UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# SIMPLESPEECH

-

# FACILITER L'ADOPTION DU TRAITEMENT AUTOMATIQUE DE LA PAROLE PAR LES COMMUNAUTÉS DE RECHERCHE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

THOMAS D. SOULAS

SEPTEMBRE 2024

## TABLE DES MATIÈRES

## TABLE DES FIGURES

## LISTE DES TABLEAUX

## ACRONYMES

**DAV** détection d'activité vocale.

**DAVSdI** détection d'activité vocale et superposition d'inférences.

**FT-FID** fréquence des termes - fréquence inverse des documents.

**IPA** interface de programmation applicative.

**IU** interface utilisateur.trice.

**JWT** JSON Web Token.

**MCV** Mozilla Common Voice.

**MFA** Montreal Forced Aligner.

**NIST** National Institute of Standards and Technology.

**OAD** objet d'accès aux données.

**RAP** reconnaissance automatique de la parole.

**SdI** superposition d'inférences.

**SdM** sacs de mots.

**SdMC** sacs de mots continu.

**SPI** superposition partielle d'inférences.

**TALN** traitement automatique du langage naturel.

**TAP** traitement automatique de la parole.

**TEC** taux d'erreur sur les caractères.

**TEM** taux d'erreur sur les mots.

**UQAM** Université du Québec à Montréal.

**XU** expérience utilisateur.trice.

## RÉSUMÉ

Ce mémoire présente SimpleSpeech, une application de transcription libre, modulaire et facilement adaptable, conçue pour simplifier l'utilisation du traitement automatique de la parole au sein des communautés de recherche.

Dans cette optique, nous explorons les outils et modèles de pointe en linguistique informatique et en traitement automatique de la parole, tout en examinant les différentes boîtes à outils disponibles pour la reconnaissance vocale. Nous passons en revue ces boîtes à outils en mettant l'accent sur les applications d'annotation manuelle et automatique. Nous analysons également les ensembles de données les plus pertinents pour le développement de cette application, et évaluons les performances des outils grâce à plusieurs métriques spécifiques.

Nous abordons ensuite les réflexions entourant la création de SimpleSpeech, en présentant les premières maquettes et architectures ainsi que les technologies employées. Nous mettons en lumière les différents modules qui composent l'application, tout en envisageant les améliorations et évolutions possibles.

Enfin, ce mémoire expose les évaluations préliminaires de SimpleSpeech, menées dans le cadre d'un projet d'évaluation d'outils d'intervention psychologique. Nous identifions également les axes de recherche futurs pour renforcer la qualité et la diversité des transcriptions offertes par SimpleSpeech.

# INTRODUCTION

La transcription de la parole date de bien avant l'antiquité où déjà des scribes étaient chargés de transcrire manuellement les contenus de sources orales ou de documents existants. La transcription manuelle fut une pratique courante dans les monastères, les universités et les chancelleries pendant des siècles. Les sténographes jouent un rôle crucial dans l'histoire de la transcription manuelle. Ils et elles sont des professionnels formés à la saisie rapide et précise de la langue parlée, souvent lors de réunions, de procès, de conférences ou d'autres événements où une transcription en temps réel est nécessaire. Encore aujourd'hui, de nombreux discours, tribunaux ou assemblées législatives emploient des sténographes alors que la transcription automatique est développée et améliorée chaque jour. Par conséquent, à mesure que les techniques de transcription automatique s'améliorent, il devient de plus en plus crucial de rendre ces outils accessibles autant à la communauté de recherche qu'au grand public.

La transcription s'inscrit dans le domaine du traitement automatique de la parole (TAP). De nombreuses tâches existent et sont couvertes par ce dernier. D'autres tâches connues sont, par exemple, l'identification des locuteurs et l'analyse du contenu linguistique. Ces tâches reflètent les besoins des personnes utilisatrices dans différents contextes professionnels, académiques, industriels ou médicaux. En effet, le TAP a de nombreux domaines d'application, changeant souvent la manière dont nous interagissons avec la technologie au quotidien. Parmi ces applications, on peut citer les assistants vocaux personnels qui aident les personnes utilisatrices à effectuer des tâches telles que la gestion de leur emploi du temps, la recherche d'informations en ligne, et le contrôle des appareils domestiques intelligents. Les interfaces automatiques de clavardage utilisant le TAP facilitent les interactions utilisateur-machine dans divers contextes, allant du service client en ligne à la réservation de billets. De plus, le TAP alimente la traduction automatique en temps réel, rendant la communication multilingue de plus en plus simple. Dans le domaine de la santé, les applications de TAP peuvent assister les professionnels médicaux dans la documentation des dossiers médicaux électroniques, la transcription des dictées cliniques et même dans le suivi des symptômes des patients. Enfin, les applications de reconnaissance vocale permettent aux conducteurs d'interagir avec leurs véhicules sans quitter la route des yeux, améliorant ainsi la sécurité routière.

Aujourd'hui, le domaine du TAP est très actif en recherche (Radford *et al.*, 2022; Gulati *et al.*,

2020) mais les travaux produits restent très peu accessibles pour des personnes qui n'ont pas de compétences en programmation (Ravanelli *et al.*, 2021; Watanabe *et al.*, 2018; Kuchaiev *et al.*, 2019). Les systèmes de TAP peuvent être nécessaires dans une variété de cas d'utilisation, tels que la transcription automatique de réunions, d'entretiens, ou encore de conférences. De plus, ces systèmes peuvent être utiles dans le cas de sous-titrage et de dictée ou encore d'interaction avec des assistants vocaux. Les personnes non expertes peuvent choisir des solutions payantes, mais ces solutions ne répondent généralement pas entièrement à leurs besoins et peuvent être peu modulaires. De plus, elles sont souvent accessibles en ligne, ce qui les rend vulnérables aux cyberattaques. En outre, leur coût peut constituer un obstacle majeur, tant pour les groupes de recherche que pour le grand public. Certaines lois couvrent la sécurité et la confidentialité des données, toutefois pour tout usager une fois que les données sont déposées sur ces solutions payantes il est impossible de savoir ce qui est réellement fait avec elles. En pratique, tout contenu partagé en ligne devient en effet utilisable par des tiers.

Il est essentiel de démocratiser l'accès aux techniques de TAP car la disponibilité de logiciels de transcription permet de rendre les contenus audio accessibles à des publics plus larges, notamment ceux qui ont des difficultés à comprendre l'oral ou qui préfèrent lire le texte. En transcrivant des contenus de divers niveaux de langues, avec ou sans particularismes locaux, nous contribuons à préserver la richesse et la diversité linguistique. Cette pratique permet de refléter la variété des expressions et des accents présents dans la francophonie. Par ailleurs, la disponibilité de transcriptions précises dans différents registres linguistiques offre aux communautés de recherche en linguistique l'opportunité d'étudier et d'analyser les langues, contribuant ainsi à une meilleure compréhension de leurs structures et de leurs utilisations dans divers contextes sociaux et culturels.

Notre objectif est de fournir une application de transcription automatique accessible, gratuite et en code source ouvert. Comme mentionné précédemment, il existe divers outils et boîtes à outils disponibles pour les communautés de recherche et les personnes développeuses spécialisées en TAP. Nous visons à choisir l'un de ces ensembles d'outils et à le mettre en œuvre pour créer une application qui réponde précisément à ces besoins. En termes de fonctionnalités, de communauté ainsi que de prise en main, SpeechBrain (Ravanelli *et al.*, 2021) est la solution que nous jugeons appropriée pour créer une telle application. Les raisons de ce choix seront expliquées dans les chapitres 1 et 2

La question de recherche traitée dans ce manuscrit est la suivante :

| Comment faciliter l'usage du traitement automatique de la parole par les communautés de recherche ? |
|---|

Pour répondre à notre question de recherche, il est essentiel de comprendre le contexte du TAP, notamment en abordant en détail les outils et les boîtes à outils utilisés. Nous présenterons au chapitre 1 une revue de littérature sur les techniques d'annotation manuelle ainsi que sur les solutions payantes dominantes sur le marché. Dans le chapitre 2, nous examinerons divers ensembles de données disponibles en TAP et évaluerons leur pertinence pour notre travail. Nous discuterons également des métriques clés pour évaluer les performances des différents outils mentionnés dans le chapitre 1. Nous présenterons dans le chapitre 3 une définition initiale de SimpleSpeech et mettrons en avant les réflexions que nous avons eu en amont du projet. Nous détaillerons SimpleSpeech dans le chapitre 4 : son architecture, les décisions stratégiques prises lors de sa conception et son processus de développement, ainsi que l'état actuel du projet. Enfin, nous conclurons en soulignant les possibilités et les évolutions envisagées pour SimpleSpeech, mettant en lumière les perspectives d'avenir de l'outil.

## AUTRES CONTRIBUTIONS

Dans le cadre de ma maîtrise, j'ai participé à l'atelier *Early Risk Prediction on the Internet* (eRisk [4]). Cet atelier propose chaque année trois tâches. Ces tâches concernent la détection de symptômes liés à certaines pathologies mentales, comme la dépression, l'automutilation ou encore la dépendance à certaines formes de jeux d'argent. La détection de symptômes est réalisée à partir de la production textuelle. J'ai pu contribuer à :

— Maupomé, D., Armstrong, M. D., Rancourt, F., **Soulas, T.** et Meurs, M.-J. (2021). Early detection of signs of pathological gambling, self-harm and depression on social media. Dans *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CLEF-WN 2021)*, volume 2936 de CEUR-WS.org, 1031–1045. CEUR Workshop Proceedings (Maupomé *et al.*, 2021)

— Saravani, S. H. H., Normand, L., Maupomé, D., Rancourt, F., **Soulas, T.**, Besharati, S., Normand, A., Mosser, S. et Meurs, M.-J. (2022). Measuring the severity of the signs of eating disorders using similarity-based approaches. Dans *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF-WN 2022)*, volume 3180 de CEUR-WS.org, 936–946. CEUR Workshop Proceedings (Saravani *et al.*, 2022)

— Maupomé, D., **Soulas, T.**, Rancourt, F., Cantin-Savoie, G., Winterstein, G., Mosser, S. et Meurs, M.-J. (2023). Lightweight methods for early risk detection. Dans *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF-WN 2023)*, volume 3497 de CEUR-WS.org, 718–726. CEUR Workshop Proceedings (Maupomé *et al.*, 2023)

Durant la période de la Covid-19, nous avons développé une preuve de concept permettant l'analyse de sentiments des usagers de X [5] en fonction de mots clefs. L'article correspondant est en cours d'évaluation dans une revue. J'ai participé à l'écriture et la relecture de ce dernier ainsi qu'à une partie du développement de l'interface.

— Rokni, S., Farmer Y., **Soulas T.**, Lancelot N., Meurs M.-J., Mosser S., Duhoux A. et Bouthillier M.-E. Using X 1 sentiment analysis to gauge public opinion regarding COVID-19 pandemic in Quebec.

4. `https://erisk.irlab.org/`

5. Anciennemment désigné sous le nom de Twitter

# CHAPITRE 1
# SYSTÈMES DE TRAITEMENT AUTOMATIQUE DE LA PAROLE

L'objectif de ce chapitre est de présenter l'état actuel des outils de transcription et d'annotation de la parole. Il est donc primordial d'introduire et définir le traitement automatique de la parole (TAP). Puis, nous examinerons les outils et modèles développés dans le cadre de la recherche en linguistique informatique et en traitement automatique du discours. Nous présenterons ensuite les boîtes à outils conçues pour répondre à diverses tâches liées au traitement de la parole. Nous tenterons également d'illustrer la complexité d'utilisation de chaque outil à l'aide d'exemples.

En dernier lieu, bien que notre intérêt principal porte sur l'automatisation de la transcription, nous soulignerons les quelques applications non-automatisées qui ont été préalablement employées et appréciées avant les initiatives d'automatisation. Nous terminerons ce chapitre en discutant des applications automatiques dédiées au traitement de la parole et en décrivant en détail leurs utilisations spécifiques.

## 1.1 Traitement automatique de la parole

Le traitement automatique du langage naturel (TALN) est le domaine de l'apprentissage automatique qui se concentre sur l'interaction entre les humains et les machines par le biais du langage parlé et écrit. Le TAP est une discipline qui emprunte certaines approches de l'apprentissage automatique dans l'objectif d'automatiquement capter, transmettre, identifier, synthétiser et comprendre la parole. Les particularités de la langue parlée ont rendu le développement du domaine difficile. Comme le TALN, celui-ci implique notamment la linguistique informatique, les mathématiques et les statistiques mais aussi le traitement du son tel que les signaux et l'acoustique.

Le TAP tel qu'on le connaît existe depuis plus de sept décennies, notamment avec le premier outil de transcription des chiffres parlés (Davis *et al.*, 1952). Le domaine du TAP ne se limite pas à la transcription de l'audio. Il existe un très grand nombre de cas d'utilisation. La reconnaissance automatique de la parole (RAP) est le sous-domaine communément appelé « Speech-to-Text (STT ou S2T) » dont l'objectif est de transcrire la parole. Il existe aussi des techniques d'identification et de vérification biométrique de la personne locutrice. La segmentation du flux audio en segments

homogènes, basée sur l'identification de chaque locuteur, est parfois indispensable. Ce processus est connu sous le nom de diarisation des locuteurs.

Dans divers contextes, l'identification de la langue, la reconnaissance des émotions, des sentiments, de l'âge, du genre ou de l'accent peut s'avérer nécessaire. Pour d'autres, détecter des mots clefs ou des silences est crucial. De plus, des tâches visant à améliorer la qualité sonore, notamment en détectant et en réduisant le bruit ou les distorsions sonores, peuvent être indispensables. Dans le cadre d'entretiens, il peut être nécessaire de faire de la RAP puis d'aligner la transcription à l'audio pour savoir quand un mot est prononcé. Par ailleurs, certaines analyses nécessitent l'extraction de mots, de phonèmes, de variations de fréquence sonore, de rythme et d'intonation à partir de la parole, voire l'identification de séquences atypiques telles que des bégaiements.

L'usage de ces diverses avancées dans la recherche peut se révéler indispensable dans de nombreux domaines. La RAP permet aux conducteurs et conductrices d'utiliser les fonctionnalités de leur véhicule en parlant, ce sans quitter la route des yeux. Les assistants virtuels tels que Siri d'Apple, Alexa d'Amazon ou encore Cortana de Microsoft utilisent les techniques du TAP pour être capables de « comprendre » les requêtes et d'y répondre. Pareillement, les outils de dictée sont de plus en plus utilisés et certaines entreprises utilisent des outils permettant de transcrire et résumer leurs réunions.

Plus simplement, le TAP peut être un moyen de réduire l'écart entre les personnes sans handicap et les personnes en situation de handicap. Que le handicap soit une déficience auditive, ou une pathologie affectant la capacité à correctement prononcer ou entendre des mots, des solutions sont envisageables par le biais du TAP. C'est le cas, par exemple avec le sous-titrage ou la création de corpus spécifiques permettant d'analyser les problématiques liées à certains handicaps (Woisard *et al.*, 2021).

Dans le cadre de ce projet, nous cherchons une solution qui permet de répondre à une multitude de tâches. Nous ferons donc la distinctions entre « outils » et « boîtes à outils ». Nous qualifierons d'outils les solutions qui répondent au maximum à deux tâches. Une solution qui permet de transcrire et de traduire est donc un outil. Nous qualifierons de boîte à outils toute solution qui permet de répondre à plus de deux tâches. Une solution qui permet de transcrire, traduire et résumer la parole

est une boîte à outils. Certains pourraient arguer qu'une boîte à outils soit considérée comme telle dès lors qu'elle contient plusieurs outils. Cependant, opter pour cette approche nous permet de nous concentrer sur les ensembles d'outils riches en fonctionnalités.

Á travers ce projet, nous souhaitons aussi mettre l'accent sur la protection des données notamment dans le cadre de données sensibles. Á cette fin, nous souhaitons nous concentrer sur les législations en vigeur dans le cadre de la protection des données.

## 1.2 Législation et protection des données

Il est très important que les applications utilisées ne constituent pas une brèche de confidentialité des données. Lorsqu'un service est disponible en ligne, il n'est pas exclu que celui-ci récupère les données et s'en serve pour améliorer ses résultats. Ceci constitue une brèche de confidentialité puisque ces données sont réutilisées par le service tiers.

Il est crucial de souligner l'importance de la sécurité et de la régulation des données. À cet égard, il est primordial de mettre en lumière certaines législations et normes :

- **Règlement Général sur la Protection des Données (RGPD)** [6] est une loi européenne visant à protéger les données personnelles des individus. Elle donne aux gens plus de contrôle sur leurs données et oblige les entreprises à les traiter de manière responsable. Cela signifie obtenir le consentement des personnes utilisatrices avant de collecter leurs données, les informer sur la manière dont leurs données seront utilisées, et leur permettre de les supprimer ou de les corriger si nécessaire.
- **Health Insurance Portability and Accountability Act (HIPAA)** [7] est une loi américaine qui protège les informations médicales des individus. Elle impose des règles strictes sur la manière dont les prestataires de soins de santé et les compagnies d'assurance peuvent utiliser et divulguer ces informations. HIPAA vise à garantir la confidentialité, l'intégrité et la sécurité des données médicales des patients.
- **Loi sur la protection des renseignements personnels et les documents électro-**

6. `https://gdpr.eu/`

7. `https://www.hhs.gov/hipaa/index.html`

**niques (LPRPDE)** [8] est une loi fédérale canadienne qui s'applique aux entreprises qui exercent des activités commerciales au niveau fédéral. La loi établit des principes fondamentaux pour le traitement des renseignements personnels tels que le consentement, la limitation de la collecte, la sécurité des données et le droit d'accès des individus à leurs renseignements personnels. Elle accorde également aux individus le droit de corriger les renseignements personnels inexacts et de déposer des plaintes auprès du Commissariat à la protection de la vie privée du Canada en cas de violation de la loi.

— **ISO 27001** [9] est une norme internationale qui établit les exigences pour un système de gestion de la sécurité de l'information. Cette norme vise à aider les organisations à protéger les informations sensibles en mettant en place des processus et des contrôles de sécurité appropriés. La norme ISO 27001 repose sur un cycle d'amélioration continue, comprenant l'établissement de politiques de sécurité, la réalisation d'évaluations des risques, la mise en œuvre de contrôles de sécurité et la surveillance continue des performances du système de gestion de la sécurité de l'information.

— **La loi 25** [10], adoptée au Québec en 2021, modernise les règles de protection des renseignements personnels pour les entreprises et les organismes publics. Elle impose de nouvelles obligations, comme la nomination d'un responsable de la protection des données et l'élaboration de politiques de gestion des renseignements personnels. La loi renforce également les droits des citoyens, notamment en leur permettant de demander l'accès, la correction ou la suppression de leurs données. Les entreprises doivent aussi signaler les incidents de sécurité et obtenir le consentement clair des individus pour collecter et utiliser leurs informations personnelles.

Même si HIPAA se concentre spécifiquement sur les données médicales, il est essentiel de le prendre en considération, car le domaine médical peut également avoir des exigences en matière de transcription du discours. Cependant, même lorsque les entreprises se conforment à ces lois, il reste possible que des fuites et des problèmes de sécurité surviennent.

---

8. Commissariat à la protection de la vie privée du Canada

9. `https://www.iso.org/standard/27001`

10. `https://www.cai.gouv.qc.ca/protection-renseignements-personnels/sujets-et-domaines-dinteret/principaux-changements-loi-25`

## 1.3 Les outils de traitement automatique de la parole

Plusieurs approches de TAP et de RAP de bout en bout [11] ont été développées. Nous faisons ici une liste non exhaustive d'approches libres d'accès et d'utilisation. Ces outils respectent donc les législation mises en place pour la protection des données puisqu'ils sont la responsabilité de la personne utilisatrice.

Par exemple, DeepSpeech (Hannun *et al.*, 2014), développé par Mozilla, est un outil permettant de transcrire la parole, que celle-ci soit dans des environnements bruyants ou non. Il prend en charge une utilisation sur appareil hors ligne, fournit une documentation complète et propose des modèles pré-entraînés pour une adoption facile. DeepSpeech a été entraîné sur 7380 heures d'audio en anglais (36880 locuteurs et locutrices différents). Il n'est plus maintenu et certaines personnes du projet ont fondé Coqui. L'équipe de Coqui a mis en avant un dépôt [12] basé sur la RAP. Cependant, celui-ci n'est plus maintenu non plus, ce qui est dû aux avancées en recherche dans le domaine.

L'une de ces avancées est notamment Whisper (Radford *et al.*, 2022), un projet d'OpenAI. Il est possible de communiquer avec l'interface de programmation applicative (IPA) d'OpenAI [13] pour recevoir des transcriptions de fichiers et le code source de Whisper est libre d'accès. Whisper permet de générer une transcription mais aussi de réaliser sa traduction vers l'anglais. Ce modèle a été entraîné sur 680 000 heures d'audio : 65% représentent une transcription de l'anglais et 35% sont des audio d'une autre langue transcrite en anglais ou dans la langue d'origine. Il est possible d'utiliser Whisper en mode hors-ligne et celui-ci détecte la langue utilisée à partir des 30 premières secondes de l'audio fourni. Il a la particularité de couper automatiquement les fichiers longs en des segments courts.

CAT (Xiang et Ou, 2019; An *et al.*, 2020) est une approche bout en bout qui se concentre sur la RAP. Elle se base sur une interprétation continue de l'audio. Son objectif est de combiner les avantages des approches hybrides et bout en bout pour atteindre de meilleures performances. Elle est aussi utilisable pour de la transcription en temps réel. Les résultats obtenus sont similaires à

11. On entend bout en bout une approche ne nécessitant pas de modification interne pour fonctionner sur d'autres données d'entrée.

12. `https://github.com/coqui-ai/STT`

13. `https://platform.openai.com/docs/models/whisper`

d'autres approches à l'état de l'art.

NeurST (Zhao *et al.*, 2021) est présenté comme une boîte à outils, mais se focalise principalement sur la transcription et traduction de la parole. Il permet l'apprentissage et la mise en place de processus de transcription et de traduction. Son développement est en arrêt depuis 2022. Pyannote (Bredin, 2023; Bredin et Laurent, 2021) et SideKit (Larcher *et al.*, 2016) sont des bibliothèques qui se concentrent particulièrement sur la séparation des personnes locutrices. Le projet ALIZÉ (Bonastre *et al.*, 2005) met aussi à disposition un outil de reconnaissance de personnes locutrices. Son développement a cependant cessé depuis 2017. Une approche possible est donc d'utiliser Pyannote pour déterminer la diarisation et de transcrire en utilisant Whisper. Utiliser une approche de diarisation permettra de segmenter l'audio par locuteur en les identifiant. SPEAR (Khoury *et al.*, 2014) est une bibliothèque Python axée sur la reconnaissance des locuteurs et des locutrices. Cette bibliothèque, basée sur Bob (Anjos *et al.*, 2012), se concentre sur la détection de l'activité vocale et l'extraction d'attributs pour atteindre cet objectif. Asteroid (Pariente *et al.*, 2020) est un outil spécialisé dans la séparation de signaux audio composés de multiples sources en signaux individuels. FASST (Salaün *et al.*, 2014) et openBliSSART (Schuller *et al.*, 2009) sont également des outils dédiés à la séparation de signaux audio. Ces outils offrent des fonctionnalités avancées pour la décomposition de signaux complexes. pyAudioAnalysis (Giannakopoulos, 2015) est une bibliothèque Python destinée à l'extraction d'attributs et à la classification de segments audio.

Certains outils de RAP se concentrent sur des langues peu dotées ou sur une seule langue afin de répondre aux besoins spécifiques de ces langues ou de leurs locuteurs. Par exemple, Vakyansh (Chadha *et al.*, 2022) est un outil qui se concentre principalement sur les langues indo-aryennes et dravidiennes. Il vise à fournir des solutions RAP adaptées à ces langues, qui peuvent souvent être sous-représentées dans les ressources et les technologies linguistiques existantes. Un autre exemple est YAST (Ferreira *et al.*, 2012), un modèle libre qui était spécialisé dans la reconnaissance de la parole pour toutes les langues sous-représentées. Le code de YAST n'est plus libre d'accès.

## 1.4 Boîtes à outils pour le traitement automatique de la parole

Les approches bout en bout citées précédemment ne répondent cependant pas à tous nos besoins. Nous nous concentrons certes sur la transcription, mais il est plus intéressant pour nous de nous

permettre une évolution vers de nombreuses tâches sur le long terme. Nous pourrions dépendre de plusieurs outils différents mais cela signifie devoir s'adapter à chaque outil et notamment dépendre de la maintenance de plusieurs outils. Cela nécessiterait des ajustements considérables dans l'infrastructure, le développement et les processus de déploiement, étant donné que chaque outil est souvent spécialisé dans une tâche spécifique. De plus, une boîte à outils a plus de chances d'être maintenue. Elle remplit en effet plus de fonctions et peut plus facilement regrouper une communauté signifiante de chercheurs et chercheuses.

Il existe plusieurs « boîtes à outils » de transcription parmi lesquelles nous pouvons retrouver Kaldi (Povey *et al.*, 2011), une boîte à outils efficace établie depuis 2010. L'objectif de Kaldi est de proposer une boîte à outils en TAP. Cette boîte à outils doit être la plus extensible et modulaire possible, tout en proposant des exemples pour construire des systèmes de TAP. Elle est très documentée et vise à être facile d'utilisation, notamment en utilisant des algorithmes qu'il n'est pas nécessaire de reconfigurer en fonction des jeux de données. Kaldi reste encore aujourd'hui l'une des boîtes à outils en TAP les plus proéminentes. Beaucoup d'outils et boîtes à outils utilisent, en effet, le même formalisme. Cependant, Kaldi est écrite en C++ et la majorité des projets en apprentissage automatique sont développés en python, ce qui rajoute donc une couche supplémentaire d'abstraction en python à réaliser. PyKaldi2 (Lu *et al.*, 2019) est une tentative d'interfaçage entre Kaldi et Pytorch mais le développement a cessé depuis 2019. Nous ne nous concentrerons donc pas sur Kaldi puisque nous souhaitons limiter les couches applicatives et le jonglage entre différents langages de programmation.

Fairseq (Ott *et al.*, 2019) est une boîte à outils qui se concentre sur les tâches de modélisation de séquences. Elle peut être mise en place pour utiliser des modèles pré-entraînés pour la traduction, le résumé automatique et la modélisation de la langue. Certains chercheurs et chercheuses l'ont aussi utilisée pour générer ou corriger des textes. Fairseq S2T (Wang *et al.*, 2022a), une extension de Fairseq, met à disposition un certain nombre de modèles à l'état de l'art pour les tâches de RAP et de traduction de la parole. Fairseq S^2 (Wang *et al.*, 2021a) est une autre extension de Fairseq, cette fois-ci permettant de faire de la synthèse de la parole.

S3PRL (Yang *et al.*, 2021; Liu *et al.*, 2020b,a) est une interface python permettant d'apprendre des modèles ou d'utiliser des modèles pré-entraînés de manière non-supervisée. S3PRL se concentre

sur de l'apprentissage non-supervisé. Le dépôt dans lequel se trouve l'outil met en avant le fait que S3PRL puisse être utilisé pour tout type de tâches. Cela dit, les auteurs et autrices de (Yang *et al.*, 2021) mettent surtout en avant les tâches de reconnaissance de mots ou de phonèmes. Pareillement, l'identification et la vérification de locuteurs et locutrices ainsi que leur séparation est aussi évoquée. Est aussi mis en avant la possibilité de détecter les intentions et les émotions.

Fairseq et S3PRL sont des boîtes à outils intéressantes. Cependant, Fairseq ne se concentre pas sur la parole mais sur les tâches génératives telles que de la traduction ou le résumé. Quant à S3PRL, la documentation est moindre et elle se concentre sur les méthodes auto-supervisées. De plus, ces deux approches proposent essentiellement l'apprentissage de modèles mais ne permet pas d'utiliser facilement des modèles pré-entraînés. Les propositions suivantes sont des approches qui permettent l'entraînement et l'utilisation de modèles pré-entraînés.

Vosk [14] est une solution de reconnaissance vocale dédiée à une utilisation hors ligne sur du matériel peu puissant comme des Raspberry Pi. D'après les personnes développant Vosk, celle-ci fournit la reconnaissance vocale pour les chatbots, les appareils ménagers intelligents et les assistants virtuels. Elle peut également créer des sous-titres pour des films ou des transcriptions pour des conférences et des entretiens. Vosk dispose de modèles dans une vingtaine de langues dont l'anglais, le français et l'allemand. Ces modèles sont mis à disposition mais sont très petits pour permettre une utilisation avec peu de ressources de calcul. Entraîner des modèles requiert la plupart du temps d'utiliser Kaldi. Il n'y a que 2 modèles français mis à disposition à l'heure actuelle. Un extrait de code illustrant une utilisation simple de Vosk est présenté figure A.1. Les modèles Vosk sont petits mais permettent la transcription continue d'un large vocabulaire, une réponse sans latence avec une IPA en continu, un vocabulaire reconfigurable et l'identification des personnes oratrices. Le déploiement d'un modèle Vosk demande la déclaration d'un modèle et de ses paramètres, ce qui est normal et même nécessaire pour permettre aux utilisateurs et utilisatrices un certain niveau de paramétrage. Cependant, pour obtenir le résultat écrit vis-à-vis d'un fichier audio, il faut soi-même détailler la façon d'alimenter le modèle avec le fichier. Quelque chose de plus simple serait de le mettre en paramètre d'appel du modèle. Vosk permet toutefois de générer l'horodatage, ce qui est, par exemple, très pertinent dans le cas de transcriptions d'entretien.

14. `https://alphacephei.com/vosk/`

ESPnet (Watanabe *et al.*, 2018; Inaguma *et al.*, 2020; Hayashi *et al.*, 2020) est une autre boîte à outils qui incorpore entre autre Asteroid et Whisper. Initialement axée sur la reconnaissance vocale de bout en bout et la synthèse vocale de bout en bout, ESPnet s'est étendue à diverses autres tâches de traitement de la parole. Les tâches possibles avec ESPnet sont nombreuses. On compte par exemple la traduction automatique de la parole, l'amélioration et séparation de la parole, la synthèse de la parole ainsi que la RAP. La version 1 de ESPnet requiert Kaldi et se repose dessus de manière générale. La version 2 [15] ne se base plus sur Kaldi. Celle-ci est encore en développement et certains outils implémentés n'obtiennent pas encore les mêmes performances que la version 1. Cependant, c'est celle qui est mise en avant aujourd'hui, notamment dans la documentation. ESPnet est d'ailleurs bien documentée et met à disposition des tutoriels. La version 2 s'utilise de manière générale presque comme la version 1, ce qui permet de ne pas trop dépayser les utilisateurs et utilisatrices qui travaillaient déjà avec avant. Lors du développement, il est nécessaire de déclarer le modèle et les paramètres utilisés. Contrairement à Vosk, il suffit de récupérer la sortie du modèle pour avoir le résultat.

NVIDIA NeMo (Kuchaiev *et al.*, 2019) est une boîte à outils accessible pour la recherche en apprentissage automatique et facilement évolutive. Elle fournit une collection d'outils en TAP et en TALN ainsi qu'en synthèse de la parole. C'est aussi le cas pour les modèles multimodaux et les grands modèles de langue. Il est donc possible d'utiliser NeMo pour de la RAP mais aussi pour corriger, traduire ou classifier les résultats obtenus par celle-ci. NeMo est libre d'accès et est développée par NVIDIA. Elle est construite pour pouvoir utiliser au maximum les ressources matérielles liées à l'apprentissage profond. Beaucoup de tutoriels sont mis à disposition pour la RAP ou le TALN [16]. Actuellement, cette boîte à outils peut être utilisée pour de la RAP, de la séparation de locuteurs et locutrices, de la reconnaissance d'orateur et de la classification de segments de parole. NeMo met en avant des modèles à l'état de l'art en RAP et la possibilité de les reprendre pour les affiner sur d'autres données. Elle est bien documentée et son utilisation est simple. Dans le cas de modèles pré-entraînés, il suffit de déclarer le modèle et les paramètres qui y sont liés. Ensuite, pour de la RAP comme dans nos exemples, une fonction prend en charge la transcription et nous renvoie directement le résultat. Il est important de noter que la fonction de transcription prend directement en paramètre une liste de fichiers audio. Ceci rend simple la transcription par lot, mais n'est pas

15. `https://espnet.github.io/espnet/espnet2_tutorial.html`

16. `https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/starthere/tutorials.html`

très intuitif dans le cas de transcriptions uniques. Ce n'est pas un inconvénient majeur mais il faut tout de même le garder en tête. Dans le cas de la figure A.3, aucun horodatage n'est généré.

