专题 多语种智能信息处理

# IMUT-MC：一个针对蒙古语语音识别的语音语料库

刘志强[1]，马志强[1,2*]，张晓旭[1]，宝财吉拉呼[1]，谢秀兰[1]，朱方圆[1]

1. 内蒙古工业大学数据科学与应用学院，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

**摘要：**蒙古语作为少数民族语言，其使用人群分布辽阔，收集标注语音数据困难，导致没有公开的大规模蒙古语语音语料库为广大研究人员提供实验支撑，阻碍了蒙古语语音识别的进一步发展。本课题组构建了一个针对蒙古语语音识别任务的语音语料库 IMUT-MC，包含 417 位说话人录制的约 212 小时的阅读语音，致力于推进蒙古语语音识别研究。课题组分别在传统语音识别模型和端到端语音识别模型上使用 IMUT-MC 进行基线语音识别实验，基于 GMM-HMM、DNN-HMM 和 Transformer 的语音识别模型在 IMUT-MC 上词错率分别为 69.90%、67.45%和 26.10%，证明了 IMUT-MC 是进行蒙古语语音识别可靠的语料库。

**关键词：**蒙古语；语音识别；语音语料库；阅读语音

## 数据库（集）基本信息简介

| | |
|---|---|
| **数据库（集）名称** | 蒙古语语音语料库 IMUT-MC |
| **数据作者** | 刘志强、马志强、张晓旭、宝财吉拉呼、谢秀兰、朱方圆 |
| **数据通信作者** | 马志强（mzq_bim@imut.edu.cn） |
| **数据时间范围** | 2017–2021年 |
| **数据类型** | 单通道阅读语音 |
| **数据来源** | 人员录音 |
| **采样率** | 16000 Hz |
| **数据量** | 25.16 GB（压缩后18.75 GB） |
| **数据格式** | *.wav |
| **数据服务系统网址** | http://www.doi.org/10.11922/sciencedb.j00001.00349 |
| **基金项目** | 国家自然科学基金（61762070，62166029）；内蒙古自然科学基金（2019MS06004）。 |
| **数据库（集）组成** | 语料库共包括 9 个数据文件，其中：(1) *.zip是数据集文件压缩包，分别是IMUT-MC1、IMUT-MC2、IMUT-MC2修正、IMUT-MC3、IMUT-MC4和IMUT-MC5数据集的文件压缩包；(2) lexicon.txt 是发音词典文件；(3) mongolian_data_prep.sh 是语音识别实验数据处理脚本；(4) README.txt 是阅读指南文件。 |

## 引 言

语音识别（Automatic speech recognition，ASR），尤其是大词汇量连续语音识别，是机器学习领域的重要课题。长期以来，隐马尔科夫高斯混合模型（GMM-HMM）[1]一直是主流的语音识别模型。随着深度神经网络的发展，深度

ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2021.0096.zh
文献 DOI：
10.11922/11-6035.csd.2021.0096.zh
数据 DOI：
10.11922/sciencedb.j00001.00349
文献分类：信息科学

收稿日期：2021-12-29
开放同评：2022-01-28
录用日期：2022-06-06
发表日期：2022-06-29

* 论文通信作者
马志强：mzq_bim@imut.edu.cn

神经网络隐马尔科夫模型（DNN-HMM）[2]和端到端模型[3-5]已经取得超越 GMM-HMM 的性能。这些语音识别模型通常需要大量高质量的语音数据来达到优异的性能。近年来，英语、汉语等大语种凭借丰富的语料资源在语音识别的不同任务中取得巨大的进步，其中一个重要原因是各种规模的公开数据集为语音识别研究人员提供开放的数据平台，如 Ted-Lium[6]、Librispeech[7]、THCHS-30[8]和 AISHELL[9]等。然而，对于许多小语种语言的语音识别，没有大规模高质量的标注语音数据已经成为阻碍它们进一步发展的关键因素。

蒙古语作为少数民族语言，其使用人群分布辽阔，收集标注语音数据困难，导致没有公开的大规模蒙古语语音语料库为广大研究人员提供实验支撑，蒙古语语音识别研究受到了极大的限制。在此之前，已有很多研究者致力于蒙古语语音识别研究[10-11]。但是，在没有普遍接受的蒙古语语音数据集的情形下，研究者都在其内部数据上进行实验并记录结果。这阻碍了实验复现和性能基准测试，从而限制了蒙古语语音识别进一步发展。为解决这个问题，课题组在国家自然科学基金（62166029，61762070）和内蒙古自然科学基金（2019MS06004）的资助下构建了蒙古语语音语料库 IMUT-MC，其中包含 417 位说话人录制的约 212 小时的阅读语音，所有说话人均为能够熟练使用蒙古语交流的蒙古族学生。

IMUT-MC 语料库主要是为蒙古语语音识别研究构建，共包含 5 期阅读语音数据集。每一期数据集构建目的不同，可用于语音识别任务下各种子任务研究，如语音表示[12]、声学建模[13]和说话人自适应[14]等相关研究。IMUT-MC 也可以用于其他与语音相关的任务，如语音合成[15]和语音翻译[16]。课题组期望 IMUT-MC 成为语音识别研究社区的宝贵资源，并成为蒙古语语音识别研究的基线数据集。IMUT-MC 语料库不仅致力于推进蒙古语语音识别研究，而且希望促进蒙古语在语音应用中的发展和使用，如消息听写、语音搜索、语音命令和其他语音控制智能设备。

