Université Hassan II de Casablanca
Faculté des Sciences Ben M'Sick
Département de Mathématiques et Informatique

RAPPORT DE MINI PROJECT

LICENCE EXCELLENCE DE INTELLIGENCE ARTIFICIELLE

Prof modèle :
Pr. EL HABIB Ben Lahmar

Group 1

Houssame Bouyous
Brahim Chbab
Chakir Abderrahmane

TEAM DeepMind

Sujet :
Text Correction Pipeline After Speech Recognition

CONTENTS

Chapitre 1-Whisper d'OpenAI : description, application et limites en reconnaissance vocale spécialisée

1.1-Description et fonctionnement du modèle Whisper d'OpenAI

Le modèle Whisper d'OpenAI est un système avancé de reconnaissance automatique de la parole (ASR) qui convertit la parole en texte écrit. Il est basé sur une architecture Transformer encodeur-décodeur et a été entraîné sur un immense corpus de 680 000 heures de données audio multilingues supervisées, ce qui lui permet de gérer une grande variété d'accents, de langues et de contextes acoustiques. Whisper segmente l'audio en spectrogrammes, puis utilise ses réseaux neuronaux pour analyser les sons et prédire la séquence de mots la plus probable, offrant ainsi une transcription précise et souvent capable d'inférer la ponctuation et le contexte.

1.2-Utilisation dans le domaine SST et adaptation non supervisée

Dans le domaine de l'adaptation aux environnements spécifiques, notamment pour la reconnaissance de la parole dans des contextes variés comme le bruit, les accents ou les situations particulières (domaine SST, par exemple), un problème majeur est que le modèle pré-entraîné peut ne pas bien fonctionner sans ajustement. Pour cela, la méthode Self-Taught Recognizer (STAR) propose une adaptation non supervisée qui utilise uniquement des données non étiquetées du domaine cible pour améliorer la robustesse de Whisper. Cette approche évalue la qualité des pseudo-étiquettes générées par le modèle sans vérité terrain, ce qui permet d'affiner Whisper pour mieux reconnaître la parole dans des environnements spécifiques sans nécessiter de données annotées supplémentaires.

1.3-Problèmes et limites de l'adaptation du modèle Whisper

Cependant, ce type d'adaptation présente aussi des défis. Par exemple, Whisper peut rencontrer des difficultés dans des domaines où la langue parlée est très informelle ou peu grammaticale, comme dans certains corpus de conversation spontanée, ce qui génère des pseudo-étiquettes de moindre qualité et limite l'efficacité de l'adaptation. De plus, il existe un risque de "catastrophique Forget Ting", où le modèle adapté oublie ses connaissances initiales sur le domaine source, bien que STAR ait montré qu'il pouvait éviter ce problème sans avoir besoin de rappeler les données originales. Enfin, l'adaptation nécessite un certain volume de données non étiquetées (moins d'une heure dans les expériences), ce qui reste un compromis pratique pour déployer Whisper dans des scénarios réels variés.

Chapitre 2-Modèle Google T5/t5-base : Description, usages et complémentarité avec Whisper

2.1-Description du modèle Google T5/t5-base

Le modèle T5 (Text-to-Text Transfer Transformer) de Google est un modèle de traitement du langage naturel (NLP) basé sur une architecture Transformer encodeur-décodeur, introduit en 2019. Sa particularité est de reformuler toutes les tâches NLP sous forme de conversion texte-à-texte, où l'entrée et la sortie sont toujours des chaînes de texte. Cette approche unifiée permet à T5 d'être extrêmement polyvalent : il peut être utilisé pour la traduction, la classification, le résumé, la génération de texte, la réponse à des questions, et bien d'autres tâches. T5 est pré-entraîné sur un très grand corpus (le Colossal Clean Crawled Corpus, C4) et peut ensuite être affiné (fine-tuning) pour s'adapter à des tâches ou domaines spécifiques, ce qui lui confère une grande capacité d'adaptation et de transfert d'apprentissage.

