategy can be applied to other temporal prediction problems as well.

### 2) REASONING TURNS TEST

In this section, we are examining how the number of multi-turn reasoning modules affects performance. Fig. 5 shows that the most effective performance is achieved using 2 dialogue-level reasoning iterations ($T_g$) and 4 speaker-level reasoning iterations ($T_s$). Performance gradually improves as $T_g$ ranges from 1 to 3 and $T_s$ ranges from 3 to 5. However, excessive increasing the number of reasoning iterations can lead to a decline in model performance due to the introduction of irrelevant emotional cues. This decline is largely caused by overfitting. Our analysis is as follows: when the number of iterative inferences is too small, the model cannot achieve optimal performance as more distant and deeper emotional cues have yet to be retrieved. Given that the attention mechanism learns emotion cues related to the current utterance from long-term memory at each step, an excess of iterative inferences would retrieve and integrate irrelevant emotional cues, leading to an overfitting problem and a subsequent decline in model performance. It is also noteworthy that increasing the number of speaker-level reasoning modules impacts model performance more than increasing the number of dialogue-level reasoning modules. This phenomenon shows that speakers' emotional inertia more significantly impacts emotional expression than their interaction with others.

## V. CONCLUSION

This research introduces a novel Dialogue Perception and Cognition Network (DialoguePCN) to simulate human conversations' emotional perception and cognitive processes. DialoguePCN employs an activation module in the perception phase to provide corresponding contextual information for each utterance to reduce interference from irrelevant contexts. In the cognition phase, the proposed dynamic long-term memory update module(DMM) is employed for the first time to utilize existing emotional reasoning results to reconstruct contextual information. This strategy makes it easier to retrieve emotions from context for subsequent utterances. In addition to addressing the cold start issue in ERC tasks, the DMM mitigates the problem of emotional cue forgetting caused by overly lengthy dialogues by iteratively preserving emotional cues.

## REFERENCES

[1] Y.-P. Ruan, S.-K. Zheng, T. Li, F. Wang, and G. Pei, "Hierarchical and multi-view dependency modelling network for conversational emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7032–7036.

[2] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7037–7041.

[3] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, Feb. 2022.

[4] Q. Li, D. Gkoumas, A. Sordoni, J.-Y. Nie, and M. Melucci, "Quantum-inspired neural network for conversational emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 15, pp. 13270–13278.

[5] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, 2021, pp. 13789–13797.

[6] K. P. Seng and L.-M. Ang, "Video analytics for customer emotion and satisfaction at contact centers," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 3, pp. 266–278, Jun. 2018.

[7] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "Caire: An end-to-end empathetic chatbot," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 9, pp. 13622–13623.

[8] G. Tu, B. Liang, D. Jiang, and R. Xu, "Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations," *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 1803–1816, 2023.

[9] F. Chen, Z. Sun, D. Ouyang, X. Liu, and J. Shao, "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1064–1073.

[10] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[11] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.

[12] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," 2021, *arXiv:2106.01978*.

[13] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6818–6825.

[14] B. Wang, G. Dong, Y. Zhao, R. Li, Q. Cao, K. Hu, and D. Jiang, "Hierarchically stacked graph convolution for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110285.

[15] X. Song, L. Zang, R. Zhang, S. Hu, and L. Huang, "Emotionflow: Capture the dialogue level emotion transitions," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8542–8546.

[16] J. Wen, D. Jiang, G. Tu, C. Liu, and E. Cambria, "Dynamic interactive multiview memory network for emotion recognition in conversation," *Inf. Fusion*, vol. 91, pp. 123–133, Mar. 2023.

[17] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, p. 2122.

[18] N. Majumder, D. Ghosal, D. Hazarika, A. Gelbukh, R. Mihalcea, and S. Poria, "Exemplars-guided empathetic response generation controlled by the elements of human communication," *IEEE Access*, vol. 10, pp. 77176–77190, 2022.

[19] F. Ren and T. She, "Utilizing external knowledge to enhance semantics in emotion detection in conversation," *IEEE Access*, vol. 9, pp. 154947–154956, 2021.

