第 38 卷　第 3 期
2023 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38　No. 3
Jun. 2023

DOI:10.13364/j.issn.1672-6510.20220189

融合情感与语义的多模态对话生成方法

张翼英，马彩霞，张　楠，柳依阳，王德龙
（天津科技大学人工智能学院，天津 300457）

摘　要：近年来，语音对话等一系列非可视化对话场景在生活中屡见不鲜，比如智能机器人的语音交互、各类客服通过语音对话了解客户需求等. 音频中往往蕴含情感信息，而文本中则包含丰富的语义层面的信息，因此基于语音文本多模态特征更能充分挖掘语义及情感信息，生成信息更加丰富的对话响应. 当前基于文本和音频的对话生成技术主要基于较传统的 Seq2Seq 模型实现，生成的响应存在多样性较低、上下文不够连贯等问题. 为此，本文提出 AT-Transformer 模型实现文本、音频多模态场景下的对话生成任务. 首先使用 WordEmbedding 对上下文和回复进行词嵌入矩阵的构建，然后使用 VGGish 对对话音频进行特征提取，接着分别将其输入 AT-Transformer 模型中，并在多模态注意力机制中实现两种模态特征的融合，最后设计目标函数旨在提高生成语句的多样性. 实验分别对情感丰富度、上下文语义相关性和句子连贯性进行评估，相较最优基准模型，情感匹配度提升 2%，语义相关性提升 0.5%.

关键词：多模态；对话生成；Transformer 模型；文本生成
中图分类号：TP389.1　　**文献标志码**：A　　**文章编号**：1672-6510(2023)03-0052-09

Multimodal Dialogue Generation Method Integrating Emotion and Semantics

ZHANG Yiying，MA Caixia，ZHANG Nan，LIU Yiyang，WANG Delong
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：In recent years，a series of non-visual dialogue scenes such as voice dialogue have become common in life，such as the voice interaction of intelligent robots and all kinds of customer service to understand customer needs through voice dialogue. Audio often contains emotional information，while text contains rich semantic information，so it is of certain research significance to integrate more comprehensive audio features in text dialogue generation task. At present，the dialogue generation technology based on text and audio is mainly based on the traditional Seq2Seq model，and the generated responses have some problems，such as low diversity and insufficient coherence of context. At-Transformer model is therefore proposed in this article to realize the dialogue generation task in multi-mode text and audio scenarios. WordEmbedding is first used to construct the WordEmbedding matrix for the context and reply，and VGGish is used to extract the features of the dialogue audio. Then the features are inputted to the AT-Transformer model proposed in the article，and the fusion of the two modal features is implemented in the multi-modal attention mechanism. Finally，the objective function is designed to improve the diversity of generated statements. The experiment evaluated the emotional richness，contextual semantic relevance and sentence coherence respectively. Compared with the optimal benchmark model，the emotion matching degree increased by 2% ，and the semantic relevance increased by 0.5% .

Key words：multimodal；dialogue generation；Transformer model；text generation

开放域对话由于应用范围广而受到产业界和学术界普遍关注[1]，语音对话具有省时、高效等优势，在开放域对话中发挥着越来越重要的作用. 语音模态是指以音频形式存储的说话内容，由于其中包括振

收稿日期：2022-07-29；**修回日期**：2022-11-22
基金项目：国家自然科学基金资助项目(61807024)
作者简介：张翼英（1973—），男，辽宁人，教授，yiyingzhang@tust.edu.cn

幅、频率等多种音频特征，这些特征包含说话人的重要信息，因此仅利用文本单模态生成对话往往无法满足要求. 如何利用音频和文本特征生成信息丰富且流畅的对话响应是值得探究的问题.

在文本单模态对话生成方面，基于门控循环单元（gate recurrent unit，GRU）构建 Seq2Seq 对话模型，其编码器将上下文文本编码至一个向量，解码器将该向量作为输入，并对信息解码，从而输出响应序列[2]. 但 Seq2Seq 不能很好地捕捉到上下文信息[3]，于是 HRED（hierarchical recurrent encoder-decoder）模型应运而生，该模型通过额外增加一个编码器对上下文建模，减少了相邻句子间的计算步骤，促进信息的传播[4]. 为了提升回复的多样性并控制回复的情感倾向，条件变分自动编码（conditional variational auto encoder，CVAE）模型往往结合注意力机制或 Seq2Seq 模型实现指定情绪的响应生成[5-6]. 但 Seq2Seq 模型和 HRED 模型对长句的生成效果较差，为解决这一问题，Google 团队于 2017 年提出自注意力机制和 Transformer 序列到序列模型[7]. 该模型能够并行提取其他位置的信息，并将信息进行加权平均化，再和当前位置进行融合，在对话生成、情感识别等多种任务上的运行效果均有较大提升；在多模态对话任务中，研究人员应用 CVAE 模型根据多模态条件和给定情感信息生成连贯的对话响应[8]，但该研究仅仅将模态间进行线性连接，未考虑不同模式之间的交叉融合，不能深入挖掘模态内部的关联. 文献[9]利用音频辅助文本进行对话生成，提出融合音频的 Audio-Seq2Seq 文本对话生成模型，将文本嵌入向量和音频向量同时输入注意力模块，探究振幅及响度对于对话生成情感的重要性. 由于该研究的基础是 Seq2Seq 模型，因此仍存在对话较为通用、多样性较差等问题. 此外，不少学者致力于研究多模态 Transformer，其中有研究[10]采用基于 Transformer 的自监督多模态表示学习框架 VATT（video-audio-text transformer）实现了多模态视频的有效监督，但该方法常被应用于图像相关的下游任务；文献[11]提出了视听场景感知对话（audio-visual scene-aware dialog，AVSD），通过引入多任务学习实现多模态对话生成，但该方法将语音模态进行了单向映射，未将音频特征与文本特征充分融合，因此生成的对话不能囊括音频特征中丰富的情感信息.

