Contributing to Speech-to-Speech Translation for African Low-Resource Languages : Study of French-Mooré Pair

Fayçal S.A. Ouédraogo[1,2], Maimouna Ouattara[1,4], Rodrique Kafando[1], Abdoul Kader Kaboré[1,4], Aminata Sabané[1,3] and **Tegawendé F. Bissyandé[1,4]**

[1] Centre d'Excellence Interdisciplinaire en Intelligence Artificielle pour le Developpement (CITADEL)
[2] Ecole Polytechnique de Ouagadougou [3] Université Joseph KI-ZERBO
[4] University of Luxembourg

Correspondence: fayouedraogo@protonmail.com, maimouna.ouattara@uni.lu, rodrique.kafando@citadel.bf, abdoulkader.kabore@uni.lu, aminata.sabane@ujkz.bf, tegawende.bissyande@uni.lu

Abstract

Most of African low-resource languages are primarily spoken rather than written and lack large, standardized textual resources. In many communities, low literacy rates and limited access to formal education mean that text-based translation technologies alone are insufficient for effective communication. As a result, speech-to-speech translation systems play a crucial role by enabling direct and natural interaction across languages without requiring reading or writing skills. Such systems are essential for improving access to information, public services, healthcare, and education. The goal of our work is to build powerful transcription and speech synthesis models for Mooré language. Then, these models have been used to build a cascaded voice translation system between French and Mooré, since we already got a French-Mooré machine translation model. We collected Mooré audio-text pairs, reaching a total audio duration of 150 hours. Then, We fine-tuned Orpheus-3B and XTTS-v2 for speech synthesis and Wav2Vec-Bert-2.0 for transcription task. After fine-tuning and evaluation by 36 Mooré native speakers, XTTS-v2 achieved a MOS of 4.36 out of 5 compared to 3.47 out of 5 for Orpheus-3B. The UTMOS evaluation resulted in 3.47 out of 5 for XTTS-v2 and 2.80 out of 5 for Orpheus-3B. The A/B tests revealed that the evaluators preferred XTTS-v2 Mooré audios in 77.8% of cases compared to 22.2% for Orpheus-3B. After fine-tuning on Mooré, Wav2Vec-Bert-2.0 achieved a WER of 4.24% and a CER of 1.11%. Using these models, we successfully implemented a French-Mooré Speech-to-Speech Translation system.

1 Introduction

Despite being classified as a low-resource language, Mooré is the most widely spoken local language in Burkina Faso, with 52.9% of the population using it as their primary language[1]. In contrast, French, which serves as a work language, dominates formal communication channels including news media, banking services, educational institutions, healthcare facilities, and government administration. This linguistic asymmetry creates a significant language barrier that marginalizes Mooré-speaking populations, limiting their access to essential services and information in their native language. Furthermore, Mooré exhibits a predominantly oral tradition, with speakers far outnumbering readers and writers of the language. This oral-first characteristic is common among many African languages, yet existing natural language processing research has focused predominantly on text-based translation systems. The gap between the oral nature of Mooré usage and text-centric NLP approaches represents a critical limitation in making translation technology accessible to the majority of Mooré speakers. As (Ouattara et al., 2025) argue, African low-resource languages urgently need speech-based systems that align with their actual patterns of use. Recognizing this need, our work aims to develop practical speech translation solutions that can genuinely serve Mooré-speaking communities in their daily interactions with French-dominated institutions and services.

In this work we have developed Automatic Speech Recognition (ASR, or speech-to-text) and Text-To-Speech (TTS, or speech synthesis) models for Mooré. These models have been joined to a French-Mooré Machine Translation (MT) model, with other models, to create a Cascaded Speech-to-Speech Translation (S2ST) system.

Our work contributed to release Mooré paired audio-text corpora of about 88.000 utterances and 150 hours. We also fine-tuned high-performance ASR and TTS models for Mooré, able to handle properly tone and accent.

