

## 蒙汉语音翻译数据集

戚肖克<sup>1,2</sup>, 特尼格尔<sup>2,3</sup>, 孙媛<sup>2</sup>, 赵小兵<sup>2\*</sup>

ISSN 2096-2223

CN 11-6035/N



文献 CSTR:

32001.14.11-6035.csd.2021.0093.zh

文献 DOI:

10.11922/11-6035.csd.2021.0093.zh

数据 DOI:

10.11922/sciencedb.j00001.00345

文献分类: 信息科学

收稿日期: 2021-12-21

开放同评: 2022-01-28

录用日期: 2022-05-17

发表日期: 2022-06-28

1. 中国政法大学, 北京 102249
2. 国家语言资源监测与研究少数民族语言中心, 北京 100081
3. 中央民族大学中国少数民族语言文学学院, 北京 100081

**摘要:** 目前, 由于缺乏公开数据集, 面向少数民族语言的语音翻译的研究较少。为此, 本文构建并公开了蒙语语音到汉语文本语音翻译数据集 (NMLR-Mon2Chs ST)。本数据集包含 36 位年龄在 20–25 岁之间的蒙古人通过手机录制的蒙语语音, 以及由专业人员标注的蒙语和汉语的文本。为保证数据质量, 对数据进行了预处理, 如去除空语音文件、重采样、归一化后, 最终得到 25 小时的高质量数据, 数据集中音频的平均时长为 4.2 秒。本数据集的建立为探索面向少数民族语言的语音翻译技术提供了一定的数据基础。

**关键词:** 语音翻译; 蒙汉; 少数民族语言; 低资源; 数据集

## 数据库(集)基本信息简介

| 数据库(集)名称 | 蒙汉语音翻译数据集 (NMLR-Mon2Chs ST)                                                                                                      |
|----------|----------------------------------------------------------------------------------------------------------------------------------|
| 数据作者     | 戚肖克, 特尼格尔, 孙媛, 赵小兵                                                                                                               |
| 数据通信作者   | 赵小兵 (nmzxb_cn@163.com)                                                                                                           |
| 数据时间范围   | 2020年                                                                                                                            |
| 地理区域     | 内蒙古自治区呼和浩特市                                                                                                                      |
| 数据量      | 1.62 GB                                                                                                                          |
| 数据格式     | *.wav, *.json                                                                                                                    |
| 数据服务系统网址 | <a href="http://www.doi.org/10.11922/sciencedb.j00001.00345">http://www.doi.org/10.11922/sciencedb.j00001.00345</a>              |
| 基金项目     | 国家语委重点项目 (ZDI135-118)                                                                                                            |
| 数据库(集)组成 | 数据集共包括2个数据文件, 其中, (1) wav.zip 是语音数据, 包含21478个音频文件, 总时长为25小时, 数据量为1.62 GB; (2) text.json是文本数据, 由音频文件名、对应的蒙语文本及汉语文本组成, 数据量为4.9 MB。 |

## 引言

语音翻译 (Speech Translation, ST), 又称为口语翻译 (Spoken Language Translation, SLT), 它的任务是将一种语言的语音转换为另一种语言的文本<sup>[1]</sup>。语音翻译是打破人类交流语言壁垒的一项关键技术, 应用较为广泛, 如电影字幕、国际会议、旅游辅助等。

\* 论文通信作者

赵小兵: nmzxb\_cn@163.com

语音翻译技术建立在自动语音识别 (Automatic Speech Recognition, ASR) 和机器翻译 (Machine Translation, MT) 技术之上。近年来, 随着计算机算力的提升、端到端神经网络方法的提出、数据的剧增等, ASR 和 MT 领域都有了显著的进展, 语音翻译也成为语音信号处理及自然语言处理领域的一个研究热点。

然而, 受公开的数据集限制, 目前 ST 方向的研究大多针对中英<sup>[2]</sup>、英德<sup>[3]</sup>、英法<sup>[4]</sup>、英日<sup>[5]</sup>等语言之间的翻译, 较少机构研究面向少数民族语言的语音翻译。为了缓解这一问题, 本文采集了年龄在 20–25 岁之间的 36 位蒙古族人员的语音, 并由蒙汉专业人员标注了每个音频对应的汉语文本。经整合和预处理后, 共得到 25 小时的有效蒙语语音数据, 形成了蒙汉语音翻译数据集 NMLR-Mon2Chs ST。本数据集不仅可供 ST 领域研究使用, 还可用于 ASR、MT、蒙语语音合成、说话人识别等方向的研究。