为了解决上述问题，本文提出文本音频 Transformer（audio text transformer，AT-Transformer）模型实现音频和文本双模态的对话生成，该模型的编码器将文本和音频双模态进行模态间和模态内部特征融合，区别于已有的线性连接方法，实现了模态间特征关联性的深入挖掘. 为了验证模型的有效性，在 IEMOCAP 数据集[12]上进行了实验，通过与基于纯文本的 Transformer 模型和基于音频、文本多模态的 Audio-Seq2Seq、VATT 和 AVSD 模型进行困惑度及生成多样性比较，并从语义相关性、流畅度和情感匹配性 3 个方面进行人工评估. 实验结果表明，本文模型能够生成内容丰富、情感适宜的响应.

本文的主要工作如下：

（1）提出多模态注意力机制，探究文本特征和音频特征之间的深入关联，使得文本生成任务能够充分融入音频特征所包含的潜在信息.

（2）从语句生成的多样性方面提升对话生成效果，避免生成通用性、无意义的回复.

（3）通过灰度对数功率谱图、Mel 频谱图、Mel 频率倒谱系数（Mel frequency cepstrum coefficient，MFCC）图与注意力热力图的对比验证了语音频率、基频、共振与注意力之间的正向关系，表明语音模态能够明显促进对话生成质量的提升.

1　相关研究

1.1　对话生成

随着深度学习技术的快速发展和算力的提升，许多学者致力于研究对话生成技术，按生成的依据可以将这些技术划分为纯文本对话生成和多模态对话生成两种方式. 纯文本对话生成通过对文本数据的分析和处理，进而生成响应的过程. 传统的 Seq2Seq 模型对上下文信息的依赖有限，生成的响应存在无意义、内容不丰富等问题，而 HRED 模型将 Seq2Seq 模型进行层次化改进，提升了对上下文信息的关注度，进而提高了多轮对话的生成效率. 变分自编码器（variational auto encoder，VAE）通过将潜在特征表述为概率分布的方式更适合对话上下文内部状态的表示，条件变分自编码器（CVAE）结合双重注意力机制能够将上下文响应和随机的潜在变量连接，有效地控制响应的情感倾向[13]. 为了解决 Seq2Seq 模型和 HRED 模型对长句及多轮对话生成效果不佳的问题，Transformer 模型通过多头注意力机制关注当前的词和句子中的其他词，可以有效获取上下文语义信息[7]. 尽管这些模型取得了较好的对话效果，但是并未考虑语音模态，可能会存在对上下文语义感知不准确的问题，故而对对话生成质量造成影响.

多模态对话生成以视频、音频、微表情、文本等多种模态特征为依据，通过模态融合建模不同模态之间的关系，进而生成适合不同场景的回复，具有广阔的研究前景[14]. Wang 等[2]通过视觉模型提取视觉特征，并将其输入序列到序列的对话生成中，学习在给定文本和视觉上下文情况下生成下一语句的概率. Chen 等[5]使用文本实体定位图像中的相关对象，建立文本与对象之间的映射，并通过跨模态注意力机制构建多模态 Transformer，从而生成与视觉和文本上下文一致的响应. 除了视频模态之外，文献[9]对音频上下文进行建模，并提出音频增强的 Seq2Seq 模型，实现对话生成任务，验证了音频特征对于对话生成的有效性. 上述研究虽然能够产生效果较好的响应，但是未对语音模态进行考虑，并且序列到序列的模型存在生成多样性较差、语义不丰富等问题. 本文工作区别于已有工作，通过应用多模态融合实现音频和文本模态间特征的深度挖掘，从而构建多模态注意力机制 AT-Transformer 模型，经验证双模态特征比纯文本特征实现了对话质量和情感匹配度的显著提升.

1.2　多模态融合

多模态融合是将音频、视频、微表情等多形态数据进行综合处理的过程，是多模态对话生成的基础[13]. 模型相关的融合方法虽然复杂性较高，但具有较强的实用性和较高的准确率. Rohanian 等[15]使用长短期记忆(long short-term memory，LSTM)网络对文本中的词汇信息和音频中的声学特征进行顺序建模，实现阿尔茨海默病的检测. Shen 等[16]通过构建 LSTM 网络交互单元，对音频和文本之间的动态交互进行建模，实现语音情感的准确分类. 由于上述研究对文本和音频特征进行顺序建模，未考虑特征间的深层交互关系，并且 LSTM 网络仍存在梯度消失及梯度爆炸问题. Saha 等[17]提出基于自身、模态间和任务间注意力机制的多模态多任务深度神经网络，实现情感和任务类别的联合学习. 该模型实现了对话行为及情感的准确分类以及模态间的深层融合，但未对主要关联部分进行探究，并且对话生成任务还需补充解码器部分. 本文通过将 Transformer 的编码器部分的多头注意力机制部分进行跨模态设计，并通过实验分析不同参数的重要性程度，促进对话生成质量的进一步提升.