In the following sections, we present in Section 2 the Mooré language and a literature review of related works. Section 3 describes our data and

[1] https://www.insd.bf/fr/resultats

collecting method, followed by our fine-tuning approach for Mooré ASR and TTS models. Section 4 presents the experimental results, followed by the conclusion in Section 5.

2 Background

In this section we first present the Mooré language, then we review some related works according to ASR, TTS and S2ST for low-resource languages.

2.1 The Mooré language

Mooré, also known as Mossi (ISO 639-3: `mos`) is a language belonging to the Niger–Congo family, more specifically to the Gur group of the Oti–Volta branch, and is widely spoken in Burkina Faso as well as in several neighboring countries (Hartell and Bendor-Samuel, 1989; Lewis et al., 2016). The language is estimated to have between 5 and 8 million native speakers.

From a linguistic perspective, Mooré is a tonal language, characterized by the use of distinctive tone levels at both the lexical and grammatical levels (Beckman and Venditti, 2010). These tones are rarely marked systematically in the standard orthography, which poses a significant challenge for speech recognition and speech synthesis systems.

The Mooré language is written using a Latin-based alphabet officially standardized in Burkina Faso. The alphabet consists of 21 consonant letters and 7 vowel letters, including both oral and nasal vowels. In addition to the five basic vowels *a, e, i, o, u*, Mooré distinguishes the mid vowels like *ɛ*, which plays a phonemic role.

2.2 Automatic Speech Recognition (ASR)

ASR for African low-resource languages faces significant hurdles due to data scarcity, linguistic complexity, and limited computational resources (Imam et al., 2025). Recent research focuses on developing comprehensive datasets, applying self-supervised learning (SSL), and utilizing end-to-end models (Imam et al., 2025). SSL and transfer learning are crucial for mitigating data scarcity. Pre-trained models such as Whisper, XLS-R, MMS, and W2v-BERT have been benchmarked across 13 African languages (Nahabwe et al., 2025). Fine-tuning wav2vec 2.0 models on minimal data yielded promising results for Ika (Nzenwata and Ogbuigwe, 2024). AfriHuBERT extended coverage to 39 African languages using over 6,500 hours of speech data, enhancing Language Identification and ASR (Alabi et al., 2024). Multilingual modeling and cross-lingual transfer also leverage high-resource language data for low-resource languages (Thangaraj et al., 2024).

End-to-end Transformer models are gaining prominence for directly mapping acoustic features to text. This architecture, exemplified by Speech-Transformer, has been applied to languages like Central Kurdish (Abdullah et al., 2024b) and Northern Kurdish (Abdullah et al., 2024a), aiming to improve performance through deep learning (Tan, 2023).

Data augmentation, including synthetic voice generation for Hausa, Dholuo, and Chichewa (DeRenzi et al., 2025), and latent mixup techniques (Bian et al., 2025), further expands training corpora. Despite progress, challenges persist, such as handling phonetic variations in African American English (Mojarad and Tang, 2025), diacritics in Hausa (Abubakar et al., 2024), and code-switching in South African languages (Biswas et al., 2022).

2.3 Text-to-Speech (TTS)

Developing TTS systems for low-resource languages faces significant difficulties primarily due to the scarcity of high-quality speech corpora and corresponding text data (Imam et al., 2025). Recent research on neural TTS for low-resource languages has focused on data-efficient learning through transfer, multilingual modeling, and self-supervised pre-training. Cross-lingual adaptation of sequence-to-sequence models such as Tacotron and FastSpeech enables leveraging high-resource languages to synthesize speech with limited paired data (Chen et al., 2019; Zhang and Lin, 2020). Multilingual and zero-shot TTS systems further reduce data requirements by learning shared acoustic and linguistic representations across languages (Casanova et al., 2024). To support these approaches, BibleTTS (Meyer et al., 2022) offers a high-fidelity, open-access speech corpus featuring up to 86 hours of aligned, studio-quality 48kHz single-speaker recordings per language across ten Sub-Saharan African languages. By providing verse-aligned audio-text pairs derived from Biblica's Open.Bible project, the dataset facilitates the development of TTS models for low-resource languages without relying on pre-existing acoustic or grapheme-to-phoneme models. Additionally, the corpus has been used to train TTS models with Coqui TTS, demonstrating its effectiveness for both in-domain and out-of-domain synthesis tasks. Self-supervised speech representations

