

ORIGINAL RESEARCH ARTICLE

Speech data collection system for KUI, a Low resourced tribal language

Subrat Kumar Nayak¹, Ajit Kumar Nayak², Smitaprava Mishra², Prithviraj Mohanty², Nrusingha Tripathy¹, Abhilash Pati^{1*}, Amrutanshu Panigrahi¹

¹ Department of Computer Science and Engineering, FET(ITER), Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar 701030, India

² Department of Computer Science and Information Technology, FET(ITER), Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar 701030, India

* Corresponding author: Abhilash Pati, er.abhilash.pati@gmail.com

ABSTRACT

A new generation of speech translation technology is being developed to enable natural cross-language communication. Research efforts must focus on large vocabulary, spontaneous speech, and speaker variances to accommodate the varying demands of speech recognition technologies. These are important issues that need to be resolved for the general application of voice recognition in realistic settings. Most languages with limited resources don't even have any speech data. Creating speech corpora is extremely difficult and time-consuming. Among all, KUI is regarded as one of the low-resource languages. In this paper, we developed the speech dataset for the KUI language to document and preserve their culture, tradition, and history for future generations. We also discuss the design, data collection procedures, and implementations and outline the different research possibilities using our KUI dataset. This paper mainly describes the GUI and method for the collection of KUI speech more quickly. In this section, the statistics of the people who helped and contributed to the collection of this KUI dataset have been provided. This study details a novel method of gathering data for any speech dataset. Using this process, we collected 60 hours of speech data sampled at 16 kHz by three different devices such as a Zoom recorder, Mobile, and Laptop from 80 different speakers. Each speaker contributed 500 sentences in the KUI language. A GUI application is designed to capture the speeches of numerous speakers in the KUI language. Several guidelines are proposed and used for the collection of the KUI speech dataset. All the guidelines are based on real-time experience gained during the data collection process by our team members.

Keywords: automatic speech recognition; speaker design; KUI dataset; corpus design; audio recording

ARTICLE INFO

Received: 7 August 2023
Accepted: 30 August 2023
Available online: 8 October 2023

COPYRIGHT

Copyright © 2023 by author(s).
Journal of Autonomous Intelligence is published by Frontier Scientific Publishing. This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).
<https://creativecommons.org/licenses/by-nc/4.0/>

1. Introduction

Human interaction can take place in many forms, including speech, sign language, gesticulation, and graphic language. Speech-based communication is regarded as the most popular and effective among these. This leads to the conclusion that oral communication should be used to handle human-computer interaction. The significance of creating an automatic speech recognition (ASR) system is thus highlighted by this fact. Increasing the capability of the speech recognition process automatically is the need for every domain-specific data. Also, using real user speech as a dataset increases the ASR's overall effectiveness. Finding an appropriate speech dataset is a very important and vital step in building an ASR system for KUI, a low-resource language. Yet, it is clear that low-resource languages do not have the necessary quantity and quality of

voice data^[1]. As a result, the equipment and methods used to gather speech data can be quite important for developing speech recognition systems. Gathering fresh speech data can be done under local circumstances, requiring the presence of speakers. The advantage of this technique and equipment is that it allows for environmental control, excluding elements like noise, disruption, and other influences on the speaker during recording^[2]. However, because it depends on the speaker's actual physical presence in the surroundings, which might have several challenges, this strategy is ineffective when there are a lot of speakers. In contrast, the application is designed to gather data remotely; it enables speakers to contribute to data collection regardless of their current location. This procedure is far more effective than the existing methods. A lot of work has been done for speech recognition. The systems developed for English and other European languages have achieved significant accuracy. However, the work in the speech domain for Indian languages is still in progress^[3].

As almost no work is done for the speech domain in the KUI language, it can lead to the development of an ASR system in the KUI language. Our motivation is to do speech recognition, which a speech data collection interface for the KUI language should lead. The research in the speech domain has attained new heights for English, other languages, and languages spoken in other developed countries. Hence, we selected to develop the speech database for the KUI language. In this work, we have tried to capture the maximum variant of the KUI language spoken in the Kandhamal district of Odisha^[4]. Speech data-gathering strategies and approaches are typically centered on websites and mobile applications as the world has become Internet-centric in terms of current trends and availability. Despite its inability to influence the existing environment in many parts of India, this method is acceptable by a large number of users, hence proving the most effective one at the present time. There are ten lakhs native KUI speakers in Odisha. Here, the first KUI-language speech recognition system construction has been attempted. We are putting forth an online database collection tool that is targeted toward low-resource language voice data collection. This website's capabilities are explored by gathering KUI speech data. The results of the explorations and analysis show that this technique makes it possible to quickly and painlessly gather the required amount of voice data. Other researchers who want to use this data collection method may find the outcomes helpful. To achieve more natural language communication, we are now creating a system for next-generation voice translation, which calls for speech recognition to be considerably more resistant to huge vocabulary, changes in speaking style, and speaker variances. These are typical issues with voice recognition in everyday settings.

Despite its benefits and impacts, machine learning calls for a lot of processing capacity, advanced computing capability, and specialized hardware like tensor processing units (TPUs) and graphics processing units (GPUs). The drawbacks of this traditional approach include significant latency, a slow workflow, large power, and high bandwidth. Only a small portion of spoken languages have used speech recognition technology worldwide^[5]. Before the development of ASR systems, this speech typically had to be curated and transcribed, and speech from a large number of speakers was typically gathered^[6]. To minimize the complexity, several tools and important guidelines are used to design the corpus for ASR technology^[7]. Here, we take care of both corpus size as well as number of speakers. In this paper, we focused on data collection methodology and deployed a speech processing system for a low-resourced language. Since 1960, computer scientists have been exploring ways to program computers to record, decipher, and comprehend human speech. Various domains, like school, home, military, medical, travel, artificial intelligence, etc., can greatly benefit from computer systems that can interpret spoken languages. To do any type of study, a researcher needs some historical information^[8-10]. Databases are typically essential for research.

We present the KUI-labeled speech dataset for the study of speech recognition. The area where the recording sessions took place was soundproof. The audio signal is gathered with the use of a directed, far-field microphone. The past ten years have seen significant growth in work for low-resource languages. In the past decade of work on creating language resources for languages that lack them, we have seen terms such as low

density and less commonly target. Large volumes of speech data are essential for speech technologies, such as speaker identification and voice recognition. This paper discusses the methodologies for collecting speech data from different speakers. The acquisition of speech corpora is a costly endeavor. The usual method of gathering data is to either call the speakers remotely or ask them to come on-site. The former offers the advantage of a controlled setting where conditions may be maintained uniformly across speakers. The latter enables users of a web-based interface to record speech.

The following sections outline how the research information in this publication is organized. Section II of the article describes the KUI language's background. In Section III, several related works are introduced. The methodology of dataset development is reported in section IV. In Section V, we briefly describe the user interface design. Evaluation and discussion of dataset preparation are reported in section VI. In Section VII, several points are described for future improvement. The conclusion part of our dataset preparation is described, along with the plans concerning the KUI dataset, which are discussed in Section VIII.

2. Background of KUI language

The indigenous people of Odisha have a very old culture. Their everyday activities, customs, and language, as well as their folklore, are unique from others. Their surroundings have an impact on their language, culture, and traditional knowledge. Even if they don't follow the crowd, they highly regard themselves. The tribal culture is evolving due to the impact of the mainstream and the growth of electronic media. The most crucial tool for integrating tribal people into society is language. The biggest number of Scheduled Tribe communities are found in Odisha, where 13 of the 62 tribal tribes are deemed Particularly Vulnerable Tribal Groups (PVTG). These tribes have 74 dialects and 21 official languages. Odia has been selected as the language of communication in dictionaries thus far, even though 7 of the 21 tribal languages have their own scripts. KUI speakers can also be found in different states of India, like Andhra Pradesh, Chhattisgarh, Madhya Pradesh, Maharashtra, and West Bengal. However, they are primarily concentrated in Odisha. Many of the 900,000+ who speak the language are Kondhs, a hunter-and-gatherer tribal group that has started adjusting to a more modern way of life. KUI is expressed in the Odia script, as there is no specific script for KUI^[4]. KUI speakers are mostly present in the state of Odisha, in which India is the primary language. The bulk of KUI-speaking Kondhs live in the mountainous forest areas of South and Central Odisha, especially in the districts of Kandhamal.