## 1 数据采集和处理方法

IMUT-MC 是一个单通道蒙古语语音语料库。语音话语是在密闭录音室内通过高保真的麦克风录制，采样为 16Khz，16-bit WAV 格式。IMUT-MC 由约 8.43 万句语音话语组成约 212 小时，其中包含 5 期语音数据集，分别是 IMUT-MC1、IMUT-MC2、IMUT-MC3、IMUT-MC4 和 IMUT-MC5。IMUT-MC1 由 8 名录音人员按照不同方式朗读 1255 句录音文本录制而成，包括 1 人录制的 1255 句语音话语和 7 人随机录制的 1255 句语音话语，共 2510 句语音话语，总时长为 1.8 小时；IMUT-MC2 由 99 名录音人员朗读相同的 200 句录音文本录制而成，共包含 1.98 万句语音话语，总时长为 23.5 小时；IMUT-MC3 是由 111 名录音人员朗读相同的 200 句录音文本录制而成，共包含 2.22 万句语音话语，总时长为 40.8 小时。不同于前三期数据集，IMUT-MC4 是由 100 名录音人员朗读固定的 200 句录音文本录制而成。其中，100 名录音人员分为 5 组，每 20 人朗读相同的 200 句录音文本进行录制，共包含 2 万句语音话语，总时长为 69.74 小时。IMUT-MC5 与 IMUT-MC4 构建方式相同，共包含 1.98 万句语音话语，总时长为 75.29 小时。录制完成后，语音话语被处理成语音识别实验要求的格式，与转录文本一一对应。IMUT-MC 语料库基本信息如表 1 所示。

**表 1 IMUT-MC 语料库基本信息**

**Table 1 Basic information of IMUT-MC corpus**

| 数据集 | 语音总句数（万） | 语音总时长（小时） | 来源 |
|---|---|---|---|
| IMUT-MC1 | 0.25 | 1.8 | 人员录音 |

| 数据集 | 语音总句数（万） | 语音总时长（小时） | 来源 |
|---|---|---|---|
| IMUT-MC2 | 1.98 | 23.5 | 人员录音 |
| IMUT-MC3 | 2.22 | 40.8 | 人员录音 |
| IMUT-MC4 | 2.0 | 69.74 | 人员录音 |
| IMUT-MC5 | 1.98 | 75.29 | 人员录音 |

## 2 数据样本描述

### 2.1 说话人信息

IMUT-MC 共有 417 名说话人参与录制，说话人的性别、电话、年龄和生活地区被记录为元数据。参与录制的说话人都具备熟练使用蒙古语进行日常交流的能力，大多数来自蒙古族的在校大学生，年龄分布在 18-25 岁，并且性别比例平衡，分别为 48%的男性和 52%的女性。IMUT-MC 语料库说话人具体信息如表 2 所示。

**表 2 IMUT-MC 语料库说话人信息**

**Table 2 Speaker information of IMUT-MC corpus**

| 数据集 | 地区数 | 说话人个数 | 男性人数 | 女性人数 |
|---|---|---|---|---|
| IMUT-MC1 | 4 | 8 | 7 | 1 |
| IMUT-MC2 | 10 | 99 | 48 | 51 |
| IMUT-MC3 | 7 | 111 | 48 | 63 |
| IMUT-MC4 | 10 | 100 | 51 | 49 |
| IMUT-MC5 | 10 | 99 | 49 | 50 |

由于蒙古族在内蒙古自治区分布情况不同，导致课题组采集语料难度有所差异。每一位说话人的生活地区信息被记录，他们大多数来自通辽、赤峰、兴安盟和锡林郭勒盟等地区。不同地域的说话人具有当地口音特色，因此 IMUT-MC 也可用于蒙古语方言语音识别。同时行政分区编码也被加入语料的处理中，用于区分说话人的口音信息。IMUT-MC 语料库说话人口音信息如表 3 所示。

**表 3 IMUT-MC 语料库说话人口音信息**

**Table 3 Voice information of speakers in IMUT-MC corpus**

| 行政分区编码 | 口音区域 | 说话人个数 | 男性人数 | 女性人数 |
|---|---|---|---|---|
| A | 呼和浩特 | 7 | 2 | 5 |
| B | 包头 | 2 | 1 | 1 |
| C | 乌海 | 0 | 0 | 0 |
| D | 赤峰 | 82 | 44 | 38 |
| E | 呼伦贝尔 | 8 | 1 | 7 |
| F | 兴安盟 | 59 | 25 | 34 |
| G | 通辽 | 152 | 87 | 65 |
| H | 锡林郭勒盟 | 72 | 30 | 42 |

| 行政分区编码 | 口音区域 | 说话人个数 | 男性人数 | 女性人数 |
|---|---|---|---|---|
| J | 乌兰察布 | 2 | 2 | 0 |
| K | 鄂尔多斯 | 14 | 5 | 9 |
| L | 巴彦淖尔 | 17 | 6 | 11 |
| M | 阿拉善盟 | 2 | 0 | 2 |

