Collecting Consented Speech in the Languages of the Sahel: Design of a Mobile Programme

Namu AI Lab
namuai-lab.com

Abstract

The main obstacle to speech technology for the languages of the Sahel is data. Our earlier explorations measured how little there is: of about 1,060 hours of public Zarma audio, about 10.7 hours are both openly licensed and transcribed, all from seven speakers; open transcribed speech in the Fulfulde of Niger comes to about 12 hours; there is no public conversational, telephone or children's speech in Zarma and no public test set with many speakers; and a recogniser fine-tuned on five speakers memorised their voices within about a thousand updates while its error on a new speaker stopped falling. We describe the programme designed in response. Adults record short spoken missions in Hausa, Zarma, Tamasheq and Fulfulde in a collection app on the phone and the web, under a versioned agreement bound to every recording. Recordings travel in small hashed units and count as received only when the server has read back the exact stored bytes. Evaluators who passed an identity check and a listening test judge each recording without knowing who spoke; two agreeing judgments decide, and a withdrawal is designed to reach every stored copy, datasets included. Speaker-balanced test sets of 100 to 110 speakers are to be recorded in assisted sessions under evaluation-only consent, and voices for synthesis in a studio under their own licence. Every set will carry a data statement and a frozen version. The programme is opening: no contributor's speech has been collected, the app has been tested only on a development deployment with synthetic accounts, and several parts are plans, which we mark. We report the design, the reasons for it, and its limits and risks.

1 Introduction

In Niger, Hausa, Zarma, Fulfulde and Tamasheq are among the languages people speak every day, and French is the language of the state and of most written material. Many people read no written language at all, and many reach the internet through a low-cost phone on a weak and costly connection. Speech in their own language is the natural interface for them, and it is the interface that current speech technology serves worst.

Our earlier exploration of Zarma [14], and a similar survey for Fulfulde, located the problem in the data. Hours of public audio exist, but almost none of them are both openly licensed and transcribed, they come from very few speakers, they are read or broadcast rather than spoken in everyday exchanges, and no public test set lets anyone measure a model on many speakers kept apart from training. Models trained on what exists learn the few voices in it. Synthesis needs something more: a voice whose owner agreed that a machine may speak with it.

Technical report, Namu AI Lab, November 2026. This report describes research systems, not products.

Namu therefore designed a programme to record the speech that is missing, with the people who speak it. Its centre is a collection app on iPhone and the web (an Android version is being ported) in which adults record short spoken missions, and other speakers listen to and transcribe each recording. Around it sit two assisted activities: recording speaker-balanced test sets with a facilitator present, and recording voices for synthesis in a studio. The programme is opening, so there are no collection results to report; in the spirit of dataset papers, we give the design, the reasons for each choice, the evidence behind it, and what is built and what is planned (Table 1).

Our contributions are the following. (1) An account, from measurements, of what public speech data for Niger's languages lacks, and the requirements that follow (Section 3). (2) The design of a mobile collection app: missions that ask people to speak rather than read, consent bound to every recording, recording and upload that survive weak networks, and blind review by people who are themselves checked (Sections 5 to 9). (3) A design for speaker-balanced, speaker-disjoint test sets recorded under evaluation-only consent (Section 10). (4) A separate route for voices for synthesis under their own licence, and the data statement, versioning and intended uses of every set (Sections 11 and 12). (5) The limits of the design and its risks, including fair pay, consent, who is represented and misuse of voices (Section 14).

2 Related work

Volunteer and community collection. Common Voice gathers read sentences from volunteers through a web app and validates each clip by the votes of other volunteers [1]. Recent large collections for African and Indian languages record spontaneous speech prompted by images and topics, and organise speakers through their communities [6, 19]. AfriSpeech-200 collected African-accented English for clinical and general recognition [17], and participatory research showed that the people who speak a language can drive the creation of its resources [15]. We borrow image and topic prompts and add, inside the same app, review by evaluators who are themselves checked and are to be paid.

Speech data for African languages. Massively multilingual recognisers extend to more than a thousand languages, often from recordings of religious texts with one reader per language [18]; synthesis corpora have been built from Bible recordings [13] and from found and newly recorded data [16]. For Niger, a large radio-based recognition dataset covers several national languages, with a small human-corrected part [11]. FLEURS reads the same sentences in about a hundred languages for evaluation [4], and our Zarma report measured what public resources allow for one language of Niger [14].

Documentation and governance. Datasheets [7] and data statements [2] ask dataset creators to record who is in a dataset, how it was collected and what it should not be used for, and accountability frameworks treat a dataset as an engineered artefact with requirements, versions and maintenance [8].

Annotation quality and evaluation. Estimating each annotator's error rates from their answers is an old idea [5]; we measure evaluators directly on clips with known answers. Bootstrap intervals for recognition error resample whole utterances [3]; when errors depend on the speaker, the speaker becomes the unit, which is why the number of test speakers matters. Listening tests follow standard protocols for mean opinion scores [9] and for comparing several systems against a hidden reference [10]. Recognisers can perform very differently across groups of speakers [12], which a test set can only show if those groups are in it.

Misuse of voices. Public challenges measure how well synthetic and replayed speech can be detected [21], and watermarks embedded in generated audio let a provider recognise its own output [20].

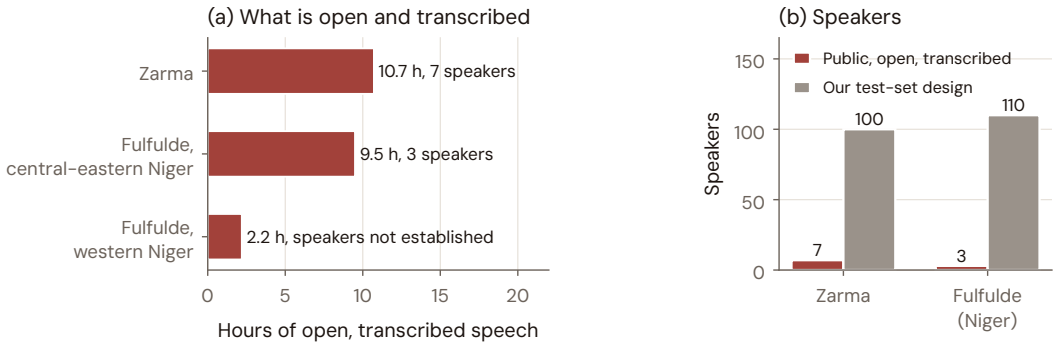


Figure 1: What exists and what we plan. (a) Hours of openly licensed, transcribed speech in Niger varieties, as surveyed in our explorations; for the western Fulfulde corpus only the usable part is counted, and its speakers were not established. (b) Speakers in those public sets (for Fulfulde, the central-eastern corpus only) against the speakers in our test-set designs (Section 10), which are plans.