2　任务定义

本文的目标是通过音频、对话上下文两种模态信息生成内容丰富、具有一定情感并且流畅的回复. 该任务定义为：D_{A}(dialogue audio)表示当前对话单位音频片段；D_{T}(dialogue text)为当前对话音频 D_{A} 所对应的文本；R(response)表示在给定对话单位音频片段 D_{A} 和对话文本 D_{T} 的前提下生成的对话响应文本，其中包含 m 个单词，即 $R=\{r_1,r_2,\cdots,r_m\}$. 则在给定对话音频片段 D_{A} 和对应文本 D_{T} 的情况下生成响应文本 R 的概率表示为

$$P(R\mid D_{\mathrm{A}},D_{\mathrm{T}};\theta)=\prod_{i=1}^{m}P(r_i\mid D_{\mathrm{A}},D_{\mathrm{T}},r^{<i};\theta)\tag{1}$$

其中：$r^{<i}$ 表示响应文本 R 中的前 $i-1$ 个单词，θ 为可训练的参数.

3　模型描述

本文提出一种基于多模态注意力机制的 AT-Transformer 模型，综合考虑文本、音频双模态，旨在探究语音模态对于对话生成效果的影响. 实验证明，利用该生成模型能够生成内容丰富、情感适宜并且流畅的对话回复. 该模型在传统 Transformer 的基础上提出多模态注意力机制，并设计情感和内容相关的目标函数，采用核采样算法提升回复的多样性，整体架构如图 1 所示.

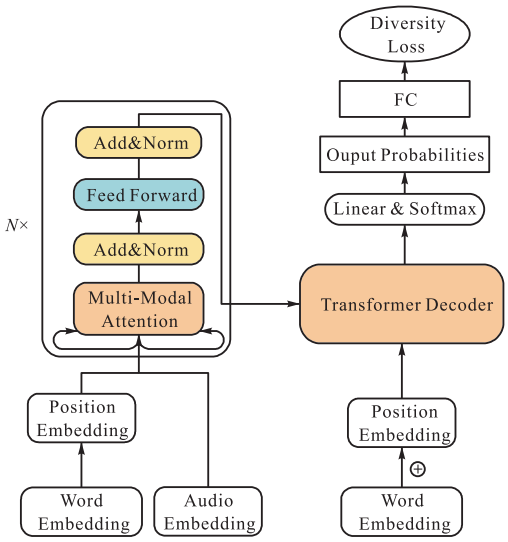


图 1　AT-Transformer 模型整体结构

Fig. 1　Overall structure of AT-Transformer model

该模型主要分为 3 个部分：第 1 部分对数据进行预处理，将音频数据缺失的数据进行过滤，通过计算 Mel 声谱，进行卷积操作获得嵌入向量，同时对文本数据设置最大单词长度，进行特征提取，然后将其进行嵌入向量表示；第 2 部分将文本嵌入向量和音频嵌

入向量输入生成模型进行训练，在训练过程中通过多模态注意力机制实现音频和文本特征的融合；第 3 部分通过多样性损失函数提升对话生成质量.

4　AT-Transformer 模型

4.1　多模态融合

对音频中的特征进行提取，需要考虑说话者的态度、情感色彩的变化、对应的声音形式、语调及说话节奏等特征[18]，而 Mel 频谱图更接近人类感知音高的方式，因此本文使用 VGGish 对该特征进行提取，并通过卷积操作获取音频向量的嵌入表示，最后使用主成分分析法进行特征降维，从而在编码器中实现特征融合.

为了使对话文本向量携带相应的顺序信息，文本表示由单词(w)嵌入和位置嵌入构成，具体如图 1 中的输入部分所示.

$$\boldsymbol{C}_{w_i}=\boldsymbol{E}_{w_i}+\boldsymbol{E}_i \tag{2}$$

其中：$\boldsymbol{C}_{w_i}$ 为第 i 个单词的文本表示，$\boldsymbol{E}_{w_i}$ 是 w_i 的单词嵌入，$\boldsymbol{E}_i$ 是第 i 个单词的位置嵌入.

4.2　多模态注意力机制

目前已有的多模态注意力机制主要是将不同的模态进行一维卷积操作，并将不同的模态进行跨模态操作并投影至同一模态，然后将该模态下的所有特征进行连接，再进行自注意力操作实现多模态特征融合[19]. 虽然该方法实现了特征的有效融合，但是这种方法进行了两次跨模态操作，计算复杂度较高.

为了使音频特征和文本特征进行有效融合，本文使用多模态注意力计算的方法，将文本和音频分别进行嵌入向量表示，并通过注意力分数体现二者之间的关系. 本文在 Transformer 模型[7]的基础上对其中的多头注意力机制进行改进，其中 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别代表注意力中的查询、键和值，多模态注意力机制的结构如图 2 所示，其中的蓝色圆形表示输入的文本向量，橙色圆形表示输入的音频向量.