and compact latent codes have also been shown to improve synthesis quality in low-resource settings (Guo et al., 2022). More recently, parameter-efficient fine-tuning and continual learning strategies have been explored to adapt large pretrained TTS models to new low-resource languages with minimal additional data (Kwon et al., 2025).

2.4 Speech-to-Speech Translation (S2ST)

S2ST for low-resource languages has seen significant methodological divergence between cascaded and direct approaches. Cascaded systems, which sequentially integrate ASR, MT, and TTS, benefit from modular design and transfer learning, enabling adaptation to low-resource settings through multilingual training and data augmentation (Krishna Paleti et al., 2024)(Pradeep Kumar et al., 2025). However, they are inherently susceptible to error propagation across stages, degrading overall performance (Gupta et al., 2025). In contrast, direct S2ST models bypass intermediate text representations by leveraging end-to-end architectures, reducing latency and improving robustness in streaming scenarios (Ochieng and Kaburu, 2024)(Fang et al., 2024). Recent advances employ self-supervised discrete units and pre-trained speech encoders to mitigate data scarcity (Communication et al., 2023), with frameworks like UnitY, SeamlessM4T and Translatotron demonstrating feasibility even without parallel speech-text pairs (Li et al., 2022)(Lin et al., 2024). Techniques such as pseudo-labeling, synthetic data generation, and cross-lingual transfer further enhance performance in under-resourced contexts (Dong et al., 2022; Popescu-Belis et al., 2025). Despite progress, challenges persist in achieving high fidelity and natural prosody, particularly for unwritten or severely low-resource languages.

3 Methodology

In this work, we collected, segmented and aligned audio-text data in order to train ASR and TTS models for Mooré. These models have been joined to other models to build a S2S Translation system based on the cascaded approach illustrated in Figure 1.

We used the cascaded S2ST approach instead of the direct one, because of the lack of paired French-Mooré audios data in sufficient quality and quantity. In addition, the cascaded approach was more convenient for us due to scarcity of aligned French-Mooré audios. We also already had a French-Mooré Machine Translation model. This approach also allows us to directly setup Text-to-Speech Translation and Speech-to-Text Translation systems without further data collection and model training.

For Mooré speech synthesis, we fine-tuned and compared two TTS models: Orpheus-3B and XTTS-v2. For the Mooré ASR task, we fine-tuned a wav2vec-Bert-2.0 transcription model. Our French-Mooré S2ST system integrates several other components: a fine-tuned NLLB-200-distilled-600M model from CITADEL for French-to-Mooré MT, Whisper Turbo for French ASR, and the XTTS-v2 base version for French TTS in the translation pipeline.

3.1 Data Collection and Preprocessing

Our Mooré Speech datasets are built from religious websites such as jw.org and bible.com, that have different translations of Bible. These websites allow users to read and listen Bible in many different languages including a lot of low-resource ones. These websites support Mooré language giving both text and audio data. We collected text and audio on jw.org using the python package *jwsoup*[2]. The data originating from bible.com website have been collected using traditional scraping techniques.

The data gathered after this step was not paired and there was a single audio for an entire chapter of the Bible. So we used fairseq toolkit (Ott et al., 2019) to automatically split audios according to verses and pair each audio sample with the related verse. In addition, audios have been converted to wav format and the sample rate has been set to 24kHz. These steps have been applied to data from jw.org and bible.com, which allowed us to create two different datasets. We initially also got a legacy audio-text Mooré dataset from CITADEL, which was small and monospeaker. All the datasets we used are described in Table 1.

