Measuring Hausa Speech Recognition: Spelling, Test Sets and What a Score Means

Namu AI Lab
namuai-lab.com

Abstract

Hausa is written with four hooked letters that most keyboards lack, ɓ ɗ ƙ ƴ, and an apostrophe that is typed in several ways, while tone and vowel length are not written at all. A Hausa error rate therefore depends on how text is cleaned before scoring and on whose spelling the reference follows. We report what we learned from measuring Hausa speech recognition with public recognisers and public test sets. We describe a text normaliser with three scores: strict, which keeps the hooked letters; folded, which removes hooks and apostrophes; and space-free. The references disagree with themselves: in the public FLEURS Hausa test split, about half of the word types written with a hook also occur without it, and the hooked spelling accounts for 46% of those words' occurrences. The split holds 621 recordings of 331 sentences; 620 are labelled female, none carries a speaker identifier, and 64 are longer than 30 seconds. On read speech, zero-shot multilingual recognisers of about 1B and 3B parameters reach 4 to 8% character error rate (CER). Folded word error rate (WER) is 2.7 to 4.9 times folded CER on every set, except where a recogniser drifts into other scripts. On one speaker's phone recordings, a recogniser with a CTC head reaches 38% CER and writes 80 of 167 phrases partly in non-Latin scripts, while one of the same size with a language-model decoder, told the language, reaches 9.5%. A public recogniser fine-tuned for Hausa has a higher median per-utterance CER than a zero-shot multilingual one on the three sets where we compared them; we computed no intervals for that comparison. Of the 502 hooked letters in 483 FLEURS references, a zero-shot recogniser writes 131 as in the reference. After fine-tuning on Zarma alone it writes 1, and its Hausa folded CER rises from 4.6 to 12.4%; with Hausa mixed back into training it writes 134, and its CER is 5.3% at the checkpoint selected on Zarma. All results are automatic measurements; no native speaker has yet reviewed the normaliser or the spelling counts.

1 Introduction

Hausa is one of the most widely spoken languages of West Africa, used across northern Nigeria and southern Niger and as a common language of trade far beyond them [7, 12]. Unlike most languages of the Sahel, it is present in public speech benchmarks and in the language lists of massively multilingual recognisers [3, 13]. Hausa recognition results are therefore published often, usually as one word error rate on one public test set.

This report is about what such a number measures. We began measuring Hausa recognition to choose recognisers for our own research, to score synthetic speech, and to check whether a model trained on a neighbouring language, Zarma, still hears Hausa [11]. Each time, the score moved for reasons

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

that had little to do with hearing: a hook on a letter, an apostrophe written three ways, a word split in two, a recording too long for the model, a reference typed without its hooked letters. We set out here what we found, so that others reporting Hausa results can see what their numbers include.

We describe the models we used by their type and scale rather than by name, and the test sets by what they contain, except for FLEURS, the public benchmark most Hausa results are reported on. Every number is an automatic measurement. Our contributions are the following. (1) A text normaliser for Hausa scoring with three forms, strict, folded and space-free, and a way to divide a recogniser's errors among spelling conventions (Section 4). (2) An account of the public FLEURS Hausa test split: who and what is in it, how its references spell, and what a score on it can and cannot show (Section 5). (3) Measurements of public recognisers on several Hausa test sets, including a recogniser fine-tuned for Hausa, and of how they write the hooked letters (Sections 6 and 7). (4) An analysis of what a Hausa word error rate is worth, and of long recordings (Sections 8 and 9). (5) The Hausa that a recogniser forgets when it is fine-tuned on Zarma, and how much of it mixing Hausa back in keeps (Section 10).

2 Related work

Multilingual speech benchmarks. FLEURS provides read speech in about a hundred languages, recorded from the sentences of the FLoRes translation benchmark, and is the most common public test set for Hausa recognition [3, 5]. Common Voice gathers read speech from volunteers and includes Hausa [1]. Massively multilingual recognisers extend coverage to more than a thousand languages by pre-training on untranscribed audio and adapting with little data per language [13]. Benchmarks of this kind usually report one number per language and leave its conditions to the reader.

Hausa speech and writing. Hausa large-vocabulary recognition was studied on read speech more than a decade ago [15], and a recent community collection recorded about 600 hours of read Hausa from 1,879 speakers [4]. Reference grammars describe the hooked consonants, the glottalised glide written *'y* in Nigeria, and the tone and vowel length that standard spelling leaves unwritten [7, 12].

Scoring and its conditions. Word error rate counts substitutions, deletions and insertions from an edit alignment [9]. Its value depends on text normalisation, and large recognisers have been evaluated with normalisers written for that purpose [14]. Benchmark error rates can differ from error rates on real speech for reasons of data and conditions [16], and bootstrap intervals give a sense of how much a score would move on another sample of the same test set [2]. Fine-tuning a network on new data can erase what it knew [8, 10]; our Zarma report measured this for Hausa recognition [11], and we analyse that result in more detail here.

3 How Hausa is written

Hausa is written in a Latin alphabet with three hooked consonants, *ɓ ɗ ƙ*, which differ in sound from *b d k*, and a glottalised glide that is written *y̓* or, as standard spelling in Nigeria does, as an apostrophe followed by *y* [12]. The apostrophe also marks a glottal stop inside words, as in *sa'a*. Tone and vowel length distinguish words but are not written [7, 12].

Real text departs from this in predictable ways. Many keyboards have no hooked letters, so writers drop the hooks, type an apostrophe before the plain letter (*'b 'd 'k*), or substitute look-alike characters. The apostrophe itself appears as at least three different characters: in the references of the FLEURS Hausa test split we count 110 right single quotation marks, 99 ASCII apostrophes and 28 left single quotation marks, some of them used as quotation marks. In the same references an apostrophe-like mark followed by *y* starts a word 50 times (twice it is a quotation mark), and *y̓* is written once.