3. Related works

The most fundamental and crucial step in creating any recognition assignment is data collection. We have identified several methods for gathering information on speech through the study of numerous books and articles.

- 1) Recording directly;
- 2) Through telephone;
- 3) Web-based;
- 4) Questionnaires response through telephone.

Direct recordings are carried out under predetermined local conditions with the assistance of those who are aware of and engaged in the direct activity collection process. Professional studios are typically able to manage and exclude undesired voices to produce a good speech. Yet, these methods frequently have a dearth of speakers and are only equipped with a few speakers. Due to this, it takes more time to collect a big dataset of voice data using this technique^[11,12].

Using the telephone, ASR systems perform well in a local setting and perform less when dealing with data from the real world. It is one of the effective methods of data gathering^[13]. This study involved gathering

data while attaching specialized recording equipment to a basic telephone operator workspace^[14,15]. This experiment, however, was carried out a very long time ago. In the present era, call centers might be used for data collection if a similar strategy were used. As far as we know, unless they are specifically intended for call-related jobs, it is practically difficult to use recordings gathered in such a setting. Developing a speech dataset using this method is not feasible as long as cell providers are unwilling to share their amassed speech recordings or refuse to grant authorization to utilize them. Telephone survey replies are considered a more expensive corpus collection method. This approach is not the most cost-effective strategy to acquire a comprehensive speech corpus because of the limited financing for the majority of studies. Even if collecting data over the phone is preferred over gathering data in studios, this method is not regarded as the most practical. This approach has several drawbacks, including the speaker's uncertain environment, speaking ability, background noise, and user identification.

Much material has been written about creating a speech corpus using online resources. Radio, Television, and different social media sites like YouTube can all be considered sources for a web-resourcing strategy^[16,17]. Although these sites have millions of audio recordings that can be analyzed, this technique is not appropriate for all domains. Moreover, a large number of utterances that many speakers must pronounce are needed to adequately train the ASR system. Assuming that every recording from these sources has been obtained, it is clear that more effort has to be made to create transcriptions of these audio and text corpora. This profession necessitates lengthy hours. Many datasets can now be crowd-sourced thanks to the development of the Internet, but there are still many challenges to solve. Another web-based strategy is using online applications to collect speech recordings^[18,19]. The transcriptions can be predefined and delivered to speakers through a web application to gather the matching recordings. Additionally, this method enables recording various speech data types from various speakers and does not necessitate the user's physical presence in the area. Yet, relying on a computer that generates a range of recordings and unavoidable loud surroundings can be seen as significant barriers. Yet, given that we live in a society with ready access to the internet, the advantages of this method shouldn't be understated. We have discovered from reading the literature that systems make data collection possible. However, these systems are exceedingly expensive and inappropriate for less common languages like the KUI language^[20].

Using mobile devices to collect data and build the corpus is a very recent and popular strategy. Mobile devices are ideal and efficient for handling the gathering of voice data due to their widespread use and convenience. The program presents sentence transcriptions so that the speaker can record them using their microphone. The fact that this application can process offline is a very intriguing feature. If the device does not have internet access, offline recordings are stored in the phone's local memory^[21]. Later, an Android-based application was released that used a similar methodology. Even though these methods have several advantages over the other systems mentioned above, some problems and challenges cannot be avoided. The user's unpredictable environment is the key issue. As it is hard to control the user's environment, it is conceivable to record noises and sounds from outside the user's environment. As a result, we listen to each tape and decide whether to keep it or delete it based on its quality because we are unable to regulate and monitor the surroundings and recording quality^[22]. Only the admin user manages everything. With the use of this skill, we can address several problems with the data collection techniques we've already mentioned. Because practically every normal person has a smartphone with an adequate internet connection, the voice data collection process becomes much quicker and more efficient. This technique's independence from any particular field is another key feature, which is achieved by including any text transcription in the process of gathering the talks^[23]. Previous studies have taken into account the use of mobile applications. However, there have been significant limitations connected to the type of phones. Our program is web-based; it is not at all dependent on the model or type of smartphone. Several speech datasets have already been developed using several methods, which are given in **Table 1**.

Table 1. Various audio datasets available with the year of publishing.

Reference	Speech/language	Dataset creation year	Reference	Speech/language	Dataset creation year
[24]	Japanese	1996	[25]	Thai speech	2003
[26]	Arabic speech	2004	[19]	South African speech data	2006
[27]	Portuguese	2010	[11]	Japanese	2011
[18]	Swahili	2012	[28]	Sinhala	2013
[29]	Marathi speech	2013	[30]	Swahili language	2013
[31]	Chinese	2015	[32]	AGH corpus speech	2016
[7]	Frisian	2016	[33]	Romanian	2017
[34]	Santali	2017	[35]	Russian	2018
[36]	Assamese	2018	[37]	Kannada	2018
[38]	Chhattisgarhi speech	2018	[39]	Russian	2019
[40]	Myanmar speech	2019	[13]	Sinhala	2019
[41]	Kazakh speech	2020	[42]	Marathi	2020
[43]	Kazakh	2020	[44]	Turkish	2020
[45]	Danish	2020	[46]	Malayalam	2020
[47]	Russian	2021	[48]	Uzbek language	2021
[49]	Telugu	2021	[50]	Uzbek	2021
[51]	Sanskrit	2021	[52]	Marathi speech	2021
[53]	Kazakh language	2021	[9]	Dutch	2021
[15]	Annotated Arabic speech	2021	[54]	Mandarin speech	2022
[55]	Awadhi	2022	[55]	Bhojpuri	2022
[55]	Braj	2022	[55]	Magahi	2022
[56]	Kurdish	2022	[57]	The native language of Peru	2022
[58]	Kazakhhtts2	2022	[59]	Romanian	2022
[60]	Chukchi	2022	[61]	Marathi	2022
[62]	Armenian	2022	[63]	HuZhouSpeech	2022
[64]	Cantonese	2022	[65]	Catalan	2022
[66]	Norwegian	2022	[67]	Bangla	2022
[14]	Massive Arabic speech	2022	[16]	Tunisian Arabish corpus	2022
[68]	Telugu	2023	[69]	Kirghiz speech	2023
[20]	Algerian corpus	2023			

4. Methodology

During the development of the dataset, we considered four types of sentence units, which are as follows:

- **Statement:** Declarative sentences or fragments are known as statements, and they are typically punctuated with a period or an exclamation mark.
- **Question:** Complete phrases that act as interrogatives and end with a question mark are called questions.
- **Incomplete:** Grammatically incorrect statements are known as incomplete sentences. Usually, this happens in one of two circumstances: Either the speaker breaks off to rearrange his speech, or another speaker cuts the speaker off.
- **Non-speech:** Non-speech sounds include silence, music, or background noise.

Our methodology is based on some rules. These rules are the basis of a set of guidelines. According to our guidelines, we designed the architecture. We will go through each step in detail to demonstrate how to use this tool.

4.1. Set of rules

It is a crucial phase because the design rules specify the goals for the completion of the data collection process. Our process was developed by a set of rules, which are given below.

