# Enhancing Automatic Speech Recognition Models for Maternal and Reproductive Health: Fine-Tuning and Real-World Evaluation in Wolof

**Ertony Basilwango[1], Yann Le Beux[1], Oche David Ankeli[1], Pierre Herve Berdys[1]**
Dhananjay Balakrishnan[1,2],
[1] YUX Design, Senegal
[2]Stanford University, USA
{yann, pierre, ertony, oche}@yux.design
dhananjb@stanford.edu

## Abstract

Automatic Speech Recognition (ASR) systems perform well for high-resource languages, but most African languages, including Wolof, remain underrepresented, particularly in maternal and reproductive healthcare. This work proposes a domain-specific approach to improving Wolof ASR under low-resource conditions, addressing limited annotated data, orthographic variability, and code-switching. We curated a dataset of 750 validated Wolof utterances covering 250 maternal health keywords and applied data augmentation to increase acoustic diversity. Pretrained models, including wav2vec 2.0 and Whisper, were benchmarked to select candidates for fine-tuning. Using parameter-efficient Low-Rank Adaptation (LoRA), a Whisper model was adapted to the maternal health domain. Evaluation using Word Error Rate (WER), Character Error Rate (CER), and Keyword Error Rate (KER), which measures medically critical term transcription accuracy, shows substantial gains, reducing WER from 46.5% to 23.2% and KER from 17% to 11%. Community-based evaluation on 1,340 real-world utterances reveals a moderate degradation, with WER increasing by 35%. These results demonstrate that lightweight domain adaptation with small, high-quality data can significantly improve ASR for low-resource healthcare applications.This work introduces one of the first Wolof ASR datasets for healthcare and presents a practical framework for developing reliable speech recognition tools in underrepresented languages, improving access to healthcare information and services

**Keywords:** Automatic Speech Recognition, Low-Resource Languages, Wolof, Maternal Health, Data Augmentation, Domain Adaptation, Real World ASR Evaluation.

## 1 Context problem statement

Oral communication plays a central role in daily life across African societies, yet speech technologies have not evolved at the same pace for local languages(Caubrière and Gauthier, 2024). Existing transcription tools offer limited or no support for the majority of African languages, restricting the development of research, digital services, and domain-specific applications in health, finance, education, and governance (Imam et al., 2025). Although the African continent is home to more than 2,000 languages, only a very small fraction has any form of automatic speech recognition (ASR) resources, and even fewer have systems that perform reliably in real-world contexts. The few available corpora often contain government speeches or religious texts, which do not reflect spontaneous, conversational speech, and therefore hinder downstream performance in other domains (Mak et al., 2024).

Wolof poses significant challenges for ASR due to its rich phonology (17 vowels and 45 consonants) (Cissé and Sadat, 2023) and high orthographic variability shaped by regional usage, French code-switching, and the lack of a standardized writing system, leading to transcription and pronunciation inconsistencies that hinder both acoustic and language modeling(Bourdeau, 2024), (Cissé and Sadat, 2023). In Wolof, such variations are further amplified by flexible orthographic conventions and morphophonological alternations, making normalization and error handling essential preprocessing steps for reliable ASR(Aliou, 2010).

This technological gap disproportionately affects maternal and reproductive health, a domain where accurate documentation and communication are critical for patient safety, continuity of care, and public health monitoring. The lack of reliable ASR systems for local languages, increasing the risk of information loss and clinical error, further marginalizes already underserved populations and prevents healthcare systems from leveraging AI-driven efficiencies.

*Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 256–263
March 28, 2026 ©2026 Association for Computational Linguistics

# 2 Related works

## 2.1 ASR for Low-Resource African Languages

Automatic speech recognition (ASR) for African languages remains challenging due to limited data, high linguistic diversity, and the lack of standardized writing systems. Most African languages do not have enough labeled speech data, making it hard to build reliable ASR systems (Hedderich et al., 2021). Available datasets often come from formal sources, such as government speeches or broadcast news, which do not reflect everyday, conversational speech (Gauthier et al., 2016). Recent initiatives like AfriSpeech-200 aim to improve coverage across African accents (Olatunji et al., 2023), but they still do not fully capture domain-specific contexts such as healthcare, where vocabulary, pronunciation, and discourse differ. Evaluating ASR in low-resource settings is also difficult because there are no standard benchmarks or consistent annotation practices. Common metrics, like Word Error Rate (WER) and Character Error Rate (CER), are widely used to measure transcription accuracy (Jurafsky and Martin, 2009). However, these metrics are very sensitive to differences in spelling and writing style, which is especially challenging for African languages with evolving orthographies.

## 2.2 Wolof ASR and Linguistic Challenges

Wolof, spoken by 10–12 million people in Senegal, The Gambia, and Mauritania, is still underrepresented in speech technology. The early datasets (Gauthier et al., 2016);(Diop, 2015);(Aliou, 2010) )are limited in size and domain. The language shows a wide spelling variation, code-switching with French, and complex sound changes, which make ASR challenging (Aliou, 2010), (Bourdeau, 2024). Recent datasets, such as Kallaama (Gauthier et al., 2024), cover agriculture, but domain mismatch remains, highlighting the need for specialized corpora and preprocessing, especially for sensitive areas such as healthcare.

## 2.3 Domain Adaptation and Fine-Tuning for ASR

To address data scarcity, researchers have adapted pretrained ASR models to low-resource languages using limited labeled data. Self-supervised models like wav2vec 2.0 (Baevski et al., 2020), Massively Multilingual Speech(MMS)(Pratap et al., 2023) and Whisper (Radford et al., 2022) learn robust speech representations that transfer well. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) allow adaptation with lower computational cost, which is important in African contexts with limited GPU access. Prior work shows that lightweight, language-specific adaptation can outperform multilingual models trained on larger but less relevant datasets (Caubrière and Gauthier, 2024). These approaches are especially useful for domain-specific tasks, such as maternal health, where collecting labeled data is challenging.

## 2.4 Data Augmentation and Robustness

Data augmentation is commonly used to improve ASR in low-resource settings. Methods like speed perturbation, noise injection, pitch shifting, and volume changes can reduce Word Error Rate (WER) (Alex et al., 2023); (Ko et al., 2015), especially when collecting real-world data is difficult, such as in healthcare. However, models trained only on augmented data may still struggle in real-world conditions (Flynn and Ragni, 2024). Using imperfect or noisy speech can also help; carefully selected community-recorded data can improve acoustic modeling when clean data is scarce. (Badenhorst and de Wet, 2019).

## 2.5 Real-World and Human-Centered Evaluation of ASR

Standard ASR evaluation on clean benchmark datasets often fails to reflect real-world performance. Models tested on a single dataset can lose 35–50% WER when deployed in different domains or acoustic conditions (Likhomanenko et al., 2021); (Shah et al., 2024); In healthcare, this is critical: even models with low overall WER can misrecognize medical terms, risking patient safety (Afonja et al., 2024). Fine-tuning on domain-specific, accented clinical speech improves recognition of medical entities, highlighting the need for domain-aware and human-centered evaluation. Moreover, speech quality metrics, like DNSMOS (Reddy et al., 2021), do not always predict ASR accuracy, emphasizing the importance of direct evaluation on real-world data.

# 3 Study Area, Data, and Methods

## 3.1 Data collection

In collaboration with domain experts and community partners, we created a Wolof dataset of 750