Tone and vowel length are a blind spot for any score computed on text. From a public pronunciation dictionary and a public online dictionary we mined candidate homograph pairs that differ only in tone (51 pairs), only in vowel length (42) or in both (54). They are candidates for native review and

Table 1: Two references from the FLEURS Hausa test split in the forms we score.

Form	Text
As written	*Waɗannan ma'aurata na iya zaɓar yin shirin tallafi ga jaririn.*
Strict	*waɗannan ma'aurata na iya zaɓar yin shirin tallafi ga jaririn*
Folded	*wadannan maaurata na iya zabar yin shirin tallafi ga jaririn*
Space-free	*wadannanmaauratanaiyazabaryinshirintallafigajaririn*
As written	*Don 'yan Springboks, ya ƙare rashin nasara a wasanni biyar.*
Strict	*don ƴan springboks ya ƙare rashin nasara a wasanni biyar*
Folded	*don yan springboks ya kare rashin nasara a wasanni biyar*

do not form a test set, and we report no recognition result on them. A recogniser that confuses such a pair writes the same letters either way, and no character or word error rate can see it.

4 A normaliser and three scores

Both texts pass through the same normaliser before any edit is counted. It applies Unicode composition, maps look-alike characters to the hooked letters, writes every apostrophe-like character as one apostrophe, turns an apostrophe followed by *y* at the start of a syllable into *ƴ*, and turns an apostrophe before *b*, *d* or *k* into the hooked letter, since a glottal stop is written only before a vowel. It then lowercases, removes punctuation and quotation marks, keeps apostrophes between letters, and drops tokens made only of digits, because recognisers write numbers in words, as digits, or not at all. This gives the *strict* form, in which the hooked letters count. The *folded* form then maps *ɓ ɗ ƙ ƴ* to *b d k y* and deletes apostrophes. The *space-free* form also removes word boundaries. Table 1 shows two FLEURS references in each form.

We compute character error rate (CER) over the whole test set, as total character edits divided by total reference characters, with spaces counted as characters, and word error rate (WER) in the same way over words. Every interval is a 95% percentile bootstrap over utterances with 2,000 resamples [2], or over speakers where the source states it; differences between systems are paired.

Because the forms are cumulative, the drop in character edits from one form to the next is the share of the strict errors that a convention explains. For a zero-shot multilingual recogniser with a CTC head [6], about 1B parameters, on 483 FLEURS recordings of at most 40 seconds whose references contain no digit, the 3,430 strict character edits fall to 2,957 when hooks are folded (13.8% of the strict edits), to 2,951 when apostrophes are removed (0.2%), and to 2,381 when word boundaries are ignored (16.6%). Strict CER is 5.3% [4.9, 5.9], folded 4.6% [4.1, 5.2] and space-free 4.5%. About three in ten of the edits a strict score counts are hooks or word boundaries; the remaining 69.4% are hearing errors and other spelling differences (Figure 5a).

The normaliser has a known failure. One FLEURS sentence, read twice in the test split, opens a quotation with a left single quotation mark directly before a word that begins with *y* (*'yanayin muhalli mai tsanani'*), and the glide rule turns the quotation mark into the hook of *ƴ*. A rule that reads typed text cannot always tell a quotation mark from an apostrophe. No native speaker has reviewed the normaliser yet.

5 Who is in the FLEURS Hausa test set

We read the metadata file of the FLEURS Hausa test split (configuration *ha_ng*, test) in the public release. It lists 621 recordings with a total of 3.34 hours (Table 2 and Figure 1). The recordings last from 6.1 to 78.9 seconds, with a median of 17.9.

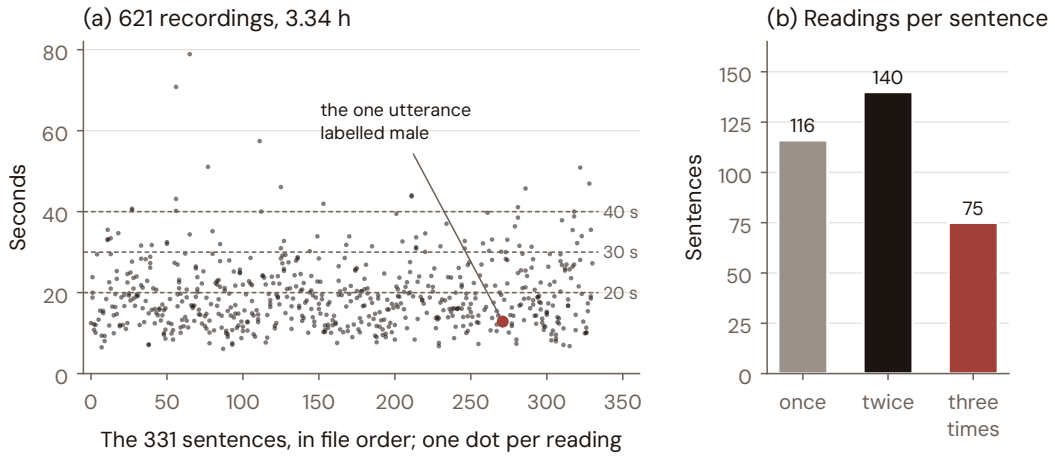


Figure 1: The FLEURS Hausa test split. (a) Every recording, one dot per reading, grouped by the sentence it reads, with its length in seconds; dashed lines mark 20, 30 and 40 seconds. (b) How many times each of the 331 sentences is read. Source: the split's public metadata file.

Table 2: The FLEURS Hausa test split, computed from its public metadata file.

Quantity	Value
Recordings	621
Audio	3.34 h
Labelled female / male	620 / 1
Speaker identifiers	none
Distinct sentences	331
Sentences read once / twice / three times	116 / 140 / 75
Length: minimum / median / maximum	6.1 / 17.9 / 78.9 s
Recordings longer than 20 / 30 / 40 s	235 / 64 / 18
Recordings whose reference contains a digit	131
Recordings of 3 to 20 s	386
Recordings of at most 40 s without a digit	483

Voices. Each recording carries a gender label: 620 are labelled female and one male. The file has no speaker identifier, so it cannot tell how many people were recorded or how many recordings each made. The benchmark paper states that each sentence was recorded by three different speakers, that it aimed at a ratio of at least 30 to 70 between the sexes where possible, and that test speakers differ from training speakers [3]. A per-speaker interval cannot be computed, and an interval over recordings is narrower than the true uncertainty if a few voices dominate. A separate score for men is impossible: a rate computed on one recording would mean nothing.

