Tok234: speech recognition for Nigerian Pidgin, tested where it is spoken

[First Author][1], [Second Author][1], [Third Author][2]

[1] Tala Labs, [City], Nigeria · [2] [Partner institution] · Correspondence: [name]@talaresearch.com

ABSTRACT

[150–250 words. One sentence each: the problem; what we built; how we tested it and where; the main result with its number and test set; what it does not do yet; where the models and data statement are published.] Example opening: Nigerian Pidgin is spoken daily by [N, cite] people, yet speech recognition systems rarely handle it. We present Tok234 Mini and Tok234 Large, trained on [N] hours of consented Pidgin speech and evaluated in [settings].

1 · INTRODUCTION

State the gap in one paragraph. Say who is affected and how you know. Avoid “revolutionary”, “state-of-the-art” or “first” unless you can cite it.

House style: claims come with evidence; results come with numbers. Pidgin examples are set in Google Sans with an English gloss, e.g. **“How far? Wetin dey happen?”** (Hi — what’s going on?).

2 · DATA

2.1 Collection and consent

Sources, regions, speakers, hours, recording devices, consent process, payment, and what speakers were told. Link the data statement.

2.2 Splits

SPLIT	HOURS	SPEAKERS	REGIONS
Train	[—]	[—]	[—]
Validation	[—]	[—]	[—]
Test	[—]	[—]	[—]

Table 1. Dataset splits. Test speakers do not appear in training.

3 · MODELS

Two sizes trained on the same data: **Tok234 Mini** for phones and laptops, **Tok234 Large** for servers. Always name models in full, size after the name: Tok234 Mini, Tok234 Large. Repositories: `tala-labs/tok234-mini`, `tala-labs/tok234-large`.

Architecture, tokenizer, training schedule, compute used and its carbon estimate.

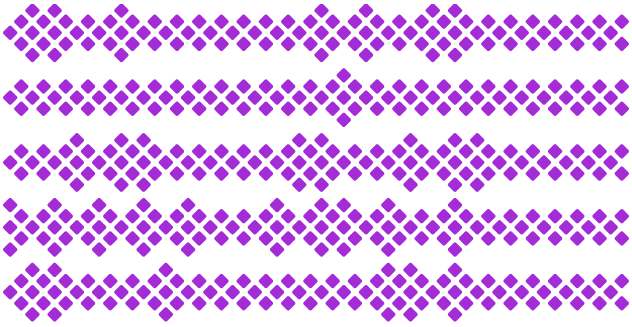

Figure 1. Amplitude envelopes of the five demo phrases, drawn as token waves (one column per bucket of speech). Use this style for any per-utterance signal. [Replace with envelopes from native-speaker recordings.]

4 · FIELD TESTING

Where and how you tested outside the lab: calls, markets, noisy rooms. Who ran the tests. What went wrong.

5 · RESULTS

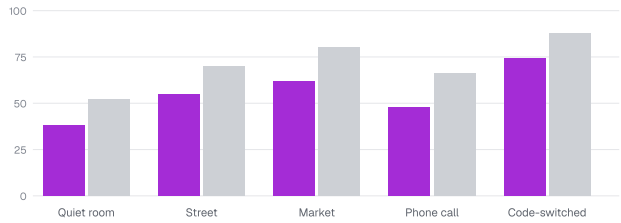

Figure 2. Word error rate (%) by condition. ■ Tok234 Large · ■ baseline. Illustrative values — replace with evaluation output. Tok234 series in Signal Violet, comparisons in grey; never use colour alone to carry meaning.

MODEL	WER ↓	CER ↓	RTF
Baseline [name]	[—]	[—]	[—]
Tok234 Mini	[—]	[—]	[—]
Tok234 Large	[—]	[—]	[—]

Table 2. Results on the Tok234 Pidgin test set. Report confidence intervals.

6 · LIMITATIONS

Accuracy drops with heavy background noise and overlapping speakers. [quantify] Code-switching into Yoruba, Igbo or Hausa is only partly covered. [quantify] Not tested for medical, legal or emergency decisions. Keep a person in the loop.

7 · WHAT COMES NEXT

What you will test next, and what would change your conclusions.

ACKNOWLEDGEMENTS

Thank the speakers first. Then funders and partners.

REFERENCES

[1] [Author, A. and Author, B.] [Year]. [Title]. In *[Venue]*, pp. [—].

[2] [Author, C.] [Year]. [Title]. *[Journal]* [vol]([issue]).

Tala.Labs