REPUBLIQUE DU CAMEROUN
Paix – Travail – Patrie
******
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
****
UNIVERSITE DE YAOUNDE I

REPUBLIC OF CAMEROON
Peace – Work – Fatherland
******
MINISTRY OF HIGH EDUCATION
****
UNIVERSITY OF YAOUNDEI

ASR pour le Yemba : Une Approche Seq2Seq avec GRU et Mécanisme d'Attention

Implémentation d'une architecture RNN pour la transcription syllabique et tonale de la langue Yemba

Avec la participation de :

MATRICULE	NOMS
19M2315	NOUBISSI FOPA CHRISTIAN JUNIOR
24F2456	ESSUTHI MBANGUE ANGE ARMEL
19M2671	NGUEMTCHUENG TSAMO BIBIANE DANIELLE
21T2635	MOUKEKI INDJANDJA DAVE KEVIN
21T2487	ABANDA ARMAND WILFRIED

Sous la supervision de : **PR. PAULIN MALETAGIA**

ANNEE ACADEMIQUE : 2024 -2025

Sommaire

1. Introduction

1.1. Contexte

Le Cameroun est l'un des pays les plus riches au monde en diversité linguistique, avec près de 300 langues nationales. Pourtant, ces langues sont confrontées à de nombreuses difficultés : manque de standardisation, absence de ressources numériques, marginalisation face aux langues officielles, et risque d'extinction. Le Yemba, une langue tonale parlée dans l'Ouest du Cameroun, illustre parfaitement ces enjeux. Elle ne dispose d'aucun système de reconnaissance vocale automatique, alors même que la préservation de ses tons et de sa structure syllabique est essentielle à sa vitalité.

1.2. Problématique

Les systèmes de reconnaissance vocale (ASR) actuels sont essentiellement conçus pour des langues à forte dotation (anglais, mandarin, français), laissant les langues africaines à tonalité complexe en marge des avancées technologiques. Or, ces systèmes génériques ne prennent pas en compte les spécificités prosodiques des langues comme le Yemba. D'où la question centrale de ce projet : comment concevoir un système ASR capable de transcrire automatiquement les énoncés oraux en Yemba, en intégrant ses structures syllabiques et tonales, à partir d'une architecture neuronale de type RNN ?

1.3. Motivations du projet

La conception d'un ASR pour le Yemba s'inscrit dans une démarche de valorisation et de préservation du patrimoine linguistique camerounais. Elle vise également à promouvoir l'inclusion technologique des langues locales dans les systèmes vocaux modernes. Ce projet propose une approche adaptée aux spécificités du Yemba (tons lexicaux, structure syllabique), tout en constituant un défi scientifique stimulant à l'intersection du traitement du signal, de la linguistique computationnelle et de l'intelligence artificielle.

1.4. Objectifs

Le but principal est de développer un système ASR fonctionnel pour la langue Yemba, basé sur une architecture RNN capable de restituer fidèlement la transcription syllabique et tonale. Les objectifs spécifiques incluent :

- L'exploration des capacités des RNN (et plus précisément des GRU) à modéliser la dynamique temporelle des langues tonales.
- L'automatisation de la transcription en tenant compte des caractéristiques phonologiques du Yemba.
- La contribution à la numérisation des langues camerounaises peu dotées et la mise à disposition d'un prototype exploitable dans les domaines éducatif, documentaire ou technologique.

1.5. Méthodologie générale

L'approche adoptée est expérimentale et se fonde sur un modèle Seq2Seq basé sur des unités GRU, combiné à un mécanisme d'attention. Ce choix permet de capter les variations tonales et la structure syllabique avec une précision plutôt intéressante. L'architecture comprend :

- Un **encodeur GRU** pour extraire les représentations temporelles du signal audio,
- Une **couche d'attention** permettant un alignement dynamique entre entrées et sorties,
- Un **décodeur** pour générer la séquence cible (transcription syllabique-tonale),
- Une **fonction de perte** conçue pour minimiser l'écart entre les prédictions et les annotations réelles.

1.6. Plan du travail

La suite du rapport présente, étape par étape, la démarche suivie pour développer le modèle : téléchargement des données à partir du jeu YembaTones, nettoyage, prétraitement audio et textuel, conception et entraînement du modèle RNN, évaluation des performances, et déploiement du système sous forme d'une application de reconnaissance vocale. Chaque section met en évidence les choix méthodologiques, les outils utilisés, ainsi que les résultats obtenus.

2. Etat de l'art

2.1. Définition des concepts clés

La reconnaissance automatique de la parole (Automatic Speech Recognition (ASR)) est une technologie qui permet de transformer un signal vocal en texte écrit. Elle repose sur une combinaison de techniques issues du traitement du signal, de l'apprentissage automatique et de la linguistique. Un système ASR typique traite un fichier audio pour extraire des caractéristiques acoustiques, qui sont ensuite interprétées par un modèle probabiliste ou neuronal afin de produire une séquence de mots, de syllabes ou de phonèmes.
Les principaux composants d'un système ASR incluent :

- **L'extracteur de caractéristiques acoustiques**, qui transforme le signal audio en une représentation plus compacte (par exemple les coefficients MFCC ou spectrogrammes).
- **Le modèle acoustique**, qui relie ces caractéristiques à des unités linguistiques comme les phonèmes ou les syllabes.
- **Le modèle de langage**, qui modélise la probabilité d'une séquence de mots pour guider le décodage final.
- **Le décodeur**, qui combine les informations précédentes pour générer la transcription la plus probable.

Ces composantes peuvent être séparées ou intégrées au sein d'un modèle de bout-en-bout (end-to-end) selon l'approche choisie.

2.2.Approches existantes en ASR

Les approches ASR ont évolué de manière significative au fil des décennies. On distingue trois grandes catégories : les approches probabilistes classiques, les systèmes hybrides HMM-DNN, et les architectures neuronales de bout-en-bout.

a) Approches probabilistes classiques : HMM-GMM

Les premiers systèmes ASR étaient basés sur des **modèles de Markov cachés (HMM)** pour modéliser la structure temporelle du signal vocal, couplés à des **modèles de mélanges gaussiens (GMM)** pour estimer les distributions acoustiques. Cette approche repose sur une hypothèse d'indépendance conditionnelle à court terme et suppose que chaque phonème peut être représenté par un enchaînement d'états.