Speechbrain (Ravanelli *et al.*, 2021) est un projet libre de droits et communautaire créé dans l'optique de faciliter la recherche mais aussi le développement de techniques et technologies de la parole. L'objectif de Speechbrain est « d'être simple, flexible, bien documenté et facilement utilisable. » Il se concentre particulièrement sur le TAP, ce qui fait de lui un candidat idéal. Avec Speechbrain il est possible de répondre à beaucoup de tâches liées au TAP. C'est le cas, notamment, pour la RAP, la reconnaissance de personnes locutrices, la séparation de personnes locutrices ou encore l'amélioration et la segmentation de la parole. Speechbrain est sorti en version 1.0 en février 2024 [17]. Cette version a apporté beaucoup de changements et d'améliorations dans l'environnement SpeechBrain. En 2024, on peut compter plus de 200 recettes différentes pour le TAP, ainsi que 100 modèles pré-entraînés et mis à disposition de la communauté. Parmi ceux-ci, on compte notamment Whisper que nous avons évoqué précédemment. Dans le cas de la figure A.4, nous déclarons deux modèles. Le premier permet de transcrire et le second permet d'aligner une transcription à un horodatage. Ensuite, nous appelons la fonction de transcription suivie de la fonction d'alignement. En quelques lignes, nous pouvons donc générer une transcription avec ou sans horodatage de manière assez simple.

PaddleSpeech (Zhang *et al.*, 2022) est une boîte à outils mettant à disposition une interface facile à utiliser. Elle peut être utilisée comme interface en ligne de commandes ou bien directement par le biais d'un langage de programmation. PaddleSpeech se concentre majoritairement sur le mandarin et l'anglais. Elle permet notamment de faire de la RAP, de la synthèse vocale, de la classification de segments audio et de la traduction de la parole. PaddleSpeech offre des modèles pré-entraînés, ainsi que des outils pour l'entraînement de nouveaux modèles sur des données personnalisées. Elle est conçue pour être facile à utiliser et efficace, offrant à la communauté de recherche un ensemble d'outils pour leurs projets de traitement de la parole. PaddleSpeech est aussi basée sur Kaldi pour certaines fonctionnalités. La documentation de PaddleSpeech est cependant complexe à naviguer. Ce n'est pas une documentation dédiée mais plutôt une liste de textes expliquant comment faire certaines choses. Il n'y a en effet pas vraiment de document expliquant les décisions et le rôle de chaque élément que l'outil propose. Quelques démonstrations, tutoriels et explications sont cependant disponibles. L'utilisation de PaddleSpeech est relativement simple. Il suffit de déclarer le

17. `https://colab.research.google.com/drive/1IEPfKRuvJRSjoxu22GZhb3czfVHsAy0s?usp=sharing`

modèle et d'appeler sa fonction pour transcrire. Puisqu'il est dédié en particulier au chinois et à l'anglais, PaddleSpeech perd en pertinence du fait que nous souhaitions couvrir un panorama plutôt multilingue.

Nous avons consigné les différentes tâches possibles dans la table 1.1 pour chacune des boîtes à outils précédemment citées. En termes de possibilités en TAP, SpeechBrain répond à beaucoup de critères. Pour le début du projet, nous avons surtout besoin de RAP, diarisation des locuteurs et possibilité d'horodatage. La plupart des boîtes à outils que nous avons citées sont capables de répondre à ce besoin. Cependant, Fairseq ne porte que très peu d'attention au TAP. S3PRL est limité en termes de possibilité d'apprentissage. Vosk est consacré aux appareils à faibles ressources computationnelles. PaddleSpeech se concentre principalement sur l'anglais et le chinois, ainsi que sur la traduction de l'un à l'autre.

ESPNet, NeMo et SpeechBrain sont toutes trois des options extrêmement intéressantes. ESPNet est la boîte à outils la plus difficile à prendre en main des trois. NeMo est simple à prendre en main mais a changé son point de concentration sur les grands modèles de langue et donc principalement le TALN. SpeechBrain est dédié entièrement au TAP et pousse ses efforts pour faciliter l'utilisation des ressources en TAP. SpeechBrain est aussi la boîte à outils la plus récente et la moins connue comme on peut le voir table 1.2. Contrairement à NeMo, ESPNet et SpeechBrain évoluent particulièrement grâce à la communauté et le support de certains partenaires. NeMo est entièrement soutenu par NVIDIA même s'il est libre et qu'il est possible de proposer des améliorations.

| | Vosk | ESPNet | NeMo | Speechbrain | PaddleSpeech | Fairseq | S3PRL |
|---|---|---|---|---|---|---|---|
| RAP | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| RAP en direct | | | ✓ | ✓ | ✓ | | |
| Reconnaissance du locuteur | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Diarisation des locuteurs | ✓ | ✓ | ✓ | | | | ✓ |
| Séparation de la parole | | ✓ | ✓ | ✓ | | | ✓ |
| Amélioration de la parole | | ✓ | ✓ | ✓ | | | |
| Traduction de la parole | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Traitement multi-microphone | | ✓ | ✓ | ✓ | | | |
| Compréhension de la parole | | ✓ | ✓ | ✓ | | | |
| Résumé de la parole | | ✓ | | | | | |
| Classification de parole/son | | | ✓ | ✓ | ✓ | | |
| Identification du langage | | | | ✓ | | | |
| Détection de l'intention | | | | ✓ | | | ✓ |
| Détection de sentiments | | | | ✓ | | | ✓ |
| Détection de silence | | | | ✓ | | | |

TABLE 1.1 – Récapitulatif des tâches possibles avec les diverses boîtes à outils.

| | ESPNet | NeMo | Speechbrain |
|---|---|---|---|
| Courbe d'apprentissage | Haute | Moyenne | Basse |
| Popularité GitHub | 7.7k | 9.6k | 7.6k |
| Date de sortie | 2018 | 2019 | 2021 |

TABLE 1.2 – Points d'intérêt des boîtes à outils (mis à jour en mars 2024).

Parmi toutes les boîtes à outils citées, nous faisons donc le choix de nous concenter sur ESPNet, NeMo et SpeechBrain. SpeechBrain et ESPNet sont équivalentes en termes de popularité et de tendance (confère 1.2 et 1.1). ESPNet est l'approche la plus ancienne et remplit la même fonction que SpeechBrain, à la différence que SpeechBrain offrent plus de possibilités. La popularité de NeMo

s'explique en partie grâce à la mise en avant des grands modèles de langue. Moitié 2023, ces trois outils avaient approximativement le même nombre d'étoiles sur GitHub.

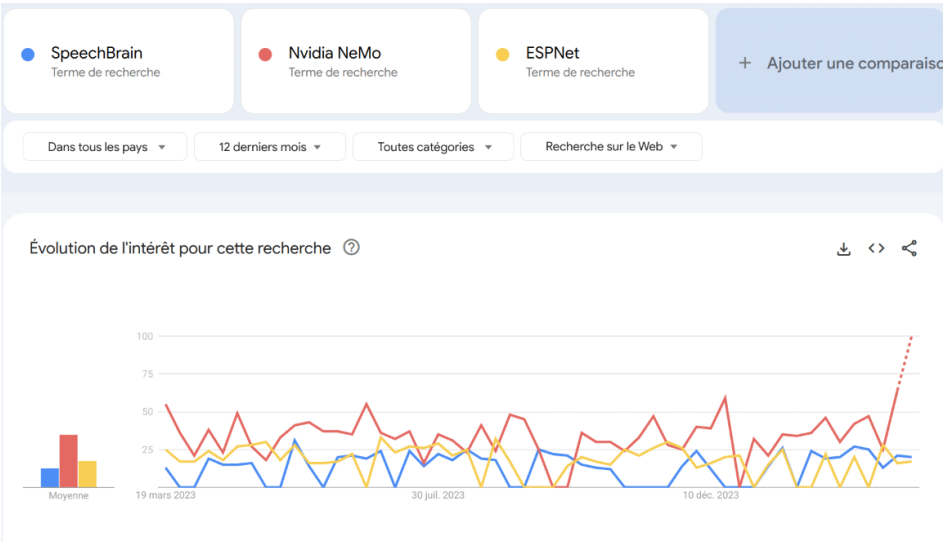


FIGURE 1.1 – Tendances Google concernant la fréquence de recherche de chaque boîte à outils sur les 12 derniers mois depuis mars 2024.

## 1.5 Applications proposées comme service

### 1.5.1 Applications libres et manuelles d'annotation d'audio

Explorer les approches non automatisées offre une occasion précieuse de mieux comprendre les exigences et les limitations de ces outils. En effet, même si notre objectif est de développer une solution entièrement automatisée, de nombreux aspects peuvent être transférés tels que les besoins, les fonctionnalités et les contraintes.

Praat (Boersma, 2001) est un logiciel conçu pour l'analyse approfondie de la parole et du langage. Il est largement utilisé par les professionnels, linguistes, phonéticiens, chercheurs en traitement du signal ou encore orthophonistes. Les fonctionnalités de Praat couvrent un large éventail d'analyses acoustiques, allant de l'analyse spectrale et temporelle à la manipulation avancée des signaux audio. Praat offre également des outils pour segmenter et annoter les enregistrements, facilitant ainsi l'analyse des différents aspects de la parole.

18. `https://www.ling.upenn.edu/courses/Fall_2019/ling001/PraatInstructions.html`

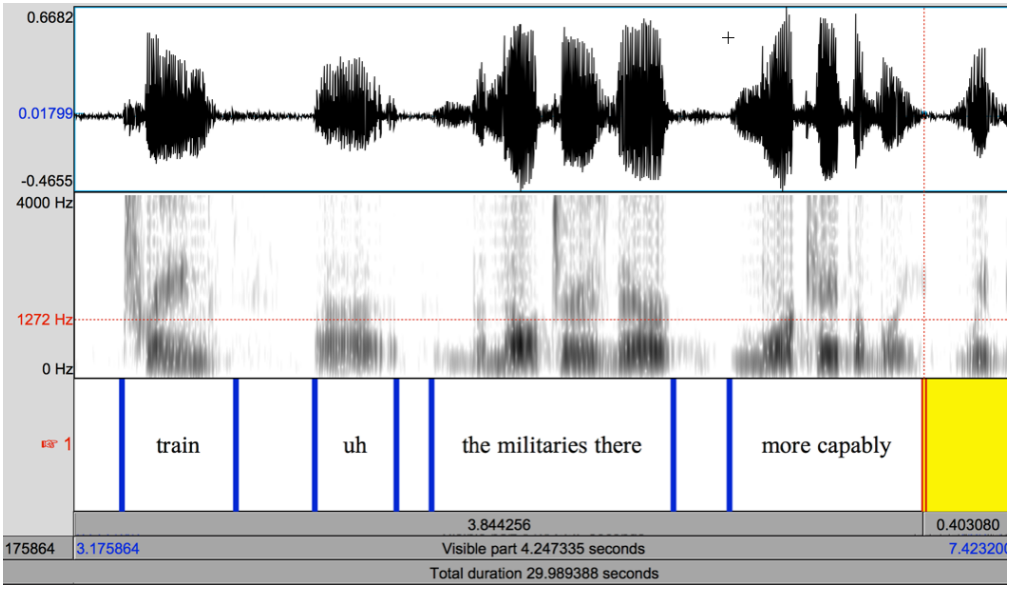


FIGURE 1.2 – Exemple d'utilisation de Praat [18].

Ces fonctionnalités avancées rendent Praat particulièrement adapté à une grande variété d'applications, telles que l'étude de la phonétique, la recherche sur les troubles de la parole, l'enseignement de la phonétique et bien d'autres encore. Praat est un outil puissant et polyvalent qui répond aux besoins de différents professionnels travaillant dans le domaine de la linguistique et de la parole, offrant des fonctionnalités avancées pour l'analyse approfondie de la parole et du langage.

Gecko (Levy *et al.*, 2019) est un outil d'annotation qui permet aux personnes utilisatrices de marquer et de commenter divers types de contenus, tels que des documents, des images ou des vidéos, avec une grande précision et flexibilité. Parmi ses principales caractéristiques, on trouve la capacité à ajouter différents types d'annotations, comme des surlignages et des commentaires textuels. Gecko permet également une collaboration en temps réel.

De plus, Gecko propose des outils avancés pour organiser et gérer les annotations, permettant de les filtrer, de les classer et de les rechercher facilement. Il peut être intégré à d'autres outils de gestion de contenu et de collaboration, ce qui en fait un choix populaire dans les environnements professionnels et éducatifs.

Label Studio (Tkachenko *et al.*, 2022) est un outil logiciel conçu pour faciliter l'annotation de données

19. `https://github.com/gong-io/gecko`

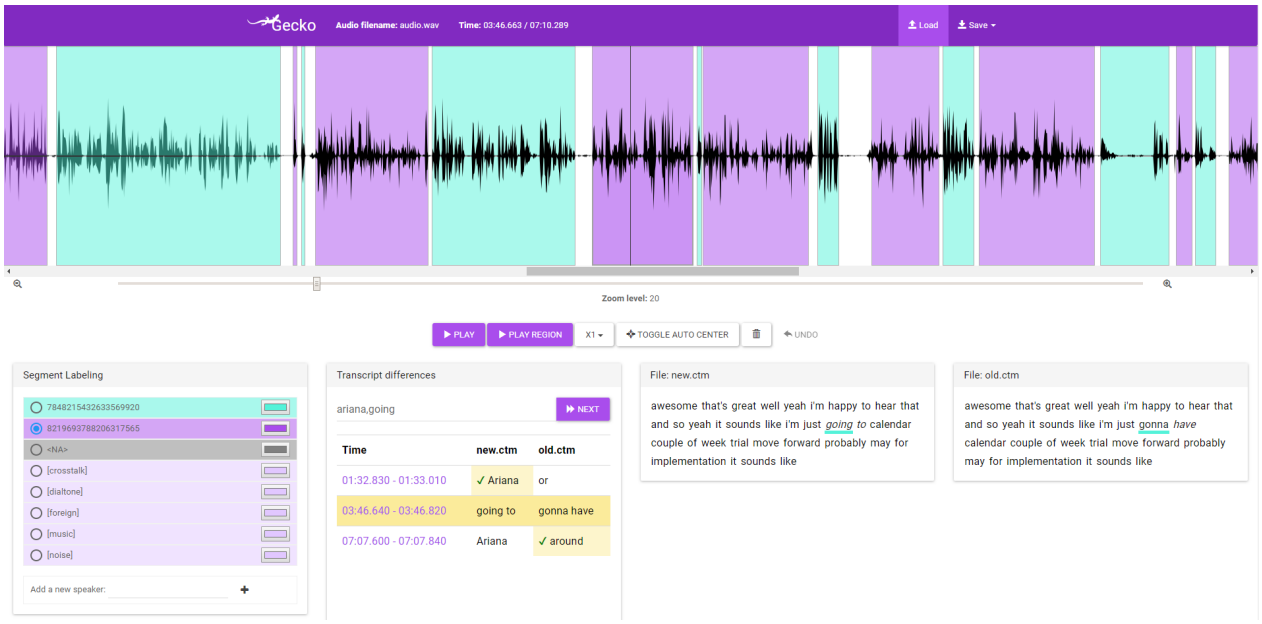


FIGURE 1.3 – Exemple d'utilisation de Gecko [19].

en vue de leur utilisation dans des projets d'apprentissage automatique. Cette plateforme offre une interface permettant d'annoter différents types de données, tels que des images, des vidéos, du texte, etc. Grâce à Label Studio, les personnes utilisatrices peuvent créer des schémas d'annotation personnalisés et collaborer avec d'autres annotateurs en temps réel, ce qui facilite la gestion et la synchronisation des annotations.

Label Studio offre une solution complète et flexible pour l'annotation de données, répondant aux exigences des chercheur.e.s, des développeur.e.s et des personnes professionnelles travaillant dans le domaine de l'apprentissage automatique et de l'analyse de données.

En résumé, de telles applications répondent généralement aux critères suivants :

- **Annotation de données** : capacité à annoter les données et à y ajouter des méta-données.
- **Personnalisation des schémas d'annotation** : possibilité de créer des schémas d'annotation personnalisés pour répondre aux besoins spécifiques des personnes utilisatrices.
- **Collaboration en temps réel** : fonctionnalité permettant à plusieurs annotateurs de travailler simultanément sur les mêmes données annotées.
- **Interface conviviale et intuitive** : interface conviviale facilitant l'annotation des données.

20. `https://github.com/HumanSignal/label-studio`

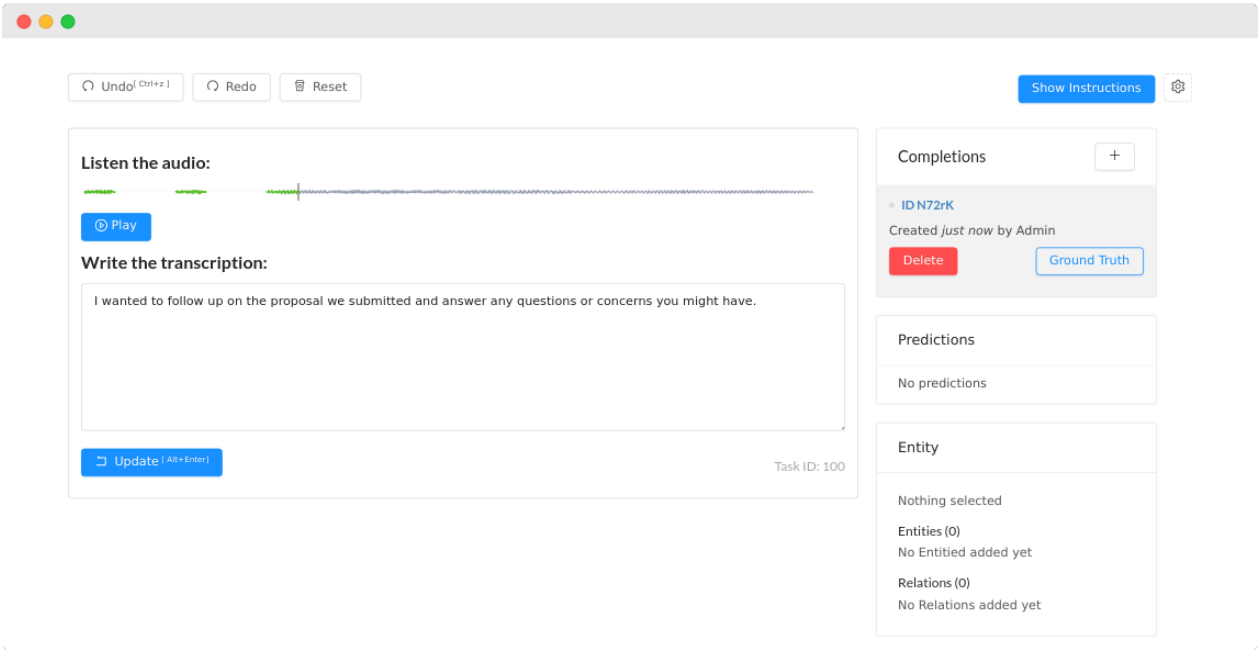


FIGURE 1.4 – Exemple d'utilisation de Label Studio [20].

— **Flexibilité et extensibilité** : capacité à personnaliser et à étendre l'outil pour s'adapter à divers cas d'utilisation.

— **Sécurité et confidentialité** : intégration de fonctionnalités de sécurité avancées pour protéger la confidentialité des données annotées.

### 1.5.2 Applications de transcription automatique

Il existe plusieurs solutions prêtes à l'emploi pour répondre à la nécessité de transcrire des enregistrements audio ou des conversations, tant pour les particuliers que pour les entreprises qui ne souhaitent pas développer leur propre système de transcription. Certaines de ces solutions sont facilement accessibles via une interface web conviviale, offrant une gamme étendue de fonctionnalités de correction et d'édition. D'autres, quant à elles, requièrent un minimum de compétences en développement et sont disponibles via une IPA.

Certaines sociétés fournissant ces applications offrent également des services de transcription effectués par des professionnels. Cependant, notre intérêt se porte sur la transcription automatique, nous ne détaillerons donc pas ce point davantage.

Nuance Dragon [21] représente une solution de pointe pour la reconnaissance vocale en temps réel, offrant la possibilité de transcrire efficacement des fichiers audio en texte, que ce soit individuellement ou en lot. Cette technologie est largement adoptée dans divers domaines tels que la santé, le droit et le domaine académique où la dictée et la transcription automatisée sont essentielles. En outre, Nuance Dragon présente des avantages significatifs en matière d'accessibilité, offrant un support précieux aux personnes ayant des besoins spécifiques.

NVivo Transcription [22] est une solution de transcription audio vers texte, permettant une conversion précise et efficace des fichiers audio en documents textuels. Outre la transcription automatique, cette plateforme propose des fonctionnalités d'édition collaborative visant à améliorer la précision des transcriptions. De plus, les transcriptions générées peuvent être annotées, analysées et exportées vers le logiciel NVivo pour une utilisation ultérieure dans des études de recherche.

Nuance Dragon et NVivo Transcription sont deux applications qui peuvent produire de la transcription automatique en local, ce qui signifie que les données restent avec la personne utilisatrice.

Les solutions telles que Vook.ai [23], Amberscript [24], Trint [25], HappyScribe [26], Descript [27], Authôt [28], et Verbit [29] sont des solutions de transcription à portée de main rapides et proposant seulement des services de transcription avec, pour la plupart, des services de correction automatique ou manuelle. Ces solutions sont toutes certifiées RGPD et/ou ISO 27001.

Rev [30] propose de la transcription ainsi que de la séparation de locuteurs et locutrice. Il offre la possibilité de corriger directement dans un éditeur. Ce service met aussi à disposition les horodatages.

21. https://www.nuance.com/fr-fr/dragon.html
22. https://lumivero.com/products/nvivo-transcription/
23. https://www.vook.ai/en
24. https://www.amberscript.com/fr/
25. https://trint.com/fr/home
26. https://www.happyscribe.com/
27. https://www.descript.com/transcription
28. https://authot.com/
29. https://verbit.ai/
30. https://www.rev.com/services/auto-audio-transcription

Otter [31] est une application d'enregistrement et de résumé de réunions. Elle peut ainsi produire à partir d'une réunion la transcription et mettre en avant les éléments clefs évoqués. Otter produit la transcription en temps réel. Elle peut aussi être intégrée à d'autres applications comme Zoom ou Microsoft Teams. Il ne semble pas possible d'utiliser Otter pour transcrire des fichiers en grosse quantité. Sonix [32] est une interface web qui permet de transcrire et proposes diverses fonctionnalités de correction et d'édition du résultat. Elle permet aussi de créer des résumés ainsi que de collaborer avec d'autres personnes. Comme Otter, Sonix peut être intégré à d'autres applications. Speechmatics [33] propose aussi un service de transcription en temps réel ainsi que de résumé. Elle propose aussi la possibilité de créer un dictionnaire personnalisé ainsi que de la transcription en grosse quantité si besoin.

f4x [34] est un service de transcription qui propose aussi des fonctionnalités d'horodatage et de détection de silence. Il est aussi possible de personnaliser un dictionnaire.

Scribe [35] est un projet basé sur le projet Vosk. Le code source de Scribe est sous licence libre et celui-ci permet seulement de transcrire des fichiers vidéo ou audio en trois langues. Il est possible de créer sa propre instance puisque le code est libre. Temi [36] simplifie le processus de transcription en récupérant les fichiers audio, en les transcrivant et en renvoyant la transcription par courriel. De plus, il offre la capacité d'identifier les locuteurs dans les enregistrements audio.

Dictation [37] est un service gratuit de transcription audio en temps réel qui fonctionne comme un outil de dictée. Il nécessite l'utilisation du navigateur web Google Chrome, utilisant les services de transcription de Google pour fournir des transcriptions précises et instantanées.

31. https://otter.ai/
32. https://sonix.ai/
33. https://www.speechmatics.com/
34. https://www.audiotranskription.de/en/f4x/
35. https://scribe.cemea.org/
36. https://www.temi.com/
37. https://dictation.io/

Amazon Transcribe [38], Google Speech to Text [39], Microsoft Azure Speech to Text [40] IBM Watson Speech to Text [41], Deepgram [42] et AssemblyAI [43] sont des services payants plus personnalisables que les solutions précédemments citées. Cependant, ils requièrent plus de compétences techniques. De même, Whisper [44] est un modèle de RAP développé par OpenAI. Bien qu'il soit possible de l'utiliser en local sur sa propre machine, la solution principale offerte est l'utilisation de l'IPA fournie par OpenAI. Cette approche est similaire à celle des autres solutions disponibles sur le marché.

Le tableau 1.3 présente un récapitulatif des différentes applications mentionnées précédemment, mettant en avant les tarifs proposés ainsi que le nombre de langues que chaque entreprise affirme prendre en charge.

38. https://aws.amazon.com/fr/transcribe/

39. https://cloud.google.com/speech-to-text?hl=en

40. https://azure.microsoft.com/en-us/products/ai-services/speech-to-text

41. https://www.ibm.com/products/speech-to-text

42. https://deepgram.com/

43. https://www.assemblyai.com/

44. https://openai.com/research/whisper

TABLE 1.3 – Aperçu des applications de transcription automatique - types P(propriétaire) et L(ibre).

| | **Type** | **Utilisation** | **Langues** | **Coût** |
|---|---|---|---|---|
| AmberScript | P | Interface web | Français, Anglais et 37 autres langues | 10$/h Abonnement 40$/mois |
| Nuance Dragon | P | Application | Français, Anglais et 13 autres langues | 999$ la licence 499$ la mise à jour |
| Nvivo Transcription | P | Application | Français, Anglais et 40 autres langues | 12-37,50$/heure |
| Trint | P | Interface web | Français, Anglais et 40 autres langues | Abonnement 50$/mois |
| HappyScribe | P | Interface web | Français, Anglais et 64 autres langues | 0,15-0,2$/min |
| Vook.ai | P | Interface web | Français, Anglais et 4 autres langues | 3$/h |
| Otter.ai | P | Interface Web | Anglais | Abonnement 12,50-37,50$/mois |
| Dictation | P | Interface Web | Français, Anglais et 123 autres langues | Gratuit |
| Authôt | P | Interface Web | Français, Anglais et 30 autres langues | 0,1$/min |
| Temi | P | Interface web | Anglais | 0,25$/min |
| Verbit | P | Interface web | Anglais et Espagnol | Personnalisé |
| Descript | P | Interface web | Français, Anglais et 20 autres langues | 12-24$/mois pour 10-30 heures |
| f4x | P | Application | Français, Anglais et 18 autres langues | 7,33-20,63$/heure |
| Rev | P | Interface web | Français, Anglais et 53 autres langues | 0,25$/min Abonnement 29,99$/mois |
| Sonix | P | Interface web | Français, Anglais et 40 autres langues | 10$/heure |

| | Type | Utilisation | Langues | Coût |
|---|---|---|---|---|
| Speechmatics | P | Interface web | Français, Anglais et 48 autres langues | 0,30-1,35$/heure |
| AssemblyAI | P | IPA | Français, Anglais et 18 autres langues | 0,0043-0,0145$/min |
| Deepgram | P | IPA | Français, Anglais et 30 autres langues | 0,37$/heure |
| IBM Watson Speech to Text | P | IPA | Français, Anglais et 10 autres langues | 0,013-0,27$/min |
| Microsoft Azure Speech to Text | P | IPA | Français, Anglais et 137 autres langues | 0,485-1,615$/heure |
| Amazon Transcribe | P | IPA | Français, Anglais et 123 autres langues | 0,01875$/min |
| Google Speech to Text | P | IPA | Français, Anglais et 123 autres langues | 0,024$/min |
| Whisper | P | IPA | Français, Anglais et 100 autres langues | 0,006$/min |
| Scribe | L | Interface web | Français, Anglais et Espagnol | Gratuit |

Certaines de ces solutions ont été évaluées avec rigueur. Louw (2021) met en avant quelques comparaisons entre Sonix et Otter appliqués à 5 audio anglais de qualités différentes, enregistrés dans des contextes différents et par des orateurs de nationalités différentes. L'auteur ne précise pas clairement ce qu'il entend par une erreur et en particulier ce qui est défini comme une "presque correspondance", c'est-à-dire une erreur dont le sens peut être facilement déduit. Pour détecter ces erreurs, l'auteur dit utiliser un outil [45] et compter le nombre d'erreurs. Cet outil est un outil de détection des différences qui ne permet pas de produire de métrique.

Néanmoins cette étude est intéressante pour pouvoir avoir un aperçu de la qualité de transcription produite par Otter et Sonix en 2020 et 2021 (confère table 1.4). Chacun des audio est d'une durée

45. `https://text-compare.com/`

de 3 mins. La longueur du texte n'est pas précisée pour chaque audio mais on peut estimer une moyenne entre 400 et 500 mots.

| | Otter | Sonix |
|---|---|---|
| Audio 1 (monologue) | **0,3250** | 0,4000 |
| Audio 2 (entretien) | **0,4653** | 0,6237 |
| Audio 3 (entretien - bruit) | **0,6117** | 0,7177 |
| Audio 4 (entretien) | **0,5267** | 0,6072 |
| Audio 5 (discussion à 3) | **0,7430** | 0,7569 |

TABLE 1.4 – Taux d'erreur de Otter et Sonix (Louw, 2021).

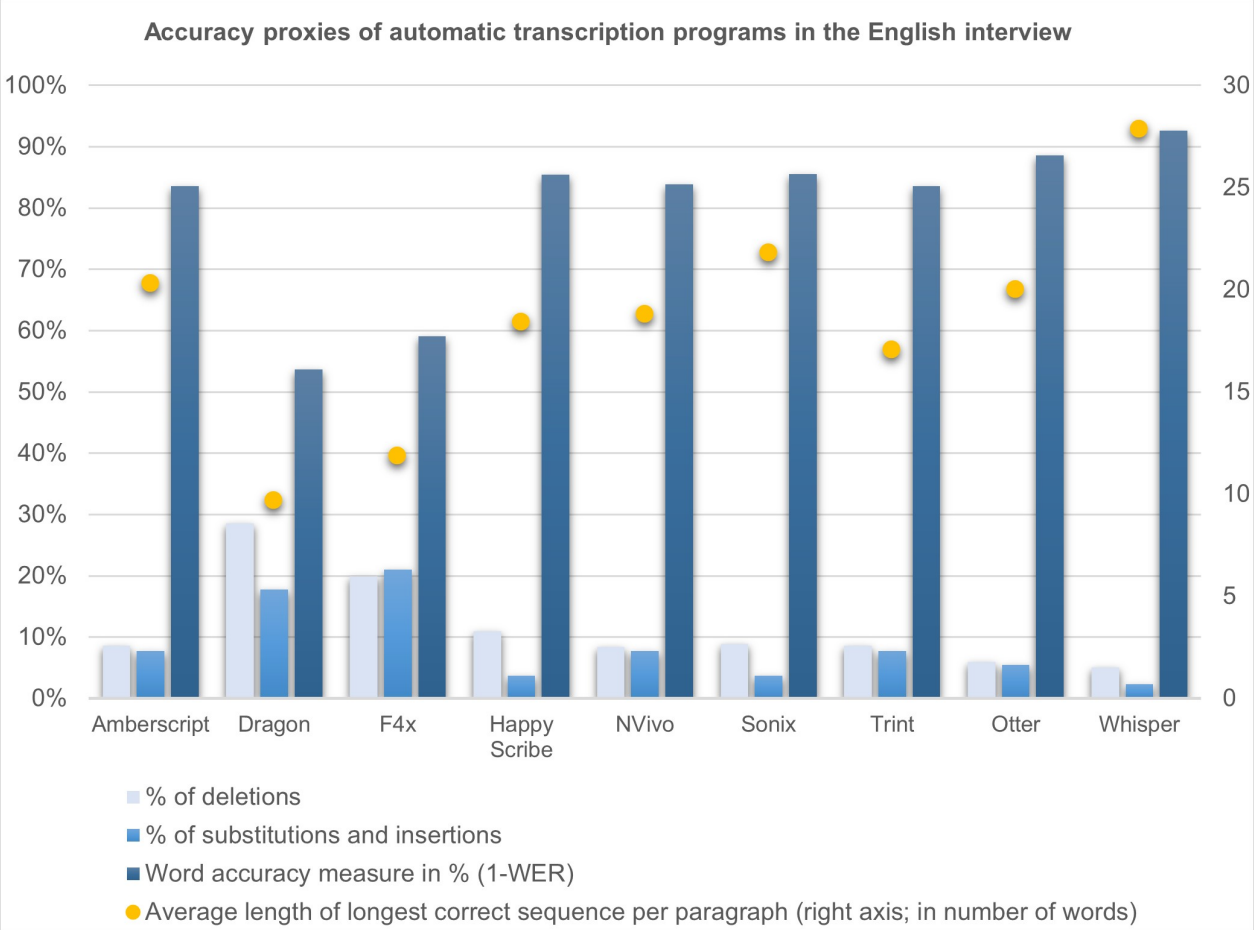


FIGURE 1.5 – Les indicateurs de précision des programmes de transcription automatique dans l'entretien en anglais, source : (Wollin-Giering *et al.*, 2024).

