One Speech Synthesis Model for Hausa, Zarma and French: Adding Languages One at a Time

Namu AI Lab
namuai-lab.com

Abstract

We grew one voice-cloning speech synthesis model from Hausa to Zarma and then to French, and measured at every step whether the new language cost the earlier ones. Starting from a public massively multilingual model, we fine-tuned it on about 128 hours of read Hausa (version 1.0), then on about 3.6 hours of Zarma from five speakers with Hausa in 40% of the training samples (version 1.1), then added French at about a fifth of the samples (version 1.2). Two recognisers of different architecture read the renders back (one in the French stage); we report folded character error rate (CER) with 95% intervals, and count Hausa hooked letters in Zarma read-backs as a proxy for a Hausa accent. Training on Hausa alone kept the base model's French but, by this proxy, carried Hausa sounds into Zarma: 46 of 700 Zarma renders contained a Hausa hooked letter, against 7 for the base model. Training on Zarma with 40% Hausa lowered Zarma CER by 1.1 to 3.2 points against the Hausa model, removed the hooked letters (3 of 700) and moved Hausa by $+0.11$ points (interval -0.12 to $+0.35$); the same Zarma data with Hausa in a quarter of the samples, started from the public base instead, ended 1.2 points behind the Hausa model on Hausa. The three-language model holds French, Zarma and Hausa within tolerances fixed before scoring, with French judged by one recogniser on a test set we repaired after the registered set failed. Several effects outside the model were as large as the effect of training: the choice of voice spread Zarma CER from 9 to 16%, digits left as numerals raised it to about 20%, silence in a reference clip became long pauses in the renders, one truncated reference transcript made every system fail on its voice, and a pitch-based estimator put women at 14% of the French hours where a more careful one found 34%. All results are automatic measurements; no native-speaker panel has rated any output.

1 Introduction

Many languages are spoken in Niger, and many of its people speak more than one. Hausa and Zarma are among the most widely spoken, and French is a working language of the state. A speech service for Niger therefore needs a voice that can speak all three. Building one model per language multiplies the models to host, update and test, and wastes what related languages could share. Building one model for all of them raises a question that every multilingual system meets: what a model keeps of the languages it already speaks when it learns a new one.

This report follows one speech synthesis model through three versions. We started from a public voice-cloning model trained on more than 600 languages, fine-tuned it on Hausa, then taught it Zarma while keeping Hausa, then practised French while keeping both. At each step we measured

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

the earlier languages as well as the new one, with rules for what counts as a loss written before any score was computed. Our companion report on Zarma speech technology summarised the Zarma and French steps in one section [17]. Here we go deeper: we describe the Hausa stage, the choices that decided each step, and the measurement that told us whether a language was kept. We describe the models we used by type and scale, and our own models by the version and what they were trained to do.

Our contributions are the following. (1) A record of growing one synthesis model through three versions, with the data described in aggregate and the decision rules that chose each step (Section 3). (2) A measurement protocol for three languages that reads speech back with two recognisers of different architecture, normalises each language's spelling, tracks a proxy for accent and applies a per-language tolerance (Section 4). (3) Evidence, from an accent proxy, that training on one language moved a neighbouring language's sounds while leaving a more distant language untouched, and that mixing the earlier language back into training kept it (Sections 5 to 7). (4) A set of factors outside the model, namely the voice, digits, the reference clip, reference transcripts and inline sound tags, whose effects were as large as the effect of training (Section 8). (5) A case in which the estimated share of women in the training data more than doubled when we changed the pitch estimator (Section 9).

2 Related work

Multilingual and zero-shot synthesis. Speech synthesis that clones a voice from a short reference clip has moved from single-language systems to models trained on many languages at once, using neural codec language models [26], flow matching [13] and multilingual zero-shot designs [4, 5]. Massively multilingual work has extended synthesis to more than a thousand languages, often from one reader of religious text per language [22]. Earlier studies showed that one model can share parameters across languages [18] and transfer a voice from one language to another, which requires separating the speaker from the language [27]. A review of multilingual data strategies surveys how data in other languages has been used to make up for scarce target data [9], and recent work has built multi-speaker synthesis of English in many African accents [21]. Our work starts from a public multilingual model and studies what happens to each language as others are added.

Forgetting and replay. Neural networks trained on a new task tend to lose earlier ones [10, 16]. Known remedies are replay, which mixes earlier data back into training [23], penalties on changing the weights that matter for earlier tasks [12], and distillation from the earlier model [14]. In multilingual translation and language modelling, the share of each language in training decides which languages improve and which degrade [2, 6]. We use the simplest of these tools, replay at a fixed share, and measure its effect per language.

Measuring synthesis with recognisers. The intelligibility of synthesised speech is commonly estimated by asking a recogniser to transcribe it and scoring the transcript, which ties the result to the recogniser's own errors; confidence intervals for such measures need care [25]. We use two recognisers of different architecture, cluster bootstrap intervals [3] and decision rules written before scoring [20]. Pitch for gender estimates comes from YIN-family trackers [8, 15] and normalised cross-correlation [24]. We keep each training mixture with a datasheet [11].

Hausa and Zarma. Hausa is a Chadic language written in a Latin alphabet with the hooked letters ɓ ɗ ƙ ƴ; tone and vowel length are not written in everyday spelling [19]. It has public read speech in collections such as Common Voice [1] and FLEURS [7]. Zarma, a Songhay language, has far less: our companion report surveys its public speech and text and finds very little that is both open and transcribed [17]. The two languages are not related, but they share a country and many loanwords, and Hausa spellings with hooked letters appear in Zarma transcripts [17].

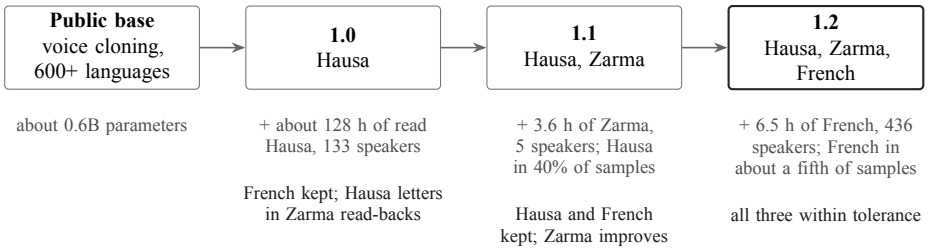


Figure 1: The three versions. Each starts from the one before; the line under each box gives the data added, and the last line what the measurements in Sections 5 to 7 found.