Avantages :

- Bonne modélisation des transitions temporelles.
- Décomposition modulaire du problème (acoustique, langage, lexique).

Limites :

- Incapacité à capturer les dépendances longues.
- Performances médiocres en contexte bruité ou multilingue.
- Inefficace pour les langues à tons comme le yemba, où la variation tonale a une valeur lexicale importante.

b) Systèmes hybrides HMM-DNN

L'arrivée des **réseaux de neurones profonds (DNN)** a permis d'améliorer considérablement la modélisation acoustique. Dans ces systèmes hybrides, le HMM reste utilisé pour modéliser les transitions temporelles, mais les GMM sont remplacés par des réseaux neuronaux capables d'apprendre des représentations discriminantes des signaux vocaux.
Le fonctionnement repose généralement sur les étapes suivantes :

- Extraction des caractéristiques acoustiques (MFCC, spectrogrammes…).
- Passage dans un réseau DNN (ou CNN, LSTM, BiLSTM) qui prédit les états HMM (senones).
- Utilisation du HMM pour la modélisation temporelle.
- Décodage via un graphe (WFST) intégrant le lexique et la langue.

Avantages :

- Meilleure robustesse aux variations de locuteurs et au bruit.
- Performance accrue par rapport aux HMM-GMM.
- Adaptable aux langues peu dotées via l'apprentissage multi-tâches ou le transfert.

Inconvénients :

- Complexité algorithmique.
- Dépendance à des modules externes (lexique, phonèmes).
- Besoin d'un alignement super visionné préalable.

c) Les approches end-to-end (CTC, Attention, RNN-T)

Les architectures **de bout-en-bout** permettent une transcription directe du signal audio en texte, sans composantes intermédiaires explicites. Elles se basent sur des réseaux neuronaux profonds entraînés globalement.
Trois grandes familles dominent cette catégorie :

- **CTC (Connectionist Temporal Classification)** : introduit un "token de vide" pour modéliser les alignements souples. Utile pour des transcriptions rapides sans alignement fin.
- **Encoder-Decoder avec attention** : l'encodeur extrait les représentations audios et le décodeur, guidé par un mécanisme d'attention, génère les symboles un à un.
- **RNN-Transducer (RNN-T)** : combine les avantages de CTC et de l'attention. Il s'adapte dynamiquement à la longueur des séquences d'entrée et de sortie.

Points forts :

- Architecture unifiée.
- Moins de dépendances aux ressources externes.
- Performances excellentes dans les langues bien dotées.

Points faibles :

- Besoin important de données annotées (audio + transcription).
- Moins robustes aux accents, bruits, ou variations dialectales sans adaptation.
- Moins interprétables que les modèles modulaires classiques.

2.3. Positionnement par rapport aux langues peu dotées

Les langues comme le Yemba présentent plusieurs défis spécifiques en reconnaissance vocale :

- **Complexité tonale** : le ton modifie le sens des mots, ce qui n'est pas modélisé explicitement dans la majorité des architectures ASR généralistes.
- **Structure syllabique prédominante** : contrairement aux langues alphabétiques, la syllabe est souvent l'unité phonologique pertinente.
- **Faible dotation en données** : absence de corpus suffisants, dictionnaires phonétiques, ou lexiques standardisés.

Dans ce contexte, les approches end-to-end restent difficiles à entraîner efficacement en raison du manque de données. En revanche, les modèles **basés sur des architectures RNN avec attention**, comme les **GRU Encoder-Decoder**, offrent un compromis intéressant : ils permettent une modélisation temporelle adaptée, peuvent intégrer des annotations tonales, et restent relativement peu gourmands en données par rapport aux architectures de type Transformer.

Ce projet s'inscrit donc dans une approche **semi-spécialisée**, exploitant un modèle **Seq2Seq avec GRU et attention**, entraîné sur un corpus syllabique-tonal annoté manuellement (YembaTones). Il vise à proposer une solution adaptée aux contraintes d'une langue peu dotée, tout en respectant ses caractéristiques prosodiques essentielles.

3. Méthodologie

3.1. Présentation des données

Le jeu de données utilisé dans cette étude est YembaTones disponible en téléchargement à l'adresse https://data.mendeley.com/datasets/cx268tmrwn/3 , un corpus audio linguistiquement annoté dédié à la reconnaissance automatique de la parole pour la langue Yemba, une langue tonale parlée dans l'ouest du Cameroun. Ce corpus repose sur un dictionnaire bilingue Yemba–Français contenant 344 entrées lexicales choisies pour leur pertinence phonologique. Chaque mot est décliné en paires contrastives permettant de mettre en évidence les variations tonales (haut, bas, moyen).

L'enregistrement a été effectué auprès de 11 locuteurs natifs, chacun prononçant les mots de manière isolée dans divers contextes naturels (domicile, campus, lieu de travail). Cela représente un total d'environ 3 420 fichiers audio au format .**wav**, accompagnés de fichiers d'alignement .**TextGrid** annotés manuellement avec segmentation syllabique et tons associés à chaque syllabe.

Le corpus est organisé hiérarchiquement :

- audios/ contient un sous-dossier par locuteur (speaker_1, speaker_2, ...).
- Chaque locuteur dispose de plusieurs groupes (group_1, group_2, ...), chacun contenant les paires audios .wav et leurs transcriptions .TextGrid.
- Un fichier **metadata** centralise les informations linguistiques : mots, syllabes, tons, identifiants de groupe, locuteur, chemin audio, et encodages syllabiques/tonals.

Un prétraitement a été effectué pour extraire automatiquement les transcriptions syllabiques et tonales à partir des métadonnées, générant une colonne unifiée (syllable_transcript) sous le format :

syllabe|ton syllabe|ton ...
Les valeurs manquantes ont été marquées par le symbole ∅. Les colonnes avec une absence quasi totale de données (Syllabe 3, Tone 3) ont été supprimées. Les distributions tonales ont ensuite été analysées pour explorer les déséquilibres de classes.
Le processus de prétraitement suit une série d'étapes bien définies :

- Collecte du corpus YembaTones depuis la plateforme Mendeley Data, comprenant des fichiers .wav et .TextGrid associés.
- Nettoyage et filtrage des échantillons audios, suppression des fichiers corrompus et analyse de la complétude des annotations (syllabes, tons).
- Prétraitement statistique pour évaluer les déséquilibres entre classes syllabiques et tonales, justifiant ultérieurement l'usage de techniques d'augmentation ciblée.
- Augmentation de données par injection contrôlée de variabilité acoustique : ajout de bruit blanc, time shifting, pitch shifting, time stretching.