En 2021, Otter était plus performant que Sonix sur ces différents types d'audio en anglais. L'étude

(Wollin-Giering *et al.*, 2024) couvre une plus grande quantité de solutions et semble confirmer les résultats cités précédemment bien que la différence entre Otter et Sonix soit plus faible. Cette étude ne met pas à disposition les résultats exacts mais permet de constater qu'en anglais Otter, Trint, Sonix, NVivo Transcription, Happy Scribe et Amberscript ont un taux d'erreur sur les mots (TEM) [46] entre 0.1 et 0.2. F4x et Dragon sont respectivement à 0.41 et 0.46. Enfin, Whisper a un TEM de 0.07.

L'entretien sur lequel les outils sont testés est constitué de 23 paragraphes où deux personnes s'expriment alternativement. En moyenne, chaque paragraphe compte 34 mots, les plus courts n'en comportant qu'un ou deux (principalement des réponses telles que "mhm"), tandis que le paragraphe le plus long contient 284 mots. Le clip de 5 minutes de l'entretien contient au total 787 mots.

Pareillement, l'étude (Loakes, 2024) porte sur la qualité des transcription lorsque les audio sont de bonne et mauvaise qualité. Le résumé est disponible dans le tableau 1.5.

| Système | Audio de bonne qualité | Audio de mauvaise qualité |
|---|---|---|
| Descript | **0** | 0,7330 |
| Sonix | 0,0400 | 0,8280 |
| Amazon Transcribe | 0,0400 | 0,9060 |
| Microsoft Azure | 0,0400 | 0,8540 |
| Google Cloud | 0,1800 | N/A |
| Assembly AI | **0** | 0,8190 |
| Deepgram | 0,1200 | 0,8800 |
| Whisper | 0,0800 | **0,5090** |

TABLE 1.5 – Taux d'erreur de certaines applications dans le cadre d'audio de bonne et mauvaise qualité.

En date de mars 2024, les outils les plus recherchés parmi ceux que nous avons cités sont Rev, Whisper, Sonix, Trint et Happy Scribe d'après Google Trends [47]. Otter est lui aussi populaire mais est plus spécialisé. En termes de performance, Whisper est la solution la plus optimale si le but est

46. Le TEM est expliqué en détails section 2.2

47. `https://trends.google.fr/trends/explore?q=Trint,Sonix,Rev,Whisper,HappyScribe&hl=fr`

seulement d'avoir une transcription. Les autres outils peuvent cependant être utiles pour d'autres situations et notamment d'autres besoins, comme pour de la diarization ou de la correction avec interface et connexion du texte à l'audio.

Dans notre situation, nous recherchons un outil capable de transcrire, tout en ayant la capacité de reconnaître les locuteurs à terme et de répondre à d'autres besoins contextuels. Nous attachons également une grande importance à la sécurité et à la confidentialité des données. Bien que les applications mentionnées précédemment respectent les normes et les lois de confidentialité, le fait que ces solutions soient pour la plupart en ligne expose potentiellement à des vulnérabilités.

# CHAPITRE 2
# DONNÉES, MÉTRIQUES ET PERFORMANCES

Ce chapitre présente des corpus disponibles pour la recherche en TAP ainsi que les métriques les plus utilisées avec leur signification, leurs avantages et inconvénients. Dans le cadre de ce mémoire, nous nous concentrons sur les corpus en anglais et en français. Nous détaillons aussi les performances de quelques boîtes à outils et outils présentés dans le chapitre précédent.

## 2.1 Données

### 2.1.1 Les corpus de données accessibles

LibriSpeech (Panayotov *et al.*, 2015) est une collection de données largement exploitée dans le domaine de la recherche en reconnaissance automatique de la parole (RAP). Celle-ci a été publiée en 2015 et est disponible au téléchargement et à l'utilisation suivant la licence « Creative Common ». Composé d'environ 1 000 heures de discours en anglais extraits de livres audio, l'ensemble de données LibriSpeech est organisé en plusieurs parties comprenant des enregistrements provenant de différents livres et narrateurs. Cette diversité permet de couvrir un large éventail de styles de discours, d'accents et de contextes linguistiques, ce qui en fait un ensemble de données adapté à différents scénarios d'application.

Mozilla Common Voice (MCV) (Ardila *et al.*, 2020) est un ensemble de données vocales multilingues en libre accès. Il suit la licence « Mozilla Public Licence 2.0 » qui autorise l'accès et l'utilisation à toutes et tous. La première version du MCV a été publiée en février 2019. Il est collecté auprès de volontaires qui lisent à haute voix des phrases fournies par le projet. Les données sont disponibles dans plusieurs langues et ont contribué au développement de modèles RAP pour diverses langues. Le projet est complètement ouvert, que ce soit pour partager des segments audio ou réaliser des transcriptions [48]. Il est aussi possible de télécharger directement le corpus [49]. Une version du corpus est publiée tous les 3 mois. À date de mars 2024, MCV contient 30 329 heures de segments audio parmi 120 langues différentes. 19 916 heures sont confirmées comme valide par rapport à leur version

48. `https://commonvoice.mozilla.org/en`

49. `https://commonvoice.mozilla.org/en/datasets`

écrite. Pour l'anglais, 3 438 heures sont disponibles (2 585 heures validées) issues de 90 474 voix différentes. Pour le français, 1 113 heures sont disponibles (989 heures validées) issues de 18 487 voix différentes. Il y a 17 versions de MCV. Chaque itération rajoute des heures d'audio dans de multiples langues.

VoxPopuli (Wang *et al.*, 2021b) est un corpus de parole multilingue à grande échelle qui fournit 400 000 heures de données vocales non étiquetées pour 23 langues, 1 800 heures de données vocales transcrites pour 16 langues et 17 300 heures de transcription puis traduction de la parole d'une langue vers une autre. Les données brutes sont collectées à partir d'enregistrements d'événements du Parlement européen de 2009 à 2020. L'ensemble de données comprend des discours transcrits dans 15 langues différentes. Celui-ci a été publié en mars 2021 sous la licence « Creative Commons ».

TED-LIUM 3 (Hernandez *et al.*, 2018) est une collection de conférences TED et de leurs transcriptions disponibles sur le site web TED. Elle a été préparée pour entraîner des modèles acoustiques pour participer à l'atelier international sur la traduction de la parole en 2011. Elle a été publiée sous licence « Creative Commons » en septembre 2018. Celle-ci contient 2 351 enregistrements audio de conférences TED, ainsi que leurs transcriptions. Ceci correspond à 452 heures d'audio en anglais. Cet ensemble de données peut aussi être utilisé pour de la transcription de longs fichiers. Il suffit alors de concaténer les extraits d'un même TED.

FLEURS (Conneau *et al.*, 2023) est un ensemble de données dérivé du corpus FLORES-101 (Goyal *et al.*, 2022). Ces deux corpus sont libres d'accès et d'utilisation suivant la licence « Creative Common »et ont été publiés respectivement en mai 2022 et juin 2021. Le corpus FLORES-101 permet l'évaluation de la traduction automatique à l'écrit. Celui-ci est constitué de 3 001 phrases anglaises traduites dans 101 langues. FLEURS est la version vocale de celui-ci, il contient approximativement 12h d'audio pour 102 langues. Ce qui le rend très intéressant pour de la RAP et de la traduction d'une langue vers une autre. Notamment dans le cadre de langues à faibles ressources.

Meanwhile [50] est un corpus de données publiés en septembre 2023 dans le cadre de Whisper (Radford *et al.*, 2022). Celui-ci est libre d'accès et d'utilisation. Il est constitué de 64 segments en anglais de *The Late Show with Stephen Colber*. La durée de ces segments est entre 30 secondes et 2 minutes.

50. `https://huggingface.co/datasets/distil-whisper/meanwhile`

Cet ensemble de données permet d'évaluer les résultats de RAP sur de longs extraits.

Rev16 (Radford *et al.*, 2022) est un sous-échantillon de l'ensemble de données en anglais publié par Rev [51] en février 2019. Rev est libre d'accès et d'utilisation. D'après (Radford *et al.*, 2022), un certain nombre d'extraits sont erronés. De ce fait Rev16 est une sélection des fichiers nommés :

3 4 9 10 11 14 17 18 20 21 23 24 26 27 29 32

Kincaid46 [52] contient 46 extraits longs en anglais. Ces extraits proviennent de Youtube. Ces derniers ont été compilés par Jason Kincaid pour réaliser une étude indépendante des outils disponibles sur le marché. Kincaid46 a été publié en septembre 2018. Il est libre d'accès et d'utilisation.

Earnings-21 (Del Rio *et al.*, 2021) est un corpus de 39 heures de conversations à propos de finance provenant de neuf secteurs financiers différents. Il contient des discours riches en entités (comme les noms de sociétés, les chiffres financiers, etc.). Il a été publié en juin 2021. Earnings-22 (Rio *et al.*, 2022) totalise 119 heures de conversations à propos de finance en anglais provenant de sociétés mondiales. Il a été publié en avril 2022. Ces deux ensembles de données ne sont pas composés des mêmes données et sont disponibles à l'usage [53] sous licence « Creative Commons ». Earnings-21 et Earnings-22 sont constitués d'extraits de 30 minutes à 1 heure 30.

Le *Corpus of Regional African American Language* (CORAAL) (Kendall et Farrington, 2023) est le premier corpus public de données sur l'anglais afro-américain. Il comprend les enregistrements audio ainsi que les transcriptions alignées dans le temps de 231 entretiens sociolinguistiques avec des locuteurs nés entre 1888 et 2005. Ces enregistrement audio sont tous d'une durée supérieure à 15 minutes. Cet ensemble de données est gratuit pour une utilisation dans le cadre de la recherche. Il a été publié en janvier 2018 puis mis à jour continuellement jusqu'à sa version actuelle qui date de juin 2023.

51. https://www.rev.com/blog/media-and-entertainment/podcast-transcription-benchmark-part-1

52. https://medium.com/descript/which-automatic-transcription-service-is-the-most-accurate-2018-2e859b23ed19

53. https://github.com/revdotcom/speech-datasets

### 2.1.2 La Commission Viens

Il est également envisageable de générer des ensembles de données à partir de données brutes collectées préalablement. Dans notre situation, l'obtention de corpus en français québécois serait particulièrement pertinente, permettant ainsi de perfectionner des modèles déjà existants.

Suite à la révélation de pratiques potentiellement discriminatoires envers les Autochtones dans les services publics au Québec, le gouvernement québécois a mis en place la Commission Viens, une initiative visant à enquêter sur ces questions et à améliorer les relations avec les Autochtones. Cette Commission vise à examiner les enjeux systémiques et formuler des recommandations pour remédier à la violence, à la discrimination systémique, et aux traitements disparates. Philippe Couillard, premier ministre du Québec de 2014 à 2018 soulignait l'urgence d'agir pour restaurer la confiance. Il promouvait des valeurs de tolérance, d'ouverture et de dignité afin de reconstruire l'harmonie entre les communautés autochtones et l'ensemble de la société québécoise [54]. La commission Viens porte le nom du juge qui l'a présidée, l'Honorable Jacques Viens. Son nom complet est « Commission d'enquête sur les relations entre les Autochtones et certains services publics au Québec : écoute, réconciliation et progrès ».

La commission Viens est constitué d'audiences, d'entretiens et de témoignages de la population autochtone au Québec. Certaines audiences sont en anglais et certaines interventions dans la langue de l'intervenant. Dans un premier temps, nous souhaitons nous concentrer sur le français québécois mais il n'est pas exclu d'utiliser ces données à terme pour permettre de mieux représenter les peuples autochtones. Après étude [55] plus précise des segments audio et transcriptions de cette Commission, nous savons que près de 658 heures d'audio sont disponibles. Ces 658 heures d'audio sont réparties sur 173 jours non consécutifs entre juin 2017 et décembre 2019. Chaque audience est transcrite et suit un format similaire à celui présenté dans la figure 2.1.

Lorsqu'il s'agit de gérer certains cas spéciaux, comme une déclaration sur honneur ou l'émission d'une pièce à conviction, les sténographes n'ont pas tous la même façon de formatter ceux-ci. Néanmoins, le reste des transcriptions suit une nomenclature fixe et bien représentée par la figure 2.1.

54. `https://www.cerp.gouv.qc.ca/index.php?id=3&L=12`

55. Cette étude a été réalisée avec Lucas Maison, chercheur au Laboratoire informatique d'Avignon (LIA), dans l'optique de créer un corpus de données se concentrant sur le Français québécois. Un article est en cours d'écriture.

**OUVERTURE DE LA SÉANCE**

**LA GREFFIÈRE-AUDIENCIÈRE:**

La Commission d'enquête sur les relations entre les Autochtones et certains services publics au Québec, présidée par l'Honorable Jacques Viens, est maintenant ouverte.

**L'HONORABLE JACQUES VIENS (LE COMMISSAIRE):**

Alors, bonjour. Bienvenue en cette autre journée de nos audiences à Val-d'Or au Conservatoire de musique en territoire anichinabé. Je vais commencer par demander aux procureurs de s'identifier pour les fins de l'enregistrement.

**Me EDITH-FARAH ELASSAL,**

**PROCUREURE POUR LA COMMISSION VIENS:**

Oui. Bonjour, Monsieur le Commissaire. Édith-FARAH Elassal pour la Commission.

FIGURE 2.1 – Exemple de fichier de transcription.

Au total, nous récupérons 173 audiences, chacune scindée en plusieurs extraits audio dont la durée varie entre quelques minutes et plusieurs heures. Ces audiences peuvent être scindées pour des raisons diverses telles qu'une pause ou un problème technique. La distribution de la longueur des extraits audio est illustrée en figure 2.2. Il y a au total 625 extraits d'audio.

L'objectif est de créer un jeu de données sur lequel il est possible d'entraîner des modèles de RAP. Ces extraits étant de durée variable, nous avons décidé de les couper en segments représentant une ligne dans la transcription, ou approximativement 25 mots dans le cas d'un discours. Ils sont aussi coupés par locuteur, il y a donc seulement un locuteur ou une locutrice par segment. Cela permet aussi d'entraîner des modèles de RAP sans risquer de dépasser la mémoire disponible à cause de fichiers trop longs. Nous avons créé ainsi approximativement 344 000 segments audio qu'il faut aligner avec leur équivalent transcrit. De ces 344 000 segments, nous en ignorons 23% (soit près de 80 000) car ils sont en anglais et nous nous sommes concentrés sur le français.

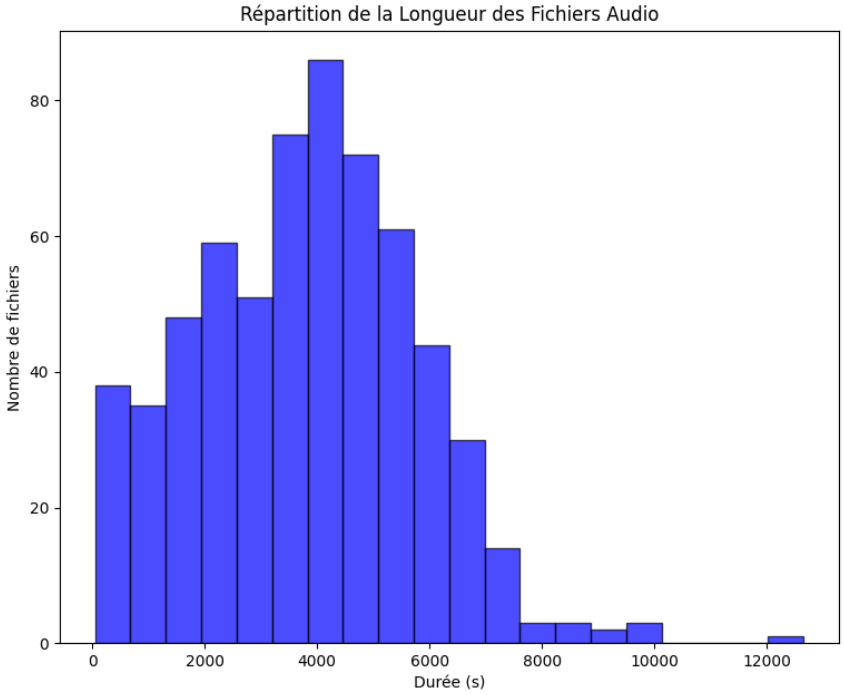


FIGURE 2.2 – Répartition des extraits audio de la Commission Viens en fonction de leur durée (figure créée par Lucas Maison).

Pour forcer l'alignement entre les segments audio et la transcription, nous avons utilisé Aeneas [56] pour les premières passes. Aeneas permet de faire l'alignement de manière très rapide, cependant nous avons remarqué lors du traitement que beaucoup des segments étaient mal alignés. Ceci nous a poussé à utiliser Montreal Forced Aligner (MFA) (McAuliffe *et al.*, 2017) sur les passages les plus difficiles. Le MFA est un outil très puissant et précis. Nous avions cependant évité de l'employer car il utilise l'algorithme de Viterbi, qui permet de trouver les séquences les plus plausibles. L'algorithme de Viterbi demande de paramétrer un « espace de recherche ». Puisque certains fichiers sont extrêmement longs, cela signifie un espace de recherche très gros et donc un temps de traitement très long. Par exemple, pour 45 minutes d'audio, MFA prend plusieurs heures alors qu'Aeneas prend quelques minutes. Sur 658h, cela représenterait près de 292 jours de traitement. Les détails à propos de la création du corpus sur la Commission Viens seront publiés dans un prochain article.

56. `https://github.com/readbeyond/aeneas`

### 2.1.3 La Commission Charbonneau

La Commission Charbonneau est une Commission québecoise d'enquête sur l'octroi et la gestion des contrats publics dans l'industrie de la construction. Celle-ci fut mise en place par le gouvernement provincial en octobre 2011 et présidée par la juge France Charbonneau. L'objectif était d'examiner les pratiques de collusion et de corruption dans l'industrie de la construction, afin de déterminer les liens potentiels avec certains partis politiques et le crime organisé. La Commission a produit approximativement 93 giga octets de données.

Générer un ensemble de données à partir des enregistrements audio et des transcriptions permettrait de créer une représentation assez diverse du français québécois. Il s'agit principalement d'auditions de témoins mais aussi de quelques experts et représentants d'organismes. Il pourrait donc être intéressant de cumuler les données de la commission Viens avec celles de la commission Charbonneau pour avoir un corpus très varié de français québécois.

Le travail sur la commission Charbonneau n'a pas encore débuté.

## 2.2 Métriques

### 2.2.1 Taux d'erreur

Le taux d'erreur sur les mots (TEM), (Word Error Rate (WER) en anglais), permet d'évaluer la précision d'un système de reconnaissance de la parole en comparant la sortie du système (les mots reconnus) avec la transcription correcte (les mots réellement prononcés). On applique celui-ci pour comparer un document de référence à un autre document. Ces documents peuvent être des phrases simples ou bien des ensembles de phrases (comme un paragraphe). La formule du TEM est définie comme suit :

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

où :

— $S$ représente le nombre de substitutions (mots incorrects),

— $D$ représente le nombre de suppressions (mots manquants dans la transcription générée),

— $I$ représente le nombre d'insertions (mots en trop dans la transcription générée),

— $N$ représente le nombre total de mots dans la transcription de référence.

Le TEM examine les différences entre les deux séquences de mots. Plus précisément, on calcule le coût minimum d'opérations de substitution, suppression et insertion nécessaires à retrouver la séquence correcte à partir de la séquence fournie. Pour ce faire, un coût élémentaire de 1 est assigné à chacune des opérations. La séquence optimale d'opérations à effectuer pour minimiser ce coût est trouvée à l'aide de l'algorithme de Levenshtein (Levenshtein, 1965).

En d'autres termes, le TEM est calculé à partir de la distance de Levenshtein entre les séquences avec un coût uniforme de 1 pour chaque opération. Il est souvent exprimé en pourcentage pour faciliter l'interprétation. Le TEM peut aller de 0 à l'infini, signifiant une infinité d'erreurs. Un TEM plus bas indique une meilleure performance du système, car cela signifie moins d'erreurs par rapport à la transcription de référence.

Considérons les phrases de référence et d'hypothèse :

**R :** J'aime les petits chats

**H :** J'adore les chats noirs

| | | | | | |
|---|---|---|---|---|---|
| R : | J'aime | les | petits | chats | - |
| H : | J'adore | les | chats | noirs | - |

— **Substitutions (S)** : 1 ("J'adore" remplace "J'aime") ;

— **Insertions (I)** : 1 (le mot "noirs" est inséré) ;

— **Suppressions (D)** : 1 (le mot "petits" est supprimé) ;

— **Nombre de mots dans la référence (N)** : 4.

Le TEM se calcule comme suit :

$$TEM = \frac{S + D + I}{N} = \frac{1 + 1 + 1}{4} = \frac{3}{4} = 0.75$$

Pour illustrer un TEM supérieur à 1, considérons une phrase hypothétique modifiée :

**Nouvelle Phrase Hypothétique :** "J'adore les grands chats noirs et mignons"

- **Insertions (I)** : 3 (les mots "noirs", "et" et "mignons" sont insérés) ;
- **Substitutions (S)** : 2 ("J'adore" remplace "J'aime", "grands" remplace "petits") ;
- **Suppressions (D)** : 0.

$$TEM = \frac{S + D + I}{N} = \frac{1 + 1 + 3}{4} = \frac{5}{4} = 1,25$$

Ainsi, le TEM est de 1,25, ce qui indique plus d'erreurs que de mots dans la phrase de référence.

Le TEM seul permet cependant une évaluation limitée car il prend en considération toutes les erreurs comme ayant la même « valeur » (Roux *et al.*, 2022). Par exemple, considérant la phrase de référence « J'aime les chats » et les phrases d'hypothèse 1 et 2 « J'aime les chaats » et « J'aime les chiens », le TEM serait le même dans les deux cas. Plus précisément, la première hypothèse substitue « chaats » à « chats » et la deuxième « chiens » à « chats ». Dans les deux cas le calcul du TEM est donc $\frac{S+D+I}{N} = \frac{1+0+0}{3} = 0,33$, pour un total de 33% de mots incorrects.

Pour pallier ce problème, il est possible d'utiliser le taux d'erreur sur les caractères (TEC), (Character Error Rate (CER) en anglais), qui est basé sur la même formule mais au niveau du caractère plutôt qu'au niveau du mot. Ceci permet d'inférer que le système de transcription tend ou non à générer des mots proches de la référence. Pour le français, par exemple, un faible TEC pourrait indiquer qu'il y a des erreurs en termes de genre ou de nombre (Roux *et al.*, 2022). Il est également envisageable d'employer des variantes des TEM/TEC basées sur la nature ou le lemme d'un mot. Cela implique de vérifier si la catégorie grammaticale ou la forme de base (lemme) de chaque mot dans l'hypothèse concorde avec celle de chaque mot dans la référence.

Toujours sur notre exemple des chats, le TEC pour notre première hypothèse est $\frac{S+D+I}{N} = \frac{1+1+0}{16} = 0,063$ soit 6.3% de taux d'erreur. Le TEC pour notre deuxième hypothèse est $\frac{S+D+I}{N} = \frac{2+0+1}{16} = 0,188$ soit 18.8% de taux d'erreur. Ce qui signifie qu'affiner la granularité permet d'affiner la détection des erreurs. Cependant, cet algorithme prend un temps quadratique en $O(n \cdot m)$ où $n$ est égal à la taille du premier texte, $m$ à la taille du second texte. Ce qui, dans le cas de très gros corpus, peut rendre le traitement très long.

### 2.2.2 Distance et similarité cosinus

Il est également possible d'utiliser la **distance cosinus** pour mesurer la similarité sémantique entre deux phrases. La distance cosinus offre une mesure quantitative de la distance sémantique en évaluant la séparation angulaire entre les vecteurs dans un espace multidimensionnel.

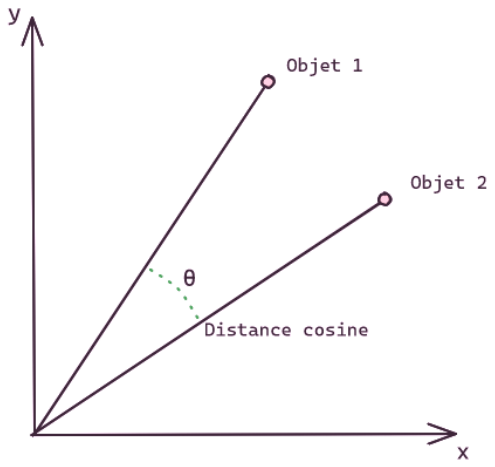


FIGURE 2.3 – Distance cosinus.

La formule de la similarité cosinus est la suivante :

$$\text{Similarité}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

où $A$ et $B$ sont tous deux des vecteurs de dimension $n$.

La dissimilarité ou distance cosinus est égale à $1 -$ Similarité.

Par exemple, pour un vecteur $A = [1, 3]$ et un vecteur $B = [2, 4]$

$$\text{Produit scalaire} = (1 \times 2) + (3 \times 4) = 14$$

$$\text{Norme de } \mathbf{A} = \sqrt{1^2 + 3^2} = \sqrt{10}$$

$$\text{Norme de } \mathbf{B} = \sqrt{2^2 + 4^2} = \sqrt{20}$$

$$\text{Similarité}(A, B) = \frac{14}{\sqrt{10} \times \sqrt{20}} = \frac{7}{5\sqrt{2}}$$

Toutefois, pour pouvoir user de ces mesures, il faut d'abord trouver une façon de vectoriser chaque phrase. Une première option est l'approche sacs de mots (SdM) (Bag-of-Words (BoW) en anglais) (Harris, 1954). Elle permet de vectoriser des phrases ou des documents en considérant uniquement la présence ou le nombre d'occurrences des mots qui les constituent. Pour ce faire, un vocabulaire est établi d'avance. Chaque mot de ce vocabulaire se voit attribuer une composante fixe de la représentation vectorielle souhaitée. Le vecteur correspondant à un document aura donc pour chacune de ces composantes le nombre d'occurrences du mot associé. Par exemple :

Phrase A : « J'aime les chats. »

Phrase B : « J'aime les chiens. »

| | J | aime | les | chats | chiens |
|---|---|---|---|---|---|
| Phrase A | 1 | 1 | 1 | 1 | 0 |
| Phrase B | 1 | 1 | 1 | 0 | 1 |

Nous avons donc les deux vecteurs $A = [1, 1, 1, 1, 0]$ et $B = [1, 1, 1, 0, 1]$. La similarité cosine entre ces deux vecteurs est ainsi la suivante :

$$\text{Produit scalaire} = (1 \times 1) + (1 \times 1) + (1 \times 1) + (1 \times 0) + (0 \times 1) = 3$$

$$\text{Norme de } \mathbf{A} = \sqrt{(1^2) + (1^2) + (1^2) + (1^2) + (0^2)} = \sqrt{4} = 2$$

$$\text{Norme de } \mathbf{B} = \sqrt{(1^2) + (1^2) + (1^2) + (0^2) + (1^2)} = \sqrt{4} = 2$$

$$\text{Similarité}(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{3}{2 \times 2} = \frac{3}{4} = 0,75$$

Il existe d'autres techniques plus avancées, comme fréquence des termes - fréquence inverse des documents (FT-FID) (Sparck Jones, 1972; Robertson, 2004), Word2Vec (Mikolov *et al.*, 2013) ou Sentence-Bert (Reimers et Gurevych, 2019).

FT-FID est une technique permettant de déterminer l'importance d'un terme en se basant à la fois sur sa fréquence dans un document donné et sur sa rareté dans l'ensemble de la collection de documents. Dans notre cas, le « terme » est un **mot** mais il peut être de nature différente selon le contexte d'application.

FT-FID se calcule comme suit :

Pour $t$, $d$, $D$ respectivement le terme, le document analysé et la collection de documents.

$$\text{FT-FID}(t, d, D) = \text{FT}(t, d) \times \text{FID}(t, D)$$

où

$$\text{FT}(t, d) = \frac{\text{Fréquence de } t \text{ dans } d}{\text{Nombre de termes dans } d}$$

$$\text{FID}(t, D) = \log\left(\frac{\text{Nombre de documents dans } D}{\text{Nombre de documents dans } D \text{ contenant } t}\right)$$

Par exemple, si on a deux documents :

D1 : « J'aime les chats. »

D2 : « J'aime les chiens. »

Alors :

$$\text{FT-FID}(\text{« chats »}, \text{D1}, \{\text{D1, D2}\}) = \frac{1}{3} \times \log\frac{2}{1} \approx 0,23$$

$$\text{FT-FID}(\text{« J'aime »}, \text{D1}, \{\text{D1, D2}\}) = \frac{1}{3} \times \log\frac{2}{2} = 0$$

Une valeur élevée FT-FID signifie que le terme est plus important ou pertinent car il est fréquent dans le document et apparaît moins fréquemment dans les autres documents. Un valeur basse FT-FID signifie que le terme est moins pertinent car il est commun à tous les documents ou qu'il n'apparaît pas ou peu dans le document analysé. Si le score FT-FID est égal à 0, cela signifie que le FT ou le FID est égal à 0 et donc soit que le terme n'apparaît pas dans le document, soit qu'il apparaît dans tous les documents du corpus. En l'occurrence « J'aime » est présent dans tous les documents, et « chats » n'est présent que dans un document.

Les modèles Word2Vec sont entraînés pour conserver le contexte des mots. Chaque mot est représenté par un vecteur dont les coordonnées expriment sa relation avec les mots du texte. Ainsi, les mots qui apparaissent dans des contextes similaires sont représentés par des vecteurs proches au sens de la distance cosinus. Pour ce faire, les auteurs de (Mikolov *et al.*, 2013) proposent deux variantes de Word2Vec : sacs de mots continu (SdMC) et skip-gram. Dans les SdMC, les représentations vectorielles des mots du contexte (c'est-à-dire des mots environnants) sont combinées pour prédire le mot du milieu. Par exemple pour « J'aime les chats » et en prenant comme contexte seulement le mot d'avant, le contexte « J'aime » a pour mot cible « les ». Dans (Mikolov *et al.*, 2013), les meilleurs résultats sont atteints en utilisant comme mots contexte les quatre mots de part et d'autre du mot cible. Le « skip-gram » opère de façon inverse : le vecteur du mot cible doit servir à prédire les mots de son contexte.

Quant à SentenceBERT, il est basé sur BERT (Devlin *et al.*, 2019) et modifié pour être capable de générer des encodages de phrase. SentenceBERT est pré-entraîné sur de grands corpus de textes en utilisant un objectif de modélisation du langage masqué. Au cours du pré-entraînement, le modèle apprend à prédire les mots masqués dans une phrase en fonction de leur contexte, de la même manière que BERT est pré-entraîné (Reimers et Gurevych, 2019). Après le pré-entraînement, le modèle SentenceBERT est affiné sur des tâches qui nécessitent des représentations au niveau de la phrase, telles que la similarité textuelle sémantique, la classification de texte et le regroupement.

Le code disponible en annexe figure B.2 utilise un modèle Word2Vec entraîné sur un corpus français [57]. Avec ce code, on obtient une similarité cosinus de 0,852 pour « J'aime les chaats » en comparaison à la phrase référence « J'aime les chats ». On obtient une similarité cosinus de 0,960 pour « J'aime les chiens » par rapport à la phrase référence. Ceci s'explique simplement, le mot « chaats » n'existe pas dans le vocabulaire utilisé par Word2Vec. En comparaison, le code disponible figure B.3 utilise un modèle SentenceBERT multilingue [58]. Avec ce code, on obtient une similarité cosinus de 0,703 pour « J'aime les chaats ». Et on obtient une similarité cosinus de 0,641 pour « J'aime les chiens ». Donc selon la façon de représenter vectoriellement les objets, la distance cosinus entre deux éléments peut être différente.

57. `https://huggingface.co/Word2vec/nlpl_43`

58. `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

## 2.3 Performance

Whisper est une alternative très intéressante pour la détection de langue ainsi que la transcription multilingue. Au besoin, il peut aussi répondre à la tâche de transcription et traduction d'un langage x à un langage y. Whisper est aussi capable de prendre en compte les fichiers longs. De ce fait, il est intéressant d'examiner les résultats rapportés.

| | Librispeech | | Common Voice 9 | | Vox Populi | | FLEURS | |
|---|---|---|---|---|---|---|---|---|
| Modèles | Anglais | Français | Anglais | Français | Anglais | Français | Anglais | Français |
| Whisper tiny | 15,7 | 36,8 | 28,8 | 49,7 | 11,6 | 32,9 | 12,4 | 41,4 |
| Whisper base | 11,7 | 26,6 | 21,9 | 37,3 | 9,5 | 24,9 | 8,9 | 28,5 |
| Whisper small | 8,3 | 16,2 | 14,5 | 22,7 | 8,2 | 15,7 | 6,1 | 15,0 |
| Whisper medium | 6,8 | 8,9 | 11,2 | 16,0 | 7,6 | 12,2 | 4,4 | 8,7 |
| Whisper large-v2 | 6,2 | 7,3 | 9,4 | 13,9 | 7,0 | 11,4 | 4,2 | 8,3 |
| Canary-1b | | | 7,8 | | 5,8 | | | |

TABLE 2.1 – TEM rapportés par (Radford *et al.*, 2022; Srivastav *et al.*, 2023) pour de la transcription multilingue.