3.2 Models fine-tuning

3.2.1 Orpheus-3B

Orpheus-3B is a state-of-the-art open-source text-to-speech model released by Canopy Labs in March 2025, built on the Llama-3B architecture and pre-trained on over 100 000 hours of primarily English speech data and billions of text tokens (Hackster.io,

[2]https://github.com/sawallesalfo/jwsoup

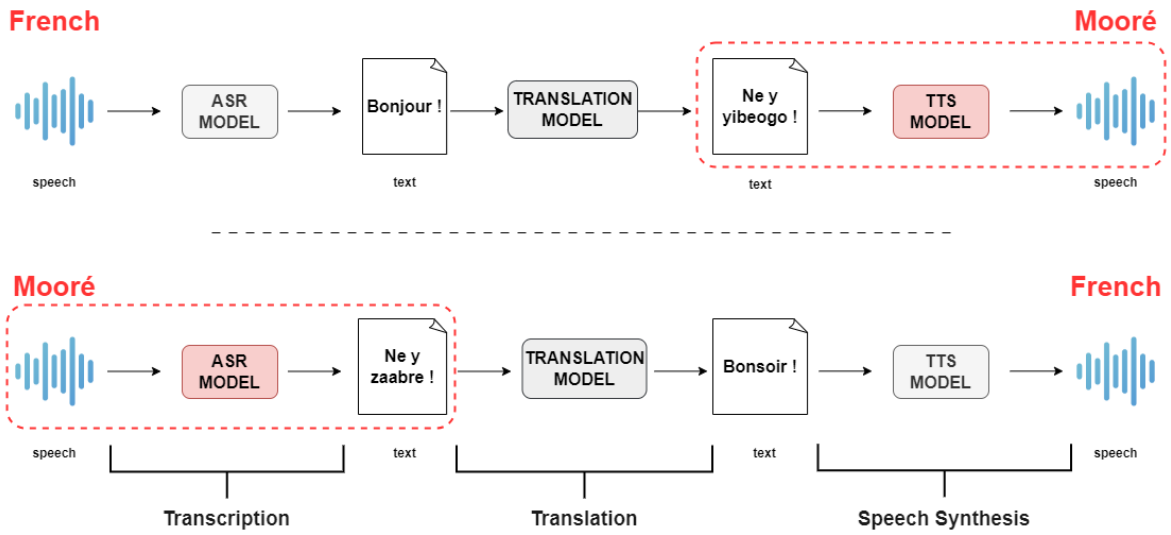

Figure 1: French-Mooré S2S Translation Architecture

Feature	Dataset 1	Dataset 2	Dataset 3
Source	Bible	JW.org	Bible.com
Size (utterances)	3 099	27 068	57 410
Multispeaker	No	Yes	Yes
Sample rate (kHz)	24	24	24
Min. length (s)	1.20	1.53	0.34
Max. length (s)	18.09	40.4	86.5
Total duration (h)	6.48	70.04	74.00

Table 1: Datasets caracteristics

Hyperparameter	Value
LoRA rank r	64
lora_dropout	0
lora_aplha	64
per_device_batch_size	1
gradient_accumulation_steps	4
max_steps	200
learning_rate	2e-4

Table 2: Orpheus-3B training hyperparameters

2025; Canopy Labs, 2025b). It enables zero-shot voice cloning and emotional speech synthesis, with high potential for multilingual adaptation even for low-resource languages (Canopy Labs, 2025a).

This model only supports english natively. We used the version of this model fine-tuned by Unsloth and made a LoRA fine-tuning on Mooré data. For this training we only use the dataset 1 described in Table 1 with the training hyperparameters in Table 2.

3.2.2 XTTS-v2

XTTS-v2, released by Coqui AI in 2023, supports 17 languages (English, Spanish, French, German, Italian, Portuguese, Polish, Turkish, Russian, Dutch, Czech, Arabic, Chinese, Japanese, Hungarian, Korean, Hindi) with no explicit African language (Casanova et al., 2024). Trained on 27 000 hours of multilingual speech data, it uses a GPT-style autoregressive architecture with convolutional encoders for zero-shot voice cloning from 6-second clips.