3 One model, three versions

Figure 1 shows the path. Each version starts from the one before and adds one language.

The public base. The starting point is a public voice-cloning synthesis model of about 0.6B parameters, built on a small language model and trained on speech in more than 600 languages, Hausa, Zarma and French among them. Given a reference clip with its transcript, a language tag and a text, it predicts masked audio codes of a neural codec over a fixed number of refinement steps and decodes them to speech in the reference voice. It also accepts inline tags for some non-speech sounds, such as laughter and sighs. Its training data included the public Zarma corpus we later train on; what French it saw is not known to us.

Version 1.0: Hausa. We fine-tuned the base on about 128 hours of read Hausa from public speech datasets, 133 speakers in all. A little over half of the hours come from a single narrator. The model was trained before the Zarma work began and was the Hausa model of our earlier voice agent work.

Version 1.1: Zarma, keeping Hausa. The Zarma data is about 3.6 hours from the five training speakers of a public corpus of prompted spontaneous speech, the only transcribed Zarma speech we could train on [17]. A filter for synthesis-grade clips (length, alignment to the transcript, signal-to-noise ratio, speaking rate) kept 3.6 of the 7.45 hours of these speakers' segments; one speaker, who answers in long lines, kept only 0.15 hours. Numbers were spelled out in Zarma before training. Hausa from version 1.0's own training data was mixed into 40% of the training samples (about 2.5 hours). Two choices were made by rules written before any score was read: the starting point (the public base or version 1.0, both tried on the same data) and the Hausa share (about a quarter first, 40% as a pre-registered fallback).

Version 1.2: French, keeping both. Version 1.2 continues from 1.1. The French data is about 6.5 hours from 436 speakers, mostly audiobook readings in standard French, about an hour of spontaneous speech and about half an hour of Nigerien-accented read speech. We had aimed for 15 to 20% Nigerien speech; the public source holds less than an hour of it, and what passed the filter came to under a tenth of the French hours. French made up about a fifth of the training samples, with Zarma and Hausa keeping version 1.1's ratio (about half and a third of the samples). A second arm with French at about a third of the samples was trained beside it, and a pre-registered tie-break chose between them.

Keeping each version's data. Version 1.0 was trained on the whole Hausa corpus. Versions 1.1 and 1.2 each name the exact clips they were trained on, stored as a numbered training mixture with a datasheet [11], so that the next version can replay them. Building this record taught us that the clips themselves must be kept as audio beside the recipe that cut them. The Zarma clips of version 1.1 had been lost with a temporary machine. Re-cutting them from the source on another GPU type moved the boundaries of 16 of 2,030 clips by 0.02 to 1.30 seconds, because low-precision alignment differs across hardware, as we had found for test sets [17]. Comparing the audio codes of both candidate cuts with the codes the model was trained on settled each case: the right cut agreed on 0.976 to 1.000

of its codes and the wrong one on at most 0.051. The French clips, re-encoded on the same GPU type, matched all 3,205 training rows exactly.

4 Measurement

Read-back intelligibility. A render is intelligible to the extent that a recogniser reads back the intended text. Judge 1 is a public multilingual recogniser with an LLM decoder, about 3B parameters; judge 2 is a CTC recogniser of the same family and size, which has no language model and cannot write fluent text that is not in the audio. For the Hausa stage we used the 1B members of the same family. In the French stage judge 1 did not fit in the training machine's memory, so judge 2 scored every system there. We report corpus CER.

Spelling. Each language has its own normaliser. For Hausa, strict scoring keeps the hooked letters and folded scoring maps them to plain letters. For Zarma, folded scoring maps *ŋ* to *n* and *ɲ* to *ny*, removes nasal marks, collapses doubled vowels and ignores hyphens and apostrophes; our companion report explains why these differences are spelling convention rather than hearing [17]. For French, folded scoring removes accents, ligatures and apostrophes and spells digits out on both sides. Folded CER is the headline.

Accent. We have no Zarma listeners yet, so we use a proxy for a Hausa accent in Zarma: the number of Zarma renders whose judge 2 transcript contains a Hausa hooked letter. Natural Zarma speech triggers it in 1 of 888 segments. For Hausa we count the reverse, the share of hooked letters that a recogniser writes as plain letters. Both are proxies; native listeners will decide.

Voice, length and failures. Speaker similarity is the cosine between speaker-verification embeddings of a render and of natural clips of the same speaker, reported relative to a ceiling from two natural clips of that speaker. We calibrated a same-speaker threshold per language at the equal error rate: 0.418 for Zarma (5.6%, seven speakers only), 0.338 for Hausa (7.6%) and 0.387 for French (0.9%, 15 speakers). Duration is seconds per letter after trimming silence, relative to the reference clip or the natural recording. A render fails if it is empty, truncated, longer than 2.5 times the expected length, loops, or reads back at more than 60% CER.

Intervals and rules. Every interval is a 95% percentile bootstrap with 2,000 resamples, clustered by sentence, document or speaker as the test requires; differences between systems are paired on identical texts, references and seeds. Rules written before any score set the tolerance: a later version passes for an earlier language if its CER is at most 0.5 points worse on the point estimate and at most 1.0 point worse at the upper end of the interval (for French, the rule fixed only the point estimate); speaker similarity fails if it drops by 0.01 or more with the interval below zero; durations must stay within 10% and failures within two of the reference system. Checkpoints were chosen on development sets only, with a guardrail on the earlier language. Table 1 lists the test sets.

5 Version 1.0: training on Hausa alone

Figure 2 summarises every language across the four models under judge 2, and Figure 5 gives each paired difference with its interval. We start with what the Hausa stage did.

Hausa on new speakers. On 258 broadcast news clips whose speakers and sentences are outside its training data, version 1.0's renders read back at 5.4% strict CER under judge 2, against 6.7% for the public base and 10.9% for the natural recordings (Figure 3a). After compression by a neural speech codec of the kind a spoken-dialogue model consumes, the gap widened to 7.4 against 11.9%. Judge 1 saw a tie (5.9 against 5.7% strict, 4.9 against 4.7% folded), but word error rate separated the two (16.9 against 19.6%). Speaker similarity was the same (0.697 against 0.696, with a ceiling of 0.774). We did not compute an interval for the difference between the two models on this set; each lies well

Table 1: Evaluation sets. Every training set was checked against each of them for shared speakers, clips and texts.