Inévitablement, puisque la quantité de données d'apprentissage en anglais est près de 40 fois supérieure à celle en français, Whisper est moins performant en français. Ceci est d'autant plus flagrant pour les plus petits modèles comme nous pouvons le constater sur le tableau 2.1. Les modèles « medium » et « large-v2 » offrent quant à eux des résultats relativement acceptables. Choisir entre ces modèles revient donc à faire un compromis entre vitesse et qualité de transcription. Le « medium » étant environ 2 fois plus rapide que le « large-v2 » [59]. Le TEM moyen inter corpus est de 7,5 en anglais et de 11,45 pour le modèle « medium ». Pour le modèle « large-v2 », il est respectivement de 6,7 et 10,23. Tous les corpus n'ont pas le même nombre de données ni les mêmes données, ces dernières valeurs sont donc à prendre avec précaution. Elles permettent toutefois d'avoir un ordre d'idée sur les performances des modèles.

En comparaison, Canary-1b, actuellement le modèle à l'état de l'art, obtient un TEM de 7,97 en anglais et 6,53 en français sur Common Voice 16.1. Ce dernier est proposé par NVIDIA et a

59. Confère `https://github.com/openai/whisper`

en moyenne sur 9 corpus anglais le TEM le plus bas. À noter qu'il est aussi beaucoup plus lent à l'inférence par rapport à d'autre alternatives. Par exemple, il est près de 2 fois plus lent que Whisper medium. Sur Common Voice 9 en anglais, il obtient 7,75 de TEM (Srivastav *et al.*, 2023).

Nous sommes aussi intéressés par la transcription de fichiers longs. Beaucoup d'approches ne permettent pas de transcrire des fichiers de plus de 30 secondes. C'est aussi le cas de Whisper, mais celui-ci incorpore dans son système une coupure par tranche lors de fichiers plus longs, ce qui permet d'avoir des résultats relativement bons comme rapportés tableau 2.2. Le stt_en_conformer_ctc_large [60] est un modèle proposé par NVIDIA Nemo et servant de base de comparaison.

| Modèles | TED-LIUM3 | Meanwhile | Kincaid46 | Rev16 | Earnings-21 | Earnings-22 | CORAAL |
|---|---|---|---|---|---|---|---|
| Whisper tiny | 6,8 | 15,5 | 16,7 | 17,0 | 18,7 | 24,4 | 33,1 |
| Whisper base | 4,8 | 12,2 | 12,2 | 14,5 | 13,5 | 18,4 | 26,9 |
| Whisper small | 4,2 | 6,9 | 10,1 | 12,1 | 11,1 | 14,3 | 22,3 |
| Whisper medium | 3,8 | 5,4 | 8,6 | 11,4 | 10,3 | 13,2 | 20,3 |
| Whisper large-v2 | 3,5 | 5,1 | 8,8 | 11,3 | 9,7 | 12,6 | 19,6 |
| stt_en_conformer_ctc_large | 4,0 | 9,8 | 13,1 | 14,5 | 12,6 | 17,6 | 25,1 |

TABLE 2.2 – TEM rapportés par (Radford *et al.*, 2022) sur la transcription de fichiers longs en anglais.

En anglais, Whisper atteint des résultats proches d'une transcription réalisée par des professionnels (Radford *et al.*, 2022). Pour 25 fichiers donnés du corpus Kincaid46, Whisper obtient un TEM de 8,81 tandis que des services de transcription professionnels obtiennent un TEM entre 8,14 et 10,5. Un service de professionnels assistés par ordinateur obtient un TEM de 7,61.

SpeechBrain reprend Whisper « medium » en le réentrainant sur des données du langage voulu [61]. Il peut donc être intéressant de le comparer avec Whisper « medium » sur des corpus français. L'anglais étant prédominant et puisqu'une partie de notre question de recherche implique l'aspect

60. `https://huggingface.co/nvidia/stt_en_conformer_ctc_large`

61. Par exemple, pour le français `https://huggingface.co/speechbrain/asr-whisper-medium-commonvoice-fr`

multilingue, les 4 modèles que nous allons analyser sont soit des modèles pensés pour le français, soit des modèles multilingues. Pour une analyse quantitative et qualitative de modèles par SpeechBrain et NeMo, nous prenons les modèles suivants :

- `speechbrain/asr-whisper-medium-commonvoice-fr` [62] : 9,65 TEM rapporté sur Common Voice 10. Raccourci `sb-whisper`.
- `speechbrain/asr-wav2vec2-commonvoice-14-fr` [63] : 10,24 TEM rapporté sur Common Voice 14. Raccourci `sb-wav2vec2`.
- `nvidia/stt_fr_fastconformer_hybrid_large_pc` [64] : 7,92 TEM rapporté sur Common Voice 12. Raccourci `nvidia-fastconformer`.
- `nvidia/stt_fr_conformer_ctc_large` [65] : 7,95 TEM rapporté sur Common Voice 7. Raccourci `nvidia-confctc`.

Nous utilisons comme base de comparaison le modèle medium de Whisper. NVIDIA NeMo met à disposition canary 1b qui est un modèle équivalent à Whisper large. Celui-ci est le modèle le plus performant en anglais d'après l'« Open ASR Leaderboard » [66]. Cependant, pour cette étude nous souhaitons comparer des modèles accessibles sur des machines de particulier. Pareillement, nous nous concentrons principalement sur la transcription bien qu'à terme d'autres tâches seraient très intéressantes à incorporer dans notre outil. De ce point de vue, SpeechBrain semble avoir le profil parfait. Nous ignorons aussi ESPNet dû à sa courbe d'aprentissage et à la quantité minime des modèles multilingues et français disponibles.

**À noter que la version actuelle de sb-whisper ne supporte pas la prise en charge de longs extraits.** Cependant, une amélioration de ce processus est actuellement en cours [67] par l'un des mainteneurs principaux. Nous récupérons donc SpeechBrain à cette version pour permettre la prise en charge de longs fichiers lors de nos tests.

62. `https://huggingface.co/speechbrain/asr-whisper-medium-commonvoice-fr`
63. `https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-14-fr`
64. `https://huggingface.co/nvidia/stt_fr_fastconformer_hybrid_large_pc`
65. `https://huggingface.co/nvidia/stt_fr_conformer_ctc_large`
66. `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`
67. `https://github.com/speechbrain/speechbrain/pull/2450`

### 2.3.1 Analyse quantitative

Les modèles que nous utilisons n'ont pas tous été évalués sur les même corpus de données. De ce fait, nous les testons sur FLEURS en français. Au total, nous avons 672 extraits, chacun d'une durée comprise entre 4 et 12 secondes. Dans l'objectif d'avoir une expérience relativement proche d'une machine d'un particulier, les expériences sont lancées sur une grappe de calcul sans l'utilisation d'une carte graphique et avec 12Go de RAM [68].

| | TEM | TEC | Similarité cosinus | Temps de calcul |
|---|---|---|---|---|
| Whisper medium | 0,12 | 0,07 | **0,96** | 81,88s |
| `sb-whisper` | 0,13 | 0,08 | 0,95 | 46,62s |
| `sb-wav2vec2` | 0,64 | 0,25 | 0,48 | 1,75s |
| `nvidia-fastconformer` | **0,11** | **0,06** | **0,96** | **1,68s** |
| `nvidia-confctc` | 0,13 | 0,07 | 0,95 | 2,81s |

TABLE 2.3 – Résultats en moyenne sur FLEURS.

La différence de temps entre Whisper medium et son équivalent SpeechBrain est étonnante. Elle peut être due à un changement dans le système d'inférence. Ce changement n'a pas l'air d'impacter les résultats de `sb-whisper` mais le fait qu'il soit affiné sur des données françaises ne semble pas non plus améliorer les résultats obtenus sur FLEURS. Cela étant dit, les modèles NVIDIA sont beaucoup plus rapides pour des résultats comparables. Le temps de traitement entre les modèles SpeechBrain et NVIDIA est dû au fait que les modèles utilisés aient des architectures différentes.

Le modèle wav2vec2 obtient un TEM très haut mais un TEC très bas, ce qui est caractéristique de fautes d'orthographes. Les résultats de ce dernier sont bas mais sont cohérents avec le fait que ce type de modèle n'est plus vraiment à l'état de l'art. De plus le wav2vec2 est généralement utilisé en parallèle avec d'autres approches pour palier les problèmes orthographiques.

Pour évaluer la performance du système de reconnaissance vocale sur un ensemble de données, nous avons calculé le TEM, le TEC, la similarité cosinus ainsi que le temps de calcul pour chaque échantillon. La figure 2.4 montre les boîte à moustaches correspondant à la distribution de ces

68. La plupart des machines ont maintenant 16 Go de RAM, l'OS (Windows, Linux, MacOS) prend généralement 4Go ou moins.

métriques sur l'ensemble de FLEURS.

La boîte centrale indique l'étendue interquartile, c'est-à-dire la plage entre le premier et le troisième quartile, qui contient 50 % des valeurs. La ligne horizontale à l'intérieur de la boîte représente la médiane, qui est la valeur centrale de la distribution. Les "moustaches" s'étendent jusqu'aux valeurs minimales et maximales qui ne sont pas considérées comme des extrêmes. Les points situés en dehors des moustaches sont des valeurs aberrantes.

La figure 2.4 nous permet de voir que les modèles sont plutôt consistants en termes de qualité de transcription. Excepté le `sb-wav2vec2`, les trois premiers quartiles de tous les modèles sont inférieurs à 0,2 de TEM. Aussi, le temps d'exécution de Whisper medium varie de manière importante contrairement à celui de `sb-whisper`.

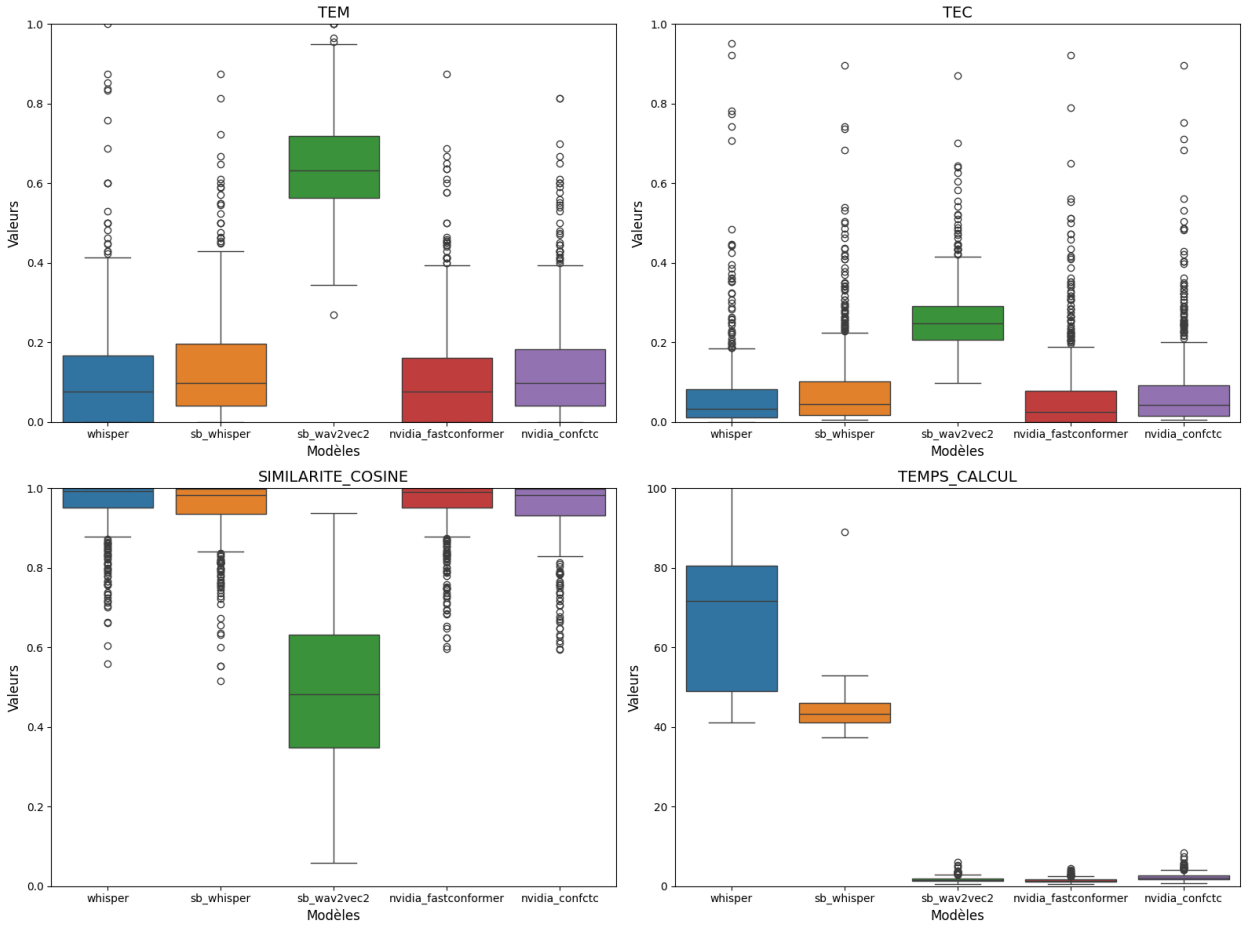


FIGURE 2.4 – Distribution des résultats de chaque modèle sur FLEURS.

### 2.3.2 Analyse qualitative

Nous nous concentrerons sur certains extraits de la commission Viens. Nous sélectionnons aléatoirement les audiences du 12 septembre 2017, du 16 janvier 2018 et du 23 octobre 2018. Pour chaque audience, 3 extraits d'une durée de 15 secondes et 3 extraits d'une durée de 30 secondes sont sélectionnés aléatoirement. La transcription de référence est récupérée manuellement à partir des fichiers de transcription mis en ligne par la commission Viens. Nous avons donc un total de 18 extraits audio. Tous les résultats sont disponibles en annexe C.

| | TEM | TEC | Similarité cosinus | Temps de calcul |
|---|---|---|---|---|
| Whisper medium | **0,19** | **0,17** | **0,95** | 23,38s |
| `sb-whisper` | 0,29 | 0,20 | 0,92 | 22,25s |
| `sb-wav2vec2` | 0,69 | 0,32 | 0,56 | 6,04s |
| `nvidia-fastconformer` | 0,34 | 0,21 | 0,88 | **4,70s** |
| `nvidia-confctc` | 0,35 | **0,19** | 0,86 | 4,91s |

TABLE 2.4 – Résultats en moyenne sur les 9 extraits de 15s de la commission Viens.

La table 2.4 nous permet de constater que le modèle obtenant les meilleurs résultats est Whisper medium. Cependant, `nvidia-conftrans` obtient le même TEC pour un temps d'exécution près de 6 fois plus court. `sb-whisper` obtient les meilleurs similarités cosines et TEM après Whisper. Son TEC est légèrement supérieur à celui de `nvidia-confctc`, ce qui peut être attribué à la marge d'erreur. Par contre, pour un TEC très proche, le `nvidia-confctc` a un TEM 15 points supérieur à Whisper medium et 6 points supérieur à `sb-whisper`. `nvidia-fastconformer` a des résultats similaires. Ceci tend à montrer que les modèles NVIDIA prédisent correctement la plupart des caractères et que les erreurs sont éparpillées entre les mots, ce qui invalide beaucoup de mots, potentiellement à cause d'un caractère incorrect. `sb-wav2vec2` est quand à lui très mauvais bien qu'il soit assez rapide.

Les figures 2.5 et 2.6 présentent des diagrammes en violon. Un diagramme en violon combine une boîte à moustaches et une estimation de la densité des données. La largeur du violon à différents niveaux indique la densité des données à cette valeur, c'est-à-dire où les valeurs sont les plus fréquentes. La première figure nous permet de visualiser les résultats pour tous les extraits pour un modèle donné. La seconde figure nous permet d'appréhender la difficulté d'un segment à être transcrit par un modèle donné ou en moyenne.

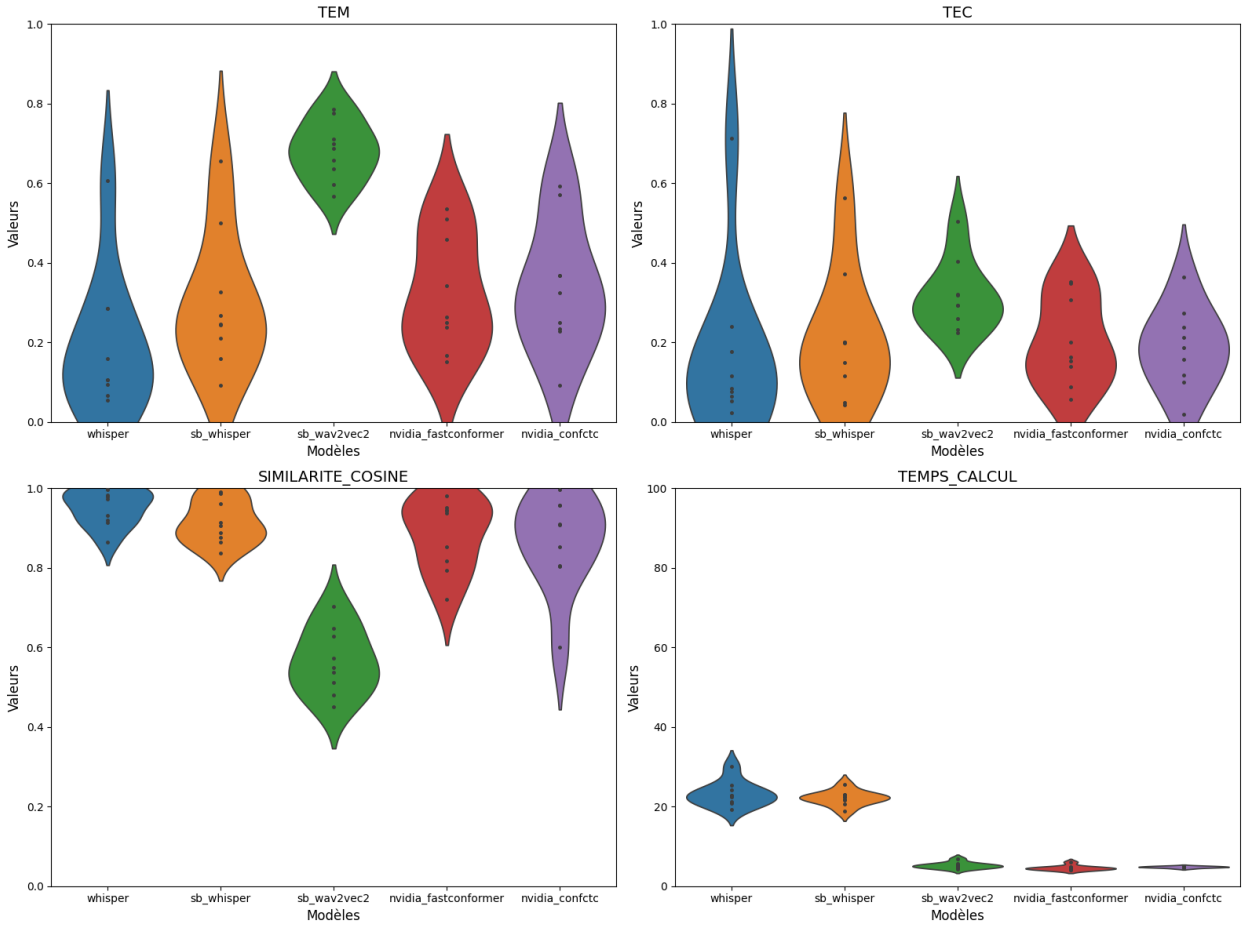


FIGURE 2.5 – Distribution des résultats de chaque modèle sur les extraits de 15s

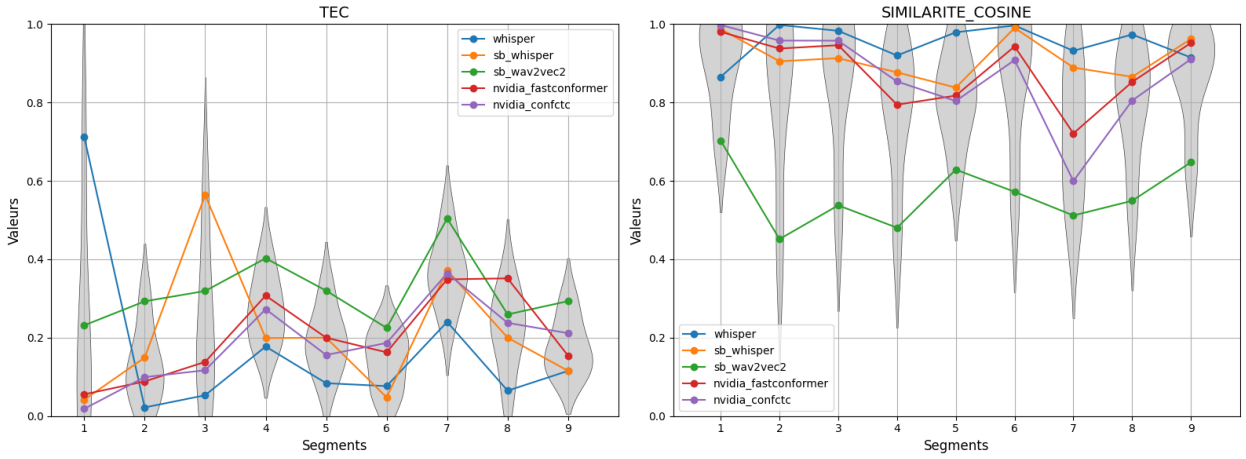


FIGURE 2.6 – Distribution par extrait de 15s des performances des modèles.

La figure 2.5 nous permet de constater si ces moyennes sont lourdement influencées ou non par des extrêmes. Pour chacun des modèles il y a toujours environ deux fichiers pour lesquels les métriques sont significativement moins bonnes. Ces fichiers ne sont cependant pas tous communs entre les différents modèles comme nous pouvons le voir sur la figure 2.6.

`sb-wav2vec2` est presque toujours le moins performant. Whisper Medium est médiocre sur le premier segment. `sb-whisper` performe significativement moins bien sur le troisième segment. Il en va de même pour les modèles NVIDIA sur le 7ième segment. De manière générale, tous les modèles performent relativement mieux sur le 6ième segment. Nos intuitions sont confirmées par la table 2.5. Whisper infère incorrectement une suite de « bonjour ». `sb-whisper` cesse de prédire trop tôt. Les modèles NVIDIA prédisent des sonorités, ce qui rend incorrect leur prédiction du mot.

| **Segment 1** | |
|---|---|
| **Référence** | **Hypothèse par Whisper Medium** |
| ...veuillez vous asseoir alors bonjour bonjour à tous | ...veuillez vous asseoir alors bonjour bonjour à tous bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour |
| **Segment 3** | |
| **Référence** | **Hypothèse par `sb-whisper`** |
| de chacune des nations et on fait réaliser aux intervenants que onze nations onze cultures différentes donc pierre vient un petit peu établir des différences culturelles et c est là qu il | leurs cultures de chacune des nations ont fait réaliser aux intervenants qu onze nations onze cultures différentes |
| **Segment 7** | |
| **Référence** | **Hypothèse par `nvidia-fastconformer`** |
| difficile là de penser que l autre il veut te prendre dans tes bras quand il veut juste t étrangler là je pense à un des gars là | l a dépensé que l autre y veut prendre tes brocas qui veut juste étrangler je pense à un digot |
| | **Hypothèse par `nvidia-confctc`** |
| | l a dépensé que l hôte y veut prendre t des brocs às qui veut juste étrangler je pense à un des gots |

TABLE 2.5 – Erreurs significatives des modèles sur les extraits de 15s.

En ce qui concerne les extraits de 30s, les problèmes de `sb-whisper` sont exacerbés. La plupart des fichiers n'ont pas de transcription complète. Nous rappelons toutefois que la gestion des fichiers longs par `sb-whisper` est en développement. Les résultats des autres modèles sont cohérents avec leurs résultats sur les extraits de 15 secondes. La table 2.6 consigne les résultats obtenus sur les

| | TEM | TEC | Similarité cosine | Temps de calcul |
|---|---|---|---|---|
| Whisper medium | **0,18** | **0,12** | **0,95** | 35,03s |
| `sb-whisper` | 0,53 | 0,45 | 0,84 | 25,25s |
| `sb-wav2vec2` | 0,64 | 0,30 | 0,64 | 8,06s |
| `nvidia-fastconformer` | 0,25 | **0,16** | 0,91 | **5,43s** |
| `nvidia-confctc` | 0,32 | 0,18 | 0,88 | 5,61s |

TABLE 2.6 – Résultats en moyenne sur les 9 extraits de 30s de la commission Viens.

extraits de 30 secondes.

Les figures 2.7 et 2.8 nous permettent de constater que les résultats de `sb-whisper` sont très épars. Les valeurs mauvaises et extrêmes sont dues au fait que le modèle cesse de prédire très tôt.

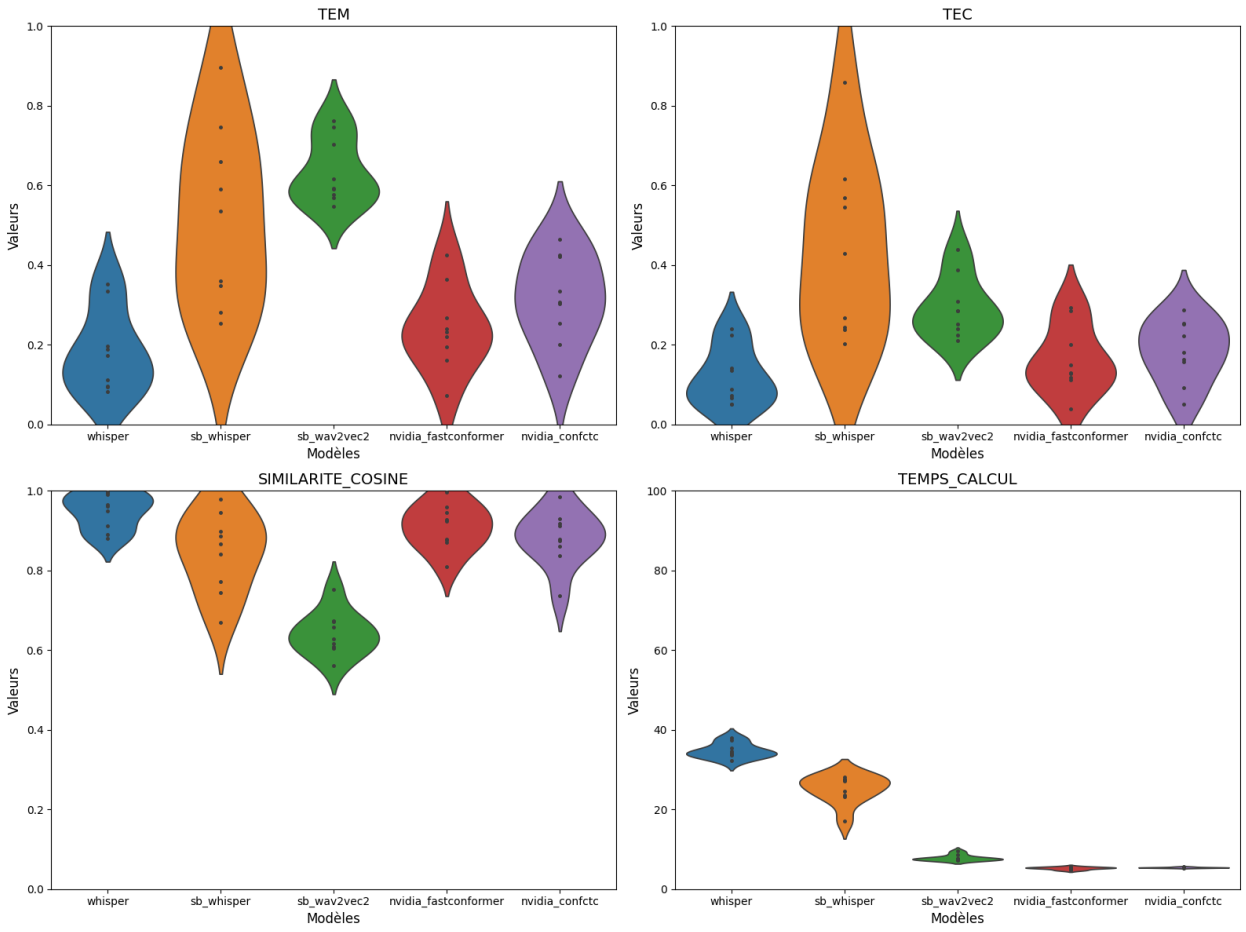


FIGURE 2.7 – Distribution des résultats de chaque modèle sur les extraits de 30s.

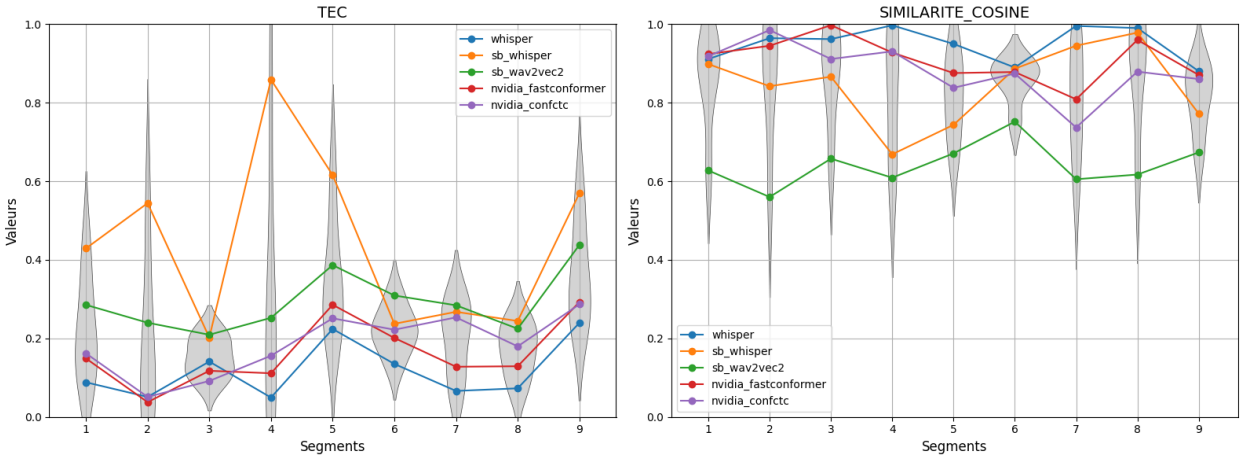


FIGURE 2.8 – Distribution par extrait de 30s des performances des modèles

# CHAPITRE 3
# RÉFLEXIONS PRÉLIMINAIRES

Nous aborderons dans le chapitre 3 les besoins, objectifs et questions auxquels nous souhaitons répondre. Nous présenterons ensuite les premières maquettes et architectures ainsi que les technologies que nous allons utiliser.

## 3.1 Objectifs et questions

Nous souhaitons développer un outil de transcription automatique. Celui-ci se distingue par trois principes centraux : la préservation de la confidentialité des personnes utilisatrices et de leurs données, son accès libre et sans restriction, ainsi que sa modularité, permettant une adaptation aisée.

Cet outil de transcription doit être capable de couper des extraits audio en portions de plus courte durée, *i.e.* segmenter les extraits audio. Il doit aussi transcrire chacune de ces portions puis être capable de les regrouper en un seul fichier texte. Il doit ensuite faire correspondre chaque mot transcrit avec l'horodatage de l'audio. De plus, il doit offrir les fonctionnalités de base présentes dans les alternatives dépourvues de modules de transcription automatique, telles que la possibilité de corriger le résultat, de sélectionner des sections et de vérifier l'alignement avec l'audio.

Cependant, l'automatisation du processus de transcription peut entraîner des problèmes, tels que des erreurs fréquentes dans la transcription de noms propres peu courants ou de mots mal prononcés. Pour résoudre ces problèmes, des fonctionnalités de correction et d'édition par lot sont envisagées pour offrir aux personnes utilisatrices une expérience plus fluide et précise.