3 Why collect: what the explorations show

Few speakers. We counted about 1,060 hours of public Zarma audio [14]. About 10.7 hours of it are both openly licensed and transcribed, roughly 1%, and all of that is prompted spontaneous speech from seven speakers. For Fulfulde, about 268 hours of open transcribed speech exist across all varieties, but only about 12 hours are varieties of Niger: about 9.5 hours of the central-eastern variety from three speakers, and about 2.2 usable hours of the western one (Figure 1). No test set for the Fulfulde of Niger exists anywhere we looked.

Missing kinds of speech. We found no openly licensed Zarma corpus recorded to the standard a synthesis model needs, no conversational or telephone speech, no speech from children, and no public test set with many speakers kept apart from training [14]. Hausa has more data, but its public evaluation speech is narrow: in the Hausa test set of FLEURS, 620 of 621 utterances are spoken by women.

Models learn the voices they hear. We fine-tuned a public medium-size recogniser on about seven hours of open Zarma speech from the five training speakers of the public corpus. Figure 2 shows what happened. Training loss fell from 125 to 37, while the loss on a held-out speaker rose from 153 to 221: the model memorised the five voices within about a thousand updates. Its error on that held-out speaker fell from 29.0% to 25.7% folded character error rate (CER), then stayed within 0.2 points from 2,000 to 5,000 updates, level with the best zero-shot recogniser (25.5%). On the test speaker the fine-tune went from 27.1% to 24.2%, which ties the zero-shot model that was best on the held-out speaker and does not beat it [14]. A synthesis model fine-tuned on 3.6 hours of read Zarma from five speakers peaked at its first checkpoint. With five speakers, longer training added little, so the programme aims first at the number of speakers.

Measurement needs many voices. Our Zarma recognition test set is one speaker of a public corpus, so its intervals describe that speaker and that transcriber, not Zarma speakers in general [14]. Checking that corpus also showed that references can be wrong: ten development references belonged to another speaker's recording, and 27 test references were shorter than their audio. Since then we check every reference against its audio with a recogniser before a set is frozen. Labels need care too: on our French training data, a common pitch-based estimator put the share of women's voices at about 14% and a more careful one at 34%, so we ask speakers rather than estimate.

Synthesis needs consented voices. A voice-cloning model copies a speaker's timbre, rhythm and even the pauses in a reference clip: our first reference clip for one consenting team member was 28% silence, and the renders paused in the middle of sentences [14]. A voice a product speaks with must

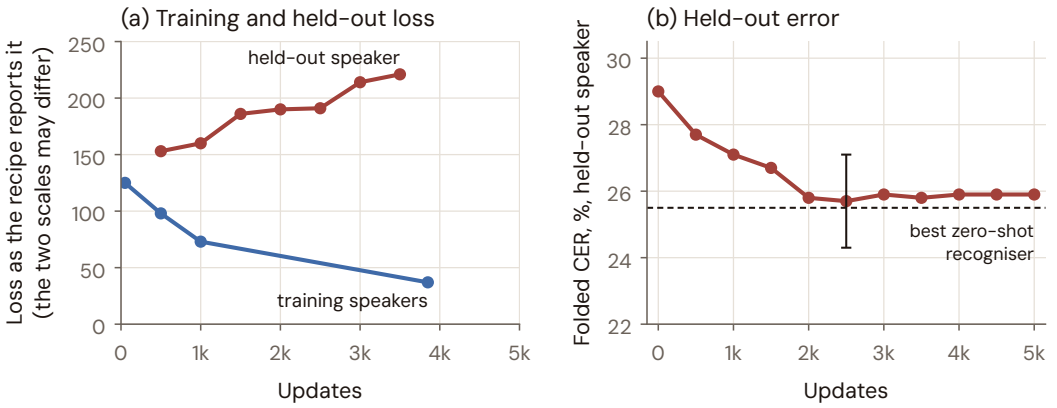


Figure 2: A public recogniser fine-tuned on about seven hours of Zarma from five speakers. (a) Training loss and loss on a held-out speaker, as the training recipe reports them; the two may be on different scales, so only their directions should be compared. (b) Folded CER on the held-out speaker (486 segments) at each saved checkpoint, with the 95% interval at the selected one and the best zero-shot recogniser on the same speaker. One held-out speaker and one transcriber: indicative only.

come from a person who agreed to exactly that use, which public corpora do not provide.

Requirements. These findings set the requirements for the programme. It must reach many speakers across regions, ages and genders (R1), and record everyday, conversational and phone speech as well as reading (R2). Test speakers must be kept apart from training from their first recording, and test sets frozen before any model is trained on the language (R3). Consent must name each use, with voices for synthesis under a separate licence (R4). Recording must work on low-cost phones and weak networks (R5). Review must be done by people who speak the language, are themselves measured, and judge without knowing whom they hear (R6). A withdrawal must reach every copy, including datasets (R7).

4 The programme in outline

Roles. Everyone has one account, the same on the phone and the web. A *contributor* is any adult who accepts the agreement and records missions. An *evaluator* qualifies in a language after an identity check and a listening test, judges other people's recordings and, with a further test, writes or corrects their transcripts. An *expert*, appointed by invitation and also identity-checked, settles recordings held by a flag, writes the final transcript when two writers disagree, audits a sample of decisions and confirms the reference clips used to measure evaluators.

Languages. Contributors record in Hausa, Zarma, Tamasheq or Fulfulde. The interface is in French, with spoken guidance in Hausa; guidance in the other languages does not exist yet. Niger comes first.

One recording's journey. Figure 3 follows one recording from the mission to a versioned dataset, and shows where a withdrawal reaches. The sections that follow take each station in turn.

What is built. Table 1 separates what runs today from what is designed. No contributor's voice has been collected, no reward amount is set, and no model has been trained on these voices.