## 1 数据采集和处理方法

### 1.1 数据采集方法

蒙汉语音翻译数据集 (NMLR-Mon2Chs ST) 包含语音和文本两部分数据。语音数据由 36 位年龄在 20–25 岁之间的蒙古族说话人通过录制得到, 这些说话人均来自于我国内蒙古自治区呼和浩特市。首先, 准备蒙语文本, 每位录音人员在安静的环境下, 通过手机朗读文本的句子, 进行录音, 朗读的每句保存为一个 wav 格式的语音文件, 文件名为朗读文本中的句序号, 每个说话人的音频放在一个单独文件夹中。之后, 由既懂蒙语又懂汉语的专业人员对每个语音文件标注对应的汉语文本。然后, 整合语音和文本文件, 并对其进行预处理, 最终得到蒙语语音翻译数据集。

### 1.2 数据预处理方法

从 36 位录音人员处收集数据, 数据的形式为每位说话人一个单独文件夹, 文件夹内为以句序号命名的 wav 文件及对应的以句序号命名的蒙文和汉语文本。将此数据集称为原始蒙汉语音翻译数据集, 对此数据集进行预处理, 经过 6 个步骤后, 可以得到最终的蒙语语音翻译数据集。具体的预处理步骤如图 1 所示。

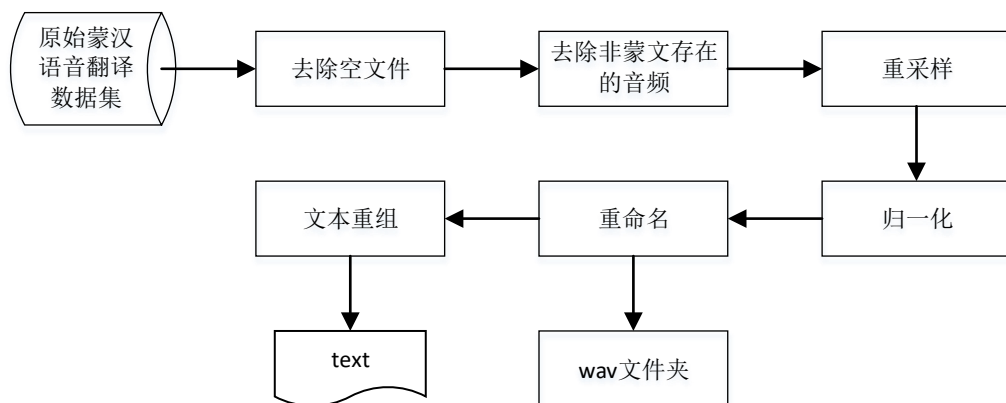


图 1 数据的预处理过程

Figure 1 The process of data preprocessing

第一步, 去除空文件。由于说话人在录制过程中, 存在误触、录制失败等问题, 导致空语音文

件的产生。因此，预处理首先要去除无语音数据的文件。方法为：设置一个阈值，当语音音频时长小于阈值时，认为该文件内不含有意义的语音数据，因此将从数据集中删除该音频文件。在本数据集中，设置阈值为 0.2 秒。

第二步，去除非蒙语存在的音频。在录制的蒙文文本中，存在非蒙文词，如 2020、King、Uncle、Roger 等。由于数量较少，在预处理时简单地将这类文本数据及对应的语音数据从数据集中删除。

第三步，重采样。由于 36 位说话人在不同的时间不同的设备上录制语音，使得数据集中不同的音频文件采样率存在区别，如存在个别音频的采样率为 44.1 kHz。为解决这一问题，对所有音频，重采样至 16 kHz。

第四步，归一化。由于说话人录音时音量高低不一致，导致不同音频信号间强弱差异较大。本文采用归一化将语音数据归于 $[-1,1]$ 范围内，即对每个音频内的值  $y=[y_1, y_2, \dots, y_N]$ ，计算幅度最大值  $m = \max_{n=1,2,\dots,N} |y_n|$ ，则归一化后的音频信号为  $\tilde{y} = y / m$ 。

第五步，按照一定格式重命名音频，具体格式描述如第 2 章所示。

第六步，文本文件重组。原始蒙汉语音翻译数据集中每个音频都对应一个文本文件，不利于数据的处理。因此，将所有音频的文本加入音频名称作为文本标记，全部整合入一个文本中，形成最终的文本文件。