Hyperparameter	Value
batch_size	8
grad_acumm	4
learning_rate	5e-6
optimizer	AdamW
weight_decay	1e-2
lr_scheduler	MultiStepLR

Table 3: XTTS-v2 training hyperparameters

Given that XTTS-v2 doesn't support Mooré natively, we extended it's vocabulary size to 4000 with additional Mooré tokens using texts from dataset 2. In addition, we applied a full fine-tuning for 20 epochs on dataset 2 and then 20 additional epochs on dataset 3. The training hyperparameters details are in Table 3.

3.2.3 Wav2Vec-Bert-2.0

Wav2Vec2-BERT-2.0, released by Meta AI in 2024, is a self-supervised speech representation model pretrained on 4.5 million hours of unlabeled multilingual audio across over 140 languages including many low-resource ones (Communication et al., 2023). It extends wav2vec 2.0 with a BERT-style masked prediction layer on top convolutional encoders and transformers blocks for contextual speech understanding, excelling in low-resource

Hyperparameter	Value
group_by_length	True
per_device_train_batch_size	8
gradient_accumulation_steps	8
eval_strategy	steps
num_train_epochs	7.0
gradient_checkpointing	True
fp16	True
learning_rate	5e-5
warmup_steps	500

Table 4: Wav2Vec-BERT-2.0 training hyperparameters

ASR (Baevski et al., 2020).

We trained Wav2Vec-BERT-2.0, that is a pre-trained model, on Mooré dataset 2 by splitting it into 90% for training and 10% for evaluation. The text of this dataset has been converted to lowercase and symbols, numbers, and other special characters has been removed. The final dataset only contained Mooré alphabet letters and some punctuation. The vocabulary used by the model also contained Mooré alphabet letters and some punctuation. The hyperparameters used for training are described in Table 4.

4 Results and Discussion

4.1 TTS models

We evaluated our TTS models using Mean Opinion Score (MOS), UTMOS and A/B tests. TTS models evaluation on subjective metrics like MOS and A/B test has been made by 36 Mooré native speakers. Each of them evaluated audios generated by our two TTS models, using non-religious texts from news, sport, tales and proverbs. In MOS evaluation, Orpheus-3B and XTTS-v2 have been used to generate Mooré audios using different utterances. For the A/B tests, the same Mooré utterance is used to generate audios using both TTS models. The evaluator listened both audios and selected which is the best one in his opinion. The results of TTS models evaluation are described in Table 5.

Orpheus-3B achieved 3.47 out of 5, which is very good considering it has only been trained on 6.48 hours of data. XTTS-v2 got 4.36 out of 5 with a much better naturalness, tone and pronunciation accuracy according to the evaluators. The UTMOS metric gave 2.80 out of 5 for Orpheus-3B and 3.47 out of 5 for XTTS-v2. That's why XTTS-v2 logically won on A/B tests by being preferred to Orpheus-3B 77.8% of time, when comparing them on the same utterances. These results show that religious data in enough quantity and quality can be used to create high performance TTS models for low-resource languages, even in non-religious domain use cases.

Model	MOS	UTMOS	A/B
Orpheus-3B	3.47	2.80	22.2%
XTTS-v2	4.36	3.47	77.8%

Table 5: TTS models evaluation results

4.2 ASR model

After training Wav2Vec-BERT-2.0 on dataset 2, with a train size of 90%, we evaluated it on the remaining 10% data, using Word Error Rate (WER) and Character Error Rate (CER) metrics. The model achieved a WER score of 4.24% and CER score of 1.11%, which we consider being very impactful results for a low-resource language. The Figure 2 illustrates evolution of WER and CER on test set through fine-tuning.

