Research on the Construction of a High-Quality Speech data set for Tibetan Dialect Identification Tasks

Journal:	*Data Intelligence*
Manuscript ID	DI-2025-0282.R1
Manuscript Type:	IMLIP 2025 and DIKS 2025
Date Submitted by the Author:	10-Jul-2025
Complete List of Authors:	CaiRang, GaZang; Tibet University, School of Information Science and Technology, Model Base for Training Innovative Talents in Tibetan Information Technology Gao, Dingguo; Tibet University, School of Information Science and Technology, Model Base for Training Innovative Talents in Tibetan Information Technology xu, songtao; Tibet University, School of Information Science and Technology, Model Base for Training Innovative Talents in Tibetan Information Technology
Keywords:	tibetan dialect identification, data set construction, speech processing, High-Quality, Computational Linguistics
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
DI Sample.tex elsarticle.cls	

Research on the Construction of a High-Quality Speech data set for Tibetan Dialect Identification Tasks

GaZang CaiRang, Gao Dingguo†, Xu Songtao

School of Information Science and Technology, Tibet University, Model Base for Training In-novative Talents in Tibetan Information Technology, Lhasa 850000, XiZang Autonomous Region, China

Abstract

Research on Tibetan dialect identification is crucial for exploring the linguistic features of various dialects, advancing intelligent speech technology, and enriching linguistic theory. The construction of a high-quality speech data set forms the foundation of such research. This study integrates insights from Tibetan linguistics and speech signal processing to propose a method for building a data set specifically designed for Tibetan dialect identification and quantification research. First, based on the Tibetan phonemic system, example words and sentences were designed to include Tibetan vowels, consonants, and the "eight-case" grammatical structure. Second, to address the problem of low-quality speech data, an automatic detection method based on signal-to-noise ratio (SNR), pitch, speech rate, and other indicators was introduced to identify and handle abnormal samples. Third, to overcome the issue of limited pronunciation styles in some dialect regions, the GPT-Sovits 4.0 voice cloning model was employed to enrich pronunciation variations. As a result, a speech data set comprising 56,760 samples and approximately 39.8 hours of speech data was developed. Comparative experiments with publicly available data sets of similar scale demonstrate that the constructed data set improves Tibetan dialect identification accuracy by 6.12% (using the k-nearest neighbor algorithm) and 12.27% (using long short-term memory networks) compared to existing public data sets. These findings indicate that the newly constructed data set more effectively captures the phonetic distinctions among Tibetan dialects and enhances performance in dialect identification tasks.

Keywords: tibetan dialect identification; data set construction; speech processing

1. Introduction

Tibetan is a language with multiple dialects, exhibiting significant differences in vocabulary, grammatical structure, and phonology. Among these, phonological variation is particularly pronounced, presenting challenges in areas such as data standardization for Tibetan speech processing and model training. The phonetic distinctions across dialects directly impact the applicability of Tibetan speech processing systems, causing a single system to perform poorly when handling multiple dialects, thereby affecting user experience. Therefore, the development of a high-quality Tibetan dialect automatic identification system holds not only great academic importance but also

†Corresponding author: Gao Dingguo (Email: gdg@utibet.edu.cn)

substantial practical value. As a typical data-driven classification task, the construction of a comprehensive speech data set encompassing multiple dialects is a fundamental prerequisite. Such a high-quality dialect speech data set can serve not only as the foundational resource for model training but also as the basis for optimizing other dialect-related applications and improving system performance. Internationally, notable progress has been made in the development of dialect identification and language recognition data sets. In 2013, T. Schultz et al. [1] released the GlobalPhone data set, which includes samples from over 20 languages and dialects, suitable for multilingual recognition tasks. In the same year, Anguera X et al. [2] introduced the SWS2013 data set for language identification, comprising data from nine low-resource languages, including Albanian.

In China, dialect identification data sets have primarily focused on various Chinese dialects and regional languages. In 1992, Beijing Language and Culture University [3] extended the "Beijing Oral Survey" project initiated in the 1980s and established the "Beijing Oral Corpus Query System" (BJKY), which links text with audio to aid researchers in analyzing the phonetic, lexical, and grammatical features of the Beijing dialect. In 2008, the National Language Committee [4] launched the construction of the Chinese Language Resource Audio Database, which collected audio data of modern Chinese languages, including both dialects and minority languages. In 2017, the Beijing Aliyun AIShell team [5] released the AIShell-1 data set, primarily designed for speaker recognition but also incorporating various Chinese dialects and colloquial speech, making it applicable to dialect identification.

In addition to these data set construction efforts, Chinese scholars have conducted extensive research on methods for building dialect identification databases. In 2008, Chen Xiaoying [6] summarized the processes of corpus design, recording, annotation, and management, proposing a comprehensive construction methodology. In 2011, Gao Yuan et al. [7] introduced design standards for dialect databases. In 2012, Zou Faxin [8] outlined a workflow for speech recording and management systems. In 2013, Du Fuqiang [9] systematically examined the theoretical foundations and implementation methods of dialect database construction, covering data collection and annotation processes.

Currently, there is no specialized Tibetan dialect identification speech data set available in academia. However, early efforts have been made by scholars to construct Tibetan dialect databases. In 1992, Bao Huaqiao et al. [10] developed an acoustic parameter database for the Lhasa Tibetan dialect. In 2007, Yu Hongzhi et al. [11] built an acoustic parameter database for the Amdo Tibetan dialect. In 2013, Kong Changqing et al. [12] constructed a large-scale continuous Tibetan telephone speech recognition database targeting the dialect of the Tibet Autonomous Region. In 2015, Lu Rongjiang et al. [13] created a Tibetan–Chinese bilingual multimodal physiological speech database based on the Tibetan dialect. In 2018, Huang Xiaohui et al. [14] established a Tibetan oral speech corpus based on the Lhasa Tibetan dialect, covering aspects such as corpus selection, recording standards, and annotation guidelines. However, most of these data sets are limited to a single dialect, making them unsuitable for dialect identification tasks, and they are not publicly accessible. Additionally, existing resources used in Tibetan dialect research, such as the Tibetan–Chinese Lhasa Oral Dictionary, Amdo Tibetan Oral Dictionary, and Tibetan–Burmese Phonology and Vocabulary, do not meet the requirements for speech recognition tasks. Public Tibetan speech data sets like TIBMD@MUC [15] and XBMU-AMDO31 [16] include multiple dialects but are primarily designed for Tibetan speech recognition and synthesis tasks. Although they can process recognition tasks involving the three main Tibetan dialects (Central Tibetan, Amdo, and Kham), they perform poorly in fine-grained dialect identification due to limited phoneme coverage and incomplete dialect representation. Moreover, constructing

a Tibetan dialect identification speech data set presents several challenges: Tibetan dialects are highly diverse, with considerable phonological differences, making it difficult to collect comprehensive and representative data. Furthermore, speech data cleaning and classification are both time-consuming and costly.

To address these challenges, this paper proposes the design of a high-quality speech data set specifically for Tibetan dialect identification, resolving issues such as incomplete dialect representation and insufficient phoneme coverage found in existing public data sets. This data set holds significant value for advancing future research in Tibetan dialect identification. The construction process is outlined as follows: 1) Data Collection: Based on linguistic research on Tibetan dialects and the Tibetan phonological system, a comprehensive data collection plan (referred to hereafter as "the plan") was formulated. According to this plan, speech data were gathered from 24 dialect regions, yielding 5,160 speech samples (2,760 words and 2,400 sentences). 2) Abnormal Sample Detection: To ensure data quality, an abnormal sample detection method based on key indicators such as signal-to-noise ratio, speech rate, and pitch was developed using box plot analysis. A total of 384 low-quality samples were identified and either removed or re-recorded. 3) Data Augmentation: To mitigate the limited number of speakers in certain dialect regions, the GPT-Sovits 4.0 voice cloning model was employed to enhance voice style variation for each speaker's data. As a result, the original data set was expanded to include 56,760 speech samples, forming the Tibetan Dialect Speech Corpus 2024 (TDSC-2024), encompassing approximately 39.8 hours of speech data. The construction process is illustrated in Fig. 1. Construction Process of a High-Quality Speech data set for Tibetan Dialect Identification.

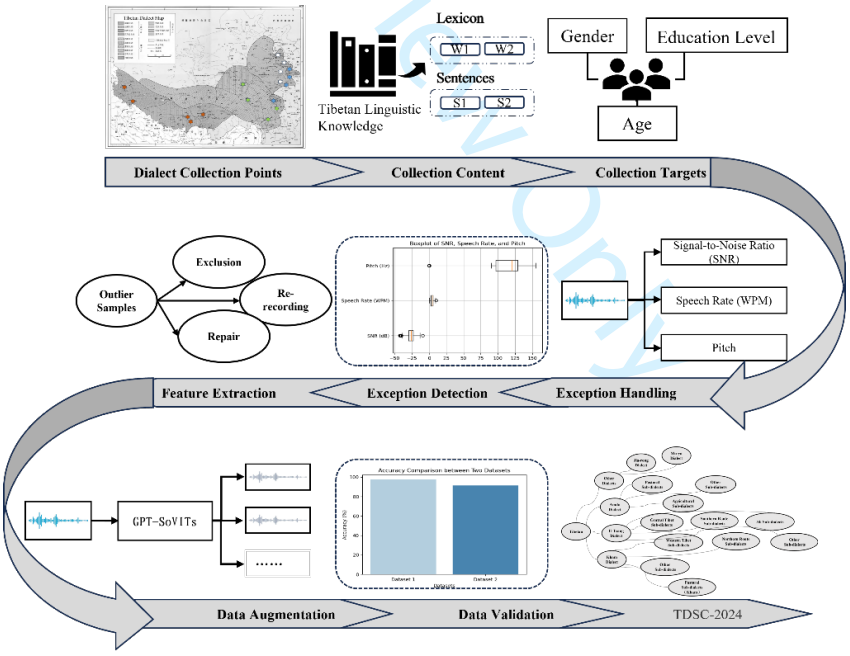


Figure 1: Construction Process of a High-Quality Speech data set for Tibetan Dialect Identification

Subsequently, to evaluate the dialect discrimination capability of TDSC-2024 compared to other publicly available, non-specialized Tibetan dialect speech data sets, automatic dialect iden-

tification experiments were conducted for the three major Tibetan dialects using both public data sets (TIBMD@MUC, XBMU-AMDO) and TDSC-2024. Mel Frequency Cepstral Coefficients (MFCC) [17] and Gamma Frequency Cepstral Coefficients (GFCC) [18] were combined as input features, and experiments were performed using both K-Nearest Neighbors (KNN) [19] and Long Short-Term Memory (LSTM) [20] models. When controlling for comparable data set sizes, TDSC-2024 achieved accuracy improvements of 6.12% (KNN) and 12.27% (LSTM) over the publicly available data sets. These results demonstrate the significant advantage of TDSC-2024 in Tibetan dialect identification tasks, as it more effectively captures and represents phonetic differences among various dialects.

2. Methodology for Constructing the Tibetan Dialect Identification Speech data set

As described above, the construction method consists of three parts: Tibetan dialect speech collection, abnormal sample detection and processing, and GPT-SoVITS-based speech style transformation.

2.1. Tibetan Dialect Speech Collection

2.1.1. Tibetan Dialect Distribution and Col-lection Point Selection

There are differing opinions within the traditional linguistics community regarding the classification of Tibetan dialects and the selection of representative local varieties. The views of scholars such as Gesang Junmian [21] and Qu Aitang [22] are widely regarded as authoritative and objective, aligning closely with the actual linguistic characteristics of Tibetan. Based on the traditional division into three main Tibetan dialects, Gesang Junmian proposed a further subdivision into seven sub-dialects, a classification that has gained broad acceptance in the academic community. This study primarily adopted this classification method when selecting dialect collection points. Additionally, relevant content from the Atlas of Chinese Languages: Volume 2, Minority Languages [23] was consulted to subdivide the Central Tibetan dialect into the Front Tibetan, Rear Tibetan, and Ali sub-dialects. Consequently, eight Tibetan sub-dialects were established in total.