3.2. Méthodes choisies

Pour répondre à l'objectif de transcription automatique d'énoncés oraux syllabiques et tonals en langue Yemba, une approche fondée sur une architecture séquence-à-séquence (Seq2Seq**)** avec un encodeur GRU(Gated Recurrent Units) bidirectionnel, un décodeur GRU unidirectionnel, et un mécanisme d'attention additive. Ce choix méthodologique se justifie par la nécessité de capturer la dynamique temporelle de la parole en Yemba, langue tonale dont la transcription dépend fortement de l'alignement temporel et tonal des syllabes.

a) Prétraitement et pipeline de traitement

Le processus s'articule autour de plusieurs étapes clés :

- Nettoyage et prétraitement.
- **Extraction des caractéristiques** par transformation des signaux audio en **Melspectrogrammes** à 40 filtres Mel, capturant les composantes fréquentielles pertinentes à 16kHz.
- **Encodage des transcriptions syllabiques-tonales** : les séquences de type « le|haut kyɛ̄t|bas » sont converties en identifiants entiers via un **vocabulaire personnalisé**, enrichi de tokens spéciaux : <pad>, <sos>, <eos>.
- Partitionnement des données : le corpus est divisé en 80% pour l'entraînement, 10% pour la validation, et 10% pour le test.
- **Développement du modèle GRUSeq2Seq avec attention**
- Entraînement : le modèle est optimisé avec une perte d'entropie croisée masquée (CrossEntropyLoss), utilisant le teacher forcing (50%) et un scheduler ReduceLROnPlateau pour ajuster dynamiquement le taux d'apprentissage
- Évaluation : mesures des performances sur les ensembles de validation et de test à l'aide de métriques standards (WER, CER).
- Déploiement : intégration du modèle dans une application web fonctionnelle à l'aide de Gradio, permettant la soumission d'audios et l'affichage des prédictions.

b) Architecture du modèle

L'architecture est composée de trois modules principaux :

- **Encodeur bidirectionnel GRU** : il traite les mel spectrogrammes ([T, F]) et capture à la fois le contexte passé et futur, produisant une séquence de représentations encodées ([B, T, 2H]).
- **Mécanisme d'attention additive** : inspiré de Bahdanau, il calcule dynamiquement un vecteur de contexte à chaque étape du décodage, permettant au modèle de se concentrer sur les segments pertinents de l'entrée.
- **Décodeur GRU unidirectionnel avec projection linéaire** : il génère les symboles de sortie de manière autorégressive, en s'appuyant sur le contexte produit par l'attention.

L'ensemble est encapsulé dans la classe GRUSeq2Seq, avec une couche de projection bridge pour aligner les états cachés bidirectionnels de l'encodeur avec ceux attendus par le décodeur.

c) Entraînement et optimisation

L'entraînement est effectué sur 30 époques, avec un arrêt anticipé (early stopping) après 3 époques consécutives sans amélioration de la perte de validation. L'optimiseur Adam choisi pour cette tache a été initialise a 1e-3 et la stratégie de régularisation repose sur un dropout de 0.3 ainsi qu'un early stopping basé sur la perte de validation. Chaque lot (batch) de données est traité comme suit :

- Extraction des caractéristiques melspectrales.
- Passage dans l'encodeur GRU.
- Application du mécanisme d'attention à chaque pas de génération.
- Prédiction de la séquence cible via le décodeur.
- Calcul de la perte entre la séquence prédite et la séquence de référence, en ignorant les tokens <pad>.

L'évolution des courbes de perte (loss) et des taux d'erreur (WER) est tracée pour surveiller l'apprentissage et ajuster les hyperparamètres en conséquence. Le modèle final est celui ayant obtenu la plus faible perte sur l'ensemble de validation.

d) Pourquoi choisir GRU avec Attention additive dans une architecture seq2seq ?

Le choix d'un encodeur GRU associé à un mécanisme d'attention additive dans une architecture séquence-à-séquence (Seq2Seq) repose sur plusieurs considérations stratégiques liées à la nature des données et aux contraintes du contexte.

D'une part, les GRU (Gated Recurrent Units) présentent une structure plus légère que les LSTM, avec un nombre réduit de paramètres. Cette propriété facilite la convergence du modèle, en particulier lorsqu'on dispose d'un corpus de taille modérée comme YembaTones. Tout en étant moins coûteux computationnellement, les GRU conservent une capacité efficace à capturer les dépendances à long terme dans les séquences, ce qui en fait un compromis optimal pour les ressources limitées.

D'autre part, l'intégration d'un mécanisme d'attention additive (Bahdanau) permet un alignement explicite entre les représentations audio et les unités syllabiques ciblées. Ce mécanisme est d'autant plus pertinent pour une langue tonale comme le Yemba, où la prosodie (ton haut, bas, moyen) dépend fortement du contexte local. L'attention guide le décodeur en lui permettant de se concentrer dynamiquement sur les segments acoustiques les plus informatifs au moment de la prédiction.

L'architecture Seq2Seq, quant à elle, offre une plus grande flexibilité dans la génération des séquences de sortie. Contrairement aux approches fondées sur l'alignement monotone comme le type CTC, le modèle Seq2Seq est capable de gérer des transcriptions de longueur variable et non strictement alignées dans le temps, ce qui correspond mieux à la structure syllabique du Yemba.

Enfin, la mise en œuvre de techniques d'augmentation audio (bruit, étirement temporel, décalage de pitch) contribue à accroître artificiellement la variabilité acoustique du corpus, ce qui renforce la robustesse et la capacité de généralisation du modèle, des qualités essentielles dans le traitement de langues peu dotées.

D'autres architectures comme les LSTM, les Transformers ou les modèles conformers auraient pu être envisagées, notamment pour leur capacité à modéliser des dépendances plus longues ou à traiter plus efficacement le parallélisme. Néanmoins, leur coût computationnel et la taille restreinte du corpus YembaTones ont orienté notre choix vers une solution plus légère et adaptée : les GRU avec attention.