Sentences. The 621 recordings read 331 distinct sentences: 116 are read once, 140 twice and 75 three times. Every reading of a sentence has the same reference text. Repeated sentences make the set smaller than it looks for anything that depends on text, such as a language model's familiarity with the words.

What it contains and what it does not. The sentences are encyclopaedic text translated from English [5], read aloud. The configuration is labelled as Hausa of Nigeria. Nothing in it is spontaneous, conversational, recorded over a phone, or labelled as Hausa of Niger. A FLEURS score therefore describes how a recogniser hears formal read Hausa from voices labelled female. It is useful for comparison with published work, and that is the limit of what it shows.

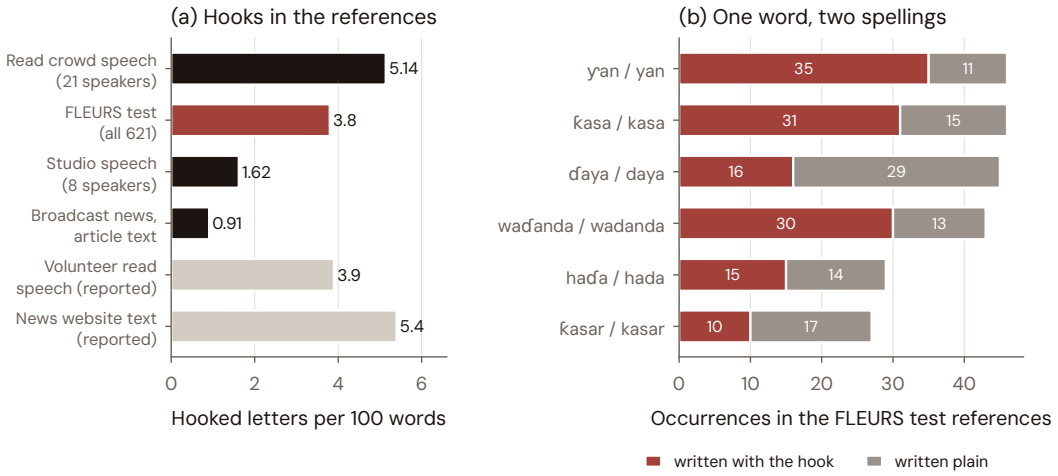


Figure 2: How Hausa test references write the hooked letters. (a) Hooked letters per 100 words in the references of four test sets after the strict normaliser; the pale rows are measurements from our earlier notes on a public volunteer test set and on news website text, not recomputed here. (b) The six most frequent words in the FLEURS test references that occur both with and without the hook.

Subsets. Published results often use a subset. A common filter of 3 to 20 seconds keeps 386 recordings (62%) and 1.52 hours; recognisers limited to 40 seconds can score 603. Scores on different subsets are not comparable: for the CTC recogniser above, folded CER is 4.6% on the 483 recordings without digits and 5.3% on all 603 of at most 40 seconds.

How the references spell. After the strict normaliser, the references hold 3.80 hooked letters per 100 words (Figure 2a). Of 147 word types written with a hook, 74 also occur with no hook at all, and among the occurrences of those 74 types the hooked spelling accounts for 46% (375 of 811). The most frequent cases are shown in Figure 2b: *ƙasa* 31 times and *kasa* 15, *ɗaya* 16 and *daya* 29, *waɗanda* 30 and *wadanda* 13. Some of these pairs may be different words, and only a native reader can tell each case; the counts show that the references do not follow one convention. The other test sets differ from FLEURS and from each other. The references of a crowd-sourced read-speech set hold 5.14 hooked letters per 100 words, a studio set recorded for synthesis 1.62, and a broadcast set whose references are the published article text 0.91, almost all of them *ƴ* from the apostrophe convention: it has 8 *ƙ* in 12,066 words. One studio reference writes *Daya bisa goma sau miliyan ɗaya*, the same word plain and hooked in one sentence. Strict scores on these sets measure, in part, agreement with each set's typing habits.

6 Recognisers on Hausa test sets

We measured public recognisers without any Hausa training of our own on seven kinds of Hausa speech: crowd-sourced read speech from 21 speakers held out from all our training, FLEURS, a studio set, broadcast news, read speech recorded on phones in communities, spontaneous speech, and one consenting team member reading a script on their own phone. Table 3 gives the main results and Figure 3 two paired comparisons. The rows come from different measurements, on different subsets, and are not paired with each other unless stated.

Clean read speech. On read speech the zero-shot recognisers reach 4 to 8% strict CER. The same 1B decoder model scores 0.61% on an English read-speech control scored in the same way, so its Hausa CER on read speech is about seven to thirteen times its English CER. Broadcast speech scores worse (9.96%), but its references are article text aligned to the audio rather than a transcription, so part of that error is paraphrase. We measured the 1B and 3B decoder models only on different subsets of the read crowd set (4.42% on 355 recordings and 4.4% [3.4, 5.6] on 334 recordings from 20 of

Table 3: Zero-shot public recognisers on Hausa test sets, CER and WER in percent with 95% intervals where measured (over utterances, or over speakers for the read crowd and studio sets). All recognisers are from one public multilingual family; the decoder models were told the language. Values keep the precision of their source. The English row is a control scored in the same way.

