# 面向 Transformer 模型的蒙古语语音识别词特征编码方法

**张晓旭** [1]**, 马志强** [1,2*]**, 刘志强** [1]**, 宝财吉拉呼** [1]

[1] 内蒙古工业大学, 呼和浩特, 010000

[2] 内蒙古自治区基于大数据的软件服务工程技术研究中心, 呼和浩特, 010000

mzq_bim@imut.edu.cn

## 摘要

针对 Transformer 模型在蒙古语语音识别任务中无法学习到带有控制符的蒙古语词和语音之间的对应关系，造成模型对蒙古语的不适应问题。提出一种面向 Transformer 模型的蒙古语词编码方法，方法使用蒙古语字母特征与词特征进行混合编码，通过结合蒙古语字母信息使 Transformer 模型能够区分带有控制符的蒙古语词，学习到蒙古语词与语音之间的对应关系。在 IMUT-MC 数据集上，构建 Transformer 模型并进行了词特征编码方法的消融实验和对比实验。消融实验结果表明，词特征编码方法在 HWER、WER、SER 上分别降低了 23.4%、6.9%、2.6%；对比实验结果表明，词特征编码方法领先于所有方法，HWER 和 WER 分别达到 11.8%、19.8%。

**关键词：** 蒙古语语音识别；Transformer；注意力机制；词编码

# Researching of the Mongolian word encoding method based on Transformer Mongolian speech recognition

**Zhang Xiaoxu**[1]**, Ma Zhiqiang**[1,2*]**, Liu Zhiqiang**[1]**, Bao Caijilahu**[1]

[1] Inner Mongolia University of Technology, Huhhot, 010000

[2] Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Huhhot, 010000

mzq_bim@imut.edu.cn

## Abstract

In view of the fact that the Transformer model cannot learn the correspondence between Mongolian words with control symbols and the speech in the Mongolian speech recognition task, which causes the model to not adapt to the Mongolian language. A Mongolian word encoding method for Transformer model is proposed. The method uses Mongolian letter features and word features for mixed encoding. By combining Mongolian letter information, the Transformer model can distinguish Mongolian words



**基金项目**：国家自然科学基金 (61762070,62166029)，内蒙古自然科学基金 (2019MS06004)

with control symbols, and learn Mongolian words and pronunciation. Correspondence between. On the IMUT-MC dataset, the Transformer model is con-structed and the ablation and comparison experiments of word feature encoding methods are carried out. The results of ablation exper-iments show that the word feature encoding method reduces HWER, WER, and SER by 23.4%, 6.9%, and 2.6%, respectively; the comparative experimental results show that the word feature encoding method is ahead of all methods, with HWER and WER reaching 11.8%, 19.8%.

**Keywords:** Mongolian speech recognition , Transformer , Attention , Word encoding

## 1 引言

蒙古语语音识别（Mongolian Speech Recognition, MSR）作为少数民族语言信息处理技术之一，是将蒙古语语音序列转换成蒙古语文本序列的过程。蒙古语作为一种黏着语，是一种表音文字语言 (清格尔泰, 1991)。蒙古语主要通过增加格后缀丰富词的表示，实现语法功能 (高莲花, 2021)。蒙古语格后缀一般指虚词，不能单独出现在句子中充当主、谓、宾、定和状语等成分，但具有强大的组合能力。例如，当格后缀与名词组合可以让名词有不同的意思和读音，类似于汉语的“的”、“和”、“以”等介词。蒙古语格后缀的复杂性体现在格后缀出现在词根后的不同位置会有不同的书写方式和读音 (莫日根, 2016)。在计算机表示中，蒙古语词采用字母拼接表示，通过空格进行词与词的划分，而对于词加格后缀是通过控制符来进行连接并控制其外形。该控制符与蒙古语空格的书面表示形式相似，而在计算机中采用不同的编码。例如在一个名词 ᠭᠡᠷ（没有）后接不同的格后缀构成相同语义的词，目的是能适应不同句子的语法要求，具体见表1所示。

表 1. ᠭᠡᠷ 接不同格后缀的词

| 格后缀的词 | 词根 | 格后缀 |
|---|---|---|
| ᠭᠡᠷ ᠦᠨ | ᠭᠡᠷ | ᠦᠨ |
| ᠭᠡᠷ ᠢᠶᠡᠨ | ᠭᠡᠷ | ᠢᠶᠡᠨ |
| ᠭᠡᠷ ᠲᠡᠭᠡ | ᠭᠡᠷ | ᠲᠡᠭᠡ |
| ᠭᠡᠷ ᠳᠡ | ᠭᠡᠷ | ᠳᠡ |
| ᠭᠡᠷ ᠲᠡ | ᠭᠡᠷ | ᠲᠡ |

对于蒙古语语音识别模型建模研究主要经历了三个阶段。首先，蒙古语语音识别任务的研究是将隐马尔可夫-高斯混合模型（Gaussian of Mixture Hidden Markov Model，GMM-HMM）作为建立蒙古语语音识别系统的研究起点。(Guanglai Gao, 2006) 等人首次将 GMM-HMM 技术引入蒙古语语音识别任务，并构建基于 GMM-HMM 的蒙古语语音识别系统。(Qilao Hasi, 2008) 等人通过对声学模型的优化提高蒙古语语音识别模型的性能。(Feilong Bao, 2013) 等人通过使用 GMM-HMM 对声学模型进一步设计，从而提高蒙古语语音识别模型的性能。其次，随着深度神经网络（Deep Neural Network，DNN）的出现，深度神经网络结合隐马尔科夫模型（Deep Neural Network of Mixture Hidden Markov Model，DNN-HMM）的组合方式也在蒙古语语音识别模型中开展研究和使用。(Hui Zhang, 2015) 等人在基于传统的混合蒙古语语音识别模型的研究中，引入基于 DNN 的声学模型，使蒙古语语音识别系统获得了显著的性能提升。(马志强,