5 Missions: what people are asked to say

Contributing should not feel like filling in a form, so the app presents collection as a journey. A contributor moves along a path of missions with a small animated companion as a guide; points mark progress, and taking a break is never penalised. The catalogue has six worlds, 18 sections, 57

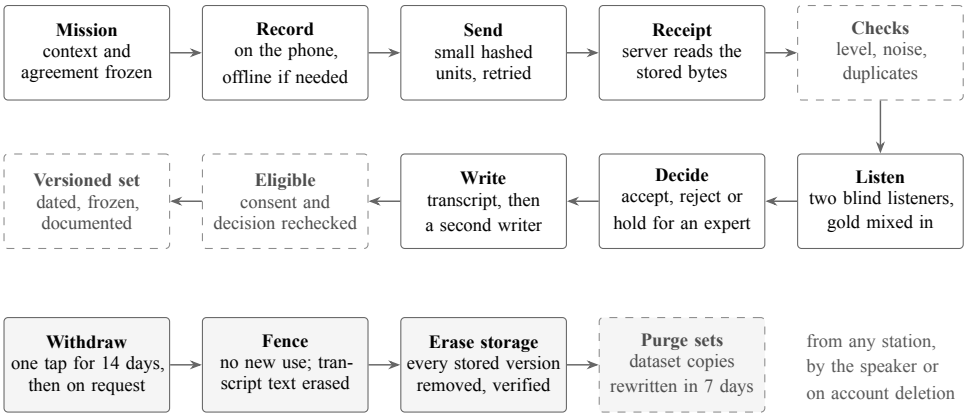


Figure 3: The life of one recording. Top two rows: from the mission to a versioned dataset; bottom row: what a withdrawal reaches. Solid boxes are built and tested on a development deployment with synthetic accounts; dashed boxes are planned.

Table 1: What is built and what is planned, as of November 2026. "Built" means implemented and tested on a development deployment with synthetic accounts and recordings; collection with the public has not started.

Component	State
iPhone app: offline recording, agreement, receipts, withdrawal, account deletion	Built; internal test build
Web app: recording, upload in small units, receipts, withdrawal	Built
Android app	Being ported
Listening review with gold clips, consensus, recusal and safety flags	Built
Evaluator levels, expert desk, home regions, shared-device links	Built
Transcription and verification; automatic drafts	Built; drafts off by default
Three tries per mission, then a pause	Built
Reward ledger (separate from points)	Built; payments off, amounts not set
Identity check by a document service	Planned; a staff step today
Automatic server checks, compressed previews for evaluators	Planned
Muting of personal information; a separate identity store	Planned
Dataset export, splits, versions and purge	Planned
Pair and group missions; assisted test-set sessions; studio voices	Planned

units and 225 missions (Figure 4; Appendix Table 6).

Spoken prompts. No published mission asks anyone to read. A mission shows a picture, a pair of objects or a short scenario and asks the speaker to talk about it in their own words: name an everyday object and say how it is used, compare two shapes, choose between two containers and say why, imagine a new and harmless use for an empty box, tell a short story. Seven formats of this kind make up the 113 missions playable by one speaker, published in Hausa, Zarma, Tamasheq and Fulfulde as 452 mission revisions. Recordings are short by design: the web app stops at two minutes, while the server and the iPhone app accept up to ten minutes for a published mission. Reading is reserved for the test sets and the studio, where the text is controlled (Sections 10 and 11).

Privacy by design of the prompt. The prompts steer away from personal information: 98 of the 225 missions ask for fictional content, and only 6 invite personal stories, always with another mission offered instead. The agreement asks speakers to share no private information, and the text of most published missions repeats it (Section 6).

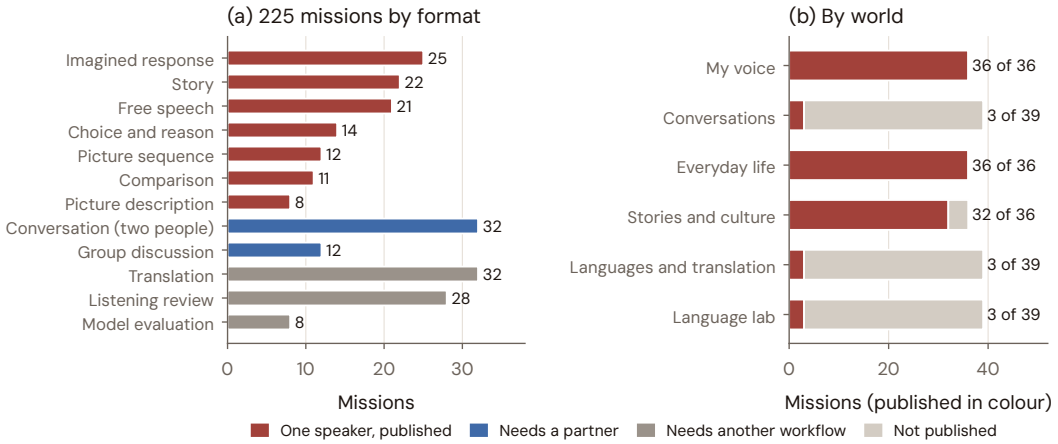


Figure 4: The mission catalogue. (a) The 225 missions by format and by what they need to run: 113 are playable by one speaker and published in four recording languages; the others need a partner, a language or review workflow, or a model, and are not yet activated. (b) Missions per world, published ones in colour.

What is not yet open. The catalogue already holds 32 two-person conversations and 12 group discussions, each requiring every participant to be a consenting adult; 32 translation missions (40 missions in all ask for two languages); 28 listening reviews; and 8 missions in which a speaker makes a request to a model and judges whether it was understood. None of these is activated yet.

Frozen context. Before the microphone opens, the app freezes the mission's full context: the catalogue version and its hash, the activity, the language, the choices the speaker made and the order of the cards they saw. A recording made offline keeps this context and is matched to the server's published mission only when the phone reconnects. Every recording therefore carries exactly what was asked.