Pour créer un outil de transcription automatique efficace, nous devons surmonter plusieurs défis. Il est crucial d'aligner l'interface utilisateur.trice (IU) avec l'audio de manière intuitive, permettant ainsi aux personnes utilisatrices de naviguer facilement dans le texte en cliquant sur des segments audio pertinents. Ensuite, nous devons trouver des solutions pour traiter les audio comprenant plusieurs langues ou dialectes, garantissant ainsi une transcription précise dans des contextes linguistiques variés. Une fois la transcription automatique réalisée, il est essentiel de permettre aux personnes utilisatrices de peaufiner les résultats pour une meilleure précision. Cela implique de

fournir des outils de correction et d'édition conviviaux, afin que les personnes utilisatrices puissent ajuster les transcriptions selon leurs besoins spécifiques. Enfin, nous devons élaborer des stratégies pour gérer les audio très longs, en les découpant sans perdre de contexte et en réalignant les segments découpés de manière cohérente.

En résumé, pour développer un outil de transcription automatique fiable et pratique, nous devons relever ces défis avec ingéniosité et mettre en place des solutions efficaces pour offrir une expérience utilisateur.trice (XU) optimale.

## 3.2 Besoins

Dans le développement de notre outil de transcription automatique, plusieurs éléments sont prioritaires. Tout d'abord, la confidentialité et la sécurité sont essentielles, avec la possibilité pour les personnes utilisatrices de sauvegarder localement leurs données sans aucune centralisation ou partage pour l'entraînement ultérieur du modèle. L'identification précise des locuteurs est également une priorité, tout comme la flexibilité permettant d'effectuer des éditions manuelles et des corrections grâce à une interface conviviale avec fonctionnalité de glisser-déposer. De plus, l'ajout d'un horodatage aux changements de locuteurs et une fonction de recherche et de remplacement des locutions similaires sont indispensables pour faciliter les corrections.

Ensuite, certains éléments sont considérés comme secondaires mais néanmoins importants. L'adaptation au français québécois est prioritaire, avec la possibilité d'ajouter d'autres langues par interlocuteur pour une plus grande diversité linguistique. La capacité à filtrer les bruits ambiants, l'identification des interjections et des redondances, ainsi que la détection des erreurs orthographiques et grammaticales sont autant d'aspects cruciaux pour améliorer la qualité globale de la transcription. De plus, la création d'un dictionnaire de jargon à reconnaître dans les enregistrements futurs, avec la possibilité d'ajouts manuels, est un point à prendre en considération pour une reconnaissance plus précise et spécialisée.

## 3.3 Définition initiale de SimpleSpeech

L'outil SimpleSpeech est un outil conjugant le traitement du discours automatique et la faculté d'annoter et de corriger dans une interface d'édition directement liée au fichier audio concerné.

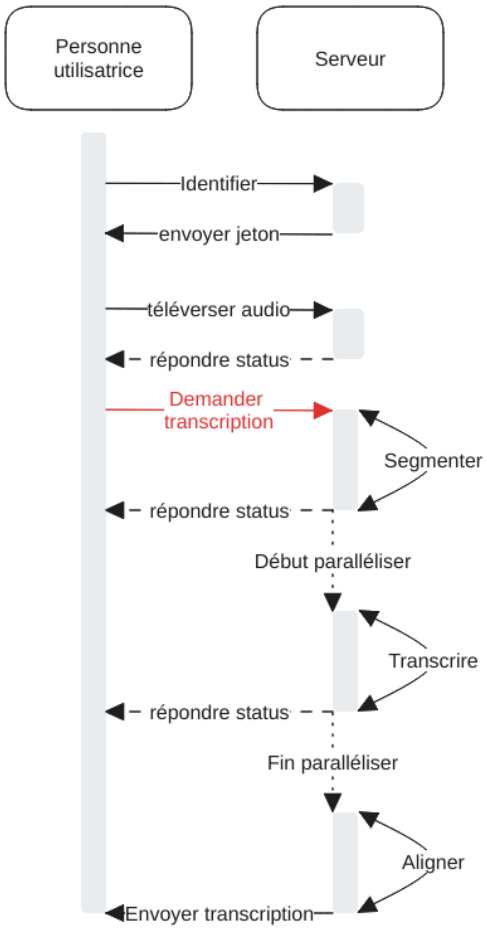


FIGURE 3.1 – Diagramme de séquences.

Suivant la figure 3.1, la personne utilisatrice s'authentifie puis demande à transcrire un ou plusieurs fichiers après les avoir téléversés. Cette transcription se fait en plusieurs étapes. Il s'agit notamment de détecter si le fichier est long, car la plupart des modèles ne sont pas capables de transcrire directement des fichiers longs. Si le fichier est long, il est alors coupé en plusieurs fichiers courts temporairement. Le ou les fichiers courts sont ensuite transcrits. Dans le cas d'un fichier long, tous les fichiers temporaires sont réalignés après transcription. Pour finir, la personne reçoit la transcription liée à son ou ses fichiers audio.

Ce processus correspond à une interaction du client vers une interface de programmation applicative (IPA). L'objectif est d'abstraire une telle interaction derrière une interface conviviale et facile d'utilisation. Néanmoins, le fonctionnement reste fondamentalement le même pour la personne utilisatrice : connexion, envoi du fichier, réception de la transcription.

À terme, l'interface permettra à la personne utilisatrice de visualiser l'état de la transcription ainsi que de voir et corriger le résultat de la transcription. Ceci est un processus itératif. Nous souhaitons d'abord créer un noyau simple auquel seront ajoutées des fonctionnalités.

### 3.3.1 Architecture

SimpleSpeech est une application qui doit pouvoir être utilisée en local totalement et ce sans requérir une connexion internet sauf dans le cas où la personne utilisatrice voudrait utiliser d'autres modèles ou fonctions. De ce fait, la figure 3.2 est un premier aperçu de l'architecture de SimpleSpeech. Sim-

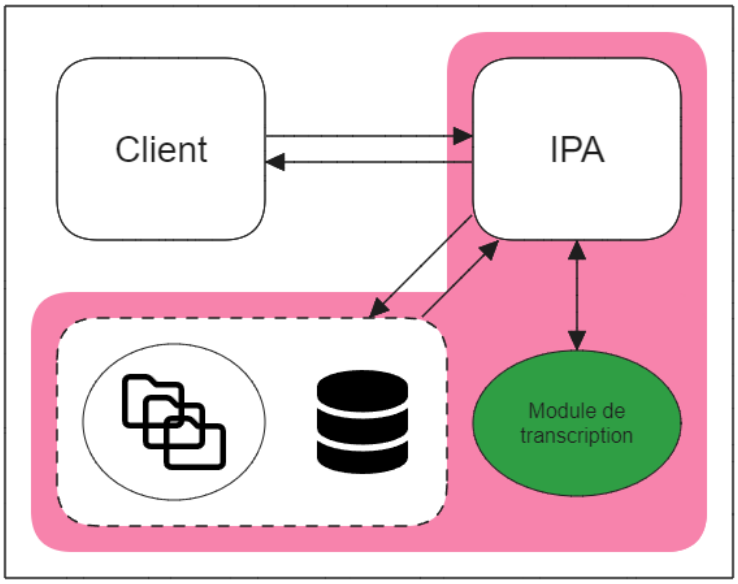


FIGURE 3.2 – Architecture en local (les icônes représentent un système de gestion de fichiers et de base de données).

pleSpeech devra être capable de stocker et analyser les extraits audio pour ensuite les transcrire et stocker cette transcription dans le but de permettre sa visualisation ainsi que son édition. Nous pouvons ainsi voir sur la figure 3.2 que SimpleSpeech est basé sur quatre modules fondamentaux. Le client est l'interface de visualisation, celle avec laquelle les personnes utilisatrices pourront se familiariser et lancer les transcriptions. C'est le module client, celui qui doit absolument être disponible sur la machine de la personne utilisatrice. Nous avons aussi une gestion de fichiers et une base de données qui permettent le traitement et la modification des fichiers. Enfin, nous avons le chef d'orchestre, l'IPA, qui permet d'actualiser les informations visibles sur le client et d'exécuter les commandes et processus souhaités. Ce module doit aussi pouvoir être totalement utilisable par

des développeurs et développeuses.

La partie rose de la figure 3.2 représente un module séparé qui peut être déployé en local sur sa propre machine ou bien sur son propre serveur si besoin. Dans le cas d'une utilisation sur serveur, les personnes utilisatrices n'ont donc besoin que d'installer le « client » et peuvent se connecter au serveur (local) souhaité.

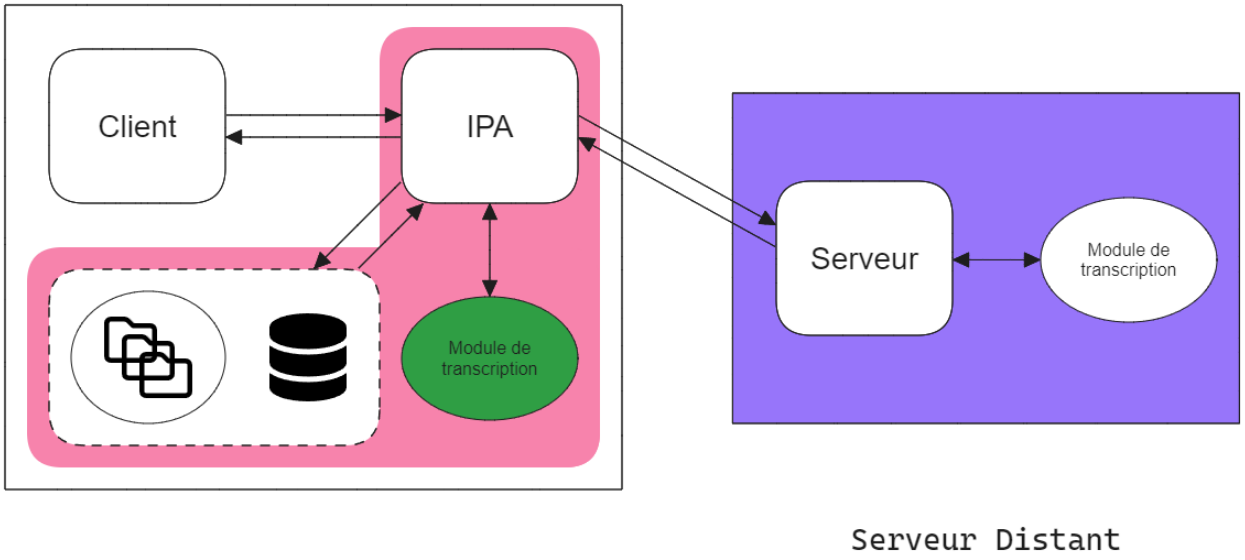


FIGURE 3.3 – Architecture globale.

La figure 3.3 démontre une contrainte supplémentaire : notre module séparable doit aussi prévoir la possibilité d'être déployé sur un serveur qui n'appartient pas à la personne utilisatrice. Ce serveur ne doit donc pas stocker les données. Cette contrainte est une contrainte à long terme dans le cas de certains modèles très performants qui ne pourraient pas tourner sur des machines pour particuliers. Dans un tel cas, la possibilité de déployer des instances spécialisées pour faire tourner ces modèles est envisageable. Dans l'optique de respecter les principes que nous avons mis en avant, l'utilisation de tels serveurs nécessite que la personne utilisatrice soit mise au courant et y consente.

### 3.3.2 Maquette

Dans un cadre totalement local, il n'est pas nécessaire d'avoir une gestion des usagers. Cependant, pour répondre à la possibilité de déployer cet outil pour plusieurs personnes et permettre la collaboration, il est nécessaire d'avoir une gestion des usagers, une gestion de projets et un panel d'administration. De ce fait, toute personne utilisatrice doit d'abord se connecter. Ensuite, elle accédera à l'interface de projet depuis laquelle il est possible de faire transcrire un nouveau fichier audio ou bien de visualiser et modifier une transcription déjà faite.

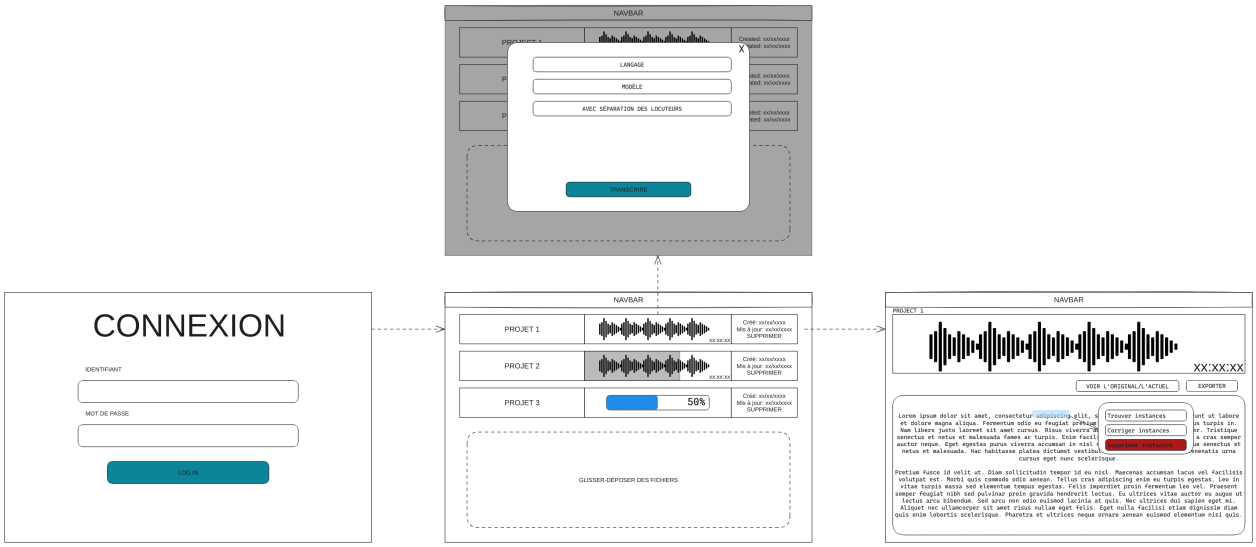


FIGURE 3.4 – Maquette de l'ensemble.

Pareillement, le niveau de choix et possibilités lors d'une demande de transcription doit être modifiable. Pour des personnes peu versées en informatique ou en linguistique, certains détails ne sont pas forcément importants et il n'est donc pas nécessaire de les afficher. En effet, présenter a priori trop d'informations « non-compréhensibles » risque de rebuter les personnes utilisatrices.

Les maquettes figures 3.4, 3.5 et 3.6 permettent seulement d'avoir une visualisation préliminaire du projet et ne sont pas des représentations figées de l'application.

La figure 3.5 est un aperçu de la gestion par projet. Nous pouvons y retrouver la possibilité de glisser-déposer des fichiers et de voir le status de téléversement ou transcription d'un projet donné. Ici, un projet est égal à un fichier audio. Il serait cependant intéressant d'envisager la possibilité d'avoir un groupe de fichiers audio par projet. Par exemple, dans le cadre d'entretiens, tous les fichiers audio

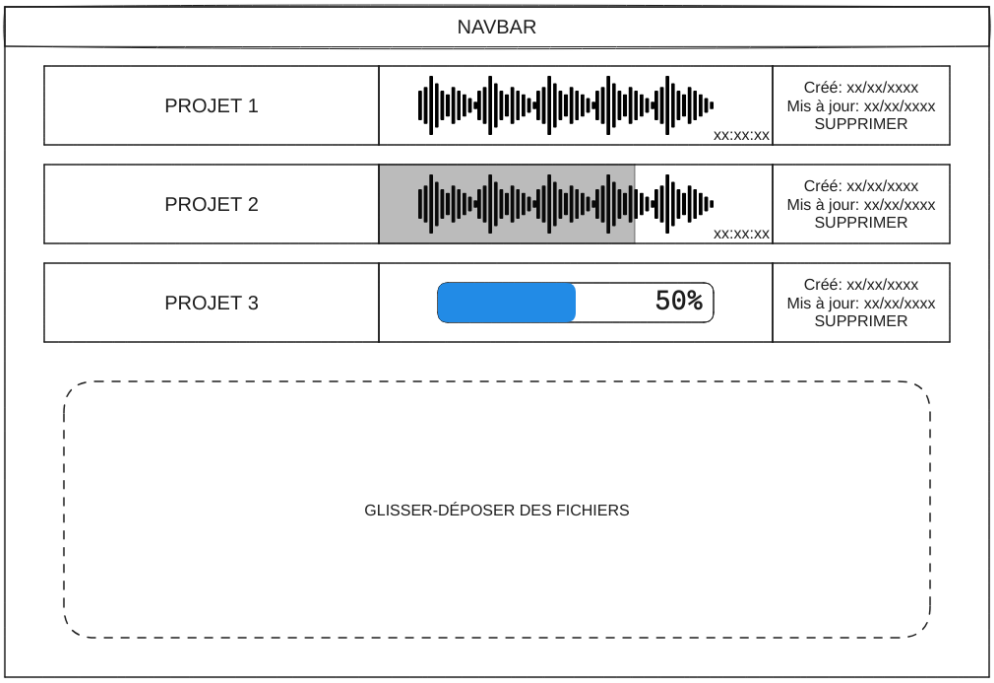


FIGURE 3.5 – Maquette de la gestion des projets à transcrire.

concernant la psychologie des enfants pourraient être disponibles dans le projet « enfants », et les fichiers concernant les troubles de l'autisme dans le projet « autisme ». Sur cette figure, on peut voir le projet 1 qui est un projet nouvellement téléversé. Si on clique dessus on aura alors accès à une interface permettant de choisir comment réaliser la transcription. Le projet 2 est un projet en cours de transcription et le projet 3 est un projet en cours de téléversement. Sur la droite de chaque projet, on peut voir quelques détails sur les dates de création et de modification ainsi que l'option de supprimer ledit projet.

Quant à la figure 3.6, nous pouvons y voir l'aperçu de la trame audio ainsi que le texte. L'objectif est que ce texte soit à la fois relié au segment audio (si on clique sur un endroit du texte, l'écoute de l'audio est directement placée à l'endroit correspondant au texte) mais aussi que les différentes versions après édition soient sauvegardées. Aussi, comme évoqué dans les objectifs et contraintes, la personne utilisatrice doit pouvoir être capable de trouver toutes les instances similaires, de les corriger ou de les supprimer. Enfin, la transcription doit être exportable.

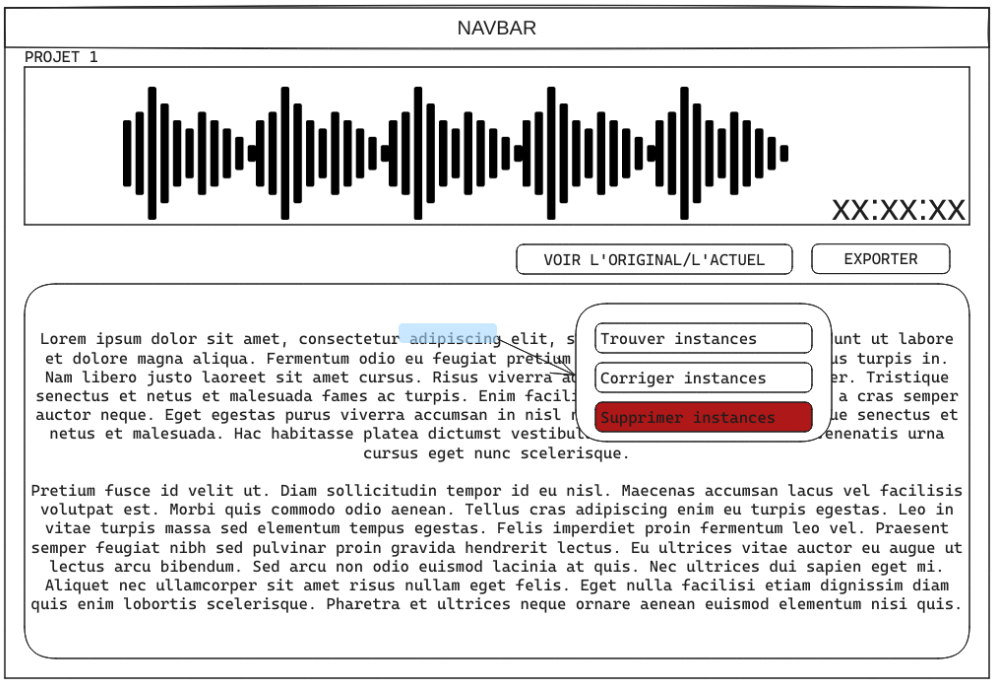


FIGURE 3.6 – Maquette de l'édition de fichier.

## 3.4 Choix technologiques

Comme nous l'avons vu dans la section 3.3.1, il est nécessaire de gérer une partie visuelle dite « client », et une partie gestion des requêtes dite « IPA » ou « serveur ». Il est aussi nécessaire de choisir un système de gestion de base de données.

### 3.4.1 Le serveur

Dans l'optique d'éviter les surcouches non-nécessaires et puisque le trafic d'utilisation de SimpleSpeech se fera de manière ponctuelle et par des petits groupes (l'application étant locale), **Python** est le langage de prédilection pour utiliser SpeechBrain qui est aussi développé en Python ainsi que pour communiquer avec le client et contrôler la base de données. En termes de développement web, il existe deux « frameworks » très connus : Django et Flask.

Django est un framework tout en un qui inclut le développement web, la gestion de serveur et l'affichage des pages. Il est très fixe dans sa forme et suit une structure peu flexible. Flask est quant à lui beaucoup plus léger et flexible mais aussi très brut. Nous choisissons **Flask** car nous n'avons

pas besoin de fonctionnalités pour gérer l'envoi de contenu web puisque nous avons un client pour cela. Nous avons besoin de flexibilité et d'extensibilité. Cela étant dit, pour bâtir sur une structure pérenne, nous décidons de suivre la structure disponible en figure 3.7.

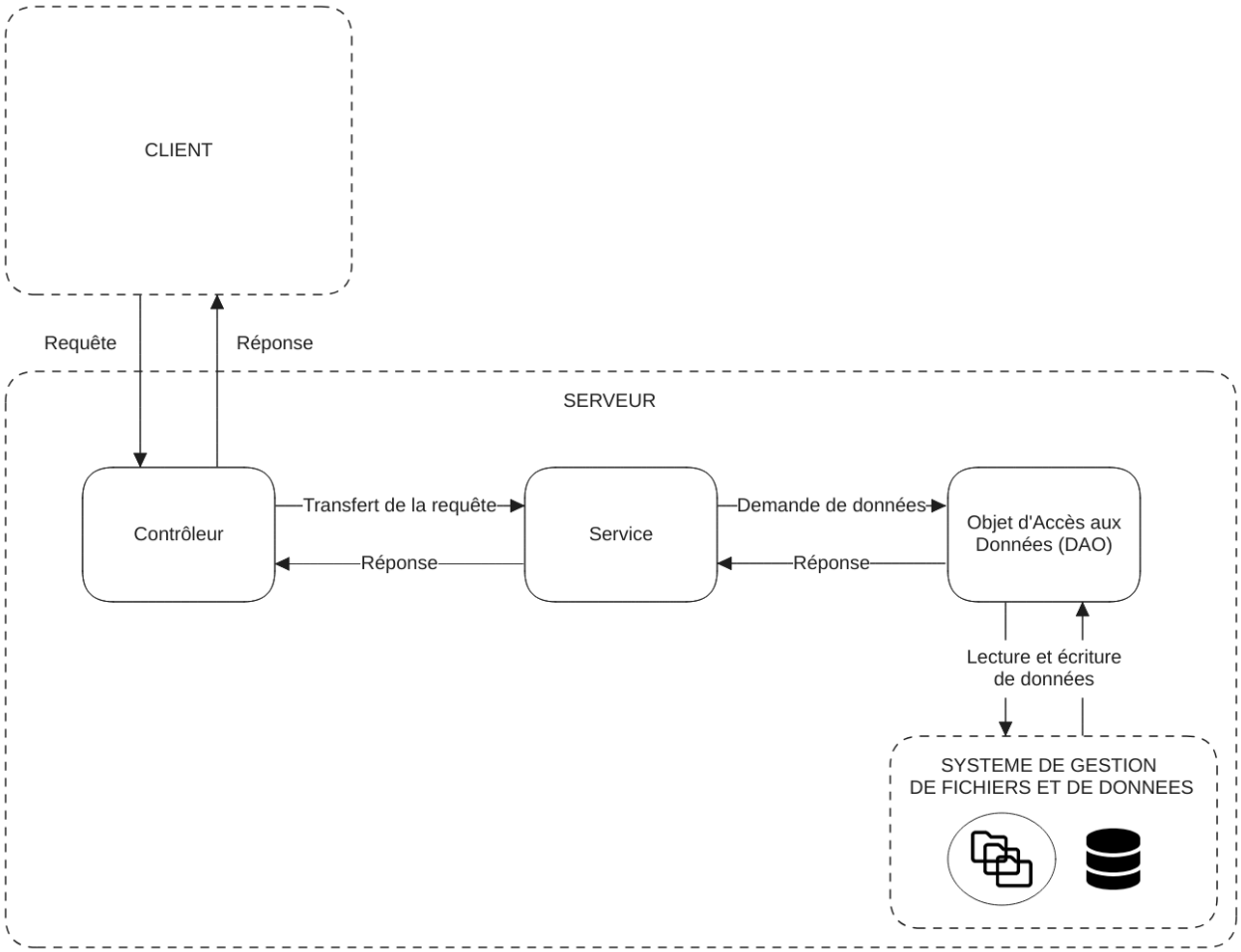


FIGURE 3.7 – Architecture de développement côté serveur.

Sur la figure 3.7, le client est abstrait et fait une requête au « contrôleur ». De l'extérieur, c'est le seul accessible. Il permet de vérifier l'authenticité de la personne se connectant ainsi que de contrôler le trafic et les requêtes. Basée sur ces informations, le service correspondant répondra en exécutant la demande et en contactant la couche « objet d'accès aux données (OAD) » au besoin. La couche « OAD » est une simple interface de communication avec le système de gestion de données et de fichiers, elle permet l'accès aux données.

Un exemple concret serait l'authentification d'une personne. Le client va requêter au contrôleur qu'il veut se connecter. Le contrôleur va faire la liaison avec le service qui permet de se connecter. Le

service fait le gros du travail en vérifiant qu'il y a bien les codes d'authentification et en exécutant d'autres fonctions au besoin. Il appelle aussi la couche OAD pour confirmer que cette personne utilisatrice existe et que son mot de passe est correct.

Dans l'optique de rendre notre serveur utilisable par des développeur.e.s, nous utilisons aussi la bibliothèque `flask_restx` qui permet de générer une documentation directement à partir du code.

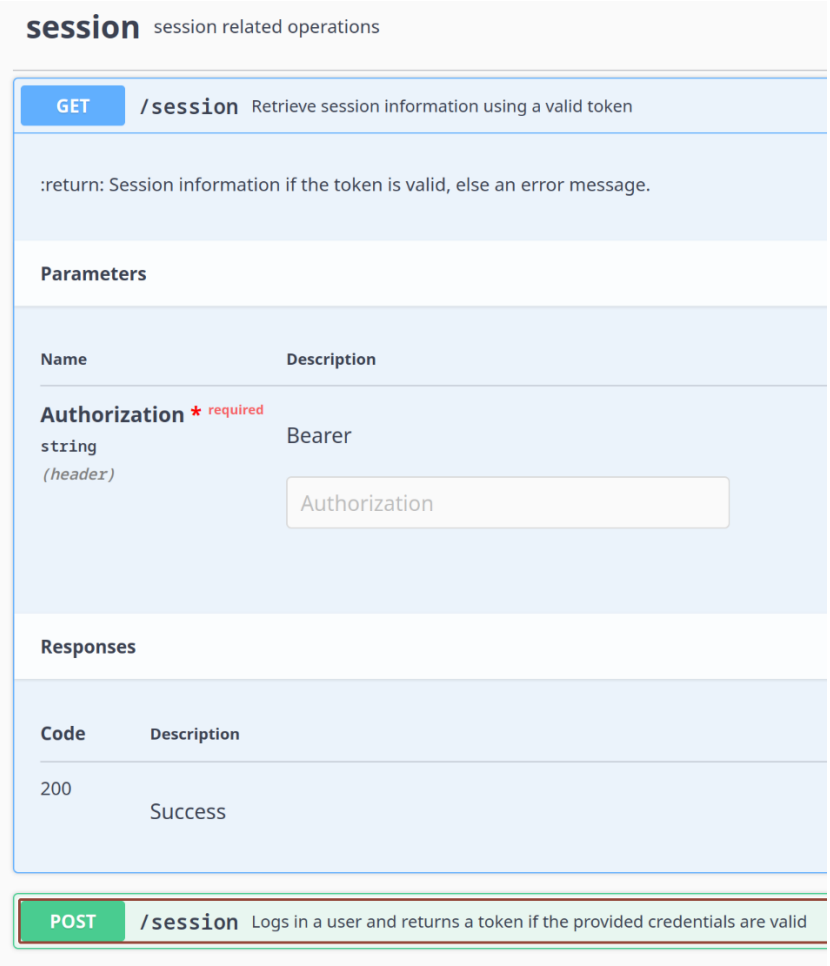


FIGURE 3.8 – Exemple de documentation générée par Flask_restx.

### 3.4.2 Le client

Par souci de portabilité, l'interface client sera une interface web. En effet, développer une application web permet de ne pas se soucier des différences entre les systèmes d'exploitation (Windows, Linux, MacOS). Deux langages sont particulièrement connus pour du développement web, PHP et

Javascript. PHP est un langage qui se concentre principalement sur l'aspect serveur, il est exécuté sur le serveur puis le résultat est envoyé au client. JavaScript est principalement utilisé pour la programmation côté client dans les navigateurs web. Puisque notre partie serveur est déjà gérée par Python, **Javascript** est le choix le plus logique. Ceci nous permettra aussi de communiquer plus facilement avec les bibliothèques Javascript permettant de générer les éléments dont nous aurons besoin. Les applications Javascript peuvent aussi être déployées en utilisant Electron, c'est à dire une application web qui est lancée en dehors du navigateur. Utiliser Electron n'étant pas une porte que nous souhaitons fermer, c'est une raison de plus pour utiliser Javascript.

Pour éviter de réinventer la roue et permettre la mise en place d'une structure de projet facilitant le travail en communauté, il est préférable d'aussi utiliser un framework Javascript plutôt que du Javascript simple. Trois frameworks sont bien connus et donc documentés : ReactJS, AngularJS et VueJS. Il existe d'autres alternatives comme ExpressJS, Svelte ou encore NextJS. Nous nous concentrerons sur les trois premiers que nous avons cités car ce sont les plus connus. ExpressJS est lui aussi très populaire mais est aussi très flexible en termes de possibilités et structures. Dans un environnement où le code est libre, il est plus facile de maintenir celui-ci si la structure est assez rigide. Vue.js est souvent apprécié pour sa simplicité et sa courbe d'apprentissage. React.js se distingue par son approche de rendu efficace basée sur le Virtual DOM, sa composabilité et son écosystème dynamique, en faisant un choix privilégié pour les grandes applications évolutives. AngularJS met en place une structure prédéfinie et des conventions strictes, ce qui le rend plus approprié pour les projets à grande échelle nécessitant une architecture robuste et une collaboration étroite entre les développeur.e.s. De fait, nous choisissons **AngularJS** pour le développement de SimpleSpeech.

AngularJS utilise les architecture de code « Modèle-Vue-Contrôleur » (MVC) et « Modèle-Vue-VueModèle »(MVVM). Ces architectures sont toutes deux des modèles couramment utilisés dans le développement d'applications web front-end. Dans le modèle MVC, l'application est divisée en trois composants : le Modèle, qui représente les données et les règles métier, la Vue, qui gère l'affichage des données à l'utilisateur, et le Contrôleur, qui agit comme un intermédiaire entre le Modèle et la Vue, traitant les actions de l'usager. En revanche, dans le modèle MVVM, la Vue est associée à un composant supplémentaire appelé VueModèle, qui agit comme une liaison de données bidirectionnelle entre la Vue et le Modèle, facilitant la manipulation de l'interface et la gestion de l'état de

l'application.

