CLEAR Global

TWB Voice: Explainer

CLEAR Global's TWB Voice initiative

Thank you for your interest in TWB Voice, a data collection platform and initiative hosted by CLEAR Global! CLEAR Global is a non-profit organization with a mission to help people get vital information and be heard, whatever language they speak. This document provides information about TWB Voice, its aims, how the voice recordings collected will be used and your data protected, and how you can get involved. If you have any questions about the initiative, are interested to get involved and need support to sign up, or would like to speak to us about our work, you can contact translators@translatorswithoutborders.org.

A short glossary of terms used in these frequently asked questions

Here are some key concepts which you will find in the explanations:

Automatic speech recognition means computer software that automatically converts spoken words into written text; it is also called speech-to-text.
CLEAR Global is a non-profit organization working to help people get information and be heard, whatever language they speak. Translators without Borders is a component of CLEAR Global; the other components are CLEAR Tech, and CLEAR Insights.
Dataset means a collection of information organized for use. A voice dataset is a collection of voice recordings and related information organized for use in research, training, or improving voice models.
Language model means a computer program capable of processing human language in text or voice; a voice model is a language model for voice.
Language technology (or language AI) is the use of computer systems and software to understand, interpret, and generate human language. It powers applications like chatbots, voice recognition, and machine translation.
License means the legal terms governing how the voice model can be used, modified, and distributed by others.
Open source means making the voice model publicly available under specific terms that allow others to view, use, and potentially modify it, while still being bound by the license terms.
Text-to-speech model is a type of voice model that is capable of generating natural-

sounding vocal output from a text either by mimicking specific voices or producing new voices.

Training means the process of teaching a computer program or voice model to recognize and reproduce patterns in voice data. This involves using large datasets of voice recordings and transcriptions to help the model improve its ability to perform tasks, such as converting text to speech or speech to text.

Transcription is the process of converting spoken words into written text.

TWB (Translators without Borders) is the non-profit organization managing the TWB Voice platform; it is part of CLEAR Global.

TWB Community is TWB's global community of over 100,000 members helping people get vital information and be heard, whatever language they speak.

Underrepresented language means a language which has little or no presence online; it is sometimes called a 'low-resource' or 'marginalized' language.

Voice data means any audio recordings, transcriptions, or related data provided by a contributor in TWB Voice.

Voice data collection is the process of collecting audio recordings, transcriptions, and related data to use to develop or improve voice models.

Voice model means a computer program used to process and generate human voice. It uses complex data stored in a way that humans can't read, and it doesn't include actual recordings of the person's voice. The voice model cannot be taken apart to recreate the original voice or the recordings of the person speaking.

Voice technology is language technology that processes or generates spoken communication.

Questions about the TWB Voice platform and what it is used for

What is TWB Voice?

TWB Voice is a platform owned by CLEAR Global/Translators without Borders, which is designed to collect voice recordings from many users. These recordings are used to create large voice datasets, which are essential for developing language models for language technology applications like text-to-speech and automatic speech recognition. The datasets, made up of recordings from tens or hundreds of voices, are also shared openly to support research and technology development for underrepresented languages.

The primary goal of TWB Voice is to support the development of voice technology for marginalized language speakers. For millions of people worldwide, language barriers stop people from accessing digital services. By developing voice technology for underrepresented languages, we want to help bridge this gap, making digital tools and

services more accessible. Central to this mission is data collection for these languages—an essential step in developing language technology. TWB Voice brings together the TWB Community, a global community of over 100,000 linguists, to contribute the data required to achieve this goal.

How do I use the TWB Voice platform?

Through TWB Voice you can complete four types of task:

1. Recording your voice
2. Rating the recordings of other community members
3. Transcribing voice recordings
4. Rating the transcriptions of other community members

Together these tasks will generate voice datasets—collections of voice recordings paired with their text and additional information. These may be published as datasets for use by others. In some cases these datasets may also be used by TWB's own team to create language technology models.