Set	Content	Caveat
Zarma test	140 held-out sentences of a public parallel set without digits, each in five voices: 700 renders	voices seen in training
Zarma dev	246 held-out sentences, five voices in rotation (100 in the French stage)	selection only
Hausa test	334 targets from 20 speakers, each cloned from another clip of the same speaker	in 1.0's training data
Hausa, new speakers	258 broadcast news clips; references from 51 speakers found by embedding clustering	Hausa stage only
Hausa dev	5 speakers × 20 sentences	selection only
French test	300 sentences of a public read-speech set [7] read by 5 held-out audiobook readers	one judge
Nigerien French	61 targets from 5 held-out speakers	small
French dev	5 held-out readers × 20	selection only

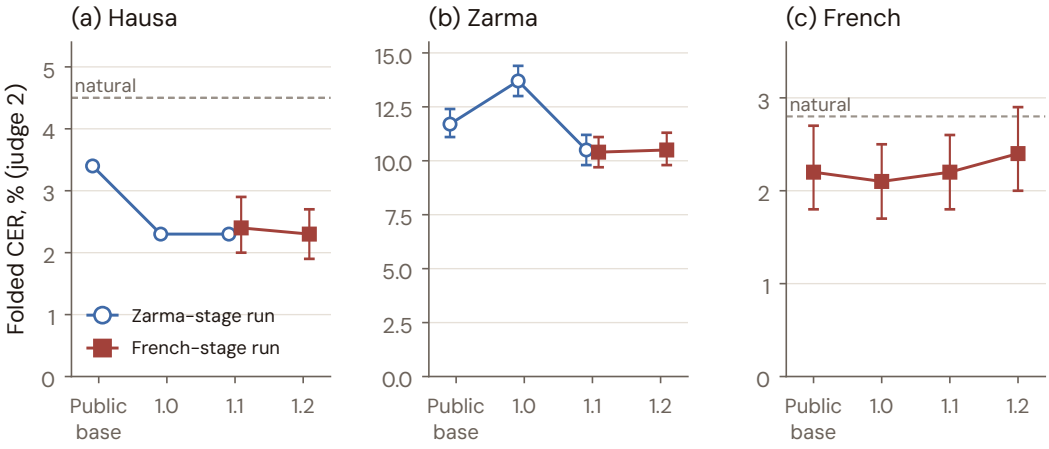


Figure 2: Each language across the public base and the three versions, folded CER read back by judge 2, with 95% intervals where we computed them. Two evaluation runs: the Zarma-stage run (open circles) and the French-stage run (filled squares), which re-rendered version 1.1 on other hardware. Dashed lines: natural recordings. Hausa is tested on speakers that are in version 1.0's training data.

below the natural recordings (by -4.1 $[-5.8, -2.8]$ and -4.3 $[-5.6, -3.0]$ points under judge 1). Every failure, one render for version 1.0 and two for the base, came on inputs longer than 60 words, so we split text into units of at most about 30 words. On sentences from its own training data, the recogniser wrote the hooked letters of version 1.0's renders as plain letters 4.6% of the time, against 9.2% for the base and 9.8% for natural speech, and the 3B judge 1 read its Hausa at 1.8% [1.5, 2.1] folded CER against 2.9% [2.6, 3.3] for the base; those numbers favour the model that trained on them.

Hausa letters in Zarma. Version 1.0 had seen no Zarma since the base. Judge 1 found no loss of Zarma intelligibility: -0.41 $[-0.84, 0.00]$ points against the base in our first measurement and -0.36 $[-0.77, +0.02]$ when we re-rendered both. Judge 2 found a loss of 1.82 $[+1.38, +2.24]$ and 1.95 $[+1.48, +2.36]$ points. The accent proxy suggests why the judges disagree: 37 to 46 of 700 version 1.0 renders contained a Hausa hooked letter, against 3 to 7 for the base (Figure 3b). We read this as a drift toward Hausa sounds rather than a loss of Zarma: version 1.0 seems to say Zarma with Hausa-like consonants, which a recogniser with a language model can repair and a CTC recogniser writes down. Only native listeners can confirm it. The same split between judges appeared on Hausa

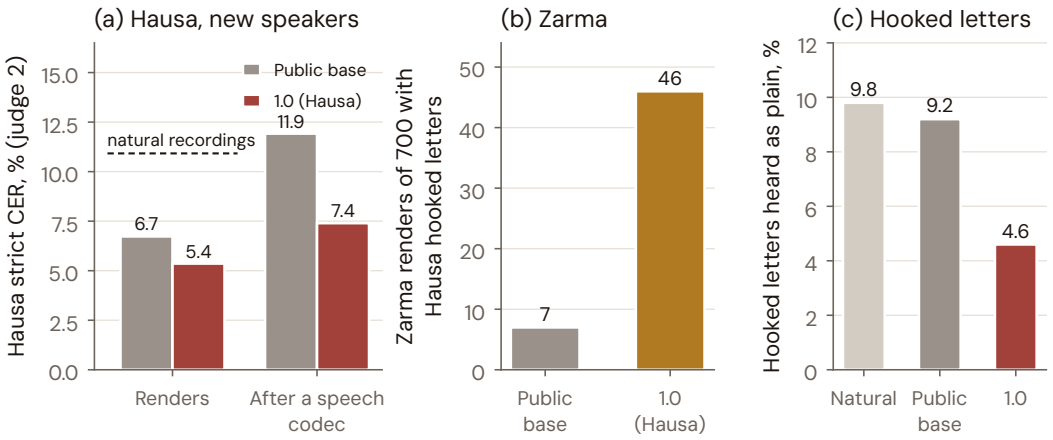


Figure 3: What training on Hausa alone did. (a) Hausa on 258 broadcast clips from speakers outside the training data, strict CER read back by a 1B CTC recogniser, as rendered and after a neural speech codec; the dashed line is the natural recordings. (b) Zarma renders, of 700, whose read-back contains a Hausa hooked letter; natural Zarma triggers this in 1 of 888 segments. (c) Share of Hausa hooked letters a 1B recogniser writes as plain letters, on sentences that are in version 1.0's training data.

itself, where judge 1 tied the two models and judge 2 did not.

The language tag. The base takes a language tag, and the Hausa stage changed how the model used it. For the base, the Zarma tag measured no better than no tag (judge 2 favoured no tag by 0.59 [+0.28, +0.91] points), while the Hausa tag cost 2.61 [+2.20, +2.98] points against the Zarma tag under judge 2 and raised the hooked-letter count to 21. Version 1.0 ignored the tag: with the Zarma or the Hausa tag its Zarma differed by −0.01 [−0.37, +0.36] points under judge 1. We never use the Hausa tag for Zarma.