将文本模态和音频模态分别表示为 c 和 a，二者的输入分别为 $\boldsymbol{X}_c\in\mathbb{R}^{T_c\times d_c}$ 和 $\boldsymbol{X}_a\in\mathbb{R}^{T_a\times d_a}$，其中 $T(\cdot)$ 和 $d(\cdot)$ 分别为序列长度和特征维度. 将注意力中的查询、键和值定义为

$$[\boldsymbol{Q}_c\quad\boldsymbol{K}_c\quad\boldsymbol{V}_c]=[\boldsymbol{X}_c;\boldsymbol{X}_c;\boldsymbol{X}_c]\begin{bmatrix}\boldsymbol{W}_{\boldsymbol{Q}_c}\\ \boldsymbol{W}_{\boldsymbol{K}_c}\\ \boldsymbol{W}_{\boldsymbol{V}_c}\end{bmatrix} \tag{3}$$

$$[\boldsymbol{Q}_a\quad\boldsymbol{K}_a\quad\boldsymbol{V}_a]=[\boldsymbol{X}_a;\boldsymbol{X}_a;\boldsymbol{X}_a]\begin{bmatrix}\boldsymbol{W}_{\boldsymbol{Q}_a}\\ \boldsymbol{W}_{\boldsymbol{K}_a}\\ \boldsymbol{W}_{\boldsymbol{V}_a}\end{bmatrix} \tag{4}$$

其中：$\boldsymbol{Q}_c$、$\boldsymbol{K}_c$ 和 $\boldsymbol{V}_c$ 分别是文本模态所对应的查询、键和值，$\boldsymbol{Q}_a$、$\boldsymbol{K}_a$ 和 $\boldsymbol{V}_a$ 为音频模态所对应的查询、键和值，权重矩阵 $\boldsymbol{W}_{\boldsymbol{Q}}\in\mathbb{R}^{d\times d_q}$、$\boldsymbol{W}_{\boldsymbol{K}}\in\mathbb{R}^{d\times d_k}$、$\boldsymbol{W}_{\boldsymbol{V}}\in\mathbb{R}^{d\times d_v}$. 模态内部和模态之间的多头注意力表示为

$$\boldsymbol{C}_c=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_c\boldsymbol{K}_c^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right) \tag{5}$$

$$\boldsymbol{A}_c=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_c\boldsymbol{K}_a^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right) \tag{6}$$

$$\boldsymbol{A}_a=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_a\boldsymbol{K}_a^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right) \tag{7}$$

$$\boldsymbol{C}_a=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_a\boldsymbol{K}_c^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right) \tag{8}$$

其中：$\boldsymbol{C}_c$ 和 $\boldsymbol{A}_a$ 分别为文本和音频模态内部计算所得注意力，$\boldsymbol{A}_c$ 和 $\boldsymbol{C}_a$ 为文本和音频两种方式的跨模态注意力，d_{k} 为输入向量的维度. 然后，将式(5)—式(8)与对应模态的值进行向量乘积，此处以 $\boldsymbol{A}_c$ 为例，赋值后 $\boldsymbol{A}_c'$ 为

$$\boldsymbol{A}_c'=\boldsymbol{A}_c\boldsymbol{V}_a \tag{9}$$

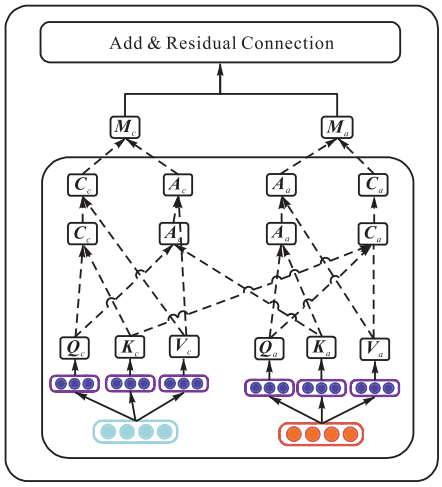


图 2　多模态注意力机制

Fig. 2　Multimodal attention mechanism

公式(9)将文本和音频两种方式的跨模态注意力分别与对应模态的值进行向量乘积，分别求取每部分的最终注意力值，实现模态之间的深度融合[7]. 同时，受文献[20]的启发，将模态之间的注意力向量进行连接，实现语音与文本之间注意力机制的深度挖

掘，保证了模态融合的完整性，公式为

$$R_{\mathrm{MM}}=\mathrm{concatenation}(\boldsymbol{M}_c,\boldsymbol{M}_a)=\mathrm{concatenation}(\boldsymbol{C}_c,\boldsymbol{A}_c,\boldsymbol{A}_a,\boldsymbol{C}_a) \tag{10}$$

式中：R_{MM} 为最终的语音文本模态注意力计算结果，$\boldsymbol{M}_c$ 和 $\boldsymbol{M}_a$ 分别为文本模态和音频模态的注意力，$\boldsymbol{C}_c$、$\boldsymbol{A}_c$、$\boldsymbol{A}_a$ 和 $\boldsymbol{C}_a$ 分别为依据式(9)进行向量乘积之后的计算结果.