5 Conclusion

In this paper we presented state of the art ASR and TTS models for Mooré, a low-resource language from Burkina Faso. These models have been trained on religious data and evaluated using non-religious Mooré texts from news, sport, tales and proverbs. XTTS-v2 required vocabulary extension with Mooré tokens, Orpheus-3B has directly been adapted to Mooré, and Wav2Vec-Bert-2.0 required a vocabulary with Mooré alphabet letters. After this out-of-domain evaluation, the results achieved by our Mooré TTS and ASR models are very promising. These models have successfully been joined to other models to create a cascaded S2S French-Mooré Translation system.

As a future pathway to improve the proposed models, synthetic speech data may be generated in order to create huge datasets to improve existing models and implement both direct and cascaded S2ST systems. In addition, we are exploring ways to apply our approach to other low-resource local languages in Burkina Faso. Direct S2ST system will also be experimented to reduce error propagation and potentially obtain a lighter and faster architecture.

Limitations

The cascaded approach used in this work for French-Mooré S2ST, while flexible, may lead to higher error rates due to error propagation between

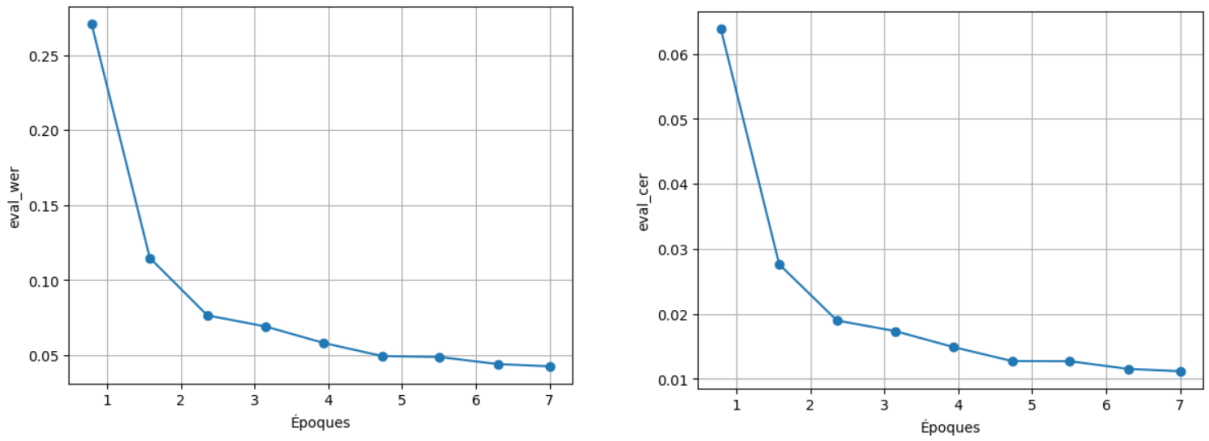

Figure 2: Mooré ASR model evaluation results through training

the different models. In addition, this approach requires important computational power for both training and inference, which is not suitable for low-resource technological environments.

Acknowledgments

This work is supported by Centre Interdisciplinaire d'Excellence en Intelligence Artificielle pour le Développpement (CITADEL). We are very grateful to the volunteers who helped with human evaluation of our TTS models.

References

Abdulhady Abas Abdullah, Shima Tabibian, Hadi Veisi, Aso Mahmudi, and Tarik Rashid. 2024a. End-to-end transformer-based automatic speech recognition for northern kurdish: A pioneering approach.

Abdulhady Abas Abdullah, Hadi Veisi, and Tarik Rashid. 2024b. Breaking walls: Pioneering automatic speech recognition for central kurdish: End-to-end transformer paradigm.

Abdulqahar Mukhtar Abubakar, Deepa Gupta, and Susmitha Vekkot. 2024. Development of a diacritic-aware large vocabulary automatic speech recognition for hausa language. *International Journal of Speech Technology*, 27(3):687–700.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages.

Alexei Baevski, Henry Zhou, Abdel-rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*.

Mary E. Beckman and Jennifer J. Venditti. 2010. *Tone and Intonation*, chapter 16. John Wiley Sons, Ltd.