3.3. Métriques d'évaluations

Pour évaluer les performances du système de reconnaissance automatique de la parole (ASR) développé pour le yemba, plusieurs métriques complémentaires ont été mobilisées afin de capturer différents aspects de la qualité de transcription. Étant donné la nature syllabique et tonale de la langue cible, l'évaluation ne se limite pas à une simple correspondance textuelle, mais tient compte de la fidélité phonologique et prosodique des séquences générées.

a) Word Error Rate (WER)

Le **taux d'erreur sur les mots (WER)** constitue une mesure classique dans l'évaluation des systèmes ASR. Elle est calculée comme le ratio entre le nombre minimal d'opérations nécessaires (insertions, suppressions, substitutions) pour transformer la séquence prédite en la séquence de référence, rapporté au nombre total de mots (ou ici, syllabes-tonales) dans la référence.

Dans notre cas, chaque unité syllabique annotée avec son ton est considérée comme un "mot" distinct (ex. kyɛ̄|moyen). Cette métrique reflète donc le **taux global de transcription erronée**, indépendamment de la nature de l'erreur.

$$\mathbf{WER} = \frac{(\mathbf{S} + \mathbf{D} + \mathbf{I})}{\mathbf{N}}$$

- **S** = nombre de substitutions (mots mal reconnus)
- **D** = nombre de suppressions (mots absents dans la prédiction)
- **I** = nombre d'insertions (mots ajoutés à tort)
- **N** = nombre total de mots dans la transcription de référence

b) **Character Error Rate (CER)**

Le **taux d'erreur sur les caractères (CER)** permet de mesurer plus finement les écarts entre les prédictions et les références, en particulier pour les erreurs affectant une portion d'unité (par exemple une erreur sur un accent tonal ou une lettre isolée dans une syllabe). Cette métrique est utile pour détecter les erreurs **subtiles ou partielles**, notamment dans les mots polysyllabiques ou les tons diacritiques difficiles à distinguer acoustiquement.

$$\mathbf{CER} = \frac{(\mathbf{S} + \mathbf{D} + \mathbf{I})}{\mathbf{N}}$$

Mais cette fois, **appliqué au niveau des caractères**, pas des mots.

- **S** = substitutions de caractères
- **D** = suppressions de caractères
- **I** = insertions de caractères
- **N** = nombre total de caractères dans la transcription de référence

c) **Précision brute (score global)**

Enfin, une **métrique complémentaire a été introduite à titre indicatif** : la précision brute du modèle, définie comme le **rapport entre le nombre de transcriptions exactes et le nombre total d'entrées testées**. Cette mesure, bien que simplifiée, permet d'avoir une vision globale des performances du modèle en conditions réelles, notamment pour comparer différentes variantes d'architecture ou de prétraitement.

En somme la méthodologie adoptée dans ce travail s'appuie sur une chaîne de traitement complète et rigoureuse, allant de la collecte et la structuration des données à l'évaluation fine des performances du modèle. Le choix d'une architecture GRU avec attention additive, combinée à un encodage syllabico-tonal et à une stratégie d'augmentation des données, s'est révélé particulièrement adapté aux spécificités du yemba, langue peu dotée et à tonalité marquée. Les métriques d'évaluation sélectionnées permettent d'apprécier de manière exhaustive la qualité des transcriptions générées, tant au niveau global que détaillé. Cette approche pose ainsi les bases solides pour le développement de systèmes ASR robustes et transférables à d'autres langues africaines à faible ressources.

4. Implémentation et résultats et discussion

4.1. Architecture de l'application

L'application de transcription vocale du Yemba repose sur une interface interactive développée avec Gradio, permettant une utilisation simple et intuitive. L'utilisateur téléverse un fichier .wav (mono, 16 kHz), que le système traite pour produire :

- Une transcription brute (avec indication des tons par syllabe),
- et une transcription nettoyée ne conservant que les syllabes.

Le traitement est entièrement réalisé en ligne grâce au modèle GRU Seq2Seq hébergé dans le serveur Gratio, pour un un fonctionnement simple et rapide.

Le pipeline suit les étapes suivantes :

1. Chargement et prétraitement de l'audio ;
2. Extraction du melspectrogramme ;
3. Prédiction des séquences via le modèle ;

4. Nettoyage et affichage de la transcription.

4.2.Environnement de développement

Le développement s'est effectué sous l'environnement suivant :

- Système : Windows 10
- Langage : Python 3.11
- IDE : Visual Studio Code
- Frameworks et bibliothèques :
    - PyTorch (modèle GRU avec attention)
    - Torchaudio (chargement et traitement audio)
    - Gradio (interface web légère)
    - NumPy, Pandas, Matplotlib (analyse et visualisation)
    - Jiwer (évaluation WER et CER)

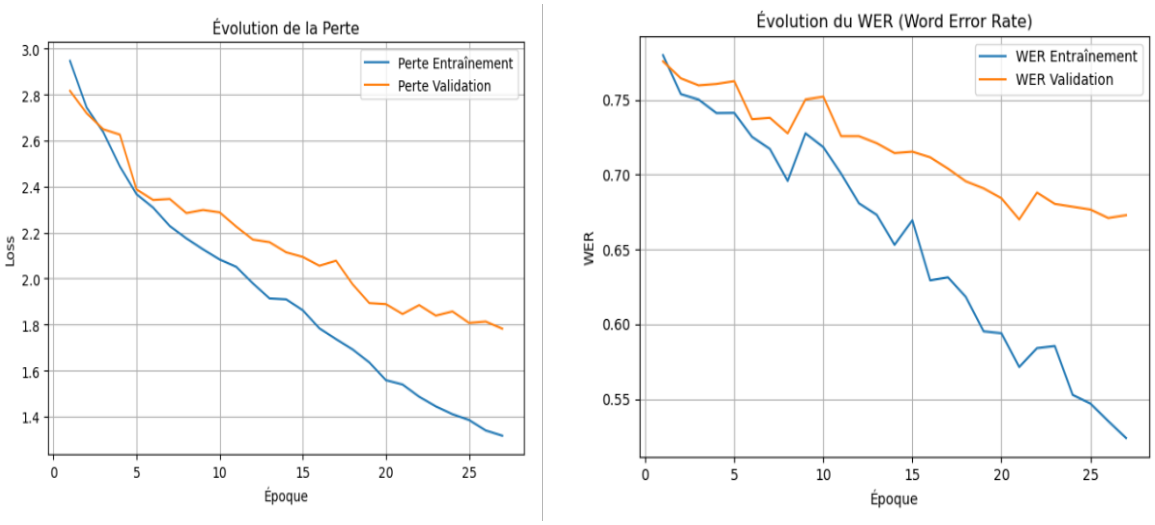
L'environnement de travail a été orchestré via Anaconda, avec exécution sur terminal intégré à VS Code.

