专题 2023 多语种智能信息处理数据集

藏汉语音翻译数据集

赵小兵[1,3]，刘佳洛[1,3]，周毛克[1,3]，江雪[1,3]，戚肖克[2*]

1. 国家语言资源监测与研究少数民族语言中心，北京 100081
2. 中国政法大学法治信息管理学院，北京 102249
3. 中央民族大学信息工程学院，北京 100081

摘要：语音翻译研究的前沿取决于可用数据集的质量和多样性。目前在探索少数民族语言的语音翻译时，由于缺乏公开的数据集，相关研究面临着诸多限制。为此，本文构建并公开藏语语音到汉语文本的语音翻译数据集。本数据集来源于微信公众平台以及已公开的藏语语音识别数据集。通过网络爬虫和机器翻译辅助采集数据，并进行人工切分与标注，最终交由专家审核和校正后得到高质量的藏汉语音翻译数据集。本数据集包含样本 7270 条，大小为 965 MB。本数据集为探索低资源藏汉语音翻译技术提供了一定的数据基础，有助于推动相关技术和算法的进步，也为语音翻译系统在少数民族语言环境下的应用提供了实质性的支持。

关键词：语音翻译；藏汉；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	TCST:藏汉语音翻译数据集
数据作者	赵小兵，戚肖克，刘佳洛，江雪
数据通信作者	戚肖克（qixiaoke@cupl.edu.cn）
数据时间范围	2020–2023年
地理区域	西藏、甘肃
数据量	965 MB
数据格式	*.wav, *.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01024
基金项目	国家语委重点项目（ZDI135-118）
数据库（集）组成	数据集共包括音频文件和文本文件，其中，(1) wav中是语音数据，包含7270个*.wav音频文件，总时长为527.1分钟，数据量为965 MB；(2) text.json是文本数据，数据量为810 KB。

引　言

随着科技和社会的进步，不同国家或地区之间的交流变得更为频繁。语音作为人际交流的一种重要方式，使用不同语言的人们迫切希望能够实现无障碍的交流。语音翻译 (Speech Translation, ST)[1]，又称为口语翻译 (Spoken Language Translation, SLT)[2-3]，是一种通过技术实现从源语言的语音转译成目标语言的文本或语音的过程[4]。作为突破人类语言交流壁障的一项关键技术，语音翻译在电影

ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2024.0023.zh

文献 DOI：
10.11922/11-6035.csd.2024.0023.zh

数据 DOI：
10.57760/sciencedb.j00001.01024

文献分类：信息科学

收稿日期：2024-01-19
开放同评：2024-05-15
录用日期：2024-11-05
发表日期：2024-12-20

* 论文通信作者
戚肖克：qixiaoke@cupl.edu.cn

字幕、国际会议、旅游辅助等领域得到了广泛应用。

传统的语音翻译系统由自动语音识别（Automatic Speech Recognition, ASR）[4-5]和机器翻译（Machine Translation, MT）[6]两个系统级联而成。由于 ASR 和 MT 都有大量的公开语料，并且算法相对成熟，级联模型通常能够取得较高的语音翻译性能。然而，级联模型存在错误传播等问题。因此，近年来，基于端到端的语音翻译系统[7-8]的研究成为研究领域的一个热点。通过端到端的方法，可以完全消除错误传播的问题。此外，对于一些不存在书面文字的语言，端到端语音翻译成为唯一的途径。然而，目前的研究还大多集中在高资源语言，如中英[9]、英德[10]、英法[11]、英日[12]。由于缺少公开的数据资源，较少机构具备研究面向藏汉的语音翻译技术。

为此，本文通过对网络上公开数据爬取及对公开的藏语语音识别数据集处理，并交由专家审核，最终经过整合及处理后，获取了包含样本 7270 条，大小为 965 MB 的高质量的藏汉语音翻译数据集。本数据集不仅有助于藏汉语音翻译的研究，还可用于 ASR、MT 等领域的研究。