2018) 等人针对 HMM 在蒙古语语音识别声学模型中不能充分学习声学特征之间相关性和独立性假设的问题，使用深度神经网络对声学模型建模进行研究，并且取得了良好的识别性能，用其构建的蒙古语语音识别系统在工业界得到了应用，以上基于 HMM 蒙古语语音识别模型属于传统的混合语音识别模型。但是传统的混合语音识别模型的独立训练与联合识别的特性导致模型参数无法达到全局最优，同时构建的复杂性也给蒙古语语音识别工业化应用带来了困难。最后，为了降低蒙古语语音识别模型构建和训练的复杂度，研究人员用端到端神经网络模型的学习过程代替工程过程，剔除了发音词典部分，降低了蒙古语语音识别模型工业化应用的门槛，使端到端蒙古语语音识别模型成为工业化应用的研究热点。但是在进行端到端蒙古语语音识别模型建模时，模型不能正确识别一些带有控制符的蒙古语词，导致蒙古语语音的识别结果与蒙古语词目标文本不一致的情况，造成模型对蒙古语的不适应性问题。因此如何使模型学习到蒙古语语音和蒙古语词的对应关系是一个重要的问题。

本文在蒙古语构词特点的基础上，基于 Transformer 模型提出一种蒙古语字母与词混合编码方法。主要贡献为：

- 基于注意力机制提出了一种面向蒙古语语音识别的蒙古语词编码方法；
- 把蒙古语字母特征与词特征进行混合编码，构建了一个基于 Transformer 的端到端蒙古语语音识别模型；
- 在 IMUT-MC 数据集上进行了验证实验，实验结果显示提出的蒙古语字母与词混合编码方法在性能上相较基线方法有显著提升。

## 2 相关工作

端到端语音识别模型一般选用适合语言特点的建模单元进行编码。比如英语选用子词作为建模单元进行编码，而汉语选用以字为建模单元进行编码。对于端到端蒙古语语音识别模型来说，蒙古语的构词是以词根、词干上连接不同的词缀构成，因此针对于蒙古语形态复杂多变的情况，模型一般选用蒙古语词作为建模单元进行编码。

基于蒙古语词的切分构成建模单元进行编码研究，(赵伟, 2010) 等人为了利用蒙古语语法信息，通过分析蒙古语词的构形特点，划分出词干和词的构形附加成分，提出一种有效的蒙古语词标注方法。实验表明该方法的词切分准确率比现有蒙古语词切分系统的准确率有明显的提高。(杨振新, 2017) 等人针对蒙汉统计机器翻译面临的形态差异大的问题，通过词素和短语两个层面编码信息的结合，实现了汉语与蒙古语语言在形态结构上的对称，通过实验结果表明该方法在 BLEU 上，比基线模型有了明显的提高，在一定程度上消解了形态差异对汉蒙统计机器翻译的影响。(Qing-Dao-Er-Ji Ren, 2020) 等人对蒙古语双语料库预处理阶段蒙古语的词干和词缀进行分割和标记。通过实验结果表明，使用该方法的端到端蒙汉神经机器翻译模型与传统统计方法和基于递归神经网络的机器翻译模型相比，在翻译质量和语义混淆方面有所提高。

基于蒙古语词特征编码研究。(包乌格德勒, 2018) 等人为探究词特征输入对蒙汉翻译系统的影响，分别采用蒙古语的词模型、切分模型和子词模型作为输入词特征。通过对比实验结果表明，子词模型在基于 CNN 和 RNN 的神经蒙汉翻译模型中可以有效地提高翻译质量。(曹宜超, 2020) 等人针对蒙古语形态复杂多变的问题，提出了一种结合词向量编码对齐的蒙汉神经机器翻译方法。通过实验结果表明，该方法的翻译效果高于基线模型。(卞乐乐, 2021) 等人为探究词向

量的质量是否影响模型的质量问题，使用三种不同的词向量生成模型对蒙古语单语数据进行词向量的生成，对不同模型生成的蒙古语单语词向量对翻译模型质量的影响进行实验。通过实验表明，将大量蒙古语单语词向量嵌入蒙汉翻译模型能够使模型效果得到显著提升。

在蒙古语语音识别任务中，由于蒙古语在构词中添加了控制符来改变词的外形，使得模型无法区分相同读音的蒙古语词。因此，本文基于蒙汉机器翻译中蒙古语构词特点，提出了蒙古语字母与词混合编码单元，使 Transformer 语音识别模型适应蒙古语的读音和构词特点，解决了模型对蒙古语的不适应问题，从而提升模型的识别准确率。