The collection points were designed to cover representative local dialects of these eight sub-dialects, along with other Tibetan dialects possessing unique features, such as the Jiayong, Zhuoni, Diebu, and Moya dialects. Specific collection points and their corresponding dialect classifications are detailed in Appendix 1. Each collection point is categorized into two hierarchical levels: dialect regions and sub-dialect regions. This multi-level data structure provides flexible support for tasks requiring different levels of dialect identification.

2.1.2. Corpus Design

The content collected for TDSC-2024 primarily follows the framework outlined in the Survey Handbook of Chinese Language Resources Audio Database: Tibetan–Burmese Language Family. The design includes two main linguistic elements: vocabulary items and grammatical example sentences. For the vocabulary component, a Swadesh core vocabulary list served as the foundation, supplemented with knowledge from Tibetan phonology and commonly used Tibetan words to form the TDSC vocabulary list. This list covers all Tibetan consonants and vowels in their written form and aims to collect speech data on the most frequently used words across various Tibetan dialects. For the example sentences, the design was informed by the handbook and integrated with the Tibetan "eight-case" grammatical structure. Emphasis was placed on

capturing the phonetic and prosodic features of different dialects, including intonation, tonal variation, and sentence cohesion. This approach was intended to better reflect the pronunciation and prosodic characteristics distinctive to each dialect, as shown in Table 1.

The collected material contributed to the development of the Tibetan Dialect Automatic Identification Recording Manual (referred to as the TDSC-2024 Manual). Tibetan was used as the primary language for the collected content, accompanied by Chinese translations to help speakers fully understand the semantic meaning before recording their utterances. This method minimized the risk of unnatural or overly formal reading that can occur when only Tibetan content is provided, ensuring that the collected samples more accurately reflected the natural pronunciation and lexical habits of each dialect.

Table 1: TDSC-2024 Content Structure

Linguistic Granularity	Quantity	Design Purpose
Words	115	To analyze the linguistic features of words
Sentences	100	To analyze the prosodic features of sentences

2.1.3. Speaker Selection

For the data collection points (Appendix 1), local residents were selected as speakers. Each speaker recorded the designated collection texts (Table 1). Most participants had low levels of formal education and had resided in the local area for extended periods, ensuring that their speech accurately represented the authentic features of the local dialect in everyday use. To comprehensively capture pronunciation characteristics across different age groups and genders, a systematic design was applied to the gender and age distribution of the speakers. Speech data were collected from a total of 24 speakers, with their gender and age distribution detailed in Table 2 and Table 3.

Table 2: Gender Composition

Gender	Male	Female
Quantity (people)	8	16
Proportion (%)	33.33	66.67

Table 3: Age Composition

Age Group	Above 18 Years	19-49 Years	50 Years and Above
Quantity (people) ZXCV	4	12	8
Proportion (%)	16.67	50	33.33

2.1.4. Labeling Structure

Speech data from 24 speakers were collected across 24 dialect collection points. After editing and organizing the recordings, a Tibetan dialect speech database comprising nearly 258 minutes and 5,160 entries was constructed. To facilitate subsequent data processing and recognition studies, a systematic four-level data labeling structure was designed. This structure enables multi-level management of the speech data, making the corpus more efficient and flexible for various applications.

The multi-level labeling system consists of four layers of information: dialect region, sub-dialect region, dialect collection point, and content identifier. For example, the word "ཁྲིབ་མ" (meaning "shadow") collected in Maqu County is labeled as "Am-duo.Amduomuqu.MaQuXian.W20." The specific labeling rules are as follows: "Amduo" (Amdo Dialect) – First-level label: Indicates the general dialect category to which the sample belongs. "Amduomuqu" (Pastoral Sub-dialect) – Second-level label: Specifies the sub-dialect classification of the sample, providing a finer distinction within the broader dialect category. "MaQuxian" (County-level administrative region) – Third-level label: Denotes the specific dialect collection point, offering detailed location information. This level can be further refined in future data expansion efforts to increase the precision of collection site identification. "W20" (Collection content) – Fourth-level label: Identifies the specific recorded content. In this example, "ཁྲིབ་མ" (meaning "shadow") corresponds to the content identifier W20 in the TDSC-2024 manual. "W" denotes words, while example sentences are labeled as "S××."

A specific example of the labeling structure is shown in Table 4.

Table 4: Specific Label Examples

Speech Phrase	Label	ID
ཁྲིབ་མ (meaning "shadow")	Amduo.Amduomuqu.MaQu xian.W20	W20
ཁྱེད་རང་ལམ་གཅིག་ཡོད། (meaning "How many people are there in your family?")	Amduo.Amduomuqu.MaQu xian.S13	S13

2.2. *Anomaly Detection and Processing*

The 24 dialect collection points included in TDSC-2024 are distributed across five provincial-level administrative regions. Among these, Qinghai Haidong and Tibet Ali are separated by nearly 3,000 kilometers, making field data collection extremely costly. Consequently, some collection points were outsourced to other locations for recording. While this approach helped reduce collection costs, it also introduced uncontrollable factors that could affect the quality of the dialect samples. To address the challenge of automatically detecting low-quality samples arising from these variations, this study proposes a boxplot-based anomaly detection method, SNR-Pitch-Words Per Minute Boxplot Outlier Detection (abbreviated as SPROD). This method evaluates three key indicators: signal-to-noise ratio (SNR), pitch, and speech rate. The specific workflow of this method is illustrated in Fig. 2.

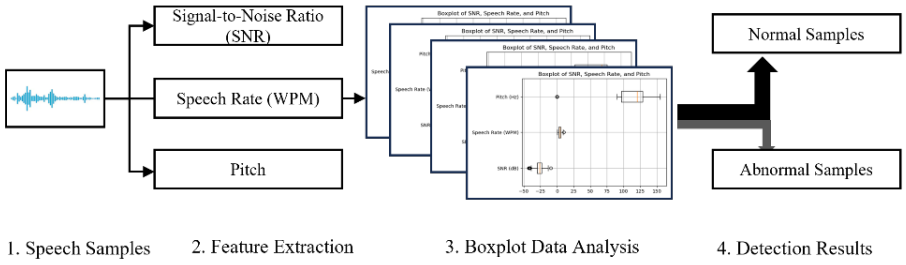


Figure 2: SPROD Anomaly Speech Sample Detection Process

2.2.1. Key Features in the SPROD Method

The SNR represents the ratio of signal power to noise power in an audio sample and serves as a key indicator of audio clarity. Signal and noise components can typically be distinguished using Power Spectral Density (PSD) analysis. Specifically, noise segments generally exhibit lower frequencies and smaller PSD values, whereas signal segments display relatively higher frequencies and larger PSD values. The PSD is calculated as follows:

$$S(f) = |X(f)|^2/T \tag{1}$$

$S(f)$ represents the power spectral density of the signal at frequency f, $X(f)$ is the Fourier transform of the signal at frequency f, and T is the time duration of the signal. Based on the average PSD values within the signal and noise frequency bands, the signal and noise power are calculated, and the SNR is determined as follows:

$$SNR = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) \tag{2}$$

P_{signal} represents the signal power, and P_{noise} denotes the noise power. Experimental results indicate that defining frequencies above 2000 Hz as the signal band and those below 500 Hz as the noise band yields higher accuracy in anomaly sample detection.

Words Per Minute (WPM) refers to the number of words or characters spoken within a given unit of time and is commonly used as a metric for speech fluency. Accurate calculation of speech rate typically relies on speech recognition technology supported by linguistic models to determine the actual word count, which in turn requires a high-precision speech recognition module. However, due to the lack of mature recognition technology for the dialect data used in this study, such an approach is difficult to implement. Therefore, a speech rate estimation method based on the Zero Crossing Rate (ZCR) is employed instead. This method estimates speech rate by analyzing short-term variations in the zero-crossing rate of the speech signal, thereby avoiding dependence on a complex speech recognition system and enabling effective estimation of speech rate even in the absence of high-precision recognition technology.

Generally, as speech rate increases, the ZCR value also rises because faster speech contains more high-frequency components. These lead to more frequent changes in the signal's sign, resulting in a higher zero-crossing rate. This relationship was validated through empirical sample comparisons. As shown in Fig. 3, the ZCR values of the Tibetan phrase "བཀྲ་ཤིས་དང་ཙེ་རིང་" (meaning "Zashi and Tshering") at different speech rates exhibit this pattern: faster speech produces denser audio fluctuations and correspondingly higher ZCR values (see Table 5).

Table 5: ZCR Values at Different Speeds (ms)

Speech Rate	ZCR Value
Slower	0.0338
Normal	0.0370
Fast	0.0394
Fastest	0.0438

Pitch can be estimated by determining the fundamental frequency of the audio signal, measured in Hertz (Hz). The calculation method is as follows:

$$f_{\text{pitch}} = \frac{1}{N}\sum_{i=1}^{N} f_0(i) \tag{3}$$

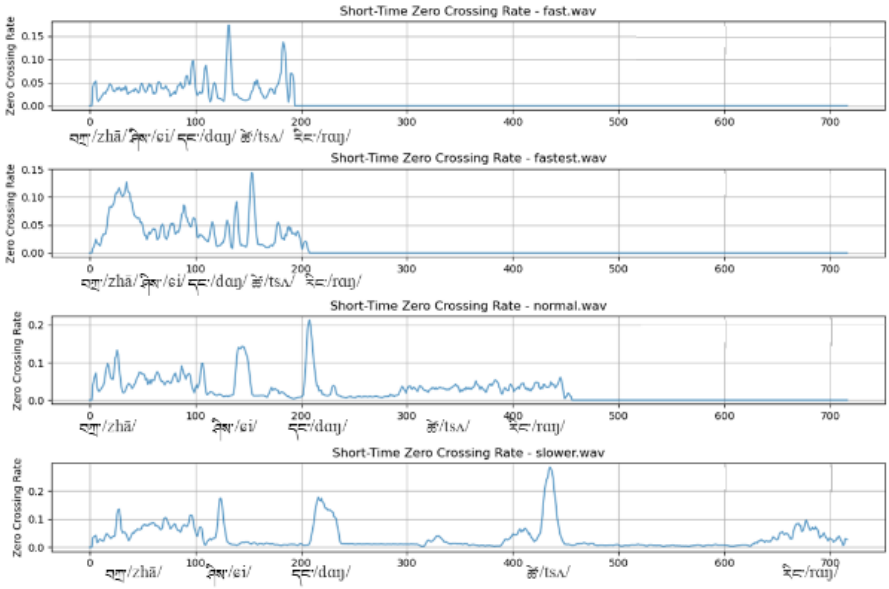


Figure 3: Comparison of Audio Waveforms at Different Speed.

f_{pitch} represents the average pitch of the audio signal; $f(i)$ is the fundamental frequency at the i-th time instance; N is the number of voiced segments, representing the total number of valid fundamental frequency values.

The SPROD method analyzes three key features of audio samples (i.e., SNR, WPM, and Pitch) to efficiently detect low-quality anomalous samples in the data set, thereby enhancing the overall quality of the speech data. SNR, speech rate, and pitch are fundamental parameters of audio signals. In this study, Boxplot analysis combined with the Interquartile Range (IQR) statistical method is employed to automatically identify anomalous data within the samples. In practical application, when the SPROD method was applied to a data set from a specific speaker, it effectively and intuitively identified anomalous samples. The effectiveness and advantages of this approach are further validated through experimental comparisons presented in the following sections.

2.2.2. *Experiment and Analysis*

To verify the practical effectiveness of the SPROD method, a series of comparative experiments were conducted. These experiments were based on the assumption that speech rate, pitch, and other relevant indicators for the same speaker should remain relatively consistent across different recording samples. Accordingly, recordings from the same speaker were selected as the detection targets. The experiments included four different anomaly detection approaches: 1) Detection using the Box Plot method (i.e., the SPROD method proposed in this study); 2) Detection based on the Z-score method; 3) Combined detection by taking the union of results from the Box Plot and Z-score methods; 4) Combined detection by taking the intersection of results from the Box Plot and Z-score methods.