You will access projects in TWB Voice through the TWB Platform. You can think of this as the "hub" of Translators without Borders, from where you can access tasks, visit our learning center to complete our free e-learning courses, and collaborate with other community members on the TWB Forum. Your TWB Voice account is linked to your TWB Platform account.

How do I sign up? Do I need to have experience in voice recording?

If you are not already a member of the Translators without Borders community you will need to first create an account in the TWB Platform and make sure to set your native language as Hausa, Kanuri, or Shuwa Arabic (the three languages of our current project).

Once you have a TWB Platform account and your native language is set, you will see a banner inviting you to sign in to the TWB Voice platform. You can also access this at twbvoice.org. Once there, you will be invited to provide additional details and sign in to your TWB Voice account where you will find available projects on the "My Projects" page.

No previous experience is required! The platform includes guidelines on how to make a quality recording, and walks new users step by step through how to use it. We are also very happy to support new users with questions or demo sessions by contacting us at translators@translatorswithoutborders.org.

How are my privacy and my data protected on TWB Voice?

Voice data collection is governed by privacy and data protection regulations. All TWB Voice users have rights under certain laws such as the General Data Protection Regulation in the EU. This means that, regardless of where you live, you have the right:

- to know what personal information we store about you
- to object to how your personal information is being processed and to ask for your data to be corrected or deleted
- to request your personal data in a format that can be transferred to another organization

Your voice is also considered personal information because it is unique to you and can reveal personal details about you such as your gender, age, or region of origin. This means it can potentially be used to identify you, similar to a name, photograph, or fingerprint. For this reason, your voice is treated with the same level of care and protection as other forms of personal information.

How will my recordings, transcriptions and ratings be used?

The aim of TWB Voice is to advance voice technology in a way that benefits marginalized language communities. By contributing to our voice data collection platform, you are helping to create valuable datasets that will support the development of voice technologies for languages that are often underrepresented.

The projects to which you will contribute in TWB Voice aim to collect voice data, with the main aim of:

1. Creating voice datasets - collections of voice recordings with transcriptions and data that researchers and developers can use to build and train language technology in these languages.
2. Using these datasets to create and train language technology models including text-to-speech and automatic speech recognition models.

The current project we are running aims to collect voice data in three languages: Hausa, Kanuri, and Shuwa Arabic. We aim to collect 50-100 hours of voice data per language, with no more than 2 hours donated by an individual contributor, to ensure diversity in the dataset. The current project will finish at the end of May-early June. With these datasets, we will then create text-to-speech and automatic speech recognition models in Hausa and

Kanuri, and publish these openly to support language technology development in these languages. We aim to do the same under a second phase of this project and also develop TTS and ASR models for Shuwa Arabic. Creating a voice dataset in Shuwa Arabic is a foundational step toward this.

While the models and datasets will be published openly for use by researchers, practitioners and communities to improve language technology, CLEAR Global is working with Hausa, Kanuri, and Shuwa following extensive work to improve multilingual communication in northeast Nigeria over the last decade and having seen the potential for voice technologies in the region. For example, in 2023 CLEAR Global developed a conversational AI chatbot to provide information to community leaders and women in northeast Nigeria. Users could ask the chatbot written questions in three local languages - Hausa, Kanuri, and Shuwa Arabic - and get written answers. By collecting voice data from community members in these languages, we can train voice models so users can ask spoken questions of the chatbot and receive spoken answers. A voice-enabled chatbot is accessible to people with lower literacy levels and can have a wider impact.

How will my voice data be published?

The voice data you contribute will be published together with voice data from other contributors as a voice dataset - a collection of voice recordings and related information, like transcriptions, used to develop technology. The data we publish will include your recordings and transcriptions, as well as the demographic information (year of birth, gender, education level, language variant) which you provide in the platform linked to each recording. The data will be anonymized within the datasets; identifiers like your name, email address and username will not be included. The voices of multiple contributors will be included within a dataset.