6 Consent, privacy and withdrawal

The agreement. Before the first recording, a signed-in adult accepts the contribution agreement; its Hausa version can be heard in the app. It states that the person takes part voluntarily and is at least 18; that they share only their own voice, with no private information and no other person's voice; that Namu keeps the recordings in the United States for its research and its artificial intelligence products, including commercial ones; that the recordings are not published automatically and have no automatic expiry; that a recording can be withdrawn directly for 14 days after receipt and afterwards by a request with a reason field; that a request excludes the recording from new uses while it is processed; that deactivating an account keeps contributions, while deleting the account includes them and shows as in progress until their deletion is verified; and that deletion is not guaranteed to undo results of models already trained. A contact address for privacy and deletion requests is given.

Bound to every recording. The agreement is stored as an immutable, published revision. Acceptance is recorded online against the exact account and revision, and each recording references the consent grant it was made under. A new use of data needs a new reviewed agreement, never a wider reading of an old one. Offline, the app records only under a consent the server has already acknowledged; no new agreement can be accepted offline, and a revoked grant stops recording.

Scopes. Table 2 lists the agreements. Today there is one, for the app. The plan adds three: evaluation-only consent for test-set speakers, a training consent for assisted collection, and a licence for studio voices. Synthesis is the use we treat with most care. Recordings made under the current app agreement will enter recognition sets but not synthesis sets; a synthesis set needs an agreement revision that names synthesis, and no consent package other than the voice licence ever allows a

Table 2: Agreements and what each allows. Only the first exists today; the others are designed for the assisted sessions and the studio.

Agreement	Who	Allows	Never
App contribution (built)	Adults recording missions	Namu research and AI products, including commercial; kept in the United States; not published automatically	Another person's voice; private information; synthesis sets
Evaluation only (planned)	Test-set speakers	Testing Namu's models in a frozen private test set	Training; a product voice
Training (planned)	Assisted collection	Training and testing, including products; multi-speaker synthesis training as a separate choice	A product voice; copying the voice
Voice licence (planned)	Studio voices	Recordings and a synthetic voice in Namu's products and API, within listed limits	Any use outside the licence

person's voice to be copied.

Who sees what. Each person has a profile with contact details and a separate pseudonymous contribution account; review and data work see only the second. Review is blind: no evaluator or expert is told who spoke, their account, region or device, or where the audio is stored. A listener who recognises a voice presses a button labelled "I know this voice": the clip goes to someone else and the pair is never matched again. Two accounts seen on the same phone or browser are linked and never review each other; the web app sends a random installation identifier (the iPhone app is to do the same), and the server keeps only a keyed hash of it. In the design, an identity check keeps no document image, birth date or address, only its outcome, a few facts about the check and a keyed hash of the document number, in a separate store the main system cannot read; staff hear nothing by default, and access to original audio needs a recorded reason and two approvers.

Sensitive content. Evaluators can flag a recording for personal information, a person at risk, a child's voice, missing consent or harmful content. A flag holds the recording for an expert with no decision; a person-at-risk flag also opens a safety case for staff. In the design, confirmed personal information is muted in the training copy, by replacing the span, padded by 150 ms, with noise at the clip's own noise floor, and the unmuted copy is erased 30 days after the decision. A confirmed child's voice, missing consent or harmful content leads to erasure. Muted clips may enter recognition sets but never synthesis sets.

Withdrawal and deletion. Table 3 shows how far a withdrawal reaches. The built parts act at once, in the request's transaction; a cleanup job then removes every stored version of the audio and verifies that none remains, leaving a zero-byte marker so that an upload already under way cannot bring it back. Account deletion is reported as complete only after that verification. The planned parts carry the withdrawal into derived files, datasets and future training runs.

7 Recording that survives weak networks

Capture. The phone records in its native format and seals the audio in independent segments, which are checked before the app counts a recording as saved; an interruption keeps what was recorded and never restarts the microphone by itself. The web app records the raw microphone with echo cancellation, noise suppression and automatic gain turned off, because these change speech in ways that harm recognition data, and writes 16-bit mono audio at 24 kHz without lossy compression.

Table 3: Where a withdrawal reaches, and when.

Target	Mechanism	State
New uses and review	Fenced in the request's transaction; audio no longer served	Built
Transcript text, all versions	Erased in the same transaction	Built
Reference copies used for evaluators	Retired at once	Built
Original audio	Every version removed and verified; a zero-byte marker stays	Built
Previews and derived files	Removed and verified, within 72 hours	Planned
Dataset versions	Member revoked at once; shards rewritten within 7 days	Planned
Training runs	Each run loads the list of revoked clips first	Planned
Weights already trained	Not undone; the agreement does not promise it	

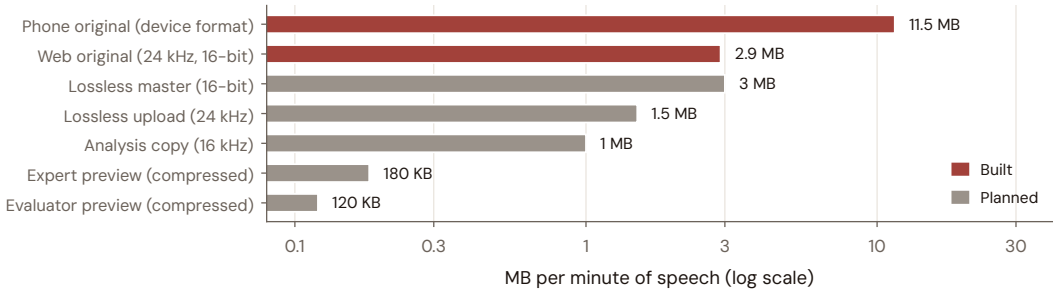


Figure 5: What one minute of speech weighs, by artefact (log scale). The phone original depends on the device format. Built: what the apps record today; planned: the formats of the processing design.

Small units, frozen first. The web app cuts each recording into units of at most 60 seconds and 4 MB, so that a failed send on a weak network repeats little. Before any network call, the app freezes the recording's identity, its mission context, its audio facts and a manifest with the SHA-256 hash of every unit, and keeps the audio on the device until the server confirms receipt. A send that fails is retried with a delay that starts at 15 seconds and doubles up to 5 minutes, and again when the page regains focus, when the network returns and on the next visit; on iPhone, each unit has up to 32 tracked attempts, each with a fresh upload authorisation that lasts five minutes. Uploads run while the app is open; we do not promise delivery while the phone has suspended it.

Receipts. A successful upload is not a receipt. The server reads back the exact stored version of each unit, checks its length and hash, decodes the audio and only then issues a receipt. The app shows four distinct states: on this phone, sending, checking receipt, and received.