2.2.3. Experimental Corpus

The experimental corpus comprised samples from four speakers collected during the preliminary stage of this study, totaling 2,624 speech samples. All samples were manually inspected and individually labeled, with the labeling format presented in Table 6.

Table 6: Speech Sample Labels

File	Ture_label
sample (1).wav	1
sample (2).wav	0
sample (3).wav	0

In this study, a True_label value of 1 indicates that the sample was manually classified as an anomalous sample. Reasons for this classification include excessively fast speech rate, abnormal pitch, or a low signal-to-noise ratio. Conversely, a True_label value of 0 indicates that the sample was deemed normal. After manual inspection, a total of 149 low-quality speech samples, accounting for 5.67% of all samples, were found not to meet the required standards.

2.2.4. Experimental Setup

The experimental environment consisted of a system equipped with 16 GB of RAM, an Intel i7 13th-generation processor, and an NVIDIA 3050 GPU. Python 3.12 was used as the programming language for all experiments.

3. Comparative Experiments and Analysis

This study presents the performance of the four detection methods using visual charts and statistical data, with the results summarized in Table 8. To enable a more detailed comparison, the baseline detection method was tested under eight different threshold combinations to assess its detection performance across various threshold settings. These threshold combinations are listed in Table 7.

Table 7: Threshold Combinations

Threshold name	SNR	SPEED_LOW	SPEED_HIGH	F0_MIN	F0_MAX
1	5	120	200	80	300
2	10	100	180	70	250
3	8	130	220	90	310
4	15	110	190	85	280
5	12	125	205	75	270
6	7	115	185	80	125
7	9	135	215	85	300
8	6	140	220	90	310

Based on the experimental results, the Z-score-based detection method achieved its best performance at Threshold 8, standing out among all threshold combinations. The detailed results are presented in Table 8 and Table 9. At this threshold, when compared with the four detection methods, including SPROD, it was observed that the SPROD method outperformed the others, achieving the highest detection accuracy, precision, and F1 score, 92%, 22%, and 16%,

Table 8: Detection Performance of Different Thresholds Based on Z-Score

Threshold name	Accuracy	Precision	Recall	F1 Score
1	0.88	0.10	0.12	0.14
2	0.86	0.17	0.11	0.13
3	0.82	0.12	0.14	0.13
4	0.87	0.14	0.15	0.14
5	0.85	0.12	0.12	0.12
6	0.86	0.15	0.08	0.10
7	0.87	0.14	0.10	0.12
8	0.90	0.19	0.11	0.15

Table 9: Effectiveness of Different Detection Methods Based on Threshold 8

Method	Accuracy	Precision	Recall	F1 Score
Z-score	0.90	0.19	0.11	0.15
Z-score & Box Plot	0.90	0.15	0.13	0.14
Z-score — Box Plot	0.89	0.14	0.12	0.13
SPROD	0.92	0.22	0.12	0.16

respectively. Additionally, the recall rate reached 12%, ranking second among all methods, thus demonstrating the comprehensive performance advantage of the SPROD method. However, the experiments also revealed that despite SPROD's excellent accuracy, its precision, recall, and F1 score remained relatively low. Upon further analysis, this was attributed primarily to the severe class imbalance in the data set: the majority of samples belonged to the negative class (normal samples), while only 5.67% of the total were positive class samples (anomalous samples). This imbalance led to a higher occurrence of false positives and false negatives when the model processed positive class samples, thereby affecting precision, recall, and F1 score metrics. Nonetheless, the overall accuracy remained high, confirming the method's effectiveness for large-scale voice sample anomaly detection tasks.

Moreover, these results indirectly validate the effectiveness of the three key indicators (i.e., SNR, speech rate, and pitch) used in the SPROD method for automatic detection and recognition of abnormal voice samples. In addition to the quantitative performance, the superior anomaly detection capability of the methods is visually demonstrated by the data clustering results shown in Fig. 4 and Fig. 5 (where blue represents normal samples and red represents anomalous samples). The Z-score-based detection method showed limited effectiveness in identifying anomalous samples, whereas the Box Plot-based SPROD method more accurately detected outlier samples that deviated significantly from the normal range.

4. Application of the SPROD Method

The SPROD method was applied to detect outliers within the 5,160 collected speech data samples, comprising 2,760-word samples and 2,400 sentence samples. The analysis was conducted on a speaker-by-speaker basis, resulting in the identification of 420 low-quality samples. Following manual verification, 384 of these samples were confirmed as invalid and subsequently discarded. To meet the actual data requirements, these discarded samples were re-collected and supplemented accordingly.

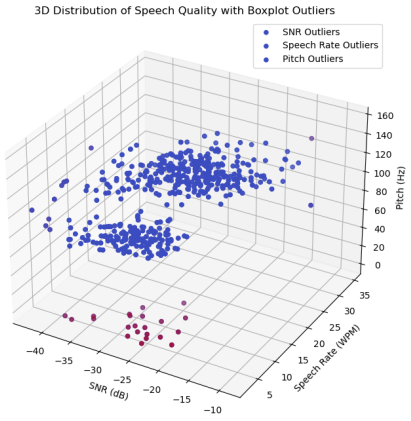


Figure 4: Detection Performance of the SPROD Method.

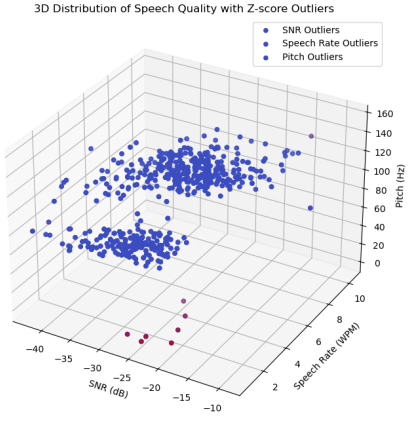


Figure 5: Detection Performance of the Z-Score Method.

4.1. Speech Style Conversion Based on GPT-SoVITS

GPT-SoVITS is a speech cloning and style transfer technology that integrates the GPT model with the SoVITS (Speech Variations and Implicit Text-to-Speech) model. This method enables the generation of speech data from various speakers using only a small amount of original speech while incorporating diverse speech style attributes such as gender, age, and accent. The core principle leverages the GPT model's language understanding capabilities alongside the SoVITS model's proficiency in capturing and processing speech characteristics. This allows for the simulation of distinct speaker features while preserving speech quality.

By applying this technology, a wide variety of speech styles can be produced from limited data, thereby improving the robustness of the speech synthesis system across different speech environments.

To reduce data collection costs, TDSC-2024 selected only one speaker per dialect region, resulting in a relatively uniform speaker structure at each collection point. However, since speech characteristics vary significantly across age groups and genders, the GPT-SoVITS speech cloning model was employed to transform each single-speaker data set into multiple speaker styles. This approach not only expanded the data set but also enhanced the model's ability to generalize across different speaker tones. The specific process is illustrated in Fig. 6.

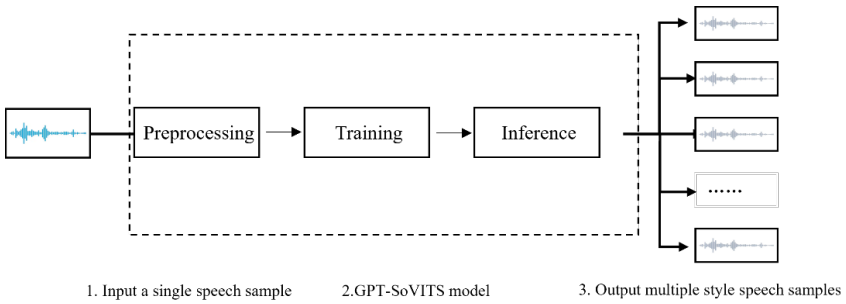


Figure 6: Speech Style Transfer Method Based on GPT-SoVITS.

In the field of speech conversion and synthesis, the Mean Opinion Score (MOS) is widely used to evaluate speech quality. Evaluators assess the speech samples based on factors such as clarity and naturalness. MOS ratings categorize speech quality into five levels, with a score of 5 representing excellent and highly natural speech, and a score of 1 indicating poor and unacceptable quality. The MOS is calculated as follows:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{4}$$

where N is the number of evaluators and x_i is the score given by the i-th evaluator. In this study, 20 randomly selected speech samples generated by GPT-SoVITS were subjected to subjective MOS evaluation, with $N = 15$. The resulting MOS score was 3.89, indicating that the generated speech was of good quality, relatively clear and natural, and generally acceptable to most evaluators. However, some issues such as background noise and a lack of naturalness were noted, with certain details requiring further refinement. Despite these shortcomings, the generated speech is suitable for use in training tasks for the subsequent dialect identification model.

In this study, GPT-SoVITS 4.0 and its publicly available 10-speaker models were employed to perform a 1×10 pitch conversion enhancement on the speech data of each original speaker. This process expanded the original data set to 56,760 speech samples, resulting in a Tibetan dialect identification speech data set (TDSC-2024) comprising nearly 39.8 hours of audio data.

4.2. Validation of TDSC-2024

To evaluate the performance of the TDSC-2024 data set in Tibetan dialect identification tasks, this study conducted comparative experiments using TDSC-2024 and other publicly available data sets for the automatic identification of the three major Tibetan dialects. In these experiments, MFCC and GFCC were combined as input features, while KNN and LSTM models were employed as classification algorithms.

4.2.1. Experimental Data

In China, the Tibetan language is generally classified into three major dialects: Ü-Tsang, Amdo, and Kham. Specifically, representative local varieties of the Ü-Tsang dialect include the Lhasa, Shigatse, and Shannan dialects; those of the Amdo dialect include the Gansu Xiahe and Qinghai Zeku dialects; and those of the Kham dialect include the Sichuan Dege, Qinghai Yushu, and Tibet Changdu dialects. The TDSC-2024 corpus further subdivides these three major dialects into eight subdialects: front Ü-Tsang, rear Ü-Tsang, Ali, Amdo pastoral, Amdo agricultural, southern Kham, northern Kham, and Kham pastoral subdialects. Additionally, the corpus includes distinctive local varieties such as Jiayang, Zhuoni, Diebu, and Muya. In contrast, publicly available data sets such as TIBMD@MUC and XBMU-AMDO31 classify Tibetan solely into the three major dialect categories without further subdivision.

To ensure a fair and noticeable comparison, this study extracted an equal number of speech samples from both the public data sets and the TDSC-2024 data set for the three major dialect identification experiments, forming Data1 and Data2, respectively. Detailed information is provided in Table 10. All samples were uniformly divided into training, validation, and test sets using an 8:2:2 ratio.

Table 10: Experimental data set Information

data set	Label	Sample Size	Dialect	Data Source
Data1	Weizang	2250	Lhasa	TDSC-2024
	Anduo	2250	Xiahe	TDSC-2024
	Kang	2250	Changdu	TDSC-2024
Data2	Weizang	2150	Shigatse	TIBMD@MUC
	Anduo	2150	Xiahehua	XBMUAMDO31
	Kang	2900	Dege, etc.	TIBMD@MUC

4.2.2. Experimental Setup

In this study, MFCC coefficients were used as the primary features, while GFCC coefficients served as auxiliary features. For feature extraction, the frame length was set to 25 ms (corresponding to 400 sampling points per frame), and the frame shift was set to 10 ms (skipping 160 sampling points each time). Specifically, the MFCC features comprised 13 dimensions per frame, while the GFCC features comprised 40 dimensions per frame, resulting in a final fused feature dimension of 53. For the classification models, the LSTM network was configured with 64 LSTM units, 64 units in the fully connected (Dense) layer, a dropout rate of 0.2, and trained for 100 epochs. The KNN classifier employed a Radial Basis Function (RBF) as the kernel, with the regularization parameter C set to 1. To evaluate model performance, four metrics were used: Accuracy, Precision, Recall, and F1-score. Additionally, a confusion matrix was utilized to further analyze the experimental results.

