Rapport d'Évolution de la Solution de Speech-to-Text en Temps Réel

EIP - EasyStream : Streamez on s'occupe du reste
modifié le: 31/08/2023

Jules **Minguet**
Dev Plugin

Romain **Bry**
Dev Plugin

Aboh Daniel Cédrick **Kakou**
Dev Plugin

Alexandre **Juan**
Dev App

Emeryck **Clerin**
Dev Full Stack

Guillaume **Robin**
Dev App Mobile

Sommaire

Introduction

Le présent rapport décrit l'évolution de notre solution de Speech-to-Text en temps réel, qui était précédemment inefficace en termes de vitesse et de précision. Nous avons mené une étude approfondie des différents modèles disponibles pour résoudre ce problème, analysé leurs avantages et inconvénients, et justifié notre choix final. Enfin, nous avons sélectionné un modèle candidat pour l'implémentation future et défini les résultats attendus de cette mise à jour.

Analyse des Modèles Étudiés

1. Speech2Text2

Avantages

- Rapidité: Speech2Text2 est un modèle de décodeur uniquement, ce qui signifie qu'il n'a pas besoin de calculer la sortie de l'encodeur de parole. Cela le rend plus rapide que les modèles de transformateurs de parole complets, qui doivent calculer la sortie de l'encodeur et du décodeur.
- Précision: Speech2Text2 est un modèle de grande taille, formé sur un ensemble de données massif. Cela lui permet d'obtenir une précision élevée sur une variété de tâches de transcription vocale.
- Flexibilité: Speech2Text2 peut être utilisé avec n'importe quel modèle d'encodeur de parole uniquement. Cela offre une grande flexibilité aux développeurs, qui peuvent choisir l'encodeur de parole qui convient le mieux à leurs besoins.

Inconvénients

- Exigence de puissance de calcul: Speech2Text2 est un modèle de grande taille, nécessitant une puissance de calcul importante pour fonctionner. Cela peut être un problème pour les applications qui doivent fonctionner sur des appareils à faible puissance, tels que les smartphones ou les appareils wearables.
- Recherche d'erreurs: Speech2Text2 peut encore commettre des erreurs, en particulier dans des conditions difficiles, telles que du bruit de fond ou des locuteurs avec un accent fort. Cela peut être un problème pour les applications qui doivent fournir une transcription vocale précise.

2. Vosk

Avantages

- Petite taille: Vosk est relativement petit, ce qui le rend adapté aux appareils à faible puissance. Cela signifie qu'il peut être utilisé sur des smartphones, des Raspberry Pi et d'autres appareils à faible puissance de calcul.
- Facile à utiliser: Vosk est facile à utiliser et à configurer, même pour les développeurs débutants. Cela le rend accessible à un large éventail d'utilisateurs.
- Multilingue: Vosk est disponible pour une variété de langages. Cela le rend adapté aux applications qui doivent être utilisées par des utilisateurs de différentes régions.

Inconvénients

- Précision: Vosk peut encore commettre des erreurs, en particulier dans des conditions difficiles, telles que du bruit de fond ou des locuteurs avec un accent fort. Cela peut être un problème pour les applications qui doivent fournir une transcription vocale précise.
- Temps de réponse: Vosk peut prendre un certain temps pour reconnaître la parole, ce qui peut être un problème pour les applications qui doivent fournir une réponse rapide (principal problème dans notre cas).
- Flexibilité: Vosk n'est pas aussi flexible que certains autres modèles de reconnaissance vocale. Il ne prend en charge qu'un nombre limité de langages et de fonctionnalités.

3. Whisper

Avantages

- Précision: Whisper est un modèle de grande taille, formé sur un ensemble de données massif. Cela lui permet d'obtenir une précision élevée sur une variété de tâches de transcription vocale. La précision de Whisper est l'un de ses principaux avantages. Il a été testé sur une variété de tâches de transcription vocale, y compris la transcription de discours, d'enregistrements audio et de conversations. Dans tous les cas, Whisper a obtenu des résultats de précision très élevés.
- Flexibilité: Whisper est disponible pour une variété de langages. Cela le rend adapté aux applications qui doivent être utilisées par des utilisateurs de différentes régions. La flexibilité de Whisper est également un avantage important. Il est disponible pour une variété de langages, y compris l'anglais, le français, l'allemand, l'espagnol et l'italien. Cela permet aux développeurs de créer des applications qui peuvent être utilisées par des utilisateurs du monde entier.
- Facile à utiliser: Whisper est disponible sous forme de bibliothèque Python, ce qui le rend facile à utiliser par les développeurs. Whisper est également facile à utiliser. Il est disponible sous forme de bibliothèque Python, ce qui le rend accessible à un large éventail de développeurs. La documentation de Whisper est complète et bien écrite, ce qui facilite la prise en main du modèle.

Inconvénients

- Exigence de puissance de calcul: Whisper est un modèle de grande taille, nécessitant une puissance de calcul importante pour fonctionner. Cela peut être un problème pour les applications qui doivent fonctionner sur des appareils à faible puissance, tels que les smartphones ou les appareils wearables. C’est l'un de ses principaux inconvénients.
- Erreurs: Bien que particulièrement précis dans des conditions optimales, Whisper peut encore commettre des erreurs, en particulier dans des conditions difficiles, telles que du bruit de fond ou des locuteurs avec un accent fort. Cela peut être un problème pour les applications qui doivent fournir une transcription vocale précise. La précision de Whisper est élevée, mais il peut encore commettre des erreurs. Ces erreurs sont plus fréquentes dans des conditions difficiles, telles que du bruit de fond ou des locuteurs avec un accent fort. Cela peut être un problème pour les applications qui doivent fournir une transcription vocale précise.