2.2-Domaines d'utilisation de T5

Dans le domaine d'utilisation, T5 s'applique à une vaste gamme de problématiques NLP, notamment la traduction automatique, le résumé automatique, la classification de texte, la génération de réponses contextuelles, et même la compréhension grammaticale. Sa capacité à traiter des tâches variées avec un seul et même modèle facilite grandement le développement d'applications complexes, comme des chatbots contextuels ou des systèmes de recherche d'information dans des documents spécialisés. De plus, sa structure permet d'intégrer des techniques avancées d'adaptation au domaine, comme le multi-task learning ou le transfert de connaissances, ce qui améliore ses performances même avec peu de données spécifiques.

2.3-Pourquoi utiliser T5 avec Whisper ?

L'utilisation conjointe de T5 avec Whisper, modèle de reconnaissance vocale d'OpenAI, est particulièrement intéressante dans des systèmes de traitement de la parole avancés. Whisper se charge de la transcription audio en texte, tandis que T5 peut ensuite traiter ce texte pour effectuer des tâches de compréhension, de génération ou de reformulation. Par exemple, dans un pipeline de conversion parole-à-parole (speech-to-speech), Whisper transcrit la parole d'entrée, T5 génère une réponse ou une reformulation textuelle adaptée au contexte, et un modèle de synthèse vocale (TTS) produit la parole de sortie. Cette complémentarité est exploitée dans des architectures comme T5-SpeechtoSpeech, qui utilise un paradigme enseignant-étudiant pour créer des systèmes efficaces et précis de conversion vocale en temps réel. Ainsi, T5 apporte la puissance du traitement du langage naturel à partir du texte produit par Whisper, permettant des applications avancées comme les assistants virtuels, la transcription enrichie ou la traduction vocale.

2.4-Test model T5

Original	Corrigé
he go to school every day.	he goes to school every day.
I has three apple	I have three apple.
they was playing football yesterday	they were playing football yesterday.

Chapitre 3 - Méthodologie

3.1-Etape de pipeline

1. Entrée Audio

- Fichier audio ou flux en temps réel.

2. Transcription avec Whisper

- Modèle Whisper (ASR) → Convertit la parole en texte brut.
- Sortie : Texte transcrit (potentiellement bruité/erroné).

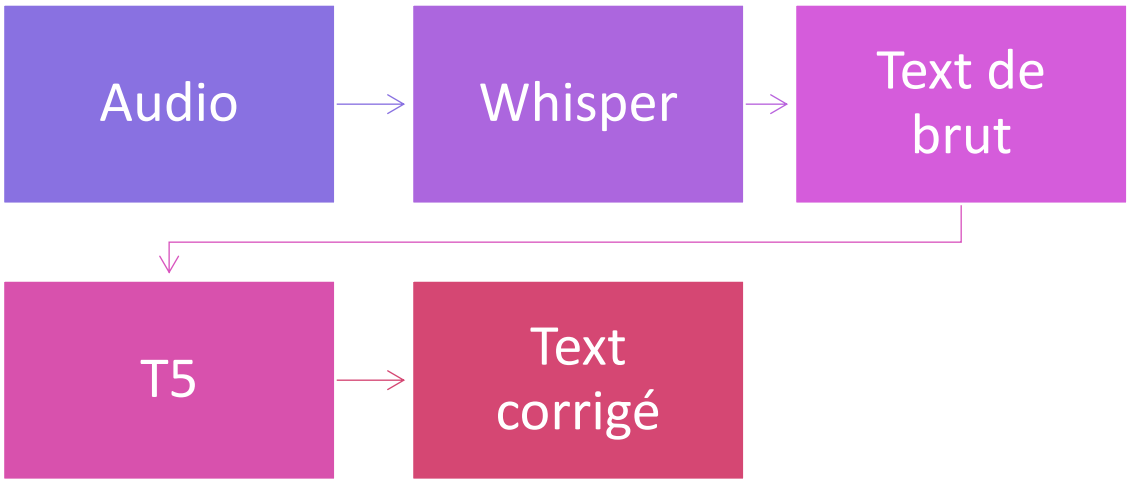
3. Correction avec T5

- Modèle T5 (NLP) → Reçoit le texte brut et le corrige.
- Tâches : Correction grammaticale, reformulation contextuelle.
- Sortie : Texte final lisible et optimisé.

4. Sortie Texte

- Transcription corrigée prête pour utilisation (affichage, traduction, etc.).

3.2-schema de Pipeline



3.3-Technologies utilisées

- **Frontend :** Next.js
- **Backend :** Flask