4.2.3. Experimental Results and Analysis

In the KNN and LSTM Tibetan dialect identification models, the data set constructed in this study (Data1; TDSC-2024) achieved accuracy rates of 97.42% and 95.33%, respectively, surpassing Data2 (the public data set) by 6.12% and 12.27%. As both data sets were of comparable size, these results validate the superior effectiveness of the TDSC-2024 data set for Tibetan dialect identification tasks. Detailed results are presented in Table 11.

Table 11: Tibetan Dialect Identification Results on Different data sets

data set	Model	Accuracy	Precision	Recall	F1
Data2	KNN	91.30%	91.30%	91.30%	91.26%
	LSTM	83.06%	83.20%	83.06%	82.94%
Data1	KNN	97.42%	97.43%	97.42%	97.41%
	LSTM	95.33%	95.46%	95.33%	95.30%

To further assess the data quality of TDSC-2024, confusion matrices were generated for different data sets and recognition methods (as shown in Fig. 7, Fig. 8, Fig. 9, Fig. 10). The values in each matrix indicate the number of times the model predicted the row label dialect category as the column label dialect category, with the diagonal values representing the number of correct predictions. These results clearly demonstrate the model's predictive ability across various dialect categories. From the confusion matrices, it is evident that TDSC-2024 achieves higher overall recognition accuracy for the three major Tibetan dialect categories, particularly for complex and easily confused dialect pairs. In these cases, TDSC-2024 significantly outperforms the public data set. This suggests that the TDSC-2024 data set offers a distinct advantage in dialect

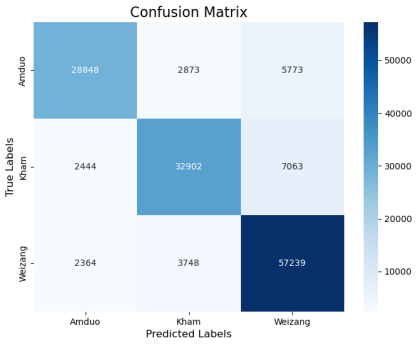


Figure 7: Confusion Matrix of Data2-LSTM Experiment

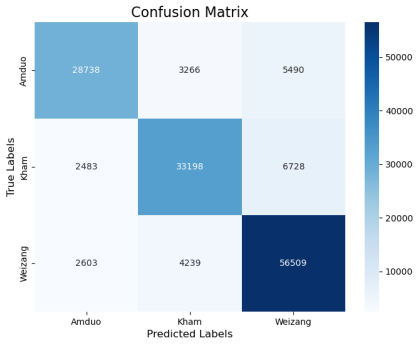


Figure 8: Confusion Matrix of Data1-LSTM Experiment

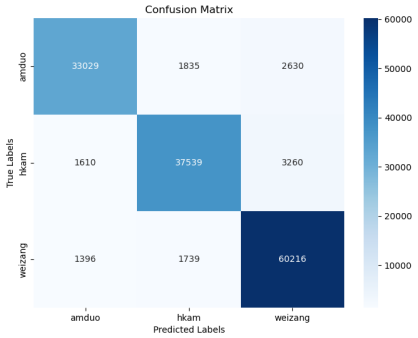


Figure 9: Confusion Matrix of Data2-KNN Experiment

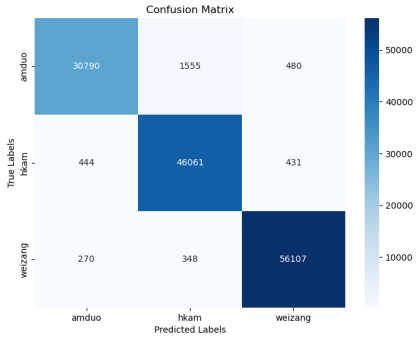


Figure 10: Confusion Matrix of Data1-KNN Experiment

differentiation, capturing phonetic, grammatical, and pronunciation differences more effectively. support for fine-grained language identification tasks involving Tibetan dialects, subdialects, and even local variations, areas where available data remain scarce.

5. Conclusion

This study constructs a high-quality speech data set, TDSC-2024, for automatic Tibetan dialect identification, grounded in Tibetan linguistic knowledge and speech signal processing techniques. Compared to existing Ti-betan dialect data sets, TDSC-2024 more accurately reflects the linguistic features of Tibetan dialects in terms of collection locations, text content, and data organization. To verify its effectiveness, KNN and LSTM models were employed for dialect recognition experiments. The results demonstrate that TDSC-2024 significantly outperforms commonly used public data sets at a comparable data scale, indicating superior dialect differentiation capability. Additionally, this study introduces the SPROD speech anomaly detection method, based on boxplot analysis, to address potential quality deviations during multi-speaker, multi-location data collection. This method effectively minimizes interference factors and enhances data reliability. For data augmentation, the GPT-SoVITS voice style conversion technique was applied,

expanding the data set's scale and diversity while reducing collection costs. With these innovations, TDSC-2024 not only supports conventional Tibetan dialect recognition tasks but also enables fine-grained identification of subdialects and local varieties. Its multi-level dialect labeling structure offers essential experimental support for dialect comparison, feature description, and related research fields.

References

[1] Schultz T, Vu N T, Schlippe T.: Glob-alphone: A multilingual text & speech database in 20 languages. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 8126-8130 (2013)

[2] Anguera X, Metze F, Buzo A, et al.: The spoken web search task (2013).

[3] Institute of Language Teaching, Beijing Language and Culture University: A large-scale language engineering project—"Beijing Oral Survey" passed expert evaluation. Language Teaching and Research. (03), 159 (1992).

[4] Li Yuming: On the construction of Chinese language resource audio databases. Chinese Language. (04), 356-363+384 (2010).

[5] H. Bu, J. Y. Du, X. Y. Na, et al.: AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline. In: 20th Conf. Oriental Chapter Int. Coordinating Comm. Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul (2017)

[6] Chen Xiaoying: Research on the design of speech corpora. Science and Technology Information. (36), 5-6 (2008).

[7] Gao Yuan, Gu Mingliang, Sun Ping, et al.: Design of a multi-purpose Chinese dialect speech database. Computer Engineering and Applications. 48(05), 118-120 (2012).

[8] Zou Faxin.: Design and implementation of a speech corpus. Guangxi Normal University (2012)

[9] Du Fuqiang.: Preliminary study on the construction of dialect databases. Ningbo University (2012)

[10] Bao Huaqiao, Xu Ang, Chen Jiayou: Lhasa Tibetan speech acoustic parameter database. Ethnic Languages. (05), 10-20+9 (1992).

[11] Yu Hongzhi, Li Yonghong, Suo Nanlengtse, et al.: Research on the Amdo Tibetan monosyllabic acoustic parameter database. In: Chinese Association for Chinese Information Processing Ethnic Language and Script Information Committee. Research on Ethnic Language and Script Information Technology—Proceedings of the 11th National Conference on Ethnic Language and Script Information. Northwest University for Nationalities, pp. 16-21 (2007)

[12] Kong Changqing, Guo Wu.: Non-specific Tibetan telephone speech database and recognition experiment. In: Chinese Association for Chinese Information Processing Speech Information Committee, Chinese Acoustical Society Language, Auditory, and Music Acoustics Branch, Chinese Linguistic Society Phonetics Committee. Proceedings of the 12th National Conference on Human-Machine Speech Communication (NCMMSC2013). Department of Electronic Engineering and Information Science, University of Science and Technology of China; National Engineering Laboratory for Speech and Language Information Processing, pp. 163-166 (2013)

[13] Lu Rongjiangcai, Wei Jianguo, Lu Wenhui, et al.: Establishment of a Tibetan-Chinese bilingual multimodal physiological speech database. In: Chinese Association for Chinese Information Processing Speech Information Committee. Proceedings of the 13th National Conference on Human-Machine Speech Communication (NCMMSC2015), pp. 578-580 (2015)

[14] Huang Xiaohui, Li Jing, Ma Rui: Design and research of Tibetan spoken speech corpus. Computer Engineering and Applications. 54(13), 231-235 (2018).

[15] Zhao, Y., Xu, X., Yue, J., Song, W., Li, X., Wu, L., & Ji, Q.: An open speech resource for Tibetan multi-dialect and multitask recognition. International Journal of Computational Science and Engineering. 22(2/3), 297-304 (2020).

[16] Li, S., Li, G., & Ning, J.: XBMU-AMDO31: An open source of Amdo Tibetan speech database and speech recognition baseline system. In: National Conference on Man-Machine Speech Communication, NCMMSC 2022 (2022)

[17] Davis, S. B., & Ermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP). 28(4), 357-366 (1980).

[18] Zhang Weiqiang, Liu Jia: Language recognition based on auditory perception features. Tsinghua University Journal (Natural Science Edition). 49(01), 78-81 (2009).

[19] T. Cover and P. Hart: Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 13 (1), 21-27 (1967).

[20] Hochreiter, S., & Schmidhuber, J.: Long short-term memory. Neural Computation. 9(8), 1735-1780 (1997).

[21] Gesang Jiumian, Gesang Yangjing.: Introduction to Tibetan Dialects. Ethnic Publishing House, (2002)

[22] Qu Aitang: On the tones of Si-no-Tibetan languages. Ethnic Languages. (06), 10-18 (1993).

[23] Institute of Linguistics, Chinese Academy of Social Sciences.: Atlas of Chinese Languages (2nd Edition) (Minority Languages Volume) (Fine). Commercial Press, (2012)

Appendix 1

Table 12: TDSC-2024 Data Collection Locations

No.	Dialect Category	Sub-dialect Category	Data Collection Location
1	U-Tsang	Central U-Tsang	Lhasa City, Chengguan District, Tibet Autonomous Region
2	U-Tsang	Central U-Tsang	Lhasa City, Chengguan District, Tibet Autonomous Region
3	U-Tsang	Central U-Tsang	Luoza County, Shannan City, Tibet Autonomous Region
4	U-Tsang	Central U-Tsang	Qusong County, Shannan City, Tibet Autonomous Region
5	U-Tsang	Lower U-Tsang	Nyalam County, Shigatse City, Tibet Autonomous Region
6	U-Tsang	Lower U-Tsang	Saga County, Shigatse City, Tibet Autonomous Region
7	U-Tsang	Lower U-Tsang	Jiangzi County, Shigatse City, Tibet Autonomous Region
8	U-Tsang	Ali	Geji County, Ali Region, Tibet Autonomous Region
9	U-Tsang	Ali	Gar County, Ali Region, Tibet Autonomous Region
10	Kham	Northern Route	Nangqian County, Yushu Tibetan Autonomous Prefecture, Qinghai Province
11	Kham	Northern Route	Dege County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province
12	Kham	Northern Route	Karuo District, Chamdo City, Tibet Autonomous Region
13	Kham	Southern Route	Batang County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province
14	Kham	Southern Route	Deqin County, Diqing Tibetan Autonomous Prefecture, Yunnan Province
15	Kham	Pastoral Area	Dingqing County, Chamdo City, Tibet Autonomous Region
16	Kham	Pastoral Area	Seini District, Nagqu City, Tibet Autonomous Region
17	Amdo	Pastoral Area	Aba County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
18	Amdo	Pastoral Area	Maqu County, Gannan Tibetan Autonomous Prefecture, Gansu Province
19	Amdo	Agricultural Area	Xiahe County, Gannan Tibetan Autonomous Prefecture, Gansu Province
20	Amdo	Agricultural Area	Xunhua County, Haidong City, Qinghai Province
21	Other	Jiarong	Ronga Township, Aba County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
22	Other	Zuo'ni	Niba Town, Zhuoni County, Gannan Tibetan Autonomous Prefecture, Gansu Province
23	Other	Diebu	Tiebu Town, Ruo'ergai County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
24	Other	Muya	Bamei Township, Daofu County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province

Response to Reviewer Comments

Dear Reviewer,

We sincerely appreciate your valuable feedback and constructive suggestions on our manuscript. We have carefully revised the paper according to your comments, with detailed modifications outlined below:

1. Data Sharing

As recommended, we have shared core sample datasets of major Tibetan dialects (including speech samples and annotations) via GitHub: https://github.com/gazangcairang3077/TDSC-2024-Small-scale-sample-.git

Note: As this dataset is essential for ongoing thesis research in our group, full public release is temporarily restricted. We confirm that all data will be progressively published through designated platforms (e.g., this GitHub repository or specialized data archives) upon thesis completion.

