Speech To Text with a Pre-Trained Model

License D'excellence en Intelligence Artificial

Module de Deep Learning

FACULTE DES SCIENCES BEN M'SICK
UNIVERSITÉ HASSAN II DE CASABLANCA

Réalisé par :

- ***BOUSSAID Salama***
- ***CHAABAN Malika***
- ***ASSOULI Fatima Zahra***

Encadré par :

- ***Prof. Mr. Ben LAHMAR El Habib***
- ***Prof . Mr. EL FAKIR Zakaria***

Introduction

La conversion de la parole en texte, aussi connue sous le nom de Reconnaissance Automatique de la Parole (Speech To Text), est une composante vitale de l'intelligence artificielle qui transforme un signal vocal en écrit. Ce projet se positionne dans ce secteur en tirant parti du modèle Whisper d'OpenAI, qui est salué pour sa compétence à transcrire des discours dans plusieurs langues avec exactitude, y compris en cas de bruit ou d'accents.

Nous avons fait appel au jeu de données « Speech Accent Archive », qui regroupe des enregistrements de personnes parlant anglais avec différentes accents, pour mesurer ses performances. Ceci facilite l'évaluation de la robustesse du modèle face à la variété des accents, tout en garantissant une constance dans le contenu linguistique. Ainsi, cette étude examine tout le processus : de la préparation des données audio à la transcription automatique, en passant par l'analyse acoustique et l'évaluation via le taux d'erreur de mots (WER).

Objectifs du Projet

Mettre en œuvre une chaîne complète de traitement de la parole, de l'audio brut à la transcription finale, via :

- ✓ Visualiser les caractéristiques audio à l'aide de représentations telles que la forme d'onde, le spectrogramme et le mel-spectrogramme.
- ✓ Prétraiter des fichiers audio issus d'un corpus d'accents variés.
- ✓ Transcrire automatiquement les discours à l'aide du modèle Whisper d'OpenAI.
- ✓ Évaluer la précision des transcriptions en calculant le taux d'erreur de mots et des caractères (WER, CER).
- ✓ Explorer l'impact des accents étrangers sur la qualité de la transcription.

Methodologies

1. Chargement et Exploration du Dataset

- ❑ Utilisation de KaggleHub pour importer le corpus Speech Accent Archive.
- ❑ Lecture d'un fichier audio au format .mp3 et de son texte de référence.

2. Visualisation Audio

- ❑ Affichage de la forme d'onde du signal brut.
- ❑ Génération de spectrogrammes classiques et mel-spectrogrammes, pour analyser les fréquences.

3. Prétraitement pour Whisper

- ❑ Conversion du signal audio en mono.
- ❑ Resampling à 16 kHz (standard du modèle Whisper).

4. Transcription Automatique

- ❑ Passage du signal dans WhisperProcessor pour le format adapté.
- ❑ Prédiction avec WhisperForConditionalGeneration.
- ❑ Décodage des ID générés en texte lisible.