[20] I. Carvalho, H. G. Oliveira, and C. Silva, "The importance of context for sentiment analysis in dialogues," *IEEE Access*, vol. 11, pp. 86088–86103, 2023.

[21] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychol. Rev.*, vol. 69, no. 5, p. 379, 1962.

[22] R. S. Lazarus and C. A. Smith, "Knowledge and appraisal in the cognition—Emotion relationship," *Cognition Emotion*, vol. 2, no. 4, pp. 281–300, Oct. 1988.

[23] J. S. B. T. Evans, "Heuristic and analytic processes in reasoning," *Brit. J. Psychol.*, vol. 75, no. 4, pp. 451–468, Nov. 1984.

[24] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 255–278, Jan. 2008.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[26] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," 2021, *arXiv:2105.12907*.

[27] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[28] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: COmmonSense knowledge for eMotion identification in conversations," 2020, *arXiv:2010.02795*.

[29] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intell. Syst.*, vol. 36, no. 2, pp. 88–95, Mar. 2021.

[30] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1034–1047, Mar. 2022.

[31] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowl.-Based Syst.*, vol. 235, Jan. 2022, Art. no. 107643.

[32] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8002–8009.

[33] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 7692–7699.

[34] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 371–378.

[35] H. Abdel-Nabi, A. Awajan, and M. Z. Ali, "Deep learning-based question answering: A survey," *Knowl. Inf. Syst.*, vol. 65, no. 4, pp. 1399–1485, Apr. 2023.

[36] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14974–14983.

[37] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1426–1439, Jul. 2022.

[38] M. W. Morris and D. Keltner, "How emotions work: The social functions of emotional expression in negotiations," *Res. Organizational Behav.*, vol. 22, pp. 1–50, Jan. 2000.

[39] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychol. Sci.*, vol. 21, no. 7, pp. 984–991, Jul. 2010.

[40] C. Navarretta, "Mirroring facial expressions and emotions in dyadic conversations," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 469–474.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[42] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

**XIAOLONG WU** received the M.S. degree in software engineering, in 2019. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Xinjiang University. His research interest includes speech emotion recognition.

**CHANG FENG** received the B.S. degree in computer science and technology, in 2019. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. Her research interest includes speech deepfake detection.

**MINGXING XU** received the B.S. degree in computer science and technology and the Ph.D. degree in computer application technology from Tsinghua University, Beijing, China, in 1995 and 1999, respectively. He is currently an Associate Professor and the Associate Director of the Auditory Intelligence Research Center, Institute of Artificial Intelligence, Tsinghua University. His research and development interests include speech recognition, speaker recognition, emotion recognition, and natural language processing.

**THOMAS FANG ZHENG** (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 1997. He is currently a Research Professor and the Director of the Center for Speech and Language Technologies, Tsinghua University. He has authored more than 250 papers. His research interest includes speech and language processing. He plays active roles in a number of communities, including the Chinese Corpus Consortium (the Council Chair), the Standing Committee of China's National Conference on Man-Machine Speech Communication (the Chair), Subcommittee 2 on Human Biometrics Application of Technical Committee 100 on Security Protection Alarm Systems of Standardization Administration of China (the Deputy Director), the Asia–Pacific Signal and Information Processing Association (APSIPA) (the Vice-President and a Distinguished Lecturer, from 2012 to 2013), Chinese Information Processing Society of China (a Council Member and the Speech Information Subcommittee Chair), the Acoustical Society of China (a Council Member), and the Phonetic Association of China (a Council Member). He was an Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing and the *APSIPA Transactions on Signal and Information Processing*. He is on the editorial board of *Speech Communication*, *Journal of Signal and Information Processing*, *SpringerBriefs in Signal Processing*, and the *Journal of Chinese Information Processing*.

**ASKAR HAMDULLA** received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2003. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. His research interests include speech recognition and synthesis, pattern recognition and image processing, natural language processing, information retrieval, and content security.