Wesley Bian, Xiaofeng Lin, and Guang Cheng. 2025. Bridging the language gap: Synthetic voice diversity via latent mixup for equitable speech recognition.

Astik Biswas, Emre Yılmaz, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2022. Code-switched automatic speech recognition in five south african languages. *Computer Speech amp; Language*, 71:101262.

Canopy Labs. 2025a. Orpheus-3b finetuned model. Accessed: 2026-01-04.

Canopy Labs. 2025b. Orpheus-3b pretrained model. Accessed: 2026-01-04.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *Preprint*, arXiv:2406.04904.

Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee. 2019. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. In *Interspeech 2019*, pages 2075–2079.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in african languages.

Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. Leveraging pseudo-labeled data to improve direct speech-to-speech translation.

Qingkai Fang, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. Can we achieve high-quality direct speech-to-speech translation without parallel speech data?

Haohan Guo, Fenglong Xie, Xixin Wu, Hui Lu, and Helen Meng. 2022. Towards high-quality neural tts for low-resource languages by learning compact speech representations. *Preprint*, arXiv:2210.15131.

Mahendra Gupta, Maitreyee Dutta, and Chandresh Kumar Maurya. 2025. Benchmarking hindi-to-english direct speech-to-speech translation with synthetic data. *Language Resources and Evaluation*, 59(3):2613–2651.

Hackster.io. 2025. Canopy labs releases orpheus. Accessed: 2026-01-04.

Rhonda L. Hartell and John T. Bendor-Samuel. 1989. *The Niger-Congo languages: a classification and description of Africa's largest language family*. University Press of America.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahamed, 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.

Praveen Sai Krishna Paleti, Bhavesh Saluru, Sainadh Vatturi, Basaraboyina Yohoshiva, and Nagendra Panini Challa. 2024. Cascaded speech-to-speech translation system focusing on low-resource indic languages. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, page 1–6. IEEE.

Ki-Joong Kwon, Jun-Ho So, and Sang-Hoon Lee. 2025. Parameter-efficient fine-tuning for low-resource text-to-speech via cross-lingual continual learning. pages 1613–1617.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2016. Ethnologue: Languages of the world, mooré. *SIL International*.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. 2022. Textless direct speech-to-speech translation with discrete speech representation.

Hsi-Che Lin, Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2024. Improving speech emotion recognition in under-resourced languages via speech-to-speech translation with bootstrapping data selection.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. In *Interspeech*. ISCA.

Hamid Mojarad and Kevin Tang. 2025. Automatic speech recognition of african american english: Lexical and contextual effects.

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking automatic speech recognition models for african languages.

Uchenna Nzenwata and Daniel Ogbuigwe. 2024. Automatic speech recognition for the ika language.

Peter Ochieng and Dennis Kaburu. 2024. Phonology-guided speech-to-speech translation for african languages.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Maimouna Ouattara, Abdoul Kader Kaboré, Jacques Klein, and Tegawendé F. Bissyandé. 2025. Bridging literacy gaps in African informal business management with low-resource conversational agents. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 193–203, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrei Popescu-Belis, Alexis Allemann, Teo Ferrari, and Gopal Krishnamani. 2025. Speech-to-speech translation pipelines for conversations in low-resource languages.

B P Pradeep Kumar, S Monishaa, Shreehari Menon, and Syeda Aayesha Aiman H. 2025. Real-time speech-to-speech translator: Analysis and implementation. In *2025 4th International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, page 1–6. IEEE.

Alice Tan. 2023. Deep learning in speech recognition for low resource languages. *International Journal for Research Publication and Seminar*, 14(5):530–537.

Harish Thangaraj, Ananya Chenat, Jaskaran Singh Walia, and Vukosi Marivate. 2024. Cross-lingual transfer of multilingual models on low resource african languages.

Haitong Zhang and Yue Lin. 2020. Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. In *Interspeech 2020*, pages 3161–3165.