## 2 数据样本描述

本数据集为蒙汉语音翻译数据集，数据集中包含 1 个 zip 压缩包和 1 个文本文件。其中，压缩包内有一个名为 wav 的文件夹，大小为 1.61 GB，未压缩时大小为 2.68 GB。wav 文件夹内包含 36 个子文件夹，每个子文件夹对应一位录音人员的语音数据，命名规则为录音人员的“姓名拼音”与“录制的音频的总时长（以分钟为单位）”。例如，子文件夹“ahei40”表示该文件夹下的音频均为“阿黑”录制，录制的语音总时长约为 40 分钟（由于预处理过程中去除了一部分无效语音，因此最终有效时长略小于此处标记的值）。子文件夹下为多个音频文件，每个文件的命名格式为“该音频所在的子文件夹名称-音频序号.wav”，如“ahei40-0001.wav”、“ahei40-0002.wav”等。对本数据集中 36 位录音人员录制的音频文件数目和音频总有效时长（以分钟为单位）进行统计，结果如表 1 所示。平均每位录音人员录制 597 句，平均有效时长 41.7 分钟。整个蒙汉语音翻译数据集中共包含 21478 个音频文件，有效时长为 25 小时。

表 1 36 位录音人员的音频数据统计表

Table 1 Audio data statistics table of 36 recordists

| 音频文件夹名称    | 音频文件数目 | 时长(分钟) | 音频文件夹名称          | 音频文件数目 | 时长(分钟) |
|------------|--------|--------|------------------|--------|--------|
| ahei40     | 589    | 38.8   | qigen46          | 589    | 44.6   |
| aliya40    | 589    | 38.8   | qilemuge56       | 597    | 55     |
| aliya42    | 589    | 41.1   | sarinuo44        | 600    | 42.1   |
| aoga55     | 595    | 53.1   | sulanga42        | 589    | 41.5   |
| arigunuo44 | 596    | 43.4   | tenggeerwurixi47 | 611    | 46.1   |

| 音频文件夹名称           | 音频文件数目 | 时长(分钟) | 音频文件夹名称        | 音频文件数目 | 时长(分钟) |
|-------------------|--------|--------|----------------|--------|--------|
| arunuo54          | 595    | 52.2   | tugusi49       | 589    | 48.5   |
| ayilahu37         | 600    | 36.2   | tuoya59        | 597    | 57.5   |
| ayisi39           | 596    | 38.3   | wenduer49      | 611    | 47.9   |
| batueerdun37      | 609    | 36.4   | wulijitu46     | 596    | 44.5   |
| bayaliga39        | 596    | 38.3   | wuniritu38     | 609    | 37.9   |
| erimujiletu38     | 587    | 36.8   | wurigumule36   | 610    | 35.1   |
| hairihan35        | 587    | 34.4   | wuyihan32      | 600    | 30.9   |
| hasihu26          | 588    | 25.2   | wuyundalai38   | 588    | 37.2   |
| honggeerdelehei38 | 594    | 36.9   | wuyunqimuge49  | 610    | 48.3   |
| hudeer32          | 593    | 31     | yirigui43      | 599    | 42.4   |
| huriwa49          | 595    | 48.1   | zhagunuo40     | 611    | 39.1   |
| jigeqi55          | 589    | 53.5   | zhalegamuji39  | 611    | 38     |
| nandibilige46     | 586    | 44.2   | zhurihentala38 | 588    | 37.1   |

数据集中的文本文件名为 text.json，大小为 4.9 MB。每个音频文件对应文本中的一个字典，字典中的键“filename”“mon”和“chs”分别表示“音频文件名”“音频对应的蒙文文本”和“音频对应的汉语文本”，示例如表 2 所示。

表 2 音频对应的文本内容示例

Table 2 Samples of text corresponding to audio

| 音频文件名称         | 对应的蒙文文本                            | 对应的汉语文本         |
|----------------|------------------------------------|-----------------|
| ahei40-0001    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 在门厅下面           |
| ahei40-0002    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我就给您拿一些         |
| ahei40-0003    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 如果您还有什么需要 尽管告诉我 |
| ahei40-0004    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 不用担心那个          |
| ahei40-0005    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我要买它 你不需要把它包起来  |
| ahei40-0006    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 你可以改改吗          |
| ahei40-0007    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 红绿灯是红的          |
| ahei40-0008    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我们想要张靠窗户的桌子     |
| ahei40-0009    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 在那边 就在游客信息的前面   |
| ahei40-0010    | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我打网球时扭伤的        |
| .....          | .....                              | .....           |
| wuyihan32-3063 | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 把他们放进篮里好吗       |
| wuyihan32-3064 | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我可以从我的外套拿钱包吗    |
| wuyihan32-3065 | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 有潜水学校吗          |
| wuyihan32-3066 | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 你能在早上七点把我的包拿下来吗 |
| wuyihan32-3067 | ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ ᠠᠨᠠᠭᠤᠨ | 我们需要付饮料费吗       |