French was kept. Measured later, before any French training, version 1.0's French read back at 2.1% [1.7, 2.5] against 2.2% [1.8, 2.7] for the base, a difference of −0.13 [−0.43, +0.17] points. On Nigerien-accented French the two differed by +0.83 [−0.36, +1.67] points, an interval too wide to read. Training on about 128 hours of one language moved Zarma, which borrows Hausa words and spellings [17], and left French in place. We can only guess why; one reading is that fine-tuning pulls the model's sounds toward the trained language most where the two languages share letters and loanwords.

6 Version 1.1: adding Zarma

Where to start. The Zarma stage asked first whether to start from the public base or from the Hausa model, since the two judges disagreed on which was better at Zarma. The pre-registered rule called it a tie and trained both on identical data (Figure 4a; Appendix Table 2). Starting from the Hausa model learned Zarma better: against the base start it was ahead by 0.79 [0.47, 1.12] points under judge 1 and 0.66 [0.38, 0.94] under judge 2, and its gain over the zero-shot base was about twice as large (−1.57 against −0.78 points under judge 1). The base start also could not reach version 1.0's Hausa: it ended 1.22 [+0.93, +1.51] points behind on the Hausa test and failed the Hausa guardrail at every development checkpoint (Figure 4c). Its Hausa stayed near the base's own level (3.0 against 2.9%). Learning Hausa first, then Zarma with Hausa kept, gave the better result in both languages.

How much Hausa to keep. With Hausa in about a quarter of the samples, the Hausa test moved by +0.27 [+0.09, +0.45] points: inside the tolerance but measurable. A second arm from version 1.0, with Hausa at 40%, had been added by a rule fixed before any score was read. That arm moved Hausa by +0.11 [−0.12, +0.35] points under judge 1 and −0.03 [−0.29, +0.21] under judge 2, no

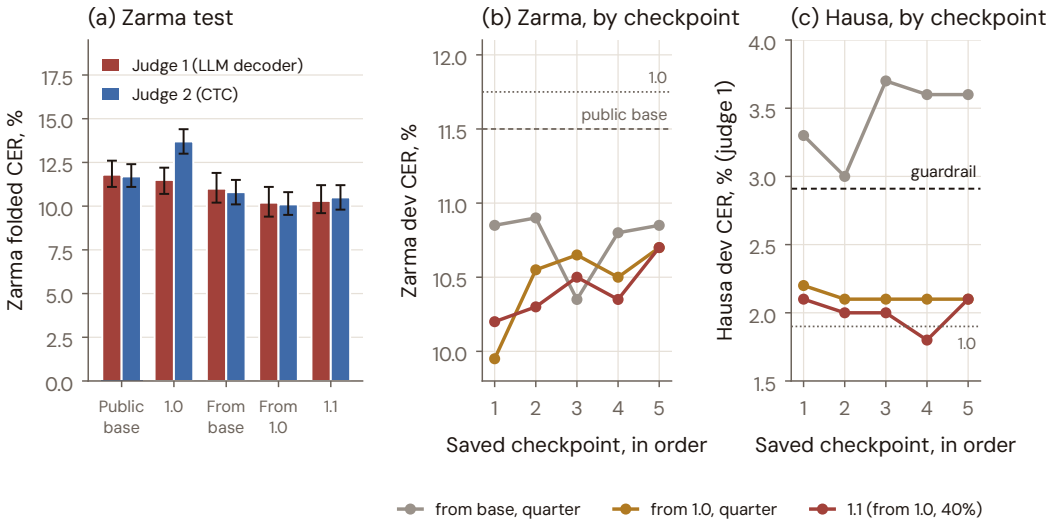


Figure 4: Adding Zarma. (a) Zarma test CER under both judges for the two zero-shot models and three fine-tunes on the same Zarma data: from the public base with Hausa in about a quarter of samples, from version 1.0 with the same share, and from 1.0 with 40% (version 1.1). (b) Zarma development CER (mean of the two judges) at each saved checkpoint, in training order; the first is after about four passes over the data. (c) Hausa development CER at the same checkpoints, with the guardrail written before any score.

measurable change, for a small Zarma cost against the quarter share ($+0.09$ $[-0.18, +0.36]$ under judge 1, $+0.35$ $[+0.10, +0.61]$ under judge 2). We took 40% as version 1.1.

Version 1.1. On the Zarma test, version 1.1 reads back at 10.3% [9.6, 11.2] under judge 1 and 10.5% [9.8, 11.2] under judge 2. Against the base that is -1.48 $[-1.84, -1.10]$ and -1.24 $[-1.56, -0.92]$ points; against version 1.0, -1.11 $[-1.55, -0.65]$ and -3.19 $[-3.60, -2.72]$. The Hausa hooked letters left its Zarma: 3 renders of 700. Its Hausa read back at 1.9% [1.6, 2.2], with speaker similarity at 0.94 of the ceiling, as for version 1.0. Its French, measured later, was unchanged from the base: -0.07 $[-0.36, +0.21]$ points. No render failed in any Zarma system.

The first checkpoint was the best. In every arm that started from version 1.0, the first saved checkpoint, after about four passes over the data, was the best on Zarma development data; later checkpoints drifted up by 0.4 to 0.8 points, within one interval of each other (Figure 4b). The base had already trained on this corpus, so more passes over the same five speakers taught nothing new. The gains are modest (about 1.5 points against the base) for the same reason, and the next gain needs new speakers recorded for synthesis.

7 Version 1.2: adding French

Step zero. We planned the French stage on the assumption that two fine-tunes without French had worn it down. We measured that assumption first, and it was wrong: version 1.1 read French back at 2.2% [1.8, 2.6], level with the base (Section 6), and so did version 1.0. The French stage therefore had nothing to recover. Its purpose became keeping French by practice, so that later versions do not depend on what survives by chance.

Choosing the share. Both French arms, at about a fifth and about a third of the samples, passed the development guardrails. The rule written before any score took the smaller share. The larger-share arm was 0.30 points better on French and on Zarma development data, both inside the development intervals, and 0.23 points worse on Hausa.