IMUT-MC 语料库存在说话人重复情况。例如，一名说话人可能同时参与了 IMUT-MC1、IMUT-MC2 和 IMUT-MC3 的构建。据统计，参与数据集 IMUT-MC1、IMUT-MC5 构建的都是独立不重复的说话人，而数据集 IMUT-MC2、IMUT-MC3 和 IMUT-MC4 存在说话人重复情况。从图 1 可知，共有 310 名说话人参与 IMUT-MC2、3、4 期的录制，其中有 143 名是独立不重复的说话人。IMUT-MC2、3、4 期数据集说话人重复具体信息如图 1 所示。

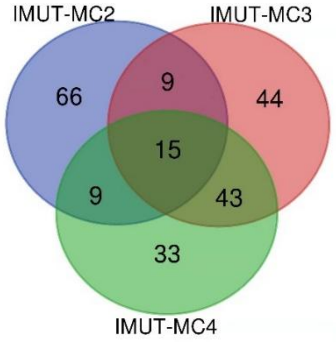


**图 1 IMUT-MC2、3、4 期数据集说话人重合情况**

**Figure 1 Speaker overlap in phase-2, phase-3 and phase-4 datasets of IMUT-MC corpus**

## 2.2 转录和词汇

IMUT-MC 语料库包含 5 期语音数据集，每期数据集构建目的不同，其转录文本来源也不同。其中 IMUT-MC1、IMUT-MC2、IMUT-MC3 针对扩充蒙古语口语知识构建，而 IMUT-MC4、IMUT-MC5 针对扩充更多的文本领域知识构建。IMUT-MC1 从蒙古语教材《蒙古语会话手册》[17]摘选 1255 句文本进行录制； IMUT-MC2 是在 IMUT-MC1 的基础上选用其中 200 句文本进行录制；IMUT-MC3 从中国新闻网（蒙语版）摘选 200 句文本进行录制；IMUT-MC4 从中国新闻网（蒙语版）中的时政、教育、体育、环境和经济等 5 个领域分别选取 200 句文本进行录制；IMUT-MC5 从中国新闻网（蒙语版）中的人文、法律、科学、技术和饮食等 5 个领域分别选取 200 句文本进行录制。同时，本文分别从文本句子数、词数和平均词个数等方面对 IMUT-MC 的各期数据集的转录文本进行对比，转录文本信息具体情况如表 4 所示。

**表 4 IMUT-MC 语料库转录文本信息**

**Table 4 Transcription text information of IMUT-MC corpus**

| 数据集 | 文本句子 | 词数 | | 平均词个数 | 文本来源 |
|---|---|---|---|---|---|
| | | 词数 | 总词数 | | |
| IMUT-MC1 | 1255 | 2237 | 24488 | 6 | 《蒙古语会话手册》 |
| IMUT-MC2 | 200 | 970 | 187700 | 10 | 《蒙古语会话手册》 |

| 数据集 | 文本句子 | 词数 | | 平均词个数 | 文本来源 |
|---|---|---|---|---|---|
| | | 词数 | 总词数 | | |
| IMUT-MC3 | 200 | 1307 | 209770 | 10 | 中国新闻网（蒙语版） |
| IMUT-MC4 | 1000 | 6591 | 427560 | 22 | 中国新闻网（蒙语版） |
| IMUT-MC5 | 1000 | 8673 | 448680 | 29 | 中国新闻网（蒙语版） |

在转录文本的制作过程中，对原始蒙古语文本语料进行如下处理：（1）对转录文本手动过滤，消除敏感政治字眼、用户隐私、色情和暴力等不适当内容；（2）转录文本句子中的一些符号，如“、”、《、》、[、]、=等均被删除；（3）所有转录文本格式采用“UTF-8”编码。转录文本处理完成后，本文对 IMUT-MC 语料库各期数据集转录文本的蒙古语单词重合情况进行了统计。其中，IMUT-MC 语料库各期数据集转录文本蒙古语单词重合情况如图 2 所示。IMUT-MC4 和 IMUT-MC5 语音数据集各子领域转录文本蒙古语单词重合情况分别如图 3 和 4 所示。