- 1) Flexibility: It is regarded as one of the necessary and significant characteristics in the case of data collection. It is avoidable to assemble a large number of people in one location. Developing a website is more productive and effective. All the recordings and translations are done from remote sites except for the final recording. Due to the quality of the audio to be preserved, the final recording should be done either in the soundproof room or the studio.
- 2) Structure and naming: As part of creating a large dataset, we must accommodate post-processing. For instance, organizing the files and folder so that speakers are separated, matching the titles of the audio files, and saving the recording in the desired format, which could be used in the future.
- 3) Capacity of recording: A maximum amount of data can be collected using the website. As KUI is a low-resourced tribal language, other methods may not be possible. We can modify the website and increase the server's space per our requirements.
- 4) Utility: The success of data collecting depends on the usability of the application. Different people can use the web-based data collection method. This needs a user-friendly interface with simple controls. Therefore, it is crucial to include instructions and Demo recordings.
- 5) Multilinguistic: A good application must support multiple languages. Our website can support several languages in the backend. It can, therefore, be utilized as a method to gather voice data in any language. The website has already been tested for other languages like English and Odia.

4.2. Architecture

The data collection architecture by the speaker or the Point of Contact (POC) is shown in **Figure 1**. Data can be collected using a desktop, laptop, or tablet. A speaker or POC makes the recording using the web application by the website given in **Figure 2**. These recordings are then transmitted through the HTTP protocol to our server. As a result, the application can be used on any platform.

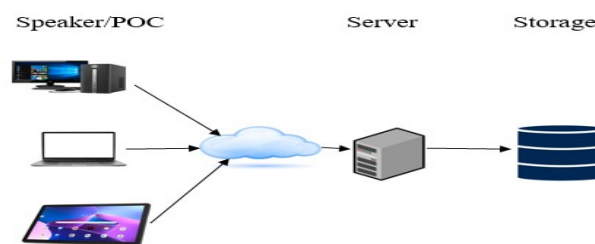


Figure 1. The architecture of KUI data collection.



Figure 2. The data collection in different environments in different places.

4.3. Technology

Our web-based application was created using modern technologies that can support almost every handheld device or computer available today. The application's backend was built using the php7.1 language

with the database MY SQL 5.0 on the client side, and the server type is MariaDB version 10.1. For service communication, we used jQuery version 3.4.1. Bootstrap3 is used here for frontend Design. For the local host webserver, we used Apache 2.4.25. We are using Hostinger-Web Hosting Servers for 1 TB of space. We also used XAMPP as it is completely free and easy to install Apache distribution containing MariaDB with PHP.

4.4. Data collection and transcription

Speech data are recorded by POC only with a digital recorder, but the speaker is recorded in 3 devices, i.e., zoom recorder, mobile, and digital recorder. There will be no physical contact between the recordings^[70]. All the collected speeches have already been translated from Odia text to KUI text. The recording condition is shown in **Table 2**. An example of file format is shown in **Table 3**.

Table 2. Recording condition specifications.

Microphone	Uni-directional (Name of the device)
Sampling frequency	kHz
Quantization	16 bits symmetric
An atmosphere of silence	Yes
Face-to-face contact	No
Topology of dialogue	One-to-one
Language	KUI (Tribal language)

Table 3. Example of a possible file attribute.

Attribute	Value	Indication
Sex	Female	F
Language	KUI	KU
Speaker regd no	A number	104
Sentence serial no	Serial no	30504
POC regd no	Regd no of POC	2
Date of recording	Actual date	19-1-2023
Time of recording	Actual time	10-30-34
Indication of file	F_KU_104_30504_2_19-1-2023_10-30-34.wav	

5. Brief description of the user interface

Our system has three crucial tasks to carry out the entire data collection process. These are admin users (only 1–2 users), Point of Contact (POC), and multiple speakers. Admin is eligible to add POC and speakers. The administrator can check to see if the recording is complete and can hear every speaker’s recording. Admin can list the recordings of all speakers. He can also notify the POC and speakers who made mistakes and delete undesirable or corrupted recordings. The administrator can download all the recordings when the recording process is complete. The downloadable zip files come with folders of speakers in a structured manner. We will go through each specific step in depth in the following section. The section consists of three parts (admin, point of contact, and speaker). First, we need to visit the URL shown in **Figure 3**.

Language Corpus Collection Link
<https://janabhasha.in/bhashasangraha/index.php>

Figure 3. Data collection link for KUI language.

After opening the website, it will ask for a user ID and password. The screen short of it is shown in **Figure 4**.

Figure 4. Login section for data collection homepage.

5.1. Admin role

Let us analyze the Admin’s Role first. After giving the correct login credentials to the admin, it will go to the admin section, shown in **Figure 5**.

Figure 5. Admin section of the data collection website.

5.1.1. POC Registration

It is not possible to contact all the speakers in the Admin. Therefore, POC takes the intermediate position between the admin and the speaker. The criteria of a POC are as follows:

- Age must be greater than 18;
- Able to read and write Odia and KUI language;
- Able to translate KUI from Odia;
- Able to speak the KUI language;
- Must be aware of operating a laptop/desktop;
- Local people from the Kandhamal district of Odisha.

A person who fulfills the above condition may be selected as a POC. There are a minimum of 4 to 5 POCs required. Initially, we selected four no of POCs. After selection, we enter their data in the database through the interface shown in **Figure 6**.

Figure 6. POC (point of contact) registration page.

The name, address, mobile no, and some other information required to register a POC. A POC must have a valid email ID and contact number for communication. After registration, a POC has a valid User ID and Password. After the POC registration, a speaker can be registered under one POC. After this step, we go for the speaker registration.

5.1.2. Speaker registration

A speaker can be registered by the Admin only under one POC. Under one POC, several numbers of speakers can be registered. A speaker cannot be registered more than once under any POC. Therefore, some validation is given when we register a speaker. The speakers are classified based on gender. A speaker must fulfill the following conditions:

- Age must be greater than 18;
- Able to read, listen, and speak the KUI language;
- Sound knowledge of computers;
- Local people from Kandhamal;
- Must be referred by one of the POC.

Speakers were given basic information about the headset used and when to speak the word. After the selection procedure, we entered the details of the speakers in the database under one POC through the interface shown in **Figure 7**. All the steps were repeated for all the speakers.

ଢୁଢ଼ା ନାମ (Speaker Full Name)*	<input type="text"/>
ଠିକଣା (Address)*	<input type="text"/>
ଲିଙ୍ଗ (Gender)*	Male
Age*	<input type="text"/>
Qualification*	Graduate
ଇ-ମେଲ (email)	<input type="text"/>
ମୋବାଇଲ (Mobile)	<input type="text"/>
Adhar Number*	<input type="text"/>
Max Limit*	<input type="text"/>
POC*	Gobinda Pradhan

Submit

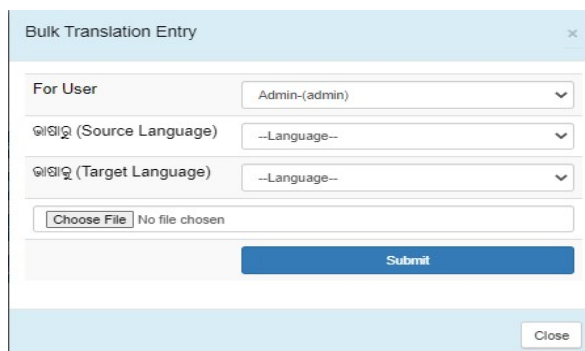
Close

Figure 7. Register for a speaker in the admin section.

The name, address, gender, age, qualification, and Aadhar number are essential for registration. A maximum number of sentences the speaker can record must also be entered here. The gender column plays a major role here. Speaker identity is validated through Aadhar no here, and it can also avoid repetition of speakers. The name of the POC is selected from the POC column. After the above steps, the speaker is registered and ready for recording.

5.1.3. Bulk translation entry

In the case of bulk translation entry, we enter the Odia text to translate it into KUI text. A person who understands both Odia and KUI would be able to translate. In the case of translation, the user may be a POC or a speaker whose interface is shown in **Figure 8**.