We will publish the datasets on open platforms which developers and researchers can access to share and use datasets to advance language technology.

We will make the TWB Voice datasets available under a Creative Commons license. This allows others to freely use, share, and build upon the data but requires them to recognize the original dataset as their source. This license also requires anyone using the TWB Voice datasets to license whatever they use the data for under the same conditions.
You can still request TWB to delete your data. See the section “How can I ask for my data to be edited or deleted?” for more information on how to do this.

How will the voice models built with my voice data be used, and will my voice be identifiable?

The datasets you contribute will be used to build models for two types of language technology: text-to-speech (TTS) and automatic speech recognition (ASR). These models help convert written text into spoken words (TTS) and spoken words into written text (ASR), enabling a range of applications to function with spoken input.

We will also publish the data and models on open platforms where developers and researchers can access them to enhance and build upon them.

Text-to-speech (TTS) models convert written text into voice. The voice is typically produced by a computer and mimics the original audio it’s trained on. It is possible that someone may recognize a voice that sounds similar to yours with this model. These models can be used in applications like voice assistants, audiobooks, screen readers, or automated phone lines.

Automatic speech recognition (ASR) models recognize spoken words and convert them into written text. In an ASR model your voice will not be heard; it will only be used to train the model to recognize someone speaking the language of your recordings. Uses of ASR models in humanitarian response might include helping emergency response communications by transcribing emergency calls in real time, and allowing speakers of marginalized languages to speak to and be understood by a computer on topics such as healthcare and receive real-time responses.

What are the benefits of participating in the TWB Voice project? Will I be compensated?

Participation in the TWB Voice project is entirely voluntary. However, to acknowledge and support the efforts of our contributors, we offer a monetary recognition scheme based on the level of your contributions. This is designed to help cover any costs you may incur while participating—such as data expenses.

Beyond financial recognition, your contributions have a meaningful impact. By donating your voice, you're helping develop open voice datasets and technologies that can improve communication marginalized communities. While these resources will be openly available for the wider tech community, CLEAR Global is working to secure funding to apply them in real-world contexts—especially for humanitarian accountability and information access in northeast Nigeria, where we’ve been supporting communities for many years.

Participants will also have opportunities to help shape the platform itself and engage in community-building initiatives, such as training sessions, voice marathons, and other events. We also welcome your feedback on other initiatives you would like to participate in or lead.

What are the risks of sharing my voice data?

Contributing your voice data involves several potential risks. While TWB will make every effort to reduce these risks, here are some risks it is important to understand before contributing your voice:

- While your voice data will be anonymized and published only with the demographic information you provide, it is possible that someone could recognize your voice or identify you through this information.
- Once a dataset is shared openly, others can use this data for their own research and technology development.
- A text-to-speech model trained on your voice data can ‘sound like your voice’. As the model automatically generates a voice, the model can be used to say things which you did not record. While TWB will take all reasonable measures to prevent misuse of any models trained on your voice, we cannot 100% guarantee prevention of misuse of the voice model by others.

How will TWB reduce the risks associated with sharing my voice data?

TWB aims to reduce these risks by:

- Protecting your personal information on secure servers
- Not publicly associating any voice models published with your identity
- Ensuring appropriate licensing terms for any voice models developed
- Technical protection measures where feasible, like collecting written information from anyone wishing to use voice models developed on their identity and the intended use, and retaining the right to revoke access

How can I help reduce the risks associated with sharing my voice data?

You can help to reduce these risks by:

- Making sure not to include any personal information in the voice recordings you make

- Finding a private place to make your voice recordings so these do not unintentionally capture any personal information relating to you or other people in the background
- Contributing to an overall commitment among the TWB Community to protect each others' privacy by always reporting issues within TWB's platforms. For example, if you rate someone else's recording and you notice this includes personal information, you should report this as an issue through the platform's report function or to your Project Officer.