Test set (items)	Recogniser	CER strict	CER folded	WER folded
Read crowd speech, 21 speakers (355)	LLM decoder, 1B	4.42 [3.0, 6.3]	4.02	17.99
Read crowd speech, 20 speakers (334)	LLM decoder, 3B	4.4 [3.4, 5.6]	3.6 [2.7, 4.8]	17.5
Read crowd speech, 20 speakers (334)	CTC, 3B		4.5	
FLEURS, 3 to 20 s (386)	LLM decoder, 1B	6.40	5.55	16.7
FLEURS, at most 40 s, no digits (483)	CTC, 1B	5.33 [4.85, 5.91]	4.60 [4.13, 5.18]	18.51 [17.36, 19.61]
Studio speech, 8 speakers (100)	LLM decoder, 1B	7.82 [6.29, 9.20]	6.62	19.6
Broadcast, article text as reference (300)	LLM decoder, 1B	9.96	8.61	23.6
Phone, one speaker (167)	LLM decoder, 1B	9.54 [7.76, 11.40]	8.38 [6.74, 10.17]	30.67 [25.42, 36.57]
Phone, one speaker (167)	CTC, 1B	38.43 [32.60, 44.78]	37.72 [31.97, 44.13]	69.12 [61.77, 76.76]
English read speech, 40 speakers (500)	LLM decoder, 1B	0.61 [0.47, 0.76]		2.11

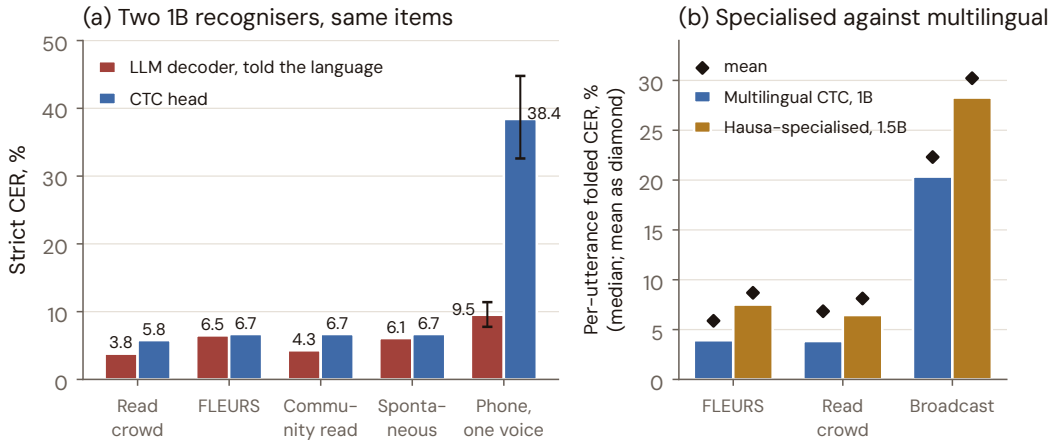


Figure 3: Paired recogniser comparisons on Hausa. (a) Two zero-shot recognisers of about 1B parameters from one public family, one with an LLM decoder told the language and one with a CTC head, on the same items: 150-item samples of three sets, 83 pieces of a spontaneous speech set (58 spontaneous, 25 read prompts) and 167 phone phrases of one consenting speaker, with 95% intervals for the phone set only. The community and spontaneous sets were reconstructed from a neural audio codec at its highest setting before recognition. (b) A public recogniser fine-tuned for Hausa against the zero-shot CTC recogniser on the same utterances (386 FLEURS, 354 read crowd and 625 broadcast), per-utterance folded CER with digits dropped: bars are medians, diamonds means; no intervals were computed.

the same speakers), so these data cannot separate a size effect from the subset.

Decoder against CTC head. On the same items, the 1B model with an LLM decoder has lower strict CER than the 1B model with a CTC head on every set (Figure 3a): 3.8 against 5.8% on read crowd speech and 6.5 against 6.7% on FLEURS. We computed intervals only for the phone set, so the smaller of these gaps may not hold on another sample. The gap is largest on the phone recordings: 9.5% [7.8, 11.4] for the decoder and 38.4% [32.6, 44.8] for the CTC head, a paired difference of 28.9 points [23.2, 35.2]. The CTC model is not told the language, and on this audio it often loses track of it: 80 of its 167 outputs contain letters from a non-Latin script, Arabic letters in 64 of them, and also Chinese, Japanese, Ethiopic, Georgian, Korean, Devanagari or Cyrillic letters. The decoder, told the language, writes none. The phone references were matched to the script lines with the help of both recognisers' transcripts, which makes both CERs on this set somewhat optimistic; it is unlikely to explain a gap of this size.

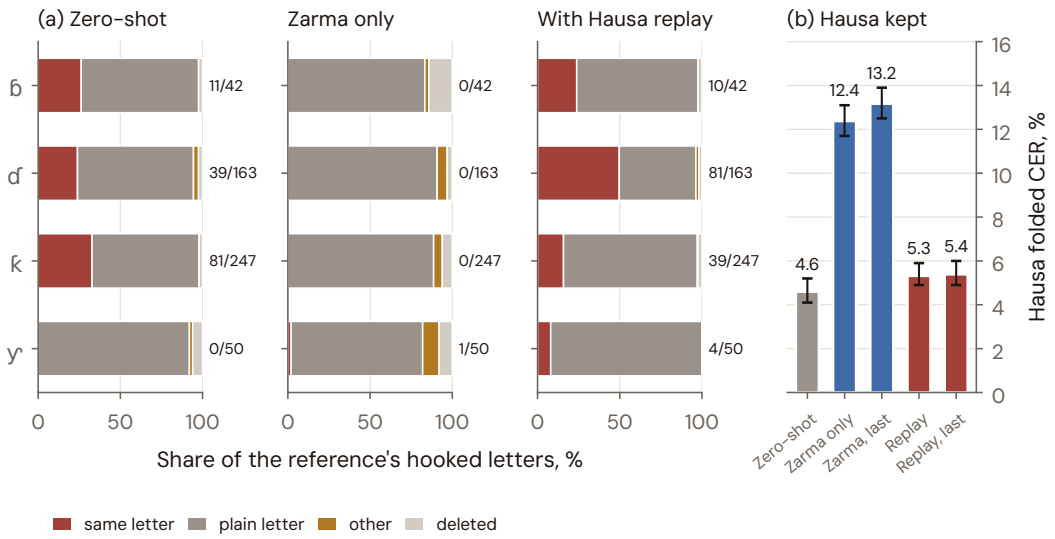


Figure 4: Hooked letters and forgetting on 483 FLEURS Hausa recordings. (a) What becomes of each hooked letter of the reference in the output of a zero-shot CTC recogniser of about 1B parameters, of the same model fine-tuned on Zarma alone, and of the same model fine-tuned with Hausa mixed in; the numbers give letters written the same out of those in the references. (b) Hausa folded CER of the same models, at the checkpoint selected on Zarma development data and at the last checkpoint, with 95% intervals.