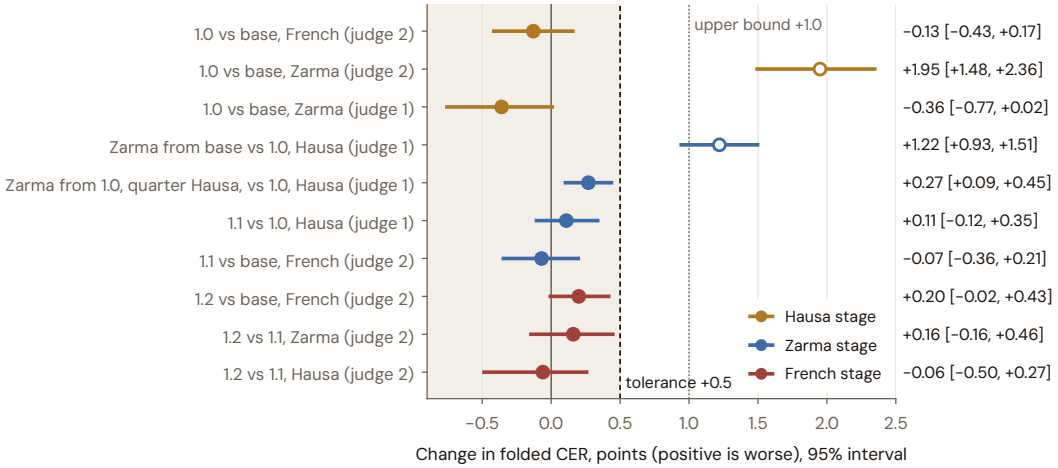


Figure 5: What each step cost each language: paired differences in folded CER, in points (positive is worse), with 95% intervals. The shaded region is the pre-registered tolerance on the point estimate; the dotted line is the limit on the upper end of the interval. Open circles mark changes outside the tolerance. Zarma: 700 renders, sentence clusters; Hausa: 334 targets, speaker clusters; French: 300 sentences, sentence clusters.

The gate. Every number in this stage comes from judge 2 alone. On the test sets, version 1.2 reads French back at 2.4% [2.0, 2.9], against 2.2% for the base (+0.20 [−0.02, +0.43]); Zarma at 10.5% [9.8, 11.3], against 10.4% for version 1.1 (+0.16 [−0.16, +0.46]); and Hausa at 2.3% [1.9, 2.7], against 2.4% (−0.06 [−0.50, +0.27]). Every difference is inside the tolerance, and no render failed in any of the three test sets (Figure 5). Speaker similarity dropped measurably in Zarma, by 0.005 [0.002, 0.008], under the 0.01 tolerance. French renders became slightly slower (duration ratio +0.019 [+0.014, +0.023]). On the five Nigerien-accented French speakers, CER moved from 5.5 to 4.7% (−0.78 [−2.62, +0.82], a set too small to decide) and speaker similarity rose by 0.034 [0.021, 0.054]: by the speaker-verification measure, the model now clones these voices more closely.

A reference cut short. The French test as first registered failed. Every system, the base included, read back at 20 to 23% CER with 49 to 55 failures, against 2.5% for the natural recordings. All failures sat in one of the five French voices, at 89 to 104% CER; the other four voices read back at 2.1 to 3.1%. That voice's reference clip continues for a second sentence that its dataset transcript leaves out, so judge 2 scored the clip against its own transcript at 169% CER, while the other nine French voices' references scored 0 to 8%. Prompted with a transcript that covers half the audio, the model lost its alignment on every sentence. The pre-registered comparison failed (+2.43 [+0.82, +4.13] points), but the difference comes from the failed renders of that one voice. We repaired the set with a rule that looks only at the reference and the recogniser, decided without looking at any render: a voice whose reference the recogniser contradicts by more than 15% CER takes the first of its held-out clips that passes. All systems re-rendered that voice's sentences, and the gate above is on the repaired set. The check existed in the Zarma stage as a column in the report but not as a filter; every evaluation voice now has to pass it before any render.

8 What else moves the result

Training moved Zarma CER by one to three points. Several things outside the model moved it as much or more (Figures 6 and 7).

The voice. Across the five Zarma voices, CER ranged from 9.2 to 15.6% for the base and from 9.1 to 15.2% for version 1.0; the two models differed by at most 1.2 points on any voice (Figure 6a). The hardest voice belongs to the speaker whose long answers mostly exceeded the clip-length limit, so that only 0.15 hours of that speaker entered training. In the fine-tune with Hausa in about a quarter of

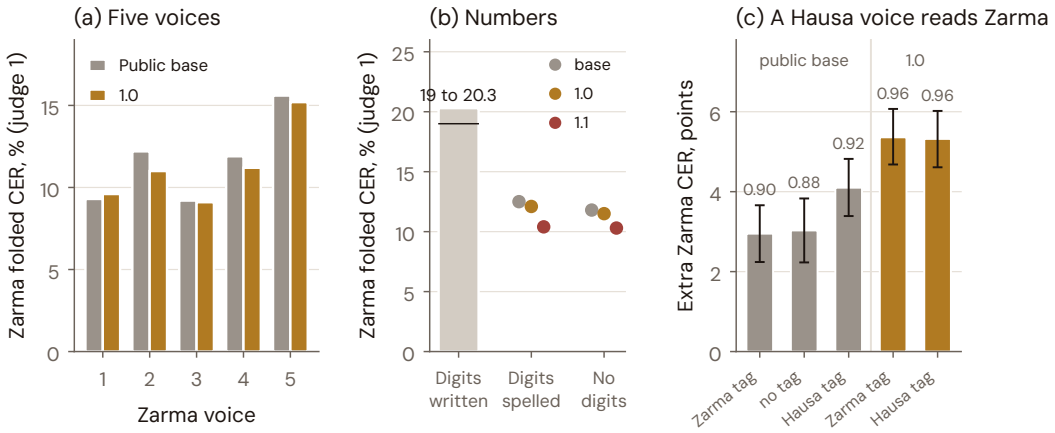


Figure 6: Effects outside the model, Zarma CER under judge 1. (a) The five Zarma voices with the public base and version 1.0. (b) Ten sentences with digits, written as numerals (range over five zero-shot systems) and spelled out in Zarma, beside the 140 sentences without digits. (c) The increase in CER when one Hausa voice reads the Zarma sentences instead of the Zarma voices, for the base and version 1.0 with different language tags; the number above each bar is the speaker similarity relative to its ceiling.

the samples, the one we have per voice, every voice improved and that one stayed the hardest (16.6 to 14.1% under judge 1).

Digits. Sentences with digits left as numerals read back at 19.0 to 20.3% for every zero-shot system, against 11 to 12% for sentences without digits. Spelled out in Zarma first, the same sentences scored 12.5% for the base, 12.1% for version 1.0 and 10.4% for version 1.1, close to the rest of the test (Figure 6b). A number-spelling front end is part of the synthesis system.