## 3 蒙古语词编码方法

### 3.1 问题描述

在蒙古语词表 $C = \{y_1, \ldots, y_i, \ldots, y_j, \ldots, y_m\}$ 中，$\exists y_i, y_j \in C$ 且 $y_i != y_j$，使 $y_i$ 与 $y_j$ 的读音相同，即 $g(y_i) = g(y_j)$ ，其中 $g(\cdot)$ 表示读音函数。在基于 Transformer 模型进行蒙古语语音识别建模时，首先输入一段蒙古语语音序列 $X = \{x_1, \ldots, x_t, \ldots, x_T\}$，通过编码器 $Encoder(\cdot)$ 得到语音特征序列 $\mathbf{h}^{enc}$。然后输入目标序列开始标签 $Y = \{sos\}$，进行词嵌入 $embedding(\cdot)$ 得到解码器输入序列 $\mathbf{h}_{Y_{sos}}$，通过解码器 $Decoder(\cdot)$ 预测 $y_l|(1 \le l \le L)$。最后目标序列 $Y = \{y_1, \ldots, y_L\}$ 经过 $L$ 步解码得到。在解码器 $Decoder(\cdot)$ 预测 $y_l|(1 \le l \le L)$ 的过程中会出现不能正确区分词表中带有控制符的蒙古语词的情况，导致 Transformer 模型对蒙古语的不适应问题。

### 3.2 Transformer 模型架构

基于 Transformer 的蒙古语语音识别模型采用编码-解码器模型结构，比其他基于注意力机制的端到端语音识别模型结构更加复杂，其编码器和解码器都是由堆叠式自注意力层和全连接层构成，模型结构如图1所示。

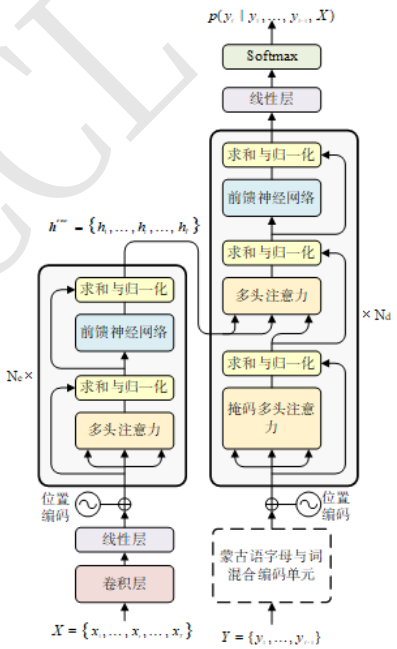



图 1. 基于 Transformer 的端到端蒙古语语音识别模型架构

基于 Transformer 的蒙古语语音识别模型中的编码器将输入的语音序列 $X$ =

$\{x_1,\ldots,x_t,\ldots,x_T\}$ 映射成为特征序列 $\mathbf{h}^{enc}=\{h_1,\ldots,h_t,\ldots,h_T\}$，然后解码器将编码器生成的特征序列 $\mathbf{h}^{enc}$ 映射成为文本序列 $Y=\{y_1,\ldots,y_l,\ldots,y_L\}$。在基于 Transformer 的端到端语音识别模型上，本文提出端到端蒙古语语音识别模型架构，其总体计算公式如下：

$$P_{att}(Y|X)=\prod_{l=1}^{L}P_{att}(y_l|y_1,\ldots,y_{l-1}) \tag{1}$$

在蒙古语语音识别任务中，首先蒙古语字母在词中位置发生一系列的变化，导致模型在进行识别时发生错误。为解决该问题，本模型考虑将蒙古语词特征序列描述为具有词特征和字母特征的混合特征序列。在蒙古语字母与词混合编码单元结构中利用注意力机制将根据字母特征生成新词特征，通过结合原有的词特征来增加蒙古语词特征的区分度，使模型的识别准确率得到提升。

### 3.3 蒙古语字母与词混合编码方法

与基于注意力机制的序列到序列模型一样，基于 Transformer 的端到端语音识别模型也是编码-解码器架构，Transformer 模型结构中的编码器和解码器是由堆叠式自注意力层和全连接层构建而成。本节主要详细介绍蒙古语字母与词混合编码单元的详细结构，主要是为基于 Transformer 的端到端语音识别模型适应蒙古语，从而提高模型的识别准确率。

基于 Transformer 的端到端语音识别模型使用的是词嵌入的方法，对于蒙古语中书写方式相似的词来说，词嵌入编码特征具有相似性，如  和 。因此，本章提出一种结合字母特征的蒙古语词编码单元，如图2所示。

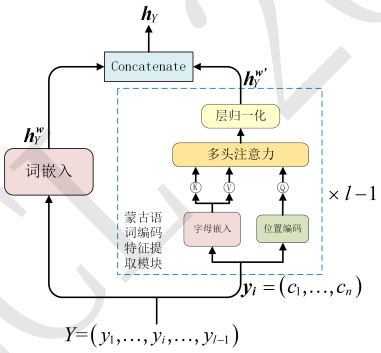



图 2. 蒙古语字母与词混合编码单元

首先，输入文本标签序列 $Y=\{y_1,\ldots,y_i,\ldots,y_{l-1}\}$ 在词级序列 $Y$ 和字母级序列 $y_i$ 上执行嵌入编码，获得词嵌入编码特征 $\mathbf{h}_Y^w$ 和字母嵌入编码特征 $\mathbf{h}_{y_i'}^{w'}$ 。然后，将获得 $\mathbf{h}_{y_i'}^{w'}$ 和字母位置编码特征 $\mathbf{pe}_{y_i'}^{w'}$ 通过注意力机制生成新的编码特征 $\mathbf{h}_Y^{w'}$。最后，将 $\mathbf{h}_Y^w$ 和 $\mathbf{h}_Y^{w'}$ 进行拼接，得到最终解码器输入特征 $\mathbf{h}_Y$。加入字母信息后，明显区分蒙古语词编码特征。注意力机制不仅用于提取单词级上下文信息，还用于对单词中的字母级信息进行编码。对蒙古语字母和词混合编码单元的功能进行分析，其计算公式如下：