A recogniser fine-tuned for Hausa. We compared a public recogniser fine-tuned for Hausa, an encoder-decoder model of about 1.5B parameters trained on at most about 100 hours of public read Hausa, with the zero-shot 1B CTC recogniser on the same utterances of three sets (Figure 3b): the 386 FLEURS recordings of 3 to 20 seconds, 354 read crowd recordings, and 625 broadcast segments, a larger selection from the same public collection than the broadcast set of Table 3. Its median per-utterance folded CER is higher on all three: 7.5 against 3.9% on FLEURS, 6.5 against 3.8% on read crowd speech, and 28.3 against 20.3% on broadcast speech, where both are scored against article text. The means order the same way. FLEURS training speech is among its training data, so it is not at a disadvantage on that set. In the same runs it wrote about half as many hooked letters per 100 words as the multilingual recogniser (*ƙ* 1.65 against 3.82), and on music or noise it sometimes wrote a sentence in another language. These comparisons have no intervals and use per-utterance medians; they show the direction, and the size only roughly.

7 How recognisers write the hooked letters

We aligned each strict reference with the strict output of the recogniser, character by character, and recorded what became of every hooked letter of the reference (Figure 4a; Appendix Table 7).

Most hooks become plain letters. The zero-shot CTC recogniser writes the same hooked letter for 131 of the 502 in the FLEURS references: 11 of 42 *ɓ*, 39 of 163 *ɗ*, 81 of 247 *ƙ* and none of 50 *y*. Most of the others become the plain letter (352 of 502), as do 46 of the 50 *y*. The recogniser does write hooks: its output holds 251 hooked letters, and 120 of them are not aligned with the same letter in the reference. Some of those may be cases where the reference dropped a hook the speaker said, given how inconsistently the references spell (Section 5); a strict score counts them as errors all the same. Without a native review of each case we cannot say how many.

Across recognisers and conditions. The 1B decoder model turns 9.8% of the hooked letters in the read crowd references into plain letters. On the phone recordings, the 35 phrases that read sentences chosen for their hooked letters or minimal pairs contain 40 hooked letters in their references; the decoder's output contains 22 and the CTC model's 18, before any alignment. The recogniser fine-

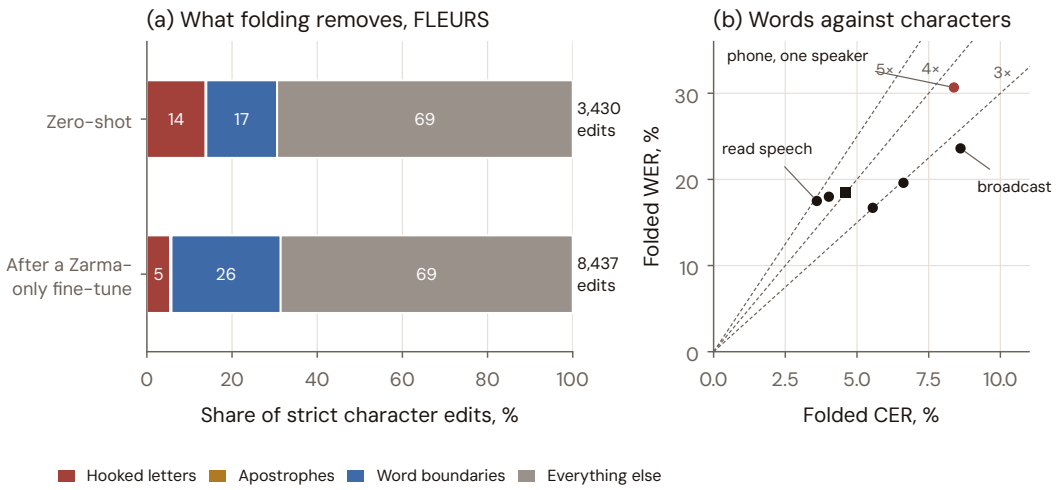


Figure 5: What a score contains. (a) Strict character edits of a CTC recogniser on 483 FLEURS Hausa recordings, divided by the convention that explains them: hooked letters, apostrophes, word boundaries, and everything else; zero-shot and after fine-tuning on Zarma alone. (b) Folded WER against folded CER for the same outputs, one point per test set and recogniser in Table 3 (squares: CTC); dashed lines mark WER three, four and five times CER. The CTC model on the phone recordings (37.7, 69.1) lies outside the panel.

tuned for Hausa writes about half as many hooks as the multilingual one (Section 6). A recogniser that rarely writes a hook loses little under folded scoring and a lot under strict scoring; one that writes hooks where references omit them loses under strict scoring too. The two scores side by side, with the letter counts, show which kind a recogniser is.

8 What a Hausa word error rate is worth

Words multiply character errors. Across the test sets of Table 3, folded WER is between 2.7 and 4.9 times folded CER (Figure 5b). The one exception is the CTC model on the phone recordings, at 1.8 times, because much of its output is not Hausa at all. The zero-shot CTC recogniser on FLEURS has 4.6% folded CER and 18.5% folded WER; with hooks kept, its WER is 21.8%. One wrong character makes a whole word wrong, and Hausa words are long: in the FLEURS references, words average about 4.5 letters. A Hausa WER near 20% can therefore come with a CER under 5%, and two systems with the same CER can differ in WER by how they place word boundaries.

