基于预训练字素模型的蒙古语 G2P 研究

顺　毅，萨仁高娃，董伟杰

（内蒙古师范大学 计算机科学技术学院，内蒙古 呼和浩特 010022）

摘　要：针对蒙古文低资源语言环境下 G2P 模型泛化能力受限的问题，将 Transformer 架构与基于字素预训练的模型引入蒙古文字素音素转换领域。通过构建蒙古文字素-音素对齐语料库，探索编码器层数、前馈网络维度对两种模型性能的影响机制。实验结果表明，在蒙古文 G2P 领域，Transformer 模型将 WER 值从传统 n-gram 基线模型的 16.3% 降至 13.64%；GBERT 注意力模型又进一步将 WER 值降至 12.84%。研究意义在于：(1)首次将 Transformer 与预训练注意力机制的模型引入蒙古文 G2P 任务；(2)构建蒙古文字素-音素对齐语料库，为低资源蒙古语研究提供数据支撑；(3)量化模型超参数与正则化策略的性能影响规律，建立可复现的实验基准。研究成果为蒙古文及同类形态复杂语言的 G2P 任务提供了理论方法与工程实践的双重参考。

关键词：字素音素转换；注意力机制；语音识别；语音合成

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-8735(2026)02-0205-09

DOI:10.3969/j.issn.1001-8735.2026.02.011

字素音素转换(grapheme-to-phoneme conversion，G2P)在语音技术领域发挥着重要作用。其本质是构建从字素(如字母、汉字等)[1]构成的序列 $G=\{g_1,g_2,\cdots,g_n\}$ 到对应音素[2]序列 $P=\{p_1,p_2,\cdots,p_n\}$ 的确定映射函数 $f:G\rightarrow P$，从而为声学模型提供底层表征。G2P 作为语音合成系统(text-to-speech，TTS)中前端文本处理模块的关键部分，需完成分词边界检测、多音字消歧等复合功能。以中文合成系统为例，该模块将输入文本如“银行行长在银行行动”精确转换为音素序列/yin1 hang2 hang2 zhang3 zai4 yin2 hang2 xing2 dong4/，其中多音字“行”的消歧准确度直接影响合成语音的语义正确性。其次在自动语音识别(automatic speech recognition，ASR)领域[3]，G2P 可将 ASR 的输出字素序列转换为音素序列，协助判断其是否符合预期发音(如多音字、专有名词)，也可对未登录词通过 G2P 推测发音，辅助生成更合理文本。

G2P 的研究经历了三次范式变革：最初基于人工语言学规则的确定性建模[4-5]，随后发展为基于统计概率的隐马尔可夫模型[6]与决策树方法[7]，目前是以深度学习架构为核心的技术体系，如 LSTM[8]，CNN[9]，Transformer[10]。近年来，国际研究聚焦于探索预训练模型[11]和大语言模型在低资源语言 G2P 任务中的应用[12-13]。

蒙古语 G2P 研究早期聚焦于基于规则的转换方法和基于联合序列模型的转换方法[14]。随后，Liu 等[15]引入了结合注意力机制的 LSTM 模型，并进一步提出融合规则与 LSTM 的策略，有效提升了转换性能。然而，近年来该领域的研究较缺乏，鉴于 Transformer 架构在自然语言处理领域取得的突破性进展，以及其应用于英语 G2P 任务取得优于 CNN 和 RNN 的效果[10]，本研究将表现优异的

收稿日期：2025-10-10

基金项目：内蒙古自治区自然科学基金资助项目“融入发音特征的神经网络蒙古语字素音素转换研究”(2023LHMS06002)；内蒙古自治区高等学校科学技术研究资助项目“面向语音识别的蒙古语发音词典建设中的关键技术研究”(NJZY22568)；内蒙古师范大学基本科研业务费资助项目“融入读音特征的神经网络蒙古语字素音素转换研究”(2022JBYJ033)。

作者简介：顺　毅(1998－)，男，在读硕士研究生。

通信作者：萨仁高娃(1978－)，女，副教授，博士，主要从事自然语言处理、机器学习研究，E-mail:ciecsrgw@imnu.edu.cn。

Transformer 模型引入蒙古文 G2P 任务,作为基线模型,旨在系统探究该架构对蒙古语复杂音系结构的适应性及其性能表现,同时探究预训练 GBERT(grapheme BERT,GBERT)模型在蒙古语 G2P 任务的性能表现。