| 音频文件名称         | 对应的蒙文文本                                  | 对应的汉语文本      |
|----------------|------------------------------------------|--------------|
| wuyihan32-3068 | би элэл эе эеэвэл                        | 我爱戈雅         |
| wuyihan32-3069 | би элэл эе эеэвэл би элэл эе             | 我渴望见到它们      |
| wuyihan32-3070 | хүүхдүү элэл эе эеэвэл эеэвэл би элэл эе | 有好的卖皮革制品的商店吗 |
| wuyihan32-3071 | хүүхдүү элэл эе эеэвэл би элэл эе        | 我的裤子上有条口子    |
| .....          | .....                                    | .....        |

### 3 数据质量控制和评估

本蒙汉语音翻译数据集由 36 位蒙古族人员在安静环境中录音的音频文件、对应的蒙语文本以及汉语文本组成，在预处理阶段对音频和文本进行了质量控制，去除了无效的音频、非蒙文的句子等，确保数据的可靠性。对音频时长区间的分布进行分析，如图 2 所示，图中的柱状图表示不同音频时长区间在所有音频中的占比，折线图表示不同音频时长区间在所有音频中的累积占比。从图中可以看出，50.7%的音频时长在 2–4 秒，97.8%的音频时长在 8 秒以内。同时，通过计算可以得出，本数据集中音频的平均时长为 4.2 秒。

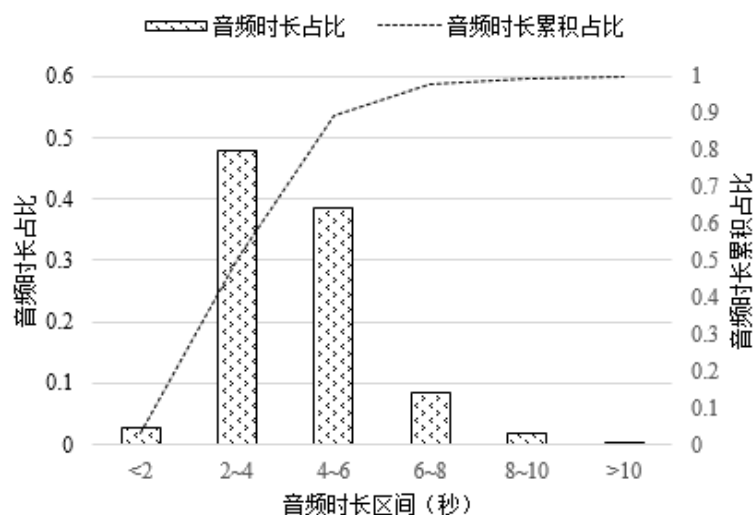


图 2 音频时长区间分布图

Figure 2 Distribution illustration of audio duration

### 4 数据价值

蒙汉语音翻译数据集中的语音来源于 36 位蒙古族人员，年龄在 20–25 岁之间，采用手机录制音频，文本由专门的人员标注，经过整合和预处理后得到 25 小时的可靠数据。本数据可为蒙汉语音翻译研究提供数据基础。此外，本数据集还可用作蒙语语音识别、语音合成、说话人识别等任务的测试集，同时也可作为训练集用于研究小样本下的任务。例如，蒙语语音和蒙文文本可用于小样本下的蒙语语音识别的研究。蒙文文本与汉语文本作为一对平行语料，可用于小样本下的蒙汉机器翻译的研究。平均每个说话人录制了约 600 句音频，可用于研究小样本蒙语语音合成或多说话人蒙语语