Word boundaries. Of the zero-shot CTC recogniser's folded word errors on FLEURS, 21.3% sit in regions where the reference words and the output words are identical once joined, so the only error is where the space falls. Human transcripts and recogniser outputs of spontaneous Hausa split and join the same words differently (we met *akaji* against *aka ji* in Section 9). Space-free CER removes this convention entirely.

Normalisation can outweigh the model. On a constructed pair of 13 words that differ only in one missing hook and one wrong word, strict WER is 15.4%, folded WER 7.7%, and WER without normalisation 46.2%, because four correct words differ from the reference only in a capital letter or a punctuation mark. The same output can thus be reported at three error rates that differ by a factor of six.

Comparability. For context, an earlier entry by our team to a public Swahili recognition challenge was scored by the challenge's organiser at 17.81% WER on its private leaderboard. That number and the Hausa WERs here look alike and cannot be compared: language, test set, speech type and normaliser all differ. A Hausa error rate is interpretable only with its test set, its subset, its normaliser

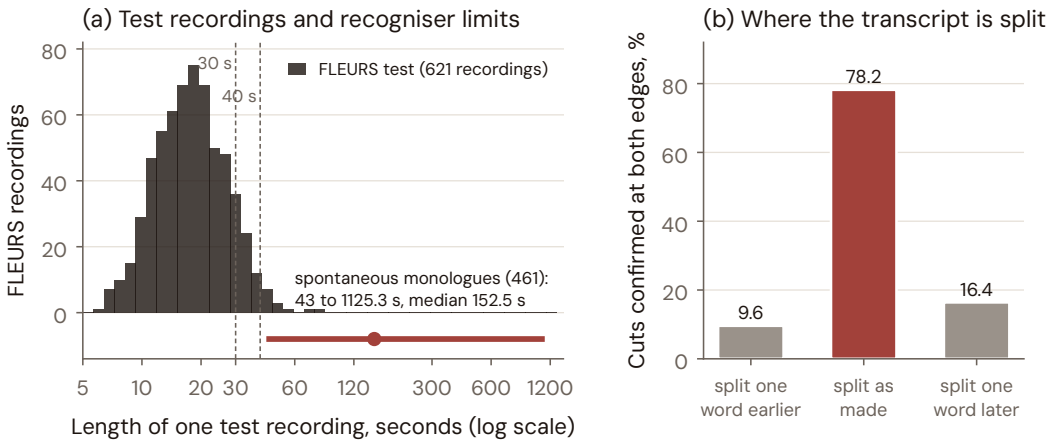


Figure 6: Long audio. (a) Lengths of the 621 FLEURS Hausa test recordings, and the range of the 461 spontaneous monologues in the held-out split of a public spontaneous Hausa corpus; dashed lines mark 30 and 40 seconds. (b) Cutting monologues into pieces: share of the 66,363 cuts between two non-trivial segments at which a recogniser hears the last word of one transcript span at the end of its segment and the first word of the next span at the start of the next segment, as made and with the split moved by one word.

and version, both strict and folded forms, and an interval.

9 Long audio

Recognisers of the family we used were trained on clips of up to 30 seconds and their reference pipeline refuses clips over 40 seconds. Test material is often longer (Figure 6a).

How much is long. The benchmark paper states that all its segments are within 30 seconds [3]. By the sample counts in the metadata file, 64 recordings of the Hausa test split are longer than 30 seconds, 0.67 of its 3.34 hours, and 18 are longer than 40 seconds. Spontaneous speech is longer still. The held-out split of a public corpus of spontaneous Hausa monologues contains 461 monologues from 22 speakers, of 43 to 1,125 seconds, with a median of 152.5 seconds; none is 30 seconds or shorter. Spontaneous Hausa cannot be scored at all by a clip-level recogniser without first cutting the audio, and every published result on such material depends on how it was cut.

Cutting at pauses. To use these monologues we cut every clip longer than 29.5 seconds in the middle of pauses, aiming at pieces of about 20 seconds and never longer than 29.5, and cut at the quietest point only where a long stretch has no pause. Across the whole corpus, 43,992 clips and 484.1 hours, 10,746 clips were cut into 78,032 segments with a median of 19.8 seconds; 1.8% of the cuts were forced. Each clip has one human transcript, which we split among its segments with one alignment over the whole clip. The split sits on the word boundary: at 78.2% of cuts a recogniser hears the last word of each span at the end of its segment and the first word of the next span at the start of the next, against 9.6% with the split moved one word earlier and 16.4% one word later (Figure 6b). Only 0.07% of words end after the next segment begins.

What the pieces show. We then compared each segment's share of the human transcript with one recogniser's output. In the corpus's training split, read prompts and spontaneous speech behave differently: the median folded CER between transcript and output is 1.8% for read segments and 6.5% for spontaneous ones, and under a fixed agreement threshold 98% of read hours and 85% of spontaneous hours agree. The spontaneous segments that disagree are mostly passages the transcriber tidied or paraphrased, and word-splitting differences such as *akaji* against *aka ji*. For spontaneous Hausa the transcriber's choices shape the reference, so a test set built from it needs a transcription convention as well as a normaliser. Cutting is also less stable than it looks: in our Zarma work,

re-cutting a test set on different hardware moved 14 of 1,568 segment boundaries [11], so a cut test set should be stored as a manifest and rebuilt from it.

10 The Hausa a Zarma fine-tune forgets

Zarma and Hausa share a country and many words, and a recogniser that serves both must keep both. In our Zarma work we fine-tuned the 1B CTC recogniser on about seven hours of Zarma from five speakers, alone and with Hausa read speech mixed in at about a fifth of the utterances, and measured Hausa on the same 483 FLEURS recordings [11]. Figure 4 and Appendix Table 8 give the results.