1 数据采集和处理方法

藏汉语音翻译数据集的构建来源于两种方法：一种方法为爬取微信公众号平台中的藏汉数据，进行切分及对齐处理；另一种方法为将公开的藏语语音识别数据集中的文本经过藏汉机器翻译转换为汉字，随后提交给专家进行人工审核校对。最终，对数据进行整合和归一化处理，得到了一个高质量藏汉语音翻译数据集。

1.1 基于爬虫的数据采集方法

基于爬虫的数据采集方法的流程如图 1 所示。整个过程分成 4 部分：搜索、网络爬虫、切分和对齐。具体步骤如下：

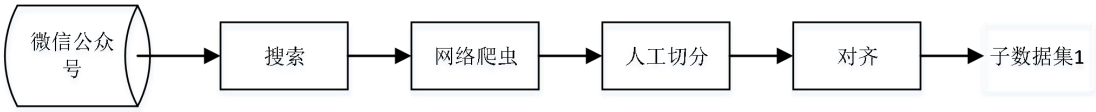

图 1　基于爬虫的数据采集方法流程图

Figure 1　Flowchart of data collection methods based on web crawling

(1) 从微信公众号平台中搜索同时包含藏语语音、藏语文本、汉语文本三种数据的藏汉语音翻译数据。

(2) 通过网络爬虫技术获取数据。具体为：首先通过 request 技术发送请求，获取目标公众号海量文章的链接地址集合，其次针对每篇文章，使用基于 splinter 的爬虫技术来获取各文章的藏语音频和对应文本。

(3) 对爬取的数据，采用 Praat 软件[13]进行句子级别的切分。Praat 是一款专业的语音学软件，主要用于对语音信号进行分析、标注、处理以及合成等。在进行音频子句标注前，首先根据藏语和汉语文本的语义对应关系，人工将长句进行分句，并将藏汉文本进行对应。如图 2 给出了一个示例，左侧为原句，右侧为人工将长句分成了 6 个短句的结果。

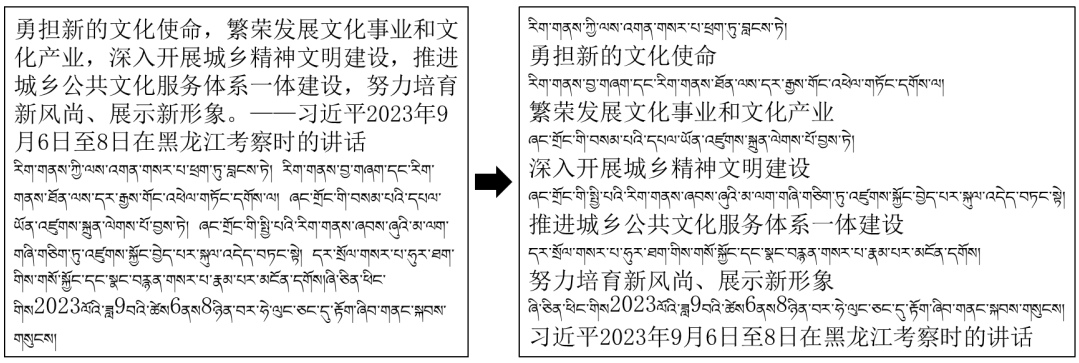

图 2　藏汉平行文本的切分

Figure 2　Segmentation of Tibetan-Chinese parallel texts

然后，根据分句文本对原长音频切分，图 3 给出了通过 Praat 切分音频的图示。该示例中原音频共 70.24s，通过 Praat 切分成了 10 个子句，对每个子句，人工标注对应的藏汉文本。Praat 切分后会生成 TextGrid 文件，里面给出了每个子句的开始和结束位置及对应的文本。编写程序，利用这些信息将长音频切分成多个短音频。