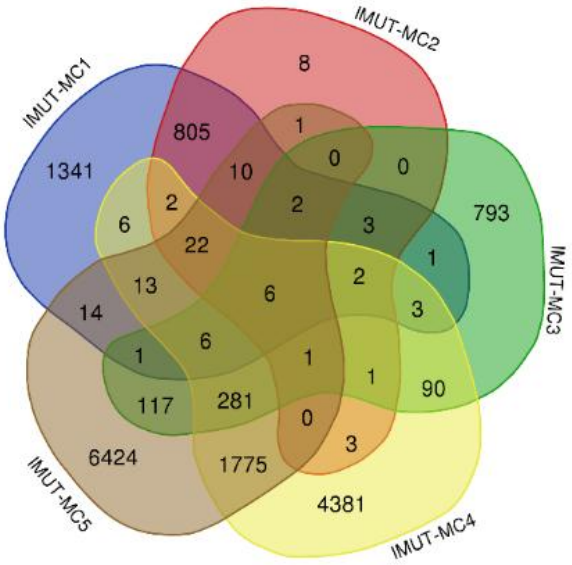


**图 2 IMUT-MC 语料库各期数据集蒙古语单词重合情况**

**Figure 2 The overlap of Mongolian words in all phases of IMUT-MC corpus**

## 2.3 发音词典

在 IMUT-MC 语料库中，IMUT-MC1、IMUT-MC2 和 IMUT-MC3 针对传统蒙古语语音识别研究构建，其涵盖的蒙古语单词发音标注已经完成。而 IMUT-MC4 和 IMUT-MC5 针对端到端蒙古语语音识别研究构建，课题组尚未对其涵盖的蒙古语单词进行发音标注。IMUT-MC 语料库的发音词典通过课题组蒙古族老师和同学对蒙古语单词的发音进行人工标注构建而成，识别基元为音素。其中，蒙古语单词的音素标注以拉丁文字母的形式表示，可直接用于传统蒙古语语音识别实验。目前，发音词典仅涵盖 IMUT-MC1、IMUT-MC2 和 IMUT-MC3 数据集中的蒙古语单词，共有 2092 个蒙古语单词的发音标注。未来，课题组会对发音词典进行扩充，完成 IMUT-MC4、IMUT-MC5 数据集中蒙古语单词的发音标注。发音词典扩充完成后将会第一时间更新。

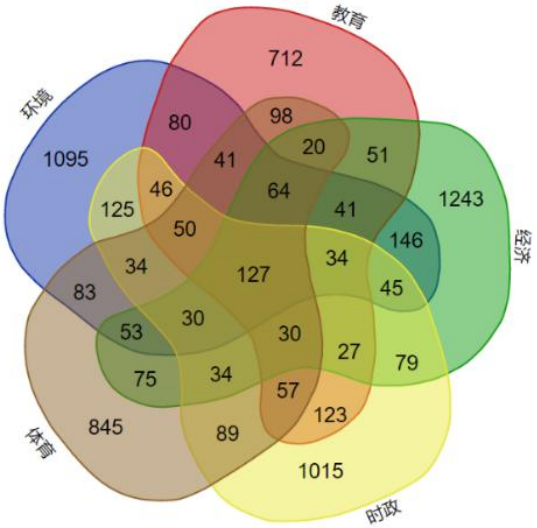


**图 3 IMUT-MC4 各子领域蒙古语单词重合情况**

**Figure 3 Coincidence of Mongolian words in each subfield of IMUT-MC4**

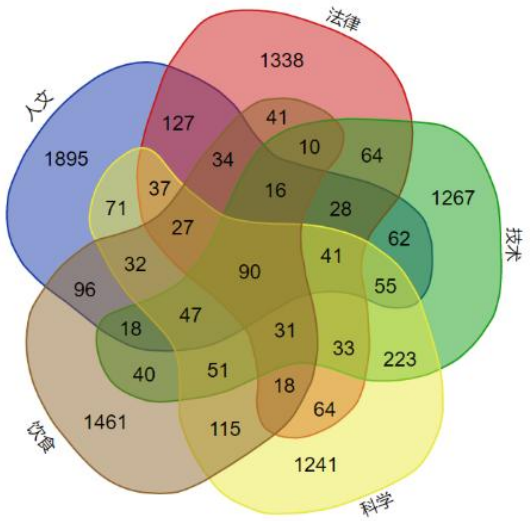


**图 4 IMUT-MC5 各子领域蒙古语单词重合情况**

**Figure 4 Coincidence of Mongolian words in each subfield of IMUT-MC5**

# 3 数据质量控制和评估

## 3.1 数据质量控制

课题组对录制好的蒙古语语音话语逐条检查完成数据质量的控制，筛选内容如下：（1）剔除含有强噪声或电流声的语音话语；（2）剔除含有明显发音错误的语音话语；（3）剔除因文件意外损坏而不能播放的语音话语。

## 3.2 数据质量评估

为了证明 IMUT-MC 语料库在蒙古语语音识别研究中的可靠性，课题组在传统语音识别模型和端到端语音识别模型上完成蒙古语语音识别基线实验。首先构建基于 GMM-HMM、DNN-HMM 的传统语音识别模型和基于 Transformer 的端到端语音识别模型，然后使用 IMUT-MC 语料库完成所有模型的训练，最后以字错率（Character Error Rate，CER）、词错率（Word Error Rate，WER）和句错率（Sentence Error Rate，SER）等评价指标完成模型评估。由于发音词典没有涵盖 IMUT-MC4、IMUT-MC5 中的蒙古语单词，因此 IMUT-MC4、IMUT-MC5 在传统语音识别模型上的可靠性还未得到验证。后续数据集的可靠性实验结果将会在发音词典扩充完成后，第一时间公布。