Errors. Fine-tuning on Zarma alone raises Hausa folded CER from 4.6 to 12.4% [11.7, 13.1] at the checkpoint selected on Zarma, a paired increase of 7.8 points [7.3, 8.3], and to 13.2% at the last checkpoint. Folded WER rises from 18.5 to 53.1%. With Hausa mixed in, Hausa CER is 5.3% at the selected checkpoint, a paired increase of 0.7 points [0.5, 1.0] over the zero-shot model, and 5.4% at the last one.

Letters. The letters show the forgetting in more detail than the scores. Of the 502 hooked letters in the references (42 ɓ, 163 ɗ, 247 ƙ and 50 ƴ), the zero-shot model writes 131 as in the reference. After training on Zarma alone it writes 1, an ƴ, and none of the 452 ɓ ɗ ƙ; its whole output holds four hooked letters. It also writes the Zarma letter ŋ where the Hausa reference has *n*, 310 times. Folding hides the first effect but not the second. With Hausa mixed in, the model writes 134 hooked letters as in the reference (10 ɓ, 81 ɗ, 39 ƙ and 4 ƴ), against 131 for the zero-shot model, but not the same ones: it writes 81 of 163 ɗ and 39 of 247 ƙ, against 39 and 81 before. The mixed-in Hausa changed which hooks the model writes, and strict scores alone would not show it.

11 Discussion

Spelling is part of the test. Every Hausa test set we examined has its own typing habits, from references with 5.14 hooked letters per 100 words to article text with 0.91, almost all of them ƴ. A strict score on such a set measures agreement with those habits as well as hearing, and a folded score does not show whether the hooked letters are right. We therefore report both, add the space-free form where word boundaries are uncertain, as in spontaneous speech, and divide the errors among conventions. The division is cheap to compute and changed how we read at least one result: a recogniser fine-tuned on Zarma with Hausa mixed in moved by 0.7 points of folded CER, and its letters showed that it writes a different set of hooks.

What FLEURS can carry. FLEURS is the shared yardstick for Hausa, and we will keep reporting it. Its Hausa test split is read formal text, labelled female in all but one recording, without speaker identifiers, with two thirds of its sentences read more than once and a fifth of its audio beyond 30 seconds. A result on it should name the subset and normaliser, and should not be read as a measure of conversational, phone or Niger Hausa.

Which recogniser. On read speech, the two 1B multilingual recognisers differ by 0.2 to 2.4 points of strict CER on the same items. On one speaker's phone recordings the difference between architectures is large: a CTC head that is not told the language drifts into other scripts, and a decoder told the language does not. A public recogniser fine-tuned on about a hundred hours of read Hausa had a higher median per-utterance CER than a zero-shot multilingual one on each of the three sets where we compared them, a comparison without intervals. For a model that must serve Hausa and a neighbouring language, mixing Hausa back into training kept Hausa folded CER within a point of the zero-shot model (5.3 against 4.6% at the selected checkpoint).

12 Limitations and broader impact

Limitations. No native speaker has reviewed the normaliser, the spelling counts, the word pairs, or any recogniser output in this report; some pairs we count as one word written two ways may be two words. Every number is an automatic measurement. Our frozen Hausa test sets hold 21 identified held-out speakers; FLEURS and the broadcast set have no speaker identifiers, so their intervals are over recordings and understate the uncertainty if a few voices dominate. The phone results come from one consenting speaker, and their references were matched to a script with the help of the recognisers under test. Rows of Table 3 use different subsets and were not paired; the size comparison is confounded with the subset; the comparison with the recogniser fine-tuned for Hausa uses per-utterance medians without intervals. Two sets were reconstructed from a neural codec before recognition. The public recognisers may have seen some of this public material in their own training. We have no Hausa test set from Niger, no conversational or telephone set with many speakers, and no test of tone or vowel length.

Broader impact. Scores decide which systems are offered to people. A Hausa score that ignores hooks counts output without them as correct, though Hausa readers would have to restore them by hand, and a score on read Nigerian Hausa from voices labelled female says little about a man in Niger speaking into a phone. We hope that reporting the conditions of each number, and both scores, makes it harder to overstate what a Hausa recogniser can do. The phone recordings in this report are of a consenting team member and are not released.

13 Conclusion

Most of the difficulty we met in measuring Hausa recognition came from deciding what counts as the same word. On FLEURS, hooks, word boundaries, digits and the choice of subset each moved a score about as much as the choice between two recognisers of the same size did on read speech, recording length decided what could be scored at all, and the main public test set, while useful, covers one kind of Hausa from voices labelled almost entirely female. The measures that held up were simple: one normaliser with strict, folded and space-free forms, errors divided by convention, intervals, letters counted one by one, and test sets stored as manifests. The next steps follow from the gaps: native review of the normaliser and of the hooked and plain word pairs; a speaker-disjoint Hausa test set that states its dialects, genders and recording conditions and includes Hausa from Niger and phone speech; listening tests for tone and vowel length; transcription conventions and evaluation for long spontaneous speech; and joint training for recognisers that serve Hausa with its neighbours.

Acknowledgements

This work relies on public resources released by others: the FLEURS benchmark and its Hausa recordings, public Hausa read, studio, broadcast and spontaneous speech corpora, public dictionaries, and open recognisers. We thank the people who recorded, transcribed 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] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of ICASSP 2004*, volume 1, pages 409–412, 2004.

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

[4] C. Emezue et al. The NaijaVoices dataset: cultivating large-scale, high-quality, culturally-rich speech data for African languages. In *Proceedings of Interspeech 2025*, pages 1338–1342, 2025.

[5] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán and A. Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[6] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*, pages 369–376, 2006.

[7] P. J. Jaggar. *Hausa*. London Oriental and African Language Library 7. John Benjamins, Amsterdam, 2001.

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

[9] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

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

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

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

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

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML 2023*, pages 28492–28518, 2023.

[15] T. Schlippe, E. G. Komgang Djomgang, N. T. Vu, S. Ochs and T. Schultz. Hausa large vocabulary continuous speech recognition. In *Proceedings of the Third Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, 2012.