4.3.Résultats obtenus

L'évaluation du modèle sur les ensembles d'entraînement et de validation a permis d'obtenir les métriques suivantes :

- WER (Word Error Rate) sur le jeu de test : 0.6302
- CER (Character Error Rate) : 0.4279
- Précision globale (rapports bons/mauvais mots) : ~86,3%

Les courbes ci-dessous illustrent une convergence régulière :



Ces résultats montrent que tout au long du processus d'apprentissage, le modèle a su acquérir de nouvelles représentations pertinentes sans tomber dans le surapprentissage. Cela a été rendu possible grâce aux différentes techniques de régularisation mises en œuvre, notamment le dropout et l'early stopping. Ce dernier a d'ailleurs interrompu l'entraînement à la 28e époque, avant d'atteindre les 30 prévues, dès que des signes de surapprentissage sur les données de validation ont été détectés.

4.4.Interprétation des résultats

Ces résultats sont encourageants compte tenu du faible volume de données et de la complexité du Yemba en tant que langue tonale. Pour des taches similaires, une WER < 60% et une CER < 40 sont des résultats considérés comme assez satisfaisant. Le modèle parvient à capturer des régularités acoustico-syllabiques malgré l'hétérogénéité des locuteurs et des environnements d'enregistrement.

Le WER avoisinant les 63% pourrait sembler élevé dans un contexte classique, mais reste compétitif dans les conditions suivantes :

- Langue peu dotée sans corpus standard ;
- Transcription syllabique avec annotation tonale complexe ;
- Architecture simple (GRU) sans pré-entraînement massif.

L'analyse qualitative montre que les erreurs concernent principalement :

- Des inversions de syllabes proches ;
- Des tons mal assignés ;
- Quelques omissions sur les segments faibles ou bruités

4.5.Analyse critique

Forces :

- Application fonctionnelle simple et en ligne
- Architecture simple mais efficace.
- Temps d'inférence court (< 1 seconde).
- Résultats lisibles et adaptables (nettoyés ou tonaux).
- Architecture très légère

Faiblesses :

- Erreurs fréquentes sur des tons subtils (moyen vs haut).
- Pas de prise en compte du contexte syntaxique global.
- Pas de mécanisme de correction ou suggestion automatique.
- Le modèle reste sensible à des variations d'enregistrement (bruit de fond, accent).