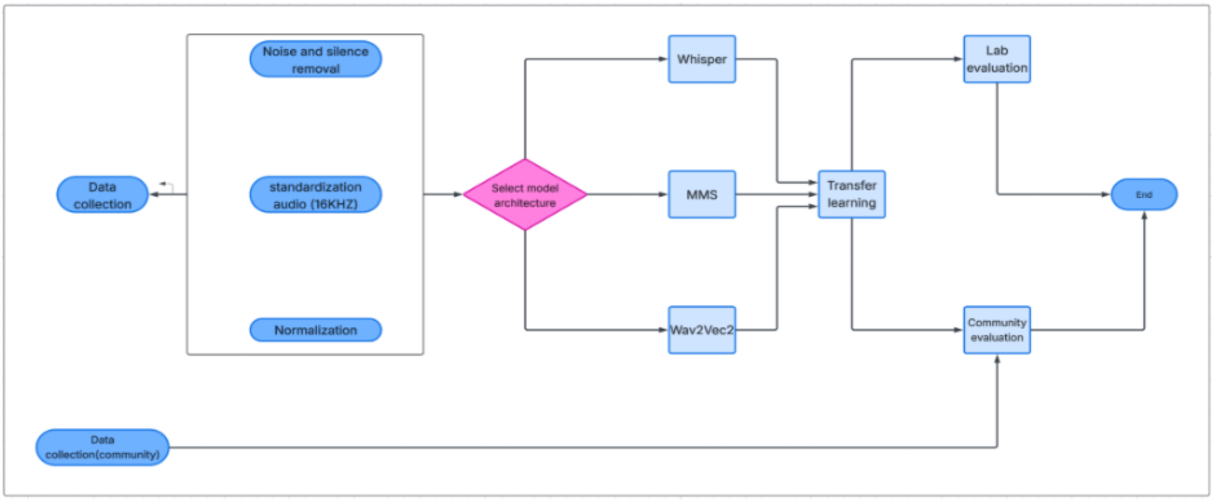
Noise and silence removal
Data collection
standardization audio (16KHZ)
Normalization
Select model architecture
Whisper
MMS
Wav2Vec2
Transfer learning
Lab evaluation
End
Community evaluation
Data collection(community)

Figure 1: High-Level Workflow

utterances (2 hours) based on 250 maternal and reproductive health keywords spanning antenatal care, delivery, complications, infections, and family planning. Phrases were translated and culturally validated, then recorded via the Looka platform, a digital data collection system that facilitates remote recording and survey participation, by 20 regionally diverse speakers with balanced gender representation. To reduce sensitivity to individual speakers or specific keywords, we prioritized speaker diversity, applied data augmentation, and performed utterance-level splits while maintaining speaker diversity across train and test sets. While some sensitivity is inevitable in low-resource settings, this design mitigates overfitting to particular speakers or lexical items. Each iteration was reviewed by linguists, health professionals, and social scientists to identify and correct potential biases.

### 3.2 Ethics, Privacy, and Deployment Constraints

Given the sensitive nature of maternal and reproductive health data, all participants provided informed consent. Audio files were anonymized and securely stored, and risk mitigation measures were implemented to ensure participation posed no harm. These steps make the dataset ethically suitable for research in healthcare contexts.

### 3.3 Pre-processing

- Normalization

Wolof shows significant lexical and orthographic variation, where multiple forms can express the same meaning depending on region, speaker, or transcription conventions (Afonja et al., 2024). To address this, we applied a multi-stage normalization pipeline, including spelling standardization, letter-case normalization, canonical representation of numbers, and removal of non-essential punctuation (Rahimi and Homayounpour, 2022).

| Mispellings | Correct wolof |
|---|---|
| dadial | dajale |
| guinaw | guinnaw |
| Mousiba | Musiba |
| Infection | infekcion |
| gnakk | ñakk |
| thiosane | cosaan |

Table 1: Wolof Spelling Normalization Examples

- Data augmentation

To improve the robustness of ASR in low-resource settings, we applied standard data augmentation techniques known to reduce WER (Ko et al., 2015; Bagchi et al., 2020). Using speed perturbation (±10%), pitch shifting (±2 semitones), volume adjustment (±3 dB), and additive background noise (SNR 20–30 dB) in the original data set resulted in approximately 10 hours of data in total. This approach complements naturally noisy speech, which is beneficial in low-resource ASR settings (Badenhorst and de Wet, 2019).