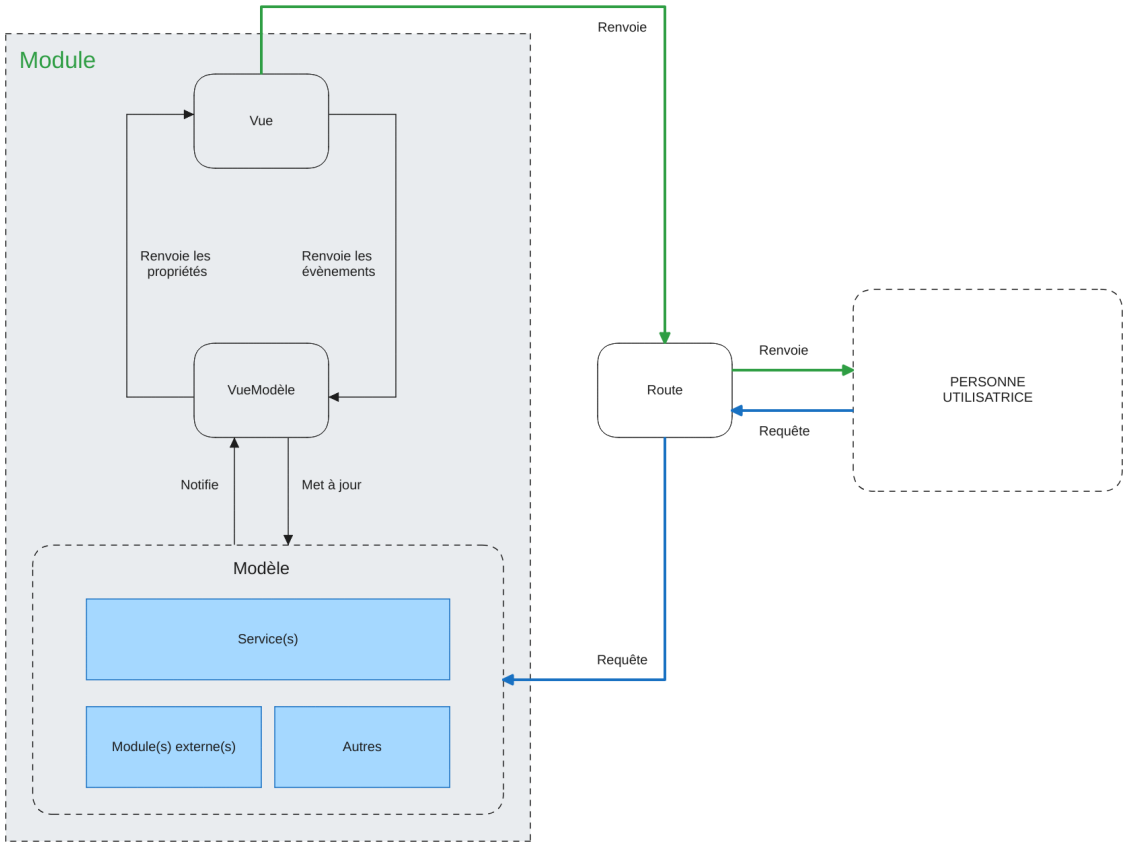


FIGURE 3.9 – Aperçu de l'architecture utilisée dans AngularJS.

### 3.4.3 La base de données

Pour la base de données, il est plus intéressant d'utiliser une base de données portative et donc légère comme Sqlite qui est une alternative mature qui fonctionne dans un fichier. Sqlite est une technologie qui supporte modérément l'insertion de données déstructurées sous la forme de JSON.

Pour ce projet, il est obligatoire d'avoir une gestion de fichiers pour les extraits audio. Un dossier sera donc attribué et lié avec le projet par le biais de la base de données.

La figure 3.10 est un premier schéma permettant d'avoir un aperçu du fonctionnement de la base de données. Nous avons 4 tables principales : `User`, `Project`, `audiofile` et `Model` qui sont respectivement les tables pour les personnes utilisatrices, les projets, les fichiers audio liés à ces projets et enfin le modèle choisi pour transcrire cet audiofile. La table `User_Project` permet de faire la

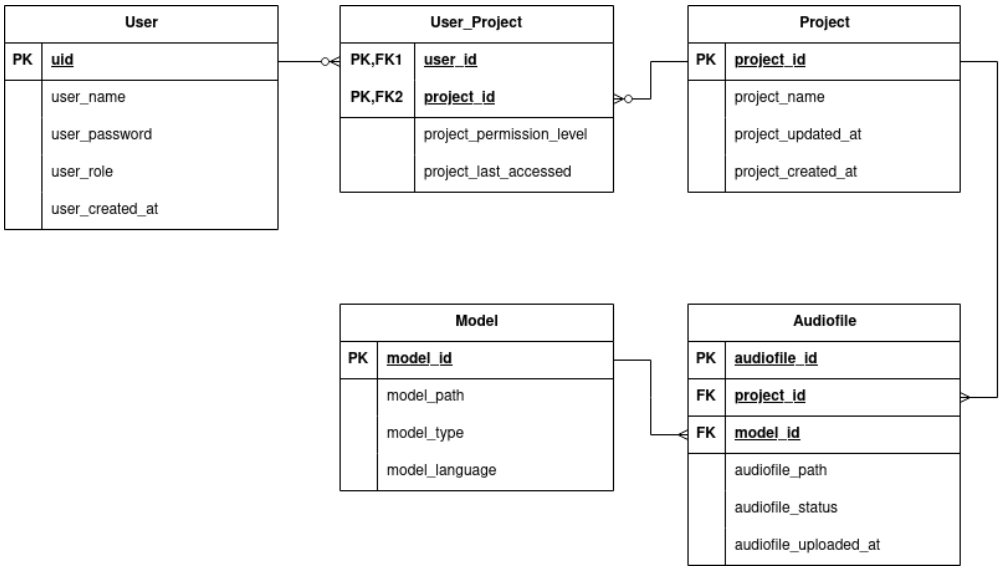


FIGURE 3.10 – Schéma préliminaire de la base de données.

connexion entre la table `User` et la table `Projet` signifiant qu'un usager peut participer à plusieurs projets et un projet peut avoir plusieurs usagers.

Nous reparlerons du schéma de la base de données plus en détails dans la partie développement. Comme on peut le remarquer la table `Audiofile` dispose d'une colonne `audiofile_path`. Cette colonne permet de pointer vers le dossier qui contiendra tous les éléments nécessaires à la transcription, notamment le fichier audio mais aussi les différentes itérations textuelles pour peut-être à terme permettre aux personnes utilisatrices d'avoir un historique des modifications.

## 3.5 Preuve de concept

Afin d'être plus conscient des difficultés et des contraintes liées au développement d'une interface de transcription de fichiers, nous avons d'abord souhaité mettre en place une preuve de concept. L'objectif est double : d'abord, appréhender la mise en place d'un tel système puis, ensuite, proposer cette preuve de concept aux communautés de recherche avant un outil plus poussé.

Ceci nous permet donc de collecter des retours sur les besoins et les difficultés liées à l'interface utilisateur.trice (IU) et l'expérience utilisateur.trice (XU). Ces points sont primordiaux puisque nous souhaitons produire les outils les plus accessibles possible. Ceci nous permet aussi de répondre,

même partiellement, à un besoin qui grandit rapidement. L'objectif est de rendre l'installation de ce prototype simple pour qu'il puisse être utilisable comme substitut à des alternatives payantes. De ce fait, nous utilisons Whisper (Radford *et al.*, 2022) pour l'identification de la langue ainsi que pour la transcription. Nous utilisons Pyannote (Bredin, 2023) pour la détection et séparation des locuteurs et locutrices.

Cette preuve de concept est réalisée en python et doit être utilisable entièrement en local. Elle ne permet que de prendre un ou plusieurs fichiers audio et de récupérer en sortie les fichiers textes associés avec ou sans séparation des locuteurs et locutrices. Ceci signifie donc que la durée du processus de transcription peut énormément varier, notamment si la machine est performante ou non, ou si elle est dôtée d'un GPU ou non. Si les machines sont trop peu performantes, il est aussi possible que le prototype ne fonctionne pas ou ralentisse beaucoup la machine lors du processus.

À titre informatif, l'utilisation du modèle large est impossible sur une machine avec les caractéristiques techniques suivantes :

```
Processeur: 16 × 12th Gen Intel® Core™ i7-1260P
Memoire: 31.1 GiB of RAM
Processeur graphique: Mesa Intel® Graphics (intégré)
```

Cependant, cette machine peut utiliser le modèle moyen.

L'interface graphique du prototype (figure 3.11) est constituée de trois sections.

La première section permet de visualiser les extraits audio dont la transcription est souhaitée. Les fichiers peuvent être sélectionnés en cliquant sur le bouton adéquat ou en les glissant déposant. Certains raccourcis sont disponibles, notamment pour la sélection rapide en utilisant les touches « contrôle » ou « majuscule ». Après avoir sélectionné plusieurs fichiers, il est possible de les supprimer en appuyant sur la touche de retour en arrière. Pour un seul fichier, double cliquer le supprime aussi de la liste.

La seconde section permet de sélectionner le dossier de destination des fichiers de transcription ainsi que la qualité voulue et l'utilisation de la diarisation. Le dossier de destination par défaut est celui

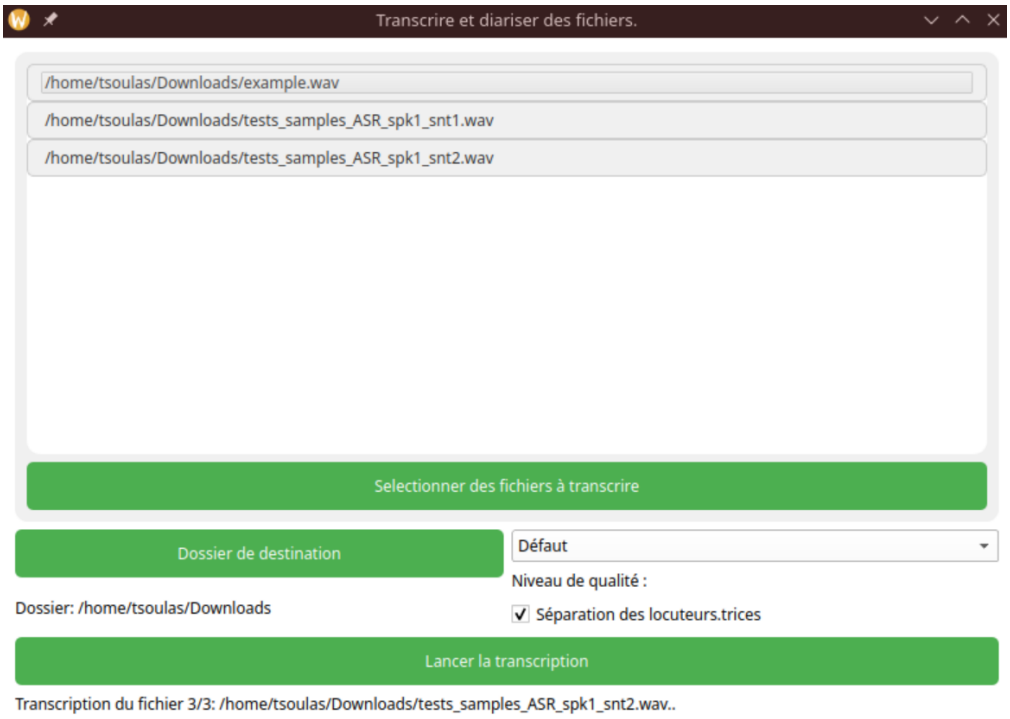


FIGURE 3.11 – Preuve de concept.

dans lequel est situé l'application. Les niveaux de qualité vont de « défaut » à « meilleure qualité », chaque niveau de qualité étant un modèle whisper différent. Le tableau 3.1 [69] représente les modèles disponibles dans le prototype et leur dénomination.

| Taille | Dénomination dans prototype | Paramètres | Modèle multilingue | VRAM requise | Vitesse relative |
|---|---|---|---|---|---|
| base | Défaut | 74M | base | ~1 GB | ~16x |
| small | Bonne qualité | 244M | small | ~2 GB | ~6x |
| medium | Haute qualité | 769M | medium | ~5 GB | ~2x |
| large | Meilleure qualité | 1550M | large | ~10 GB | 1x |

TABLE 3.1 – Taille et spécification des modèles Whisper.

Enfin, la troisième section permet de lancer le processus de transcription et de voir le statut actuel du processus. Lors de la première utilisation d'un modèle, celui-ci doit être téléchargé. La première transcription peut donc prendre un peu plus de temps que les suivantes.

69. Informations tirées de `https://github.com/openai/whisper` (excepté la dénomination dans le prototype)

Ci-dessous, voici le type de sortie auquel la personne utilisatrice peut s'attendre :

```
0.00 3.20 SPEAKER_00  The birch canoe slid on the smooth planks.


0.00 3.00 The child almost hurt the small dog.
```

Un extrait plus conséquent est disponible en annexe D.

# CHAPITRE 4
# SIMPLESPEECH

Le chapitre 4 présente SimpleSpeech, l'application de traitement automatique de la parole (TAP) que nous développons. Nous y aborderons les différents modules qui constituent SimpleSpeech ainsi que les améliorations et les possibilités d'évolution envisagées. Nous parlerons aussi du statut de développement de l'application.

## 4.1 Prototypage

Avant de commencer le développement de SimpleSpeech, nous avons d'abord tenté de réadapter notre preuve de concept au format web. Ce format web permet notamment d'utiliser un système de conteneurisation permettant à des programmes sous linux de tourner sur d'autres systèmes d'opération [70].


FIGURE 4.1 – Prototype web.

70. Nous utilisons Docker

Cette interface est une simple adaptation de notre preuve de concept au format web. Ce prototype est cependant plus portable et peut être déployé sur un serveur au besoin. Contrairement à SimpleSpeech il n'y a cependant pas d'authentification donc les fichiers téléversés ne sont pas protégés. Il est conseillé d'utiliser ce prototype sur sa machine ou bien sur un réseau privé. Nous utilisons les mêmes mécanismes que pour notre preuve de concept donc les mêmes résultats devraient être obtenus pour les deux applications.

## 4.2 Gestion de l'authentification

Puisque l'objectif est que cette application soit déployable en local et en serveur, il faut donc penser à un système d'authentification. Si l'application est déployée en local, alors il est plutôt simple de le détecter et de, soit désactiver l'authentification, soit créer un compte avec des identifiants simples. Pour une expérience fluide, il serait préférable de ne pas avoir à s'identifier. Pour permettre la migration sur un serveur au besoin, il est tout de même nécessaire d'attribuer les projets et autres données à un compte. L'objectif est donc que ceci soit invisible pour la personne utilisatrice mais ait tout de même bien lieu en fond pour permettre la prise en charge de ce genre de cas. En effet, il est fort possible qu'une personne utilisatrice commence avec un besoin assez limité, mais, que très rapidement, ce besoin soit trop conséquent pour permettre une exécution locale.

Ceci signifie qu'indépendamment du fait que l'application soit en local ou à distance, il est nécessaire d'avoir une gestion d'utilisateurs et utilisatrices. L'authentification a lieu côté IPA. Pour une authentification sécurisée et une bonne expérience utilisateur.trice (XU), deux éléments sont à prendre en compte : la sécurisation des mots de passe et la gestion de "session".

"Sécuriser les mots de passe" signifie empêcher ou réduire la possibilité qu'une personne malveillante puisse récupérer les mots de passes dans le cas d'une brèche. Dans le cas d'une application telle que celle-ci, l'ajout d'une telle sécurité peut être discutable. Cependant il ne faut pas oublier que beaucoup de personnes utilisatrices utilisent le même mot de passe à travers plusieurs applications en dépit du fait que cela soit une très mauvaise pratique. Le standard de protection des mots de passe est de "hacher" le mot de passe. À la différence du chiffrement, le hachage est unidirectionnel. Un texte chiffré peut être déchiffré. Un texte haché ne peut pas être "déhaché". Le protocole HTTPS, par exemple, permet de sécuriser nos recherches internet en chiffrant les échanges entre notre navigateur

et les sites web sur lesquels nous naviguons. Le hachage est quant à lui généralement utilisé pour vérifier l'intégrité des données. Typiquement, dans le cas de vérification de mot de passe, ce n'est pas le mot de passe que nous vérifions. nous stockons la version haché et lors de chaque connexion nous comparons le hache du mot de passe donné avec celui stocké.

Pour que la personne utilisatrice puisse naviguer dans l'application sans avoir à se connecter très souvent, il faut que l'application soit capable de reconnaître celui-ci. Il est possible de transférer les identifiants à chaque requête mais ce n'est pas du tout sécurisé. En effet, en cybersécurité le risque zéro n'existe pas et plus un élément est transféré souvent, plus il risque d'être récupéré par une personne malveillante même si il est chiffré. L'objectif en cybersécurité n'est pas d'avoir une sécurité inébranlable mais surtout d'avoir une sécurité assez importante pour dissuader les attaquants en rendant le coût de l'attaque supérieur aux gains potentiels. Il est aussi possible de demander à la personne utilisatrice de donner ses identifiants à chaque connexion ou bien de faire stocker les informations dans le navigateur. Dans le premier cas, ce n'est pas confortable en termes d'XU, et dans le deuxième cas cela ajoute une vulnérabilité.

Un bon compromis est de communiquer avec le serveur pour avoir un identifiant temporaire, généralement appelé « jeton ». Ce jeton peut être décodé par le serveur et contient des informations utiles mais non compromettantes. Comme par exemple, l'id de la personne et le temps avant expiration du jeton.

In fine, une personne utilisatrice donne ses identifiants pour que ceux-ci soient vérifiés. D'abord en vérifiant que le compte existe puis en vérifiant la version haché du mot de passe. Ensuite, si les vérifications sont passées, un jeton temporaire est fourni pour une durée donnée. Pour toute action, ce jeton est à fournir. L'application vérifie constamment la véracité du jeton et, si celui-ci périme ou n'est pas valide, le client redirigera automatiquement à la page de connexion tandis que l'IPA refusera toute requête.

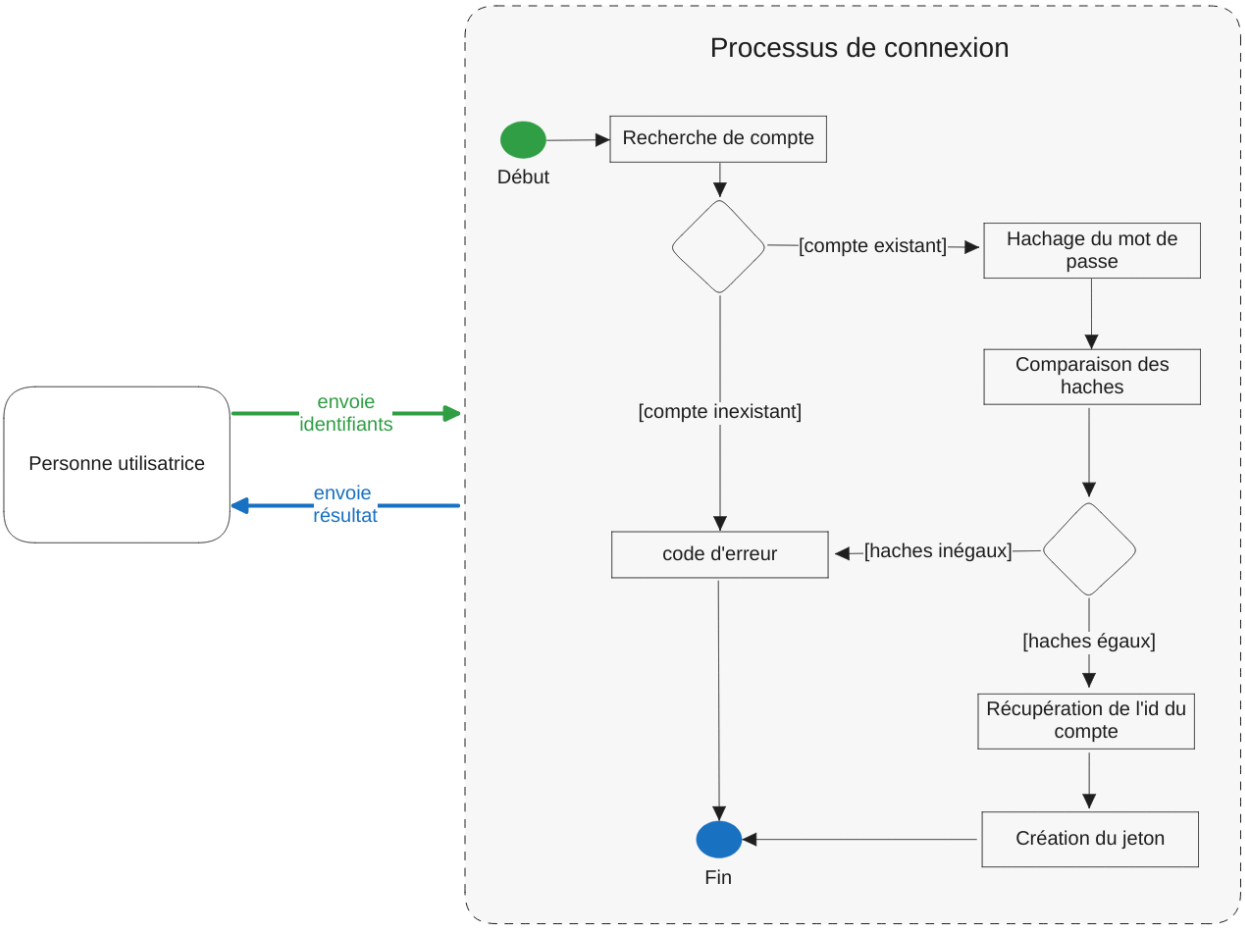


FIGURE 4.2 – Diagramme d'interactions lors de la connexion.

### 4.2.1 Hiérarchisation des projets

SimpleSpeech est une application qui peut être utilisée en local pour une personne ou en communauté. Si elle est déployée sur un serveur, elle doit permettre l'authentification par personne mais aussi la gestion de multiples projets. En tant que personne utilisatrice, cela permet l'organisation suivante :

```
Projet 1 (accessible par : personne A, personne B)
    |   --    Audio A
    |           |   -- Transcription A1
    |   --    Audio B
    |           |   -- Transcription B1
```

```
Projet 2 (accessible par : personne A, personne C, personne D)
    |   --    Audio C
    |           |   -- Transcription C1
    |           |   -- Transcription C2
```

Un projet peut être créé par n'importe quelle personne qui en a les droits. La personne créant ce projet en devient l'administratrice et peut ainsi ajouter des collaborateurs et collaboratrices. Un sous-système de droit est mis en place pour permettre à chaque projet d'avoir ses propres personnes administratrices. Le rôle administrateur global a cependant la priorité sur le rôle administrateur projet. Les projets ne sont visibles et accessibles que si le droit est donné, ce qui permet de cloisonner les projets si nécessaire. C'est notamment intéressant dans le cas d'une instance de SimpleSpeech partagée par plusieurs groupes. Toutefois, si le serveur est accessible par l'un des groupes ou une personne tierce, cela est à prendre en compte dans le cadre d'un certificat éthique d'un autre groupe.

Pour chaque audio, nous souhaitons aussi permettre l'ajout et la consultation d'un arbre des changements ou même de plusieurs versions. Ceci serait à l'image des outils de contrôle de version tels que Git [71]. Ceci n'est cependant pas une priorité de développement et n'est donc pas un point sur lequel nous avons travaillé en détails.

71. https://git-scm.com/

## 4.3 Gestion de l'édition manuelle

Toute l'interface d'édition est encore en phase de réflexion. Cette interface doit permettre à la personne utilisatrice de voir le résultat de la transcription et de le corriger ou même l'éditer. De ce fait, la personne utilisatrice doit pouvoir accéder à l'audio et à sa transcription en parallèle pour pouvoir comparer les deux et réécouter les extraits incertains. Chaque mot doit être aligné à son équivalent dans l'extrait audio dont il est question. Quand la personne utilisatrice clique sur un mot alors elle est redirigée à la seconde où il est dit. Lorsqu'elle clique sur un passage de l'extrait alors elle est redirigée au mot écrit relatif à cet extrait. Il faudrait aussi permettre la correction de multiples instances, notamment dans le cas d'entités nommées (comme les noms propres) qui, si elles sont incorrectes, risquent souvent d'être incorrectes de la même façon.

En termes d'interface, l'objectif serait d'atteindre un résultat proche de la figure 4.3 [72]. D'abord sans la gestion des étiquettes (orateur 1, orateur 2, bruit, ...). Ensuite avec, une fois que le serveur supportera la diarisation.


FIGURE 4.3 – Interface d'annotation de StageZero.

La différence est cependant que nous souhaitons être capables d'avoir l'horodatage au niveau des mots, ce qui veut dire qu'on ne peut pas se baser entièrement sur une interface telle que celle-ci

72. Source : `https://stagezero.ai/audio-annotation-tool/`

puisque faire un segment par mot serait illisible. Une option serait de reprendre cette interface et que SimpleSpeech s'aligne automatiquement sur des segments d'une durée donnée, potentiellement choisie par la personne utilisatrice. Ces segments seraient donc alignés sur $x$ secondes et les mots de ces segments seraient connectés à leurs périodes de temps respectives. Une limitation conséquente provient des potentiels conflits entre la transcription donnée par SimpleSpeech et la correction donnée par l'outil. Il est notamment important de savoir quel niveau de liberté est donné à la personne utilisatrice, puisque plus le niveau de liberté est élevé, plus la cohérence de la transcription relève des décisions prises par l'usager. C'est à dire que pour un segment de 4 secondes :

> « J'aime les chats et les chiens. »

La personne utilisatrice peut très bien changer pour

> « J'aime les chats, les chiens ainsi que le chocolat et le lait d'amande. »

Et cette correction sera renvoyée à SimpleSpeech qui doit maintenant être capable de réagir correctement à cette correction. Par réagir correctement, on entend réattribuer un horodatage correct ou bien choisir d'attribuer à cet ensemble de mots l'horodatage du mot qui a été remplacé. Une autre option est de permettre ou de forcer la personne utilisatrice à choisir un horodatage pour cet ensemble de mots.

Enfin, l'interface permet l'édition et la correction mais elle doit aussi permettre l'obtention du résultat dans un ou plusieurs formats de fichiers à télécharger. Les fichiers VTT, SRT et TXT sont des formats couramment utilisés. Permettre d'avoir le résultat au format PDF ou DOCX peut être aussi intéressant.

## 4.4 Gestion de la transcription

Le module de transcription est le module qui requiert le plus de travail et de recherche. Il est central à l'application. Ce module est de prime abord le module prenant en entrée un fichier audio pour produire en sortie une transcription de ce dernier. Il doit cependant être évolutif et permettre l'ajout de fonctionnalités. L'une des fonctionnalités les plus importantes à terme est la possibilité de détecter et différencier les locuteurs.

Nous nous concentrons sur la gestion de la transcription et les possibilités du module dans sa version actuelle. En plus de la transcription, il est important de mettre à disposition la possibilité de corriger par un modèle de langue et de gérer les extraits longs. Ces fonctionnalités sont surtout utiles côté serveur et sont invisibles pour la personne utilisatrice.

Pour le client, l'alignement de l'audio avec le texte est nécessaire pour offrir une expérience de correction et d'édition optimale.

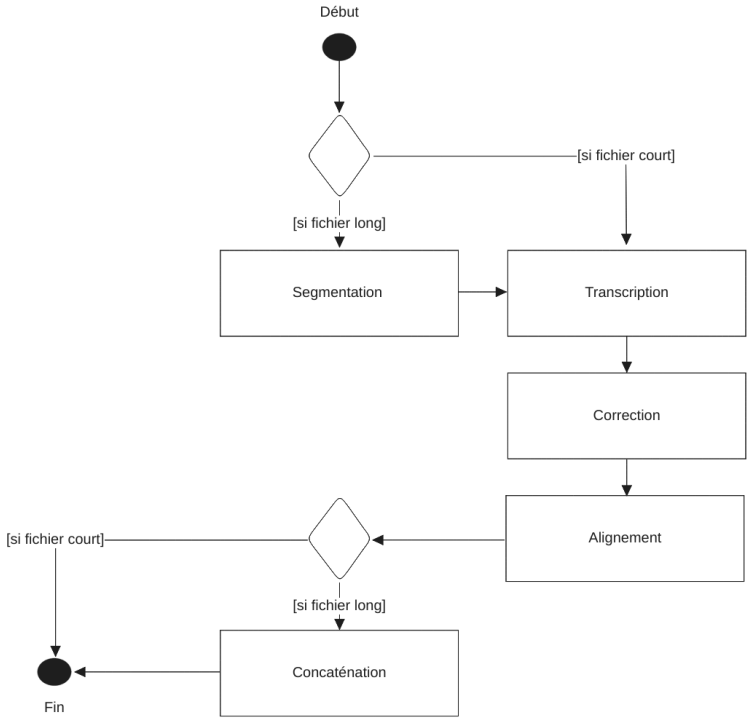


FIGURE 4.4 – Diagramme d'activité du module de transcription.

## 4.5 Transcription

Pour la transcription, nous devons être capable de proposer des modèles à l'état de l'art pour permettre aux personnes utilisatrices d'avoir la meilleur transcription possible. Bien que certains modèles ne soient pas adaptés par SpeechBrain ou NeMo, ces deux derniers proposent une interface et une liaison avec HuggingFace qui permet la mise en place et l'utilisation des modèles de transcription de manière simple. Utiliser Hugging Face est d'autant plus intéressant que cela donne accès à la liste des modèles proposés [73] [74]. Par exemple, tous les modèles SpeechBrain pour le TAP sont nommés `asr-X` où `X` est le nom du modèle.

Dans le cadre du développement, nous nous concentrons cependant sur Whisper et les versions Whisper de SpeechBrain qui sont modifiées pour être optimales pour une langue donnée. L'outil doit être utilisable sans connaissance préalable. Il n'est donc pas nécessaire de surcharger l'interface d'informations inutiles. Par contre, il est nécessaire d'être capable de suivre l'état de l'art. De plus, les modèles proposés par défaut sont des modèles qui doivent fonctionner en local sur des machines à faible puissance de calcul. Dans le cadre où SimpleSpeech serait déployé sur des instance avec de plus grandes ressources de calcul, il faudrait alors être capable, par exemple, d'utiliser `canary-1b` [75] ou la dernière version large de Whisper.

Aujourd'hui SimpleSpeech implémente SpeechBrain et donc les modèles proposés par ce dernier. Il est aussi possible d'aisément implémenter Whisper puisque nous l'avons déjà implémenté dans le cadre de notre preuve de concept.

73. `https://huggingface.co/speechbrain`

74. `https://huggingface.co/nvidia`

75. `https://huggingface.co/nvidia/canary-1b`

### 4.5.1 Modèles de langue

Il peut être intéressant de tenter de corriger la transcription obtenue. Une méthode simple et un peu naïve est de comparer chaque mot à un dictionnaire ou lexique de mots. Si un mot n'est pas dans le lexique, alors on tente de trouver le plus proche à l'aide d'une distance d'édition comme la distance de Levenshtein [76] évoquée section 2.2. Cette approche a cependant ses limites, la limite la plus contraignante étant le manque de contexte. En effet, dans la phrase « J'aime lo chat de Tom. », « lo » est incorrect. Suivant cette technique, le mot correct peut alors être « le » ou « la » par exemple. À l'aide du contexte nous pouvons déterminer que le mot correct est « le », on peut aussi constater que « le » est phonétiquement relativement proche de « lo ». Cependant l'approche par distance de Levenshtein ne pourra pas déterminer la bonne proposition.

Une piste intéressante pour la correction de transcription est d'utiliser des modèles de langue. Un modèle de langue est un modèle qui permet de prédire la distribution de symboles distincts (comme des lettres, des phonèmes ou des mots) dans une langue naturelle. Ces modèles sont entre autres utilisés pour de la traduction automatique d'une langue écrite à une autre ou pour identifier une langue écrite. Dans le cadre du TAP, ces modèles de langue aident à estimer la probabilité des différentes séquences de mots, ce qui permet de choisir la séquence de mots la plus probable. Certains outils de transcription utilisent déjà des modèles de langues, Whisper en est un bon exemple (Radford *et al.*, 2022). SpeechBrain (Ravanelli *et al.*, 2021) et NeMo (Kuchaiev *et al.*, 2019) permettent tous deux d'incorporer un modèle de langue dans le processus de transcription. Le travail sur l'ajout de modèles de langue est encore en cours de réalisation.

76. confère `https://norvig.com/spell-correct.html`

### 4.5.2 Alignement mot au texte

L'alignement forcé est une technique employée dans le cadre du TALN pour aligner des éléments linguistiques entre deux textes ou corpus. Dans le contexte de la parole, ceci se réfère à la synchronisation d'un enregistrement audio avec sa transcription en alignant chaque segment de parole prononcée avec les mots correspondants dans la transcription.

Le Montreal Forced Aligner (MFA) (McAuliffe *et al.*, 2017) permet de synchroniser l'audio avec sa transcription au mot voire au phonème près. Dans le cadre de SimpleSpeech, une synchronisation au mot est suffisante puisque nous utilisons l'alignement seulement pour permettre aux personnes utilisatrices de se repérer entre l'extrait audio et sa transcription. Gentle [77] et NeuFA (Li *et al.*, 2022) sont aussi des outils d'alignement forcé mais ceux-ci se concentrent sur l'anglais. MFA supporte plusieurs dizaines de langues et obtient des résultats à l'état de l'art pour beaucoup d'entre elles. MFA garantit son évolutivité en se basant sur Kaldi (Povey *et al.*, 2011) et SpeechBrain (Ravanelli *et al.*, 2021).