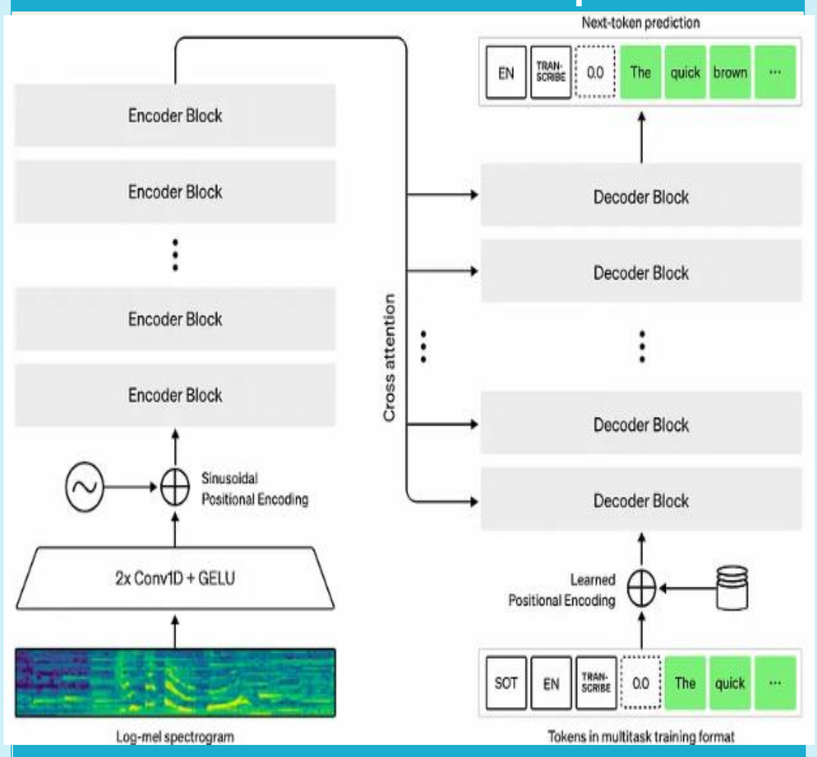
5. Évaluation

- ❑ Comparaison entre la transcription générée et le texte original via WER (Word Error Rate) et CER (Character Error Rate).

Technologies Utilisées

- **Traitement des données & Visualisation :** NumPy, seaborn, matplotlib, Pandas
- **Chargement & Lecture des fichiers audio:** librosa, IPython

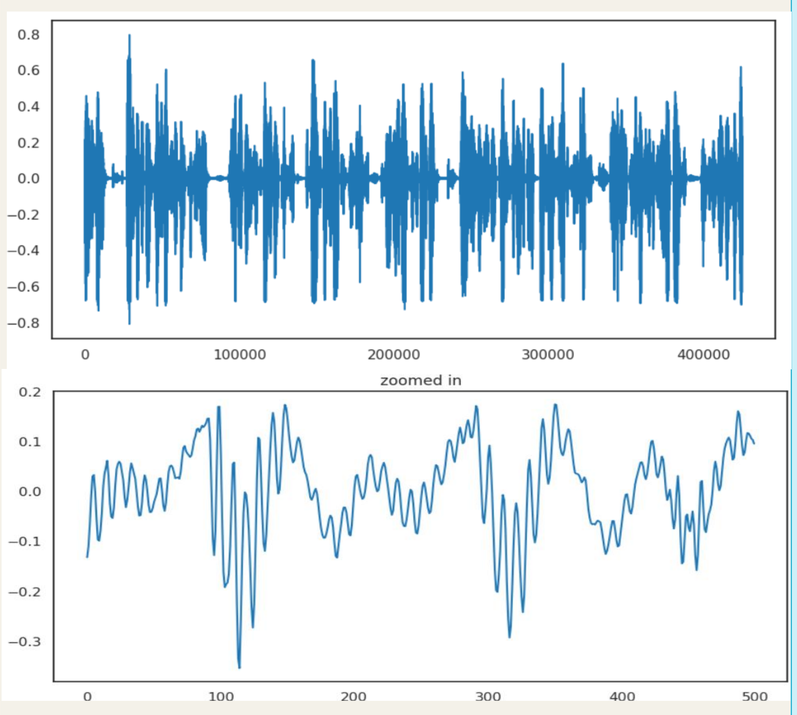
Architecture du Whisper



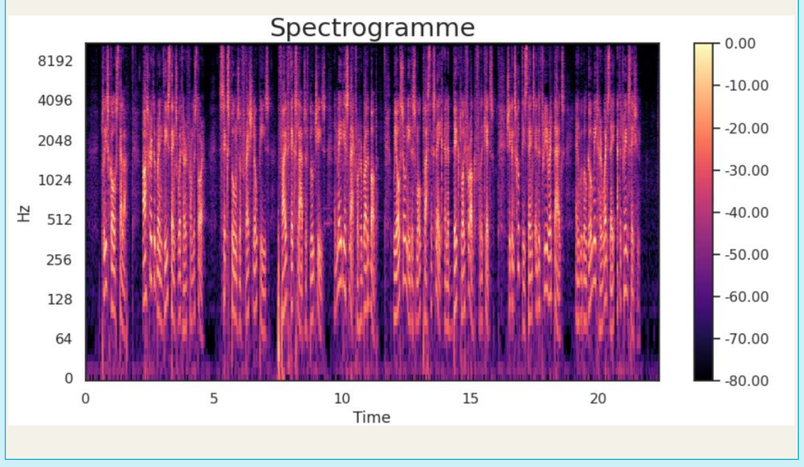
Résultats

Des représentations audio ont été créées afin de mieux saisir la structure acoustique du signal :