A voice from the other language. One Hausa voice reading the Zarma test sentences lost 2.95 [2.24, 3.66] points against the Zarma voices with the base and 5.36 [4.68, 6.07] with version 1.0, while keeping 0.90 and 0.96 of its speaker-similarity ceiling (Figure 6c). The Hausa model kept the Hausa speaker most faithfully and was the least intelligible in Zarma. By the speaker embedding, a voice stays the same person in the other language, but it is harder to understand there; for now we give each language its own voice.

Silence in the reference clip. A voice-cloning model copies the reference's rhythm along with its timbre: renders ran at 0.95 of their reference's seconds per letter. An earlier pair of reference clips for a consenting team member's voice was 28% silence in Zarma and 10% in Hausa, and the renders paused in mid-sentence and ran about 30% longer than the other voices'. Trimming the leading silence and cutting pauses over 0.15 seconds to 0.12 seconds shortened the clips from 6.7 to 4.8 seconds (Zarma) and from 8.7 to 7.0 seconds (Hausa). Over four sentences in each language, long silences in the renders fell from 9.8 to 3.3 seconds in Zarma and from 6.0 to none in Hausa (Figure 7a). Read-back CER moved from 8.2 to 10.1% in Zarma and from 0.9 to 1.5% in Hausa, within the noise of four sentences. A reference clip should be chosen like a recording for a studio voice: its pace, its pauses and its transcript all reach the output.

Inline sound tags. The base model accepts tags for laughter and sighs. We tested them on version 1.1 in the same team member's voice with one render per condition, an observation and not a measurement of the model's ability (Figure 7b). In Hausa, a laughter tag between two short sentences lengthened the render from 2.32 to 3.12 seconds and added about 0.8 seconds of sound. In Zarma, the same tag lengthened it from 2.20 to 2.73 seconds but added no sound: the extra time was silence, including a pause of about 0.35 seconds where the tag sits. A sigh tag at the start of a Hausa sentence that begins with a voiced consonant gave it about 0.6 seconds of unvoiced sound before its first voiced sound; for the Zarma sentence, which begins with a voiceless consonant, our measure could

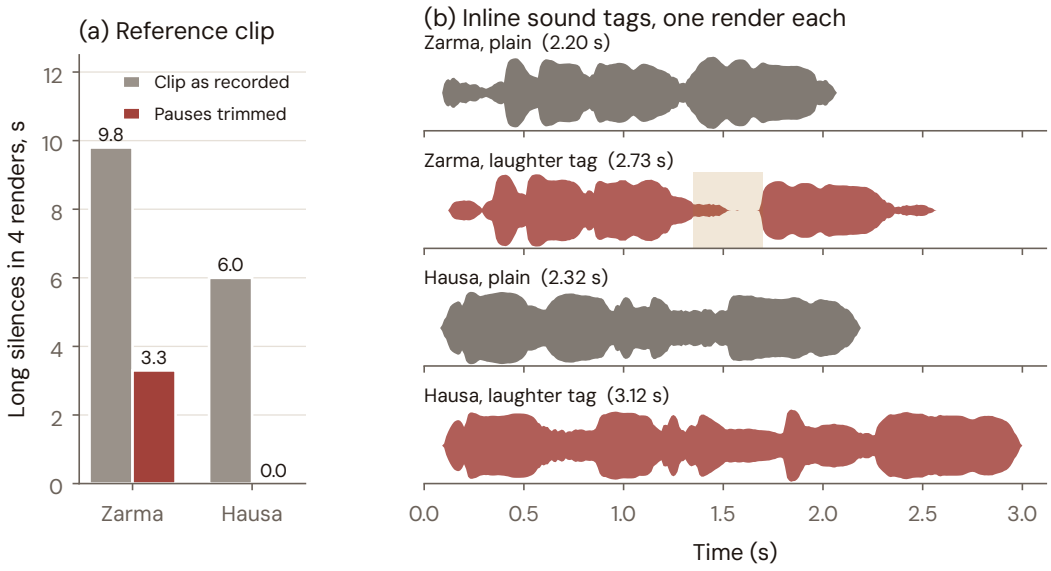


Figure 7: What the reference clip and the text carry, version 1.1 in a consenting team member's voice. (a) Total length of long silences in four Zarma and four Hausa renders, with the reference clip as recorded and with its pauses trimmed. (b) Amplitude envelopes of one render each of a short greeting, *Fofo! Mate ni go? Ay go baani.* in Zarma and *Sannu! Yaya kuke? Ina lafiya.* in Hausa, plain and with a laughter tag between the second and third sentence; the shaded span marks a pause of 0.35 s.

not separate a sigh from the consonant. With one render each we cannot say why; one possibility is that a tag works better in the language with more training data. Expressive speech in Zarma will need recordings that contain laughter and sighs.

9 Who is in the data

We estimate speaker gender from pitch, because no source labels it for every clip. In the French stage, a fast detector from a general audio library, based on normalised cross-correlation [24], with a median pitch of 165 Hz or more counted as female, put women at about 14% of the French training hours: 50 of 436 speakers (Figure 8). Three of the five French test voices are women by the same pitch rule, so this looked like a mismatch between training and test.

When we prepared the training data as a versioned record, we checked the estimator against pitch estimates from a phonetics toolkit for 64 Hausa speakers. It disagreed on 20 of them, each time calling male a speaker whom the toolkit placed as female. A probabilistic YIN tracker [15] on up to five clips per speaker, with a gap between the male band (below 160 Hz) and the female band (above 170 Hz) and the speakers in between counted as unknown, agreed on all 64. With it, the French pool has 142 women, 269 men and 25 speakers of unknown gender, and women hold 33.9% of the hours: the careful tracker calls women 92 speakers that the fast detector called men. The Zarma data is 32.6% women by hours (two of five speakers), and the Hausa mixed into training only 9.2%, because two thirds of it is one male narrator.

The French result itself did not single out the women's voices. Per voice, version 1.2 read the three women's voices back at 2.54, 3.11 and 1.97% against 2.40, 2.99 and 1.47% for the base, and the two men's voices at 1.49 and 3.04% against 1.75 and 2.58%, with an overall change of $+0.20$ points. The lesson concerns reporting. A claim about balance in a speech dataset is a claim about the estimator as much as about the data, so the estimator should be checked on labelled speakers, its unknowns reported, and its name stated next to the number.