$$\mathbf{h}_Y^w=embedding(y_1,\ldots,y_{l-1}) \tag{2}$$

$$\mathbf{h}_{y_i'}^{w'}=embedding(c_1,\ldots,c_n) \tag{3}$$

$$\mathbf{pe}_{y_i'}^{w'}=PE_{(pos,j)} \tag{4}$$

$$\mathbf{h}_{y_i}^{w'} = attention(\mathbf{h}_{y_i'}^{w'}, \mathbf{h}_{y_i'}^{w'}, \mathbf{pe}_{y_i'}^{w'}) \tag{5}$$

$$\mathbf{h}_Y = \mathbf{h}_Y^w \oplus \mathbf{h}_{y_i}^{w'} \tag{6}$$

$$p(y_l|y_1, \ldots, y_{l-1}) = Decoder(\mathbf{h}^{enc}, \mathbf{h}_Y) \tag{7}$$

## 4 实验

### 4.1 实验设置

#### 4.1.1 实验数据

本实验使用的语料库为 IMUT-MC，由 (刘志强, 2021) 等人构建的一个针对蒙古语语音识别任务的语音语料库。其中，具体选用语料库中 IMUT-MC2、IMUT-MC3 和 IMUT-MC4 数据集进行实验。IMUT-MC2 和 IMUT-MC3 的数据来自于蒙古语的日常对话，文本语句都比较简短。IMUT-MC4 的数据来源于蒙古语新闻，包括环境、教育、经济、时政和体育等领域，文本语句较长。IMUT-MC 的基本信息如表2所示。

表 2. 语料库 IMUT-MC 基本信息

| 数据集 | 时长 | 总句数 | 总词数 | 平均词数 | 说话人个数 | 来源 |
|---|---|---|---|---|---|---|
| IMUT-MC2 | 23.5h | 19800 | 970 | 10 | 99 | |
| IMUT-MC3 | 40.8h | 22200 | 1307 | 10 | 111 | 人员录音 |
| IMUT-MC4 | 69.7h | 20000 | 6591 | 22 | 100 | |
| 总计 | 134h | 62000 | 8868 | 13 | 310 | |

将 IMUT-MC 数据集，按照当前大部分研究神经网络训练的 8:1:1 比例进行划分训练集、验证集和测试集。训练集用来使模型进行学习；验证集用来确定神经网络结构或控制模型复杂程度的参数；测试集用来检验最终选择最优的模型的性能。

#### 4.1.2 实验环境

本实验平台使用两个高性能计算机，包括个人工作站和深度学习服务器。在硬件方面，个人工作站的环境配置包括 Intel i7-9700 CPU、NVIDIA RTX-2060 GPU；深度学习服务器的环境配置包括 Intel Xeon 6130 CPU、NVIDIA Tesla P100 GPU。在软件方面，使用 Ubuntu 操作系统，安装 Python 环境和 PyTorch 深度学习框架，搭建 Kaldi 和 Espnet 语音识别平台，支持 GPU 并行计算。所有模型的实验工作均在上述计算设备上开展。

#### 4.1.3 模型参数

基于 Transformer 的端到端蒙古语语音识别模型采用的超参数设计如下：编码器层数为 12，编码器单元为 2048；解码器层数为 6，解码器单元为 2048；注意力机制的维度为 256，注意力头数为 4；丢弃率为 0.1，使用 Adam 优化算法，学习率为 1.0；批处理大小为 16，迭代训练 20 轮得到最终的模型；模型的特征处理部分采用 1 层卷积层，其中卷积核大小为 3，步移为 2，下采样率为 4；字母与词混合编码单元中的多头注意力机制的维度为 256，注意力头数为 4。

### 4.2 评价指标

蒙古语语音识别模型的评价指标包括：HWER（Homophone Word Error Rate，同音词错误率）、WER（Word Error Rate，词错误率）、SER（Sentence Error Rate，句错率）。各评价指标的含义如下：

（1）HWER 是指所有错误同音词的和所占总同音词数的百分比，其公式为：

$$HWER = \frac{S_h + D_h + I_h}{N_h} * 100\% \tag{8}$$

$S_h$ 为替换错误的同音词数，$D_h$ 为删除错误的同音词数，$I_h$ 为插入错误的同音词数，$N_h$ 表示数据集中同音词总词数。

（2）WER 是指所有错误词的和所占总词数的百分比，其公式为：

$$WER = \frac{S_w + D_w + I_w}{N_w} * 100\% \tag{9}$$

$S_w$ 为替换错误的词数，$D_w$ 为删除错误的词数，$I_w$ 为插入错误的词数，为数据集中的总词数，其中，为数据集中非同音词的总词数。

(3) SER 是指所有识别结果与对应文本不能正确匹配的测试音频所占总音频数的百分比，其公式为：

$$SER = \frac{N_error}{N_s} * 100\% \tag{10}$$

$N_error$ 为识别错误的蒙古语音频的个数，$N_s$ 为数据集中蒙古语音频的总个数。

### 4.3 实验结果与分析

#### 4.3.1 语种对比实验

基于 Transformer 语音识别模型在英语、汉语和蒙古语上的识别结果，如表3所示。

表 3. 不同语种在 Transformer 语音识别模型上的识别结果

| 语言 | 数据集 | 总时间 | 总词数 | 平均时间 | 平均词数 | WER |
|---|---|---|---|---|---|---|
| 汉语 | Aishell-1 | 178h | 4229 | 4.45s | 15 | 9.7 |
| 英语 | LibriSpeech(train-clean-100) | 100h | 6967 | 12.8s | 20 | 9.8 |
| 蒙古语 | IMUT-MC | 134h | 8868 | 7.8s | 13 | 26.7 |