1 方法

1.1 Transformer G2P

Transformer 采用典型的编码器解码器架构,如图 1 所示[16],其核心模块多头注意力机制(图 2)与 LSTM 中的注意力机制相比,具有如下三点优势:(1)通过并行化子空间建模捕捉多维上下文依赖,加速了训练进程;(2)允许序列中任意两个位置直接建立联系,可有效捕捉长距离音素依赖关系;(3)通过将输入投影到不同的子空间,每个头可以学习并关注输入序列的不同方面,使得模型对输入的理解更加全面和深入,明显提升表征能力。鉴于上述优势思考,Transformer 架构适合处理蒙古语复杂的形态音系结构。

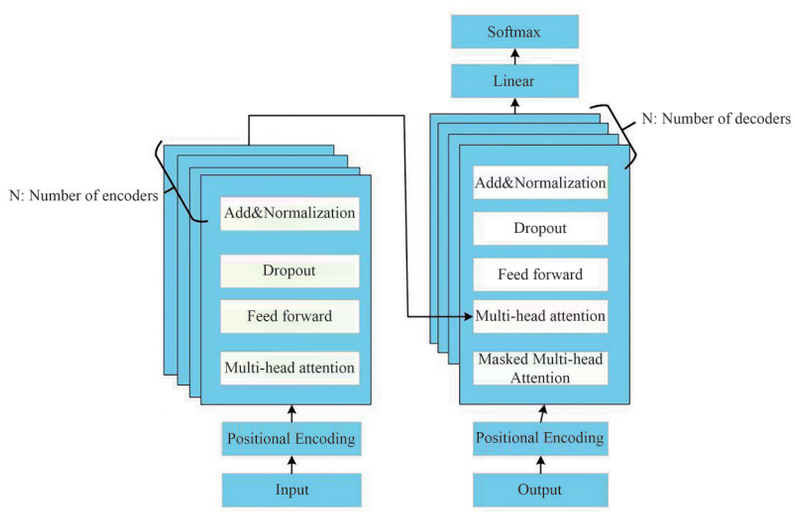

图 1 Transformer 的结构图

Fig. 1 The framework of the Transformer

1.2 GBERT 集成 Transformer 的 G2P

受 BERT 启发的预训练字素模型(GBERT)在 SIGMORPHON 2021 任务[17]中对最具挑战性的 4 种语言(荷兰语、塞尔维亚-克罗地亚语、保加利亚语和韩语)进行了实验,结果表明,基于 GBERT 的 G2P 模型性能均优于 Transformer 基线模型。相较于 Transformer 模型,该方法凭借自监督学习机制,能够从未标注的蒙古文文本中自动学习字素间的上下文依赖关系,无需依赖大规模标注数据即可捕捉复杂的音变规则,这一特性契合了蒙古文 G2P 研究中面临的数据资源匮乏难题。

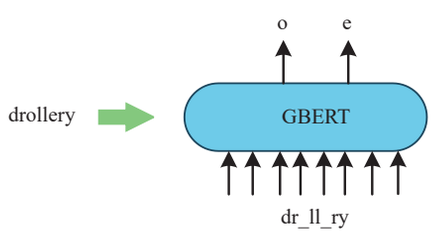

图 2 掩蔽字素预训练 GBERT 的案例

Fig. 2 The case of masked grapheme pre-training GBERT

1.2.1 GBERT 预训练 GBERT 这一范式继承于 BERT[18],其输入单元由传统 BERT 的词元(wordpiece)序列调整为单词内部的字符序列。GBERT 使用 Transformer 编码器通过双向掩码实现上下文感知,如图 2 所示。由于在下游 G2P 任务的输入始终为完整的字符序列,为避免预训练与微调阶段的输入分布差异,采用了动态掩码策略:对输入字符实施 80% 的掩码符“_”替换、10% 的随机字符替换以及 10% 的原字符保留。

1.2.2 Fine-tuning 为实现 GBERT 与现有 Transformer-G2P 框架的融合,有两种集成范式。第一种

方法采用编码器微调策略，具体流程包含三个关键阶段。(1)架构扩展：在预训练的 GBERT 编码器基础上，新增随机初始化的 Transformer 解码器模块，构建完整的序列到序列架构。(2)联合训练：通过端到端优化同步更新编码器与解码器参数，其中编码器采用较低学习率(通常设置为解码器学习率的 10%～20%)以保留预训练知识。(3)表征传递：编码器输出的上下文向量经跨模态注意力机制与解码器交互，逐步生成目标音素序列。该方法的优势在于通过差异化的学习率调控，既能充分利用预训练模型的语义表征能力，又能有效适应 G2P 任务的音系转换特性。