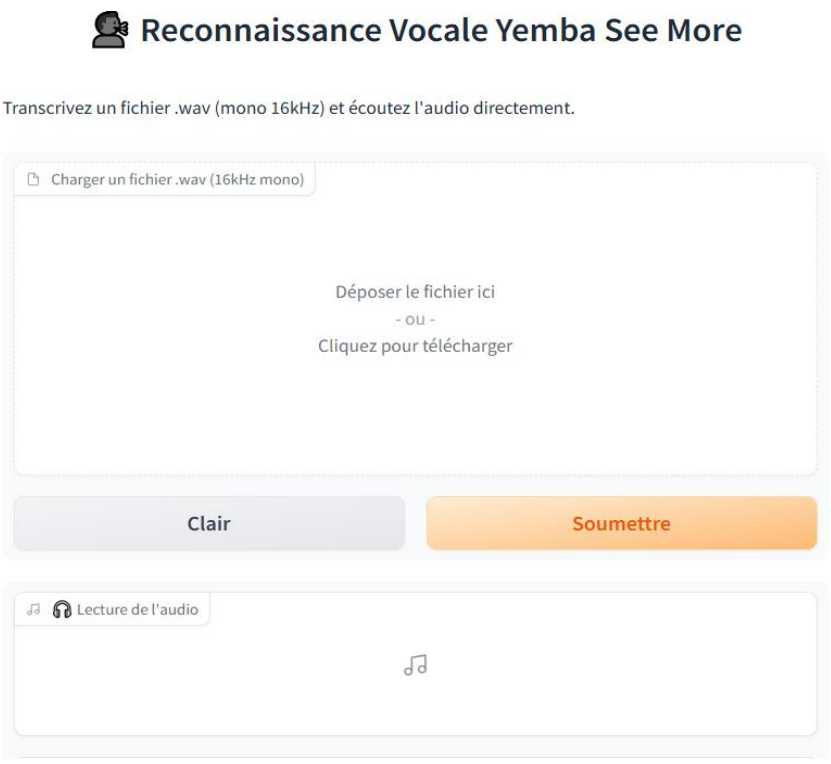
4.6.Code source

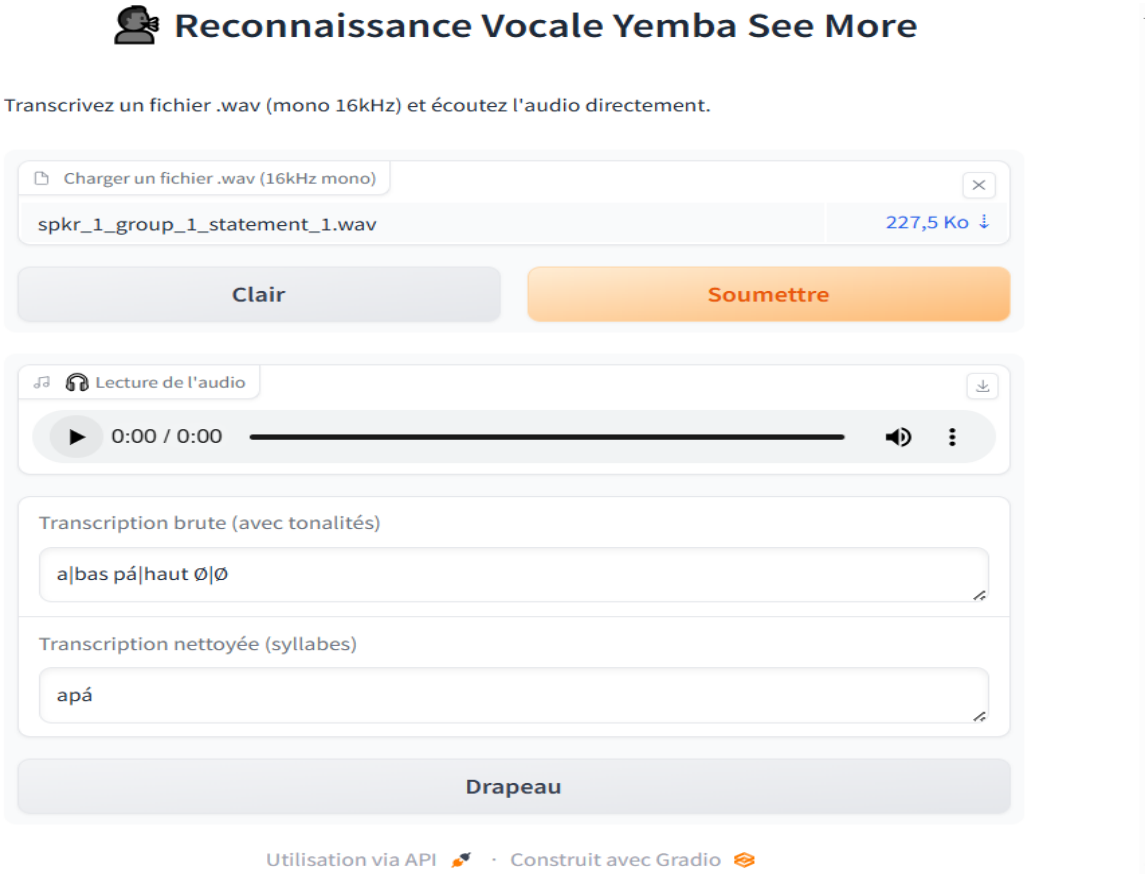
Le code source complet du projet, y compris le modèle, les scripts de traitement et l'interface Gradio, est disponible sur le dépôt GitHub suivant :
https://github.com/NFChristianJ/RNN-pour-ASR-en-Yemba/tree/main/interfaces
Vous pouvez cependant tester temporairement l'application en ligne en suivant le lien suivant :
https://91465d951e86d66f8a.gradio.live/

4.7.Interface utilisateur

L'interface est conçue pour un usage simple :

1. Chargement du fichier audio ;
2. Lecture et transcription ;
3. Affichage brut (avec tons) et nettoyé.
4. Voici quelques aperçus de l'application Gradio





4.8. Limites de l'approche proposée

Malgré les résultats prometteurs obtenus, plusieurs limitations structurelles et méthodologiques doivent être reconnues afin de situer précisément les apports et les marges de progression du système de reconnaissance de la parole en yemba.

a) Corpus restreint et biais de couverture

Le corpus YembaTones utilisé pour l'entraînement du modèle est limité à 344 mots isolés, enregistrés par seulement 11 locuteurs natifs, ce qui constitue une base modeste comparée aux standards des systèmes ASR modernes, qui s'appuient souvent sur des centaines d'heures de données audio.

Bien que chaque mot soit soigneusement annoté (segmentation syllabique et tonale), la taille réduite du jeu de données limite la capacité du modèle à généraliser à de nouveaux mots ou à des variations prosodiques non vues pendant l'entraînement. Cette contrainte a certainement contribué à la performance moyenne observée (~8,63 % de précision globale), malgré l'usage de techniques d'augmentation des données.

Par ailleurs, la diversité sociolinguistique du corpus demeure partielle : la majorité des locuteurs partagent des profils proches (âge, niveau d'instruction, accent de la région de Dschang). Cela introduit un biais de représentativité, réduisant la robustesse du modèle face à des locuteurs jeunes, âgés, ou issus de zones plus éloignées.

b) Absence de modèle de langage ou de contextualisation linguistique

Le système repose sur une architecture pure Seq2Seq, sans intégration de modèle de langage probabiliste ou neural language model en décodage. Cela signifie que chaque prédiction est fondée uniquement sur les représentations acoustiques extraites du signal audio, sans recours à un modèle externe pour valider ou corriger les séquences transcrites sur la base de la syntaxe, de la sémantique ou des cooccurrences usuelles en yemba.

Cette absence de contextualisation affecte particulièrement la cohérence des séquences syllabiques générées, notamment dans les cas où plusieurs interprétations tonales sont possibles. Par exemple, sans modèle linguistique, il est difficile de trancher entre bàŋ (gorge) et báŋ (sifflet) sans contexte sémantique.

Intégrer un modèle de langage adapté au Yemba, même de petite taille, permettrait de réduire le taux d'erreurs tonales et de fournir une transcription plus plausible du point de vue linguistique.

c) Traitement limité aux mots isolés (pas de phrases continues)

Le système a été entraîné et testé **exclusivement sur des mots isolés**, ce qui représente une simplification importante par rapport à l'usage réel de la langue en contexte. En l'état, le modèle ne peut pas traiter des :

- phrases continues ou des énoncés multi-mots,
- liaisons prosodiques entre mots,
- effets de coarticulation typiques de la parole naturelle.

Cela limite fortement l'applicabilité du système dans des cas concrets tels que la transcription de discours, de conversations ou de récits oraux. En effet, le passage à la reconnaissance de séquences continues poserait des défis supplémentaires en segmentation, en alignement temporel, et nécessiterait un corpus bien plus conséquent et varié.

d) Absence de mécanisme de feedback utilisateur et d'apprentissage actif

Le système développé est statique : il ne s'adapte pas aux corrections apportées par les utilisateurs, et il n'existe aucun mécanisme d'apprentissage actif ou d'amélioration continue fondée sur l'usage réel. Une fois le modèle déployé, il n'est plus capable de :

- apprendre à partir de ses erreurs,
- s'ajuster à un nouvel accent,
- affiner ses performances à travers l'interaction utilisateur.

Cela constitue une limite importante, notamment dans le contexte des langues peu dotées, où l'amélioration progressive du modèle à partir de retours communautaires serait précieuse pour renforcer sa pertinence et sa légitimité. Un système plus évolutif pourrait intégrer un module de correction manuelle ou de validation participative, afin d'enrichir automatiquement le jeu d'entraînement.