### 3.4 Benchmarking of ASR Models on Maternal and Reproductive Health Data

Before fine-tuning, open-source models were benchmarked on 120 cleaned samples from our maternal health dataset using Word Error Rate

| Models | WER | CER | Latency(s) |
|---|---|---|---|
| **Alwaly/whisper** | 0.464 | 0.172 | 1.394 |
| facebook/mms | 0.526 | 0.181 | 0.55 |
| bilalfaye/wav2vec2 | 0.533 | 0.186 | 0.55 |
| CAYTU/whisper | 0.544 | 0.228 | 1.626 |
| cibfaye/whisper | 0.546 | 0.224 | 0.463 |

Table 2: Benchmarking of ASR models on maternal and reproductive health Wolof data.

(WER) and Character Error Rate (CER) (Rahimi and Homayounpour, 2022). Models with WER below 50% were selected for domain-specific adaptation, as they provide a sufficient baseline for learning from limited in-domain data. The threshold therefore serves as a computationally efficient screening step rather than an optimal boundary. This pre-selection follows established ASR transfer learning practice, where models with reasonable initial performance adapt more effectively than poorly performing ones (Baevski et al., 2020); (Hu et al., 2021)

- Model Notation

These models are referred to by their abbreviated names in the tables and figures throughout the paper
- Alwaly/whisper: Alwaly/whisper-medium-wolof
- CAYTU/whisper: CAYTU/whisper-large-v2
- Facebook/mms: Facebook/mms-1b-fl102
- bilalfaye/wav2vec2: bilalfaye/wav2vec2-large-mms-1b-wolof

### 3.5 Fine-tuning Approach

We fine-tuned three different speech recognition models on Wolof medical conversations using Low-Rank Adaptation (LoRA), a lightweight fine-tuning method that updates only a small part of each model (Hu et al., 2021). This approach enables the models to capture local speech patterns, accents, and domain-specific pronunciations even with a small amount of annotated data, while keeping computational requirements low and training times short. By updating only a subset of parameters, LoRA provides an efficient way to specialize large pre-trained models to underrepresented languages and healthcare-specific speech without the need for extensive hardware or massive datasets.

The maternal and reproductive health data set was divided into 80% for training (10 hours of audio) and 20% for testing (2 hours). Fine-tuning was performed for 20 epochs with evaluation at each epoch, using mixed-precision training (fp16) and the AdamW optimizer (Loshchilov and Hutter, 2019). Training was conducted on an NVIDIA A100-SXM4 GPU (40 GB) with CUDA 12.4.

The results show that the Alwaly model achieved the lowest Word Error Rate (WER)

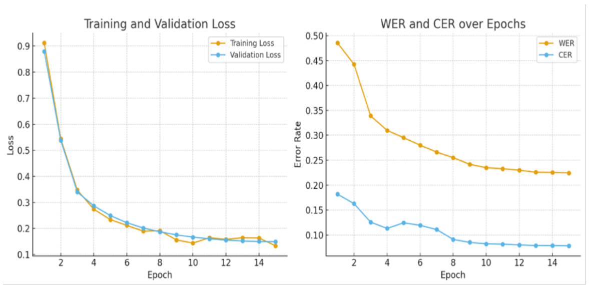

Figure 2: Training and validation loss curves of the Alwaly model after fine-tuning

The first graph shows the training and validation loss, both decreasing over epochs, indicating successful model convergence. The second graph illustrates the reduction in WER (orange) and CER (blue), reflecting improvements in the model's speech transcription accuracy, particularly for maternal and reproductive health in Wolof.

| Models | WER | CER |
|---|---|---|
| **Alwaly/whisper** | 0.23 | 0.172 |
| CAYTU/whisper | 0.374 | 0.181 |
| Facebook/mms | 0.406 | 0.308 |

Table 3: Model performance after fine-tuning on maternal and reproductive health data

In addition to the main Whisper-based model, we also fine-tuned CAYTU/Whisper-large-v2 and Facebook/MMS. Although both benefited from LoRA adaptation, the performance gains were smaller and their higher inference latency makes them less suitable for lightweight deployment in low-resource healthcare settings. These results confirm that selecting a pretrained model with a reasonable baseline WER is crucial for effective domain adaptation and demonstrate that our conclusions are not limited to a single model architecture.

