Zarma Speech Technology from Scarce Data: An Exploration

Namu AI Lab
namuai-lab.com

Abstract

Zarma is spoken by about six million people, most of them in Niger, yet it has almost no public speech technology and very little data to build it from. We report what we learned from building and measuring speech recognition, speech synthesis, translation and a spoken voice agent for Zarma, starting from public resources and a small amount of our own data. Four findings stand out. First, the data is scarcer than it looks: of about a thousand hours of public Zarma audio, roughly 1% is both openly licensed and transcribed, and most downloadable Zarma text is one translation of the Bible. Second, measurement is the hard part: about a third of what a recogniser is charged as error is spelling convention, word error rate is not usable, standard language identifiers call Zarma speech Hausa about 80% of the time, and some public reference transcripts belong to another speaker. Third, models adapt well across related languages when trained jointly: a single synthesis model learned Zarma, kept its Hausa and later added French, while a recogniser fine-tuned on Zarma alone lost most of its Hausa. Fourth, in a voice agent the weakest link is not hearing but translating what was heard, and that link also decides whether a request for help survives. Our best agent answers 63 of 111 test turns correctly through every stage, with a median reply delay of 6.8 seconds. All results are automatic measurements; no native-speaker panel has rated any output yet.

1 Introduction

Zarma, a Songhay language, is spoken along the Niger River in western Niger and is one of the main languages of Niamey. Ethnologue counts about six million speakers [6]. For a language of that size it is almost invisible to speech technology. It has no collection in Common Voice [1], no test set in FLEURS [4], and no adapter in widely used massively multilingual recognisers [16]. Large language models do not write it. Many of its speakers read and write French or no written language at all, and many reach the internet through a phone on a weak and costly connection. Speech is the natural interface for them, and Zarma is the language many of them would choose.

This report describes an exploration. We wanted to know how far public resources go for Zarma, where the real obstacles are, and what a first spoken assistant in Zarma looks like when every stage of it is measured. We built and tested recognisers, synthesis models and translation models, and connected them into a voice agent that listens in Zarma, reasons in French and answers in Zarma. We kept frozen test sets, fixed our decision rules before scoring, and report 95% confidence intervals throughout. We describe the models we used by their type and scale rather than by name, and our own models by what they were trained to do; the point of the report is what the measurements show.

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

Our contributions are the following. (1) A survey of public Zarma speech and text with their terms of use, which shows that usable data is scarce in a specific way: hours exist, but almost none are both open and transcribed (Section 4). (2) An account of what makes Zarma hard to measure, including spelling variation, wrong references and failed language identification, with the practices that held up (Sections 5 and 6). (3) Recognition results showing that scale and fine-tuning on a few speakers buy little, and that training on Zarma alone erases Hausa (Section 7). (4) Synthesis results showing that one model can learn Zarma, keep Hausa and add French (Section 8). (5) Evidence that no open language model writes Zarma, and a measured voice agent in which translating recogniser output is the weakest stage, for quality and for safety (Sections 9 and 10).

2 Related work

Multilingual speech at scale. Public speech collections and models now cover hundreds or thousands of languages. Common Voice gathers read speech from volunteers [1], FLEURS provides read speech in about a hundred languages for evaluation [4], and massively multilingual recognisers extend coverage to more than a thousand languages by pre-training on large amounts of untranscribed audio and adapting to each language with little data [16]. Zarma is absent from the first two and has no adapter in the widely used releases of the third. Coverage at this scale often rests on religious recordings with one reader per language, which says little about everyday speech.

Zarma language resources. Recent work has begun to fill the gap. A French–Zarma parallel corpus with a glossary and a translation model was released for research [10], and work on grammatical error correction has treated Zarma as a case study for low-resource languages [11]. A large speech recognition dataset for the languages of Niger, built from radio, offers several hundred hours of Zarma, a small part of them human-corrected [12]. For translation, professionally translated English–Zarma sentences [3] and a French–Zarma evaluation benchmark [2] now exist, alongside older dictionaries, including one with tone-marked pronunciations [7, 14]. Linguistic descriptions document the dialects of Zarma in Niger [9], its official orthography [13] and its grammar [8]. To our knowledge, no previous report measures recognition, synthesis, translation and a spoken agent for Zarma together.

Evaluation with automatic judges. Character-level metrics are the usual choice where word segmentation is unstable; we use character error rate for speech and chrF++ [15] for translation. Intelligibility of synthesised speech is commonly estimated by having a recogniser read it back, which ties the measurement to that recogniser's own errors. We address this with two recognisers of different architecture as judges and report where they disagree.

3 The language and how it is written

In Niger, Zarma is usually described in several dialect areas: Zarmatarey around Dosso and Niamey, Zarmaganda to the north, Kaado along the river near Tillabéri and Téra, the Wogo and Kurtey communities of the river islands, and Dendi in the south around Gaya [9]. One written standard serves these varieties, and the Dosso variety underlies most school and literacy material. Speakers in Niger understand texts from Gao in Mali, written in Koyraboro Senni, only partly: intelligibility tests report 28 to 64% for a Gao text against 92 to 100% for Niger texts [9]. This matters for data, because much text labelled "Songhay" on the web is not Zarma.

The official orthography was fixed by decree in 1999 [13]. It uses a Latin alphabet of 30 letters, including *ŋ* (as in *kaŋ*) and *ɲ*, and five nasal vowels written with a tilde below the letter. Tone is not written, although the language has two tonal levels that distinguish words [8]. Vowel length is written by doubling, except at the end of most words.

Real text departs from this standard in ways that matter for every model we trained. We found no corpus that writes nasal vowels with the tilde below; the most widely available text writes a tilde above instead. Word boundaries vary: transcribers often join particles to the next word, writing

Table 1: Two spellings of one frequent word across Zarma sources. The standard writes *kaŋ*; spoken-language transcripts mostly write *kan*.