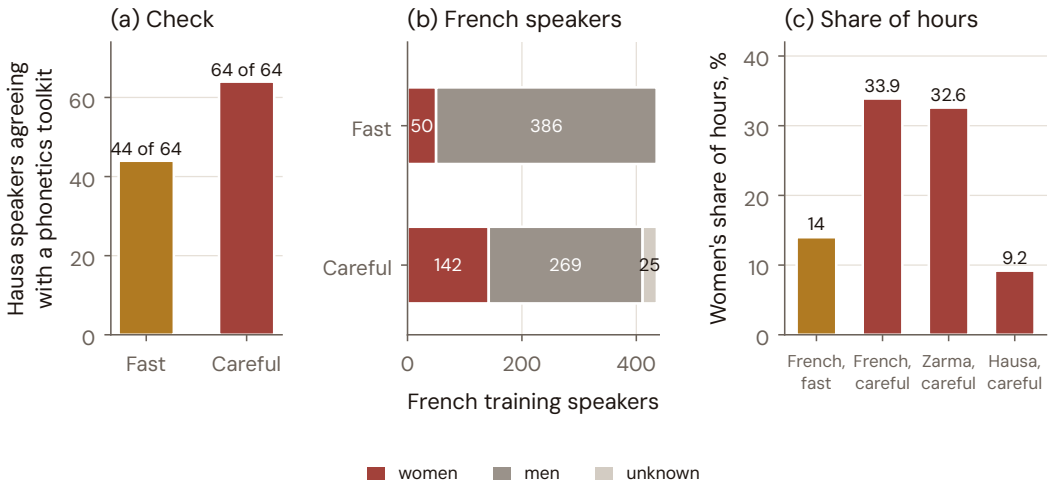


Figure 8: Gender estimated from pitch. (a) Of 64 Hausa speakers, how many each estimator assigns the same gender as a phonetics toolkit's pitch estimates. (b) The French training speakers by estimator. (c) Women's share of training hours, by estimator and language; the Hausa share is for the Hausa mixed into versions 1.1 and 1.2. Neither estimator is a human label.

10 Discussion

Practise every language in every batch. Replay at a fixed share was enough here. With 40% of samples in Hausa, adding Zarma left Hausa unmeasurably changed; with about a quarter, the loss was measurable though inside the tolerance. Once all three languages were in every batch, adding French cost neither of the others a measurable change in CER. A model that grows one language at a time needs the earlier languages' data kept as audio and replayed, which is why versions 1.1 and 1.2 name the exact clips they trained on.

The order mattered. Learning Hausa first helped Zarma. Starting the Zarma stage from the Hausa model learned Zarma about twice as well as starting from the base, from the same data, and kept Hausa; starting from the base could not reach version 1.0's Hausa with the same replay. One reading is that 128 hours of read speech taught the model things that transfer, such as clean reading at a steady pace, and that Zarma's few hours of prompted spontaneous speech could not teach these on their own. We have not tested other orders.

Measure before planning. Our French plan assumed a loss that step zero showed did not exist; our accent concern for Zarma came from a disagreement between judges that one judge alone would have hidden. Both were found by measuring every language at every step and by reading two judges of different architecture side by side. A recogniser with a language model can repair sounds that a CTC recogniser writes down, so a synthesis gate that uses only the first can miss an accent.

The data is the limit. The Zarma fine-tunes that started from version 1.0 were best at their first saved checkpoint. The five Zarma speakers are ones the base had already heard, and more passes over their 3.6 hours added nothing. The next versions need new, consented Zarma speakers recorded for synthesis, and a French pool balanced by a checked gender estimate.

11 Limitations and broader impact

Limitations. No native-speaker panel has rated any output in this report. Listening packets for Zarma and Hausa are built and have not been run; a French panel is designed and not built; a short listening check by one team member is not a panel. For the Hausa model, an earlier listening packet

reached native listeners, but their verdict came back as an informal report without per-item scores, so we report no result from it. Every number is an automatic measurement. Read-back CER measures intelligibility under a recogniser; it does not measure naturalness, tone or vowel length, and the accent measure is a proxy. The Zarma test voices are the five speakers we trained on, which the base had also seen: the sentences are held out, the voices are not, and a new voice may gain less. The Hausa test speakers are in version 1.0's training data, which favours versions 1.0 to 1.2 alike. French is judged by one recogniser, and what French the base saw is unknown. The French test as first registered failed on a reference cut short, and version 1.2 passes on the repaired set (Section 7). Each arm was trained with one seed, and differences between checkpoints of under a point are inside the noise. The inline tag results rest on one render per condition. None of the test texts is conversational.

Broader impact. A voice that speaks Hausa, Zarma and French could carry information to people who do not read French, in the languages they speak at home. The same model can imitate a person from a few seconds of their speech. Every evaluation voice in this report stays inside evaluation, the one other voice belongs to a consenting team member, and we release no model or cloned voice. Any public use of synthesised speech should be labelled as such.

12 Conclusion

One synthesis model now speaks Hausa, Zarma and French, and each step that added a language was checked against every language already there. The method has four parts: start from the model that already speaks a neighbouring language, keep the earlier languages in every batch, measure every language before and after, with two judges where both fit, and decide by rules fixed in advance. The measurements also showed how much depends on things around the model: the voice, the reference clip, the transcripts of evaluation voices, number spelling and the estimator behind a claim about balance. The next steps follow from the limits: native listening panels in all three languages with decision rules written first; new Zarma speakers recorded and consented for synthesis; a French pool balanced by a checked estimator, with more Nigerien speech; and recordings of laughter, sighs and other expressive speech in Zarma and Hausa, to test whether the model can learn what its tags do not yet produce.

Acknowledgements

This work relies on public resources released by others: speech corpora in Hausa, Zarma and French, a parallel text set, evaluation sets, and public models for synthesis, recognition and speaker verification. We thank the people who recorded, transcribed and published them, the speakers whose voices made them possible, and the consenting team member who lent a voice to this work.

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] N. Arivazhagan et al. Massively multilingual neural machine translation in the wild: findings and challenges. *arXiv:1907.05019*, 2019.

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

[4] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge and M. A. Ponti. YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of ICML 2022*, 2022.

[5] E. Casanova et al. XTTS: a massively multilingual zero-shot text-to-speech model. In *Proceedings of Interspeech 2024*, 2024. *arXiv:2406.04904*.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettle-

moyer and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, 2020.

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

[8] 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.

[9] P. Do, M. Coler, J. Dijkstra and E. Klabbers. A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages. In *Proceedings of Interspeech 2021*, pages 16–20, 2021.

[10] R. M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

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

[12] J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[13] M. Le et al. Voicebox: text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems 36*, 2023. *arXiv:2306.15687*.

[14] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[15] M. Mauch and S. Dixon. pYIN: a fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of ICASSP 2014*, pages 659–663, 2014.