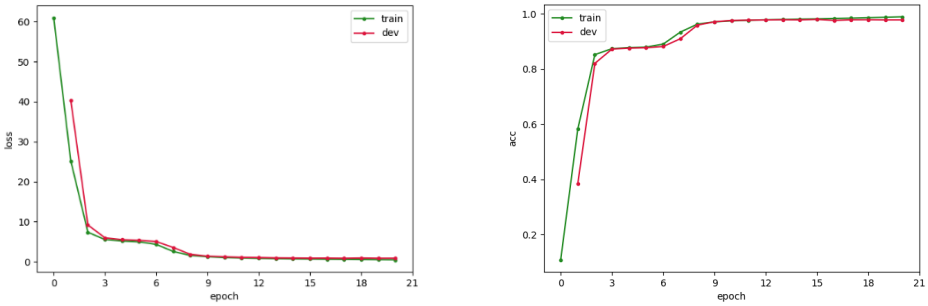
从实验结果可以看出，基于注意力机制的 Transformer 语音识别模型在蒙古语上的识别效果远不如汉语和英语的识别效果。对模型转录的蒙古语语音数据结果进行分析，大部分识别错误的词是组合词，即具有词根和词缀的蒙古语词。因此对蒙古语构词特点进行分析，蒙古语的格后缀与名词结合时，会组成新蒙古语词。在组成时，随着位置不同书写形式也不相同，会形成一些异形同音词等组合词，模型对这些组合词的词特征区分度不高。

#### 4.3.2 收敛性实验

训练过程在 IMUT-MC 数据集上对基于 Transformer 的端到端蒙古语语音识别模型开展，收敛情况如图3所示。为了保证基于 Transformer 的端到端蒙古语语音识别模型的收敛效果，实验采用训练集和验证集的损失值和准确率来验证模型收敛。

由图3 a）中可知，基于 Transformer 的端到端蒙古语语音识别模型训练集和验证集上的损失函数呈下降趋势，并于 9 轮左右使 Loss 趋向于零且处于平缓状态，表明模型能够收敛。由图3 b）可知，在训练集和验证集上的准确率不断上升，呈现出两次先陡后缓的趋势，最终在 9 轮左右趋于平稳缓和且验证集的准确率高于 90%，可以认为模型已经充分收敛，学习到了训练集的数据分布特征。

第二十一届中国计算语言学大会论文集，第333页-第343页，南昌，中国，2022年10月14日至16日。




(a) 训练的 loss 值的变化　　(b) 训练过程中识别准确率

图 3. 基于注意力的蒙古语语音识别模型的收敛情况

### 4.3.3 蒙古语字母与词混合编码单元消融实验

为了说明在蒙古语语音识别任务中，基于 Transformer 语音识别模型在增加蒙古语字母与词混合编码单元的有效性，对蒙古语字母与词混合编码单元和蒙古语词编码特征提取模块中的位置编码进行消融实验。

表 4. 编码方法消融实验结果

| 编码方法 | 位置编码 | HWER | WER | SER |
|---|---|---|---|---|
| + 词嵌入 | - | 35.2 | 26.7 | 36.4 |
| + 字母特征提取 | - | 35.6 | 28.8 | 37.4 |
| | 绝对位置编码 | 23.4 | 24.9 | 35.7 |
| | 相对位置编码 | 24.5 | 25.5 | 36.1 |
| + 词嵌入 + 字母特征提取 | - | 13.4 | 23.8 | 34.3 |
| | 绝对位置编码 | 11.8 | 19.8 | 33.8 |
| | 相对位置编码 | 11.5 | 20.7 | 33.1 |

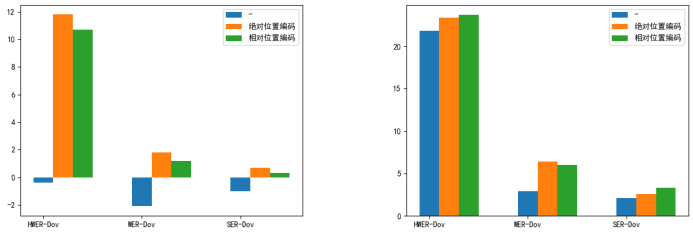
根据表4的实验结果数据分析可知，本章在基于 Transformer 的端到端语音识别模型上进行蒙古语字母与词混合编码单元的消融实验，验证了单元可以使模型获得更好的识别准确率，表明蒙古语字母与词混合编码单元能够使基于 Transformer 语音识别模型适应蒙古语的特点。并且通过探索不同位置编码对蒙古语词编码特征提取模块的影响，证明了绝对位置编码能使基于 Transformer 语音识别模型的识别准确率达到最佳。