Source	*kaŋ*	*kan*
1989 Bible translation	26,452	0
An open Zarma–French dictionary	1,425	2
A public English–Zarma parallel set	1,735	454
A large public web crawl	1,090	443
Transcripts of a public speech corpus (dev and test)	59	968

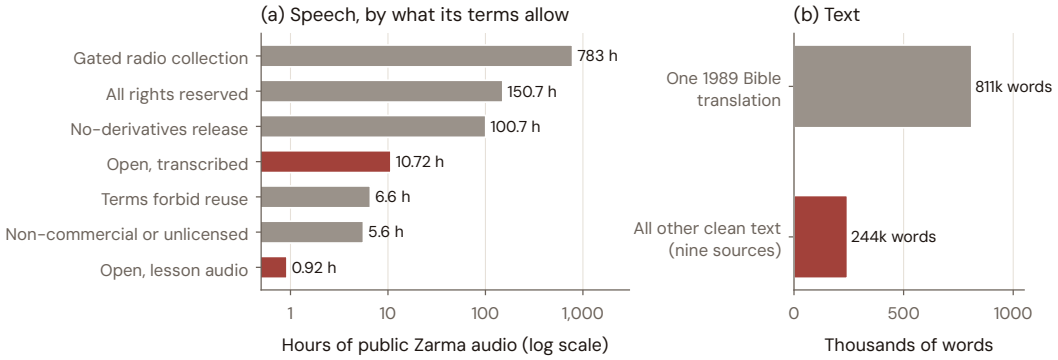


Figure 1: Public Zarma data. (a) Speech by what its terms allow, in hours, measured or reported by the source; overlaps between sources are small and ignored. (b) Downloadable Zarma text: one translation of the Bible against everything else that is clean and not a copy of it.

aga where an edited text writes *a ga*. Table 1 counts two spellings of one common word across sources. The 1989 Bible translation never writes *kan*; the transcripts of one public speech corpus almost always do, and one of its transcribers writes both forms on consecutive lines.

Zarma and Hausa are not related, but they share a country and many words. Hausa loans such as *gaskiya*, *saboda* and *lokaci* occur in everyday Zarma, and Hausa spellings with hooked letters (*ɓ ɗ ƙ ƴ*) appear in Zarma transcripts. Zarma words are short: 3.5 letters on average, against 4.5 for Hausa. Code-switching with French is common in radio and in Niamey, and we found no transcribed example of it.

4 The data landscape

We surveyed every public source of Zarma speech and text we could find, checked its terms, and measured what we could. Figure 1 summarises the result.

Speech. We counted about 1,060 hours of public Zarma audio. About three quarters of it sits in one radio collection built for speech recognition research [12], of which about 60 hours are human-corrected and access is by manual approval. Most of the rest is broadcast news and recorded Bible readings with all rights reserved, or academic releases whose terms forbid derivatives. Roughly 11 hours are openly licensed, and about 10.7 of those are transcribed: a corpus of prompted spontaneous speech from seven speakers. That is about 1% of the audio. We found no openly licensed Zarma corpus recorded and transcribed 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.

Text. One text dominates. The 1989 Zarma Bible, about 811,000 words, is the whole of two well-known parallel collections and most of a third, and the training text of a popular text language identifier is mostly that Bible. Outside it we found about 15,100 clean, unique sentences (about 244,000

words) across nine sources, and close to half of those words are transcripts of the same seven speakers. Large web crawls contain almost nothing: one keeps 46 Zarma documents, and in another, about a quarter of the words labelled Zarma are Koyraboro Senni. For translation, the best open English–Zarma resource has 2,472 sentences and 4,001 terms [3]. Open French–Zarma material adds up to about 2,850 sentence pairs from two dictionaries [7, 14] and a set of illustrated children's stories, and one public benchmark provides 854 French–Zarma test sentences for evaluation only [2]. One open dictionary also gives 6,909 tone-marked pronunciations.

What we assembled. From these sources we built an openly licensed Zarma text set of about 134,000 training words, without the Bible and with every evaluation sentence removed. It is small, and most of it is either transcribed speech or text translated from English. We froze evaluation sets for recognition, synthesis and translation (Section 5 and Appendix A), and designed a recording programme for a test set of 100 speakers balanced by gender and region, with consented voices for synthesis. None of that new audio has been collected yet; every result in this report rests on public data and a small amount of internal data.

5 Measuring Zarma speech

Evaluation for a language like Zarma is not a solved preliminary, and several of our most useful results are about measurement.

Test sets. Our recognition test set comes from a public corpus of prompted spontaneous speech, with one speaker for development and another for test: 1,060 test segments, 1.5 hours. A confidence interval on it describes this voice and this transcriber, not Zarma speakers in general. Synthesis is tested on 140 held-out sentences of a public parallel set (those without digits), each rendered in five voices for 700 renders, and on a Hausa set of 334 targets from 20 held-out speakers. Translation is tested on the held-out sentences of the English–Zarma set, on the public French–Zarma benchmark, and on 401 held-out French–Zarma dictionary examples. All sets were frozen before any score was computed, and every training set was checked against them for shared text, recordings and speakers.

Frozen means rebuilt, not re-cut. Re-cutting the recognition test set on a different GPU moved 14 of 1,568 segment boundaries, because low-precision arithmetic in the aligner differs across hardware. Rebuilding the set from its stored manifest reproduced all 1,568 segments exactly. A frozen test set has to be stored as a manifest and rebuilt from it.

Character error rate, in three forms. We score recognition and synthesis by character error rate (CER) under three normalisations: *strict*, which lowercases and removes punctuation; *folded*, which also maps *ŋ* to *n*, removes nasal marks and collapses doubled vowels; and *no-space*, which also ignores word boundaries. We report folded CER as the headline. Word error rate is not usable: on the development speaker it is 77 to 90% while CER is 28 to 35%, because transcribers join and split particles freely. Strict CER can even reward the wrong model. A large recogniser forced to decode Zarma as Hausa scores *better* on strict CER, because it stops writing *ŋ*, and the test transcriber writes *kan* 703 times against 26 for *kaŋ*.

How much error is spelling. We attributed the character edits of the best recogniser on the test set to their cause (Table 2). Word boundaries account for about a fifth, vowel length for 8%, *ŋ* and *ɲ* for 5%. About a third of what is counted as error is a spelling convention rather than a mishearing.

References can be wrong. While checking the public corpus for overlap between its splits, we found eleven training transcripts of one speaker that repeat development transcripts of another, six of them word for word, although the audio differs. A recogniser settled which side was wrong: on the development speaker's audio these references score 53 to 70% CER, about as badly as unrelated text, while on the training speaker's audio they score 16 to 36%. Ten development references therefore describe someone else's recording. We report development results without them; including them

Table 2: Where the best zero-shot recogniser's errors come from, as shares of strict-form character edits on the test speaker.