2. Speaker Enhancement Methodology

The speaker enhancement technique employs the open-source**GPT-SOVITS 4.0**voice conversion module (Project:https://github.com/RVC-BosS/GPT-SOVITS). Most enhanced samples were generated by this model. Given its public accessibility and space limitations, technical details were not expanded in the main text. Enhanced samples (speaker1–speaker10) are included in the shared dataset for verification.

3. Figure Formats

Figures 1, 2, and 6 have been replaced with high-resolution **vector formats (.eps/.svg)** as requested. Remaining figures (algorithmically/output-generated) could not be converted to vectors due to technical constraints. We ensured these retain sufficient clarity in high-DPI raster formats (.tif/.png) for print/digital readability. Should specific format requirements arise during production, we will promptly comply.

4. Literature Supplement

Two recent references ([26], [27]) have been added to relevant sections to strengthen contextual background.

5. SPROD Performance (Precision/F1)

We acknowledge your concerns about SPROD's precision/F1 scores. Analysis shows robust performance on majority-class samples but higher *missed-detection risk* for critical minority-class anomalies. To address this:

Targeted optimization: Enhance minority-class sample collection/annotation.

Threshold tuning: Explore sensitivity improvement (with false-alarm trade-off evaluation).

Fusion strategy: Integrate SPROD with specialized minority-class detection techniques.

Deployment protocol: Rigorous risk-tolerance assessment per application scenario. *High-risk deployments should adopt fusion strategies or sensitivity-tuned thresholds with false-alarm handling.*

6. Speaker Annotation Clarification

speaker0: Original recordings from collection sites.

speaker1–speaker10: GPT-SOVITS 4.0-enhanced samples.

speaker1–speaker5: Enhanced female voices

speaker6–speaker10: Enhanced male voices

Your expert review has significantly improved this work's rigor and clarity. We have incorporated all suggestions and welcome further feedback. Thank you for your time and invaluable contributions.

Sincerely,
GaZangCaiRang
2025.7.10

Research on the Construction of a High-Quality Speech data set for Tibetan Dialect Identification Tasks

Kalsang Tsering, Gao Dingguo, Xu Songtao

ABSTRACT

Research on Tibetan dialect identification is crucial for exploring the linguistic features of various dialects, advancing intelligent speech technology, and enriching linguistic theory. The construction of a high-quality speech data set forms the foundation of such research. This study integrates insights from Tibetan linguistics and speech signal processing to propose a method for building a data set specifically designed for Tibetan dialect identification and quantification research. First, based on the Tibetan phonemic system, example words and sentences were designed to include Tibetan vowels, consonants, and the "eight-case" grammatical structure. Second, to address the problem of low-quality speech data, an automatic detection method based on signal-to-noise ratio (SNR), pitch, speech rate, and other indicators was introduced to identify and handle abnormal samples. Third, to overcome the issue of limited pronunciation styles in some dialect regions, the GPT-Sovits 4.0 voice cloning model was employed to enrich pronunciation variations. As a result, a speech data set comprising 56,760 samples and approximately 39.8 hours of speech data was developed. Comparative experiments with publicly available data sets of similar scale demonstrate that the constructed data set improves Tibetan dialect identification accuracy by 6.12% (using the k-nearest neighbor algorithm) and 12.27% (using long short-term memory networks) compared to existing public data sets. These findings indicate that the newly constructed data set more effectively captures the phonetic distinctions among Tibetan dialects and enhances performance in dialect identification tasks.

Keywords: Tibetan dialect identification, data set construction, speech processing

1. Introduction

Tibetan is a language with multiple dialects, exhibiting significant differences in vocabulary, grammatical structure, and phonology. Among these, phonological variation is particularly pronounced, presenting challenges in areas such as data standardization for Tibetan speech processing and model training. The phonetic distinctions across dialects directly impact the applicability of Tibetan speech processing systems, causing a single system to perform poorly when handling multiple dialects, thereby affecting user experience. Therefore, the development of a high-quality Tibetan dialect automatic identification system holds not only great academic importance but also substantial practical value. As a typical data-driven classification task, the construction of a comprehensive speech data set encompassing multiple dialects is a fundamental prerequisite. Such a high-quality dialect speech data set can serve not only as the foundational resource for model training but also as the basis for

optimizing other dialect-related applications and improving system performance. Internationally, notable progress has been made in the development of dialect identification and language recognition data sets. In 2013, T. Schultz et al. [1] released the GlobalPhone data set, which includes samples from over 20 languages and dialects, suitable for multilingual recognition tasks. In the same year, Anguera X et al. [2] introduced the SWS2013 data set for language identification, comprising data from nine low-resource languages, including Albanian.

In China, dialect identification data sets have primarily focused on various Chinese dialects and regional languages. In 1992, Beijing Language and Culture University [3] extended the "Beijing Oral Survey" project initiated in the 1980s and established the "Beijing Oral Corpus Query System" (BJKY), which links text with audio to aid researchers in analyzing the phonetic, lexical, and grammatical features of the Beijing dialect. In 2008, the National Language Committee [4] launched the construction of the Chinese Language Resource Audio Database, which collected audio data of modern Chinese languages, including both dialects and minority languages. In 2017, the Beijing Aliyun AIShell team [5] released the AIShell-1 data set, primarily designed for speaker recognition but also incorporating various Chinese dialects and colloquial speech, making it applicable to dialect identification.

In addition to these data set construction efforts, Chinese scholars have conducted extensive research on methods for building dialect identification databases. In 2008, Chen Xiaoying [6] summarized the processes of corpus design, recording, annotation, and management, proposing a comprehensive construction methodology. In 2011, Gao Yuan et al. [7] introduced design standards for dialect databases. In 2012, Zou Faxin [8] outlined a workflow for speech recording and management systems. In 2013, Du Fuqiang [9] systematically examined the theoretical foundations and implementation methods of dialect database construction, covering data collection and annotation processes.

Currently, there is no specialized Tibetan dialect identification speech data set available in academia. However, early efforts have been made by scholars to construct Tibetan dialect databases. In 1992, Bao Huaqiao et al. [10] developed an acoustic parameter database for the Lhasa Tibetan dialect. In 2007, Yu Hongzhi et al. [11] built an acoustic parameter database for the Amdo Tibetan dialect. In 2013, Kong Changqing et al. [12] constructed a large-scale continuous Tibetan telephone speech recognition database targeting the dialect of the Tibet Autonomous Region. In 2015, Lu Rongjiang et al. [13] created a Tibetan-Chinese bilingual multimodal physiological speech database based on the Tibetan dialect. In 2018, Huang Xiaohui et al. [14] established a Tibetan oral speech corpus based on the Lhasa Tibetan dialect, covering aspects such as corpus selection, recording standards, and annotation guidelines. However, most of these data

sets are limited to a single dialect, making them unsuitable for dialect identification tasks, and they are not publicly accessible. Additionally, existing resources used in Tibetan dialect research, such as the Tibetan-Chinese Lhasa Oral Dictionary, Amdo Tibetan Oral Dictionary, and Tibetan-Burmese Phonology and Vocabulary, do not meet the requirements for speech recognition tasks. Public Tibetan speech data sets like TIBMD@MUC [15] and XBMU-AMDO31 [16] include multiple dialects but are primarily designed for Tibetan speech recognition and synthesis tasks. Although they can process recognition tasks involving the three main Tibetan dialects (Central Tibetan, Amdo, and Kham), they perform poorly in fine-grained dialect identification due to limited phoneme coverage and incomplete dialect representation. Moreover, constructing a Tibetan dialect identification speech data set presents several challenges: Tibetan dialects are highly diverse, with considerable phonological differences, making it difficult to collect comprehensive and representative data. Furthermore, speech data cleaning and classification are both time-consuming and costly.

To address these challenges, this paper proposes the design of a high-quality speech data set specifically for Tibetan dialect identification, resolving issues such as incomplete dialect representation and insufficient phoneme coverage found in existing public data sets. This data set holds significant value for advancing future research in Tibetan dialect identification. The construction process is outlined as follows: 1) Data Collection: Based on linguistic research on Tibetan dialects and the Tibetan phonological system, a comprehensive data collection plan (referred to hereafter as "the plan") was formulated. According to this plan, speech data were gathered from 24 dialect regions, yielding 5,160 speech samples (2,760 words and 2,400 sentences). 2) Abnormal Sample Detection: To ensure data quality, an abnormal sample detection method based on key indicators such as signal-to-noise ratio, speech rate, and pitch was developed using box plot analysis. A total of 384 low-quality samples were identified and either removed or re-recorded. 3) Data Augmentation: To mitigate the limited number of speakers in certain dialect regions, the GPT-Sovits 4.0 voice cloning model was employed to enhance voice style variation for each speaker's data. As a result, the original data set was expanded to include 56,760 speech samples, forming the Tibetan Dialect Speech Corpus 2024 (TDSC-2024), encompassing approximately 39.8 hours of speech data. The construction process is illustrated in **Figure 1.** Construction Process of a High-Quality Speech data set for Tibetan Dialect Identification.

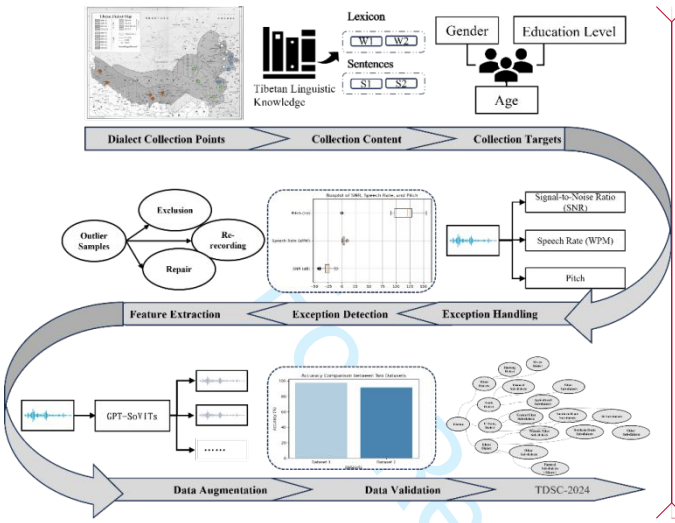


Figure 1. Construction Process of a High-Quality Speech data set for Tibetan Dialect Identification

Commented [e1]: 已按意见更换为矢量图

Subsequently, to evaluate the dialect discrimination capability of TDSC-2024 compared to other publicly available, non-specialized Tibetan dialect speech data sets, automatic dialect identification experiments were conducted for the three major Tibetan dialects using both public data sets (TIBMD@MUC, XBMU-AMDO) and TDSC-2024. Mel Frequency Cepstral Coefficients (MFCC) [17] and Gamma Frequency Cepstral Coefficients (GFCC) [18] were combined as input features, and experiments were performed using both K-Nearest Neighbors (KNN) [19] and Long Short-Term Memory (LSTM) [20] models. When controlling for comparable data set sizes, TDSC-2024 achieved accuracy improvements of 6.12% (KNN) and 12.27% (LSTM) over the publicly available data sets. These results demonstrate the significant advantage of TDSC-2024 in Tibetan dialect identification tasks, as it more effectively captures and represents phonetic differences among various dialects.

2. Methodology for Constructing the Tibetan Dialect Identification Speech data set

As described above, the construction method consists of three parts: Tibetan dialect speech collection, abnormal sample detection and processing, and GPT-SoVITS-based speech style transformation.