Bulk Translation Entry

For User: Admin-(admin)

Source Language: --Language--

Target Language: --Language--

Choose File: No file chosen

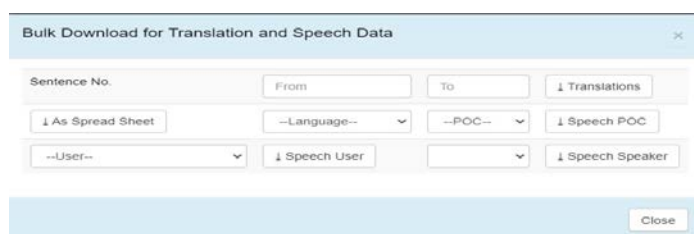
Submit

Close

Figure 8. Bulk translation entry level by admin.

5.1.4. Bulk download

After the completion of translation or recording, the admin can download both audio as well as text, which the POC or speaker records. The interface for the same is shown in **Figure 9**.



Bulk Download for Translation and Speech Data

Sentence No. From To Translations

As Spread Sheet --Language-- --POC-- Speech POC

--User-- Speech User Speech Speaker

Close

Figure 9. Dataset downloads from a specific range.

In the bulk download, first, we have to give the starting and ending sentence numbers for which the admin wants to download. If we click on translation, then all the .txt files are downloaded in different folders in ascending order. But if we click on a spreadsheet, then a file will be downloaded in .xlsx format where all the sentences are translated; a screenshot of the output is shown in **Figure 10**.

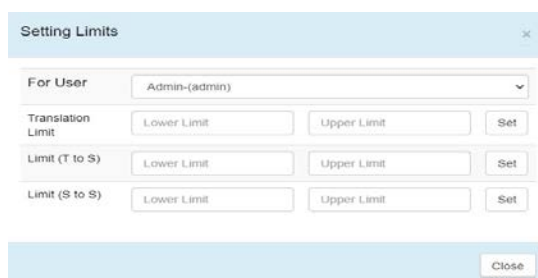
Sentence No	English	Odia	Kui
1	A buffalo is playing in the mud	ମଜଷିଟିଏ କାଦୁଅରେ ଖେଳୁଛି ।	ଉଷେ ବୋରୁ ଗେଦେଗାଳି କାହାଜମାନେ ।
2	A bull hit the cart	ଗୋଟିଏ ଗଞ୍ଜ ଶରତ ଗାଡ଼ିକୁ ଧକ୍କା ଦେଲା।	ଉଷେ ଗଞ୍ଜକୋଟି ଶରତ ଗାଡ଼ିକି ଦୁଆଡ଼ତେ ।
3	A cold wind was blowing	ଫୁଣ୍ଟା ପବନ ପ୍ରବାହିତ ହେଉଥିଲା।	ବିଲମାଦାରି ବିଲୁ ଗୁଞ୍ଜେଇ ଆକମାସେ।
4	A dead animal is floating in the river	ଗୋଟିଏ ମଲା ପଶୁ ନଦୀରେ ଭାସୁଛି	ଉଷେ ସାକାମାରି କଟ ଚଟିଦାରି ଦେଇଜାନ୍ନେ
5	A dust storm is coming	ଗୋଟିଏ ଧନିଷ୍ଟ ଆସୁଛି	ଉଷେ ଦୁନିଗାଡ଼ୁ ବାଜନ୍ନେ

Figure 10. The screenshot of the output of the download.

The audio files recorded by the POC can be downloaded by clicking the Speech POC download button. The audio of the speaker can be downloaded by selecting the POC and the speaker name and then clicking Speech Speaker Download at the bottom. It will download as a Zip file where all the audio files are present in .wav format in sequential order.

5.1.5. Setting limits

Admin can also set the limit of both translation and recording, as shown in **Figure 11**.



Setting Limits

For User: Admin-(admin)

Translation Limit: Lower Limit Upper Limit Set

Limit (T to S): Lower Limit Upper Limit Set

Limit (S to S): Lower Limit Upper Limit Set

Close

Figure 11. Setting the limit for recording.

To set the limit, we have to select the user's name, and after that, the lower limit and upper limit number will be entered, and then the set bottom will be clicked. After reaching the upper limit, it will automatically stop. A user cannot record beyond that.

5.1.6. Set POC ≥ Speaker

A POC can be changed for certain circumstances. A new POC can be assigned using this option and an authorization letter from the old POC. This option is sometimes very crucial from the Admin's point of view. The interface to change POC is shown in **Figure 12**.

Figure 12. Change the POC of a speaker.

5.1.7. Additional options

Several other options are present in the admin part, like Missing Audio files, Progress, Graph, Bulk Translation Delete, and Download survey. Due to some problem, if the audio cannot be recorded, it will come under the missing audio file. The statistics of the progress can be shown graphically under the progress option, where the admin can know the status of the work. Another very important option is Bulk Translation Delete. In this option, the sentences can be deleted due to wrong translations. After deleting, it will automatically come again for re-translation. The interface for this is shown in **Figure 13**.

Figure 13. Bulk translation deletes for a specific user.

Because of the above options available for the admin, it is easier to verify and validate the accuracy of the data-gathering process. Also, the administrator can interact with the speaker directly and point out any problems or mistakes.

5.1.8. Client role

Compared to the admin user, the speaker and POC have limited privileges. The quality of the translation depends upon the POC. The POC does the translation task, i.e., from Odia text to KUI text. The translation and recording process steps are shown below in **Figure 14**.



Figure 14. Home page for the client.

5.2. Translation

A POC can do the login using his login credentials. The admin previously uploaded the Odia text to the server, which will be translated into KUI text. After clicking the Translation button, a sentence will come up, and the POC translate it and submit it to fetch the next sentence. The process continues until the sentence's end as assigned by the admin. After the end of the translation phase, all the sentences are available in KUI text.

5.3. Text-to-speech

Once the translation process is done, the POC will be redirected from KUI text to KUI speech.

The text-to-speech part can be done either by the POC himself or by any speaker who knows the KUI language very well through the interface shown in **Figure 15**. In our work, the POC himself does all the text-to-speech conversion. The POC recorded the sentences one by one in ascending order. All the data are stored in the database. The POC recorded the sentences at the ease of their home. Hence, the speech samples are noisy. This process will be lengthy. If we include more POC, then the process will be faster. After some parts of the recording are completed, we go for speech to the speech conversion process, which is our final goal.

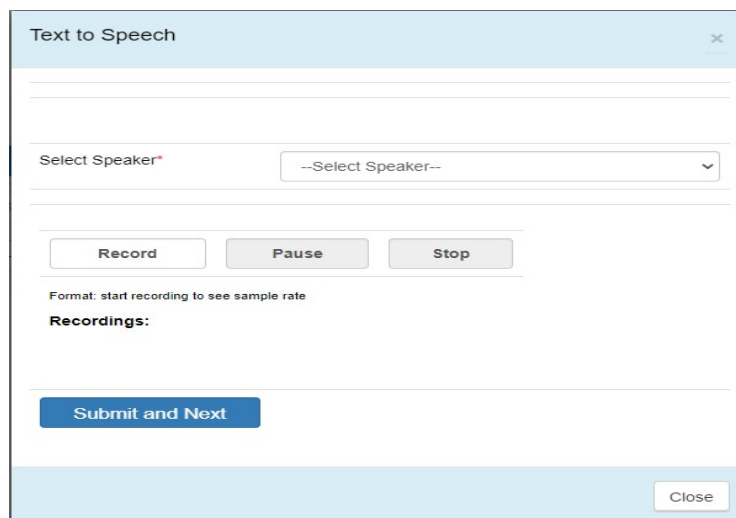


Figure 15. Text-to-speech conversion.

5.4. Speech-to-speech

Once the translation process is done, the POC will be redirected from KUI text to KUI speech. After the completion of the POC part, we go for the final recording. The recording is carried out in a noise-free room, either in a studio or noiseless room. Some data are recorded in different parts of the Kandhamal district. All the speech-to-speech conversion process is carried out in the presence of the POC or any authorized person by the Admin, who can handle the different challenges encountered during the recording. The speech-to-speech recording interface is shown below in **Figure 16**.