Data cost. People pay for mobile data, so we budget the bytes of each artefact (Figure 5). The web original is about 2.9 MB per minute; a phone original in its device format is about 11.5 MB. A planned lossless format at 24 kHz would bring uploads to about 1.5 MB per minute. Uploads stay lossless: in a small internal check on only two real Hausa recordings, compressing the audio to 24 kbit/s changed 17% of the words a recogniser returned, and 16 kbit/s changed 27%; on clean synthetic speech the change was about 3% at either rate. Evaluators will receive compressed previews of about 120 KB per minute, fetched once per task; before that rate is fixed, a 50-clip native check against a higher rate will listen for the hooked consonants of Hausa and the nasal vowels of Zarma.

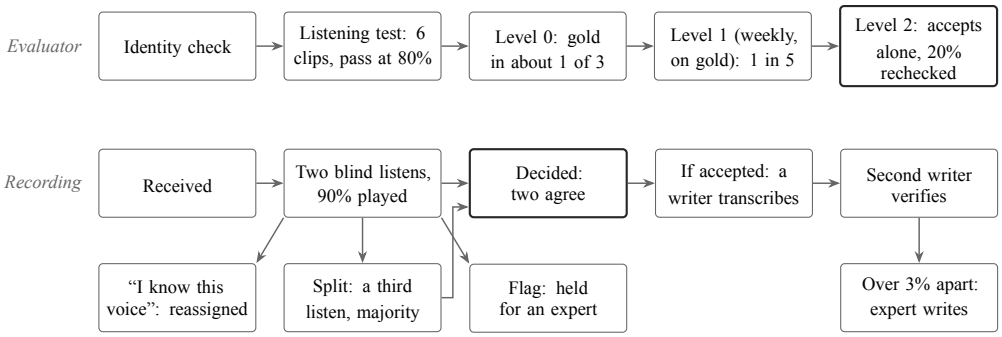


Figure 6: Review. Top: how a person becomes an evaluator and gains trust from clips with known answers (gold). Bottom: how one recording is decided and transcribed. A rejection always needs two people. Built and tested on a development deployment.

8 Review by people

Every recording is judged by people who speak its language, with three goals: decisions fair to the speaker, evaluators measured continuously, and review that stays private (Figure 6).

Becoming an evaluator. A candidate's government identity document is checked (by staff today, by a document service in the design, with in-person checks planned for people without documents), and they apply in a language. They then take a listening test of six gold clips, two of which should be refused. A pass needs at least five answers and 80% right; answering "usable" to everything scores four of six and fails. After a failure, the next test opens 24 hours later with clips the person has not heard. The tests are unpaid.

Work batches. An evaluator takes a batch of up to eight recordings with gold clips mixed in, about one for every three recordings at the first level, and never a batch of gold alone; a batch expires after four hours, and a person holds one batch at a time. Some filters are never relaxed: the right language and level, a recording that is still received and undecided, never one's own voice, never a recording heard before, and never a speaker one has recused or is linked to. Within these, a soft score prefers speakers from a distant region, using a self-chosen home region (one of Niger's eight, or abroad; no location is read). The server serves the stored audio only to the person holding the task, stops serving it at once if the recording is withdrawn, and checks that at least 90% of each clip was played and that the judged audio fits in the time since the batch began.

Decisions. Two agreeing judgments decide. A split adds a third evaluator, and the majority decides. A safety flag holds the recording for an expert with no decision. A rejection carries one of six reasons (too short, inaudible, wrong language, off task, cut, noise). After three rejections of the same mission, that mission rests for seven days for that person; nothing earned is taken back, and a pattern of pauses opens a support mode. Contributors see each decision and its reason.

Measuring evaluators. Gold clips measure two numbers: sensitivity, the share of defective clips a person catches, and specificity, the share of good clips they accept. Both are required, because on mostly good gold a person who accepts everything looks accurate and catches nothing. Levels change at most once a week, from gold judged before the week began, so that nobody can tell live which clip was gold (Table 4). As trust grows, gold becomes rarer: about one clip in three at level 0, one in five at level 1 and one per batch at level 2. At level 2 an acceptance decides alone, except for a 20% sample that still gets a second listen; a rejection always needs two people, so one person can never sink a recording.

Experts. The expert desk shows one item at a time: a held recording with the evaluators' answers and flags but not their names, where the expert accepts, rejects or excludes it from every dataset at

Table 4: How evaluator levels move, measured on gold clips judged in work batches since the person last passed the test.

Rule	Window	Bar
Pause	Last 50, at least 20 gold	Accuracy under 70%, or sensitivity under 60% with at least 5 defective clips
Level 0 to 1	Last 50	Accuracy at least 90%; sensitivity and specificity at least 80%
Level 1 to 2	Last 100	Accuracy at least 95%; sensitivity and specificity at least 90%
One level down	Last 50	Accuracy under 85%

no cost to the speaker; a blind 2% sample of decisions as an audit; and decided, unflagged recordings that two evaluators accepted, which an expert may turn into new gold. A gold copy made from a real recording is retired after 30 people have heard it, or at once if the speaker withdraws. On the development deployment, gold clips are the app's own spoken guidance and near-silent clips.

Transcripts. Writing is a second skill. A listener in a language may take a writing test of four gold clips and passes with three written within 10% folded CER. An accepted recording gets a transcription task; a second writer verifies the text, and if the two versions are more than 3% apart in folded CER, an expert writes the version that is kept. Every version is kept, and withdrawal erases the text of all of them. The convention has two forms: a strict form, stored and shown, and a folded form for scoring, which ignores hooked-letter spelling, apostrophe shapes, case and punctuation, and in Zarma also folds ŋ, ɲ, nasal marks and doubled vowels. Numbers are written as words, nine tags mark events such as noise and personal information, and other languages inside a recording are marked as spans; in our Zarma work about a third of what recognisers were charged as error was spelling convention [14].

Automatic drafts. A recogniser can draft the transcript for the writer to correct. Drafts are off by default; when on, the stored audio goes to Namu's own recognition service under a project that stores nothing, because contributors consented to collection, not to a second use of their voice. Gold clips then get drafts with errors seeded per assignment, so that errors cannot be shared between people. In the design, drafts are used only where a language's rolling draft error is 20% or lower, 5% of training clips are transcribed blind to measure how much drafts anchor writers, and test-set speakers never get a draft (Section 10).