2.1 Tibetan Dialect Speech Collection

2.1.1 Tibetan Dialect Distribution and Collection Point Selection

There are differing opinions within the traditional linguistics community regarding the classification of Tibetan dialects and the selection of representative local varieties. The views of scholars such as Gesang Junmian [21] and Qu Aitang [22] are widely regarded as authoritative and objective, aligning closely with the actual linguistic characteristics of Tibetan. Based on the traditional division into three main Tibetan dialects, Gesang Junmian proposed a further subdivision into seven sub-dialects, a classification that has gained broad acceptance in the academic community. This study primarily adopted this classification method when selecting dialect collection points. Additionally, relevant content from the Atlas of Chinese Languages: Volume 2, Minority Languages [23] was consulted to subdivide the Central Tibetan dialect into the Front Tibetan, Rear Tibetan, and Ali sub-dialects. Consequently, eight Tibetan sub-dialects were established in total.

The collection points were designed to cover representative local dialects of these eight sub-dialects, along with other Tibetan dialects possessing unique features, such as the Jiayong, Zhuoni, Diebu, and Moya dialects. Specific collection points and their corresponding dialect classifications are detailed in Appendix 1. Each collection point is categorized into two hierarchical levels: dialect regions and sub-dialect regions. This multi-level data structure provides flexible support for tasks requiring different levels of dialect identification.

2.1.2 Corpus Design

The content collected for TDSC-2024 primarily follows the framework outlined in the Survey Handbook of Chinese Language Resources Audio Database: Tibetan-Burmese Language Family. The design includes two main linguistic elements: vocabulary items and grammatical example sentences. For the vocabulary component, a Swadesh core vocabulary list served as the foundation, supplemented with knowledge from Tibetan phonology and commonly used Tibetan words to form the TDSC vocabulary list. This list covers all Tibetan consonants and vowels in their written form and aims to collect speech data on the most frequently used words across various Tibetan dialects. For the example sentences, the design was informed by the handbook and integrated with the Tibetan “eight-case” grammatical structure. Emphasis was placed on capturing the phonetic and prosodic features of different dialects, including intonation, tonal variation,

and sentence cohesion. This approach was intended to better reflect the pronunciation and prosodic characteristics distinctive to each dialect, as shown in

Table 1 .

The collected material contributed to the development of the Tibetan Dialect Automatic Identification Recording Manual (referred to as the TDSC-2024 Manual). Tibetan was used as the primary language for the collected content, accompanied by Chinese translations to help speakers fully understand the semantic meaning before recording their utterances. This method minimized the risk of unnatural or overly formal reading that can occur when only Tibetan content is provided, ensuring that the collected samples more accurately reflected the natural pronunciation and lexical habits of each dialect.

Table 1 TDSC-2024 Content Structure

Linguistic Granularity	Quantity	Design Purpose
Words	115	To analyze the linguistic features of words
Sentences	100	To analyze the prosodic features of sentences

2.1.3 Speaker Selection

For the data collection points (Appendix 1), local residents were selected as speakers. Each speaker recorded the designated collection texts (

Table 1). Most participants had low levels of formal education and had resided in the local area for extended periods, ensuring that their speech accurately represented the authentic features of the local dialect in everyday use.To comprehensively capture pronunciation characteristics across different age groups and genders, a systematic design was applied to the gender and age distribution of the speakers. Speech data were collected from a total of 24 speakers, with their gender and age distribution detailed in

Table 2 and **Table 3**.

Table 2 Gender Composition

Gender	Male	Female
Quantity (People)	8	16
Proportion (%)	33.33	66.67

Table 3 Age Composition

Age Group	Above 18 Years	19-49 Years	50 Years and Above
Quantity (Peopl	4	12	8
Proportion (%)	16.67	50	33.33

2.1.4 Labeling Structure

Speech data from 24 speakers were collected across 24 dialect collection points. After editing and organizing the recordings, a Tibetan dialect speech database comprising nearly 258 minutes and 5,160 entries was constructed. To facilitate subsequent data processing and recognition studies, a systematic four-level data labeling structure was designed. This structure enables multi-level management of the speech data, making the corpus more efficient and flexible for various applications.

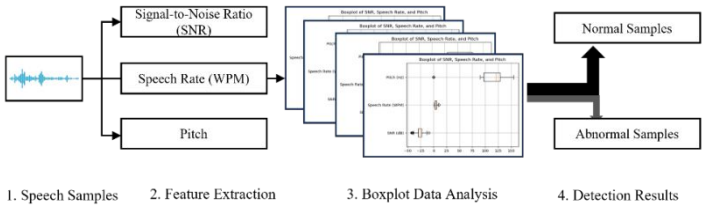
The multi-level labeling system consists of four layers of information: dialect region, sub-dialect region, dialect collection point, and content identifier. For example, the word "ཁ་ཟུག་མ" (meaning "shadow") collected in Maqu County is labeled as "Amduo.Amduomuqu.MaQuXian.W20." The specific labeling rules are as follows: "Amduo" (Amdo Dialect) – First-level label: Indicates the general dialect category to which the sample belongs. "Amduomuqu" (Pastoral Sub-dialect) – Second-level label: Specifies the sub-dialect classification of the sample, providing a finer distinction within the broader dialect category. "MaQuxian" (County-level administrative region) – Third-level label: Denotes the specific dialect collection point, offering detailed location information. This level can be further refined in future data expansion efforts to increase the precision of collection site identification. "W20" (Collection content) – Fourth-level label: Identifies the specific recorded content. In this example, "ཁ་ཟུག་མ" (meaning "shadow") corresponds to the content identifier W20 in the TDSC-2024 manual. "W" denotes words, while example sentences are labeled as "S××."
A specific example of the labeling structure is shown in

Table 4 .

Table 4 Specific Label Examples

Speech Phrase	Label	ID
ཁ་ཟུག་མ (meaning "shadow")	Amduo.Amduomuqu.MaQu xian.W20	W20
ཁཁྱེད་ཀྱི་མི་གྲངས་ཇི་ཙམ་ཡོད། (meaning "How many people are there in your family?")	Amduo.Amduomuqu.MaQu xian.S13	S13

2.2 Anomaly Detection and Processing



Commented [e2]: 已按意见改为矢量图。

Figure 2. SPROD Anomaly Speech Sample Detection Process

The 24 dialect collection points included in TDSC-2024 are distributed across five provincial-level administrative regions. Among these, Qinghai Haidong and Tibet Ali are separated by nearly 3,000 kilometers, making field data collection extremely costly. Consequently, some collection points were outsourced to other locations for recording. While this approach helped reduce collection costs, it also introduced uncontrollable factors that could affect the quality of the dialect samples. To address the challenge of automatically detecting low-quality samples arising from these variations, this study proposes a boxplot-based anomaly detection method, SNR-Pitch-Words Per Minute Boxplot Outlier Detection (abbreviated as SPROD). This method evaluates three key indicators: signal-to-noise ratio (SNR), pitch, and speech rate. The specific workflow of this method is illustrated in **Figure 2**.

2.2.1 Key Features in the SPROD Method

The SNR represents the ratio of signal power to noise power in an audio sample and serves as a key indicator of audio clarity. Signal and noise components can typically be distinguished using Power Spectral Density (PSD) analysis. Specifically, noise segments generally exhibit lower frequencies and smaller PSD values, whereas signal segments display relatively higher frequencies and larger PSD values. The PSD is calculated as follows:

$$S(f) = |X(f)|^2 / T \qquad (1)$$

$S(f)$ represents the power spectral density of the signal at frequency f, $X(f)$ is the Fourier transform of the signal at frequency f, and T is the time duration of the signal. Based on the average PSD values within the signal and noise frequency bands, the signal and noise power are calculated, and the SNR is determined as follows:

$$SNR = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) \quad (2)$$

P_{signal} represents the signal power, and P_{noise} denotes the noise power. Experimental results indicate that defining frequencies above 2000 Hz as the signal band and those below 500 Hz as the noise band yields higher accuracy in anomaly sample detection.

Words Per Minute (WPM) refers to the number of words or characters spoken within a given unit of time and is commonly used as a metric for speech fluency. Accurate calculation of speech rate typically relies on speech recognition technology supported by linguistic models to determine the actual word count, which in turn requires a high-precision speech recognition module. However, due to the lack of mature recognition technology for the dialect data used in this study, such an approach is difficult to implement. Therefore, a speech rate estimation method based on the Zero Crossing Rate (ZCR) is employed instead. This method estimates speech rate by analyzing short-term variations in the zero-crossing rate of the speech signal, thereby avoiding dependence on a complex speech recognition system and enabling effective estimation of speech rate even in the absence of high-precision recognition technology.

Generally, as speech rate increases, the ZCR value also rises because faster speech contains more high-frequency components. These lead to more frequent changes in the signal's sign, resulting in a higher zero-crossing rate. This relationship was validated through empirical sample comparisons. As shown in **Figure 3**, the ZCR values of the Tibetan phrase "ཟླ་ཤི་དེ་ཚེ་བཀྲ་ཤིས་དོན་" (meaning "Zashi and Tshering") at different speech rates exhibit this pattern: faster speech produces denser audio fluctuations and correspondingly higher ZCR values (see **Table 5**).

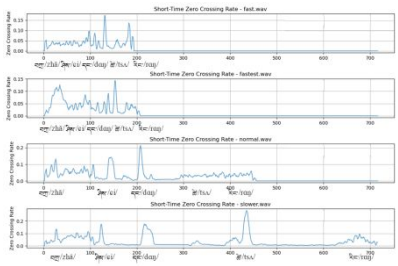


Figure 3. Comparison of Audio Waveforms at Different Speed.

Table 5 ZCR Values at Different Speeds (ms)

Speech Rate	ZCR Value
Slower	0.0338
Normal	0.0370
Fast	0.0394
Fastest	0.0438

Pitch can be estimated by determining the fundamental frequency of the audio signal, measured in Hertz (Hz). The calculation method is as follows:

$$f_{\text{pitch}} = \frac{1}{N}\sum_{i=1}^{N} f_0(i) \quad (3)$$

f_{pitch} represents the average pitch of the audio signal; $f(i)$ is the fundamental frequency at the i-th time instance; N is the number of voiced segments, representing the total number of valid fundamental frequency values.

The SPROD method analyzes three key features of audio samples (i.e., SNR, WPM, and Pitch) to efficiently detect

low-quality anomalous samples in the data set, thereby enhancing the overall quality of the speech data. SNR, speech rate, and pitch are fundamental parameters of audio signals. In this study, Boxplot analysis combined with the Interquartile Range (IQR) statistical method is employed to automatically identify anomalous data within the samples. In practical application, when the SPROD method was applied to a data set from a specific speaker, it effectively and intuitively identified anomalous samples. The effectiveness and advantages of this approach are further validated through experimental comparisons presented in the following sections.

2.2.2 Experiment and Analysis

To verify the practical effectiveness of the SPROD method, a series of comparative experiments were conducted. These experiments were based on the assumption that speech rate, pitch, and other relevant indicators for the same speaker should remain relatively consistent across different recording samples. Accordingly, recordings from the same speaker were selected as the detection targets. The experiments included four different anomaly detection approaches: 1) Detection using the Box Plot method (i.e., the SPROD method proposed in this study); 2) Detection based on the Z-score method; 3) Combined detection by taking the union of results from the Box Plot and Z-score methods; 4) Combined detection by taking the intersection of results from the Box Plot and Z-score methods.

2.2.3 Experimental Corpus

The experimental corpus comprised samples from four speakers collected during the preliminary stage of this study, totaling 2,624 speech samples. All samples were manually inspected and individually labeled, with the labeling format presented in

Table 6 .

Table 6 Speech Sample Labels

File	Ture_label
sample (1).wav	1
sample (2).wav	0
sample (3).wav	0

In this study, a True_label value of 1 indicates that the sample was manually classified as an anomalous sample. Reasons for this classification include excessively fast speech rate, abnormal pitch, or a low signal-to-noise ratio. Conversely, a True_label value of 0 indicates that the sample was deemed normal. After manual inspection, a total of 149 low-quality speech samples, accounting for 5.67% of all samples, were found not to meet the required standards.