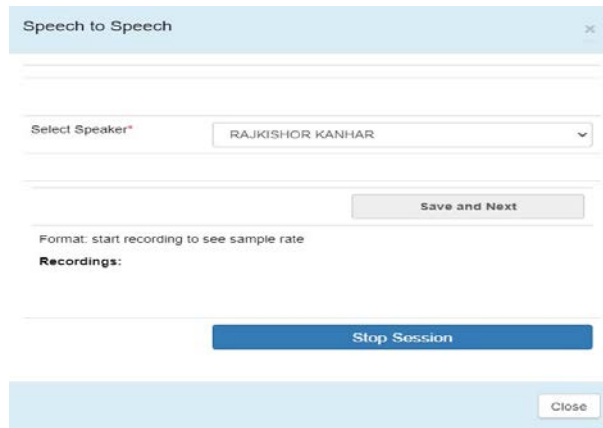


Figure 16. Speech-to-speech conversion of KUI language.

There is no special login for speakers. A speaker can log in through the account of the POC under whom he is registered. After login in, the speaker selects his name from the selection box. The sentence no, KUI text with the audio of the text given by the POC will come one by one on a page whose screenshot is shown in **Figure 17**.

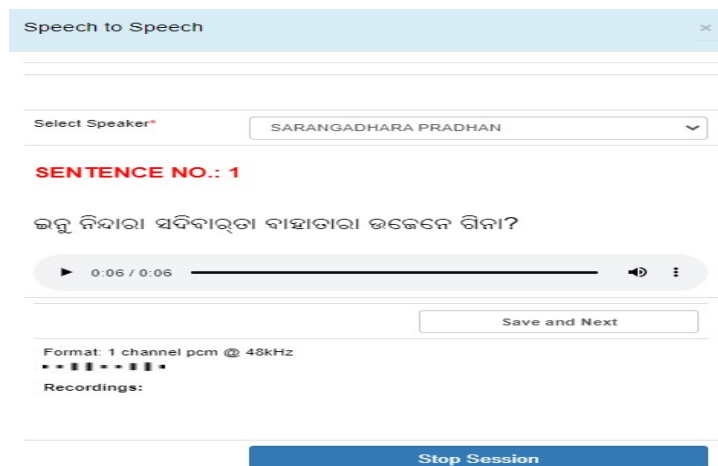


Figure 17. Example of speech-to-speech conversion.

After listening to the voice recording of the given sentence, there will be a beep sound. After the beep sound, the speaker will start reading. After that, the speaker must click the save and the next button. Then, the next sentence will come. This process will continue till the last sentence. The session will automatically stop when the number of sentences reaches the assigned number by the admin. In between, a speaker may take rest some time by clicking the stop session. If any problem arises between recordings, there is an option on the home page in which the speaker can listen to his recording; if it is not up to the mark, it can be deleted. After deletion, the sentence will appear again for recording. The audio file as .wav and transcription as .txt are stored in the database via an HTTP request. The administrator user can then review the recorded sentences and assess their quality.

6. Evaluation and discussion

We have gathered over 60 hours of voice data with the abovementioned method. This process featured about 80 speakers, with 45 men and 35 women. Five hundred sentences, on average, were supplied to each speaker. Male speeches outnumbered female speeches, as seen in the statistics on the admin page. As KUI is a tribal language, the females are not willing to give their voices. We obtained 40,000 clear audio recordings and related transcriptions. Each audio clip lasts about 5 seconds on average. The duration of the recording is 60

hours. The team members manually reviewed and edited the collected data as well as discarded unnecessary and damaged recordings. We found that the error percentage of recording is less than 2%. Initially, the error percentage was 8%. However, after we analyzed it and gave a demo recording option on the login page, the error percentage decreased significantly. In the demo, a speaker is not required for login, but he can try the recording. The demo recording will not be stored in the database, but it will help the speaker practice and be useful for final recordings.

Due to the popularity of web applications' researchers have tried to construct speech datasets using different web-based methods. One of the efficient methods for creating speech datasets is through web-based crowdsourcing. Speaking samples from smartphones are used by several voice-recognition devices. The primary goal of KUI speech data collection is to produce a dataset and train neural network models. Regardless of socioeconomic, geographic, cultural, or other circumstances, we wanted as many KUI speakers to participate. The development of extensive corpora is challenging for tribal languages with limited resources, such as KUI. KUI language in the web resources suffers in existence as compared to other languages. Even though there are few places where one can obtain data, it almost certainly costs a significant financial cost. Almost all people use smartphones with internet facilities having microphones. KUI speech is a fresh dataset for research on speech recognition. We have explained how this dataset was created. In the beginning, we thought of many people involved in having smartphones. However, letting people record on their smartphones without providing instructions or ground rules is utterly inefficient. Therefore, we went for an intermediate person whom we call a POC. The sentences and their quantity should be given to the speakers as hard copies for clean and clear audio recording. Using our tool, all the issues were resolved for speech collection. Such technology can be useful for researchers because it gives them a solid foundation in a short period. Additionally, it offers a high-quality audio recording of speech since one person can evaluate the entire process while sitting at the ease of his home. For the KUI language, a few guidelines and norms are presented. But it can be used for other tribal languages with limited resources, too. Once the platform is ready, we will record any number of sentences in various languages. We also used some instruments and environments for clear audio recordings.

6.1. Microphone

MAONO AU-903 studio-quality USB microphone is used here for recording. The quality of the microphone is good, the stand is heavy, and the bottom is made up of rubber, as shown in **Figure 18**. The sound captured by the microphone can be cardioid or omnidirectional. It is compact and has all the features of an expensive microphone. The microphone is excellent for voiceovers and podcasting; best of all, it is simple to use. This microphone can be adjusted in several ways to provide crystal-clear sounds.



Figure 18. MAONO AU-9.0.3. Microphone.

Here, the built-in MUTE function is there. When someone wants to mute, he only has to press this button. A button with an indicator light on the base can be used to turn on/off the mic easily. Cardioid rhythms are good at reducing noise. For multi-person use, the omnidirectional pickup mode is better. With the polarity

switch on the back, the user can alter the pickup mode for the microphone. The microphone can respond to the frequency range of 20 Hz to 20 KHz.

6.2. Recorder

Another recorder is also used for recording. The only portable device featuring five built-in microphones and four different recording modes, including X/Y, mid-side, 2-channel surround, and 4-channel surround, is the Zoom H2N handy recorder, as illustrated in **Figure 19**. It records directly to SD and SDHC cards of up to 32 GB. It supports up to 24-bits/96 kHz wav audio as well as a variety of mp3 formats. The H2N allows us to easily convert WAV files to MP3 format; we can use its internal mixer to mix down four-channel surround files to stereo, with independent control over level and panning.



Figure 19. Zoom recorder for speech data collection.

The H2N provides a range of inputs and outputs for recording versatility. There is a stereo line headphone jack with a dedicated volume control and a Line connector that can accept two channels of mic and line-level signals. X/Y recording provides a great way to cover a wide area while still capturing sound sources in the center with clarity and definition. The H2N's built-in X/Y microphone provides two matched unidirectional mic elements set at a 90-degree angle relative to one another. The “Side” microphone in the H2N's MS microphone captures noises from the left and right, while the “Mid” microphone in the H2N's MS microphone picks up signals immediately in front. We can then modify the stereo width while still perfectly maintaining mono compatibility by adjusting each relative level, either during recording or post-production. With the help of the H2N, we can combine the signals from the X/Y and MS microphones to produce two or four-channel surround sound recordings that capture all of the sounds we hear, not just those that are directly in front of us.

6.3. Flexible stand

We used the pTron Mount LD3 flexible stand for a zoom recorder and one mobile here. The adjustable long arm and rotatable clip allow us to view our device at any angle, as shown in **Figure 20**. Soft silicone pads hold the device securely and protect it from scratches.



