



International Journal of Neural Networks and Artificial Intelligence

Journal homepage: <https://trpub.org/journal.php?id=bR2aXTjoAZ>

TRPub

IJNNAI

International Journal of
Neural Networks and
Artificial Intelligence

Vol. 1 Issue 1 2026

Methodology Article

Open Access

Adopting a Speech Recognition Model for a Low Resource Language in East Africa: A Proposed Methodological Framework

Aditya Do^{1*}, Siddharth Cho², Chloe Karlsson³

¹Department of Electrical Electronics, Tamiraparani Research Institute of Electrical Electronics, United Kingdom

²Department of Electrical Electronics, Orinoco Graduate Institute of Electrical Electronics, Switzerland

³Department of Electrical Electronics, Ebro Graduate Institute of Electrical Electronics, Uganda

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijnnai.2026.v1.i1.pp25-30](https://doi.org/10.62715/ijnnai.2026.v1.i1.pp25-30)

Received: 15 November 2025

Accepted: 18 January 2026

Published: 19 March 2026

Abstract

Speech recognition has improved most for the languages that already had the most data. Many languages of East Africa, though spoken by millions, have only a few hours of transcribed speech and little digital text. Multilingual models pretrained on large quantities of audio have changed what is feasible, because they can be adapted to a new language with a small labelled set. Published guidance on how a local team should go about that adaptation is scattered across papers written for other purposes. This article proposes a five-stage framework for adopting a pretrained speech recognition model for a low resource East African language: profile the language and agree terms with its speakers; assemble and document the data; select a base model and an adaptation method suited to the hardware at hand; fine-tune with augmentation and decode with a language model; and evaluate on speakers and domains kept apart from training before releasing the result. For each stage the article states the decisions to be taken, the reasons for them and the principal risk. The framework draws on published work on self-supervised speech models, low resource recognition and dataset documentation. It is proposed and has not been applied or validated here, and the closing sections state what remains to be demonstrated.

Keywords: Automatic speech recognition, Low resource languages, East African languages, Self-supervised learning, Fine-tuning, Dataset documentation

© 2026 Aditya Do et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A speech recogniser built in the conventional way needs hundreds or thousands of hours of transcribed audio, a pronunciation lexicon and a large body of text. Few of the world's languages have these. Besacier *et al.* (2014), surveying recognition for under-resourced languages, listed the usual obstacles: little transcribed speech, limited electronic text, no agreed lexicon, sometimes no stable orthography, and few trained specialists among speakers. That description still fits much of East Africa, where languages with large speaker populations are absent from commercial voice products.

* Corresponding author: Aditya Do, Department of Electrical Electronics, Tamiraparani Research Institute of Electrical Electronics, United Kingdom. E-mail: aditya.do@tamiraparani.edu

© 2026 Aditya Do et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The technical situation has shifted. Models pretrained on unlabelled audio learn representations of speech that transfer across languages, and fine-tuning them on a small transcribed set can produce a usable recogniser (Baeovski *et al.*, 2020; Conneau *et al.*, 2021). A team at a university or a community organisation can now attempt what once required a national programme. What such a team lacks is an ordered account of the decisions involved, from the first conversation with speakers to the evaluation of the finished system.

This article proposes such an account as a framework in five stages. It is written for a single target language and a small team with modest computing resources. The framework is a proposal built from published methods. It is not a report of a project: no model was trained and no data were collected for this article, and nothing here should be read as a result.

2. Background

Two things must be understood before the stages are set out: what the available pretrained models are, and what is particular about the languages and data of the region.

2.1. Pretrained speech models

The models fall into two groups. The first is self-supervised. wav2vec 2.0 masks spans of a latent representation of the raw waveform and learns to identify the correct content among distractors; a linear output layer is then added and the network is fine-tuned with connectionist temporal classification (Baeovski *et al.*, 2020; Graves *et al.*, 2006). Pretraining the same architecture on many languages at once improves results for languages with little labelled data (Conneau *et al.*, 2021), and XLS-R extended the approach to 128 languages (Babu *et al.*, 2022). The Massively Multilingual Speech project went further, using recordings of religious texts to cover more than a thousand languages and adding small language-specific adapter modules (Pratap *et al.*, 2024).