为了更加直观地突出蒙古语字母与词混合编码单元对基于 Transformer 语音识别模型的影响，定义评价指标的下降值（Drop-out Value，Dov），见公式 (11)，用于描述基线评价指标 Pa 与增加单元结构的评价指标 Pb 之间的差距。正值表示提高了基线实验的评价指标值，负值表示降低了基线实验的评价指标值。

$$Dov = Pa - Pb \tag{11}$$

将表4中数据使用 Dov 指标对使用的编码特征进行比较，其中 HWER-Dov、WER-Dov 和 SER-Dov 分别表示同音词错误率下降值、词错误率下降值和句错率下降值。

从图4中可以看出，蒙古语编码特征使用字母与词的混合特征，三个评价指标没有负值产生，因此本章提出的蒙古语字母与词混合编码单元是有效地。尽管位置编码对于蒙古语字母与词混




(a) 使用字母特征　　(b) 使用字母与词的混合特征

图 4. 消融实验的 HWER、WER 和 SER 的下降值

合编码单元中注意力机制是有效地，但是绝对位置编码和相对位置编码的方式对当前的数据集的影响并不明显。通过对当前蒙古语数据集进行分析，得出不明显的原因是：数据集中的蒙古语词的长度可以使注意力机制获取的上下文关系，绝对位置编码信息可以满足当前的蒙古语数据集，后期当蒙古语数据集增加时，可以继续探究相对位置编码对该单元的影响。

#### 4.3.4　蒙古语输入文本特征编码对比实验

基于 Transformer 的端到端蒙古语语音识别模型的输入文本特征编码实验旨在探索基于 Transformer 语音识别模型对于蒙古语语音识别任务的适应性，主要是从字母、子词、词、词根与词缀划分和字母与词相结合的特征编码上进行对比实验，结果如表5所示。

表 5. 蒙古语输入特征编码对比实验

| 输入特征编码 | HWER | WER | SER |
|---|---|---|---|
| 字母特征编码 | 40.3 | 33.6 | 41.6 |
| 子词特征编码 | 18.3 | 24.2 | 35.2 |
| 词特征编码 | 35.2 | 26.7 | 36.4 |
| 词根与词缀特征编码 | 16.5 | 22.3 | 34.1 |
| 字母与词特征编码 | 11.8 | 19.8 | 33.8 |

根据表5的实验结果数据分析可知，基于 Transformer 的端到端语音识别模型在蒙古语语音识别任务中，以字母特征编码的识别效果不佳；以子词特征编码可以使基于 Transformer 的端到端语音识别模型具有良好的识别效果，但是子词是基于频率进行统计获得，对语料具有很强的依赖性；以词特征编码的识别效果也有所提升，但是也具有对语料的依赖性；以词根与词缀特征编码作为解码器输入，可以解决模型对蒙古语不适应性的问题，但是模型在识别时，会出现一些不正确的组词的错误；以字母与词特征编码作为解码器输入，可解决模型对蒙古语不适应性的问题，并使模型的识别效果达到最优。既可以缓解建模单元对语料的依赖，又可以使模型充分利用两个量级的知识，更适合当前的蒙古语语料数据。

#### 4.3.5　蒙古语端到端语音识别模型对比实验

为了说明在蒙古语语音识别任务中，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型具有良好的识别性能，因此对多种端到端语音识别模型进行对比实验研究，结果如表6所示。

根据表6的实验结果数据分析可知，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型在蒙古语语音识别任务上比当前其他端到端语音识别模型具有更好的识别准确率，

表 6. 蒙古语端到端语音识别模型对比实验

| 模型 | HWER | WER | SER |
|---|---|---|---|
| CTC | 36.7 | 30.3 | 58.2 |
| Transformer | 35.2 | 26.7 | 36.4 |
| RNN-Transducer | 33.6 | 25.1 | 44.8 |
| 字母与词特征编码 + Transformer | 11.8 | 19.8 | 33.8 |

HWER、WER 和 SER 分别达到了 11.8%、19.8% 和 33.8%。因此，蒙古语端到端语音识别模型对比实验验证了增加蒙古语字母与词混合单元的 Transformer 语音识别模型更加适应蒙古语语音识别任务，具有良好的识别性能。

### 4.4 案例分析

表7展示了蒙古语语音识别的样例，该样例从 IMUT-MC 语料库随机选出。编号 1-4 表示是从 IMUT-MC-TEST1 和 IMUT-MC-TEST2 的测试数据集中选出，主要是针对于日常对话的数据集，其中语音数据的时长集中在 3-5s 之间；编号 5-8 表示是从 IMUT-MC-TEST4 的测试数据集中选出，主要针对环境、教育、经济、时政和体育等领域的数据，其中语音数据的时长集中在 8-10s 之间。数据来源的场景不同，所以蒙古文词的使用也不尽相同。