Figure 20. Flexible stand for mobile and zoom recorder.

The p-Tron Mount LD3 flexible stand with a 138 cm longer neck allows a huge range of movement and

the bracket for 360-degree free rotation. To have the most comfortable experience, we might use our smartphone at the appropriate distance and angle. Wide device compatibility with a 4.72 width, stronger ABS plastic and metal hose, a flexible platform clamp that adjusts from 0 mm to 80 mm, and the ability to be fixed securely to a metal frame, workstation, kitchen cabinet, or any metal plate of varying depths. Anti-slip rubber covers the clip and base clamp, preventing any scuffs or marks while enhancing the grip.

6.4. Pop-shield filter

The pop-shield filter supports getting the greatest vocal recordings that are simple to comprehend and ensures our message's loudness and clarity. To effectively eliminate the feared hissing and lipping sounds produced by overeager performers, a 360-degree flexible gooseneck holder is composed of smooth, bendable steel material and can be adjusted for accurate alignment.

It is appropriate for control rooms, chat rooms, broadcasting rooms, stages, and personal as well as commercial recording studios. Any mic stand, table, desk, shelf, counter, etc., can be fixed securely with an adjustable clamp like the one in **Figure 21**.



Figure 21. Pop-shield filter required for data collection.

6.5. Laptop, mobile, and speaker

We used one advanced version of the laptop and a mobile phone. We also used a speaker which is connected to the laptop for better sound quality. We took a Zebronic wireless Bluetooth portable speaker with a support carry handle, as shown in **Figure 22**. It is a small, useful portable speaker with multiple connecting choices, including wireless BT, USB, micro-SD, and AUX. The frequency response is 120 Hz-15 kHz. It's charging time is 2.5 hours. Its Playback time is approximately 10 hours.



Figure 22. Zebronic wireless Bluetooth speaker is needed for recording.

7. Future improvements

Working with people who cannot understand languages except the KUI language is very difficult. We have faced many difficulties in the process of collecting the data. After facing problems, our team members tried to solve them. This information must be taken care of for the new researchers who try to work using this method.

- Speaker interaction: We came across the issue of being unable to communicate with speakers during the audio recording procedure. We faced difficulty in alerting a specific user while monitoring and checking

the recording. Despite having the option to contact them directly, it may take some time to see the warning and correct the problem in subsequent audio. Ensure that the user detects the problem right away and takes steps to improve the standard of his subsequent recording. We, therefore, use POC, who is intermediate between speaker and admin. In the whole process of recording a speaker, a team member was present to avoid these types of problems.

- Duration of the sentence: Each person's speaking and talking styles are different. Therefore, it is difficult to maintain the recording speeds. We discovered that some of the recordings were extremely brief, and some were very extensive. We intended to impose restrictions on audio length to manage this problem. We know this approach won't eliminate the issue, but at least it will help us cut down the number of audio clips with unnecessary lengths. The quality of the audio samples is crucial for neural network models.
- Noisy environment: Background noise has been one of the most regular issues while gathering data. As our University provides an FM radio station, we record the audio of the speaker. However, the POC recorded the entire recording in a noisy environment. It is very difficult to gather individuals in a conducive environment in one place. The cost is more of asking every speaker to travel to a specific location far away. With the use of a script or application, we intend to execute noise removal techniques. A model's performance can be greatly improved by improving the quality of the post-processing stage.
- Public awareness: We came to know that collecting voice data is an irksome and time-consuming operation. According to our experience, recording 500 sentences is not a very big task. It will take 3 hours. But convincing a KUI speaker is very difficult. According to our statistics, more than 20% of registered users haven't even added a single sentence. After public meetings and awareness, most individuals were motivated by this innovation and became conscious of our help in speech processing in their tribal language. Therefore, public meetings and awareness not only aid in the collection of more information but also directly impact the caliber of audio clips.

8. Conclusion and future plans

Many speech datasets for popular languages are available online, which enables researchers to conduct adequate experiments on speech recognition. For low-resourced tribal languages like KUI, it is not available. Due to lack of data, ASR systems for such languages do not exist. Researchers might use different procedures to generate their speech corpus for any language. In low-resourced tribal languages, the existing methods or procedures are incredibly difficult and ineffective. To overcome these issues, we created a tool that collects speech data via a web application. Our website is also usable through the smartphone screen, as smartphones are integral to our everyday lives. The use of this method addresses two key issues: quality speech recording and large datasets. The audio snippets are of high quality because of our appropriate monitoring system and recording technology. Users who get engaged in the process easily contribute a lot of audio data. Using our technique, we gathered 60 hours of audio recording. There were 80 participants, out of which 45 were males and 35 were females. They were almost all the residents of Kandhamal District. All the speakers aged between 18 to 70. We have 40,000 recordings after the full monitoring and checking procedure. For the data collection procedure, we took six months. The method we used to build the KUI dataset was a great success, and we intend to repeat this study shortly for other languages. We believe that this data-gathering approach will benefit any expert dealing with a dearth of information for tribal languages with limited resources. The ability to combine a huge volume of data with little effort is tremendously fascinating to researchers. The dataset's goal is to make the data collection accessible for research work. Before a wider release, we must finish data collection and process the data by ethical and privacy considerations. We are sure that gathering this data will open up new options for linguistic, sociological, and spoken language technologies. This dataset is meant to be accessible and useful for a wide range of users. According to their study, the researchers are free to use the most effective approaches for the dataset. The main purpose of this work was to create a KUI dataset, as there

is no existing KUI speech dataset for research work. Future work also involves collecting data from new places and increasing the size of the dataset.

Data collection is an ongoing process. We expect the dataset to be used widely for research purposes. We are planning to start the public release of them. Collections of this kind can be altered in a variety of ways, either generally or specifically. A dataset of this kind might be expanded to show changes in speaking speed, voice effort, articulation accuracy, etc. Speakers may be chosen to represent a range of ages, gender identities, ethnic backgrounds, or locations. The effects of gender, age, and other factors can be measured by other very important potential research that can be conducted on the recently produced dataset. We conducted some pilot research and found that women are more adept at speaking properly than men. Future research can be conducted in more detail.

Author contributions

Conceptualization, SKN; methodology, SKN, AKN, and SM; software, PM; validation, NT, AP (Abhilash Pati) and AP (Amrutanshu Panigrahi); formal analysis, AP (Abhilash Pati) and AP (Amrutanshu Panigrahi); investigation, SKN, and AKN; resources, SM and PM; data curation, NT; writing—original draft preparation, SKN; writing—review and editing, AKN, SM, PM, and NT; visualization, AP (Amrutanshu Panigrahi) and AP (Abhilash Pati); supervision, AKN, SM, and PM. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The transcription of speech data mentioned in this research was done collaboratively with IIT Dharwad, IIT Hyderabad, and IIT Bhubaneswar. Authors respect them for their outstanding work. The writers acknowledge Dr. S. R. Mahadeva Prasanna for his ongoing assistance and support. Also, we want to thank everyone who contributed to the project titled “Speech-to-Speech Translation for Tribal Language Using Deep Learning Framework”. This activity is currently conducted with external support by the Ministry of Electronics & Information Technology, Government of India.

Conflict of interest

The authors declare no conflict of interest.