The second group is trained with supervision on a very large scale. Whisper is an encoder-decoder Transformer trained on several hundred thousand hours of audio paired with transcripts gathered from the web, in many languages (Radford *et al.*, 2023). Its authors report that accuracy in a given language tracks the amount of training audio in that language, so its starting point for an East African language may be weak even though the model is strong overall. Both kinds of model can be fine-tuned, and the choice between them is taken up in Stage 3.

2.2. Languages and data in East Africa

The region's languages belong to several families with different demands. Bantu languages such as Luganda and Kinyarwanda are agglutinative: a single written word may carry subject, tense, object and verb root, so vocabularies are very large and unseen word forms are common. Many Bantu and Nilotic languages use tone to distinguish words or grammatical forms, and the standard orthography often leaves tone unmarked. Amharic and Tigrinya are written in the Ethiopic script, in which each character stands for a syllable. Speakers routinely mix in Swahili, English or French, depending on the country.

Open data exist for some of these languages. Mozilla's Common Voice collects read sentences from volunteers under a public-domain dedication and validates each clip by listener vote (Ardila *et al.*, 2020), and it includes collections for several East African languages. Earlier work built Swahili resources from broadcast and web sources (Gelas *et al.*, 2012). Read speech recorded on telephones is nevertheless a narrow sample of how a language is spoken, a point that shapes the evaluation stage below.

3. The Proposed Framework

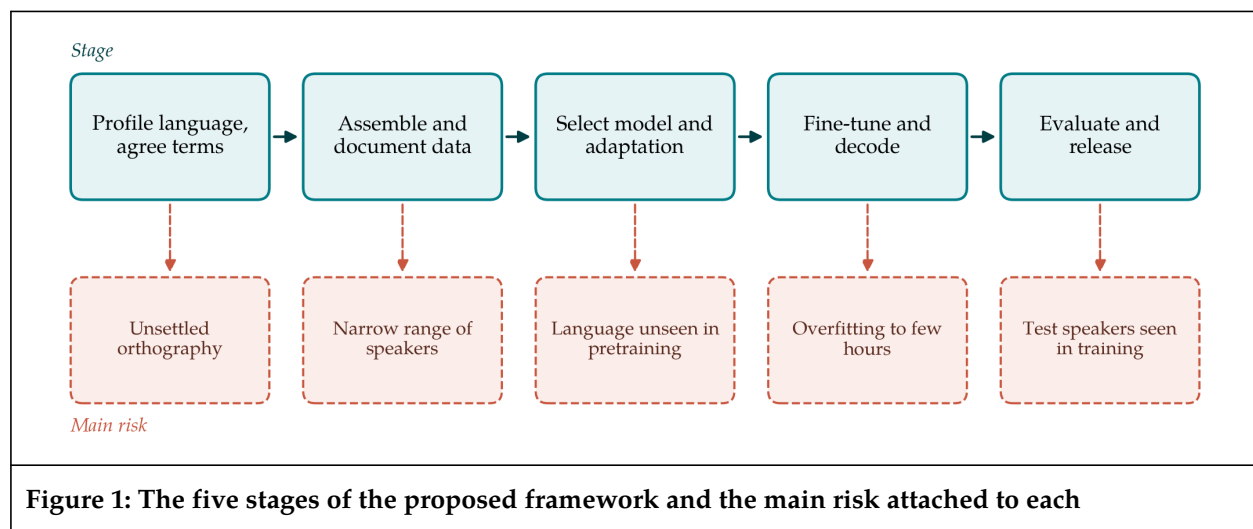
The framework has five stages. Table 1 gives the decisions and outputs of each, and Figure 1 pairs each stage with the risk most likely to undermine it. The stages are intended to be followed in order, though findings at evaluation will often send a team back to collect more data.