Pay. Payment is part of the design, built as a ledger separate from points; payments are not switched on and no amount is set. Points never become money. Contributors are rewarded only for accepted recordings, in steps along the path at counts of accepted recordings, and a withdrawal never takes back what was earned. Evaluators are paid per minute of audio judged, at a price shown and reserved before the batch starts, and every judged clip is paid the same whatever the verdict, so that pay never depends on accepting or rejecting.

9 Automatic checks

Before review. In the design, the phone measures duration, level, clipping and the share of speech within a second of stopping and offers two buttons, try again or keep it anyway; it never blocks a recording. The server then runs signal checks and model checks on every received recording, each stored with its version (Appendix Table 7). Recordings with a signal-to-noise ratio between 5 and 15 dB are kept for recognition only, and those at 25 dB or more become candidates for synthesis. Language identification only raises a notice or an evaluator flag, because standard identifiers call Zarma Hausa most of the time [14], and suspected synthetic speech goes to an expert, never to automatic rejection. All thresholds are starting values, to be re-tuned after 500 native-reviewed clips

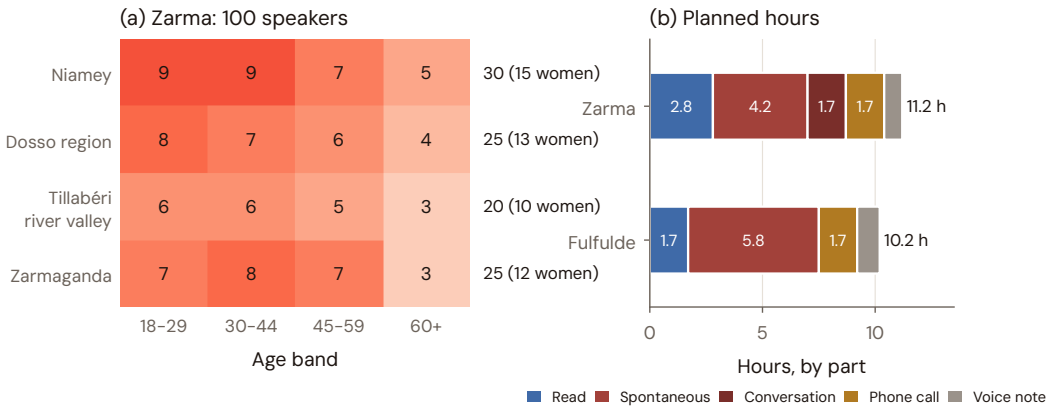


Figure 7: The planned Zarma test set. (a) 100 speakers by place of upbringing and age band, half of them women in every region. (b) Planned hours by part for the Zarma and Fulfulde designs; the optional Fulfulde conversations are not counted. Nothing has been recorded.

per language.

Reading back with a recogniser. The check we rely on most is the one our Zarma work taught us: reading the audio back with a recogniser and comparing it with the text it is supposed to match [14]. It finds references that belong to another recording, transcripts that stop before the audio does, and read sentences that were not read as written. Our rule is that no set is frozen before every reference has been checked against its audio in this way; in the studio, every file is read back by a recogniser not trained on that voice and also heard in full by a second native speaker.

10 Test sets

A test set must be recorded and frozen before any model in its language is trained, never trained on, and changed only by a new version. The app alone cannot guarantee this, because it cannot check who is speaking or keep a person who is also a contributor out of training. Test sets are therefore recorded in assisted sessions, with a facilitator present, under evaluation-only consent.

Balance. The Zarma design has 100 adults, 50 women and 50 men, in four age bands, by place of upbringing: Niamey 30, the Dosso region 25, the Tillabéri river valley 20 and Zarmaganda 25 (Figure 7). Speakers name their own variety, and speakers of the Kaado variety of the river valley form a labelled slice, never pooled silently. The Fulfulde design has 110 adults, 55 women and 55 men, labelled by variety: 55 of the western variety, 45 of the central-eastern one (including at least 12 who call their speech Wodaabe) and 10 from Diffa, whose variety the speakers will name; its floor is 70 speakers. Because errors cluster by speaker, interval width depends more on speakers than on utterances, so targets are set in speakers; the Zarma floor is 60. Fulfulde speakers who have recorded for another speech project are left out, because their voices may be in public training data.

Parts. Each Zarma speaker reads 20 newly written sentences (2,000 utterances, about 2.8 hours) and speaks for about 2.5 minutes from photographs taken in Niger and open questions (about 4.2 hours). Ten pairs hold 10-minute conversations recorded on separate channels (about 1.7 hours), 50 speakers make an ordinary phone call whose received side is recorded (about 1.7 hours), and 50 record a short voice note in the messaging app they normally use, kept in its original file (about 0.8 hours). The total is about 11.2 hours. The Fulfulde design has the same parts, with the conversations optional and longer spontaneous speech for people who do not read: about 10.2 hours without the conversations.

Disjointness. Test speakers never record training data, and anyone with accepted training audio never joins a test set. Test sentences are written fresh, refused by every other campaign and checked against every text corpus we might train on, and every training manifest is checked against the test set for shared audio, text and speakers. For app-collected data, the design applies the same rules: a speaker's split is fixed at their first accepted clip and a test speaker stays a test speaker for life; splits target 90, 5 and 5% of hours, stratified by language, home region and age band; and every build checks that no test audio hash or folded text hash appears in training.

Transcription and freeze. Test references never start from an automatic draft, so that they do not inherit a model's errors. Two transcribers work independently, checking read sentences against their text and writing everything else from scratch; their agreement is measured before adjudication, a senior linguist of the language adjudicates every disagreement and marks unresolved segments as uncertain, and a fourth person re-hears 5%. The targets for word disagreement before adjudication, in folded form, are at most 3% for read speech, 8% for spontaneous speech and 10% for conversation and phone speech. The frozen set holds the audio, both raw transcripts, the adjudicated transcript, the list of accepted spelling variants and a manifest with a hash for every file, with a data statement. A withdrawal from a test set produces a new version, and models are re-scored on it.