Cause of the character edit	Share of edits (%)
Word boundaries	21.2
Vowel length	8.1
ŋ and *ɲ*	5.2
Hyphens and apostrophes	0.9
Hearing, and spelling variation not covered above	64.6

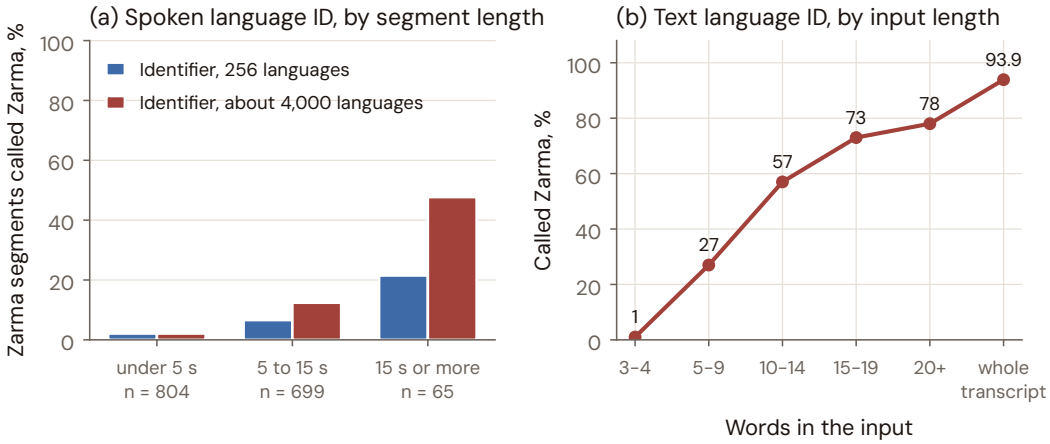


Figure 2: Language identification on Zarma. (a) Share of 1,568 Zarma speech segments that two public spoken-language identifiers label as Zarma, by segment length. (b) Share of Zarma transcript lines that a public text identifier labels as Zarma, by length in words.

moves CER by about one point. Separately, 27 test references are shorter than their audio, and the six we inspected each omit a repeated phrase. We met the same problem later in a public French corpus, where one reader's reference stops halfway and every system fails on that voice. Our rule since then is to check every reference against its audio with a recogniser before freezing a set.

Judges. A render counts as intelligible if a recogniser reads back the intended text. Because one recogniser can invent plausible speech, every synthesis result uses two recognisers of different architecture as judges, and we flag renders where they disagree sharply. The two judges sometimes disagree on which system is better, and we report both.

Intervals and selection. Every interval is a 95% bootstrap with 2,000 resamples (1,000 for translation), clustered by recording, sentence, document or speaker as the test requires; differences between systems are paired. We select on development data and report on test. A single development speaker proved too noisy to choose checkpoints: in two recognition fine-tunes, the last checkpoint beat the one chosen on development by 0.5 to 1.7 points on test. We now fix the checkpoint rule in advance.

6 Language identification

A spoken assistant that serves several languages usually starts by detecting the language. For Zarma this does not work (Figure 2).

Two public spoken-language identifiers, covering 256 and about 4,000 languages, label 82% and 78% of our 1,568 Zarma segments as Hausa. Zarma is their first choice for 5 to 9% of segments, rising only for segments longer than 15 seconds. On whole recordings of about 87 seconds the larger identifier reaches 68%. A widely used multilingual recogniser that has no Zarma spreads Zarma

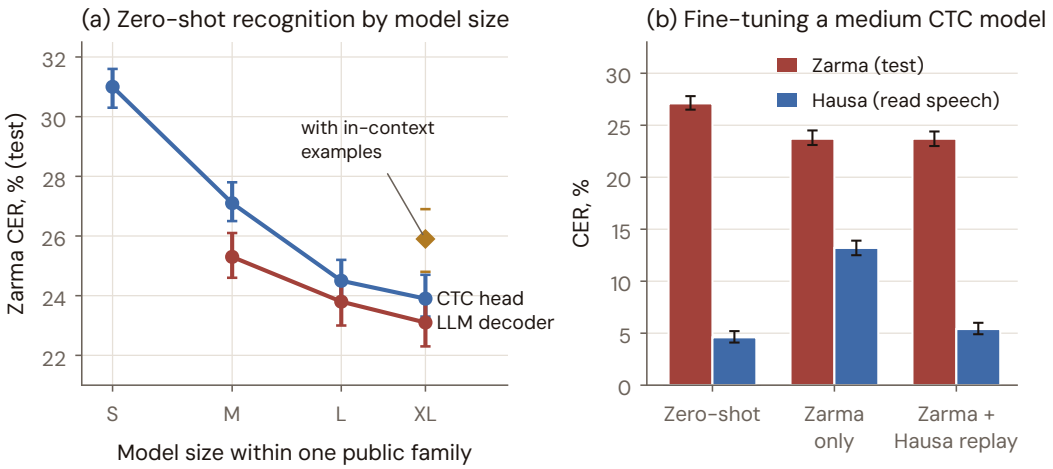


Figure 3: Zarma speech recognition, folded CER on the test speaker with 95% intervals. (a) One public multilingual family in four sizes, with a CTC head and with an LLM decoder; the diamond adds in-context examples to the largest decoder model. (b) The medium CTC model before and after fine-tuning on five Zarma speakers, alone and with Hausa mixed in, at the same training step. Hausa is measured on a public read-speech test set.

speech over Swahili, English and Sinhala.

Text identification after recognition does not rescue short turns. A public text identifier covering about 2,000 languages recognises 94% of whole transcripts but only 1% of lines of three or four words and 27% of lines of five to nine words; its usual mistakes are other Songhay languages and Hausa. On the clean sentences of a public parallel set it recognises 55%. In the voice agent, a single real Zarma turn was identified as Zarma 80% of the time. We therefore let the user choose the language at the start of a session and never switch on identification alone.

For building text data we wrote a simple identifier from Zarma function words and markers of Koyraboro Senni. Calibrated on text of known origin, it accepts 99% of documents from the parallel set and none of 47 Koyraboro Senni documents, and it removed 34 Koyraboro Senni documents from one web crawl. It separates Zarma from its Songhay neighbours but not from Hausa: it accepts 38 of 150 Hausa sentences. We claim no precision for it until native speakers label a sample.

7 Speech recognition

We tested public multilingual recognisers without any Zarma training, then fine-tuned one on Zarma. Figure 3 shows both; Appendix Table 4 gives every system.