Tableau comparatif des modèles candidats

Caractéristique	Speech2Text2	Vosk	Whisper
Type de modèle	Transformateur de parole	Transformateur de parole	Transformateur de parole
Précision	Haute	Bonne	Haute
Exigences de puissance de calcul	Élevées	Faibles	Élevées
Recherche d'erreurs	Peut encore commettre des erreurs	Peut encore commettre des erreurs	Peut encore commettre des erreurs
Avantages	Rapide, flexible, disponible pour plusieurs langages	Léger, disponible hors ligne, facile à utiliser	Précis, flexible, disponible pour plusieurs langages
Inconvénients	Peut nécessiter une puissance de calcul importante	Peut être lent, moins flexible	Peut nécessiter une puissance de calcul importante

Choix du Modèle Candidat: *Whisper*

Nous avons choisi le modèle Whisper pour notre solution en raison de plusieurs raisons convaincantes. Tout d'abord, Whisper se distingue par sa précision élevée, résultant de sa grande taille et de son entraînement sur un ensemble de données massif. Cette précision est essentielle pour fournir des transcriptions vocales de haute qualité, répondant ainsi aux besoins de nos utilisateurs en matière de précision.

De plus, une caractéristique essentielle de Whisper est la possibilité de choisir la taille du modèle, ce qui nous permet de trouver un équilibre optimal entre rapidité d'inférence et précision. Cette adaptabilité nous permet de personnaliser notre solution en fonction de nos besoins spécifiques, garantissant ainsi une expérience utilisateur fluide tout en maintenant une qualité de transcription correcte.

Comparé à Speech2Text2, qui a été annoncé en 2023, Whisper est plus ancien, ayant été introduit en 2022, et il offre une plus grande variété de bibliothèques qui simplifient son processus d'implémentation.

L'un des défis potentiels que nous pourrions rencontrer réside dans le fait que le modèle est exécuté en local sur l'ordinateur de l'utilisateur. Dans le cas d'une configuration matérielle insuffisante, il pourrait ne pas fonctionner de manière optimale.

Objectifs de l'Implémentation

Nous attendons les résultats suivants de l'implémentation de Whisper sur notre solution:

1. **Amélioration de la Précision**: Nous prévoyons une nette amélioration de la précision de la reconnaissance de la parole, réduisant ainsi le nombre d'erreurs de transcription.

2. **Accélération du Traitement**: Whisper étant open-source, nous pourrons le charger et l'utiliser en local afin de réduire le temps entre la soumission des données audio à transcrire et l'obtention des résultats, ce qui permettra une communication plus fluide et une expérience utilisateur améliorée.

Résultats obtenus

Après implémentation de la nouvelle solution, nous avons obtenu les évolutions suivantes:

Précision de la transcription

Le modèle précédemment utilisé présentait une imprécision notable en raison des contraintes liées à la transcription des données audio. Nous étions contraints de transmettre des fichiers audio, ce qui parfois nous obligeait à envoyer des fichiers avec des phrases incomplètes. En adoptant Whisper, nous avons désormais la flexibilité de soumettre les données audio pour transcription à notre convenance, évitant ainsi de transmettre des phrases tronquées ou ayant la possibilité de le faire avant que la phrase complète ne soit captée par le microphone, ce qui nous permet d'obtenir une transcription intermédiaire. De plus, il est évident que Whisper présente une précision supérieure à celle du modèle que nous utilisions précédemment.

Ancienne solution
Pourcentage de mots correctement transcrits: 22.6%

Nouvelle solution
Pourcentage de mots correctement transcrits: 74.2% (+51.6%)

Rapidité de traitement

Le modèle de transcription que nous utilisions auparavant était uniquement accessible via une API, ce qui signifiait que nous devions ajouter le temps nécessaire à l'envoi des données audio et à la réception des résultats à la durée totale du traitement des données. Grâce à l'utilisation locale de Whisper, nous avons pu éliminer ces délais de transmission. De plus, il semble que lorsque la taille de Whisper est adaptée à la configuration de l'ordinateur de l'utilisateur, la durée de transcription des données audio est simplement plus courte que celle de notre solution précédente.

Ancienne solution
Moyenne du nombre de mots transcrits par seconde: 1.429

Nouvelle solution
Moyenne du nombre de mots transcrits par seconde: 13.276 (+929.04%)

Conclusion

En conclusion, suite à l'implémentation de notre nouvelle solution basée sur le modèle Whisper, nous avons observé des améliorations significatives dans deux domaines clés : la précision de la transcription et la rapidité de traitement. En ce qui concerne la précision, nous avons constaté une augmentation considérable de 51,6% dans le pourcentage de mots correctement transcrits, passant de 22,6% avec l'ancienne solution à 74,2% avec la nouvelle solution.
De plus, en termes de rapidité de traitement, nous avons atteint une augmentation spectaculaire de 929,04% dans la moyenne du nombre de mots transcrits par seconde, passant de 1.429 mots par seconde avec l'ancienne solution à 13.276 mots par seconde avec la nouvelle solution. Ces résultats démontrent clairement les avantages indéniables de l'adoption de Whisper pour notre solution de transcription vocale en temps réel, offrant à nos utilisateurs une expérience plus précise et plus rapide.