2.2.4 Experimental Setup

The experimental environment consisted of a system equipped with 16 GB of RAM, an Intel i7 13th-generation processor, and an NVIDIA 3050 GPU. Python 3.12 was used as the programming language for all experiments.

3. Comparative Experiments and Analysis

This study presents the performance of the four detection methods using visual charts and statistical data, with the results summarized in **Table 8**. To enable a more detailed comparison, the baseline detection method was tested under eight different threshold combinations to assess its detection performance across various threshold settings. These threshold combinations are listed in

Table 7 .

Table 7 Threshold Combinations

Threshold name	SNR	SPEED_LOW	SPEED_HIGH	F0_MIN	F0_MAX
1	5	120	200	80	300
2	10	100	180	70	250
3	8	130	220	90	310
4	15	110	190	85	280
5	12	125	205	75	270
6	7	115	185	80	125
7	9	135	215	85	300
8	6	140	220	90	310

Table 8 Detection Performance of Different Thresholds Based on Z-Score

Threshold name	Accuracy	Precision	Recall	F1 Score
1	0.88	0.10	0.12	0.14
2	0.86	0.17	0.11	0.13
3	0.82	0.12	0.14	0.13
4	0.87	0.14	0.15	0.14
5	0.85	0.12	0.12	0.12
6	0.86	0.15	0.08	0.10
7	0.87	0.14	0.10	0.12
8	0.90	0.19	0.11	0.15

Table 9 Effectiveness of Different Detection Methods Based on Threshold 8

Method	Accuracy	Precision	Recall	F1 Score
Z-score	0.90	0.19	0.11	0.15
Z-score & Box Plot	0.90	0.15	0.13	0.14
Z-score | Box Plot	0.89	0.14	0.12	0.13
SPROD	0.92	0.22	0.12	0.16

Based on the experimental results, the Z-score-based detection method achieved its best performance at Threshold 8, standing out among all threshold combinations. The detailed results are presented in **Table 8** and **Table 9**. At this threshold, when compared with the four detection methods, including SPROD, it was observed that the SPROD method outperformed the others, achieving the highest detection accuracy, precision, and F1 score, 92%, 22%, and 16%, respectively. Additionally, the recall rate reached 12%, ranking second among all methods, thus demonstrating the comprehensive performance advantage of the SPROD method. However, the experiments also revealed that despite SPROD's excellent accuracy, its precision, recall, and F1 score remained relatively low. Upon further analysis, this was attributed primarily to the severe class

imbalance in the data set: the majority of samples belonged to the negative class (normal samples), while only 5.67% of the total were positive class samples (anomalous samples). This imbalance led to a higher occurrence of false positives and false negatives when the model processed positive class samples, thereby affecting precision, recall, and F1 score metrics. Nonetheless, the overall accuracy remained high, confirming the method's effectiveness for large-scale voice sample anomaly detection tasks.

Moreover, these results indirectly validate the effectiveness of the three key indicators (i.e., SNR, speech rate, and pitch) used in the SPROD method for automatic detection and recognition of abnormal voice samples. In addition to the quantitative performance, the superior anomaly detection capability of the methods is visually demonstrated by the data clustering results shown in **Figure 4** and **Figure 5** (where blue represents normal samples and red represents anomalous samples). The Z-score-based detection method showed limited effectiveness in identifying anomalous samples, whereas the Box Plot-based SPROD method more accurately detected outlier samples that deviated significantly from the normal range.

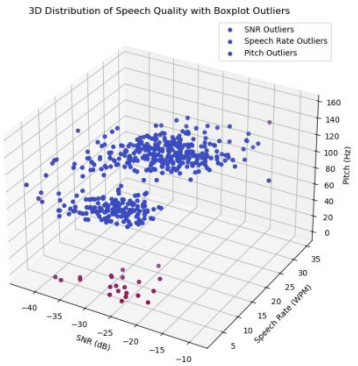


Figure 4. Detection Performance of the SPROD Method

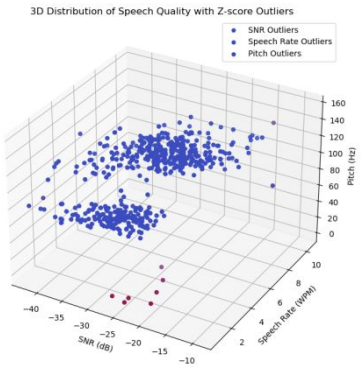


Figure 5. Detection Performance of the Z-Score Method

4. Application of the SPROD Method

The SPROD method was applied to detect outliers within the 5,160 collected speech data samples, comprising 2,760-word samples and 2,400 sentence samples. The analysis was conducted on a speaker-by-speaker basis, resulting in the identification of 420 low-quality samples. Following manual verification, 384 of these samples were confirmed as invalid and subsequently discarded. To meet the actual data requirements, these discarded samples were re-collected and supplemented accordingly.

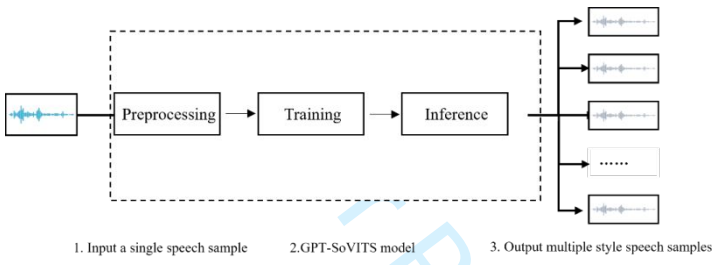
4.1 Speech Style Conversion Based on GPT-SoVITS

GPT-SoVITS is a speech cloning and style transfer technology that integrates the GPT model with the SoVITS (Speech Variations and Implicit Text-to-Speech) model. This method enables the generation of speech data from various speakers using only a small amount of original speech while incorporating diverse speech style attributes such as gender, age, and accent. The core principle leverages the GPT model's language understanding capabilities alongside the SoVITS model's proficiency in capturing and processing speech characteristics. This allows for the simulation of distinct speaker features while preserving speech quality.

By applying this technology, a wide variety of speech styles can be produced from limited data, thereby improving the robustness of the speech synthesis system across different speech environments.

To reduce data collection costs, TDSC-2024 selected only one speaker per dialect region, resulting in a relatively uniform speaker structure at each collection point. However, since speech characteristics vary significantly across age groups and genders,

the GPT-SoVITS speech cloning model was employed to transform each single-speaker data set into multiple speaker styles. This approach not only expanded the data set but also enhanced the model's ability to generalize across different speaker tones. The specific process is illustrated in **Figure 6**.



Commented [e3]: 已按要求改为矢量图

Figure 6. Speech Style Transfer Method Based on GPT-SoVITS

In the field of speech conversion and synthesis, the Mean Opinion Score (MOS) is widely used to evaluate speech quality. Evaluators assess the speech samples based on factors such as clarity and naturalness. MOS ratings categorize speech quality into five levels, with a score of 5 representing excellent and highly natural speech, and a score of 1 indicating poor and unacceptable quality. The MOS is calculated as follows:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} S_i \qquad (4)$$

where N is the number of evaluators and x_i is the score given by the i-th evaluator. In this study, 20 randomly selected speech samples generated by GPT-SoVITS were subjected to subjective MOS evaluation, with N = 15. The resulting MOS score was 3.89, indicating that the generated speech was of good quality, relatively clear and natural, and generally acceptable to most evaluators. However, some issues such as background noise and a lack of naturalness were noted, with certain details requiring further refinement. Despite these shortcomings, the generated speech is suitable for use in training tasks for the subsequent dialect identification model.

In this study, GPT-SoVITS 4.0 and its publicly available 10-speaker models were employed to perform a 1×10 pitch conversion enhancement on the speech data of each original speaker. This process expanded the original data set to 56,760 speech samples, resulting in a Tibetan dialect identification speech data set (TDSC-2024) comprising nearly 39.8 hours of audio data.

4.2. Validation of TDSC-2024

To evaluate the performance of the TDSC-2024 data set in Tibetan dialect identification tasks, this study conducted comparative experiments using TDSC-2024 and other publicly available data sets for the automatic identification of the three major Tibetan dialects. In these experiments, MFCC and GFCC were combined as input features, while KNN and LSTM models were employed as classification algorithms.

4.2.1 Experimental Data

In China, the Tibetan language is generally classified into three major dialects: Ü-Tsang, Amdo, and Kham. Specifically, representative local varieties of the Ü-Tsang dialect include the Lhasa, Shigatse, and Shannan dialects; those of the Amdo dialect include the Gansu Xiahe and Qinghai Zeku dialects; and those of the Kham dialect include the Sichuan Dege, Qinghai Yushu, and Tibet Changdu dialects. The TDSC-2024 corpus further subdivides these three major dialects into eight subdialects: front Ü-Tsang, rear Ü-Tsang, Ali, Amdo pastoral, Amdo agricultural, southern Kham, northern Kham, and Kham pastoral subdialects. Additionally, the corpus includes distinctive local varieties such as Jiayang, Zhuoni, Diebu, and Muya. In contrast, publicly available data sets such as TIBMD@MUC and XBMU-AMDO31 classify Tibetan solely into the three major dialect categories without further subdivision.
To ensure a fair and noticeable comparison, this study extracted an equal number of speech samples from both the public data sets and the TDSC-2024 data set for the three major dialect identification experiments, forming Data1 and Data2, respectively. Detailed information is provided in

Table 10 . All samples were uniformly divided into training, validation, and test sets using an 8:2:2 ratio.

Table 10 Experimental data set Information

data set	Label	Sample Size	Dialect	Data Source
Data1	Weizang	2250	Lhasa	TDSC-2024
	Anduo	2250	Xiahe	TDSC-2024
	Kang	2250	Changdu	TDSC-2024
Data2	Weizang	2150	Shigatse	TIBMD@MUC
	Anduo	2150	Xiahehua	XBMUAMDO31
	Kang	2900	Dege, etc.	TIBMD@MUC

4.2.2 Experimental Setup

In this study, MFCC coefficients were used as the primary features, while GFCC coefficients served as auxiliary features. For feature extraction, the frame length was set to 25 ms (corresponding to 400 sampling points per frame), and the frame shift was set to 10 ms (skipping 160 sampling points each time). Specifically, the MFCC features comprised 13 dimensions per frame, while the GFCC features comprised 40 dimensions per frame, resulting in a final fused feature dimension of 53. For the classification models, the LSTM network was configured with 64 LSTM units, 64 units in the fully connected (Dense) layer, a dropout rate of 0.2, and trained for 100 epochs. The KNN classifier employed a Radial Basis Function (RBF) as the kernel, with the regularization parameter C set to 1. To evaluate model performance, four metrics were used: Accuracy, Precision, Recall, and F1-score. Additionally, a confusion matrix was utilized to further analyze the experimental results.

4.2.3 Experimental Results and Analysis

In the KNN and LSTM Tibetan dialect identification models, the data set constructed in this study (Data1; TDSC-2024) achieved accuracy rates of 97.42% and 95.33%, respectively, surpassing Data2 (the public data set) by 6.12% and 12.27%. As both data sets were of comparable size, these results validate the superior effectiveness of the TDSC-2024 data set for Tibetan dialect identification tasks. Detailed results are presented in

Table 11 .

Table 11 Tibetan Dialect Identification Results on Different data sets

data set	Model	Accuracy	Precision	Recall	F1
Data2	KNN	91.30%	91.30%	91.30%	91.26%
	LSTM	83.06%	83.20%	83.06%	82.94%
Data1	KNN	97.42%	97.43%	97.42%	97.41%
	LSTM	95.33%	95.46%	95.33%	95.30%

To further assess the data quality of TDSC-2024, confusion matrices were generated for different data sets and recognition methods (as shown in **Figure 7**, **Figure 8**, **Figure 9**, **Figure 10**). The values in each matrix indicate the number of times the model predicted the row label dialect category as the column label dialect

category, with the diagonal values representing the number of correct predictions. These results clearly demonstrate the model's predictive ability across various dialect categories. From the confusion matrices, it is evident that TDSC-2024 achieves higher overall recognition accuracy for the three major Tibetan dialect categories, particularly for complex and easily confused dialect pairs. In these cases, TDSC-2024 significantly outperforms the public data set. This suggests that the TDSC-2024 data set offers a distinct advantage in dialect differentiation, capturing phonetic, grammatical, and pronunciation differences more effectively. support for fine-grained language identification tasks involving Tibetan dialects, subdialects, and even local variations, areas where available data remain scarce.