Zero-shot. The best zero-shot recogniser, a large model with an LLM decoder, reaches 23.8% folded CER on the test speaker, with an interval of 23.0 to 24.6. Scale helps less than one might hope. Within one family, growing the CTC model from the small to the extra-large size takes CER from 31.0 to 23.9%, but going from the large to the extra-large decoder model gains 0.6 points on test and loses 0.5 on development, so the sign is not settled. Giving the largest model Zarma examples in its prompt made it worse by 2.2 points and left 20 outputs empty. An older recogniser with adapters for related Songhay languages reaches 36 to 50%, about one and a half times the error of the best model. The CTC and decoder models of the same size have nearly the same CER, but the decoder's output is about six points better in folded word error rate: its gain is in placing word boundaries. The language tag matters at small scale: decoding Zarma with the Hausa tag costs 5.2 points at the medium size, and the model stops writing *ŋ*.

Fine-tuning. We fine-tuned the medium CTC model on about seven hours of speech from the five training speakers of the public corpus. Its test CER fell from 27.1 to 24.2%, which ties the best

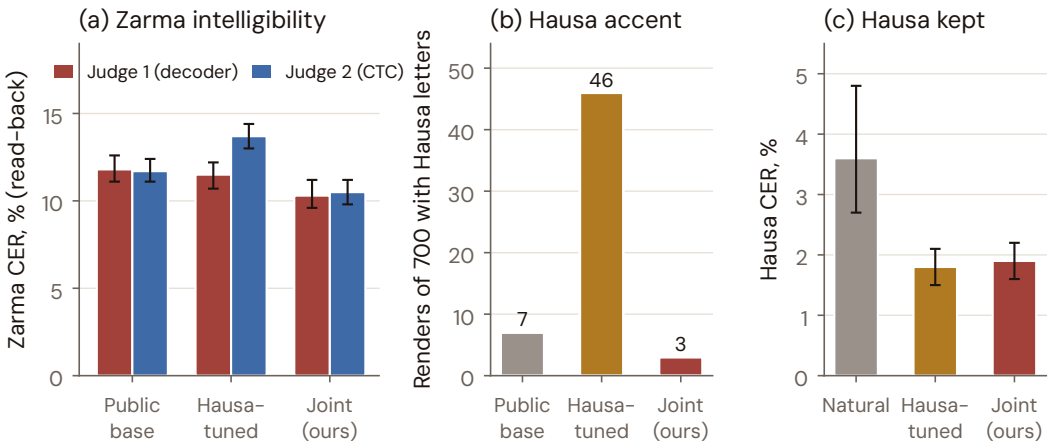


Figure 4: Zarma speech synthesis. (a) Intelligibility of 700 renders of held-out sentences in five voices, as CER read back by two recognisers. (b) Renders whose read-back contains a Hausa hooked letter, a proxy for Hausa-sounding consonants; natural Zarma speech triggers it in 1 of 888 segments. (c) The same models on a Hausa test set, against the natural recordings.

zero-shot model but does not beat it. It also forgot Hausa. On a public Hausa read-speech test set its CER rose from 4.6 to 12.4%, and it almost stopped writing the hooked letters of Hausa: of the 502 hooked letters in the reference transcripts, it wrote 131 as in the reference before fine-tuning and 1 after. Mixing Hausa back into training, at about a fifth of the utterances, kept Hausa at 5.4% at the last checkpoint (5.3% at the checkpoint selected on Zarma). At the same training step the Zarma result with and without Hausa was identical (23.7%). The five training voices were memorised within about a thousand updates.

With five speakers, more training steps and larger models buy little; the next gains will come from more speakers. When a model must serve both languages, joint training costs nothing measurable, and training on Zarma alone costs most of the Hausa.

8 Speech synthesis

We measured how intelligible synthesised Zarma is by asking two recognisers to read it back, and whether it sounds like Hausa by counting Hausa hooked letters in what the CTC judge hears. That count is only a proxy for accent; native listeners will decide. Figure 4 shows the main comparison and Appendix Table 5 every system.

Zero-shot. We started from a public massively multilingual synthesis model that clones a voice from a short reference clip, and from our own Hausa model, fine-tuned from it on about 128 hours of Hausa. On held-out Zarma text both reach 11 to 14% CER, and the two judges disagree on which is better. The Hausa model, however, carries its Hausa into Zarma: 37 to 46 of 700 renders contain Hausa hooked letters, against 3 to 7 for the base model. Three other findings shaped everything after. The voice matters more than the model: CER ranges from 9 to 16% across the five voices. Digits must be written out: sentences with digits are read back at 19 to 20% CER, against 10.4 to 12.5% when the same numbers are spelled out in Zarma first. And a voice taken from one language loses intelligibility in the other: a Hausa voice reading Zarma costs 3 to 5 points while still sounding like the same speaker. Renders are easier for a recogniser than natural spontaneous speech; on a speaker the base model had seen, natural recordings score 23.4% and renders 11 to 13%. These numbers describe intelligibility under a recogniser, not naturalness.

One model for two languages. We fine-tuned the Hausa model on about 3.6 hours of prompted Zarma speech from five speakers, with Hausa mixed in at 40% of samples. Zarma CER fell to 10.3%

(10.5% with the second judge), 1.1 and 3.2 points below the Hausa model's zero-shot result. The Hausa letters disappeared: 3 renders of 700. Hausa was unaffected: the difference from the Hausa model on the Hausa test set is $+0.11$ points, with an interval from -0.12 to $+0.35$. Starting from the public base instead of the Hausa model, with Hausa in a quarter of the samples, learned Zarma about half as well and lost Hausa by more than a point. With 25% Hausa the Zarma result was the same and Hausa slipped slightly; 40% kept it. The best checkpoint was the first one we saved, after about four passes over the data, and later checkpoints were slightly worse. The public base model had already seen Zarma of this kind in its own training, so the next gain needs new speakers recorded for synthesis.

A third language. French is the language of Niger's institutions and the agent's internal language, so we tested whether the joint model could speak it too. Before any French training it had kept the base model's French (2.2% CER on a public French read-speech set, the same as the base). We then continued training with French at about a fifth of samples, from audiobooks, spontaneous speech and Nigerien-accented read speech. The three-language model holds every language within our pre-set tolerances: French 2.4% against 2.2% for the base, Zarma 10.5% against 10.4%, Hausa 2.3% against 2.4%. On held-out speakers with a Nigerien accent its French moved from 5.5 to 4.7%, but that set of 61 targets is too small to show a difference (interval -2.6 to $+0.8$). One public French reader's reference was cut short, which failed every system on that voice until we checked references against audio. The French data taught one more lesson: a common pitch-based estimator put the share of women's voices at about 14%, and a more careful one at 34%. Gender balance claims depend on the estimator.