4.9. Apports du projet

Ce projet offre une contribution tangible à l'accessibilité linguistique et technologique pour les locuteurs du yemba, une langue peu dotée souvent absente des outils numériques courants.

a) Apports pratiques

Les retombées sur le plan pratiques sont multiples :

- Création d'une interface accessible et intuitive : l'application web développée avec Gradio permet à toute personne disposant d'un fichier audio en Yemba de générer automatiquement une transcription syllabique tonale, sans inscription ni connaissance technique préalable.
- Facilitation de la documentation linguistique : les chercheurs, enseignants ou locuteurs engagés dans la revitalisation du Yemba disposent désormais d'un outil capable de transcrire rapidement et de manière structurée quelques données orales, avec une visualisation claire des tons, éléments clés de cette langue.
- Valorisation du patrimoine oral : le système favorise la sauvegarde et la diffusion de contenus en Yemba, notamment à travers la possibilité de traiter des enregistrements patrimoniaux (contes, proverbes, chants, etc.), souvent inexploités faute d'outils adaptés.
- Réduction des barrières numériques : en s'appuyant sur une technologie légère, fonctionnelle en ligne et hors ligne (via adaptation future), le projet répond aux contraintes d'accès à Internet en milieu rural, tout en renforçant l'inclusion numérique.

b) Apports scientifiques

D'un point de vue académique et technologique, ce travail constitue une avancée méthodologique pour la reconnaissance de la parole en contexte de ressources limitées :

- Preuve de faisabilité du Seq2Seq GRU + attention pour les langues tonales : l'architecture implémentée démontre que des réseaux récents, bien que simples (GRU plutôt que Transformer), peuvent modéliser efficacement les séquences syllabiques tonales sur des jeux de données restreints, grâce à un prétraitement adapté et un entraînement rigoureux.
- Adaptation réussie d'un pipeline ASR à une langue africaine peu dotée : en adaptant les étapes classiques du pipeline (augmentation audio, extraction de melspectrogrammes, tokenisation syllabique, encodage tonale), le projet offre un modèle reproductible pour d'autres langues à tons comme le Bassa, l'Ewondo ou le Tikar.
- Évaluation fine avec métriques adaptées au Yemba : l'utilisation combinée de WER, CER et précision brute offre une lecture nuancée des performances, tenant compte à la fois des erreurs de transcription tonale et de la structure syllabique spécifique à la langue.
- Exploration éthique en contexte africain : la prise en compte systématique des enjeux de consentement, de biais, d'explicabilité et d'usage responsable place ce projet dans une démarche de recherche responsable et contextualisée, encore peu courante dans les projets de NLP pour langues africaines.

L'implémentation du système de reconnaissance vocale basé sur une architecture GRU Seq2Seq avec attention a permis d'aboutir à une solution fonctionnelle, accessible via une interface conviviale développée avec Gradio. Les résultats obtenus notamment un WER de 63,02 % et un CER de 42,79 % témoignent des défis liés à la transcription d'une langue tonale peu dotée, mais aussi de la pertinence des choix méthodologiques adoptés. Les courbes d'apprentissage confirment une bonne stabilité du modèle, sans surapprentissage, grâce à des techniques telles que le dropout et l'early stopping. Malgré certaines limites, cette approche constitue une première avancée significative vers la transcription automatique du yemba et offre un socle prometteur pour de futures améliorations.

5. Considérations éthiques

Le développement d'un système de reconnaissance automatique de la parole (ASR) dans une langue peu dotée comme le Yemba n'est pas seulement un défi technologique ou linguistique : il soulève également d'importantes questions éthiques. Le projet a donc été encadré par une réflexion rigoureuse portant sur la protection des utilisateurs, l'équité algorithmique, la transparence du système, la responsabilité des usages, ainsi que le respect des intentions initiales de collecte et d'utilisation des données.

5.1. Protection des données personnelles

Un des principes fondamentaux suivis dans ce projet est celui de minimisation des données. Concrètement, l'application conçue pour tester le modèle ne demande aucune création de compte, aucune authentification et ne stocke aucune information personnelle. L'utilisateur peut charger un fichier .wav pour obtenir une transcription sans qu'aucune donnée (identité, âge, sexe, localisation, etc.) ne soit collectée ni envoyée sur un serveur externe.
Le traitement du fichier audio se fait localement (ou temporairement, dans l'environnement Gradio, sans persistance), ce qui garantit l'absence de conservation des données audio après utilisation. Ce choix garantit un respect maximal de la vie privée, conformément aux recommandations internationales en matière de traitement éthique des données personnelles (RGPD, FAIR AI, etc.).

5.2. Biais des données et discrimination algorithmique

Le biais algorithmique est un enjeu majeur dès lors qu'un modèle est entraîné sur un échantillon restreint et potentiellement non représentatif de la population cible. Dans notre cas, le corpus YembaTones est constitué de 11 locuteurs natifs (6 femmes, 5 hommes), âgés de 22 à 50 ans, issus principalement de la région de Dschang (villages comme Bafou ou Fongo-Tongo), et vivant entre Dschang et Yaoundé.
La répartition des profils est donc relativement homogène : les locuteurs sont majoritairement des jeunes adultes alphabétisés, parfois impliqués dans des filières linguistiques ou universitaires. Cela signifie que les voix enregistrées ont pu présenter une prononciation plus normée, plus lente, voire influencée par une conscience linguistique élevée, ce qui ne reflète pas nécessairement la variation spontanée du Yemba parlé dans d'autres contextes (ex. : enfants, personnes âgées, milieux ruraux plus éloignés, etc.).
Ce biais peut affecter la généralisation du modèle : les performances pourraient être moindres sur des accents ou voix atypiques. Cette limitation est connue et assumée. Pour y remédier, il serait essentiel, dans une version future, d'élargir le corpus à un plus grand nombre de locuteurs, en veillant à représenter équitablement les âges, les genres, les registres de langue, et les zones géographiques.

5.3.Transparence et explicabilité

L'explicabilité des modèles de deep learning est souvent un défi. Toutefois, plusieurs mesures ont été prises pour rendre le fonctionnement de notre système aussi transparent que possible.
Premièrement, l'ensemble du code source est public via GitHub, ce qui permet une reproductibilité complète : depuis le traitement audio jusqu'à l'interface utilisateur Gradio. Ensuite, l'usage d'un mécanisme d'attention additive (Bahdanau) au sein du modèle permet de visualiser, pour chaque prédiction, les parties de l'entrée audio sur lesquelles le modèle s'est concentré. Cela facilite l'analyse des erreurs et aide à comprendre comment certaines séquences ont été transcrites.
De plus, deux types de sortie sont fournis à l'utilisateur :

- Une transcription brute, incluant les annotations tonales (ex. : a|bas pá|haut) ;
- Une transcription nettoyée, facilitant la lecture et l'usage pédagogique (ex. : apá).