4.3　多样性损失函数

对话生成任务通常以 softmax 交叉熵作为损失函数，倾向于从候选集中生成频率最高的语句作为响应，从而出现生成的语句无意义、重复性较高等问题. 为了提高生成语句的多样性，在原损失函数的基础上考虑了单词的频率，同时通过动态调整参数实现对目标单词索引权重的配置，进而控制损失函数 L_{e} 的收敛速度，其中该部分模型架构图 1 中的全连接层 FC，公式为

$$L_{\mathrm{e}}=w_t L_{\mathrm{s}} \tag{11}$$

$$w_t=\frac{1}{\mathrm{e}^{\lambda f(t_t)}} \tag{12}$$

$$L_{\mathrm{s}}=\log\left(\frac{\exp(x_t)}{\sum_i^{|V|}\exp(x_i)}\right) \tag{13}$$

其中：L_{s} 为 softmax 交叉熵损失函数，x 是 softmax 层之前预测层的输出，x_i 是 x 集合（$x\in\mathbb{R}^{|V|}$）中的第 i 个单词，t 是目标单词的索引. w_t 是 t 所对应的权重，t_t 是 t 所对应的单词，$f(t_t)$ 是 $token_t$ 在训练集中出现的频率，λ 为控制频率影响大小的超参数. 在公式(12)中，由于 e^{λ} 能够通过调整 λ 的大小控制权重 w_t 的变化速度，进而控制损失函数的收敛速度，同时当 $\lambda=0$ 时，该损失函数与 softmax 交叉熵损失函数相同.

5　实　验

5.1　数据集

本研究使用 IEMOCAP 作为数据集，该数据集包含 12 h 的试听数据，参与者在其中进行即兴表演或根据脚本场景表演，其中包含 5 个会话. 由于该数据集包含文本和音频双模态并且具有情感标签，探讨音频特征对于对话生成文本是否具有情感因素方面的作用有一定的帮助，本研究将后 4 个 session 作为训练集，session1 作为测试集，IEMOCAP 数据集的初始对话数、预处理后的对话数和词汇大小见表 1.

5.2　数据预处理

首先对 IEMOCAP 中的不规范文本数据及相对应的音频数据进行过滤，然后对不完整的音频数据及对应的文本数据进行过滤，通过观察音频数据的时长及文本特征长度，将特征维度进行对齐，其中文本数据的维度为 90，音频数据的维度为 90×128，学习率设置为 1×10^{-4}.

表 1　IEMOCAP 数据集的初始对话数、预处理后的对话数和词汇大小

Tab. 1　Initial utterance number, preprocessed utterance number and vocabulary size of the IEMOCAP dataset

类别	训练集	测试集
预处理后的对话数	4 053	1 106
词汇大小	8 104	1 063
初始对话数	4 111	1 121

5.3　实验评估

5.3.1　困惑度和多样性评估

开放域对话生成任务的自动评估方法一直以来都面临着挑战，而人工评估方法成为一个较为可靠的评估标准.

本实验主要进行了困惑度（perplexity，PPL）[21]和多样性两方面的自动评估. 对于一个由词语序列组成的句子，困惑度计算公式为

$$P_P(s)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i\mid w_1w_2\cdots w_{i-1})}} \tag{14}$$

其中：s 为候选句子，N 为候选句子 s 的长度，$P(w_i)$ 为第 i 个词的概率，第一个词为 $P(w_1\mid w_0)$，w_0 为句子开始占位符. 该方法用于估算模型的信息密度，检测对话生成语句相对于已有参考语句的平均生成质量，困惑度越小，语言模型越好.

回复多样性作为影响对话质量的关键要素之一，在开放式对话任务中备受关注，传统的 Seq2Seq 模型产生的回复往往会面临回复语句单一、枯燥乏味等问题，这严重影响用户体验，而 Transformer 模型能够在一定程度上缓解这一问题. 本实验主要采用 Distinct[22]方法对回复多样性 D_n 进行评估.

$$D_n=N_{\mathrm{D}}/N_{\mathrm{S}} \tag{15}$$

其中：N_{D} 为回复语句中不重复的 n-gram 的数量，N_{S} 为回复语句中 n-gram 词语的总数目. 式(15)分别对回复中不同的单个单词和两个单词进行统计，并将该数目分别除以各自相对应的总数，本实验中主要采用 D_1 和 D_2 计算回复中的内容多样性.

为了验证所提模型的对话生成质量及对话多样性效果，本文选取模型 Transformer、Audio-Seq2Seq、VATT、AVSD 进行实验比较，分别对其进行困惑度及

多样性评估，其中 Transformer 模型未考虑音频模态，而 Audio-Seq2Seq、VATT、AVSD 模型均考虑了文本和音频模型，具体比较数据见表 2. 实验结果表明，基于 AT-Transformer 模型相较于传统 Transformer 及其他各类多模态模型在困惑度和多样性均有一定提升，与表 2 中标红的其他模型的最佳实验结果相比，困惑度降低了 0.2%，D_1 和 D_2 分别提升了 0.06 和 2.7%. 总体来看，本文提出的模型在困惑度和 D_2 上的性能提升较为明显.

表 2　不同模型在 IEMOCAP 数据集上的困惑度和多样性比较

Tab. 2　Comparison of AT-Transformer's perplexity and diversity on IEMOCAP data sets with different models

模型	困惑度	D_1	D_2
Transformer	28.58	0.072 97	0.339 7
Audio-Seq2Seq	28.63	0.067 85	0.341 2
VATT	29.98	0.072 93	0.323 1
AVSD	28.79	0.073 01	0.342 1
AT-Transformer	28.51	0.073 51	0.352 6

5.3.2　人工评估

由于情感相关性与内容相关性很难通过自动评估方法进行考量，为了验证文中模型的效果，从语义相关性、流畅度和情感匹配度三方面进行评估[13]，邀请 5 名具有对话相关工作经验的人进行评分，根据语句与人工回复语句的近似程度及内容丰富性评分，其中分数分别为 0、1、2，如果语句中包含的内容较多但不偏离核心话题，那么其得分越高；对于流畅度，人工主要依据其可读性将其分数判定为 0、1、2；情感匹配度主要是由评分者判定生成语句情感和对话数据本身情感是否匹配，如果强匹配则评分为 2，若情感倾向一致，但有一点偏离，则评分为 1，若情感倾向完全不同则评分为 0. 回复在语义相关性、流畅度和情感匹配度中的达标程度见表 3.