La tâche d'alignement est une tâche qui est très rapide et qui peut être parallélisée même sur une machine à faibles performances. Les segments sont donc transcrits puis la transcription est utilisée pour aligner ces segments avec MFA.

MFA produit des fichiers `TextGrid` dont le contenu ressemble à ceci :

```
        intervals [15]:
            xmin = 3.8
            xmax = 3.91
            text = "et"
        intervals [16]:
            xmin = 3.91
            xmax = 4.27
            text = "maintenant"
        intervals [17]:
```

77. `http://lowerquality.com/gentle/`

```
        xmin = 4.27
        xmax = 4.51
        text = ""
    intervals [18]:
        xmin = 4.51
        xmax = 4.79
        text = "pilippe"
    intervals [19]:
        xmin = 4.79
        xmax = 5.25
        text = "gagnier"
```

où `xmin` et `xmax` sont les horodatages en secondes de début et de fin d'un intervalle donné. Un intervalle correspond à un mot ou à un silence. Ce fichier contient les mêmes informations au phonème près. Ces informations sont suffisantes pour notre usage.

### 4.5.3 Gestion des fichiers longs

Les modèles généralement les plus performants en transcription d'audio vers du texte sont ceux qui prennent en compte le contexte. Prendre en compte le contexte nécessite d'allouer de la mémoire à ce contexte pour être capable d'y faire référence lors de la prédiction de la transcription. De ce fait, plus la séquence est longue, et plus la mémoire allouée est importante.

Or, pouvoir faire référence au contexte est important mais il n'est pas nécessaire d'avoir l'entièreté du contexte pour pouvoir transcrire de manière correcte. De plus, les corpus d'entraînement sont généralement constitués de fichiers relativement courts (5-15 secondes), tenter de transcrire de longues séquences avec des modèles qui ont été entraînés sur de courtes séquences crée une divergence. Il est donc nécessaire de trouver une façon de couper ces fichiers.

À noter que Whisper est utilisable pour des fichiers longs et va automatiquement prendre des fenêtres d'une taille correcte tout en récupérant un peu de contexte avant et après. Ceci fait de Whisper un système intéressant à retenir pour le traitement d'enregistrements audio dans SimpleSpeech.

Trois types d'approche peuvent pallier ce problème. D'abord, l'approche naïve consiste à couper les séquences toutes les x secondes sans regard pour leur contenu ou pour ce qui pourrait être perdu dans le processus. Cette approche est l'approche la moins coûteuse en calcul mais aussi la plus propice à la perte d'information. Une autre approche, superposition d'inférences (SdI), peut être de séquencer en commençant à la moitié du segment précédent de sorte à pouvoir superposer les séquences (Chiu *et al.*, 2019; Huang *et al.*, 2022). Cette approche requiert cependant deux fois la puissance de calcul requise pour l'approche naïve puisqu'il s'agit essentiellement de transcrire deux fois le fichier de manière différente. Et enfin, une dernière approche peut être de couper seulement lorsqu'on ne détecte plus de voix pendant un instant (Huang *et al.*, 2022). On utilise les principes de détection d'activité vocale (DAV). La détection de la voix pose cependant problème lorsque le segment devient trop long à cause d'un flux ininterrompu de parole.

De ces approches découlent deux autres possibilités :

— la superposition partielle d'inférences (SPI)(Kang *et al.*, 2021),
— la détection d'activité vocale et superposition d'inférences (DAVSdI)(Wang *et al.*, 2022b).

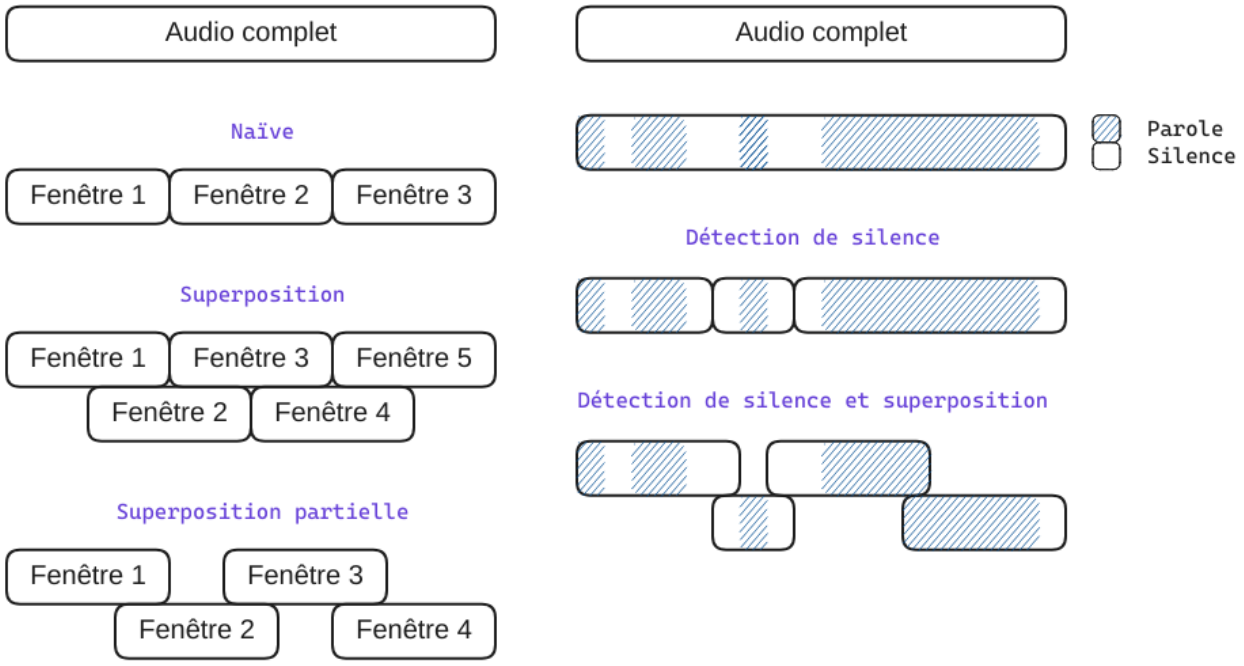


FIGURE 4.5 – Approches de segmentation de longs fichiers.

D'après (Wang *et al.*, 2022b), DAVSdI produit de meilleurs résultats que SPI si 30% de la durée totale de la fenêtre en cours a déjà traitée par la fenêtre précédente. On parle alors ici de pourcentage de superposition. Lorsque la fenêtre est superposée à 50% avec celle qui la précède, les résultats sont légèrement moins bons avec DAVSdI. Il est aussi montré que la SPI est plus intéressante que la SdI tant en termes de temps de calcul qu'en termes de résultats. Toutefois, la différence de résultats est minime.

Quelle que soit l'approche choisie, après avoir segmenté et transcrit les fichiers il faut être capable de les re-concaténer. Dans le cas du DAV, il n'est pas nécessaire de regarder le contexte précédent puisque les segments ne sont normalement pas coupés au milieu d'un mot. Dans le cas d'une SdI, SPI ou DAVSdI, il est nécessaire de vérifier la cohérence avec la fenêtre précédente.

L'article (Kang *et al.*, 2021) met en avant une technique de re-concaténation des transcriptions utilisant le principe de distance de Levenshtein. Nous avons ainsi la $i^{\text{ième}}$ fenêtre dénotée $d_i$ ainsi que la distance $C(j,k)$ où $j$ est le $j^{\text{ième}}$ caractère $d_i$ et $k$ est le $k^{\text{ième}}$ caractère $d_{i+1}$ et la longueur de $d_i$ notée $n_i$. Les poids des différentes actions sont modifiés comme suit :

— Insertion et suppression sans correspondance préalable $w_{ins,non}, w_{del,non} = 0$

— Insertion et suppression $w_{ins}, w_{del} = 2$

— Substitution $w_{sub} = 1$

— Correspondance $w_{match} = -2$

Avoir une correspondance est récompensé par l'algorithme. Puisque ne pas avoir de correspondance préalable est vu comme neutre, lorsqu'il y a une correspondance on attribue une valeur négative. $C(j, 0)$ et $C(0, k)$ sont définis comme suit :

$$C(j, 0) = w_{del,non} \times j, \qquad 0 \leq j \leq n_i$$
$$C(0, k) = w_{ins} \times k, \qquad 1 \leq k \leq n_{i+1}$$

Puisque partiellement superposé implique que la fenêtre $d_{i+1}$ commence après le début de la fenêtre $d_i$, il est nécessaire de ne pas pénaliser les erreurs au début de la fenêtre. De plus, nous augmentons la pénalité si la fenêtre $d_i$ dépasse la fenêtre $d_{i+1}$.

Similaire à la distance de Levenshtein, le schéma récursif est le suivant :

$$e_{sub}(j, k) = \begin{cases} w_{match}, & \text{si } d_i(j) = d_{i+1}(k) \\ w_{sub}, & \text{sinon} \end{cases}$$

$$C(j, k) = min \begin{cases} C(j-1, k) + w_{del}, \\ C(j, k-1) + w_{ins}, \\ C(j-1, k-1) + e_{sub}(j, k), \end{cases} \qquad 1 \leq j < n_i, 1 \leq k \leq n_{i+1}$$

$$C(j, k) = min \begin{cases} C(j-1, k) + w_{del}, \\ C(j, k-1) + w_{ins,non}, \\ C(j-1, k-1) + e_{sub}(j, k), \end{cases} \qquad j = n_i, 1 \leq k \leq n_{i+1}$$

Le schéma récursif ci-dessus est conditionné pour prendre en compte le fait que $j$ ait atteint ou non la fin de la fenêtre. S'il l'a atteint, alors il n'y a pas de pénalité.

Les valeurs de la distance de Levenshtein peuvent être négatives et on cherche à concaténer les textes après avoir trouvé la distance la plus petite. Cette distance est retrouvée après avoir utilisé

un algorithme de retour en arrière. Le tableau 4.1 permet de visualiser le calcul pour chaque lettre des mots « cognition » et « speech recognize » (respectivement $d_{i+1}$ et $d_i$).

Après avoir trouvé la distance la plus courte, on concatène à partir du moment où $k$ et $j$ sont strictement supérieurs à 0. S'il n'y a pas de mot commun dans la superposition, alors la priorité est donnée à $d_{i+1}$ et toute correspondance est ignorée.

Dans le cas du tableau 4.1, le mot concaténé est donc « speech recognition ».

| | | **c** | **o** | **g** | **n** | **i** | **t** | **i** | **o** | **n** |
|---|---|---|---|---|---|---|---|---|---|---|
| | **0** | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
| **s** | **0** | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | 17 |
| **p** | **0** | 1 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |
| **e** | **0** | 1 | 2 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
| **e** | **0** | 1 | 2 | 3 | 4 | 6 | 8 | 10 | 12 | 14 |
| **c** | **0** | -2 | 0 | 2 | 4 | 5 | 7 | 9 | 11 | 13 |
| **h** | **0** | 0 | -1 | 1 | 3 | 5 | 6 | 8 | 10 | 12 |
| | **0** | 1 | 1 | 0 | 2 | 4 | 5 | 7 | 9 | 11 |
| **r** | **0** | 1 | 2 | 2 | 1 | 3 | 5 | 7 | 8 | 10 |
| **e** | **0** | 1 | 2 | 2 | 1 | 3 | 5 | 7 | 8 | 9 |
| **c** | 0 | **-2** | 0 | 2 | 4 | 4 | 3 | 5 | 7 | 9 |
| **o** | 0 | 0 | **-4** | -2 | 0 | 2 | 4 | 4 | 3 | 5 |
| **g** | 0 | 1 | -2 | **-6** | -4 | -2 | 0 | 2 | 4 | 4 |
| **n** | 0 | 1 | 0 | -4 | **-8** | -6 | -4 | -2 | 0 | 2 |
| **i** | 0 | 1 | 2 | -2 | -6 | **-10** | -8 | -6 | -4 | -2 |
| **z** | 0 | 1 | 2 | 0 | -4 | -8 | **-9** | -7 | -5 | -3 |
| **e** | 0 | 0 | 0 | 0 | -2 | -6 | -7 | **-8** | **-8** | **-8** |

TABLE 4.1 – Table de calcul de Levenshtein modifiée entre "cognition" et "speech recognize".

Cette technique n'est pas infaillible et il peut être très intéressant d'utiliser ensuite un modèle de langue sur le résultat.

## 4.6 Synthèse

Les codes de SimpleSpeech et ses prototypes sont disponibles dans les dépôts de code suivants :

**Preuve de concept** - `https://gitlab.labikb.ca/ikb-lab/simplespeech/POC-Whisker/-/tree/main`

**Premier Prototype** - `https://gitlab.labikb.ca/ikb-lab/simplespeech/POC-Whisker/-/tree/f_webgui`

**SimpleSpeech** - `https://gitlab.labikb.ca/ikb-lab/simplespeech`

Ces derniers sont sous licence « Creative Commons Attribution-NonCommercial »(CC BY-NC). Ce qui signifie qu'ils sont gratuits, modifiables et libre d'accès mais doivent être crédités et ne peuvent être utilisés à des fins commerciales.

SimpleSpeech est actuellement en cours de développement et toutes les fonctionnalités ne sont pas encore disponibles. Il est possible de s'identifier et de gérer des projets par groupe.

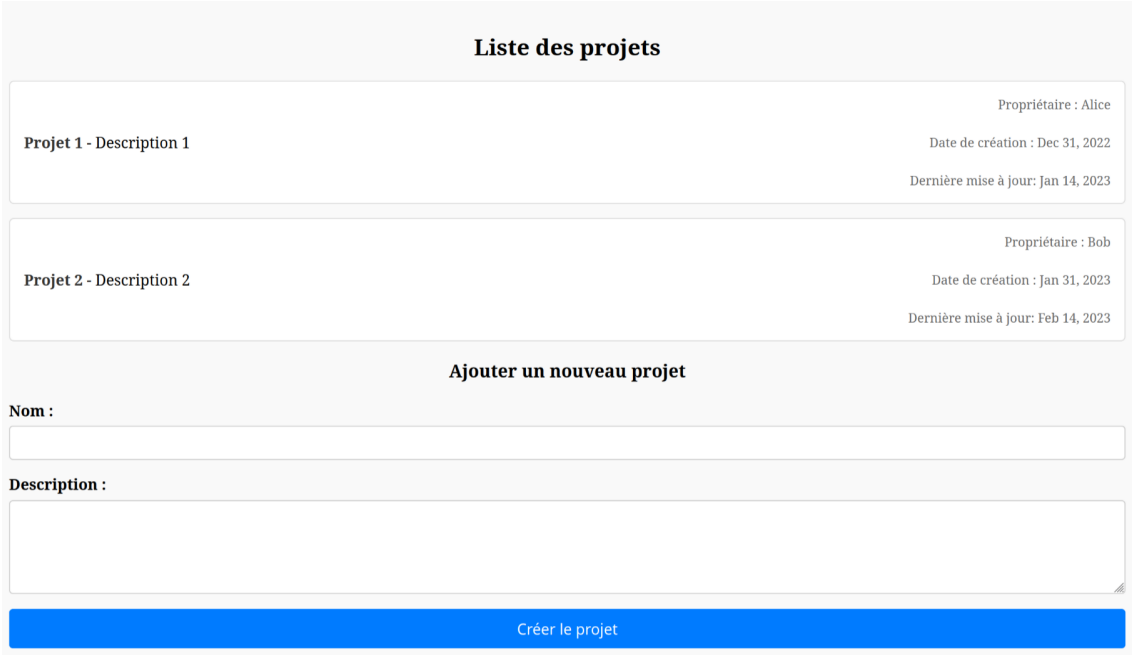


FIGURE 4.6 – Interface préliminaire de gestion des projets.

Un projet permet de mettre un ou plusieurs fichiers audio et de demander la transcription correspondante.

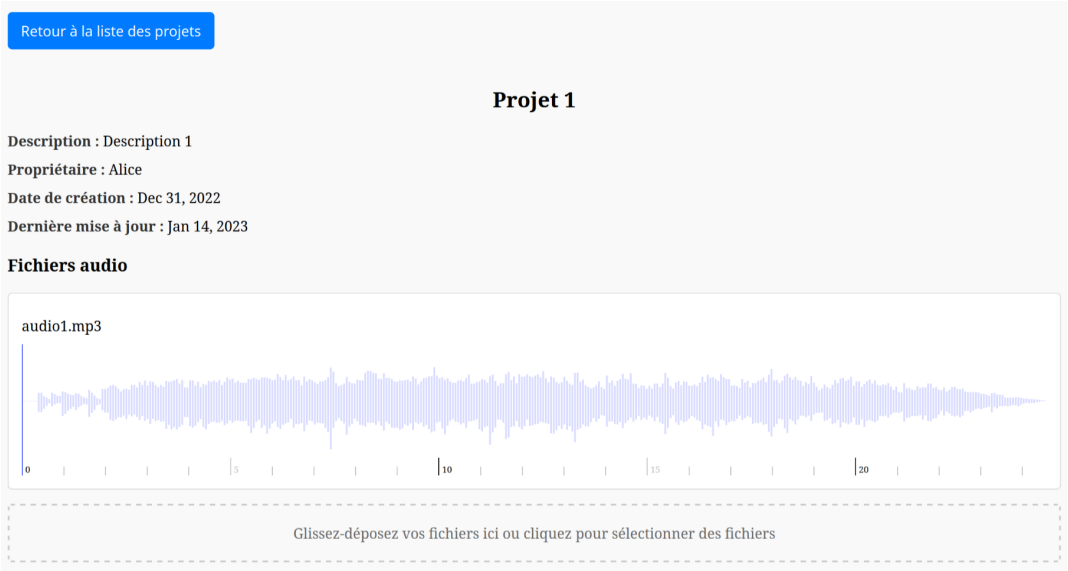


FIGURE 4.7 – Interface préliminaire de gestion des audios.

Les transcriptions peuvent être obtenues en communiquant par le serveur. Le système de transcription est mis en place mais toute la partie cliente n'est pas encore prête. En effet, l'interface permettant de faire les modifications et d'accéder aux fichiers audio est en cours de développement. Enfin, SimpleSpeech peut être déployé dans un conteneur. Ceci permet de limiter les potentiels bogues au déploiement et de faciliter son déploiement sur différents systèmes d'exploitations.

L'IU est elle aussi à raffiner et nécessite une certaine réflexion en termes de forme. Il est notamment important que l'interface de correction et d'édition soit la plus complète et facile d'utilisation possible. Pour cela, nous comptons sur les retours sur la preuve de concept que nous avons réalisée ainsi que sur les retours futurs sur les premières version de SimpleSpeech. Toutefois, la majorité du travail sur les fondations de SimpleSpeech est fait. Les modules de gestion des projets et de gestions des personnes utilisatrices sont terminés. De nombreux éléments de transcription sont aussi prêts à être utilisés. La communication client vers serveur et vice versa est elle aussi prête et notre serveur dispose d'une documentation dans le cas où des développeurs souhaiteraient travailler seulement avec le serveur et créer un client différent. La première version de SimpleSpeech et toutes versions qui suivront seront en libre accès.

Les prochaines étapes de développement sont de préparer SimpleSpeech pour du développement communautaire et de terminer l'interface de correction et d'édition.

# CONCLUSION

Dans le cadre du développement de SimpleSpeech, nous avons pu faire tester la preuve de concept ainsi que le premier prototype par une équipe de recherche de l'UQAM. Ces premiers tests ont été effectués dans le cadre d'un projet d'évaluation d'outils d'intervention psychologique. Les enregistrements sont constitués de groupes de discussion et d'entretiens individuels avec des personnes intervenantes et des personnes utilisatrices en français québécois. Nous avons utilisé une machine avec les caractéristiques suivantes :

```
Processeur :12th Gen Intel® Core™ i7-12650H × 16
Mémoire : 16 Go
```

En activant l'identification de locuteur, le temps de traitement pour chaque extrait est d'approximativement 1 minute de traitement pour 30 secondes d'audio.

| Enregistrement | Durée | Temps de calcul |
|---|---|---|
| entretien | 0h35 | 0h50 |
| entretien | 0h34 | 0h52 |
| groupe | 1h41 | 4h15 |
| groupe | 1h44 | 3h49 |

TABLE 4.2 – Temps de transcription pour des entretiens

La personne utilisatrice nous a rapporté que l'outil avait des difficultés à orthographier correctement les entités nommées, telles que des prénoms ou des noms de lieu. Celle-ci nous a aussi fait remarquer que l'identification des locuteurs était régulièrement erronée. L'outil lui est cependant très utile et lui permet de faire son travail en 3 ou 4h de temps pour 1h d'entrevue, alors que cette équipe prévoit en général 6h pour 1h d'entrevue. Ce sont des résultats encourageants.

SimpleSpeech est en cours de développement et certaines fonctionnalités nécessitent d'être raffinées. Le module de transcription est notamment un module qui doit prendre en charge un très grand nombre de possibilités, ce qui le rend aussi sujet à de nombreuses contraintes. Notamment sur la gestion de multiples modèles ainsi que la segmentation et la reconcaténation. SimpleSpeech ne pourra pas répondre à tous les besoins. Certains algorithmes que SimpleSpeech utilise doivent être

implémentés à partir de la production scientifique à l'état de l'art, ce qui ajoute un temps de développement et de maintenance qui ne permet pas à SimpleSpeech d'être toujours à l'état de l'art en temps réel. La gestion des extraits longs est par exemple un sujet dans lequel diverses solutions sont proposées. Certaines de ces solutions sont plus simples à implémenter et moins coûteuses en ressources de calcul mais obtiennent des résultats moins bons que d'autres approches qui sont plus coûteuses. Un compromis est donc nécessaire. Ce compromis est autant valable pour les personnes développant que pour les personnes utilisant l'application.

Nous souhaitons pour l'instant consolider les fondations de SimpleSpeech et nous nous concentrons donc sur la transcription. Le cœur de SimpleSpeech est d'être utilisable par toutes et tous en local ou sur un serveur donné pour obtenir des résultats de transcription sur des extraits audio longs ou courts. À terme, nous aimerions ajouter à la transcription d'autres techniques de traitement automatique de la parole. La séparation et l'identification des locuteurs sont notamment très demandées lors de l'usage d'applications de transcription. Certaines personnes utilisatrices pourraient aussi avoir besoin de résumer un audio plutôt que d'avoir accès à sa transcription.

Rajouter toutes ces fonctionnalités soulève toutefois un point important. Chaque personne utilisatrice n'aura pas les même besoins. Est-il sage de faire télécharger à une personne tout un module de résumé quand celle-ci souhaite seulement une transcription ? L'approche logique serait de permettre à l'application de télécharger et d'utiliser des modules au choix et au besoin des personnes utilisatrices.

Nous travaillons aussi sur la mise à disposition de ressources québécoises, notamment grâce aux commissions Viens et Charbonneaux. Ces dernières permettront à la communauté québécoise d'être plus représentée dans le cadre d'apprentissage de modèles de TAP. Nous souhaitons aussi affiner ces modèles en continuant leur apprentissage sur une tâche de transcription afin de pouvoir les utiliser dans SimpleSpeech. Ceci permettrait de s'adapter aux particularismes du français québécois et de proposer une transcription plus fidèle. Nous pourrions, par exemple, reprendre les travaux de SpeechBrain avec Whisper et l'affiner sur la commission Viens.

# ANNEXE A

# EXEMPLE DE CODE UTILISÉ DANS LE CADRE DE CERTAINES BOÎTES À OUTILS

```python
# Tiré de https://github.com/alphacep/vosk-api/blob/master/python/example/test_simple.py
from vosk import Model, KaldiRecognizer
import wave

# Récupère le fichier audio
audio_file = "example.wav"
wf = wave.open(audio_file, "rb")

# Déclare le modèle
model = Model(lang="en-us")
rec = KaldiRecognizer(model, wf.getframerate())
rec.SetWords(True)
rec.SetPartialWords(True)

# boucle sur la sortie jusqu'à obtenir le résultat final
while True:
    data = wf.readframes(4000)
    if len(data) == 0:
        break
    if rec.AcceptWaveform(data):
        rec.Result()
    else:
        rec.PartialResult()

# Affiche le résultat
print(rec.FinalResult())
```

FIGURE A.1 – Transcription avec horodatage d'un fichier audio en utilisant Vosk.

```python
# Tiré de https://espnet.github.io/espnet/notebook/espnet2_asr_realtime_demo.html
import string
import soundfile
from espnet_model_zoo.downloader import ModelDownloader
from espnet2.bin.asr_inference import Speech2Text

# déclarations
tag = 'Shinji Watanabe/spgispeech_asr_train_asr_conformer6_n_' + \
    'fft512_hop_length256_raw_en_unnorm_bpe5000_valid.acc.ave'
file_name = "example.wav"
d = ModelDownloader()
speech2text = Speech2Text(
    **d.download_and_unpack(tag),
    minlenratio=0.0,
    maxlenratio=0.0,
    ctc_weight=0.3,
    beam_size=10,
    batch_size=0,
    nbest=1
)


# Lecture et transcription de l'audio
speech, rate = soundfile.read(file_name)
nbests = speech2text(speech)
text, *_ = nbests[0]

print(f"ASR hypothesis: {text}")
```

FIGURE A.2 – Transcription sans horodatage d'un fichier audio en utilisant ESPnet.

```python
import nemo.collections.asr as nemo_asr

# Configuration du modèle et déclaration du fichier à transcrire
asr_model = nemo_asr.models.EncDecCTCModel.from_pretrained(model_name='QuartzNet15x5Base-En', strict=False)
files = ['example.wav']

# Transcription
transcript = asr_model.transcribe(paths2audio_files=files)[0]

# Affiche le résultat
print(transcript)
```

FIGURE A.3 – Transcription sans horodatage d'un fichier audio en utilisant NeMo.

```python
from speechbrain.pretrained import EncoderDecoderASR
from speechbrain.alignment.ctc_segmentation import CTCSegmentation

source_model="speechbrain/asr-transformer-transformerlm-librispeech"
audio_file = "example.wav"

# Déclare les modèles
asr_model = EncoderDecoderASR.from_hparams(source=source_model)
aligner = CTCSegmentation(asr_model, kaldi_style_text=False)

# Transcris et récupère le texte pour s'en servir comme base
text = asr_model.transcribe_file(audio_file).split(" ")
# Aligne le texte text avec l'audio pour générer un timestamp
segments = aligner(audio_file, text, name="example1")

# Affiche le résultat
print(segments)
```

FIGURE A.4 – Transcription avec horodatage d'un fichier audio en utilisant SpeechBrain.

```python
from paddlespeech.cli.asr.infer import ASRExecutor

# Déclarer le modèle et transcrire
asr = ASRExecutor()
result = asr(audio_file="zh.wav")

# Afficher le résultat
print(result)
```

FIGURE A.5 – Transcription sans horodatage d'un fichier audio en utilisant PaddleSpeech.

# ANNEXE B

# ENCODAGE DE PHRASES

```python
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Exemple de deux phrases
phrase1 = "J'aime les chats."
phrase2 = "J'aime les chiens."

# Création du vecteur Bag of Words
vectorizer = CountVectorizer()
X = vectorizer.fit_transform([phrase1, phrase2]).toarray()

# Normalisation des vecteurs
X_normalized = X / np.linalg.norm(X, axis=1)[:, np.newaxis]

# Calcul de la distance cosinus
cosine_sim = cosine_similarity(X_normalized)

print("Similarity between the two phrases:", cosine_sim[0, 1])
```

FIGURE B.1 – Encodage de phrases utilisant un SdM pour ainsi calculer la distance cosinus.

Le SdM généré dans B.1 est généré seulement à partir des deux phrases fournies, ce qui signifie un jeu de données très petit et donc des observations très peu pertinentes.

```python
from gensim.models import KeyedVectors
from sklearn.metrics.pairwise import cosine_similarity
from huggingface_hub import hf_hub_download
import numpy as np

# Load pre-trained Word2Vec model for French
word2vec_model = KeyedVectors.load_word2vec_format(
    hf_hub_download(repo_id="Word2vec/nlpl_43", filename="model.bin"),
    binary=True,
    unicode_errors="ignore"
)

# Tokenization
sentence1 = "J'aime les chats".lower().split()
sentence2 = "J'aime les chaats".lower().split()

# Sentence Embedding
sentence1_embeddings = [word2vec_model[word] for word in sentence1 if word in word2vec_model]
sentence2_embeddings = [word2vec_model[word] for word in sentence2 if word in word2vec_model]

sentence1_embedding = np.mean(sentence1_embeddings, axis=0)
sentence2_embedding = np.mean(sentence2_embeddings, axis=0)

print(sentence1_embedding)
print(sentence2_embedding)

# Cosine Similarity Calculation
similarity = cosine_similarity([sentence1_embedding], [sentence2_embedding])[0][0]

print("Cosine Similarity:", similarity)
```

FIGURE B.2 – Encodage de phrases utilisant Word2Vec pour ainsi calculer la distance cosinus.

Comme nous pouvons le voir dans B.2, l'encodage se fait à partir du mot. On crée ensuite une matrice à partir de ces encodages. Dans B.3 par contre, l'encodage est au niveau de la phrase.

```python
from sentence_transformers import SentenceTransformer
from sklearn.metrics.pairwise import cosine_similarity

# Load a pre-trained SBERT model
model = SentenceTransformer('paraphrase-multilingual-mpnet-base-v2')

# Define the sentences
sentences = ["J'aime les chats", "J'aime les chaats"]

# Generate sentence embeddings
sentence_embeddings = model.encode(sentences)

# Calculate cosine similarity
similarity = cosine_similarity([sentence_embeddings[0]], [sentence_embeddings[1]])[0][0]

print("Cosine Similarity between the sentences:", similarity)
```

FIGURE B.3 – Encodage de phrases utilisant SentenceBERT pour ainsi calculer la distance cosinus.