1.2.3　Fusing GBERT　第二种集成方法由 Zhu 等[19]提出，如图 3 所示。在 Transformer 的编码器-解码器层级结构中分别嵌入可调节的 GBERT 注意力模块(GBERT-Enc/GBERT-Dec)。包含三个关键设计：(1)跨模态注意力交互，每个编码层通过多头注意力机制建立原始输入表征与 GBERT 编码输出的动态关联；(2)门控融合机制，在解码层设置可训练的门控权重，自适应调节 GBERT 解码表征与任务特定表征的信息融合比例；(3)残差连接优化，引入层级残差连接确保梯度有效传播，同时缓解深层网络退化问题。这种分层增强机制不仅实现了预训练知识的渐进式融合，还通过参数隔离策略(仅训练新增模块)明显提升模型的训练效率，在保证性能提升的同时避免了传统微调方法可能引发的灾难性遗忘现象。

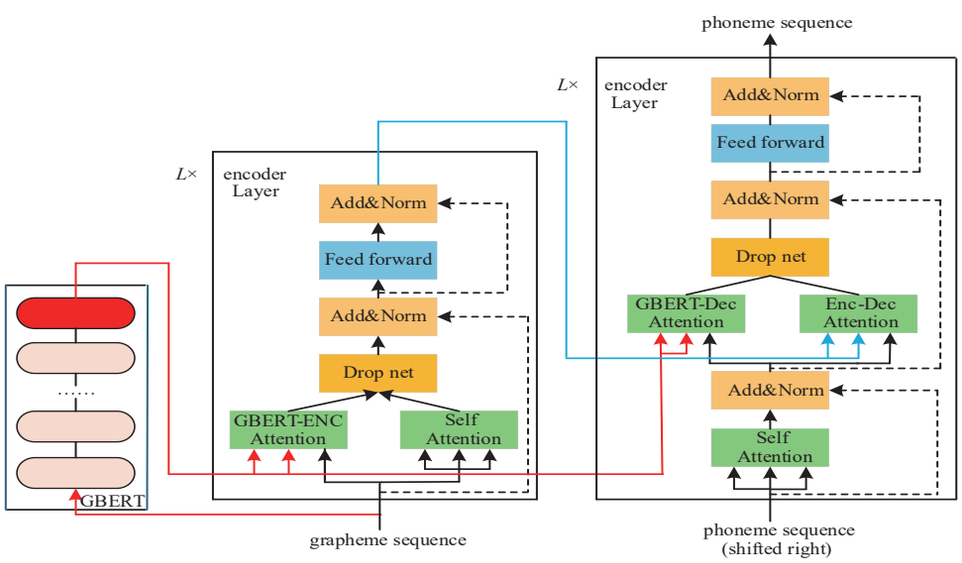

图 3　GBERT 融合模型的结构

Fig. 3　The architecture of GBERT-fused model

输入序列 x 首先由 GBERT 编码成 $H_G=\mathrm{GBERT}(x)$，H_G 是 GBERT 最后一层的输出。对于 Transformer 编码器，第 l 层的隐藏表示为 H^l，输入序列中第 i 个单元在第 l 层的隐藏表示为 h_i^l，在第 l 层($l\in[L]$)，通过式(1)方法将 GBERT 通过注意力机制集成到编码器。

$$\tilde{h}_i^l=\frac{1}{2}\left(\mathrm{Attn}_{\mathrm{B}}\left(h_i^{l-1},H_G,H_G\right)+\mathrm{Attn}_{\mathrm{S}}\left(h_i^{l-1},H^{l-1},H^{l-1}\right)\right),\tag{1}$$

其中 $\mathrm{Attn}_{\mathrm{S}}$ 代表编码器自注意力模块(图 3 编码器右下绿色部分)，$\mathrm{Attn}_{\mathrm{A}}$ 代表 GBERT-Enc 注意力模块(图 3 编码器左下绿色部分)。$\tilde{h}_i^l$ 再经过前馈网络 FNN(•)得到第 l 层的输出 H_E^l。在这一过程中注意力机制让编码器的每一层都能自适应地从 GBERT 输出中获取信息，将其与自身的表示融合。

对于 Transformer 的解码器，第 l 层在时间步 t 之前的隐藏状态为：$S_{<t}^l=\left(s_1^l,s_2^l,\cdots,s_{t-1}^l\right)$，在第 l 层，先计算 $\hat{s}_t^l=\mathrm{Attn}_{\mathrm{S}}\left(s_t^{l-},S_{<t+1}^{l-1},S_{<t+1}^{l-1}\right)$，通过式(2)得到融合表示。

$$\tilde{s}_t^l=\frac{1}{2}\left(\mathrm{Attn}_{\mathrm{B}}\left(\hat{s}_t^l,H_G,H_G\right)+\mathrm{Attn}_{\mathrm{E}}\left(\hat{s}_t^l,H^l,H^l\right)\right),\tag{2}$$