### 3.2.1 实验设置

**基于 GMM-HMM 的语音识别模型：**课题组使用 Kaldi 实验平台[18]进行实验，选用音素作为建模单元构建 GMM-HMM 蒙古语声学模型，特征提取使用梅尔频谱倒谱系数（Mel Frequency Cepstral Coefficient，MFCC）技术。MFCC 提取特征的参数设置如下：声学特征维度为 40 维，三角滤波器数量为 40 个，倒谱数量为 40，低截止频率为 40 Hz，高截止频率为 -200 Hz。GMM-HMM 蒙古语声学模型训练时，HMM 状态数（HMM-state）为 2500。

**基于 DNN-HMM 的语音识别模型：**课题组使用 Kaldi 实验平台[18]进行实验，选用音素作为建模单元构建 DNN-HMM 蒙古语声学模型，特征提取使用梅尔频谱倒谱系数技术。蒙古语 DNN-HMM 模型由输入层，隐藏层和输出层组成，每个隐藏层有 850 个节点，每个隐藏层在长度归一化后得到该隐藏层的激活输出，并作为下一层的输入。MFCC 提取特征的参数设置如下：声学特征维度为 40 维，三角滤波器数量为 40 个，倒谱数量为 40，低截止频率为 40Hz，高截止频率为-200Hz。在 DNN 网络训练时，超参数设置如表 5 所示，HMM-state 为 2500 个，Batch_size 为 512，初始学习率为 0.0015，最终学习率为 0.00015，不使用 i-vector。解码使用集束搜索（Beam Search）算法，Beam-size 为 11。

**表 5 DNN 网络训练超参数设置**

**Table 5 Hyper-parameter settings of DNN network training**

| 参数名 | 参数说明 | 具体设置 |
|---|---|---|
| Batch_size | 一次训练所选取的样本数 | 512 |
| Beam-size | 集束宽 | 11 |
| Initial_effective_lrate | 初始学习率 | 0.0015 |
| Final_effective_lrate | 最终学习率 | 0.00015 |
| HMM-state | HMM 状态数量 | 2500 |

**基于 Transformer 的语音识别模型：**课题组在 Espnet[19]实验平台进行实验，使用单词作为建模单元来构建基于 Transformer 的蒙古语语音识别模型。基于 Transformer 的蒙古语语音识别模型，编码器具有 12 层，解码器具有 6 层，注意力头为 4 个，维度为 256 维。输入语音特征为 80 维 FBank 特征，在 25 ms 窗口内每 10 ms 计算一次基音，三角滤波器数量为 80 个，采样频率为 16000Hz。模型训练时的超参数设置如表 6 所示，Batch_size 为 16，Dropout-rate 为 0.1，Mtlalpha 为 0.3，优化器使用 noam 优化器，解码使用集束搜索算法，Beam-size 为 10。

表 6 基于 Transformer 的语音识别模型训练超参数设置

Table 6 Hyper-parameter settings Transformer-based speech recognition model training

| 参数名 | 参数说明 | 具体设置 |
|---|---|---|
| Batch_size | 一次训练所选取的样本数 | 16 |
| Dropout-rate | 丢弃率 | 0.1 |
| Optimizer | 优化器 | Noam |
| Attn-dropout-rate | 注意力模块丢弃率 | 0.0 |
| Mtlalpha | CTC 权重设置 | 0.3 |
| Beam-size | 集束宽 | 10 |

### 3.2.2 评价指标

在实验过程中，GMM-HMM 和 DNN-HMM 蒙古语语音识别声学模型使用音素作为建模单元，选用 CER 作为评价指标来评价蒙古语声学模型对音素预测的准确率，选用 WER 作为评价指标来评价蒙古语语音识别的准确率。基于 Transformer 的端到端语音识别模型使用单词作为建模单元，选用 WER 和 SER 作为评价指标来评价蒙古语语音识别的准确率。自动评价指标包括： CER 、WER 和 SER。评价指标的含义如下：

CER 指已知标注文本与解码的结果，将解码结果中错误字符的累计个数除以标注中总的字符数，其公式为：

$$CER=\frac{s+d+i}{n} \tag{1}$$

式中， $s$ 为替换错误的字符数， $d$ 为删除错误的字符数， $i$ 为删除错误的字符， $n$ 为总字符数。

WER 指已知标注文本与解码的结果，将解码结果中错误词的累计个数除以标注中总的词数，其公式为：

$$WER=\frac{S+D+I}{N} \tag{2}$$

式中， $S$ 为替换错误的词数， $D$ 为删除错误的词数， $I$ 为插入错误的词数， $N$ 为总词数。

SER 指已知标注文本与解码结果，将句子识别错误的句子个数除以总的句子个数，其公式为：

$$SER=\frac{S_{incorrect}}{S} \tag{3}$$

式中， $S_{incorrect}$ 为识别错误的句子数， $S$ 为总句数。

### 3.2.3 实验结果

基于 GMM-HMM 和 DNN-HMM 的语音识别模型在 IMUT-MC 语料库上的实验结果如表 7、8 所示。从表中得知，数据集 IMUT-MC1、IMUT-MC2 在 GMM-HMM 和 DNN-HMM 上的字错率和词错率较高。对于 IMUT-MC1，其数据量较小造成模型对于蒙古语语音数据的分布拟合不够充分，导致模型出现欠拟合现象。通过对 IMUT-MC2 中语音数据逐条检查，课题组发现该数据集中存在部分弱噪声数据。而基于 GMM-HMM 和 DNN-HMM 的传统语音识别模型对复杂场景下语音识别的适应效果不高，对 IMUT-MC2 的识别精度较低。同时，课题组剔除 IMUT-MC2 中的弱噪声数据组成 IMUT-MC2 修正数据集，重新进行传统蒙古语语音识别实验，在 GMM-HMM、DNN-HMM 上识别错误率 CER 分别下降 66.3%和 69.2%，WER 分别下降 62.6%和 65.5%。