音合成算法。语音数据按照说话人分别存储在不同的文件夹下，因此，本数据集也可用于小样本下的说话人识别研究。

## 致 谢

获取本数据集得到呼和浩特民族学院包乌格德勒、斯日古楞的大力支持，在此表示感谢。

## 数据作者分工职责

戚肖克（1985—），女，山东省菏泽市人，博士，副教授，研究方向为语音信号处理、自然语言处理。主要承担工作：数据集的预处理和整合、论文撰写。

特尼格尔（1990—），男，内蒙古自治区呼和浩特市人，博士研究生，研究方向为计算语言学。主要承担工作：数据采集与质量控制。

孙媛（1979—），女，山东省滨州市人，博士，副教授，研究方向为自然语言处理。主要承担工作：数据集前期整合。

赵小兵（1967—），女，内蒙古自治区呼和浩特市人，博士，教授，研究方向为自然语言处理。主要承担工作：数据质量控制与综合管理。

## 参考文献

- [1] SPERBER M, PAULIK M. Speech translation and the end-to-end promise: taking stock of where we are[C]/Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. DOI:10.18653/v1/2020.acl-main.661.
- [2] ZHANG R, WANG X, ZHANG C, et al. BSTC: A large-scale Chinese-English speech translation dataset[J]. arXiv preprint arXiv:2104.03575, 2021.
- [3] CATTONI R, DI GANGI M A, BENTIVOGLI L, et al. MuST-C: a multilingual corpus for end-to-end speech translation[J]. Computer Speech & Language, 2021, 66: 101155. DOI: 10.1016/j.csl.2020.101155.
- [4] KOCABIYIKOGLU A C, BESACIER L, KRAIF O. Augmenting librispeech with French translations: a multimodal corpus for direct speech translation evaluation[EB/OL]. 2018: arXiv: 1802.03142[cs.CL]. <https://arxiv.org/abs/1802.03142>.
- [5] TOHYAMA H, MATSUBARA S, KAWAGUCHI N, et al. Construction and utilization of bilingual speech corpus for simultaneous machine interpretation research[C]/Interspeech 2005. ISCA: ISCA, 2005. DOI:10.21437/interspeech.2005-463.

## 论文引用格式

戚肖克, 特尼格尔, 孙媛, 等. 蒙汉语音翻译数据集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-25). DOI: 10.11922/11-6035.csd.2021.0093.zh.



## 数据引用格式

戚肖克, 特尼格尔, 孙媛, 等. 蒙汉语音翻译数据集[DS/OL]. 中国科学数据, 2022. (2022-01-28). DOI: 10.11922/sciencedb.j00001.00345.

## A dataset of Mongolian-Chinese speech translation

QI Xiaoke<sup>1,2</sup>, BORJIGIN B. Teniger<sup>2,3</sup>, SUN Yuan<sup>2</sup>, ZHAO Xiaobing<sup>2\*</sup>

1. China University of Political Science and Law, Beijing 102249, P. R. China

2. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, P. R. China

3. School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, P. R. China

\*Email: nmzxb\_cn@163.com

**Abstract:** Due to the lack of public datasets, few researches focus on speech translation in minority languages. Therefore, in this paper we constructed a dataset of Mongolian-Chinese speech translation, named “NMLR-Mon2Chs ST”. The dataset consists of Mongolian speech, Mongolian and Chinese texts. First, the Mongolian speech were recorded from 36 Mongols aged between 20 and 25 by recording the audio on their mobile phones. Then, the corresponding Chinese texts were annotated by professionals. In order to ensure the quality of the dataset, we preprocessed the data in it, such as removing the quiet speech, resampling, and normalization. As a result, a total of 25 hours of high-quality data are obtained, and the average duration of audio in the dataset is 4.2 seconds. This dataset is expected to provide certain data support for the research on the speech translation from minority languages to other languages.

**Keywords:** speech translation; Mongolian-Chinese; minority languages; low resource; dataset

### Dataset Profile

|                                  |                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Title</b>                     | A dataset of Mongolian-Chinese speech translation                                                                                                                                                                                                                                                                                                                                   |
| <b>Data corresponding author</b> | ZHAO Xiaobing (nmzxb_cn@163.com)                                                                                                                                                                                                                                                                                                                                                    |
| <b>Data authors</b>              | QI Xiaoke, BORJIGIN B. Teniger, SUN Yuan, ZHAO Xiaobing                                                                                                                                                                                                                                                                                                                             |
| <b>Time range</b>                | 2020                                                                                                                                                                                                                                                                                                                                                                                |
| <b>Geographical scope</b>        | Hohhot, Inner Mongolia                                                                                                                                                                                                                                                                                                                                                              |
| <b>Data volume</b>               | 1.62 GB                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Data format</b>               | .zip (.wav), .json                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Data service system</b>       | < <a href="http://www.doi.org/10.11922/sciencedb.j00001.00345">http://www.doi.org/10.11922/sciencedb.j00001.00345</a> >                                                                                                                                                                                                                                                             |
| <b>Source of funding</b>         | National Language Commission Project (ZDI135-118)                                                                                                                                                                                                                                                                                                                                   |
| <b>Dataset composition</b>       | The dataset consists of 2 subsets in total. The subsets are recorded as “wav.zip” and “text.json”. The former is made up of audio data, with 21,478 files, a duration of 25 hours, and a data volume of 1.62GB; the latter is made up of text data, with a data volume of 4.9MB, which consists of the name of each speech file, the corresponding Mongolian text and Chinese text. |