表 3　回复在语义相关性、流畅度和情感匹配度中的达标程度

Tab. 3　Degree to which the response meets the criteria for semantic relevance, fluency and emotional matching

模型	语义相关性/%	流畅度/%	情感匹配度/%
Transformer	33.2	29.20	44.2
Audio-Seq2seq	32.9	29.30	45.2
VATT	32.4	28.60	45.5
AVSD	32.9	26.89	44.8
AT-Transformer	33.7	29.19	47.5

实验数据表明，相较于最优基准模型，本文模型在情感匹配度上提升 2%，在流畅度及语义相关性方面与纯文本特征生成的回复效果基本持平，表现为语义相关性提升 0.5%，而流畅度则下降 0.11%. 由此可看出音频特征的增加对于提升对话的情感匹配度有一定的作用，而文本特征嵌入向量具有充分的表示能力，因此增加音频特征之后并不能使流畅度显著提升. Transformer 模型与 AT-Transformer 模型生成的回复对比见表 4.

表 4　Transformer 模型和 AT-Transformer 模型在系统中生成的回复对比

Tab. 4　Comparison of responses generated by Transformer model and AT-Transformer model

上下文	Transformer 模型生成的回复	AT-Transformer 模型生成的回复
I want something to turn out the way it is suppose to turn out.	Do you think this is what I had in mind?	You welcome to the human race. Do you think this is what I had in mind when I proposed? That four years down the road，we'd be at the beach yelling at each other over fish?
I'm trying to work this backwards. But I can't seem to pinpoint the time where everything went wrong here.	You were fine during the phone call from marge.	You were fine during the phone call from marge. Let's see，I took a shower. Uh，you watched the news. you said the grunions are running tonight. I said great，let's go. I feel lucky tonight.
You have to tell me.	It's kind of all mixed up with all these other things.	I'm embarrassing you. I didn't want to tell it to you here. I wanted some place we'd never been before. A place where we'd be brand new to each other.
No，because if she- he's not coming back then I'll kill myself. Laugh! Laugh at me!	Laugh at me. but why did that happen the very night she came here?	But why did that happen the very night she came here? She sleeps in his room and his memorial breaks to pieces. Just look. Look at it.

在表 4 中分别针对 4 组上下文将 Transformer 模型和 AT-Tranformer 模型生成的回复进行对比，从中可以看出本文所提出的模型包含的内容较为丰富，例如，针对第 4 组上下文 Transformer 模型产生的响应对于上文进行了同义表述然后进行了反问，而 AT-Transformer 模型在承接上文语义之后，又进行了话题的延展，增强了内容丰富性，同时情感与上文较为一致.

5.3.3　对话音频频率对注意力机制的影响

音频特征在一定程度上能够体现说话者所强调的语义重点及情绪特征，对于生成语义契合、情感匹配的回复具有一定的意义. 为了探究音频频率在对

话生成中的作用，选用 session1 中的第 5 个会话中的音频片段，其对应表述为“Okay. But I didn’t tell you to get in this line if you are filling out this particular form.”，其中图 3—图 5 分别为该语句所对应的灰度对数功率谱图、Mel 频谱图和 MFCC 图. 综合 3 个图可以看出，在 0.5 ~ 4 s 之间的频率较高，与此同时该音频对应的音频–文本和文本–音频注意力强度如图 6 所示.

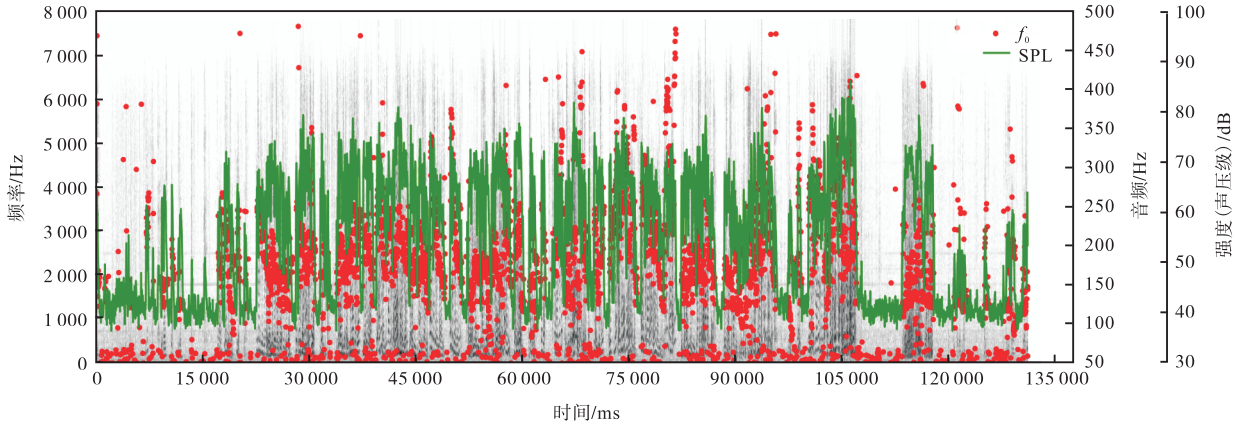


图 3　“Okay. But I didn’t tell you to get in this line if you are filling out this particular form.”音频对应的灰度对数功率谱图

Fig. 3　Gray logarithmic power spectrum corresponding to “Okay. But I didn’t tell you to get in this line if you are filling out this particular form.”