其中 $\mathrm{Attn}_{\mathrm{B}}$ 是 GBERT-Dec 模块(图 3 解码器左中绿色)，$\mathrm{Attn}_{\mathrm{E}}$是Enc－Dec Attention 模块(图 3 解码器

右中绿色)。$\tilde{s}_t^l$ 经过 FFN$(\tilde{s}_t^l)$ 得到 s_t^l,最终通过线性变换和 softmax 得到预测音素。这使得 Transformer 解码器在生成音素序列时可以结合 GBERT 的信息以及编码器的输出,更好地进行音素预测。

编码器与解码器都采用 drop net 策略防止网络过度依赖某一种注意力模型。具体策略为在训练过程中,针对模型的每一层,以一个预先设定的 drop net 率 $P_{\text{net}}\in[0,1]$ 为基础进行操作。在每次训练迭代时,从[0,1]中均匀采样一个随机变量 U^t。以编码器的计算为例,所有的 $\tilde{h}_i^l$ 会依据 U^t 的取值来计算。当 $U^t<\frac{P_{\text{net}}}{2}$ 时,只使用自注意力 $A\text{ttn}_{\text{S}}$;当 $U^t>1-\frac{P_{\text{net}}}{2}$ 时,只使用自注意力 $A\text{ttn}_{\text{B}}$;当 $\frac{P_{\text{net}}}{2}\leqslant U^t\leqslant 1-\frac{P_{\text{net}}}{2}$ 时,同时使用 $A\text{ttn}_{\text{S}}$ 和 $A\text{ttn}_{\text{B}}$。

2 实验设计

2.1 数据集

本研究基于蒙古语发音词典构建音素数据集,数据项由三部分构成:第一列为采用 Unicode 编码的传统蒙古文字母(如ᠠᠶᠢᠮᠠᠭ),第二列为蒙古文转写为对应的拉丁字母(如 aijam),第三列为国际音标标注的拉丁文发音的语音细节(如[æːdʒɪm])。依据蒙古发音词典建设[20]得到 25 079 对蒙古文和其拉丁转写以及对应的音素序列。为解决字素与音素之间复杂的映射关系,为蒙古文发音提供一种机器可读的文字,参考 CMUdict① 设计了符号映射体系,见表 1。采用周期抽样策略进行数据集划分,21 318 例(85%)用作训练集,1 253 例(5%)用作验证集,2 508 例(10%)用作测试集。

表 1 蒙古文音素符号映射表

Tab. 1 Mongolian phoneme-symbol mapping table

类型	音素	符号	类型	音素	符号
短元音	[ɑ][æ][ə][i][ɪ][ɔ][œ][ʊ][ʏ][o][u][y][e][ɭ]	AA、AE、AH、IY、IH、AO、OE、UH、UW、OW、UW、Y、EY、Z	复合元音	[ɑi][ɔi][ʊi][ʊæ][ʊɑi][oi][ui][iɑ][iə][iɔ][iʊ][io][iu][ʊɑ][ie][ou][əi][iɛ][uə][ɑʊ][əʳ]	AI、OI、UI、UAE、UAI、OY、UWIY、IA、IE、IO、IU、IW、IYUW、UA、IQ、OU、AY、IYEH、UWAH、AU、ELR
长元音	[ɑː][əː][iː][ɪː][eː][æː][œː][øː][yː][ɭː][ɻː]	AL、EL、IL、IX、OL、UL、OWL、UU、EH、AX、ER、OA、YL、SL、ZL	依附元音	ɑˇ、ɔˇ、əˇ、ŏ	AS、OS、ES、OWS
基本辅音	[n][ŋ][b][p][x][g][m][l][s][ʃ][t][d][tʃ][dʒ][j][r][w]	N、NG、B、P、KH、G、M、L、S、SH、T、D、CH、JH、J、R、W	借词辅音	[f][k][ts][dz][ɬ][dz̡][tʂ][ʂ][z̡]	F、K、TS、DZ、LH、DZH、TSH、SB、ZH
			鄂化辅音	[nj][lj][xj][gj]	NY、LY、XY、GS

2.2 训练过程

用于训练和推理的硬件是一台搭载了 Intel(R) Xeon(R) Silver 4314 CPU@2.40 GHz 处理器的服务器,服务器都配备了 4 块 NVIDIA GeForce RTX 3090 GPU 显卡(每块显卡配备 24 GB 显存)。