Safety of field work. Facilitators do not travel into insecure areas. Speakers from the Tillabéri region, including Zarmaganda, are recruited where they live in Niamey or in towns the local team judges safe, and the data statement will say that rural speech from these areas is under-represented.

11 Voices for synthesis

Contributor and radio voices are never product voices. A voice for synthesis is recorded in a studio by a professional speaker under a written licence.

Voices and script. The Zarma design records six voice talents, three women and three men: two main voices of about 12 hours each and four supporting voices of about 4 hours each, about 40 hours in all, with the supporting voices drawn from different regions. Candidates read the same 30 sentences at audition, and a blind native panel chooses. The script, about 12,000 sentences and 110,000 words, is written new by native writers, reviewed by a second writer, and never machine-translated. Its core of 2,000 sentences, read by every talent, covers every sound, nasal and long vowel and tone sequence; the rest adds minimal pairs, numbers and dates, names, conversational sentences, everyday domains such as health information, farming and markets, and short paragraphs for prosody across sentences. Test-set sentences never enter the script, and it avoids political, security and ethnic content.

Recording. Our recording protocol asks for 48 kHz, 24-bit mono files, peaks between -12 and -6 dBFS with the gain fixed per voice, no noise reduction, compression, equalisation or normalisation, room tone at the start of every session and a spoken slate on every file. A script director follows every line and asks for an immediate retake of any misread tone, nasal or cluster. Before a batch is accepted, each file is checked for clipping, noise floor and drift against the voice's first session, and aligned to its text.

The licence. The licence term sheet, to be reviewed by counsel, covers what is licensed (the recordings, training on them, and a synthetic voice in Namu's products and API), the languages (any language other than the one recorded is a separate choice, off by default), plainly stated commercial use, a list of prohibited uses (political content, religious rulings, impersonation of real people, scams, unsolicited bulk calls and others), the talent's right to hear samples before launch and refuse once with a reason, revocation, and a recorded spoken statement in which the talent says they know their voice will be used to create a synthetic voice. Talents are paid for their sessions whether or not the voice ships. Recordings made under the current app agreement are not used for synthesis (Section 6).

Table 5: What each kind of data is for. Of the recordings, only those in the first row can be collected under the current agreement.

Data	Uses
Spoken missions from the app	Training and testing recognition (after transcription); language identification; drafts for the next round of transcripts
Assisted test sessions	Frozen test sets only, reported per slice with speaker-level intervals
Pair and group recordings (planned)	Conversational recognition and turn-taking, from separate channels
Studio voices under licence	Synthesis voices; reference clips; tests of synthesis by native listeners
Transcripts and their versions	Training targets; measuring draft quality and correction effort; gold clips
Judgments and flags	Decisions; measuring evaluators; the funnel in each data statement; never exported with identities
Native ratings of model output	Naturalness, tone and translation quality; decisions to release a model

12 Data statements, versions and uses

Eligibility. In the design, a recording enters a dataset only after a recheck at export: it is received and not withdrawn, its consent stands, the account is active and not under a fraud hold, its latest decision is accepted, any redaction is done, the agreement revision covers the dataset's use, and its speaker is not in another split. Recognition and synthesis sets also need a final transcript.

Versions. In the design, datasets are kept in private repositories, by family (recognition, synthesis, evaluation, a weekly working set and language identification) and language. Each version is dated and frozen; a purge after withdrawals creates a new version with the reason recorded. Each version holds a card, data shards by split, a manifest with a hash per clip and a lineage file (parent, recipe, code version and the funnel from received to included). Shards follow speaker groups, so a withdrawal rewrites one shard per split, and old versions keep their manifests but not withdrawn audio.

The data statement. Each card will follow the data statement format [2] and adds the orthography and normaliser version, recording conditions, consent terms with counts, evidence of speaker and text disjointness, the funnel with its thresholds, muting rates, deletions since the parent version, and the sentence "removal does not retrain earlier models". Per clip it records keyed identifiers for clip and speaker, the language tag, code-switch and muted flags, signal-to-noise ratio, mission format, transcription mode and capture month. Region and age band are reported only in cells of at least 20 speakers, because regional origin can be sensitive. Gender and variety are self-reported (Section 3).

What each kind of data is for. Table 5 maps each kind of data to its uses. Native listeners evaluate model output with offline rating packets that run in a browser without a network: mean opinion scores, comparisons against a hidden reference, A/B/X and pairwise preference, tone identification, text vetting and transcription. Stimuli are anonymised and shuffled per rater, every sound must be played, results carry confidence intervals and inter-rater agreement, and raters are screened as in standard listening tests [10], for example when they rate the hidden reference under 90 on more than 15% of items. Packets for Zarma synthesis are built and not yet returned [14].

13 Discussion

Speakers before hours. Training longer on the same few voices stopped helping early, so the programme is built around the number and spread of speakers: short missions anyone can record, test

sets defined in speakers, and data statements that report who is in each set. The app can reach many people but cannot verify who is speaking or control the room, so it supplies training data, facilitators supply test sets, and studios supply voices.

Fairness inside review. Several rules protect contributors from unfair review, at some cost: a rejection needs two people, evaluators are paid the same whatever they decide, a recognised voice goes to someone else, linked accounts never review each other, and three rejections lead to a pause rather than a penalty. Between reproducibility and deletion the design chooses deletion: a withdrawn voice is to leave every future dataset version and training run.

14 Limitations and broader impact

Limitations. This report contains no results from the programme itself, because nothing has been collected. Every component labelled built has run only on a development deployment with synthetic accounts and recordings, and many parts are plans whose details may change when they meet real contributors. The thresholds of the automatic checks, the evaluator levels and the test-set targets are starting values, not measured ones. The evidence behind the design is limited: the memorisation result comes from one recogniser, five training speakers and one held-out speaker, and the compression check from two real recordings. The app reaches only adults with a smartphone and a data connection, its interface is in French, and spoken guidance exists so far only in Hausa and has not yet been reviewed by an independent panel of native speakers; people who do not read French will rely on that guidance or on facilitators. Field teams will not go to insecure areas, so rural speech from parts of the Tillabéri region will be under-represented. Gender, age, region and variety are self-reported. The identity check, which protects review, may exclude people without documents until in-person checks exist. No children are recorded, so no set will support children's speech.