3.1. Stage 1: Profile the language and agree terms

Work begins with people, not audio. The team records which variety of the language is to be served, which dialects are in scope, and how the language is written in practice as well as officially. A normalisation document fixes the treatment of numerals, loanwords, punctuation, tone marks and code-switched passages; without it, two transcribers will write the same utterance differently and the model will be penalised for their disagreement. Speakers and community bodies should be partners in these choices. [Nekoto et al. \(2020\)](#) describe a participatory approach to machine translation for African languages in which speakers take part as creators and evaluators, and the same principle applies here. Consent, ownership and the licence of the eventual model are agreed in writing at this point.

Table 1: Stages of the proposed framework with the decisions taken and the output of each

Stage	Decisions to take	Output
1. Profile and agree	Variety and dialects covered; orthography and normalisation rules; consent and licence	Language profile; written agreement
2. Assemble data	Sources of speech and text; speaker balance; speaker-disjoint splits	Corpus with a data statement
3. Select model	Self-supervised or supervised base; full or parameter-efficient tuning	Justified choice of checkpoint and method
4. Fine-tune and decode	Output units; augmentation; stopping rule; language model for decoding	Trained model and configuration
5. Evaluate and release	Metrics; out-of-domain test; subgroup breakdown; terms of release	Evaluation report; model card



3.2. Stage 2: Assemble and document the data

Three kinds of material are sought: transcribed speech for fine-tuning, a separate transcribed set for testing, and text for a language model. Existing open collections are used first, then supplemented by recording, with attention to the balance of sex, age, region and recording device. Test speakers must not appear in training, and if the system is meant for conversation or radio then some test material should come from those settings and not only from read prompts. The corpus is released with documentation of who the speakers are and how the data were gathered, following the data statements of [Bender and Friedman \(2018\)](#) or the datasheets of [Geburu et al. \(2021\)](#).

3.3. Stage 3: Select a base model and adaptation method

The first question is whether the target language, or a close relative, was present in a candidate model's

pretraining data, which the model's documentation normally states. A self-supervised multilingual encoder with a new output layer is a sound default when the language was unseen, because nothing in the model is tied to a fixed vocabulary. A supervised encoder-decoder may be preferable when it already produces rough transcripts in the language. The second question is how much of the model to update. Full fine-tuning needs the most memory. Adapter modules inserted between frozen layers (Houlsby et al., 2019) and low-rank updates to existing weight matrices (Hu et al., 2022) train a small fraction of the parameters and suit a single graphics card.

3.4. Stage 4: Fine-tune and decode

Output units are chosen to match the script: characters for alphabetic orthographies, syllable characters for Ethiopic. With only a few hours of speech the chief danger is overfitting, and augmentation is the standard defence. SpecAugment masks bands of time and frequency in the input features (Park et al., 2019), and perturbing the speed of recordings creates additional training variants (Ko et al., 2015). Training stops on the error rate of a validation set and not after a fixed number of steps.

For models with a classification output layer, decoding with an external language model usually helps, since the acoustic model has seen little text in the language. An n-gram model estimated with a toolkit such as KenLM is cheap to build and query (Heafield, 2011). For agglutinative languages the model is better estimated over sub-word units than whole words, a strategy examined for Swahili by Gelas et al. (2012), and the text must pass through the same normalisation as the transcripts.

3.5. Stage 5: Evaluate and release

Word error rate is reported because it is the common currency, and character error rate is reported beside it. In a language where one long word encodes a clause, a single wrong morpheme makes the whole word an error, so word error rate overstates the difficulty relative to English and the two measures should be read together. Results are given separately for the in-domain and out-of-domain test sets and broken down by sex, age band and region. Koenecke et al. (2020) found large differences in the error rates of commercial recognisers between groups of speakers of one language, and a new system for a small language is not exempt. The model is released with a statement of its training data, measured performance, known weaknesses and the terms agreed at Stage 1.

4. Fitting the Framework to Local Conditions

The framework assumes little. Fine-tuning with parameter-efficient methods can be run on one consumer graphics card or on rented time, and all of the tools named are openly available. The scarcer resources are human. Transcription by trained native speakers is slow, and a realistic plan budgets more time for Stages 1 and 2 than for the modelling stages. Intermittent connectivity argues for keeping data and checkpoints on local storage, and for a final model small enough to run without a network connection where it will be used.