首先系统评估 Transformer 在蒙古文 G2P 任务上的性能,对关键架构配置开展消融实验。基线模型采用 3 层编码器-解码器架构,模型维度 $d_{\text{model}}=256$,前馈网络隐藏层维度 $d_{\text{ff}}=256$,多头注意力头数 $h=4$,正则化策略 dropout=0.2[21],参数调整范围见表 2。在双 NVIDIA RTX 3090 GPU(24 GB GDDR6X)上实施分布式训练,使用 Adam 优化器更新模型权重,学习率初值设置为 0.001,当训练达

① Carnegie Mellon University. Carnegie Mellon University Pronouncing Dictionary. Pittsburgh: Carnegie Mellon University, 2011. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

到 400 周期时终止训练。

对于基于 GBERT 的 G2P 模型训练过程，首先采用 20% 掩码比例使用只含字素的训练集在相同服务器上预训练 400 周期。为严格对齐实验条件，沿用 Transformer 模型的同一套数据集划分比例及相同超参数配置，在双 RTX 3090 服务器上对 GBERT 分别进行微调与混合训练，均完整迭代 400 epoch，以比较二者在蒙古文 G2P 任务上的鲁棒性差异。然后根据较优的方法进行超参数优化，参数调整范围见表 3。为提高评估质量，两种实验方法均采用 beam search[22]搜索算法。

表 2　Transformer 模型训练参数

Tab. 2　Training parameters of Transformer

参数	数值
编码器层数	3/4/5
解码器层数	3/4/5
前馈网络隐藏层维度	256/512/1 024
Dropout	0.2
初始学习率	0.001
注意力头数	4

表 3　基于 GBERT 模型的训练参数调整范围

Tab. 3　Training parameter adjustment range based on GBERT model

参数	数值
GBEERT 掩码率	15%/20%/30%
编码器解码器层数	3/4/5
前馈网络隐藏层维度	256/512/1 024
Dropout	0.1/0.2/0.3
学习率	0.000 5/0.001/0.000 1
GBERT 参数冻结	0/1

3　实验结果与分析

3.1　评估指标

本实验使用 G2P 任务中常用的词错误率(word error rate，WER)和音素错误率(phoneme error rate，PER)评估指标评价模型的性能，词错误率(WER)是指预测的音素序列与参考发音不完全匹配单词的百分比[10]，表示为

$$R_{\mathrm{WE}}=\frac{N_{\mathrm{errors}}}{N_{\mathrm{words}}}\times 100\%, \tag{3}$$

其中，N_{errors}是预测的音素序列与参考发音序列之间不完全匹配的单词数量，N_{words}是参考发音中的单词总数。WER 表示预测的音素序列与参考发音序列在单词层面上的不匹配情况，以百分比形式表示错误率。WER 越低，表明系统在词汇层面上的准确性越高。

音素错误率(PER)作为 G2P 系统性能评估的量化指标，理论源自语音识别领域的序列对齐算法。其数学定义是预测音素序列与参考音素序列的最小编辑距离(minimum edit distance，MED)，也称 Levenshtein 距离与参考序列长度的归一化比值[10]，表示为

$$R_{\mathrm{PE}}=\frac{D(P_{\mathrm{pred}},P_{\mathrm{ref}})}{N_{\mathrm{ref}}}, \tag{4}$$

其中，P_{pred}表示模型的输出音素序列，P_{ref}为专家标注的参考音素序列，$D(P_{\mathrm{pred}},P_{\mathrm{ref}})$为基于 Levenshtein 距离的编辑操作函数，N_{ref}表示参考音素序列的长度。该函数通过动态规划算法计算实现序列对齐所需的最小编辑操作次数，其中归一化处理使得 PER 能够适用不同长度的音素序列评估。

3.2　Transformer 实验结果及与基线模型对比分析

基于 Transformer 模型的实验结果见表 4。整体上，3 层模型在 d_{ff}=1 024 时达到最优词错误率(WER 为 13.64%)和音素错误率(PER 为 3.47%)，表明适度的层数结合更大的前馈网络维度能有效捕捉蒙古文音素转换规律。5 层模型在 d_{ff}=1 024 时 PER 进一步降低至 3.44%，但 WER 略升至 13.68%。前馈网络维度的提升(如从 256 增至 1 024)对 3 层和 5 层模型效果明显，但在 4 层模型中反而导致性能下降，表明层数与维度的平衡至关重要。实验还发现，模型容量扩展(如 5 层＋1 024 维度)虽能提升音素精度，但计算成本较高，而 3 层＋1 024 维度在 WER 和计算效率间更具实用性。