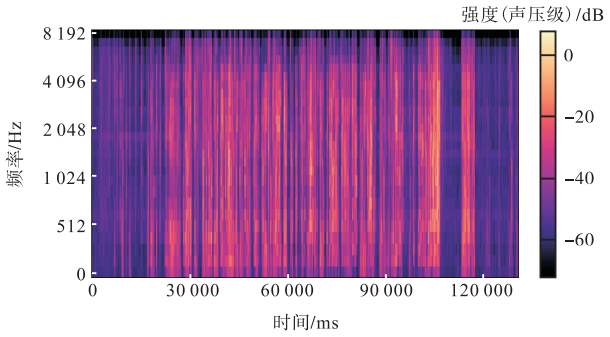


图 4　Mel 频谱图

Fig. 4　Mel spectrogram

图 3 中的红色圆点为基频 f_0，绿色区域为每帧语音在空气中的声压级（SPL 为对数功率谱），将图 3 与图 6 对比可以发现绿色及红色原点部分越密集，图 6 中注意力分数越高，这也就证明了模型的注意力与基频 f_0、声压级具有对应关系. Log-Mel Spectrogram 特征通过构建 Mel 频率的维度和时间帧长度，实现了不同时频下音频信号特征表示（图 4）. 将图 4 和图 6 对比可以发现文本–音频注意力机制与 Mel 时频的变化趋势较为一致.

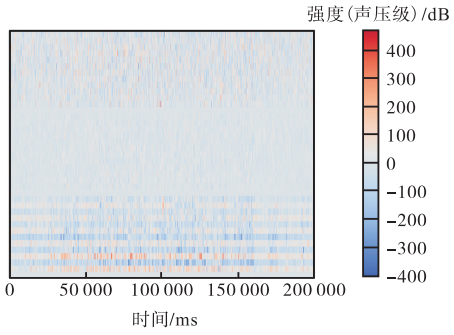
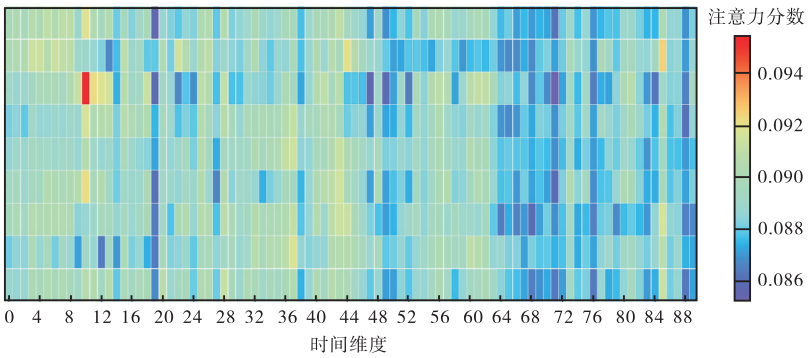


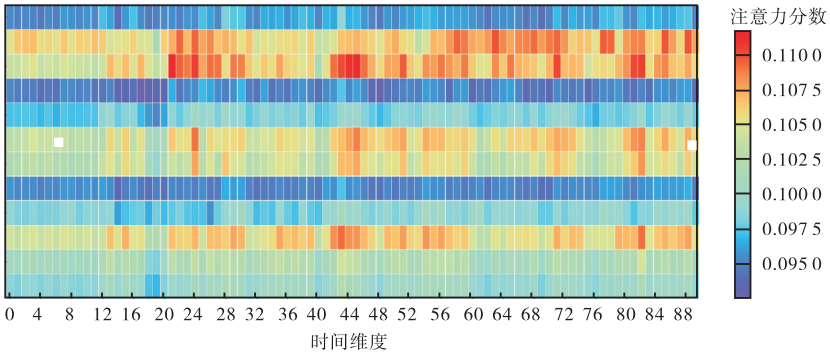
图 5　MFCC 图

Fig. 5　MFCC diagram

由图 5 可知：MFCC 特征能够充分体现说话者的声音特点，考虑到人耳对不同频率的感受程度，常用于语音辨识.



（a）音频–文本注意力



（b）文本–音频注意力

图 6　基于 AT-Attention 模型的音频–文本注意力和文本–音频注意力示意图，该图所对应的语句为“Okay. But I didn’t tell you to get in this line if you are filling out this particular form.”

Fig. 6　Audio-context attention and context-audio attention schematic diagram based on AT-Attention model , the corresponding sentence of which is “Okay. But I didn’t tell you to get in this line if you are filling out this particular form.”

MFCC 特征包括音高、过零率、共振峰等，能够在一定程度上体现说话者的情感特点，比如开怀大笑时声音会高一些，而心情不好则声音低迷. 通过对比图 5 与图 6，可以发现文本–音频注意力机制能够捕捉 MFCC 所体现的这些特征.