表 7. 蒙古语测试样例数据识别结果

| 编号 | 语音数据编号 | 标签数据 | 备注 |
|---|---|---|---|
| 1 | 00001365-F-M-19-13.wav | ᠬᠦᠮᠦᠨ ᠬᠦᠷᠢᠶᠡᠨ ᠬᠡᠪ᠎ᠡ ᠪᠠᠢᠭᠤᠯᠭ᠎ᠠ ᠬᠡᠪᠦᠦᠨ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ | 让（他）走进内蒙古地区的旗县 |
| 2 | 00000118-F-W-20-03.wav | ᠴᠠ ᠵᠤᠭᠰᠤᠭᠠᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪ᠎ᠡ ᠬᠡᠯᠡᠵᠤ ᠪᠤᠢ᠎ᠠ ᠤᠤ | 您不稍微往前站吗？我想从这里抓住 |
| 3 | 00000128-D-M-20-09.wav | ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠬᠡᠷᠡᠭ ᠴᠠ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪᠠᠰᠤᠭᠠᠷᠠᠢ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ | 我不太懂，您向别人问吧 |
| 4 | 00001228-F-W-20-03.wav | ᠴᠠ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ | 您想喝点啥，您想吃点啥 |
| 5 | 00001307-F-W-22-39.wav | ᠡᠨᠡᠨ᠎ᠡ ᠬᠠᠭᠠᠨᠠᠴᠠ᠎ᠠ ᠬᠤᠭᠤᠤᠨᠤᠴᠠ᠎ᠠ ᠬᠡᠪᠡᠷ ᠪᠢ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪ ᠬᠡᠯᠡᠵᠤ᠎ᠡ ᠨ ᠠᠯᠳᠠᠭᠳᠠᠵᠤ | 不知哪一天，从他（她，它）那里得到事情的真理 |
| 6 | 00000075-H-W-23-63.wav | ᠪᠢ᠎ᠡ ᠭᠡ ᠬᠥᠬᠡ ᠬᠠᠭᠠ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠤ᠎ᠡ ᠰᠤᠷᠤᠭ᠎ᠠ ᠪᠠᠭᠰᠢᠮᠠᠯ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠨ᠎ᠤ ᠭᠠ᠎ᠡ ᠵᠥᠪᠬᠡᠨ᠎ᠡ ᠪᠡᠷᠬᠦ᠎ᠡ ᠴᠠ᠎ᠠ ᠵᠠᠮᠴᠢᠨ᠎ᠤ ᠬᠡᠷᠡᠭᠲᠡᠢ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪᠡᠷ ᠪᠢ | 我们有蓝，黄，绿，粉和红色的袍，您需要什么颜色的 |
| 7 | 00002424-D-W-20-90.wav | ᠡᠨᠡᠬᠦ ᠪᠢ ᠪᠤᠳᠤᠯᠠᠭᠠᠨ ᠪᠢ ᠴᠢᠵᠢᠶᠠᠨ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠨᠢ ᠪᠠᠭᠠ ᠪᠠᠢᠵᠤ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ | 所以，以说教为主的家庭教育比普通的教育效率低 |
| 8 | 00001373-F-M-19-13.wav | ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢ ᠲᠠᠨ᠎ᠤ ᠬᠡᠪᠡᠯᠡᠯ᠎ᠡ ᠵᠢᠷᠤᠭᠯᠠᠩ᠎ᠤ ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢᠶᠠᠨ ᠲᠤ᠎ᠡ ᠬᠡᠭᠡᠨ᠎ᠡ ᠭᠡ ᠤᠭᠰ ᠭᠡ | 难道这叫做让世间的所有生活的资格吗 |

由表8可以看出，基于 Transformer 的端到端蒙古语语音识别模型能够较好地识别蒙古语语音，且具有一定的准确度，对带有控制符的词识别准确率更高。其中具有下划线的词、“***” 和双下划线的词分别表示替换错误词、删除错误词和插入错误词。

表 8. 蒙古语测试样例数据

| 编号 | 识别标签结果 | 备注 | 替换词错误 | 删除词 | 插入词错误 |
|---|---|---|---|---|---|
| 1 | ᠬᠦᠮᠦᠨ ᠬᠦᠷᠢᠶᠡᠨ ᠬᠡᠪ᠎ᠡ ᠪᠠᠢᠭᠤᠯᠭ᠎ᠠ ᠬᠡᠪᠦᠦᠨ ᠵᠤᠭᠰᠠᠭᠠᠵᠤ | 走进了内蒙古地区的旗县 | 1 | 0 | 0 |
| 2 | ᠴᠠ ᠵᠤᠭᠰᠤᠭᠠᠵᠤ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪ᠎ᠡ ᠬᠡᠯᠡᠵᠤ ᠪᠤᠢ᠎ᠠ ᠤᠤ | 您不稍微往前站吗？我想从这里抓住 | 0 | 0 | 0 |
| 3 | ᠪᠢ ᠮᠡᠳᠡᠬᠦ ᠬᠡᠷᠡᠭ ᠴᠠ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪᠠᠰᠤᠭᠠᠷᠠᠢ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ | 我不太懂，您向别人问吧 | 0 | 0 | 0 |
| 4 | ᠴᠠ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠵᠤᠮᠠ ᠬᠤᠭᠤᠯᠤᠭᠰᠠᠨ ᠪᠠᠢᠨ᠎ᠠ | 您想喝点啥，您想吃点啥呢 | 2 | 0 | 1 |
| 5 | ᠡᠨᠡᠨ᠎ᠡ ᠬᠠᠭᠠᠨᠠᠴᠠ᠎ᠠ ᠬᠤᠭᠤᠤᠨᠤᠴᠠ᠎ᠠ ᠬᠡᠪᠡᠷ ᠪᠢ ᠬᠡᠭᠡᠨ᠎ᠡ ᠪ ᠬᠡᠯᠡᠵᠤ᠎ᠡ ᠨ ᠠᠯᠳᠠᠭᠳᠠᠵᠤ | 不知哪一天，从他（她，它）那里得到事情的真理 | 1 | 0 | 0 |
| 6 | ᠪᠢ᠎ᠡ ᠭᠡ ᠬᠥᠬᠡ ᠬᠠᠭᠠ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠤ᠎ᠡ ᠰᠤᠷᠤᠭ᠎ᠠ ᠪᠠᠭᠰᠢᠮᠠᠯ᠎ᠠ ᠬᠤᠷᠤᠭᠤᠨ᠎ᠤ ᠭᠠ᠎ᠡ ᠵᠥᠪᠬᠡᠨ᠎ᠡ ᠪᠡᠷᠬᠦ᠎ᠡ ᠴᠠ᠎ᠠ ᠵᠠᠮᠴᠢᠨ᠎ᠤ ᠬᠡᠷᠡᠭᠲᠡᠢ ᠪᠠᠢᠨ᠎ᠠ ᠤᠤ ᠬᠡᠪᠡᠷ ᠪᠢ | 我们有蓝，黄，绿，粉和红色的袍，您需要什么颜色的 | 0 | 0 | 0 |
| 7 | ᠡᠨᠡᠬᠦ ᠪᠢ ᠪᠤᠳᠤᠯᠠᠭᠠᠨ ᠪᠢ ᠴᠢᠵᠢᠶᠠᠨ ᠪᠠᠶᠢᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠨᠢ ᠪᠠᠭᠠ ᠪᠠᠢᠵᠤ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠠᠢᠭᠠᠭᠤᠯᠯᠠᠭ᠎ᠠ ᠲᠠᠢ ᠵᠢᠨᠬᠡᠨᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ ᠪᠢᠴᠢᠭᠴᠢ ᠪᠢᠨ᠎ᠡ | 所以，以说教为主的家庭教育比普通的教育效率低 | 0 | 0 | 0 |
| 8 | ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢ ᠲᠠᠨ᠎ᠤ ᠬᠡᠪᠡᠯᠡᠯ᠎ᠡ ᠵᠢᠷᠤᠭᠯᠠᠩ᠎ᠤ ᠲᠤᠭᠤᠭᠤᠯᠤᠭᠴᠢᠶᠠᠨ ᠲᠤ᠎ᠡ ᠬᠡᠭᠡᠨ᠎ᠡ ᠭᠡ ᠤᠭᠰ ᠭᠡ | 难道这叫做让世间的所有生活的资格吗 | 1 | 0 | 0 |