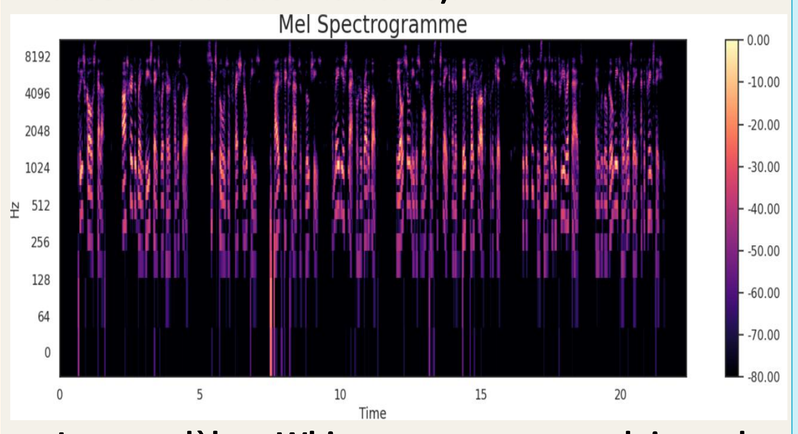
- **Onde (illustration temporelle du signal audio) :**



- **Spectrogramme (analysé selon la méthode classique du temps-fréquence):**



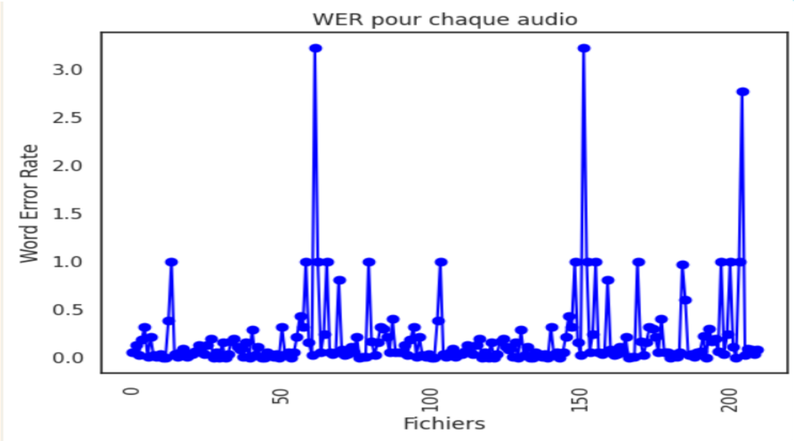
- **Mel-spectrogramme (une analyse perceptive tirée de l'audition humaine):**



Le modèle Whisper a su produire des transcription que nous avons évaluée pour savoir la robustesse de model face au différents accents.

- **WER (Word Error Rate):** Le WER mesure la différence entre la transcription générée par un modèle (texte prédit) et la transcription correcte (texte de référence).
- **WER <0.4 = acceptable:**

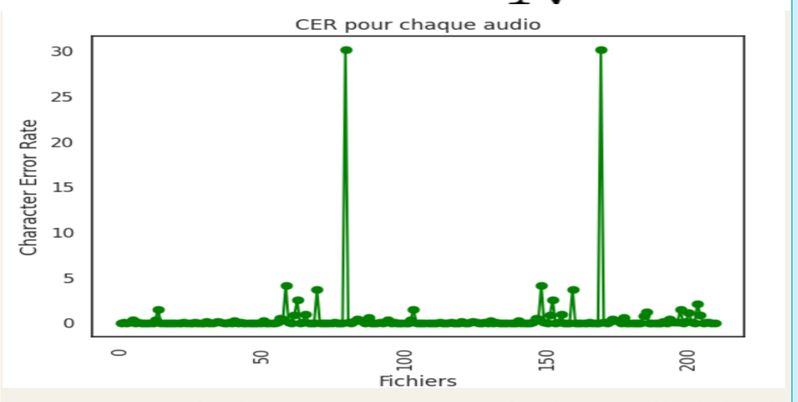
$$WER = \frac{S+D+I}{N}$$



caractères erronés (insertion, suppression, substitution) entre le texte de référence et le texte prédit .

- **CER< 20% = acceptable.**

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



Ces résultats mettent en évidence l'habileté du modèle à traiter les variations d'accent de manière efficace, du moins dans le cas examiné.

Conclusion / Perspectives

Le modèle Whisper affiche une remarquable aptitude à la transcription, même en tenant compte de divers accents. Ce projet, grâce à une procédure de prétraitement audio minutieuse et une évaluation neutre basée sur le WER, atteste l'efficacité de Whisper pour des missions de transcription multilingue à grande échelle. Ces résultats sont encourageants pour des utilisations dans la reconnaissance de la parole, l'accessibilité et même l'enseignement des langues.

References

Article Decoding Whisper

Decoding Whisper: An In-Depth Look at its Architecture and Transcription Process