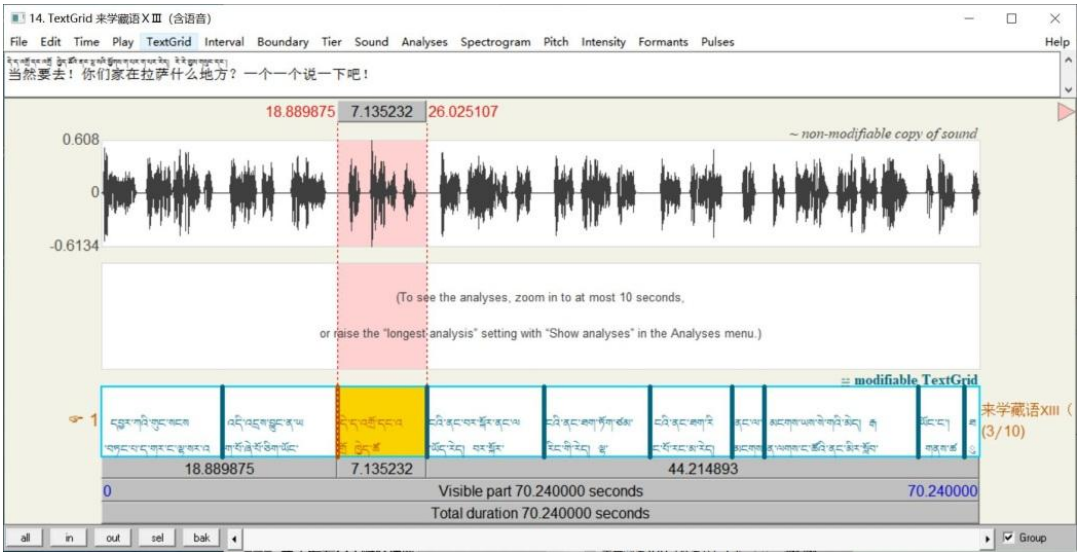

图 3　Praat 切分音频及标注示例

Figure 3　Example of audio segmentation and annotation using Praat

(4) 将所有切分后的语音句与对应的藏语文本和汉语文本进行对齐与整合，形成子数据集 1，命名为 TCST_1_ori。

1.2　基于机器翻译的数据采集方法

本部分数据集语料来自清华大学发布的藏语语音识别数据集，包含藏语语音和对应的藏语文本。通过机器翻译辅助获取藏汉语音翻译的数据集的构建流程如图 4 所示。整个过程分成 3 部分：机器翻译、人工校对及对齐，具体步骤如下：

图 4　基于机器翻译的数据构建流程图

Figure 4　Flowchart of data construction based on machine translation

(1) 通过藏汉机器翻译系统将藏语文本转换为汉语文本。从藏语语音识别数据集中选择男女共 15 人，每人抽取约 400 条数据，共获得 6064 条数据。

(2) 将藏汉平行文本交给专家审核与校对。

(3) 将所有藏语语音句与对应的藏语文本和汉语文本进行对齐与整合，形成子数据集 2，命名为 TCST_2_ori。

1.3　数据处理

TCST_1_ori 和 TCST_2_ori 数据集包含音频和文本文件，两个数据集内音频类型不一致，有 mp3、wav 两种，音频采样率包含 44.1 kHz、16 kHz 不等，且音频信号幅度不统一，数据集文本格式不统一。为了解决这些问题，需要对两个数据集内的数据进行处理。图 5 给出了处理的步骤，具体如下：

