État d’art de Whisper

Pour ce projet nous avons du réaliser des recherches afin de trouver un système de reconnaissance vocale. Nous avons trouvé des systèmes tels que Alexa d’Amazon ou encore CMU Sphinx, un système de reconnaissance vocale développé à l'Université Carnegie Mellon. Nous avons cependant décidé d’utiliser Whisper car c’était l’outil le plus simple à utiliser au sein de notre projet.

cmd : `whisper src/resources/Audio/RecordAudio.wav --language French –model base -o src/resources/Audio/ -ftxt`

Whisper est système de reconnaissance vocale automatique développé par l’entreprise OpenAI en 2022. Le système a été entraîné avec 680 000 heures de données collectées sur le web en plusieurs langues différentes. Cette quantité d’informations a permis de renforcer la fiabilité des résultats lorsque le système est confronté à des accents différents, un bruit de fond important ou encore des langages plus difficiles.
Whisper possède également la capacité de retranscrire un fichier audio en un fichier texte dans de nombreux langages. Il peut également traduire ces langages en anglais.

L’architecture Whisper est implémentée comme un encodeur-décodeur. L’audio qui est donné en entrée du programme est découpé en plusieurs parties de 30 secondes. Puis il est transformé en spectogramme qui est passé dans un encodeur. Il fait ensuite appel à un décodeur qui est entraîné pour prédire le texte correspondant. Tout ceci est mélangé à des tokens spéciaux qui dirigent le modèle afin de réaliser certaines tâches telles que l’identification du langage, les indicateurs de durée des phrases, les retranscriptions en plusieurs langues ou la traduction en anglais.

Il y a à disposition 5 types de modèles différents, dont 4 avec des versions spécifiques à l’anglais. Ces versions spécifiques ont de meilleures performances pour 2 des 4 modèles. Ces modèles offrent une précision de retranscription et une vitesse d’exécution différente.
Il existe donc les modèles suivants :

- tiny qui utilise 39 M de paramètres, ne requiert que 1 GB de mémoire vive et est très rapide mais ne garanti pas des résultats précis.
- base qui utilise 74 M de paramètres, ne requiert que 1 GB de mémoire vive et est assez rapide, il garantit des résultats satisfaisants.
- small qui utilise 244 M de paramètres, requiert 2 GB de mémoire vive et est plutôt lent, il garantit des résultats très satisfaisants.
- medium qui utilise 769 M de paramètres, requiert 5 GB de mémoire vive et est assez lent, il garantit des résultats très satisfaisants aussi.
- large qui utilise 1550 M de paramètres, requiert 10 GB de mémoire vive et est lent, il garantit de très bons résultats.

Les pourcentages d’erreur de retranscription varient selon les langages. Par exemple le Français possède un taux d’erreur de mot de 8,3 % en utilisant le modèle large, ce qui est plus que satisfaisant.

Nous avons utilisé le modèle base car c’est celui qui nous donnait les meilleurs résultats par rapport à son temps d’exécution afin de garantir un jeu assez fluide.