本实验的基线模型是基于联合序列模型(joint-sequence model)的蒙古文 G2P 转换系统。对于蒙古文 Sequitur G2P(基于联合序列模型 G2P),其在不同模型阶数下的性能表现见表 5[11]。通过数据对比发现 Transformer 模型全面超越传统联合序列模型方法,尤其在词错误率上实现突破性提升。传统模型因过拟合和长程依赖建模缺陷,难以适应蒙古文的语言特性;而 Transformer 通过全局注意力机制,明显增强了音素到词的映射鲁棒性。若实际场景需部署轻量级模型,3 层 Transformer(d_{ff}=1 024)是更优选择。

表 4 基于 Transformer 的实验结果

Tab. 4 Experimental results based on Transformer

Num layers	d_{ff}	Numheads	WER/%	PER/%
3	256	4	13.88	3.72
3	512	4	14.15	3.66
3	**1 024**	**4**	**13.64**	**3.47**
4	256	4	14.00	4.05
4	**512**	**4**	14.35	3.71
4	1 024	4	14.75	3.81
5	256	4	14.55	3.7
5	512	4	14.27	3.65
5	1 024	4	13.68	3.44

注:加粗部分表示模型达到最优性能时的参数组合。下同。

3.3 基于 GBERT 模型实验结果

GBERT 预训练模型结构超参数配置为 6 层 Transformer,隐藏层维度为 256,注意力的头数为 4,前馈层的维度为 1 024,训练配置等效批大小为 1 024(梯度累计),训练 400 周期,学习预热+衰减。正则化 Dropout=0.1,标签平滑(系数 0.1),无权重衰减,不同的掩码比例训练结果见表 6。无论是训练集还是验证集,准确率均随掩蔽率提升而上升,表明适当增加掩蔽比例可增强模型对蒙古文字素的上下文推理能力,但同时由于蒙古语低资源数据,20% 的掩蔽率下验证集准确率仅 62.12%。

表 5 基于 Sequitur G2P 模型的结果

Tab. 5 Results based on Sequitur G2P model

Model	The test set		The train set	
	WER/%	PER/%	WER/%	PER/%
Model order 6	16.4	3.4	4.1	0.8
Model order 7	16.4	3.4	3.4	0.7
Model order 8	16.3	3.6	3.3	0.7
Model order 9	16.3	3.2	2.9	0.5
Model order 10	16.3	3.2	2.9	0.5
Transformer(3,1 024,4)	13.6	3.5	3.4	1.2

第一种方法微调 GBERT 的编码器默认超参数配置与上面相同,而解码器的超参数配置为 3 层解码器,多头注意力头数为 4,前馈层维度为 1 024,Dropout=0.2。训练超参数梯度累计前批次大小为 256,梯度累计后批次大小为 1 024,编码器的学习率为 0.000 03。第二种方法通过注意力机制将 GBERT 集成到基于 Transformer 的 G2P 模型。冻结 GBERT 参数,GBERT 注意力层的 Dropout 比率为 0.5。编码器与解码器层数为 3,隐藏层维度都为 256,注意力头数为 4,前馈层维度为 1 024,Dropout=0.2。训练参数初始学习率为 0.001,训练总轮次为 400 次,实际训练的批次大小为 128,使用 Relu 代替 GELU 作为激活函数。Transformer 基线模型配置与第二种方法的 Transformer 参数配置相同,训练周期为 400,实验结果见表 7。

表 6 GBERT 掩蔽字素预训练准确率

Tab. 6 GBERT masked grapheme pre-training accuracy %

Mask ratio	train mask acc	valid mask acc
15	64.02	60.19
20	64.86	62.12

表 7 基于 GBERT 的 G2P 模型的实验结果

Tab. 7 Experimental results of G2P model based on GBERT

Model	WER/%	PER/%
Transformer	13.64	3.47
GBERT fine-tuning	14.41	3.60
GBERT attention	13.56	3.50

基于相同架构的对比实验显示,GBERT 微调模型在词级和音素级性能均出现下降,

GBERT 注意力模型通过直接复用预训练阶段的注意力权重，在保留预训练语言理解能力的同时，WER 较基准模型（Transformer）降低 0.08%，PER 升高了 0.03%，表明预训练注意力机制对蒙古文音素转换的长程依赖建模具有正向作用。这一结果揭示在蒙古语 G2P 任务下，直接迁移预训练注意力模式比全参数微调更有效。针对 GBERT 注意力模型进行参数微调，实验结果见表 8。