What the reference clip carries. A voice-cloning model copies more than timbre. Our first reference clip for one consenting team member's voice was 28% silence, and the renders paused in the middle of sentences. Trimming pauses in the reference to 0.12 seconds cut long silences in the output from 9.8 to 3.3 seconds over four test sentences, with no measurable change in intelligibility. Figure 5 shows a natural reading and a rendering of the same Zarma text in that voice. The careful natural reading takes 4.9 seconds and leaves long gaps between phrases; the model says the same words in 2.9 seconds. Its pitch, estimated with YIN [5], sits around 134 Hz, close to the voice's reference clip (about 119 Hz) and higher than this slower reading (about 90 Hz).

9 Translation and the language model

A voice agent needs a component that understands and answers. We first tested whether open language models can do this in Zarma. They cannot.

No open language model writes Zarma. Asked to translate English into Zarma, a public massively multilingual translation model writes Fulfulde for most of 150 test sentences (88 by a text identifier, and only 4 in Zarma), reaching 8.5 chrF++. A 9B open language model further trained on 50 African languages, Zarma not among them, writes Hausa for 115 of the 150 and reaches 14.6. Both are below two trivial floors: scoring the Hausa reference as if it were Zarma gives 16.0, and copying the English source gives 14.1. Five worked examples in the prompt change nothing. Asked to answer everyday questions in Zarma, the same language model answered in French, Hausa or Fulfulde, or looped, and wrote Zarma in none of 30 cases.

The model's difficulty is measurable without generation. On the same 150 sentences it needs 3.34 bits per character to encode Zarma against 1.48 for Hausa, 2.26 times as much. Its training on other African languages lowered this cost by 53% for Hausa and by 22% for Zarma, without any Zarma data (English fell by 16% in the same training): related exposure helps, but not enough. On a four-way choice of which English sentence matches a Zarma one, it is right 77% of the time, against 58% for a character-overlap baseline and 100% for the same sentences in Hausa. The model understands some Zarma and writes none. This settled the design: the language model reasons in a language it writes well, and translation carries Zarma in and out.

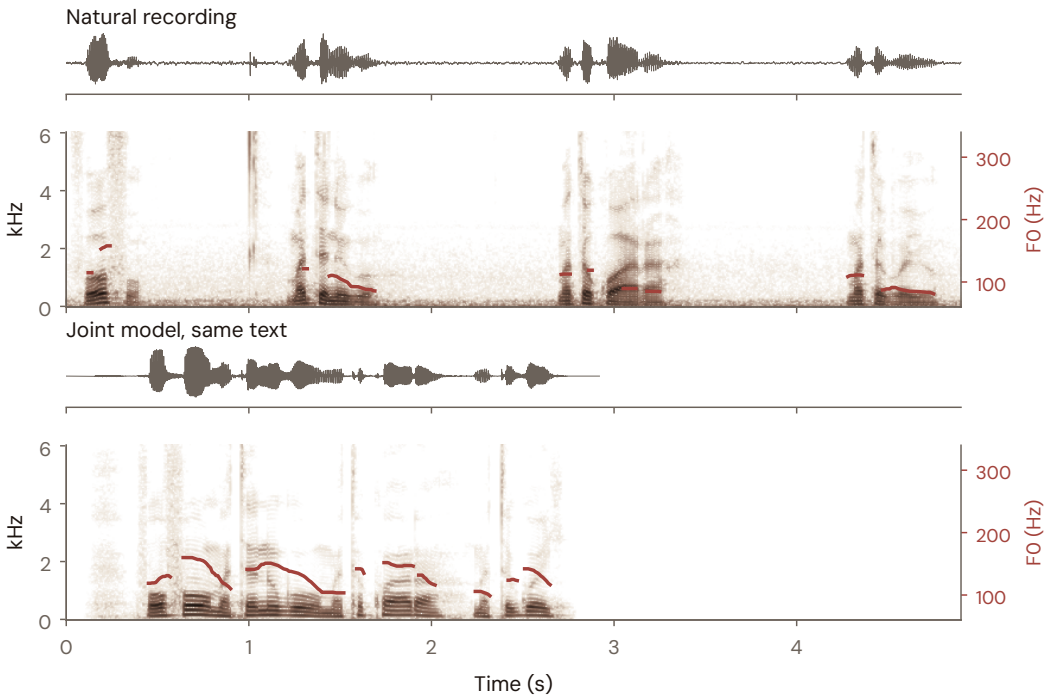


Figure 5: The same four Zarma greetings, *Fofo! Mate ni go? Mate kwaara? Mate goy?*, read naturally by a consenting speaker (top) and synthesised by the joint model in that speaker's voice (bottom), on one time scale. Each panel shows the waveform, a spectrogram up to 6 kHz and the pitch track (red).

Translation. We fine-tuned a public multilingual translation model on the open English–Zarma parallel set, with Hausa mixed in. On held-out sentences, English to Zarma rose from 8.5 to 41.7 chrF++ and Zarma to English from 21.2 to 40.4, and the Hausa directions improved as well. A round of training on back-translated text did not help. Because the agent translates what the recogniser heard, not clean text, we then trained on recogniser-style input; on transcripts of unseen voices, Zarma to English rose from 26.2 to 31.9, of which about half came from handling missing case and punctuation.

The agent reasons in French, so we trained a direct French–Zarma model from the English one with the open French–Zarma material (Figure 6; Appendix Table 6). From Zarma to French it reaches 29.2 chrF++ on the public benchmark, against 26.2 through English and 23.8 for the best public system we could find. From French to Zarma it reaches 24.7, level with the route through English (24.4) and with the best public system (24.5). Our pre-registered rule required it to beat the English route in both directions, and it did not. We adopted it anyway, as a decision we state as made after the fact, because it halves translation time and because on recogniser transcripts it is far ahead: 27.0 chrF++ from Zarma to French against 17.7 through English. No system we measured passes 25 from French to Zarma on the benchmark. What would move it is human French–Zarma translation in a conversational register, which does not yet exist in the open.

10 A first Zarma voice agent

We connected these parts into a voice agent (Figure 7). It hears Zarma, translates it into French, lets a language model answer in French inside a set of safety rules we call the care guard, translates the answer back into Zarma, spells out numbers, and speaks with the joint synthesis model. Hearing goes through a transcript; this is a cascade, not an end-to-end speech model.