[16] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: the sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[17] Namu AI Lab. Zarma speech technology from scarce data: an exploration. Technical report, November 2026. `https://www.namuai-lab.com/papers/zarma-speech-from-scarce-data.pdf`

[18] T. Nekvinda and O. Dušek. One model, many languages: meta-learning for multilingual text-to-speech. In *Proceedings of Interspeech 2020*, 2020. *arXiv:2008.00768*.

[19] P. Newman. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press, New Haven, 2000.

[20] B. A. Nosek, C. R. Ebersole, A. C. DeHaven and D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[21] S. Ogun, A. T. Owodunni, T. Olatunji et al. 1000 African voices: advancing inclusive multi-speaker multi-accent speech synthesis. In *Proceedings of Interspeech 2024*, 2024. *arXiv:2406.11727*.

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

[23] D. Rolnick, A. Ahuja, J. Schwarz, T. P. Lillicrap and G. Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems 32*, 2019.

[24] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier, 1995.

[25] J. Taylor and K. Richmond. Confidence intervals for ASR-based TTS evaluation. In *Proceedings of Interspeech 2021*, pages 2791–2795, 2021.

[26] C. Wang et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv:2301.02111*, 2023.

[27] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg and B. Ramabhadran. Learning to speak fluently in a foreign language: multilingual speech synthesis and cross-language voice cloning. In *Proceedings of Interspeech 2019*, 2019. *arXiv:1907.04448*.

A Detailed results

Table 2: The Zarma stage. Zarma: 700 test renders, folded CER (%) with intervals over sentences; renders of 700 with a Hausa hooked letter in the judge 2 read-back; speaker similarity relative to the ceiling. Hausa: 334 targets from 20 speakers in version 1.0's training data, folded CER with intervals over speakers; natural recordings score 3.6 [2.7, 4.8] (judge 1) and 4.5 (judge 2). All systems re-rendered in one run.

	Zarma				Hausa		
System	Judge 1	Judge 2	Hooked	Similarity	Judge 1	Judge 2	Similarity
Public base, zero-shot	11.8 [11.1, 12.6]	11.7 [11.1, 12.4]	7	0.91	2.9 [2.6, 3.3]	3.4	0.89
Version 1.0, zero-shot	11.5 [10.7, 12.2]	13.7 [13.0, 14.4]	46	0.90	1.8 [1.5, 2.1]	2.3	0.94
From base, Hausa about a quarter	11.0 [10.2, 11.9]	10.8 [10.1, 11.5]	2	0.92	3.0 [2.7, 3.4]	3.3	0.90
From 1.0, Hausa about a quarter	10.2 [9.4, 11.1]	10.1 [9.5, 10.8]	4	0.92	2.1 [1.8, 2.4]	2.5	0.92
Version 1.1: from 1.0, Hausa 40%	**10.3** [9.6, 11.2]	**10.5** [9.8, 11.2]	**3**	0.91	**1.9** [1.6, 2.2]	2.3	0.94

Table 3: The Hausa stage, on 258 broadcast clips from speakers outside version 1.0's training data, read back by recognisers of about 1B parameters. Strict CER keeps hooked letters. Speaker similarity: render against the natural recording of the same sentence; the ceiling is 0.774. Failures are of 258 renders.

Measure	Natural	Public base	Version 1.0
Judge 2 strict CER, %	10.92	6.73	**5.36**
Judge 2 strict CER after a speech codec, %		11.92	**7.40**
Judge 1 strict CER, %	9.96	5.70	5.87
Judge 1 folded CER, %	8.61	4.74	4.88
Judge 1 folded WER, %	23.6	19.6	**16.9**
Speaker similarity		0.696	0.697
Failures		2	1

Table 4: The French stage, judge 2 only, folded CER (%) with intervals over sentences (French, Zarma) or speakers (Nigerien French, Hausa). French is the repaired test suite. Natural French recordings are another speaker reading each sentence (240 sentences). Failures: French / Nigerien French / Zarma / Hausa.

System	French (300)	Nigerien French (61)	Zarma (700)	Hausa (334)	Failures
Natural recordings	2.8 [2.4, 3.2]	11.3		4.5 [3.5, 5.8]	
Public base	2.2 [1.8, 2.7]	5.3 [2.2, 7.3]			0 / 1 / – / –
Version 1.0	2.1 [1.7, 2.5]	6.1 [2.6, 8.3]			0 / 2 / – / –
Version 1.1	2.2 [1.8, 2.6]	5.5 [2.5, 7.8]	10.4 [9.7, 11.1]	2.4 [2.0, 2.9]	0 / 3 / 0 / 0
Version 1.2	**2.4** [2.0, 2.9]	**4.7** [2.3, 6.3]	**10.5** [9.8, 11.3]	**2.3** [1.9, 2.7]	0 / 1 / 0 / 0

Table 5: Checkpoint and arm selection on development data, folded CER (%). Zarma stage: Zarma dev (judge 1 / judge 2) and Hausa dev (judge 1) at the five saved checkpoints, in training order; guardrail on Hausa dev 2.91. French stage, judge 2 at four checkpoints: French dev against the base's 2.10, Zarma and Hausa dev against version 1.1's 10.97 and 2.69. Selected checkpoints in bold.

Arm	Set	1	2	3	4	5
From 1.0, Hausa about a quarter	Zarma	**10.2 / 9.7**	10.6 / 10.5	10.8 / 10.5	10.6 / 10.4	11.0 / 10.4
	Hausa	**2.2**	2.1	2.1	2.1	2.1
From base, Hausa about a quarter	Zarma	10.8 / 10.9	11.1 / 10.7	**10.6 / 10.1**	11.0 / 10.6	10.9 / 10.8
	Hausa	3.3	3.0	**3.7**	3.6	3.6
From 1.0, Hausa 40% (1.1)	Zarma	**10.3 / 10.1**	10.5 / 10.1	10.7 / 10.3	10.5 / 10.2	10.7 / 10.7
	Hausa	**2.1**	2.0	2.0	1.8	2.1
		1	2	3	4	
French about a fifth (1.2)	French / Zarma / Hausa	2.18 / 11.56 / 2.71	**2.12 / 10.82 / 2.73**	2.25 / 11.40 / 2.64	2.10 / 11.74 / 2.27	
French about a third	French / Zarma / Hausa	2.14 / 11.07 / 2.81	**1.82 / 10.52 / 2.96**	2.13 / 11.56 / 2.37	2.15 / 11.13 / 2.52	