表 7 基于 GMM-HMM 的语音识别模型基线实验结果

Table 7 Baseline experimental results of speech recognition model based on GMM-HMM

| 评价指标 | 数据集 | | | | |
|---|---|---|---|---|---|
| | IMUT-MC(123) | IMUT-MC1 | IMUT-MC2 | IMUT-MC2 修正 | IMUT-MC3 |
| Dev_CER | 65.49 | 61.28 | 62.45 | 19.45 | 35.32 |
| Test_CER | 67.47 | 62.91 | 64.82 | 21.89 | 37.24 |
| Dev_WER | 68.62 | 63.79 | 66.13 | 24.35 | 38.51 |
| Test_WER | 69.90 | 67.27 | 68.03 | 26.14 | 40.57 |

表 8 基于 DNN-HMM 的语音识别模型基线实验结果

Table 8 Baseline experimental results of speech recognition model based on DNN-HMM

| 评价指标 | 数据集 | | | | |
|---|---|---|---|---|---|
| | IMUT-MC(123) | IMUT-MC1 | IMUT-MC2 | IMUT-MC2 修正 | IMUT-MC3 |
| Dev_CER | 62.45 | 60.32 | 55.49 | 16.51 | 32.21 |
| Test_CER | 63.52 | 62.89 | 58.87 | 18.73 | 33.99 |
| Dev_WER | 66.48 | 64.37 | 59.32 | 20.32 | 35.56 |
| Test_WER | 67.45 | 65.35 | 63.24 | 22.47 | 37.12 |

基于 Transformer 的语音识别模型在 IMUT-MC 语料库上的实验结果如表 9 所示。基于 Transformer 的端到端语音识别模型对复杂场景下语音识别的适应效果较好，对 IMUT-MC2 展现出不错的识别结果。但端到端语音识别模型是基于大数据量建模的概率模型，模型性能随着可用训练数据量的减少而显著下降。由于数据集 IMUT-MC1 数据量规模太小，在基于 Transformer 的语音识别模型下比 GMM-HMM 和 DNN-HMM 等传统语音识别模型下识别精度更差，WER 和 SER 分别为 77.40 和 80.47。数据集 IMUT-MC4、IMUT-MC5 在基于 Transformer 的语音识别模型上的识别错误率高于 IMUT-MC2、IMUT-MC3，是由于它们的转录文本句子较长导致模型在推理预测时一定程度上受到长句依赖问题的影响。

表 9 基于 Transformer 的语音识别模型基线实验结果

Table 9 Baseline experimental results of Transformer-based speech recognition model

| 评价指标 | 数据集 | | | | | |
|---|---|---|---|---|---|---|
| | IMUT-MC | IMUT-MC1 | IMUT-MC2 | IMUT-MC3 | IMUT-MC4 | IMUT-MC5 |
| Dev_WER | 22.2 | 75.32 | 14.92 | 21.12 | 25.61 | 29.73 |
| Test_WER | 26.1 | 77.40 | 15.71 | 21.63 | 27.64 | 30.91 |
| Dev_SER | 25.3 | 79.34 | 13.46 | 18.01 | 30.53 | 33.52 |
| Test_SER | 30.2 | 80.47 | 13.73 | 18.24 | 32.47 | 34.45 |

综上所述，IMUT-MC 语料库是进行端到端蒙古语语音识别研究的可靠语料库。IMUT-MC1、IMUT-MC2 和 IMUT-MC3 数据集是进行传统蒙古语语音识别研究的可靠数据集。IMUT-MC4、IMUT-MC5 在传统蒙古语语音识别模型上的可靠性还未得到验证。后续数据集的可靠性实验结果将会在发音词典扩充完成后，第一时间公布。

## 4 数据使用方法和建议

IMUT-MC 语料库中各期数据集使用方法如下：IMUT-MC1 分为特定说话人文件夹跟非特定说话人文件夹。其中，特定说话人文件夹包含 1 名说话人录制的 1255 句语音话语，非特定说话人包含 7 名说话人随机录制 1255 句语音话语。对于其他数据集的语音数据存放形式，每个说话人作为一个文件夹，文件夹下面包含该说话人的语音数据文件，说话人的手机号作为说话人的唯一标识。每个说话人文件夹命名格式为“录音日期-手机号-地区编号-性别-年龄-录音序号”，语音文件命名格式为“转录文本编号-地区编码-年龄-录音序号”。对于 IMUT-MC 中的转录文本，编码格式为 UTF-8，存放形式为 txt 文本文件。具体内容为“转录文本编号 蒙古语文本”，每个蒙古语单词之间以空格符隔开，例如“00001243 ᠪᠢ ᠪᠣᠯᠣᠨ᠎ᠠ ᠵᠢᠨ᠎ᠠ ᠮᠠᠷᠭᠠᠰᠢ ᠶᠢᠨᠡᠭᠡ ᠵᠢᠨᠡᠭᠡ ᠨᠢ ᠪᠠᠭᠲᠠᠭᠳᠠᠭ ᠦᠭᠡᠢ ᠬᠦᠮᠦᠨ᠎ᠡ ᠵᠣᠬᠢᠶᠠᠨ ᠪᠢᠴᠢᠨ᠎ᠡ”。