To assess the impact of data augmentation, we conducted an additional experiment fine-tuning the selected Alwaly/Whisper model model using only the original 2-hour dataset without augmentation. This setting resulted in a WER of 29%, compared to 23% WER when training with augmented data. This confirms that data augmentation plays a significant role in improving robustness under low-resource conditions. While all these augmentation techniques were applied collectively, we did not separately quantify the individual contribution of each method.

### 3.6 Evaluation of medical terms

Recent research has highlighted weaknesses in relying solely on the word error rate (WER) for morphologically rich languages (K et al., 2025). To capture domain-specific accuracy, we used the Keyword Error Rate (KER), focusing on maternal and reproductive health terms. KER evaluates transcription quality at the character level by aligning predicted and reference keywords. It is calculated using dynamic programming to estimate the edit distance of Levenshtein (Levenshtein, 1966), which calculates the difference between two strings by counting the insertions, deletions, and substitutions needed to transform one string into another.

Standard metrics such as WER and CER often fail to capture errors in domain-critical terms. To address this, we evaluated the keyword error rate (KER), a metric that quantifies the precision of recognition of a predefined set of maternal and reproductive health keywords, thereby assessing the model's ability to capture domain-critical terminology. Keywords were detected using fuzzy string matching, Fuzzy matching compares strings by measuring edit distance, that is, the minimum number of character-level insertions, deletions, or substitutions required to transform one string into another, allowing minor differences between recognized and reference text. A sliding-window alignment was applied to match short segmented text with the target keywords. Fine-tuning reduced KER from 0.169 to 0.11, demonstrating improved recognition of medically important terms and highlighting the value of domain-sensitive evaluation in healthcare ASR.

$$\text{KER} = \frac{S + D + I}{N}$$

Where S, D, I and N correspond to the number substitutions, deletions and insertions in the total number of keyword found. To identify keywords in ASR outputs, we applied fuzzy string matching with text normalization and a sliding-window search, selecting the closest match based on Levenshtein similarity.

**Sliding-Window** Search is used to find matching substrings between the expected keyword and the ASR output by comparing small portions of both strings and selecting the best match as the window moves across the text. **Levenshtein Similarity** then calculates the inverse of the Levenshtein distance, quantifying the similarity between the two strings. This method played a key role in the model's fine-tuning, which resulted in a significant reduction in the Keyword Error Rate (KER), dropping from 0.17 to 0.11. This improvement highlights the model's increased ability to capture domain-specific, medically relevant terminology.

| Model | WER | CER | KER |
|---|---|---|---|
| Alwaly | 46.46% | 17% | 17% |
| Alwaly(fine-tuned) | 23.16% | 7.83% | 11% |

Table 4: Model performance evaluation based on the KER metric (Medical Keyword Error Rate)

### 3.7 Evaluation on F1-score

The F1-score was used to evaluate performance due to its effectiveness in health-related assessments, especially for measuring accurate identification of medical terminology. An F1-score of 0.809 reflects good model performance, indicating reliable recognition of medical keywords. (Sokolova and Lapalme, 2009)

### 3.8 Human-Centered Evaluation

Standard ASR benchmarks often fail to reflect real-world conditions, with performance degrading on noisy, out-of-domain speech (Likhomanenko et al., 2021). Following recommendations for community-grounded evaluation (Khan et al., 2024), we assessed our Wolof ASR using community-collected recordings capturing natural variation in accent, pronunciation, and recording environments.