表 8　基于 GBERT 注意力模型的实验结果

Tab. 8　Experimental results of GBERT attention

Num layers	d_{ff}	GBERT attention		Transformer	
		WER/%	PER/%	WER/%	PER/%
3	256	13.76	3.54	13.88	3.72
3	**512**	**13.52**	**3.59**	**14.15**	**3.66**
3	1 024	13.56	3.5	13.64	3.47
4	256	13.60	3.45	14.00	4.05
4	512	13.76	3.55	14.35	3.71
4	1 024	13.84	3.55	14.75	3.81
5	256	13.52	3.54	14.55	3.7
5	512	13.72	3.49	14.27	3.65
5	1 024	13.72	3.45	13.68	3.44

从表 8 可知，基于 GBERT 注意力机制的模型在蒙古文 G2P 任务中整体优于标准 Transformer 架构，尤其在深层网络中展现出更强的鲁棒性。当模型层数从 3 层增至 5 层时，GBERT 注意力模型的词错误率（WER）和音素错误率（PER）波动明显小于 Transformer。此外，GBERT 注意力模型在多数配置下呈现更稳定的维度扩展性——3 层 512 维时 WER（13.52%）明显优于 Transformer（14.15%），而 5 层 1 024 维时 PER（3.45%）与 Transformer（3.44%）基本持平，表明其通过迁移预训练的语言理解能力，在中等参数规模下即可实现性能饱和。在探究 dropout 参数对 GBERT attention 鲁棒性影响时，对最优配置（3,512,4）进行消融实验，结果见表 9。实验结果表明 dropout 率对蒙古文 G2P 模型性能具有明显影响，且低 dropout 率（0.1）效果最优。当 dropout 率从 0.1 增至 0.3 时，词错误率从 12.84% 逐步上升至 14.39%，音素错误率从 3.22% 增至 3.81%，反映出过高的正则化会损害模型的学习能力。

表 9　dropout 对模型性能影响的实验结果

Tab. 9　Experimental results on the impact of dropout on model performance

dropout	WER/%	PER/%
0.1	**12.84**	**3.22**
0.2	13.52	3.59
0.3	14.39	3.81

以上实验结果表明：通过预训练注入语言先验知识，可提升低资源语言 G2P 任务的模型鲁棒性；且模型性能并非随网络深度或维度单调提升，需根据任务特性寻找最优参数配置点。对低资源蒙古文 G2P 任务，优先探索轻量化架构＋预训练知识迁移的组合，而非盲目地增加模型复杂度。值得注意的是，正则化强度的设置需遵循动态适配原则，即与模型容量、数据分布特性形成精准匹配。实验数据证实，过度正则化（如 dropout＞0.2）会抑制预训练知识的表达迁移，因此在微调阶段应采用弱正则化策略（dropout≤0.1）。

4　结论

本研究首次将 Transformer 和基于预训练字素的模型架构引入蒙古文 G2P 转换任务，通过构建的蒙古文字素音素对齐语料库，探索了模型编码器和解码器层数、前馈网络中间层维度等超参数配置对蒙古文 G2P 性能的影响。实验结果表明，基于 Transformer 的模型较联合序列模型提升 16.3%，其自注意力机制有效捕捉了蒙古文黏着语的长程依赖特性。其次，引入预训练的 GBERT 注意力机制可进一步提升深层模型鲁棒性，将 WER 降至 12.84%，且正则化策略需谨慎设计，低 dropout 率（0.1）明显优化性能。研究表明，在蒙古文 G2P 任务中，轻量化模型（3 层＋1 024d_{ff}）结合预训练注意力和适度正

则化(dropout=0.1)是实用的最优解。本文在以下方面做出了贡献:验证了 Transformer 模型与基于预训练字素的模型对蒙古文 G2P 任务的敏感性规律;构建了蒙古文字素-音素对齐语料库;为后续蒙古文 G2P 研究提供了可复现的基线参照与参数调优依据。

参考文献:

[1] MELETIS D. The grapheme as a universal basic unit of writing[J]. Writing Systems Research, 2019, 11(1): 26-49.

[2] 邢福义,吴振国. 语言学概论[M]. 武汉:华中师范大学出版社,2002:77-87.

[3] 金秀丽. 端到端语音识别算法研究与实现[D]. 兰州:兰州交通大学,2023.

[4] ELOVITZ H, JOHNSON R, MCHUGH A, et al. Letter-to-sound rules for automatic translation of English text to phonetics[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24(6): 446-459.

[5] DAMPER R I, MARCHAND Y, ADAMSON M J, et al. Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches[J]. Computer Speech & Language, 1999, 13(2): 155-176.