Questions about privacy and data storage

What personal information will be collected in TWB Voice?

Your TWB Voice account is linked with the account you created on the TWB Platform. The username, email address, and information on native language provided in your TWB Platform account will be imported into TWB Voice.

In addition to the information already given you will be asked to provide some additional demographic information on TWB Voice including your year of birth, gender, education level, and language variant.

How will my personal information collected in TWB Voice be used?

This information will be published within datasets including your voice. This demographic data can help us and other researchers to improve the quality of the language technology we create by ensuring the data represents a diversity of speakers. In any datasets published your data will be anonymized with just your voice and demographic information provided.

How is my personal information stored and protected?

All information you provide to us is stored on secure servers. TWB is committed to take all reasonable steps necessary to ensure your personal data is processed securely and in line with TWB's Privacy Policy. Unfortunately, no data storage system can be guaranteed to be 100% secure. If you have reason to believe that your interaction with us is no longer secure, please contact us right away at the address translators@translatorswithoutborders.org.

Will I receive email notifications about TWB Voice?

You may receive notifications from the TWB Community Team and TWB Project Officers about TWB Voice. These emails may include:

- information about new projects and tasks available in TWB Voice,
- questions or feedback about tasks you have completed,
- invitations to provide feedback,
- information about changes to the platform, or supporting resources we think you might like to know about!

We will notify you if there is a data breach that affects your personal information.

Will other TWB Voice users see any of my personal information?

Other community members completing tasks in TWB Voice will be able to listen to your voice recordings when reviewing these. No other personal information will be displayed during these review tasks. Your username may also appear in a leaderboard published in the TWB Voice platform, which shows the most active community members in the platform.

Only community members who have also agreed to our policies and TWB staff will be able to access TWB Voice and view tasks and the TWB Voice leaderboard.

How can I ask for my data to be edited or deleted?

Your participation in TWB Voice is voluntary. You have a right to request the deletion of your data, including your voice recordings, at any time.

Please contact us at info@translatorswithoutborders.org if you would like to ask us to edit or delete your personal information, object to the processing of your personal information, or to request an electronic copy of this information in a format that can be transferred to another organization. We will respond to your request consistent with applicable law.

If you request us to delete your data, you can choose to:

1. request to delete your TWB Voice account and any record of your username or email address on TWB Voice. This means that any tasks you have completed and the demographic data (age, gender, language variant, education level) attached to this task will remain on TWB Voice and in any datasets published. After deleting your account it will not be possible to request deletion of your voice data at a later stage

as we will not retain your personal information to trace this.

2. request to delete your TWB Voice account, as well as any tasks you have completed on the platform. If the dataset which includes your voice has already been published, we will also remove your data from the published dataset. We cannot guarantee that anyone who has already downloaded the dataset will delete your data. If you choose to create a new account on TWB Voice after requesting this, your points and task contributions will start again from zero.

If your request for data deletion includes other TWB platforms beyond TWB Voice, please specify this in your email.

How can I report copyright infringement?

If you believe that any content included in TWB Voice infringes your or some else's copyright, please email us at info@translatorswithoutborders.org. You should include your contact information, a description of the copyrighted work, a screenshot or other evidence of where you found the copyrighted work in the platform. We will review your email and may request further information before taking appropriate action, such as removing the content.

How can I report misconduct?

TWB takes accountability seriously. If you have concerns about the conduct of any TWB staff member you can make a report. Visit this page on reporting misconduct and fraud for details of how to make a report.

How can I access these explanations and when will they be updated?

If you sign up to TWB Voice, these explanations will be available within the platform. We may also share these explanations with potential contributors by email.

We may update these questions from time to time. Any changes will become effective when we post the revised questions on TWB Voice.

Who can I contact if I have questions?

If you have questions about these explanations, you can speak to us at translators@translatorswithoutborders.org.