References

1. Magueresse A, Carles V, Heetderks E. Low-resource languages: A review of past work and future challenges. *arXiv* 2006; arXiv:2006.07264. doi: 10.48550/arXiv.2006.07264
2. Larcher A, Lee KA, Ma B, Li H. Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication* 2014; 60: 56–77. doi: 10.1016/j.specom.2014.03.001
3. Singh A, Kadyan V, Kumar M, Bassan N. ASRoIL: A comprehensive survey for automatic speech recognition of Indian languages. *Artificial Intelligence Review* 2019; 53(5): 3673–3704. doi: 10.1007/s10462-019-09775-8
4. Nayak SK, Nayak AK, Mishra S, Mohanty P. Deep learning approaches for speech command recognition in a low resource KUI language. *International Journal of Intelligent Systems and Applications in Engineering* 2022; 11(2): 377–386.
5. Ranathunga S, de Silva N. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. *arXiv* 2022; arXiv:2210.08523. doi: 10.48550/arXiv.2210.08523
6. Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 2014; 56: 85–100. doi: 10.1016/j.specom.2013.07.008
7. Ghyselen AS, Breitbarth A, Farasyn M, et al. Clearing the transcription hurdle in dialect corpus building: The corpus of southern Dutch dialects as case study. *Frontiers in artificial intelligence* 2020; 3: 10. doi: 10.3389/frai.2020.00010
8. Yilmaz E, van den Heuvel H, Dijkstra J, et al. Open source speech and language resources for Frisian. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016); 8–12 September 2016; San Francisco, USA. pp. 1536–1540.

9. Lee KA, Wang G, Ng KP, et al. The reddots platform for mobile crowd-sourcing of speech data. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015); 6–10 September 2015; Dresden, Germany.
10. Hinskens F, Grondelaers S, van Leeuwen D. Sprekend Nederland, a multi-purpose collection of Dutch speech. *Linguistics Vanguard* 2021; 7(s1): 20190024. doi: 10.1515/lingvan-2019-0024
11. Schultz T, Vu NT, Schlippe T. Globalphone: A multilingual text & speech database in 20 languages. In: Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013); 26–31 May 2013; Vancouver, Canada. pp. 8126–8130.
12. Masumura R, Hahm S, Ito A. Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association 2011 (INTERSPEECH 2011); 27–31 August 2011; Florence, Italy.
13. de Vries NJ, Davel MH, Badenhorst J, et al. A smartphone-based ASR data collection tool for under-resourced languages. *Speech communication* 2014; 56: 119–131. doi: 10.1016/j.specom.2013.07.001
14. Buddhika D, Liyadipita R, Nadeeshan S, et al. Voicer: A crowd sourcing tool for speech data collection. In: Proceedings of the 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer 2018); 26–29 September 2018; Colombo, Sri Lanka. pp. 174–181.
15. de Silva N. Survey on publicly available sinhala natural language processing tools and research. *arXiv* 2019; arXiv:1906.02358. doi: 10.48550/arXiv.1906.02358
16. Al-Fetyani M, Al-Barham M, Abandah G, et al. MASC: Massive Arabic speech corpus. In: Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT 2022); 09–12 January 2023; Doha, Qatar. pp. 1006–1013.
17. Salama A, Bouamor H, Mohit B, Oflazer K. YouDACC: The YouTube dialectal Arabic commentary corpus. In: Calzolari N, Choukri K, Declerck T, et al. (editors). Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014); 26–31 May 2014; Reykjavik, Iceland.
18. Mubarak H, Hussein A, Chowdhury SA, Ali A. QASR: QCRI aljazeera speech resource—A large scale annotated Arabic speech corpus. *arXiv* 2021; arXiv:2106.13000. doi: 10.48550/arXiv.2106.13000
19. Gugliotta E, Dinarelli M. TARc: Tunisian Arabish corpus first complete release. *arXiv* 2022; arXiv:2207.04796. doi: 10.48550/arXiv.2207.04796
20. Barnard E, Davel M, van Heerden C. ASR corpus design for resource-scarce languages. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009); 6–10 September 2009; Brighton, United Kingdom.
21. Gelas H, Besacier L, Pellegrino F. Developments of Swahili resources for an automatic speech recognition system. In: Beermann D, Besacier L, Sakti S, Soria C (editors). Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages; May 2020; Marseille, France.
22. de Wet F, Louw P, Niesler T. *The Design, Collection and Annotation of Speech Databases in South Africa*. Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech); 2006.
23. Zergat KY, Selouani SA, Amrouche A, et al. The voice as a material clue: A new forensic Algerian Corpus. *Multimedia Tools and Applications* 2023; 82: 29095–29113. doi: 10.1007/s11042-023-14412-2
24. Nakamura A, Matsunaga S, Shimizu T, et al. Japanese speech databases for robust speech recognition. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996); 3–6 October 1996; Philadelphia, PA, USA. pp. 2199–2202.
25. Kasuriya S, Sornlertlamvanich V, Cotsomrong P, et al. Thai speech corpus for Thai speech recognition. In: Proceedings of The Oriental COCODA 2003 International Coordinating Committee on Speech Databases and Speech I/O System Assessment; 1–3 October 2003; Singapore. pp. 54–61.
26. Al-Diri B, Sharieh A, Hudaib T. An Arabic speech corpus: a database for Arabic speech recognition. *Dirasat: Pure Sciences* 2004; 31(2): 208–219.
27. Calado A, Freitas J, Silva P, et al. Yourspeech: Desktop speech data collection based on crowd sourcing in the internet. In: Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language; 27–30 April 2010; Porto Alegre/RS, Brazil.
28. Nadungodage T, Welgama V, Weerasinghe R. Developing a speech corpus for sinhala speech recognition. In: Alonso JM, Bugarín A, Reiter E (editors). Proceedings of the 10th International Conference on Natural Language Processing (ICON 2013); 18–20 December 2013; Noida, India.
29. Gaikwad S, Gawali B, Mehrotra S. Creation of Marathi speech corpus for automatic speech recognition. In: Proceedings of the 2013 International Conference Oriental COCODA held jointly with the 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCODA/CASLRE 2013); 25–27 November 2013; Gurgaon, India. pp. 1–5.
30. Oirere AM, Deshmukh RR, Shrishrimal PP. Development of isolated numeric speech corpus for Swahili language for development of automatic speech recognition system. *International Journal of Computer Applications* 2013; 74(11): 20–22. doi: 10.5120/12929-9841
31. Wang D, Zhang X. Thchs-30: A free Chinese speech corpus. *arXiv* 2015; arXiv:1512.01882. doi: 10.48550/arXiv.1512.01882