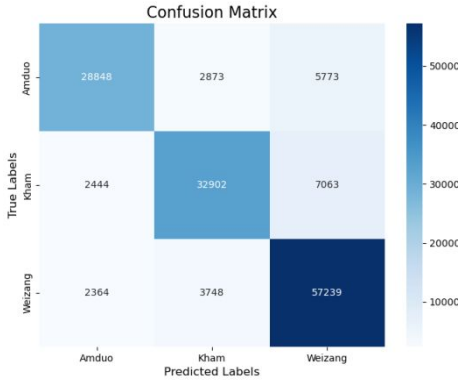


Figure 7. Confusion Matrix of Data2-LSTM Experiment

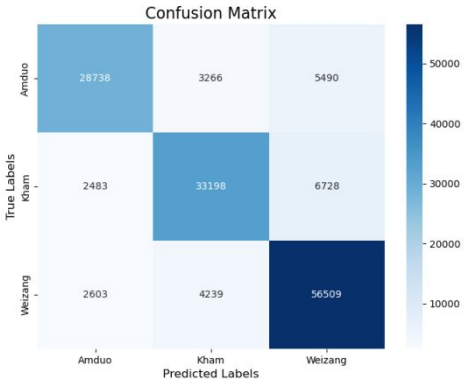


Figure 8. Confusion Matrix of Data1-LSTM Experiment

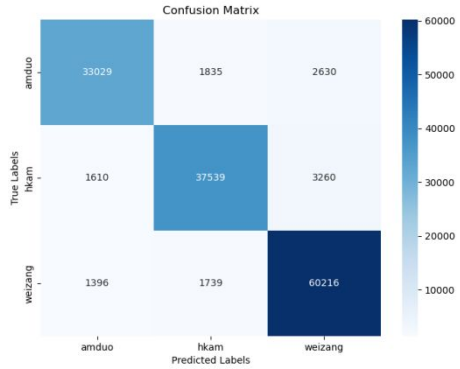


Figure 9. Confusion Matrix of Data2-KNN Experiment

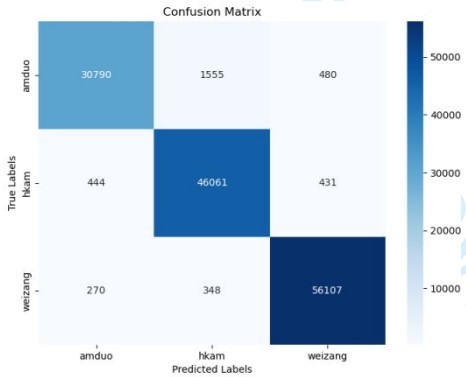


Figure 10. Confusion Matrix of Data1-KNN Experiment

5. Conclusion

This study constructs a high-quality speech data set, TDSC-2024, for automatic Tibetan dialect identification, grounded in Tibetan linguistic knowledge and speech signal processing techniques. Compared to existing Tibetan dialect data sets, TDSC-2024 more accurately reflects the linguistic features of Tibetan dialects in terms of collection locations, text content, and data organization. To verify its effectiveness, KNN and LSTM models were employed for dialect recognition experiments. The results demonstrate that TDSC-2024 significantly outperforms commonly used public data sets at a comparable data scale, indicating superior dialect differentiation capability. Additionally, this study introduces the SPROD speech anomaly detection method, based on boxplot analysis, to address potential quality deviations during multi-speaker,

multi-location data collection. This method effectively minimizes interference factors and enhances data reliability. For data augmentation, the GPT-SoVITS voice style conversion technique was applied, expanding the data set's scale and diversity while reducing collection costs. With these innovations, TDSC-2024 not only supports conventional Tibetan dialect recognition tasks but also enables fine-grained identification of subdialects and local varieties. Its multi-level dialect labeling structure offers essential experimental support for dialect comparison, feature description, and related research fields.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Schultz T, Vu N T, Schlippe T. Globalphone: A multilingual text & speech database in 20 languages [C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 8126-8130.
[2] Anguera X, Metze F, Buzo A, et al. The spoken web search task [J]. 2013.
[3] A large-scale language engineering project—"Beijing Oral Survey" passed expert evaluation [J]. Language Teaching and Research, 1992, (03): 159.
[4] Li Yuming. On the construction of Chinese language resource audio databases [J]. Chinese Language, 2010, (04): 356-363+384.
[5] H. Bu, J. Y. Du, X. Y. Na, et al., "AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline," Proc. 20th Conf. Oriental Chapter Int. Coordinating Comm. Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, 2017.
[6] Chen Xiaoying. Research on the design of speech corpora [J]. Science and Technology Information, 2008, (36): 5-6.
[7] Gao Yuan, Gu Mingliang, Sun Ping, et al. Design of a multi-purpose Chinese dialect speech database [J]. Computer Engineering and Applications, 2012, 48(05): 118-120.
[8] Zou Faxin. Design and implementation of a speech corpus [D]. Guangxi Normal University, 2012.
[9] Du Fuqiang. Preliminary study on the construction of dialect databases [D]. Ningbo University, 2012.
[10] Bao Huaqiao, Xu Ang, Chen Jiayou. Lhasa Tibetan speech acoustic parameter database [J]. Ethnic Languages, 1992, (05): 10-20+9.
[11] Yu Hongzhi, Li Yonghong, Suo Nanlengtse, et al. Research on the Amdo Tibetan monosyllabic acoustic parameter database [C]//Chinese Association for Chinese Information Processing Ethnic Language and Script Information Committee. Research on Ethnic Language and Script Information Technology—Proceedings of the 11th National Conference on Ethnic Language and Script Information. Northwest University for

Nationalities, 2007: 16-21.
[12] Kong Changqing, Guo Wu. Non-specific Tibetan telephone speech database and recognition experiment [C]//Chinese Association for Chinese Information Processing Speech Information Committee, Chinese Acoustical Society Language, Auditory, and Music Acoustics Branch, Chinese Linguistic Society Phonetics Committee. Proceedings of the 12th National Conference on Human-Machine Speech Communication (NCMMSC2013). Department of Electronic Engineering and Information Science, University of Science and Technology of China; National Engineering Laboratory for Speech and Language Information Processing, 2013: 163-166.
[13] Lu Rongjiangcai, Wei Jianguo, Lu Wenhui, et al. Establishment of a Tibetan-Chinese bilingual multimodal physiological speech database [C]//Chinese Association for Chinese Information Processing Speech Information Committee. Proceedings of the 13th National Conference on Human-Machine Speech Communication (NCMMSC2015), 2015: 578-580.
[14] Huang Xiaohui, Li Jing, Ma Rui. Design and research of Tibetan spoken speech corpus [J]. Computer Engineering and Applications, 2018, 54(13): 231-235.
[15] Zhao, Y., Xu, X., Yue, J., Song, W., Li, X., Wu, L., & Ji, Q. (2020). An open speech resource for Tibetan multi-dialect and multitask recognition. International Journal of Computational Science and Engineering, 22(2/3), 297-304.
[16] Li, S., Li, G., & Ning, J. (2022). XBMU-AMDO31: An open source of Amdo Tibetan speech database and speech recognition baseline system. In Proceedings of the National Conference on Man-Machine Speech Communication (NCMMSC 2022).
[17] Davis, S. B., & Ermelstein, P. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP), 28(4), 357-366.
[18] Zhang Weiqiang, Liu Jia. Language recognition based on auditory perception features [J]. Tsinghua University Journal (Natural Science Edition), 2009, 49(01): 78-81.
[19] T. Cover and P. Hart, "Nearest neighbor pattern classification," in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967.
[20] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
[21] Gesang Jiumian, Gesang Yangjing. Introduction to Tibetan Dialects [M]. Ethnic Publishing House, 2002.
[22] Qu Aitang. On the tones of Sino-Tibetan languages [J]. Ethnic Languages, 1993, (06): 10-18.
[23] Institute of Linguistics, Chinese Academy of Social Sciences. Atlas of Chinese Languages (2nd Edition) (Minority Languages Volume) (Fine) [M]. Commercial Press, 2012.
[24] Tukey, J. W. Exploratory Data Analysis. Addison-Wesley Publishing Company, 1977.
[25] https://github.com/RVC-Boss/GPT-SoVITS
[26] Honcharenko L K ,Bohuta H, Ivaniush A , et al.A Dataset of Real and Synthetic Speech in Ukrainian.[J].Scientific data,2025,12(1):745.
[27] Kong S ,Li C ,Fang C , et al. Building a Speech Dataset and Recognition Model for the Minority Tu Language[J].Applied Scienc-es,2024,14(15):6795-6795.

Data acquisition address

https://github.com/gazangcairang3077/TDSC-2024-Small-scale-sample-.git。

Commented [e4]: 按照审稿人意见，新添加两篇与本人方向接近且有借鉴意义的前沿文献。

Commented [e5]: 按照审稿人要求，已将部分数据共享至以下地址，包括藏语三大方言各典型土语的全部数据。同时，因本文数据是用于后续的学位论文研究，暂不全部对外开放，望审稿人谅解。

Appendix

Table 12

TDSC-2024 Data Collection Locations

No.	Dialect Category	Sub-dialect Category	Data Collection Location
1	U-Tsang	Central U-Tsang	Lhasa City, Chengguan District, Tibet Autonomous Region
2	U-Tsang	Central U-Tsang	Lhasa City, Chengguan District, Tibet Autonomous Region
3	U-Tsang	Central U-Tsang	Luoza County, Shannan City, Tibet Autonomous Region
4	U-Tsang	Central U-Tsang	Qusong County, Shannan City, Tibet Autonomous Region
5	U-Tsang	Lower U-Tsang	Nyalam County, Shigatse City, Tibet Autonomous Region
6	U-Tsang	Lower U-Tsang	Saga County, Shigatse City, Tibet Autonomous Region
7	U-Tsang	Lower U-Tsang	Jiangzi County, Shigatse City, Tibet Autonomous Region
8	U-Tsang	Ali	Geji County, Ali Region, Tibet Autonomous Region
9	U-Tsang	Ali	Gar County, Ali Region, Tibet Autonomous Region
10	Kham	Northern Route	Nangqian County, Yushu Tibetan Autonomous Prefecture, Qinghai Province
11	Kham	Northern Route	Dege County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province
12	Kham	Northern Route	Karuo District, Chamdo City, Tibet Autonomous Region
13	Kham	Southern Route	Batang County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province
14	Kham	Southern Route	Deqin County, Diqing Tibetan Autonomous Prefecture, Yunnan Province
15	Kham	Pastoral Area	Dingqing County, Chamdo City, Tibet Autonomous Region
16	Kham	Pastoral Area	Seini District, Nagqu City, Tibet Autonomous Region
17	Amdo	Pastoral Area	Aba County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
18	Amdo	Pastoral Area	Maqu County, Gannan Tibetan Autonomous

			Prefecture, Gansu Province
19	Amdo	Agricultural Area	Xiahe County, Gannan Tibetan Autonomous Prefecture, Gansu Province
20	Amdo	Agricultural Area	Xunhua County, Haidong City, Qinghai Province
21	Other	Jiarong	Ronga Township, Aba County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
22	Other	Zuo'ni	Niba Town, Zhuoni County, Gannan Tibetan Autonomous Prefecture, Gansu Province
23	Other	Diebu	Tiebu Town, Ruo'ergai County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province
24	Other	Muya	Bamei Township, Daofu County, Ganzi Tibetan Autonomous Prefecture, Sichuan Province