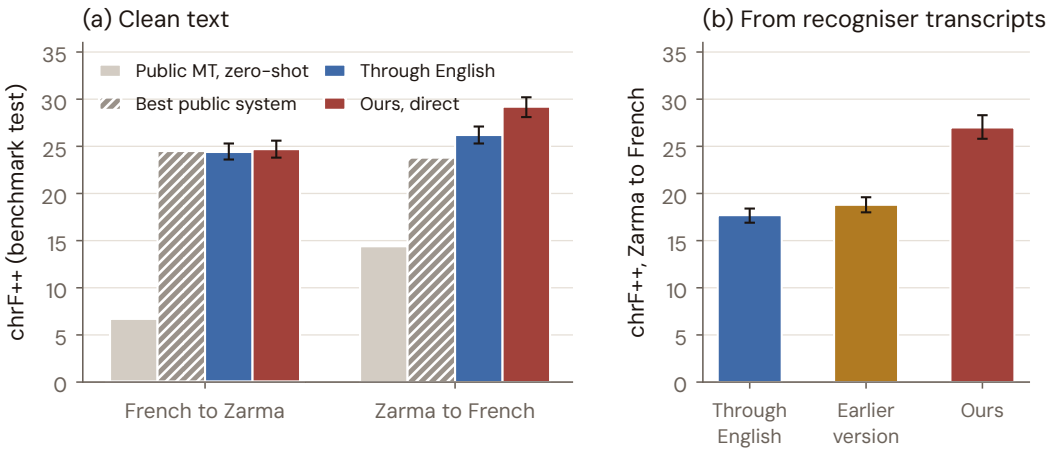


Figure 6: French–Zarma translation, chrF++ with 95% intervals where we computed them. (a) The public benchmark's test set: a public translation model without Zarma training, the best public system on the benchmark's leaderboard, translation through English with our earlier models, and our direct French–Zarma model. Leaderboard values do not state their settings and are indicative. (b) Zarma to French from recogniser transcripts of unseen voices, on 802 held-out sentences.

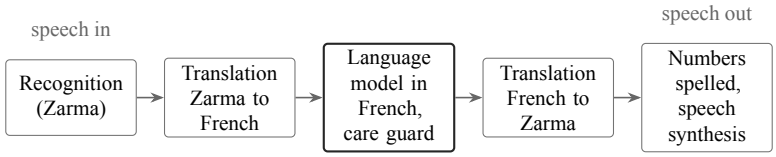


Figure 7: The voice agent. Each sentence of the answer is translated and spoken as soon as the language model finishes it.

How we test it. We test the agent with 111 spoken turns that mix everyday questions, requests for health and safety help, and harmful requests, mostly synthesised from translated text in held-out voices and partly real recordings of one speaker, plus silence and noise. A turn passes when every stage does its job: the transcript is close enough, the meaning survives translation, the answer is appropriate and short, and the reply is spoken. When a turn fails we record the first stage that failed.

Three versions. The first loop passed 23 of 111 turns (Figure 8a), and in most failures the first stage to fail was translation into the pivot language. A recogniser output with under 5% CER was enough to drop translation quality from 56 to 41 chrF++, and the mistranslations were not harmless: a three-week cough became a sentence about the size of a breast, and a pregnant woman's bleeding drew the reply that it sounded like a fun idea. The second version, with recogniser-robust translation, the care guard and shorter answers, passed 41. The third, which reasons in French with the direct French–Zarma model and a general-purpose 9B open language model chosen among six by a pre-registered rule, passes 63 and speaks on 110 of 111 turns. It routes 21 of 27 requests for help to the care reply and points to help without minimising the problem in 24 of 27. It also asks the user to repeat more often, and about six of its replies refuse ordinary statements.

Safety. The care guard is a fixed layer around the language model. It sends health and distress turns to a care reply, removes any reply that minimises a symptom, refuses harmful requests, and strips phone numbers, because no emergency number we could verify for Niger is in the system. It never sends a fixed Zarma care line through translation; those lines are left for native speakers and a health worker to write and record. Three findings from building it are worth stating. Every remaining care failure in the third version happens before the language model sees the turn: translation turned pregnancy into "I ate", rat poison into spoiled chicken, and diarrhoea into an exam. A guard can make a better model look worse: in our selection, the strongest language model lost because the

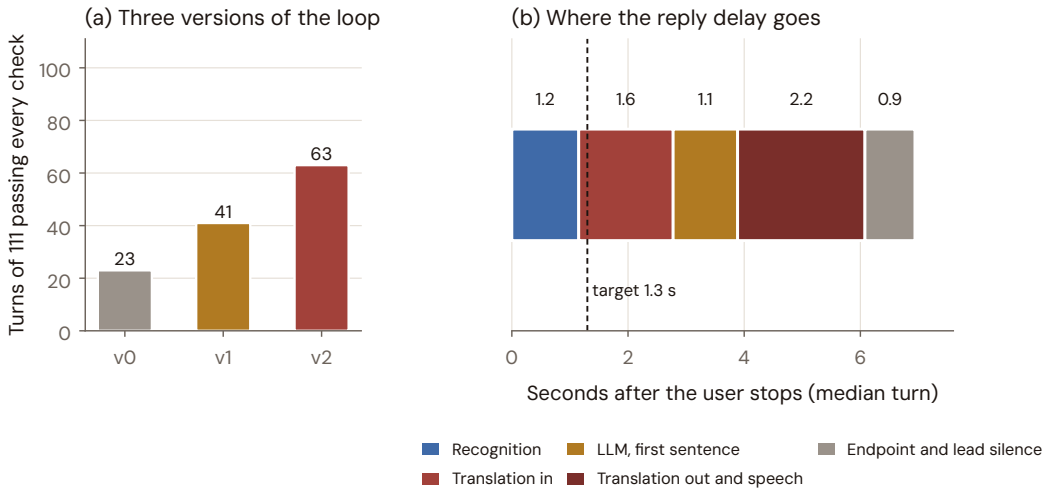


Figure 8: The voice agent. (a) Test turns passing every check in three versions of the loop: the first, through English; the second, with improved translation and the care guard; the third, through French. (b) Median reply delay of the third version, hosted on one cloud GPU, split by stage over 13 warm turns with real Zarma input.

guard replaced its correct refusals with a care line. And speed can cost safety: a faster configuration of the first loop halved its reply delay but, with a faster and weaker recogniser, routed only 6 of 26 requests for help instead of 11. We also learned that probes for distress and self-harm must be written by people, with clinical guidance, and not generated; a safety classifier stopped our attempt to translate them automatically. Those probes do not exist yet, so the care gate is not passed and the agent is not shown to the public.