32. Żelasko P, Ziółko B, Jadczyk T, Skurzok D. AGH corpus of polish speech. *Language Resources and Evaluation* 2016; 50(3): 585–601.
33. Stan A, Dinescu F, Țiple C, et al. The SWARA speech corpus: A large parallel Romanian read speech dataset. In: *Proceedings of the 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpED 2017)*; 6–9 July 2017; Bucharest, Romania. pp. 1–6.
34. Akhtar AK, Sahoo G, Kumar M. Digital corpus of Santali language. In: *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI 2017)*; 13–16 September 2017; Udupi, India. pp. 934–938.
35. Iakushkin OO, Fedoseev GA, Shaleva AS, Sedova OS. Building corpora of transcribed speech from open access sources. In: *Proceedings of the VIII International Conference “Distributed Computing and Grid-technologies in Science and Education” (GRID 2018)*; 10–14 September 2018; Dubna, Moscow region, Russia. pp. 475–479.
36. Deka B, Chakraborty J, Dey A, et al. Speech corpora of under resourced languages of North-East India. In: *Proceedings of the 2018 Oriental COCODA-International Conference on Speech Database and Assessments*; 7–8 May 2018; Miyazaki, Japan. pp. 72–77.
37. Unnibhavi AH, Jangamshetti D. Development of Kannada speech corpus for continuous speech recognition. *International Journal of Computer Applications* 2018; 179(53). doi: 10.5120/ijca2018917255
38. Londhe ND, Kshirsagar GB. Chhattisgarhi speech corpus for research and development in automatic speech recognition. *International Journal of Speech Technology* 2018; 21: 193–210. doi: 10.1007/s10772-018-9496-7
39. Gabdrakhmanov L, Garaev R, Razinkov E. Ruslan: Russian spoken language corpus for speech synthesis. In: Salah AA, Karpov A, Potapova R (editors). *Lecture Notes in Computer Science, 11658*, *Proceedings of the Speech and Computer: 21st International Conference, SPECOM 2019*; 20–25 August 2019; Istanbul, Turkey. pp. 113–121.
40. Mon AN, Pa WP, Ye KT. UCSY-SC1: A Myanmar speech corpus for automatic speech recognition. *International Journal of Electrical and Computer Engineering* 2019; 9(4): 3194–3202. doi: 10.11591/ijece.v9i4.pp3194-3202
41. Khassanov Y, Mussakhojayeva S, Mirzakhmetov A, et al. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. *arXiv* 2009; arXiv:2009.10334. doi: 10.48550/arXiv.2009.10334
42. Abraham B, Goel D, Siddarth D, et al. Crowdsourcing speech data for low-resource languages from low-income workers. In: Calzolari N, Béchet F, Blache P, et al. (editors). *Proceedings of the 12th Language Resources and Evaluation Conference*; 11–16 May 2020; Marseille, France. pp. 2819–2826.
43. Beibut A, Darkhan K, Olimzhan B, Madina K. Development of automatic speech recognition for Kazakh language using transfer learning. *International Journal of Advanced Trends in Computer Science and Engineering* 2020; 9(4): 5880–5886. doi: 10.30534/ijatcse/2020/249942020
44. Polat H, Oyucu S. Building a speech and text corpus of Turkish: Large corpus collection with initial speech recognition results. *Symmetry* 2020; 12(2): 290. doi: 10.3390/sym12020290
45. Kirkedal A, Stepanović M, Plank B. FT speech: Danish parliament speech corpus. *arXiv* 2020; arXiv:2005.12368. doi: 10.21437/Interspeech.2020-3164
46. Lekshmi KR, Jithesh VS, Sherly E. Malayalam speech corpus: Design and development for dravidian language. In: Jha GA, Bali K, Sobha L, et al. (editors). *Proceedings of the 5th Workshop on Indian Language Data: Resources and Evaluation (WILDRE5 2020)*; 11–16 May 2020; Marseille, France. pp. 25–28.
47. Karpov N, Denisenko A, Minkin F. Golos: Russian dataset for speech research. *arXiv* 2021; arXiv:2106.10161. doi: 10.48550/arXiv.2106.10161
48. Akmuradov B, Khamdamov U, Djuraev O, Mukhamedaminov A. Developing a database of Uzbek language concatenative speech synthesizer. In: *Proceedings of the 2021 International Conference on Information Science and Communications Technologies (ICISCT 2021)*; 3–5 November 2021; Tashkent, Uzbekistan. pp. 1–5.
49. Mirishkar GS, Naroju MD, Maity S, et al. CSTD-Telugu Corpus: Crowd-Sourced Approach for Large-Scale Speech data collection. In: *Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2021)*; 14–17 December 2021; Tokyo, Japan. pp. 511–517.
50. Musaev M, Mussakhojayeva S, Khujayorov I, et al. USC: An open-source Uzbek speech corpus and initial speech recognition experiments. In: Karpov A, Potapova R (editors). *Lecture Notes in Computer Science Book 12997*, *Proceedings of the Speech and Computer: 23rd International Conference, SPECOM 2021*; 27–30 September 2021; Petersburg, Russia. Springer; 2021. pp. 437–447.
51. Adiga D, Kumar R, Krishna A, et al. Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. *arXiv* 2021; arXiv:2106.05852. doi: 10.48550/arXiv.2106.05852
52. Tiwari SA, Kanke RG, Maheshwari A. Marathi speech database standardization: A review and work. *International Journal of Computer Science and Information Security (IJCSIS)* 2021; 19(7): 92–97. doi: 10.5281/zenodo.5501910
53. Kuanyshbay D, Baimuratov O, Amirgaliyev Y, Kuanyshbayeva A. Speech data collection system for Kazakh language. In: *Proceedings of the 2021 16th International Conference on Electronics Computer and Computation (ICECCO 2021)*; 25–26 November 2021; Kaskelen, Kazakhstan. pp. 1–8.
54. Wang D, Wang L, Wang D, Qi H. DTZH1505: Large scale open source mandarin speech corpus. *Journal of Computer Engineering & Applications* 2022; 58(11): 295–301. doi: 10.3778/j.issn.1002-8331.2112-0333

55. Kumar R, Singh S, Ratan S, et al. Annotated speech corpus for low resource Indian languages: Awadhi, Bhojpuri, Braj and Magahi. *arXiv* 2022; arXiv:2206.12931. doi: 10.48550/arXiv.2206.12931
56. Veisi H, Hosseini H, MohammadAmini M, et al. Jira: A central Kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation* 2022; 56(3): 917–941. doi: 10.1007/s10579-022-09594-4
57. Zevallos R, Camacho L, Melgarejo N. Huqariq: A multilingual speech corpus of native languages of Peru for speech recognition. *arXiv* 2022; arXiv:2207.05498. doi: 10.48550/arXiv.2207.05498
58. Mussakhoyayeva S, Khassanov Y, Varol HA. Kazakh TTS2: Extending the open-source Kazakh TTS corpus with more data, speakers, and topics. *arXiv* 2022; arXiv:2201.05771. doi: 10.48550/arXiv.2201.05771
59. Avram AM, Nichita MV, Bartusica RG, Mihai MV. RoSAC: A speech corpus for transcribing Romanian emergency calls. In: Proceedings of the 2022 14th International Conference on Communications (COMM 2022); 16–18 June 2022; Bucharest, Romania. pp. 1–5.
60. Safonova A, Yudina T, Nadimanov E, Davenport C. Automatic speech recognition of low-resource languages based on Chukchi. *arXiv* 2022; arXiv:2210.05726. doi: 10.48550/arXiv.2210.05726
61. Lahoti P, Mittal N, Singh G. A survey on NLP resources, tools, and techniques for Marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing* 2022; 22(2): 1–34. doi: 10.1145/3548457
62. Baghdasaryan VH. ArmSpeech: Armenian spoken language corpus. *International Journal of Scientific Advances* 2022; 3(3): 454–459. doi: 10.51542/ijscia.v3i3.25
63. Wang Y, Wu M, Zheng B, Zhu S. HuZhouSpeech: A Huzhou dialect speech recognition corpus. In: Proceedings of the 2022 5th International Conference on Information Communication and Signal Processing (ICICSP 2022); 26–28 November 2022; Shenzhen, China. IEEE; 2023. pp. 153–157.
64. Yu T, Frieske R, Xu P, et al. Automatic speech recognition datasets in cantonese: A survey and new dataset. *arXiv* 2022; arXiv:2201.02419. doi: 10.48550/arXiv.2201.02419
65. Kulebi B, Armentano-Oller C, Rodríguez-Penagos C, Villegas M. ParlamentParla: A speech corpus of catalan parliamentary sessions. In: Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference; 20–25 June 2022; Marseille, France. pp. 125–130.
66. Solberg PE, Ortiz P. The Norwegian parliamentary speech corpus. *arXiv* 2022; arXiv:2201.10881. doi: 10.48550/arXiv.2201.10881
67. Kibria S, Samin AM, Kobir MH, et al. Bangladeshi Bangla speech corpus for automatic speech recognition research. *Speech Communication* 2022; 136: 84–97. doi: 10.1016/j.specom.2021.12.004
68. Mirishkar GS, Raju V VV, Naroju MD, et al. IIITH-CSTD corpus: Crowd-sourced strategies for the collection of a large scale Telugu speech corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing* 2021; 22(7): 1–26. doi: 10.1145/3600228
69. Mamtimin I, Du W, Hamdulla A. M2ASR-KIRGHIZ: A free Kirghiz speech database and accompanied baselines. *Information* 2023; 14(1): 55. doi: 10.3390/info14010055
70. Sun X, Cai K, Chen B, et al. Application of voice recognition interaction and big data Internet of Things in urban fire fighting. *Journal of Location Based Services* 2022; 16: 1–22. doi: 10.1080/17489725.2022.2096937