Participants from multiple regions of Senegal produced their own maternal and reproductive health phrases and recorded them in everyday settings. The resulting dataset contains 1,340 unaugmented recordings ( 3 hours of speech), balanced by gender (50% female) and dialectally diverse. Compared to controlled data, real-world evaluation revealed a performance drop (WER: 0.23 → 0.31; CER: 0.08 → 0.12), driven by ambient noise, regional accents, and spontaneous speech. These results demonstrate that human-centered, ecologically valid evaluation is essential for deploying ASR systems in low-resource contexts.

| Evaluation setting | WER | CER |
|---|---|---|
| Internal test set (clean) | 23% | 8% |
| Real-world evaluation | 31% | 12% |

Table 5: Human Centered Evaluation

### 3.9 Error Analysis and Linguistic Observations

Qualitative analysis of Wolof ASR outputs reveals that many residual "errors" correspond to orthographic or phonetic variants rather than true recognition failures. These variants often arise from natural speech patterns, regional accents, and code-switching, while still preserving the intended meaning. As a result, conventional WER and CER metrics may overestimate error rates in such contexts. Table 6 provides representative examples of these variants.

These observations motivate the use of Keyword Error Rate (KER) alongside WER and CER for domain-specific ASR evaluation, and normalization prior to computing evaluation metrics or fine-tuning the model on Wolof data.

## 4 Experiments and Results

### 4.1 Benchmarking of Pretrained ASR Models

We evaluated several open-source ASR models on 120 Wolof maternal and reproductive health utterances using Word Error Rate (WER), Character Error Rate (CER), and inference latency. Domain-adapted and Wolof-specific models consistently outperformed general multilingual systems. Alwaly/whisper-medium-wolof achieved the best baseline performance (WER: 46.4%, CER: 17.2%) and was selected for further adaptation, while models with WER above 50% were excluded.

### 4.2 Fine-Tuning with LoRA

The selected Whisper model was fine-tuned using Low-Rank Adaptation (LoRA) on approximately 10 hours of augmented maternal health speech. Fine-tuning substantially reduced error rates (Table 4), confirming the effectiveness of lightweight domain adaptation for medical ASR. To reflect real-world usage, evaluation was performed with consistently reduced CER across models, indicating improved robustness to orthographic variation in Wolof.

To better capture domain-critical errors, we evaluated Keyword Error Rate (KER) on maternal and reproductive health terminology using normalized text and fuzzy string matching. Fine-tuning reduced KER from 0.169 to 0.11, demonstrating improved recognition of medically important terms and highlighting the limitations of WER and CER for domain-specific evaluation.

### 4.3 Human-Centered Real-World Evaluation

We further evaluated the model on 1,340 community-recorded utterances collected in everyday environments across Senegal. Compared to controlled test data, performance degraded (WER: 0.23 → 0.31; CER: 0.08 → 0.12), reflecting the impact of noise, regional accents, and spontaneous speech. This confirms that benchmark-only evaluation overestimates real-world ASR performance

## 5 Key contributions

Our work further demonstrates that, even with a small, carefully curated dataset, effective ASR performance can be achieved through data augmentation and LoRA fine-tuning, while evaluation on a larger, real-world dataset ensures robustness. This shows that in low-resource, domain-specific contexts, strategically combining high-quality limited data with larger evaluation corpora allows for reliable ASR deployment, making it possible to support critical applications such as maternal and reproductive healthcare in Wolof-speaking communities.

## 6 Limitations and Perspectives

This study relies on a small curated dataset and a predefined set of maternal health keywords, designed to prioritize linguistic validity, domain relevance, and speaker diversity rather than scale. As a result, findings may not fully generalize to unrestricted conversational speech. Future work will explore larger speaker pools, broader vocabularies, and systematic ablation studies. We plan to improve Wolof ASR using multilingual and cross-lingual strategies, leveraging related languages to enhance performance in extremely low-resource healthcare settings. Additional training data will include more speakers, regions, everyday environments, and advanced augmentation techniques to better reflect real-world variability. Future efforts will also focus on developing population-specific lexicons tailored to education level, region, and

| **Reference** | **Hyphothesis** | **Error category** |
|---|---|---|
| **Klinik** | Kilinig | accent / code-switching |
| Pediatre | Pejiatre | accent / accent / code-switching |
| Dagnuy | Danu | standard variant |
| infection | infekcion | code-switching variant |

Table 6: Qualitative analysis of model errors

health context, enabling finer-grained evaluation and adaptation. Finally, we aim to expand linguistic and dialectal coverage to include additional Wolof dialects and other low-resource African languages.