由于构建目的不同，IMUT-MC 语料库中各期数据集具有不同的特点，可用于语音识别任务下各种子任务研究。根据各期数据集不同特点，本文给出的数据使用建议如下：数据集 IMUT-MC1、IMUT-MC2 和 IMUT-MC3 具有包含蒙古语单词量较少、转录文本句子较短和说话人个数较多的特点，不但可用于小词汇量蒙古语语音识别研究，而且可用于说话人自适应和说话人识别研究。数据集 IMUT-MC4、IMUT-MC5 具有包含蒙古语单词量较多、转录文本句子较长和说话人个数较多的特点，不但可用于大词汇量蒙古语语音识别研究和说话人自适应研究，而且可用于语音表示研究。IMUT-MC 语料库也可以用于其他与语音相关的任务，如语音合成和语音翻译。

## 致 谢

IMUT-MC 语料库的构建获得内蒙古工业大学蒙古族同学的支持与帮助，在此表示由衷感谢！

## 数据作者分工职责

刘志强（1998—），男，山西省忻州人，在读硕士，研究方向为深度学习、语音识别。主要承担工作：数据论文撰写，数据处理和语料库的整理。

马志强（1972—），男，内蒙古自治区托克托县人，硕士，教授，研究方向为多媒体信息处理、自然语言处理、语音识别、对话生成等。主要承担工作：组织实施语料库的构建，语料库格式规范化。

张晓旭（1997—），男，山东省潍坊人，在读硕士，研究方向为深度学习、语音识别。主要承担工作：最终数据质量控制。

宝财吉拉呼（1983—），男，内蒙古自治区通辽人，博士，讲师，研究方向为机器学习、计算机视觉处理、生物信号处理、多媒体信息处理、自然语言处理等。主要承担工作：语料库格式规范化。

谢秀兰（1979—），女，内蒙古自治区兴安盟人，硕士，讲师，研究方向为蒙古语数据语义。主要承担工作：数据采集和处理。

朱方圆（1997—），男，山东省枣庄人，在读硕士，研究方向为深度学习、语音识别。主要承担工作：数据采集和处理。

## 参考文献

[1] PUJOL P, POL S, NADEU C, et al. Comparison and combination of features in a hybrid HMM/MLP

and a HMM/GMM speech recognition system[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(1): 14–22. DOI:10.1109/TSA.2004.834466.

[2] YU D, DENG L. Deep learning and its applications to signal and information processing[J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154. DOI:10.1109/MSP.2010.939038.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.

[4] SALAZAR J, KIRCHHOFF K, HUANG Z H. Self-attention networks for connectionist temporal classification in speech recognition[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK. IEEE, 2019: 7115–7119. DOI:10.1109/ICASSP.2019.8682539.

[5] LI J Y, ZHAO R, HU H, et al. Improving RNN transducer modeling for end-to-end speech recognition[C]//2019 IEEE Automatic Speech Recognition and Understanding Workshop. Singapore. IEEE, 2019: 114–121. DOI:10.1109/ASRU46091.2019.9003906.

[6] ROUSSEAU A, DELEGLISE P, ESTEVE Y. TED-LIUM: an Automatic Speech Recognition dedicated corpus[C]. Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012.

[7] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane, QLD, Australia. IEEE, 2015: 5206–5210. DOI:10.1109/ICASSP.2015.7178964.

[8] WANG D, ZHANG X. Thchs-30: A free chinese speech corpus[J]. arXiv preprint arXiv:1512.01882, 2015.

[9] BU H, DU J Y, NA X Y, et al. AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, Korea (South). IEEE, 2017: 1–5. DOI:10.1109/ICSDA.2017.8384449.

[10] 马志强, 李图雅, 杨双涛.基于深度神经网络的蒙古语声学模型建模研究[J].智能系统学报,2018,13(03):486-492. DOI:10.11992/tis.201710029.[MA Z Q, LI T Y, YANG S T, et al. Mongolian acoustic modeling based on deep neural network[J]. CAAI Transactions on Intelligent Systems, 2018, 13(3): 486–492. DOI:10.11992/tis.201710029.]

[11] 王勇和, 飞龙, 高光来. 基于 TDNN-FSMN 的蒙古语语音识别技术研究[J]. 中文信息学报, 2018, 32(9): 28 – 34. DOI:10.3969/j.issn.1003-0077.2018.09.006.[WANG Y H, FEILONG, GAO G L. Mongolian speech recognition based on TDNN-FSMN[J]. Journal of Chinese Information Processing, 2018, 32(9): 28–34. DOI:10.3969/j.issn.1003-0077.2018.09.006.]