Latency. Our target is a median reply delay of 1.3 seconds from the end of the user's speech to the first sound of the answer. No version meets it. Hosted on one cloud GPU, the third version has a median delay of 6.8 seconds and a 95th percentile of 8.0 seconds over 13 warm turns with real Zarma input, and the first turn after the service has been idle waits about three and a half minutes for the models to load. Figure 8b and Appendix Table 7 show where the time goes: recognition takes about 1.2 seconds, translation into French 1.6, the language model's first sentence 1.1, and translating and speaking that sentence 2.2. Translation in both directions is the largest share. The fastest turn of an earlier version answered in 1.6 seconds, so the target is not out of reach, but each speed-up has to be checked against the care results before it is accepted.

11 Discussion

Speakers, not steps. Every model we trained on Zarma stopped improving early. The recogniser memorised its five training voices within a thousand updates and only tied the best zero-shot model; the synthesis model peaked at its first checkpoint. Larger models and prompting added little. The binding constraint is the number of distinct speakers with consented, transcribed recordings, which is why a balanced, many-speaker recording programme comes next.

Measurement before modelling. Spelling conventions, wrong references, a defective public reference set, unstable re-cutting across hardware, an unreliable language identifier and a misleading pitch estimator each changed our conclusions at some point, and none of them is a modelling problem. For a language without a settled written form in practice, much of the work of evaluation is deciding what counts as the same word. The measures that held up were simple: folded CER with intervals, two independent judges, frozen manifests, checks of references against audio, and decision rules written before the scores.

Joint models across neighbours. Training on one language at a time erased the other: a recogniser trained on Zarma alone lost most of its Hausa, and a synthesis model trained only on Hausa carried a Hausa accent into Zarma. Training jointly avoided both at no measurable cost, and one synthesis model now holds Zarma, Hausa and French. For the languages of the Sahel, which share speakers, words and countries, one model per family of related needs looks more practical than one model per language.

12 Limitations and broader impact

Limitations. No native-speaker panel has rated any output in this report. Listening packets for synthesis, a review of spelling conventions, a sample of translations and a spot-check of the text identifier are built and not yet returned; an internal listening check by one team member is not a panel. Every number is an automatic measurement. Recognition is tested on one speaker, so its intervals describe that speaker and that transcriber. Some public evaluation speech may have been part of the public models' own training; we know this is so for the synthesis base model, and we therefore never use that corpus to test synthesis. Synthesis intelligibility is judged by recognisers, which find synthetic speech easier than natural speech, and the Hausa-accent measure is a proxy. The agent is tested mostly with synthetic users. Leaderboard comparisons in translation are indicative because the leaderboard does not state its settings. All our Zarma test material is read or prompted speech in a few voices; we have not measured conversational, telephone or children's speech, or the Zarma of different regions.

Broader impact. Speech technology in Zarma could give people who do not read French access to information and services in their own language. The same systems carry risks we take seriously. Voice cloning can be used to impersonate people; the agent speaks in one voice whose owner agreed to it, and we release no cloned voice. An assistant that answers health questions can do harm if it misunderstands or minimises a symptom; for that reason the care route is built as a fixed layer with native-written lines, and the agent is not offered to the public until its distress probes, written by native speakers with clinical guidance, are passed.

13 Conclusion

Zarma shows what building speech technology looks like when public data is thin. The obstacles we met were less about model capacity than about data rights, spelling, test references, language identification and translation of imperfect transcripts. Joint training across related languages worked well for both recognition and synthesis, and a first voice agent runs end to end, slowly and with gaps we can name. The next steps follow from those gaps: native-speaker evaluation with decision rules fixed in advance; a 100-speaker test set and consented speakers recorded for synthesis; native care lines and distress probes before any public demonstration; streaming and shorter translation paths for latency; and, further on, one model that hears and speaks Hausa, Zarma and French directly.

Acknowledgements

This work relies on public resources released by others: Zarma speech and text corpora, dictionaries, parallel sets, benchmarks and open models. We thank the people who recorded, transcribed, translated and published them, and the speakers whose voices made them possible.

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] P. Andrews et al. BOUQuET: dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of EMNLP 2025*, pages 27503–27523, 2025. https://aclanthology.org/2025.emnlp-main.1400/

[3] I. Caswell et al. SMOL: professionally translated parallel data for 115 under-represented languages. *arXiv:2502.12301*, 2025.

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

[5] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[6] D. M. Eberhard, G. F. Simons and C. D. Fennig, editors. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, 2023.

[7] C. Enguehard, M. Mangeot et al. Computerization of African languages–French dictionaries. In *Proceedings of the LREC 2014 Workshop on Collaboration and Computing for Under-Resourced Languages*, 2014. *arXiv:1405.5893*.

[8] H. Seydou Hanafiou. *Grammaire comparée français/soŋay-zarma*. IRSH, Université Abdou Moumouni, Niamey, 2007.

[9] A. Harrison, D. Harrison and M. J. Rueck. *Southern Songhay Speech Varieties in Niger*. SIL Electronic Survey Report 1999-004, 1999.

[10] M. K. Keita, E. Ibrahim, H. Alfari and C. Homan. Feriji: a French-Zarma parallel corpus, glossary and translator. In *Proceedings of the ACL 2024 Student Research Workshop*, 2024. *arXiv:2406.05888*.

[11] M. K. Keita et al. Grammatical error correction for low-resource languages: the case of Zarma. *arXiv:2410.15539*, 2024.

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

[13] République du Niger, Ministère de l'Éducation nationale. Arrêté n° 0215/MEN/SP-CNRE du 19 octobre 1999 fixant l'orthographe de la langue soŋay-zarma. Niamey, 1999.

[14] Peace Corps Niger. *Zarma Dictionary: Zarmaciine–English–Français*, revised edition, 2001.

[15] M. Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, 2017.

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

A Evaluation sets

Table 3: The frozen evaluation sets used in this report. Every training set was checked against each of them for shared text, recordings and speakers before training.