Tone deserves a deliberate decision. If the orthography does not mark it, the recogniser will not output it, and for dictation that may be acceptable to readers who are used to resolving it from context. For applications that feed a downstream system, such as translation, the team should decide with speakers whether unmarked output is adequate before data are transcribed, since adding tone marks afterwards means transcribing again.

5. Limitations and What Remains to Be Shown

The framework has not been applied. Its stages are drawn from methods that have each been published and used, but their combination in this order for an East African language has not been tested here, and it cannot be claimed that following it yields a recogniser of any particular accuracy. Three questions would need answers from actual projects: how much transcribed speech is required before the output is useful to speakers, how the choice between a self-supervised and a supervised base model plays out for languages of different families, and whether the participatory first stage changes the quality or uptake of the result.

The framework is also silent on matters that a full deployment must face. It treats one language at a time and does not address joint training across related languages, which may help where neighbours share vocabulary and sound systems. It does not cover languages with no written form. And it takes existing pretrained models as given, although their training data, drawn heavily from read religious texts or from the web, carry biases of speaker and style that fine-tuning may not remove.

6. Conclusion

Pretrained multilingual models have lowered the cost of building speech recognition for languages that were long left out, and the remaining difficulty lies less in modelling than in the choices that surround it. The five-stage framework proposed here places agreement with speakers and careful documentation ahead of training, matches the adaptation method to modest hardware, and insists on evaluation with unseen speakers and realistic speech. It remains a proposal, and its worth can be judged only when teams working on East African languages try it and report what they find.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F.M. and Weber, G. (2020). [Common Voice: A massively-multilingual speech corpus](#). *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218-4222.
- Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A. and Auli, M. (2022). [XLS-R: Self-supervised cross-lingual speech representation learning at scale](#). *Proceedings of Interspeech 2022*.
- Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020). [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). *Advances in Neural Information Processing Systems*, 33: 12449-12460.
- Bender, E.M. and Friedman, B. (2018). [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6: 587-604. https://doi.org/10.1162/tacl_a_00041
- Besacier, L., Barnard, E., Karpov, A. and Schultz, T. (2014). [Automatic speech recognition for under-resourced languages: A survey](#). *Speech Communication*, 56: 85-100. <https://doi.org/10.1016/j.specom.2013.07.008>
- Conneau, A., Baevski, A., Collobert, R., Mohamed, A. and Auli, M. (2021). [Unsupervised cross-lingual representation learning for speech recognition](#). *Proceedings of Interspeech 2021*.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K. (2021). [Datasheets for datasets](#). *Communications of the ACM*, 64(12): 86-92. <https://doi.org/10.1145/3458723>
- Gelas, H., Besacier, L. and Pellegrino, F. (2012). [Developments of Swahili resources for an automatic speech recognition system](#). *Proceedings of the Third Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*.
- Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. (2006). [Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks](#). *Proceedings of the 23rd International Conference on Machine Learning*, 369-376. <https://doi.org/10.1145/1143844.1143891>
- Heafield, K. (2011). [KenLM: Faster and smaller language model queries](#). *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187-197.

- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M. and Gelly, S. (2019). [Parameter-efficient transfer learning for NLP](#). *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97: 2790-2799.
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022). [LoRA: Low-rank adaptation of large language models](#). *Proceedings of the International Conference on Learning Representations*.
- Ko, T., Peddinti, V., Povey, D. and Khudanpur, S. (2015). [Audio augmentation for speech recognition](#). *Proceedings of Interspeech 2015*, 3586-3589.
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D. and Goel, S. (2020). [Racial disparities in automated speech recognition](#). *Proceedings of the National Academy of Sciences*, 117(14): 7684-7689. <https://doi.org/10.1073/pnas.1915768117>
- Nekoto, W., Marivate, V., Matsila, T. et al. (2020). [Participatory research for low-resourced machine translation: A case study in African languages](#). *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D. and Le, Q.V. (2019). [SpecAugment: A simple data augmentation method for automatic speech recognition](#). *Proceedings of Interspeech 2019*, 2613-2617. <https://doi.org/10.21437/Interspeech.2019-2680>
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A. and Auli, M. (2024). [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25.
- Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2023). [Robust speech recognition via large-scale weak supervision](#). *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202: 28492-28518.