[12] CHOROWSKI J, WEISS R J, BENGIO S, et al. Unsupervised speech representation learning using wavenet autoencoders[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(12): 2041-2053. DOI:10.1109/TASLP.2019.2938863.

[13] ZHANG Y W, LU X M. A speech recognition acoustic model based on LSTM-CTC[C]//2018 IEEE 18th International Conference on Communication Technology. Chongqing, China. IEEE, 2018: 1052–1055. DOI:10.1109/ICCT.2018.8599961.

[14] 朱方圆，马志强，陈艳，张晓旭，王洪彬，宝财吉拉呼. 语音识别中说话人自适应方法研究综述[J]. 计算机科学与探索, 2021, 15(12): 2241-2255. DOI:10.3778/j.issn.1673-9418.2104068.[ZHU F Y, MA Z Q, CHEN Y, et al. Survey of speaker adaptation methods in speech recognition[J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(12): 2241–2255. DOI:10.3778/j.issn.1673-9418.2104068.]

[15] PRENGER R, VALLE R, CATANZARO B. Waveglow: a flow-based generative network for speech synthesis[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK. IEEE, 2019: 3617–3621. DOI:10.1109/ICASSP.2019.8683143.

[16] ZHOU Z H, CHEN K, LI X S, et al. Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays[J]. Nature Electronics, 2020, 3(9): 571–578. DOI:10.1038/s41928-020-0428-6.

[17] 阿迪雅. 蒙古语会话手册[M]. 赤峰: 内蒙古科学技术出版社, 2009.[A D Y. Mongolian Phrasebook [M]. Chifeng: Inner Mongolia Science and Technology Press, 2009.]

[18] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]. IEEE 2011 workshop on automatic speech recognition and understanding. Big Island USA, 11-15 December 2011.

[19] WATANABE S, HORI T, KARITA S, et al. Espnet: end-to-end speech processing toolkit[C]//Interspeech 2018, ISCA, Hyderabad, India, 2018. DOI:10.21437/interspeech.2018-1456.

## 论文引用格式

刘志强, 马志强, 张晓旭, 等. IMUT-MC：一个针对蒙古语语音识别的语音语料库[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-26). DOI: 10.11922/11-6035.csd.2021.0096.zh.

## 数据引用格式

刘志强, 马志强, 张晓旭, 等. 蒙古语语音语料库 IMUT-MC[DS/OL]. Science Data Bank, 2022. (2022-06-28). DOI: 10.11922/sciencedb.j00001.00349.

# IMUT-MC: a speech corpus for Mongolian speech recognition

**LIU Zhiqiang[1], MA Zhiqiang[1,2*], ZHANG Xiaoxu[1], BAO Caijilahu[1], XIE Xiulan[1], ZHU Fangyuan[1]**

1. College of data science and Application, Inner Mongolia University of Technology, Huhhot, P.R. China

2. Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Huhhot, P.R. China

*Email: mzq_bim@imut.edu.cn

**Abstract:** There is a lack of large-scale speech corpora of Mongolian (a minority language) accessible to researchers for experimental reference, because its users are scattered and it is difficult to collect and label the speech sounds, which hinders the further development of Mongolian speech recognition. Our research group has constructed a speech corpus IMUT-MC for Mongolian speech recognition tasks, which contains

about 212 hours of reading speech recorded by 417 speakers, and we are committed to advancing Mongolian speech recognition research. The research group used IMUT-MC to conduct baseline speech recognition experiments on traditional speech recognition models and end-to-end speech recognition models respectively. The speech recognition models based on GMM-HMM, DNN-HMM and Transformer have word error rates on IMUT-MC, respectively. 69.90%, 67.45% and 26.10%, which proves that IMUT-MC is a reliable corpus for Mongolian speech recognition.

**Keywords:** Mongolian; speech recognition; speech corpus; reading speech

## Dataset Profile

| | |
|---|---|
| **Title** | IMUT-MC: a speech corpus for Mongolian speech recognition |
| **Data corresponding author** | MA Zhiqiang (mzq_bim@imut.edu.cn) |
| **Data authors** | LIU Zhiqiang, MA Zhiqiang, ZHANG Xiaoxu, Bao Caijilahu, XIE Xiulan, ZHU Fangyuan |
| **Time range** | 2017 – 2021 |
| **Type of data** | Single-channel reading speech |
| **Data Sources** | Speech recording |
| **Sampling Rate** | 16000Hz |
| **Data volume** | 25.16 GB (After compression 18.75 GB) |
| **Data format** | *.wav |
| **Data service system** | <http://www.doi.org/10.11922/sciencedb.j00001.00349> |
| **Sources of funding** | National Natural Science Foundation of China (61762070, 62166029); Natural Science Foundation of Inner Mongolia (2019MS06004). |
| **Dataset composition** | The corpus includes 9 data files: (1) *.zip is a compressed package of dataset file, compressed files for IMUT-MC1, IMUT-MC2, IMUT-MC2 corrections, IMUT-MC3, IMUT-MC4 and IMUT-MC5 datasets respectively; (2) lexicon.txt is a pronunciation dictionary file; (3) mongolian_data_prep.sh is a speech recognition experimental data processing script; (4) README.txt is a reading guide file. |