图 6 中的文本–音频注意力热力图直观地显示出不同时间的注意力强度变化，对比图 3—图 6 可发现注意力分数与音频的振幅、基频、共振峰相关特征、MFCC 系数均有关系，并且随着时间变化，注意力分数与 Mel 频谱图中的频率和对数功率谱呈明显的正向关系，与音频强度和 MFCC 具有一定的正向对应关系.

6　结　论

本文提出了基于 AT-Transformer 的语音文本多模态对话生成模型，该模型通过 VGGish 实现对对话上下文的音频特征进行提取，并通过 WordEmbedding 计算文本嵌入向量，通过将其与位置编码进行加和融入位置信息，并将二者作为模型的输入. 在编码阶段，通过多头注意力机制对文本和语音模态内、模态间关系计算，实现模态之间关系的深入挖掘，实验表明文本–音频注意力分数更能反映音频上下文的重要性程度. 语音模态对于感知对话上下文的语义重要性有着不可或缺的作用，并且从多样性方面提升对话生成质量，与纯文本对话生成任务相比较，生成语句的流畅度基本持平，情感匹配度和语义相关性均有一定的提升. 此外，由于现实生活中的音频数据具有时长差异性较大、不均衡的特点，如何对信息量较小的音频段进行过滤，实现高效的音频特征处理是下一步值得研究的问题.

参考文献：

[1] CHEN H，LIU X，YIN D，et al. A survey on dialogue systems：recent advances and new frontiers[J]. ACM SIGKDD Explorations newsletter，2017，19(2)：25–35.

[2] WANG S，MENG Y，SUN X，et al. Modeling text-visual mutual dependency for multi-modal dialog generation [EB/OL]. [2022–07–01]. https://arxiv.org/abs/2105.14445.

[3] SHAO L，GOUWS S，BRITZ D，et al. Generating high-quality and informative conversation responses with sequence-to-sequence models[C]//ACL. 2017 Conference on Empirical Methods in Natural Language Processing. New York：Association for Computational Linguistics，2017：2210–2219.

[4] SERBAN I，SORDONI A，BENGIO Y，et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C]//AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. California：The AAAI Press，2016：3776–3783.

[5] CHEN F，MENG F，CHEN X，et al. Multimodal incremental transformer with visual grounding for visual dialogue generation[C]//ACL. Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021. New York：Association for Computational Linguistics，2021：436–446.

[6] QIU M，LI F L，WANG S，et al. Alime chat：a sequence to sequence and re-rank based chatbot engine[C]//ACL.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. New York : Association for Computational Linguistics, 2017: 498–503.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000–6010.

[8] FIRDAUS M, CHAUHAN H, EKBAL A, et al. EmoSen: generating sentiment and emotion-controlled responses in a multimodal dialogue system[C]//IEEE. IEEE Transactions on Affective Computing. New York: IEEE, 2020: 9165162.

[9] YOUNG T, PANDELEA V, PORIA S, et al. Dialogue systems with audio context[J]. Neurocomputing, 2020, 388: 102–109.

[10] AKBARI H, YUAN L, QIAN R, et al. VATT: transformers for multimodal self-supervised learning from raw video, audio and text[C]//NIPS. 35th Conference on Neural Information Processing Systems. Montreal: NIPS, 2021: 11178.

[11] LI Z, LI Z, ZHANG J, et al. Bridging text and video: a universal multimodal transformer for audio-visual scene-aware dialog[J]. IEEE/ACM Transactions on audio, speech, and language processing, 2021, 29: 2476–2483.

[12] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language resources and evaluation, 2008, 42(4): 335–359.

[13] XU W R, GU X S, CHEN G. Generating emotional controllable response based on multi-task and dual attention framework[J]. IEEE Access, 2019, 7: 93734–93741.

[14] 陈晨, 朱晴晴, 严睿, 等. 基于深度学习的开放领域对话系统研究综述[J]. 计算机学报, 2019, 42(7): 1439–1466.

[15] ROHANIAN M, HOUGH J, PURVER M. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech[EB/OL]. [2022–07–01]. https://doi.org/10.21437/Interspeech.2020-2721.

[16] SHEN G, LAI R, CHEN R, et al. WISE: Word-level interaction-based multimodal fusion for speech emotion recognition[C]//ISCA. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 369–373.

[17] SAHA A, KHAPRA M, SANKARANARAYANAN K. Towards building large scale multimodal domain-aware conversation systems[EB/OL]. [2022–07–01]. https://doi.org/10.48550/arXiv.1704.00200.

[18] 张会云, 黄鹤鸣, 李伟, 等. 语音情感识别研究综述[J]. 计算机仿真, 2021, 38(8): 7–17.

[19] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//NIH. Proceedings of the Conference. Association for Computational Linguistics Meeting. Bethesda: NIH Public Access, 2019: 6558–6569.

[20] SIRIWARDHANA S, KALUARACHCHI T, BILLINGHURST M, et al. Multimodal emotion recognition with transformer-based self supervised feature fusion [J]. IEEE Access, 2020, 8: 176274–176285.

[21] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Journal of machine learning research, 2000, 3: 1137–1155.

[22] LIU C W, LOWE R, SERBAN I V, et al. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation[EB/OL]. [2022–07–01]. https://arxiv.org/pdf/1603.08023.pdf.

责任编辑：郎婧