[16] P. Szymański, P. Żelasko, M. Morzy, A. Szymczak, M. Żyła-Hoppe, J. Banaszczak, L. Augustyniak, J. Mizgajski and Y. Carmiel. WER we are and WER we think we are. In *Findings of EMNLP 2020*, pages 3290–3295, 2020.

A Test sets

Table 4: The Hausa test material used in this report. Frozen sets were stored as manifests with checksums before any score was computed.

Set	Content	References
Read crowd speech	21 speakers held out from all our training; 355 recordings of 2 to 12 s, 0.56 h	Read prompts
FLEURS test, 3 to 20 s	386 recordings, 1.52 h, 245 distinct sentences	Translated sentences
FLEURS test, at most 40 s	603 recordings, 3.10 h; 483 without digits, 2.39 h	Translated sentences
Studio speech	8 speakers, 10 to 13 recordings each; 100 recordings of 3 to 30 s, 0.34 h	Read prompts
Broadcast news	300 segments of 5 to 30 s, 1.20 h, from two configurations of a public broadcast collection, kept where the collection's own alignment score was good	Published article text
Community read speech	150 recordings sampled from 40 held-out speakers recorded on phones, reconstructed from a neural codec	Read prompts
Spontaneous speech	83 pieces from 25 held-out speakers of a spontaneous speech corpus (58 spontaneous pieces from 22 speakers, 25 read prompts from 3), reconstructed from a neural codec	Human transcripts and read prompts
Phone, one speaker	183 phrases read by one consenting team member on their own phone; 167 with a reference	Script lines
English control	40 speakers, 500 recordings of 3 to 20 s, 1.12 h	Read prompts

B Detailed results

Table 5: The zero-shot 1B CTC recogniser on FLEURS Hausa, in percent with 95% intervals over utterances, on the two subsets used in this report.

Subset	CER strict	CER folded	CER space-free	WER strict	WER folded
At most 40 s, no digits (483)	5.33 [4.85, 5.91]	4.60 [4.13, 5.18]	4.49 [4.01, 5.08]	21.84 [20.68, 22.92]	18.51 [17.36, 19.61]
At most 40 s (603)	5.96 [5.51, 6.47]	5.25 [4.81, 5.75]	5.16 [4.68, 5.68]	22.49 [21.49, 23.46]	19.33 [18.31, 20.31]

Table 6: The phone recordings of one consenting speaker (167 phrases, 489 reference words), re-scored from stored transcripts, in percent with 95% intervals over phrases; the last row is the paired difference, CTC minus decoder.

Recogniser	CER strict	CER folded	CER space-free	WER strict	WER folded
LLM decoder, 1B	9.54 [7.76, 11.40]	8.38 [6.74, 10.17]	6.81 [5.32, 8.44]	32.72 [27.22, 38.80]	30.67 [25.42, 36.57]
CTC, 1B	38.43 [32.60, 44.78]	37.72 [31.97, 44.13]	37.49 [31.20, 44.22]	70.55 [63.24, 77.98]	69.12 [61.77, 76.76]
Difference	+28.89 [+23.15, +35.20]	+29.35 [+23.87, +35.58]	+30.68 [+24.70, +37.19]	+37.83 [+29.48, +45.95]	+38.45 [+29.93, +46.28]

Table 7: What becomes of the hooked letters of 483 FLEURS Hausa references: written the same, written as the plain letter, written as another character, or deleted, for the zero-shot 1B CTC recogniser and its two Zarma fine-tunes (selected checkpoints).

Model	Letter	In references	Same	Plain	Other	Deleted
Zero-shot	ɓ	42	11	30	0	1
	ɗ	163	39	115	5	4
	ƙ	247	81	161	0	5
	y̓	50	0	46	1	3
Zarma only	ɓ	42	0	35	1	6
	ɗ	163	0	148	10	5
	ƙ	247	0	219	13	15
	y̓	50	1	40	5	4
With Hausa replay	ɓ	42	10	31	0	1
	ɗ	163	81	76	3	3
	ƙ	247	39	201	1	6
	y̓	50	4	46	0	0

Table 8: Hausa after fine-tuning the 1B CTC recogniser on Zarma, on 483 FLEURS Hausa recordings, in percent with 95% intervals over utterances; differences are paired against the zero-shot model.

Model	CER strict	CER folded	WER folded	Folded CER difference
Zero-shot	5.33 [4.9, 5.9]	4.60 [4.1, 5.2]	18.51 [17.4, 19.6]	
Zarma only, selected checkpoint	13.11 [12.4, 13.9]	12.38 [11.7, 13.1]	53.08 [51.6, 54.5]	+7.78 [+7.33, +8.26]
Zarma only, last checkpoint	13.90 [13.2, 14.6]	13.17 [12.5, 13.9]	56.00 [54.4, 57.5]	+8.57 [+8.08, +9.07]
With Hausa replay, selected checkpoint	6.07 [5.6, 6.6]	5.32 [4.9, 5.9]	21.62 [20.5, 22.8]	+0.72 [+0.48, +0.95]
With Hausa replay, last checkpoint	6.13 [5.6, 6.7]	5.39 [4.9, 6.0]	22.46 [21.3, 23.6]	+0.79 [+0.57, +1.03]

Table 9: How the references of each test set write the hooked letters, after the strict normaliser. Variable types are hooked word types whose fully plain spelling also occurs in the same set; the hooked share is that of their occurrences. The FLEURS *y̓* counts include two quotation marks read as *y̓* (Section 4).

Set	Words	ɓ	ɗ	ƙ	y̓	Per 100 words	Variable types, hooked share
Read crowd speech (355)	3,385	17	61	93	3	5.14	19 of 101, 46.9%
FLEURS test, all (621)	15,759	47	205	294	53	3.80	74 of 147, 46.2%
FLEURS test, 3 to 20 s (386)	8,200	23	80	167	28	3.63	52 of 109, 43.0%
Studio speech (100)	1,482	3	8	8	5	1.62	7 of 18, 28.9%
Broadcast, article text (300)	12,066	0	5	8	97	0.91	12 of 26, 43.8%