## 7 Conclusion

This study shows that lightweight LoRA fine-tuning of pretrained ASR models, combined with small but carefully curated domain-specific datasets and real-world evaluation, can substantially improve Wolof ASR performance for maternal and reproductive health. Although benchmark results show strong gains, performance degradation in naturalistic conditions underscores the need for human-centered and domain-aware evaluation when developing ASR systems for low-resource healthcare settings. By focusing on Wolof maternal health, this work serves as a case study demonstrating how targeted data collection and lightweight adaptation can yield effective domain-specific ASR. These findings provide a practical foundation for extending similar approaches to other low-resource languages and healthcare domains in future research.

## References

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant ASR Models for Medical Entities in Accented Speech. In *Interspeech 2024*, pages 2315–2319.

Ashish Alex, Lin Wang, Paolo Gastaldo, and Andrea Cavallaro. 2023. Data augmentation for speech separation. *Speech Communication*, 152:102949.

Diuf Aliou. 2010. Some morphonological phenomenons in suffixations of wolof language. *Russian Journal of Linguistics*, (3):50–53.

Jaco Badenhorst and Febe de Wet. 2019. The usefulness of imperfect speech data for asr development in low-resource languages. *Information*, 10(9):268.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Corentin Bourdeau. 2024. The wolof basic clause and its information-structural derivatives. *Linguistique et langues africaines*, 10(2).

Antoine Caubrière and Elodie Gauthier. 2024. Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien. In *Actes des 35èmes Journées d'Études sur la Parole*, pages 163–172, Toulouse, France. ATALA and AFPC.

Thierno Ibrahima Cissé and Fatiha Sadat. 2023. Automatic spell checker and correction for under-represented spoken languages: Case study on Wolof. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

T. I. Cissé and F. Sadat. 2023. Automatic spell checker and correction for under-represented spoken languages: Case study on wolof. *arXiv*.

Faje Fatu Diop. 2015. Active processes in the senegalese linguoculture and the wolof language in the course of globalization. *RUDN Journal of Language Studies, Semiotics and Semantics*, 3:89–94. English translation of Russian original.

Robert Flynn and Anton Ragni. 2024. How much context does my attention-based asr system need? In *Proceedings of Interspeech 2024*, Kos, Greece. ISCA.

Elodie Gauthier, Aminata Ndiaye, and Abdoulaye Guissé. 2024. Kallaama: A transcribed speech dataset about agriculture in the three most widely spoken languages in Senegal. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 10–19, Torino, Italia. ELRA and ICCL.

Etienne Gauthier, Laurent Besacier, Stephanie Voisin, Mekuriaw Melese, and Ulrich P. Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: A case study of wolof. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2545–2568, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.

Hania Khan, Aleena Fatima Khalid, and Zaryab Hassan. 2024. Transcending controlled environments assessing the transferability of asrrobust nlu models to real-world applications. *Preprint*, arXiv:2401.09354.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3589.

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

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2021. Rethinking evaluation in asr: Are our models robust enough? *Preprint*, arXiv:2010.11745.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Franco Mak, Avashna Govender, and Jaco Badenhorst. 2024. Exploring asr fine-tuning on limited domain-specific data for low-resource languages. *Journal of the Digital Humanities Association of Southern Africa*, 5(1):—.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Preprint*, arXiv:2310.00274.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Zahra Rahimi and Mohammad Mehdi Homayounpour. 2022. The impact of preprocessing on word embedding quality: A comparative study. *Language Resources and Evaluation*.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*.

Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkila, Bhiksha Raj, and Nicolas Kourtellis. 2024. Speech robust bench: A robustness benchmark for speech recognition. *Preprint*, arXiv:2403.07937.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.