Set	Content	Used for
Recognition, test	One speaker of a public prompted-speech corpus; 1,060 segments, 1.5 h	Recognition
Recognition, dev	Another speaker; 486 segments, 1.1 h, without 22 segments whose references belong to another speaker	Selection
Zarma synthesis	140 held-out sentences without digits, 5 voices, 700 renders	Synthesis
Hausa synthesis	334 targets from 20 held-out speakers	Synthesis (Hausa kept)
French synthesis	300 sentences of a public read-speech set, 5 readers; 61 targets from 5 held-out Nigerien-accented speakers	Synthesis (French)
English–Zarma	150 held-out sentences of a public parallel set (69 single sentences, 81 from 23 documents)	Translation, synthesis text
French–Zarma benchmark	854 test sentences, 198 paragraphs; evaluation only	Translation
French–Zarma dictionary	401 held-out examples from two open dictionaries, grouped by headword	Translation
Agent	111 spoken turns: everyday, care and harmful requests; 25 real recordings of one speaker; silence and noise	Voice agent

B Detailed results

Table 4: Zero-shot recognition, folded CER (%) with 95% bootstrap intervals; strict CER on test for reference. Sizes are relative within one public family (S, M, L, XL). The older recogniser uses adapters trained for related Songhay languages and, as a control, for Hausa.

System	Size	Test	Dev	Test, strict
CTC head	S	31.0 [30.3, 31.6]	30.9 [29.6, 32.3]	33.7
CTC head	M	27.1 [26.5, 27.8]	29.1 [27.5, 30.6]	29.6
CTC head	L	24.5 [23.9, 25.2]	26.1 [24.8, 27.5]	26.7
CTC head	XL	23.9 [23.3, 24.7]	26.4 [25.0, 27.7]	26.7
LLM decoder	M	25.3 [24.6, 26.1]	28.1 [26.5, 29.7]	28.2
LLM decoder	L	**23.8** [23.0, 24.6]	**25.8** [24.2, 27.5]	26.9
LLM decoder	XL	23.1 [22.3, 23.9]	26.0 [24.4, 27.7]	25.9
LLM decoder, in-context examples	XL	25.9 [24.8, 26.9]	28.7 [27.2, 30.2]	28.5
Older recogniser, Koyra Chiini adapter		36.1 [35.5, 36.8]	36.6 [35.3, 37.9]	38.4
Older recogniser, Koyraboro Senni adapter		36.7 [36.0, 37.3]	35.9 [34.6, 37.2]	40.6
Older recogniser, Dendi adapter		49.7 [49.1, 50.4]	54.2 [53.1, 55.4]	50.9
Older recogniser, Hausa adapter (control)		44.0 [43.4, 44.7]	40.1 [38.8, 41.4]	45.1

Table 5: Synthesis. Zarma: folded CER (%) of 700 renders read back by the two judges, with intervals over sentences. Hausa-letter renders: renders of 700 whose CTC read-back contains a Hausa hooked letter. Hausa: folded CER on 334 Hausa targets; natural recordings score 3.6 [2.7, 4.8].

System	Zarma, judge 1	Zarma, judge 2	Hausa-letter renders	Hausa
Public base, zero-shot	11.8 [11.1, 12.6]	11.7 [11.1, 12.4]	7	2.9 [2.6, 3.3]
Hausa model, zero-shot	11.5 [10.7, 12.2]	13.7 [13.0, 14.4]	46	1.8 [1.5, 2.1]
Zarma from Hausa model, 25% Hausa	10.2 [9.4, 11.1]	10.1 [9.5, 10.8]	4	2.1 [1.8, 2.4]
Zarma from public base, 25% Hausa	11.0 [10.2, 11.9]	10.8 [10.1, 11.5]	2	3.0 [2.7, 3.4]
Joint, 40% Hausa	**10.3** [9.6, 11.2]	**10.5** [9.8, 11.2]	**3**	**1.9** [1.6, 2.2]

Table 6: Translation, chrF++ with 95% intervals. Top: 150 held-out English–Zarma sentences. Bottom: French–Zarma on the public benchmark (854 sentences), on 401 dictionary examples, and from recogniser transcripts of unseen voices (802 sentences). Values without intervals were not computed by us.

	English–Zarma		English–Hausa (control)	
System	en→dje	dje→en	en→ha	ha→en
Public MT, zero-shot	8.5 [7.4, 9.7]	21.2 [19.0, 24.2]	33.0	45.4
9B LLM (African languages), zero-shot	14.6 [13.4, 15.6]	18.8 [17.3, 20.9]	44.6	51.5
Ours, first version	41.7 [36.4, 51.4]	40.4 [36.2, 45.9]	40.7	48.0
Ours, recogniser-robust	42.4 [37.3, 51.1]	40.8 [37.0, 45.8]	40.0	46.5
Ours, French–Zarma	43.1 [37.8, 52.8]	42.0 [37.9, 47.3]	41.5	46.6
	Benchmark		Dictionary	
System	fr→dje	dje→fr	fr→dje	dje→fr
Public MT, zero-shot	6.7	14.4	4.5	10.8
Best public system (leaderboard)	24.5	23.8		
Through English	24.4 [23.6, 25.3]	26.2 [25.3, 27.1]	26.1 [25.1, 27.1]	20.3 [19.1, 21.5]
Ours, recogniser-robust, direct	24.7 [23.9, 25.6]	24.8 [23.8, 25.8]	26.4 [25.4, 27.5]	22.0 [20.8, 23.2]
Ours, French–Zarma	**24.7** [23.8, 25.6]	**29.2** [28.1, 30.2]	**28.2** [27.1, 29.4]	**33.4** [31.8, 35.2]
From recogniser transcripts, dje→fr: through English 17.7 [16.9, 18.4]; recogniser-robust 18.8 [18.0, 19.6]; ours 27.0 [25.8, 28.3]				

Table 7: Hosted voice agent, third version: time per stage over 13 warm turns with real Zarma input, in seconds. Stage times are differences between consecutive events; the reply delay adds an assumed 0.3 s end-of-speech allowance and the leading silence of the first audio. The first turn after idle took 202 s.

Stage	Median	95th percentile	Min	Max
Recognition	1.15	1.71	0.57	1.76
Translation into French	1.63	3.21	1.03	3.26
Language model, first sentence	1.11	1.91	1.00	1.91
Translation into Zarma and speech, first sentence	2.19	2.41	1.46	2.50
End-of-speech allowance and leading silence	0.86	0.88	0.46	0.88
Reply delay	6.80	7.97	5.92	8.14