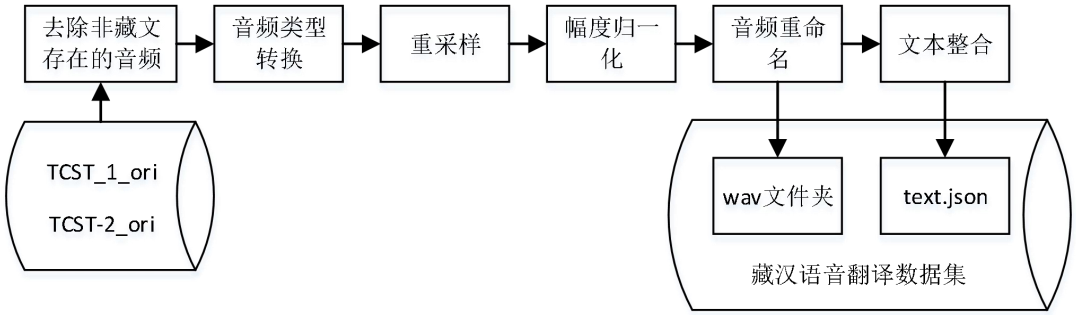

图 5　数据处理过程

Figure 5　Data processing procedure

(1) 去除非藏语存在的音频。在数据集中，存在非藏语词，如 2022、APP 等。在处理时，首先将这类数据从数据集中删除。

(2) 音频类型转换。数据集中存在 mp3 和 wav 格式的音频，将所有音频类型统一转换为 wav 类型。

(3) 音频重采样。由于数据来源不同，不同音频文件的采样率存在区别。因此，对所有音频进行重采样至 16 kHz。

(4) 幅度归一化。不同来源的音频信号间强弱的差异较大，采用归一化将数据幅度规整到-1 与+1 之间。即，对每个音频，$s=[s_1,s_2,...,s_L]$，首先计算幅度的最大值 $M=\max_{l=1,2,...,L}|s_l|$，然后，归一化过程可表示为 $\hat{s}_l=s_l/m$，其中，$l=1,2,...,L$。

(5) 音频文件重命名。以“说话人标识_音频序号.wav”的格式对每个说话人包含的音频进行重命名，其中，音频序号包含三个数字字符，如‘m25-La28_001.wav’，‘m25-La28_002.wav’。

(6) 文本整合。将音频文件、对应的藏语文本和对应的汉语文本数据整合，每条样本对应一个字典，将所有数据写入 json 文件中，形成最终的文本文件。

经过处理后，形成了包含 wav 格式的音频文件和 json 格式文本文件的藏汉语音数据集 TCST。

2 数据样本描述

本藏汉语音翻译数据集包含 1 个 wav 文件夹和 1 个文本文件。其中，wav 的文件大小为 965 MB。wav 文件夹内包含三个子文件夹，U-Tsang、Amdo 和 Kham，分别表示文件夹内的数据属于卫藏方言、安多方言或康巴方言。文件夹 U-Tsang、Amdo 和 Kham 内分别包含 17 个、5 个和 1 个子文件夹，其中每个子文件夹对应一位说话人的音频数据，文件夹名字唯一标识说话人。语音子文件夹下包含多个音频文件，每个音频文件的命名格式为“说话人标识_音频序号.wav”，其中音频序号字符宽度为 3。对本数据集说话人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。整个藏汉语音翻译数据集中共包含 7270 个样本，有效时长为 527.1 分钟。

表 1　音频数据统计表

Table 1　Audio data statistics table

说话人标识	音频文件数目	时长(分钟)	说话人标识	音频文件数目	时长(分钟)
bodad	212	8.3	L_F_0_02	410	29.3
bodkb	212	19.6	L_F_0_05	410	31.6
bodwz	212	10.4	L_F_0_10	409	30.9
cuoxiang	207	12.5	L_F_0_13	410	28.5
maqufa	114	12.7	L_M_0_06	407	22.6
maqufb	70	8.0	L_M_0_10	409	22.2
maqufc	151	18.9	L_M_0_13	409	25.4
maqumd	28	3.5	L_M_0_18	412	22
f21-La41	395	35.2	L_M_0_20	408	27.5
f58-La68	397	24.8	m20-La40	397	40
f6-La11	397	29.3	m25-La28	397	26
f71-La16	397	37.9	**总计**	**7270**	**527.1**

数据集中的文本文件名为 text.json，大小为 2.57 MB。文件内每个样本为一个字典，包含音频文件路径、该音频文件对应的藏语文本及汉语文本，数据格式为：

```
“说话人标识-音频文件名”: {
    ‘audio’: 音频文件路径,
    ‘text’: {
      ‘Tibetan’: 音频文件对应的藏语文本
      ‘Chinese’: 音频文件对应的汉语文本
        }
    }
```