## 5 结论

针对 Transformer 模型对蒙古语构词特点的不适应性问题，将蒙古语字母特征与词特征进行混合编码，提出一种基于 Transformer 的蒙古语词编码方法，并构建端到端蒙古语语音识别模型。通过字母特征进一步构建词特征能提高 Transformer 模型对带有控制符的蒙古语词识别准确率。实验结果表明，在 IMUT-MC 蒙古语语音识别数据集下，增加蒙古语字母与词混合编码单元的 Transformer 语音识别模型的 WER 降低了 6.9%，表明该单元对 Transformer 模型在蒙古语语音识别任务中具有一定的作用。

## 参考文献

清格尔泰. 1991. 蒙古语语法. 内蒙古人民出版社.

高莲花. 2021. 蒙古语的后置词短语, 民族语言, (05):108–113.

莫日根. 2016. 基于规则的传统蒙古语句法分析研究. 内蒙古大学.

Guanglai Gao, Biligetu, Nabuqing and Shuwu Zhang . 2006. *A Mongolian Speech Recognition System based on HMM*. International Conference on Intelligent Computing, Springer-Verlag, 2006: 667-676.

Qilao Hasi, and Guanglai Gao. 2008. *Researching of Speech Recognition Oriented Mongolian Acoustic Model*. Conference on Pattern Recognition, 2008: 1-6.

Feilong Bao, Guanglai Gao, Xueliang Yan and Weihua Wang. 2013. *Segmentation-based Mongolian LVCSR Approach*. International Conference on Acoustics, Speech, and Signal Processing, 2013: 1-5.

Hui Zhang, Feilong Bao, and Guanglai Gao. 2015. *Mongolian Speech Recognition Based on Deep Neural Networks*. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 2015: 180-188.

马志强, 李图雅, 杨双涛和张力. 2018. 基于深度神经网络的蒙古语声学模型建模研究, 智能系统学报, 13(03):486–492.

赵伟, 侯宏旭, 从伟和宋美娜. 2010. 基于条件随机场的蒙古语词切分研究, 中文信息学报, 24(05):31–35+84.

杨振新, 李淼, 陈雷, 卫林钰, 陈晟和孙凯. 2017. 一种基于词素媒介的汉蒙统计机器翻译方法, 中文信息学报, 31(04):57–62+69.

Qing-Dao-Er-Ji Ren, Yila Su, and Nier Wu. 2020. *Research on Mongolian-Chinese machine translation based on the end-to-end neural network*, International Journal of Wavelets, Multiresolution and Information Processing, 18(01):1941003.

包乌格德勒和赵小兵. 2018. 基于 *RNN* 和 *CNN* 的蒙汉神经机器翻译研究, 中文信息学报, 32(08):60–67.

曹宜超, 高翊, 李淼, 冯韬, 王儒敬和付莎. 2020. 基于单语语料和词向量对齐的蒙汉神经机器翻译研究, 中文信息学报, 34(2):27–32.

卞乐乐. 2021. 基于单语语料与强化学习的蒙汉神经机器翻译的研究. 内蒙古工业大学.

刘志强, 马志强, 张晓旭, 宝财吉拉呼, 谢秀兰和朱方圆. 2021. *IMUT-MC*: 一个针对蒙古语语音识别的语音语料库, 中国科学数据.