Broader impact. Speech technology in the languages of the Sahel could give people who do not read French access to information and services in their own language, and the programme is designed to pay the people whose speech and judgment build it. Pay must be fair and known before the work; amounts are not yet set, and how to set a fair reward in Niger is still open. Consent must be understood, which is why the agreement can be heard in Hausa, withdrawal is one tap for 14 days, and the uses, including commercial ones, are written in plain words. Recordings are kept by a company for its research and products and are not published, so the benefit to communities depends on the products that follow and on the choices Namu makes about access. Voices can be misused for impersonation and fraud: no contributor's voice becomes a product voice, studio voices carry a licence with prohibited uses and a right to refuse, and the design adds checks for synthetic or replayed speech and a watermark detector for Namu's own synthetic output once it is embedded [20, 21].

15 Conclusion

The public speech of Niger's languages comes from too few speakers and too few situations, and models trained on it learn the speakers rather than the language. We designed a programme to record speech with the people who speak it: spoken missions in an app that works on weak networks, consent bound to each recording and to each use, blind review by measured evaluators, deletion that reaches datasets, speaker-balanced test sets recorded in assisted sessions, and licensed studio voices for synthesis. The next steps follow from the plan: native review of the spoken guidance and the agreement in each language, a first Hausa pilot by invitation, the server checks and compressed previews, the first dataset version with its data statement, the two test sets and the studio voices.

Acknowledgements

This work relies on public speech and text corpora, benchmarks, open models and published work on data collection. We thank the people who made them and the speakers whose voices they hold.

References

[1] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers and G. Weber. Common Voice: a massively-multilingual speech corpus. In *Proceedings of LREC 2020*, pages 4218–4222, 2020.

[2] E. M. Bender and B. Friedman. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[3] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of ICASSP 2004*, volume 1, pages 409–412, 2004.

[4] A. Conneau et al. FLEURS: few-shot learning evaluation of universal representations of speech. *arXiv:2205.12446*, 2022.

[5] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):20–28, 1979.

[6] C. Emezue et al. The NaijaVoices dataset: cultivating large-scale, high-quality, culturally-rich speech data for African languages. In *Proceedings of Interspeech 2025*, 2025. *arXiv:2505.20564*.

[7] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[8] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes and M. Mitchell. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In *Proceedings of FAccT 2021*, pages 560–575, 2021.

[9] ITU-T. Recommendation P.800: methods for subjective determination of transmission quality. International Telecommunication Union, Geneva, 1996.

[10] ITU-R. Recommendation BS.1534-3: method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union, Geneva, 2015.

[11] M. K. Keita, C. Homan, E. Prud'hommeaux, S. Sako and A. Diallo. SALAN: a massive ASR dataset for the languages of Niger. In *Proceedings of LREC 2026*, pages 5821–5827, 2026.

[12] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[13] J. Meyer, D. I. Adelani, E. Casanova, A. Öktem, D. Whitenack, J. Weber et al. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Proceedings of Interspeech 2022*, 2022. *arXiv:2207.03546*.

[14] Namu AI Lab. Zarma speech technology from scarce data: an exploration. Technical report, Namu AI Lab, November 2026.

[15] W. Nekoto et al. Participatory research for low-resourced machine translation: a case study in African languages. In *Findings of EMNLP 2020*, pages 2144–2160, 2020.

[16] P. Ogayo, G. Neubig and A. W. Black. Building African voices. In *Proceedings of Interspeech 2022*, 2022. *arXiv:2207.00688*.

[17] T. Olatunji et al. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023.

[18] V. Pratap et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25, 2024. *arXiv:2305.13516*.

[19] S. Pulikodan et al. VAANI: capturing the language landscape for an inclusive digital India. *arXiv:2603.28714*, 2026.

[20] R. San Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran and H. Elsahar. Proactive detection of voice cloning with localized watermarking. In *Proceedings of ICML 2024*, 2024. *arXiv:2401.17264*.

[21] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans et al. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Computer Speech and Language*, 64:101114, 2020.

A The mission catalogue

Table 6: Mission formats in the catalogue of 25 October 2026, with the number of missions (n) and one example prompt each, translated from the French. "Published" formats are playable by one speaker and published in four recording languages.

Format	n	Example prompt	State
Imagined response	25	Imagine a new, harmless use for an empty box.	Published
Story	22	Tell an ordinary moment you like to share, or choose another mission.	Published
Free speech	21	Say one small thing you do with an ordinary object.	Published
Choice and reason	14	To carry a light object, which of two containers would you choose, and why?	Published
Picture sequence	12	Imagine a path between three landmarks and explain it in your chosen order.	Published
Comparison	11	Compare these two objects: what changes in their shape?	Published
Picture description	8	Choose an illustrated object and say what you call it.	Published
Conversation	32	Two consenting adults choose a familiar object and ask each other about its use.	Not yet
Group discussion	12	Several consenting adults propose three imaginary outings and discuss them.	Not yet
Translation	32	In two languages you speak, describe the same object separately.	Not yet
Listening review	28	Listen to an authorised example and name its language, or pass.	Not yet
Model evaluation	8	In an authorised test, make an imaginary request to a model and check it understood.	Not yet

B Automatic checks

Table 7: Planned server checks with their starting thresholds. Each result is stored with the check's version; thresholds are re-tuned after 500 native-reviewed clips per language. A missing model result counts as a warning, never as a pass.

Check	Starting threshold and action	Contributor sees
Duration	Under 2 s or over the mission's maximum: reject	A little short
Level	Active speech below -40 dBFS: try again	Bring the phone closer
Clipping	Over 0.5% of speech clipped: try again; 0.1 to 0.5%: label	A little less loud
Speech	Under 1.5 s of speech or a speech ratio under 0.25: try again; silence over 15 s: label	No voice heard
Noise	SNR under 5 dB: reject; 5 to 15 dB: recognition only; 25 dB or more: synthesis candidate	Too much noise
Exact duplicate	Same decoded audio as any account or the exclusion list: reject	Not retained
Near duplicate	Matches another recording or broadcast audio over 5 s: hold	Under review
Language	French or English with high confidence: notice; confusion among Sahel languages: evaluator flag	This mission is in …
Speaker count	A second voice over 10% of a solo recording: to evaluators	Nothing
Synthetic or replayed	Most suspect 1% until calibrated: to an expert; never rejected	Under review
Namu watermark	Any detection: hold	Under review