一些样本示例如图 6 所示。


```
"bodad-001": {
  "audio": "Amdo/bodad/bodad_001.wav",
  "text": {
    "Tibetan": "དེ་ཁ་ཉར་རྒེད་མ་འདུག་ཡ།",
    "Chinese": "别老是玩手机"
  }
},
"bodad-002": {
  "audio": "Amdo/bodad/bodad_002.wav",
  "text": {
    "Tibetan": "ཁྱོས་དཀའ་ལས་རྫོང་གོང་འ།",
    "Chinese": "你辛苦了"
  }
},
"bodad-003": {
  "audio": "Amdo/bodad/bodad_003.wav",
  "text": {
    "Tibetan": "ཁྱོས་ཅི་ཞིག་དགོས་ན་ངས་འཐུར་ཡོང་།",
    "Chinese": "你需要什么？我拿给你"
  }
},
```


图 6　文本文件中的样本示例

Figure 6　Sample example in text file

3　数据质量控制和评估

本藏汉语音翻译数据集来自两部分，一部分为微信公众平台，在音频切分阶段同步校对文本，保证数据的准确性；另一部分来源于对藏语语音识别数据集机器翻译的结果，由于机器翻译不能完全准确地进行翻译，因此，邀请了藏语专家进行人工审核并校对，以保证数据的质量。本数据集内容来源于新闻、日常对话、书籍等，应用面较广，保证数据的覆盖范围。

4　数据价值

由于缺少公开的藏汉语音翻译数据集，藏汉语音翻译技术的研究进展较小。本文构建的数据集缓解了这一问题，为藏汉语音翻译的研究提供一定的数据基础。本文采用的两种数据集采集方法：网络爬虫、机器翻译辅助，也为大规模藏汉语音翻译数据集的构建提供思路。同时，由于本数据集中每个样本均包含藏语音频、藏语文本和汉语文本，所以除了用于研究级联藏汉语音翻译系统和端到端藏汉语音翻译系统之外，本数据集还可用于藏汉机器翻译、藏语语音识别的研究。

作者分工职责

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。

刘佳洛（2001—），男，江西省赣州市人，硕士研究生，研究方向为语音识别和语音翻译。主要承担工作：数据采集、数据标注与质量控制。

周毛克（1993—），女，甘肃甘南夏河人，博士研究生，研究方向为自然语言处理。主要承担工作：数据审核与质量控制。

江雪（1998—），女，河北省保定市人，硕士研究生，研究方向为语音翻译。主要承担工作：数据集整合。

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。

致　谢

获取本数据集得到中央民族大学陈波老师、俄见才让、王子豪、常润等的大力支持，在此表示感谢。

参考文献

[1] ORTEGA J E, JOEL ZEVALLOS R, SAID AHMAD I, et al. QUESPA submission for the IWSLT 2024 dialectal and low-resource speech translation task[C]//Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024). Bangkok, Thailand (in-person and online). Stroudsburg, PA, USA: Association for Computational Linguistics, 2024: 125 – 133. DOI: 10.18653/v1/2024.iwslt-1.17.

[2] SHOBA S, SASITHRADEVI A, DEEPA S. Spoken Language Translation in Low - Resource Language[J]. Automatic Speech Recognition and Translation for Low Resource Languages, 2024: 445-459.

[3] HAN X Q, ZHANG W Z. Finding better segmentation granularity for Tibetan-Chinese bidirectional neural machine translation[C]//2024 International Conference on Asian Language Processing (IALP). Hohhot, China. IEEE, 2024: 303 – 308. DOI: 10.1109/IALP63756.2024.10661111.

[4] CHEN J L, LIU Y, XIONG Y Y, et al. Examining sources of language production switch costs amongst Tibetan-Chinese-English trilinguals[J]. International Journal of Bilingual Education and Bilingualism, 2024, 27(8): 1153 – 1167. DOI: 10.1080/13670050.2024.2348559.

[5] FENG S Y, HALPERN B M, KUDINA O, et al. Towards inclusive automatic speech recognition[J]. Computer Speech & Language, 2024, 84: 101567. DOI: 10.1016/j.csl.2023.101567.