Cela permet à l'utilisateur de vérifier la cohérence linguistique des résultats, et de repérer les éventuelles erreurs du modèle. Néanmoins, comme dans tout réseau de neurones profond, certaines décisions internes restent partiellement opaques, notamment dans les cas d'erreurs inattendues.

5.4.Responsabilité et usage

Le modèle ne constitue ni un outil de décision automatique, ni un système devant être utilisé dans des contextes sensibles (justice, santé, droits sociaux). Il s'agit avant tout d'un outil de recherche, d'expérimentation et de documentation linguistique. L'utilisateur conserve l'entière responsabilité de l'interprétation et de l'usage des résultats fournis par le modèle.
Aucune décision ne doit être automatisée à partir des transcriptions, en particulier dans les cas où la tonalité peut avoir un impact sémantique fort dans une langue comme le Yemba (où des mots identiques segmentalement peuvent changer de sens en fonction du ton).

5.5.Consentement et finalité

Les données utilisées proviennent d'un corpus publiquement accessible, mis à disposition sur la plateforme Mendeley Data dans un but explicitement académique et scientifique. Les locuteurs ayant participé aux enregistrements l'ont fait dans un cadre de consentement informé, à des fins de recherche linguistique.
L'utilisation de ce corpus dans le présent projet s'inscrit pleinement dans cette finalité de valorisation des langues camerounaises. Aucun usage commercial non autorisé n'est envisagé. De même, toute réutilisation du modèle ou des données associées devra respecter les licences ouvertes en vigueur, en conservant la visée éducative, patrimoniale ou scientifique du projet.

6. Conclusion générale

Ce travail avait pour ambition de concevoir un système de reconnaissance automatique de la parole (ASR) pour la langue Yemba, une langue camerounaise peu dotée, en s'appuyant sur une architecture de type Seq2Seq avec GRU et mécanisme d'attention additive. À travers une démarche méthodique, alliant traitement du signal, apprentissage profond et respect des contraintes linguistiques du Yemba, plusieurs objectifs essentiels ont pu être atteints.

6.1.Récapitulatif des objectifs atteints

Le projet a permis de :

- Mettre en place une chaîne de traitement complète, de la collecte des données audio au déploiement du modèle via une interface interactive.
- Construire et entraîner un modèle GRU avec attention, adapté à la structure syllabique et tonale du Yemba.
- Développer une application fonctionnelle, capable de prédire en temps réel la transcription syllabico-tonale à partir d'un fichier audio.
- Évaluer objectivement les performances du système grâce aux métriques WER, CER et précision globale, tout en intégrant des techniques de régularisation (dropout, early stopping).
- Assurer une approche éthique du projet, en respectant la vie privée des locuteurs, en reconnaissant les biais de corpus et en préservant l'objectif non commercial du système.

6.2.Bilan global

Le modèle a affiché une précision encourageante, avec un WER de 63,02 % et un CER de 42,79 %, des résultats considérés acceptables dans un contexte de ressources limitées. Il démontre que des architectures simples comme GRU + attention peuvent offrir une alternative robuste pour des langues à tonalité complexe et sous-représentées. L'approche adoptée s'est révélée pertinente tant sur le plan scientifique que pratique, en posant les bases d'un système reproductible et extensible.
L'application développée est fonctionnelle, légère et accessible, ce qui en fait un outil utile pour les chercheurs, linguistes et acteurs de la documentation linguistique. Malgré les contraintes inhérentes à la taille du corpus, à l'absence de modèle de langage ou au traitement mot à mot, le système remplit sa fonction principale : fournir une transcription automatique annotée en tons, contribuant ainsi à la numérisation du Yemba.

6.3.Ouvertures possibles / suites à donner

Plusieurs perspectives peuvent prolonger et enrichir ce travail :

- Élargissement du corpus à un nombre plus important et plus diversifié de locuteurs, couvrant différents âges, genres, et zones géographiques.
- Intégration d'un modèle de langage pour améliorer la cohérence globale des séquences produites, notamment en contexte de phrases continues.
- Traitement de phrases entières et ajout de mécanismes de segmentation automatique des énoncés complexes.
- Implémentation d'un système de feedback utilisateur ou d'apprentissage actif, permettant une amélioration continue du modèle en conditions réelles.
- Extension à d'autres langues camerounaises de la même famille linguistique, en adaptant la même architecture à de nouveaux jeux de données.

En somme, ce projet constitue une étape fondatrice pour la reconnaissance vocale en langues bantoues peu dotées. Il démontre qu'avec des moyens limités mais une approche ciblée, il est possible de développer des outils technologiques concrets au service de la diversité linguistique.

Bibliographie

1. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473.
2. Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (pp. 369-376).
3. Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022). Glottolog 4.7. Max Planck Institute for Evolutionary Anthropology. https://glottolog.org/resource/languoid/id/yemb1255
4. Jiwer (2020). A simple and fast python package to evaluate an ASR system's performance using the Word Error Rate (WER). https://github.com/jitsi/jiwer
5. Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).
6. Mbah, E., Kouokam, J. P., Tchokouaha, F. D. S., & Ngué Um, E. (2022). YembaTones: A Tonal Corpus for Syllabic Speech Recognition in an African Bantu Language. Mendeley Data, V3. https://data.mendeley.com/datasets/cx268tmrwn/3
7. PyTorch (2024). An open source machine learning framework that accelerates the path from research prototyping to production deployment. https://pytorch.org
8. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (NeurIPS).
9. Gradio (2024). Gradio: Build Machine Learning Web Apps in Python. https://www.gradio.app
10. Torchaudio (2024). Audio I/O and signal processing for PyTorch. https://pytorch.org/audio/stable/index.html
11. Hugging Face (2024). Automatic Speech Recognition models on the Hub. https://huggingface.co/models?pipeline_tag=automatic-speech-recognition
12. Mozilla. (2020). DeepSpeech Mandarin Model – Chinese ASR with DeepSpeech. https://github.com/PaddlePaddle/DeepSpeech
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45). Association for Computational Linguistics. https://huggingface.co/docs/transformers