[6] CHENG S Y, ZHU P C, LIU J T, et al. A survey of grapheme-to-phoneme conversion methods[J]. Applied Sciences, 2024, 14(24): 11790: 1-20.

[7] HÄKKINEN J, SUONTAUSTA J, RIIS S, et al. Assessing text-to-phoneme mapping strategies in speaker independent isolated word recognition[J]. Speech Communication, 2003, 41(2-3): 455-467.

[8] RAO K, PENG F C, SAK H, et al. Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks [C]//2015 IEEE International Conference on Acoustics, Speech and Signal Proces sing (ICASSP). South Brisbane: IEEE, 2015: 4225-4229.

[9] YOLCHUYEVA S, NÉMETH G, GYIRES-TÓTH B. Grapheme-to-phoneme conversion with convolutional neural networks[J]. Applied Sciences, 2019, 9(6): 1143-1162.

[10] YOLCHUYEVA S, NÉMETH G, GYIRES-TÓTH B. Transformer based grapheme-to-phoneme conversion [C]// Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019: 2095-2099.

[11] DONG L, GUO Z Q, TAN C H, et al. Neural grapheme-to-phoneme conversion with pre-trained grapheme models [C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6202-6206.

[12] NOVAK J R, MINEMATSU N, HIROSE K. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework[J]. Natural Language Engineering, 2016, 22(6): 907-938.

[13] QHARABAGH M F, DEHGHANIAN Z, RABIEE H R. LLM-powered grapheme-to-phoneme conversion: Benchmark and case study [C]//2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025: 1-5.

[14] 飞龙,高光来,闫学亮. 蒙古文字母到音素转换方法的研究[J]. 计算机应用研究, 2013, 30(6): 1696-1700.

[15] LIU Z N, BAO F L, GAO G L, et al. Mongolian grapheme to phoneme conversion by using hybrid approach [M]// ZHANG M, NG V, ZHAO D, et al. Natural Language Processing and Chinese Computing. Cham: Springer, 2018: 40-50.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 30 (NIPS 2017). Red Hook: Curran Associates Inc., 2017: 5998-6008.

[17] ASHBY L F E, BARTLEY T M, CLEMATIDE S, et al. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion [C]// Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Stroudsburg: Association for Computational Linguistics, 2021: 115-125.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language

understanding [C]// BURSTEIN J, DORAN C, SOLORIO T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[19] ZHU J, XIA Y, WU L, et al. Incorporating BERT into neural machine translation [C]// 2020 International Conference on Learning Representations (ICLR). Addis Ababa: Open Review Net, 2020: 1-17.

[20] 萨仁高娃 . 蒙古语发音词典建设及其语音识别的应用研究[D]. 呼和浩特：内蒙古大学, 2021: 45-78.

[21] SRIVASTAVA N, HINTON G, KRIZHEVSKY A. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(56): 1929-1958.

[22] HUANG L, ZHAO K, MA M. When to finish? Optimal beam search for neural text generation (modulo beam size) [D]. Ithaca: Cornell University, 2018.

G2P Research on Mongolian Based on Pre-trained Grapheme-to-Phoneme (G2P) Models

Shunyi, Sarengaowa, DONG Weijie

(*College of Computer Science and Technology*, *Inner Mongolia Normal University*, *Hohhot* 010022, *China*)

Abstract: To address the issue of limited generalization capability of G2P models in the low-resource language environment of Mongolian, this study introduced the Transformer architecture along with pre-trained grapheme-based models into the field of Mongolian grapheme-to-phoneme (G2P) conversion. By constructing a Mongolian grapheme-phoneme alignment corpus, the study explored the impact mechanisms of encoder layer numbers and feed-forward network dimensions on the performance of two types of models. Experimental results show that in the field of Mongolian G2P, the Transformer model reduces the Word Error Rate (WER) from 16.3% in the traditional n-gram baseline model to 13.64%. The GBERT attention model further lowers the WER to 12.84%. The significance of this study is as follows: (1) It is the first time that Transformer and pre-trained attention mechanism models have been applied to the Mongolian G2P task. (2) A Mongolian grapheme-phoneme alignment corpus is constructed, providing data support for low-resource Mongolian language research. (3) The impact patterns of model hyperparameters and regularization strategies on performance are quantified, establishing reproducible experimental benchmarks. The research findings offer dual references for theoretical methods and engineering practices in G2P tasks for Mongolian and other morphologically complex languages.

Key words: grapheme-to-phoneme conversion; attention mechanism; speech recognition; speech synthesis

【责任编辑　张颖娟】