[6] TAYIR T, LI L. Unsupervised multimodal machine translation for low-resource distant language pairs[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(4): 1 – 22. DOI: 10.1145/3652161.

[7] YELLAMMA P, VARUN P R, NARAYANA N C N L, et al. Automatic and multilingual speech recognition and translation by using google cloud API[C]//2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI). Lalitpur, Nepal. IEEE, 2024: 566 – 571. DOI: 10.1109/ICMCSI61536.2024.00089.

[8] SPERBER M, PAULIK M. Speech translation and the end-to-end promise: taking stock of where we are[EB/OL]. 2020: 2004.06358.https://arxiv.org/abs/2004.06358v1

[9] ZHANG R Q, WANG X Y, ZHANG C Q, et al. BSTC: a large-scale Chinese-English speech translation dataset[EB/OL]. 2021: 2104.03575.https://arxiv.org/abs/2104.03575v4

[10] 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.

[11] KOCABIYIKOGLU A, BESACIER L, KRAIF O. Augmenting librispeech with French translations: a multimodal corpus for direct speech translation evaluation[C]. LREC, Miyazaki, Japan, 2018.

[12] 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.

[13]BOERSMA P. Praat, a system for doing phonetics by computer[J]. Glot International, 2001, 5(9): 341-345.

论文引用格式

赵小兵, 刘佳洛, 周毛克, 等. 藏汉语音翻译数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-20). DOI: 10.11922/11-6035.csd.2024.0023.zh.

数据引用格式

赵小兵, 戚肖克, 刘佳洛, 等. TCST:藏汉语音翻译数据集[DS/OL]. V2. Science Data Bank, 2024. (2024-10-24). DOI: 10.57760/sciencedb.j00001.01024.

A dataset of Tibetan-Chinese speech translation

ZHAO Xiaobing[1,3], LIU Jialuo[1,3], ZHOU Maoke[1,3], JIANG Xue[1,3], QI Xiaoke[2*]

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

Beijing 100081, P. R. China

2. School of Information Management for Law, China University of Political Science and Law, Beijing 102249, P. R. China

3. School of Information Engineering, Minzu University of China, Beijing 100081, P. R. China

*Email: qixiaoke@cupl.edu.cn

Abstract: The advancement of research frontiers in speech translation relies upon the quality and diversity of available datasets. Currently, the exploration of speech translation for minority languages is subject to numerous constraints due to the limited availability of publicly accessible dataset datasets. To address this gap, this paper aims to construct and release a dataset of speech translation from Tibetan speech to Chinese text. The dataset is derived from the WeChat public platform and publicly available Tibetan speech recognition datasets. We collected the data with the assistance of web scraping and machine translation and performed manual segmentation and annotation. And the data underwent expert review and correction to

ensure its accuracy and quality, resulting in a high-quality dataset of Tibetan-Chinese speech translation. The dataset comprises 7,270 entries with a total size of 965 MB. The dataset can not only provide a foundational data framework for exploring Tibetan-to-Chinese speech translation, but also contribute to the advancement of relevant technologies and algorithms. Moreover, it is expected to offer substantial support for the application of speech translation systems in the context of minority languages.

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

Dataset Profile

Title	A dataset of Tibetan-Chinese speech translation
Data corresponding author	QI Xiaoke (qixiaoke@cupl.edu.cn)
Data authors	ZHAO Xiaobing, QI Xiaoke, LIU Jialuo, JIANG Xue
Time range	2020–2023
Geographical scope	Xizang, Gansu
Data volume	965 MB
Data format	*.wav, *.json
Data service system	<https://doi.org/10.57760/sciencedb.j00001.01024>
Source of funding	National Language Commission Project (ZDI135-118).
Dataset composition	The dataset comprises audio files and text file. The wav folder contains audio data, totaling 7,270 files with a cumulative duration of 527.1 minutes with a data volume of 965 MB. The text.json subset consists of text data with a data volume of 810 KB.