# ANNEXE C

## RÉSULTATS DE TRANSCRIPTION AVEC DIFFÉRENTS MODÈLES


```
{
    "output_segments_15s/A1_12-09-17_0_15.wav": {
        "hypothese_nvidia_conftrans_trancribe": "la commission d enquête sur les
        ↪  relations entre les autochtones et certains services publics du québec
        ↪  présidés par l honorable jacques vien est maintenant ouverte vuillez
        ↪  vous asseoir alors bonjour bonjour à tous",
        "hypothese_nvidia_fastconformer_trancribe": "la commission d enquête sur
        ↪  les relations entre les autochtones et certains services publics du
        ↪  québec présidés par l honorable jacques vient est maintenant ouverte
        ↪  payez vous asseoir alors bonjour bonjour athos ",
        "hypothese_sb_wav2vec2_trancribe": "la commission d enute sur ls rtions
        ↪  ntre les autoctones et certains srices puics u uc présié ar l onorae
        ↪  acues vien est vaintenant ouverte a paévisatoiresleurs bonours onjour à
        ↪  tous",
        "hypothese_sb_whisper_trancribe": "la commission d enquête sur les
        ↪  relations entre les autochtones et certains services publics du québec
        ↪  présidée par l honorable jacques vien est maintenant ouverte payez vous
        ↪  assoir alors bonjour bonjour à tous",
        "hypothese_whisper_transcribe": " la commission d enquête sur les relations
        ↪  entre les autochtones et certains services publics du québec présidée
        ↪  par l honorable jacques vient est maintenant ouverte veuillez vous
        ↪  asseoir alors bonjour bonjour à tous bonjour bonjour bonjour bonjour
        ↪  bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour bonjour
        ↪  bonjour bonjour bonjour bonjour bonjour bonjour ",
        "reference": "la commission d enquête sur les relations entre les
        ↪  autochtones et certains services publics du québec présidée par l
        ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
        ↪  alors bonjour bonjour à tous ",
        "similarite_cosine_nvidia_conftrans_trancribe": 0.9976472854614258,
        "similarite_cosine_nvidia_fastconformer_trancribe": 0.980069637298584,
        "similarite_cosine_sb_wav2vec2_trancribe": 0.7025892734527588,
        "similarite_cosine_sb_whisper_trancribe": 0.9875296950340271,
        "similarite_cosine_whisper_transcribe": 0.8654600381851196,
        "tec_nvidia_conftrans_trancribe": 0.018518518656492233,
        "tec_nvidia_fastconformer_trancribe": 0.0555555559694767,
        "tec_sb_wav2vec2_trancribe": 0.23148147761821747,
        "tec_sb_whisper_trancribe": 0.0416666679084301,
        "tec_whisper_transcribe": 0.7129629850387573,
        "tem_nvidia_conftrans_trancribe": 0.09090909361839294,
        "tem_nvidia_fastconformer_trancribe": 0.1515151560306549,
        "tem_sb_wav2vec2_trancribe": 0.6363636255264282,
        "tem_sb_whisper_trancribe": 0.09090909361839294,
        "tem_whisper_transcribe": 0.6060606241226196,
        "temps_calcul_nvidia_conftrans_trancribe": 5.108181953430176,
        "temps_calcul_nvidia_fastconformer_trancribe": 6.041552305221558,
        "temps_calcul_sb_wav2vec2_trancribe": 6.907730579376221,
        "temps_calcul_sb_whisper_trancribe": 22.426573038101196,
        "temps_calcul_whisper_transcribe": 30.094895601272583
    },
    "output_segments_15s/A1_12-09-17_1113_1128.wav": {
```

```
    "hypothese_nvidia_conftrans_trancribe": "du québec les autres donc les dix
    ↪  neuf autres caps policiers desservent une communauté seulement la
    ↪  sûreté du québec nous on dessert en dessertes régulières de gendarmerie
    ↪  parce que ",
    "hypothese_nvidia_fastconformer_trancribe": "du québec les autres donc les
    ↪  dix neuf autres cas policiers desservent une communauté seulement la
    ↪  sûreté du québec nous on dessert en desserte régulière de gendarmerie
    ↪  parce qu on",
    "hypothese_sb_wav2vec2_trancribe": "ls sud du uec les autrs ont les i nuf
    ↪  autrs corps poiciers esservent une comunut seunt la sortie du uec nous
    ↪  on dessert un sserte réuires e narrie pqu",
    "hypothese_sb_whisper_trancribe": "les autres donc les dix neuf autres cas
    ↪  policiers desservent une communauté seulement la sûreté du québec nous
    ↪  ont dessert en desserte régulière de gendarmerie parce qu on",
    "hypothese_whisper_transcribe": " sud du québec les autres donc les 19
    ↪  autres corps policiers desservent une communauté seulement la sûreté du
    ↪  québec nous on dessert en dessert régulière de gendarmerie parce que ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.9579039812088013,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.9376769065856934,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.45152589678764343,
    "similarite_cosine_sb_whisper_trancribe": 0.9051503539085388,
    "similarite_cosine_whisper_transcribe": 0.9978482723236084,
    "tec_nvidia_conftrans_trancribe": 0.09944751113653183,
    "tec_nvidia_fastconformer_trancribe": 0.0883977934718132,
    "tec_sb_wav2vec2_trancribe": 0.2928176820278168,
    "tec_sb_whisper_trancribe": 0.14917127788066864,
    "tec_whisper_transcribe": 0.0220994483679533,
    "tem_nvidia_conftrans_trancribe": 0.23333333432674408,
    "tem_nvidia_fastconformer_trancribe": 0.1666666716337204,
    "tem_sb_wav2vec2_trancribe": 0.699999988079071,
    "tem_sb_whisper_trancribe": 0.2666666805744171,
    "tem_whisper_transcribe": 0.06666667014360428,
    "temps_calcul_nvidia_conftrans_trancribe": 4.796341180801392,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.369383096694946,
    "temps_calcul_sb_wav2vec2_trancribe": 4.936501502990723,
    "temps_calcul_sb_whisper_trancribe": 21.897401809692383,
    "temps_calcul_whisper_transcribe": 21.304750680923462
},
"output_segments_15s/A1_12-09-17_3850_3865.wav": {
    "hypothese_nvidia_conftrans_trancribe": "chacune des nations et on fait
    ↪  réiser aux intervenants que onze nations onze cultures différentes don
    ↪  pierre vient pt peu établir des différences culturelles et c est",
    "hypothese_nvidia_fastconformer_trancribe": "de chacune des nations et ont
    ↪  fait réaliser aux intervenants que onze nations onze cultures
    ↪  différentes donc pierre vient peu établir des différences culturelles
    ↪  et",
    "hypothese_sb_wav2vec2_trancribe": "ltrelle de caun des natons et on fait
    ↪  raiser aux intrvenant ue one nations ne cultures ifrnts donc pierre vis
    ↪  peut éir ds s irncs uurelles et sestlaqui",
    "hypothese_sb_whisper_trancribe": "leurs cultures de chacune des nations
    ↪  ont fait réaliser aux intervenants qu onze nations onze cultures
    ↪  différentes",
    "hypothese_whisper_transcribe": " de chacune des nations et ont fait
    ↪  réaliser aux intervenants que 11 nations 11 cultures différentes donc
    ↪  pierre vient un petit peu établir des différences culturelles et c est
    ↪  là qu il ",
```

```
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.9576514363288879,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.9461686015129089,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.537629246711731,
    "similarite_cosine_sb_whisper_trancribe": 0.9129605293273926,
    "similarite_cosine_whisper_transcribe": 0.98284512758255,
    "tec_nvidia_conftrans_trancribe": 0.11702127754688263,
    "tec_nvidia_fastconformer_trancribe": 0.13829787075519562,
    "tec_sb_wav2vec2_trancribe": 0.3191489279270172,
    "tec_sb_whisper_trancribe": 0.563829779624939,
    "tec_whisper_transcribe": 0.05319149047136307,
    "tem_nvidia_conftrans_trancribe": 0.25,
    "tem_nvidia_fastconformer_trancribe": 0.25,
    "tem_sb_wav2vec2_trancribe": 0.6875,
    "tem_sb_whisper_trancribe": 0.65625,
    "tem_whisper_transcribe": 0.09375,
    "temps_calcul_nvidia_conftrans_trancribe": 4.680254220962524,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.67010498046875,
    "temps_calcul_sb_wav2vec2_trancribe": 4.227119445800781,
    "temps_calcul_sb_whisper_trancribe": 20.64364981651306,
    "temps_calcul_whisper_transcribe": 20.861705541610718
},
"output_segments_15s/A1_16-01-18_1285_1300.wav": {
    "hypothese_nvidia_conftrans_trancribe": "qui que ce soit même pour les
    ↪  avocats même pour les procureur à rien ainsi comme gens v me laissz à l
    ↪  abandon pour a dire on voulà dire en lui ces sont importance on vous
    ↪  conçenez un autre chose puis jusqu attend que",
    "hypothese_nvidia_fastconformer_trancribe": "qui que ce soit même pôle les
    ↪  avocats même on les procura rien oh c était comme gens on me laissait à
    ↪  l abandon pour dire on v la dire en lui c est sans dépendance on vous
    ↪  connaissait un autre chose puis jusqu à teint que",
    "hypothese_sb_wav2vec2_trancribe": "ui u ce soit même pour les oats mêe
    ↪  pour les prcureurrin aussi come jean me laissar à l aandon pourésir
    ↪  avladi en lui caissant l importanc on voucrsi un un autre s puis usu
    ↪  attaque",
    "hypothese_sb_whisper_trancribe": "qui que ce soit même pour les avocats
    ↪  même pour les procureurs rien ainsi comme jean m a laissé à l abandon
    ↪  pour dire en voulant dire ah lui c est sans importance on va continuer
    ↪  un autre chose puis jusqu à temps que",
    "hypothese_whisper_transcribe": " qui que ce soit même pas les avocats même
    ↪  pas les procuraires à rien c est comme ça qu on m a laissé à l abandon
    ↪  pour dire en voulant dire lui c est ça l importance on va continuer une
    ↪  autre chose puis jusqu à ce que ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8535985350608826,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.7944662570953369,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.4805169403553009,
    "similarite_cosine_sb_whisper_trancribe": 0.876924991607666,
    "similarite_cosine_whisper_transcribe": 0.9198471307754517,
    "tec_nvidia_conftrans_trancribe": 0.27272728085517883,
    "tec_nvidia_fastconformer_trancribe": 0.3073593080043793,
    "tec_sb_wav2vec2_trancribe": 0.4025973975658417,
    "tec_sb_whisper_trancribe": 0.1991342008113861,
```

```
    "tec_whisper_transcribe": 0.17748917639255524,
    "tem_nvidia_conftrans_trancribe": 0.5918367505073547,
    "tem_nvidia_fastconformer_trancribe": 0.5102040767669678,
    "tem_sb_wav2vec2_trancribe": 0.7755101919174194,
    "tem_sb_whisper_trancribe": 0.3265306055545807,
    "tem_whisper_transcribe": 0.2857142984867096,
    "temps_calcul_nvidia_conftrans_trancribe": 4.663280010223389,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.6798996925354,
    "temps_calcul_sb_wav2vec2_trancribe": 4.672489881515503,
    "temps_calcul_sb_whisper_trancribe": 22.3236665725708,
    "temps_calcul_whisper_transcribe": 24.19691801071167
},
"output_segments_15s/A1_16-01-18_1408_1423.wav": {
    "hypothese_nvidia_conftrans_trancribe": "c est pas ma sou qui m ét aarvée
    ↪  merci donc ll oœil a eu l arrêt des procédures en deux mille vous l
    ↪  avez appris dans les journaux après ce que vous venez de nous dire oui
    ↪  pleu environ dix ans plus tard à peu près neuf ans plus tard en fait en
    ↪  deux mesges",
    "hypothese_nvidia_fastconformer_trancribe": "c est pas ma ça qui m est
    ↪  arrivé merci ton cle y a eu l arrêt des procédures en deux mille sept
    ↪  vous l aviez appris dans les journaux après ce que vous venez de nous
    ↪  dire pleu environ dix ans plus tard à peu près neuf ans plus tard fait
    ↪  en",
    "hypothese_sb_wav2vec2_trancribe": "il y a cepenant ceux ui me anrvit merci
    ↪  tant qu ilil a eu l arrt des procéures en ux mille spt vous l avez
    ↪  appris ans les urnau aprs ce ue vous venez de nous dire pleut niron ix
    ↪  ans pus tard à peu prs nuf ans us tar en fait un deux melsaire",
    "hypothese_sb_whisper_trancribe": "mais c est pas marsal qui m est arroué
    ↪  merci donc là il y a eu l arrêt des procédures en deux mille sept vous
    ↪  l avez appris dans les journaux après ce que vous venez nous dire plus
    ↪  à environ dix ans plus tard à peu près neuf ans plus tard en fait en
    ↪  deux mille seize",
    "hypothese_whisper_transcribe": " et c est pas ma seule qui m est arrivée
    ↪  merci donc là il y a eu l arrêt des procédures en 2007 vous l avez
    ↪  appris dans les journaux après ce que vous venez de nous dire oui puis
    ↪  là environ 10 ans plus tard à peu près 9 ans plus tard en fait en 2016
    ↪  ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8037802577018738,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.8176032304763794,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.6287626028060913,
    "similarite_cosine_sb_whisper_trancribe": 0.8379687070846558,
    "similarite_cosine_whisper_transcribe": 0.9790683388710022,
    "tec_nvidia_conftrans_trancribe": 0.15600000321865082,
    "tec_nvidia_fastconformer_trancribe": 0.20000000298023224,
    "tec_sb_wav2vec2_trancribe": 0.3199999928474426,
    "tec_sb_whisper_trancribe": 0.20000000298023224,
    "tec_whisper_transcribe": 0.08399999886751175,
    "tem_nvidia_conftrans_trancribe": 0.22807016968727112,
    "tem_nvidia_fastconformer_trancribe": 0.2631579041481018,
    "tem_sb_wav2vec2_trancribe": 0.5964912176132202,
    "tem_sb_whisper_trancribe": 0.24561403691768646,
    "tem_whisper_transcribe": 0.10526315867900848,
    "temps_calcul_nvidia_conftrans_trancribe": 4.390341520309448,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.7602455615997314,
    "temps_calcul_sb_wav2vec2_trancribe": 4.817273378372192,
```

```
    "temps_calcul_sb_whisper_trancribe": 25.61246132850647,
    "temps_calcul_whisper_transcribe": 25.3188054561615
},
"output_segments_15s/A1_16-01-18_2008_2023.wav": {
    "hypothese_nvidia_conftrans_trancribe": "demander à ce qu il c était
    ↪  possible d avoir des explications avec un procureur quelconque pouvoir
    ↪  rien à ce qui s était passé n où par entre temps on n avait été
    ↪  chercher les coupures de journaux",
    "hypothese_nvidia_fastconformer_trancribe": "demander à ce qui s était
    ↪  possible d avoir des explications avec un procureur quelconque pouvoir
    ↪  rien à ce qui s était passé nous par entretemps on avait été chercher
    ↪  les coupures de journaux",
    "hypothese_sb_wav2vec2_trancribe": "nos andez est ce ue c était pssie d air
    ↪  ds pitons ac un proureur uelconue puvoir riennement ce ue c était
    ↪  pensér nous par entretemps o n avait ét cercer les coupurs de ournaux
    ↪  pouvoirs",
    "hypothese_sb_whisper_trancribe": "on a demandé est ce qu il était possible
    ↪  d avoir des explications avec un procureur quelconque pouvoir
    ↪  réellement ce qu il s était passé nous par entre temps on avait été
    ↪  chercher les coupes de journaux pour voir",
    "hypothese_whisper_transcribe": " on a demandé si c était possible d avoir
    ↪  des explications avec un procureur quelconque pour voir réellement ce
    ↪  qui s était passé nous par entre temps on avait été chercher les
    ↪  coupures de journaux pour voir ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.9082249402999878,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.9429950714111328,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.5722060203552246,
    "similarite_cosine_sb_whisper_trancribe": 0.9902830123901367,
    "similarite_cosine_whisper_transcribe": 0.9964490532875061,
    "tec_nvidia_conftrans_trancribe": 0.1866028755903244,
    "tec_nvidia_fastconformer_trancribe": 0.16267941892147064,
    "tec_sb_wav2vec2_trancribe": 0.2248803824186325,
    "tec_sb_whisper_trancribe": 0.04784689098596573,
    "tec_whisper_transcribe": 0.07655502110719681,
    "tem_nvidia_conftrans_trancribe": 0.3684210479259491,
    "tem_nvidia_fastconformer_trancribe": 0.34210526943206787,
    "tem_sb_wav2vec2_trancribe": 0.6578947305679321,
    "tem_sb_whisper_trancribe": 0.15789473056793213,
    "tem_whisper_transcribe": 0.10526315867900848,
    "temps_calcul_nvidia_conftrans_trancribe": 4.757572174072266,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.022562265396118,
    "temps_calcul_sb_wav2vec2_trancribe": 5.504959583282471,
    "temps_calcul_sb_whisper_trancribe": 22.908292531967163,
    "temps_calcul_whisper_transcribe": 22.737285614013672
},
"output_segments_15s/A1_23-10-18_1750_1765.wav": {
    "hypothese_nvidia_conftrans_trancribe": "l a dépensé que l hôte y veut
    ↪  prendre t des brocs às qui veut juste étrangler je pense à un des
    ↪  gots",
    "hypothese_nvidia_fastconformer_trancribe": "l a dépensé que l autre y veut
    ↪  prendre tes brocas qui veut juste étrangler je pense à un digot ",
    "hypothese_sb_wav2vec2_trancribe": "ililllors de penser ue lodit veut prada
    ↪  tes broca ui veut uste être anglill jetans un rdgotnppaa",
    "hypothese_sb_whisper_trancribe": "d impasser que l autre veut prendre tes
    ↪  bras quand qu il veut juste l étrangler je pense un rédicot",
```

```
    "hypothese_whisper_transcribe": " de penser que l autre veut te prendre
    ↪  dans tes bras quand il veut juste t étrangler je pense un vrai délicat
    ↪  ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.5995006561279297,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.7211630344390869,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.5118095874786377,
    "similarite_cosine_sb_whisper_trancribe": 0.8891557455062866,
    "similarite_cosine_whisper_transcribe": 0.9318276643753052,
    "tec_nvidia_conftrans_trancribe": 0.3643410801887512,
    "tec_nvidia_fastconformer_trancribe": 0.3488371968269348,
    "tec_sb_wav2vec2_trancribe": 0.5038759708404541,
    "tec_sb_whisper_trancribe": 0.3720930218696594,
    "tec_whisper_transcribe": 0.24031007289886475,
    "tem_nvidia_conftrans_trancribe": 0.5714285969734192,
    "tem_nvidia_fastconformer_trancribe": 0.5357142686843872,
    "tem_sb_wav2vec2_trancribe": 0.7857142686843872,
    "tem_sb_whisper_trancribe": 0.5,
    "tem_whisper_transcribe": 0.2857142984867096,
    "temps_calcul_nvidia_conftrans_trancribe": 4.878227472305298,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.016992092132568,
    "temps_calcul_sb_wav2vec2_trancribe": 4.914344549179077,
    "temps_calcul_sb_whisper_trancribe": 18.747769594192505,
    "temps_calcul_whisper_transcribe": 19.301992654800415
},
"output_segments_15s/A1_23-10-18_2207_2222.wav": {
    "hypothese_nvidia_conftrans_trancribe": "que ça on plus l esprit que c est
    ↪  plus vert surtout les jeunes p les vieux parce que les vieux plus de
    ↪  soiante d ans y ont comme connubin les pensionnaaures y sont comme",
    "hypothese_nvidia_fastconformer_trancribe": "que ça offre plus l esprit que
    ↪  vert surtout les jeunes pilés vieux parce que les vieux plus de
    ↪  soixante dix ans y ont comme les pensionnops y sont",
    "hypothese_sb_wav2vec2_trancribe": "taarouv plus l esprit ue c est
    ↪  plusouverts surtout les junes pilers vieux parc ue les vieux pus de
    ↪  socantétiseurs ont comme conubens les pensionopes y sont comme
    ↪  fruistes",
    "hypothese_sb_whisper_trancribe": "que ça ouvre plus l esprit que c est
    ↪  plus ouvert surtout les jeunes et les vieux parce que les vieux plus de
    ↪  soixante dix ans y ont comme connu les pensions nobles et sont comme
    ↪  christians",
    "hypothese_whisper_transcribe": " que ça ouvre plus l esprit que c est plus
    ↪  ouvert surtout les jeunes puis les vieux parce que les vieux plus de 70
    ↪  ans ils ont comme connu les pensionnats puis ils sont comme ",
    "reference": "la commission d enquête sur les relations entre les
    ↪  autochtones et certains services publics du québec présidée par l
    ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
    ↪  alors bonjour bonjour à tous ",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8050655722618103,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.8523585796356201,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.5492920875549316,
    "similarite_cosine_sb_whisper_trancribe": 0.8654332160949707,
    "similarite_cosine_whisper_transcribe": 0.9734499454498291,
    "tec_nvidia_conftrans_trancribe": 0.23783783614635468,
    "tec_nvidia_fastconformer_trancribe": 0.3513513505458832,
    "tec_sb_wav2vec2_trancribe": 0.2594594657421112,
    "tec_sb_whisper_trancribe": 0.20000000298023224,
```

```
        "tec_whisper_transcribe": 0.0648648664355278,
        "tem_nvidia_conftrans_trancribe": 0.3243243098258972,
        "tem_nvidia_fastconformer_trancribe": 0.45945945382118225,
        "tem_sb_wav2vec2_trancribe": 0.5675675868988037,
        "tem_sb_whisper_trancribe": 0.2432432472705841,
        "tem_whisper_transcribe": 0.054054055362939835,
        "temps_calcul_nvidia_conftrans_trancribe": 4.953022241592407,
        "temps_calcul_nvidia_fastconformer_trancribe": 4.44670295715332,
        "temps_calcul_sb_wav2vec2_trancribe": 5.323650360107422,
        "temps_calcul_sb_whisper_trancribe": 21.691405296325684,
        "temps_calcul_whisper_transcribe": 22.42435097694397
    },
    "output_segments_15s/A1_23-10-18_829_844.wav": {
        "hypothese_nvidia_conftrans_trancribe": "après l alle à l interrogatoire au
        ↪  pachapin jgais gazmard voulut les inviter chez nous à vi manger une
        ↪  bouchée mais ils étaient trop froids ils se plaignaient tout le temps
        ↪  qu ils étaient comme taannés et le comme",
        "hypothese_nvidia_fastconformer_trancribe": "l interrogatoire ou le chapeau
        ↪  qu izmavait voulu les inviter chez moi à venir manger une bouchée mais
        ↪  ils étaient trop froids ils se plaignaient tout le temps qu ils étaient
        ↪  comme tannés comme",
        "hypothese_sb_wav2vec2_trancribe": "inut aprs l iterrogation bachapo g
        ↪  gigismar eut vouu les initer cez nous à venir maner une bouchée mais is
        ↪  taint trop froids ss plaignait tout le tamp u is étaient comme talés le
        ↪  comm pose",
        "hypothese_sb_whisper_trancribe": "après l interrogato bein je sais pas j
        ↪  ai quasiment voulu les inviter chez nous à venir manger une boucher
        ↪  mais ils étaient trop froids ils plaignaient tout le temps qu ils
        ↪  étaient comme tannés là comme pfff ah",
        "hypothese_whisper_transcribe": " après la lettre à gâteau je sais pas j ai
        ↪  quasiment voulu les inviter chez nous à venir manger une bouchée mais
        ↪  ils étaient trop froid ils se plaignaient tout le temps qu ils étaient
        ↪  comme tannés là comme ",
        "reference": "la commission d enquête sur les relations entre les
        ↪  autochtones et certains services publics du québec présidée par l
        ↪  honorable jacques viens est maintenant ouverte veuillez vous asseoir
        ↪  alors bonjour bonjour à tous ",
        "similarite_cosine_nvidia_conftrans_trancribe": 0.910554051399231,
        "similarite_cosine_nvidia_fastconformer_trancribe": 0.9519875049591064,
        "similarite_cosine_sb_wav2vec2_trancribe": 0.647348165512085,
        "similarite_cosine_sb_whisper_trancribe": 0.9620628356933594,
        "similarite_cosine_whisper_transcribe": 0.914682924747467,
        "tec_nvidia_conftrans_trancribe": 0.21153846383094788,
        "tec_nvidia_fastconformer_trancribe": 0.1538461595773697,
        "tec_sb_wav2vec2_trancribe": 0.29326921701431274,
        "tec_sb_whisper_trancribe": 0.11538461595773697,
        "tec_whisper_transcribe": 0.11538461595773697,
        "tem_nvidia_conftrans_trancribe": 0.3684210479259491,
        "tem_nvidia_fastconformer_trancribe": 0.2368421107530594,
        "tem_sb_wav2vec2_trancribe": 0.7105262875556946,
        "tem_sb_whisper_trancribe": 0.21052631735801697,
        "tem_whisper_transcribe": 0.15789473056793213,
        "temps_calcul_nvidia_conftrans_trancribe": 5.030459403991699,
        "temps_calcul_nvidia_fastconformer_trancribe": 4.297870635986328,
        "temps_calcul_sb_wav2vec2_trancribe": 5.6741180419921875,
        "temps_calcul_sb_whisper_trancribe": 23.030489444732666,
        "temps_calcul_whisper_transcribe": 22.77544617652893
    },
    "similarite_cosine_moyen_nvidia_conftrans_trancribe": 0.8659918838077121,
```

```
    "similarite_cosine_moyen_nvidia_fastconformer_trancribe": 0.8827210134930081,
    "similarite_cosine_moyen_sb_wav2vec2_trancribe": 0.5646311442057291,
    "similarite_cosine_moyen_sb_whisper_trancribe": 0.9141632517178854,
    "similarite_cosine_moyen_whisper_transcribe": 0.951275454627143,
    "tec_moyen_nvidia_conftrans_trancribe": 0.17807456851005554,
    "tec_moyen_nvidia_fastconformer_trancribe": 0.19532553851604462,
    "tec_moyen_sb_wav2vec2_trancribe": 0.3099610507488251,
    "tec_moyen_sb_whisper_trancribe": 0.19866444170475006,
    "tec_moyen_whisper_transcribe": 0.17417918145656586,
    "tem_moyen_nvidia_conftrans_trancribe": 0.3391812741756439,
    "tem_moyen_nvidia_fastconformer_trancribe": 0.3274853825569153,
    "tem_moyen_sb_wav2vec2_trancribe": 0.6754385828971863,
    "tem_moyen_sb_whisper_trancribe": 0.28947368264198303,
    "tem_moyen_whisper_transcribe": 0.19005848467350006,
    "temps_moyen_nvidia_conftrans_trancribe": 4.913363880581326,
    "temps_moyen_nvidia_fastconformer_trancribe": 4.699437618255615,
    "temps_moyen_sb_wav2vec2_trancribe": 5.335069523917304,
    "temps_moyen_sb_whisper_trancribe": 22.25408559375339,
    "temps_moyen_whisper_transcribe": 23.383294132020737,
    "temps_total_nvidia_conftrans_trancribe": 44.22027349472046,
    "temps_total_nvidia_fastconformer_trancribe": 42.294936656951904,
    "temps_total_sb_wav2vec2_trancribe": 48.015623807907104,
    "temps_total_sb_whisper_trancribe": 200.28676843643188,
    "temps_total_whisper_transcribe": 210.44964480400085
}
```


Listing C.1 – Résultats des modèles sur les segments de 15 secondes

```
{
    "output_segments_30s/A1_12-09-17_2615_2645.wav": {
        "hypothese_nvidia_conftrans_trancribe": "une seule nation il faisait comme
        ↪  celui de manawan s occupait également de annesce attaqué donc il
        ↪  devenait pas bon avec juste une nation il y était tout le temps à
        ↪  cheval entre les deux dans le même souffle la communauté qui faisait
        ↪  une demande à la sortée du québec pouvait recevoir une réponse x de
        ↪  montréal une réponse y grecque de troisrivières puis une réponse z du
        ↪  signy lac saint jean et cette nation là un grand conseil de la nation
        ↪  donc un momentanée se parlait pis eben voyant j ai trois",
        "hypothese_nvidia_fastconformer_trancribe": "une seule nation et faisait
        ↪  comme celui de manawan s occupait également de connaisses attaquées
        ↪  donc il devenait pas bon avec juste une nation et y était tout le temps
        ↪  à cheval entre les deux dans le même souffle la communauté qui faisait
        ↪  une demande à la sortie du québec pouvait recevoir une réponse x de
        ↪  montréal une réponse y de tourivières puis une réponse z du saguenay
        ↪  lac saint jean et cette nation là a un grand conseil de la nation donc
        ↪  un moment d annexe parlait puis disait benvoyant j ai trois",
        "hypothese_sb_wav2vec2_trancribe": "pour un sule nation il isait co cui de
        ↪  manauan s ocupait éaemnt de uanesattaqué onc il enait propbon ac justen
        ↪  nation tait toule tain a chval entre les euxdans le mme soufle la
        ↪  cmmunut ui sait une deane à la sortie du uec puvait roir une rponse x
        ↪  de montréal une rponse sgrecue de toit riviar pi une rponse du sagné l
        ↪  ac saint ean et cette ntion l un rand onseil de la nation onc n moment
        ↪  donné parlait pi disait bien voyage trois répns",
        "hypothese_sb_whisper_trancribe": "il n avait pas une seule nation il
        ↪  faisait comme celui de manawan s occupait également de kenesatake donc
        ↪  il devenait pas bon avec juste une nation il était tout le temps à
        ↪  cheval entre les deux dans le même soufle la communauté qui faisait une
        ↪  demande à la sûreté du québec pouvait recevoir une réponse x de
        ↪  montréal",
        "hypothese_whisper_transcribe": " n avait pas une seule nation il faisait
        ↪  comme celui de manawan s occupait également de knesset haké donc il ne
        ↪  devenait pas bon avec juste une nation il était tout le temps à cheval
        ↪  entre les deux dans le même souffle la communauté qui faisait une
        ↪  demande à la sûreté du québec pouvait recevoir une réponse x de
        ↪  montréal une réponse y de trois rivières puis une réponse z du saguenay
        ↪  lac saint jean et cette nation là un grand conseil de la nation qui à
        ↪  un moment donné se parlait et disait bien voyons j ai trois réponses ",
        "reference": "n avaient pas une seule nation ils faisaient comme celui de
        ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
        ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
        ↪  deux dans le même souffle la communauté qui faisait une demande à la
        ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
        ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
        ↪  cette nation là a un grand conseil de la nation donc un moment donné
        ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
        "similarite_cosine_nvidia_conftrans_trancribe": 0.9184536933898926,
        "similarite_cosine_nvidia_fastconformer_trancribe": 0.9238539934158325,
        "similarite_cosine_sb_wav2vec2_trancribe": 0.6275895833969116,
        "similarite_cosine_sb_whisper_trancribe": 0.8983628153800964,
        "similarite_cosine_whisper_transcribe": 0.9112611413002014,
        "tec_nvidia_conftrans_trancribe": 0.1620626151561737,
        "tec_nvidia_fastconformer_trancribe": 0.14917127788066864,
        "tec_sb_wav2vec2_trancribe": 0.285451203584671,
        "tec_sb_whisper_trancribe": 0.42909759283065796,
        "tec_whisper_transcribe": 0.0883977934718132,
        "tem_nvidia_conftrans_trancribe": 0.30693069100379944,
        "tem_nvidia_fastconformer_trancribe": 0.2673267424106598,
```

```
    "tem_sb_wav2vec2_trancribe": 0.7029703259468079,
    "tem_sb_whisper_trancribe": 0.5346534848213196,
    "tem_whisper_transcribe": 0.1881188154220581,
    "temps_calcul_nvidia_conftrans_trancribe": 5.424461841583252,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.679977655410767,
    "temps_calcul_sb_wav2vec2_trancribe": 9.490350723266602,
    "temps_calcul_sb_whisper_trancribe": 28.097974061965942,
    "temps_calcul_whisper_transcribe": 37.93682885169983
},
"output_segments_30s/A1_12-09-17_4023_4053.wav": {
    "hypothese_nvidia_conftrans_trancribe": "sensibles de les des pensionnantts
    ↪  et que ces gens là l ont peut être vécu directement ou indirectement on
    ↪  trouvait que c était peut être on n était peut être pas encore prêt à
    ↪  les intégrer à la formation mais assurément que les policiers
    ↪  aulochtones qui eux ont à se préparer à travaillant méi autochtone
    ↪  cette formation là leur convient bien là donc oui on commence à avoir
    ↪  de la rétroaction l de de gens de la communauté donc c est la première
    ↪  journée de",
    "hypothese_nvidia_fastconformer_trancribe": "sensible des pensionnats et
    ↪  que ces gens là l ont peut être vécu directement ou indirectement on
    ↪  trouvait que c était peut être on était peut être encore prêt à les
    ↪  intégrer à la formation mais assurément que les policiers autochtones
    ↪  qui eux ont à se préparer à travailler en milieu autochtone cette
    ↪  formation là leur convient bien là donc oui on commence à avoir de la
    ↪  rétroaction lors de de gens de la communauté donc c est la première
    ↪  journée",
    "hypothese_sb_wav2vec2_trancribe": "ts sensibles de les dépensionnemant et
    ↪  ue ces ens l l ont peut êtr vécu irnt ou inirtnt on truvait u c éaint
    ↪  peut tre on étint peut tre pas enre prêts les intrr à la formaton mais
    ↪  assurément ue les poiciers aloctone ui eux ont à se prré travailler un
    ↪  milli autochtone cette oration lui convient bien onc oui on commnce
    ↪  avoir la rtroaction sd agent de la cmmunauté ssc est la rire ourne de
    ↪  forme",
    "hypothese_sb_whisper_trancribe": "on était peut être pas encore prêt à les
    ↪  intégrer à la formation mais assurément que les policiers alochtones
    ↪  qui eux ont à se préparer à travailler en milieu autochtone cette
    ↪  formation là lui convient bien",
    "hypothese_whisper_transcribe": " sensibles des pensionnats et que ces gens
    ↪  là l ont peut être vécu directement ou indirectement on trouvait que c
    ↪  était peut être on n était peut être pas encore prêt à les intégrer à
    ↪  la formation mais assurément que les policiers à l octonne qui eux ont
    ↪  à se préparer à travailler en milieu autochtone cette formation là lui
    ↪  convient bien donc oui on commence à avoir de la rétroaction de gens de
    ↪  la communauté donc c est la première journée de forme ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.984783411026001,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.9449913501739502,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.5600559115409851,
    "similarite_cosine_sb_whisper_trancribe": 0.8419494032859802,
    "similarite_cosine_whisper_transcribe": 0.9643429517745972,
    "tec_nvidia_conftrans_trancribe": 0.05111110955476761,
    "tec_nvidia_fastconformer_trancribe": 0.03777777776122093,
```

```
    "tec_sb_wav2vec2_trancribe": 0.23999999463558197,
    "tec_sb_whisper_trancribe": 0.5444444417953491,
    "tec_whisper_transcribe": 0.05111110955476761,
    "tem_nvidia_conftrans_trancribe": 0.1204819306731224,
    "tem_nvidia_fastconformer_trancribe": 0.07228915393352509,
    "tem_sb_wav2vec2_trancribe": 0.5903614163398743,
    "tem_sb_whisper_trancribe": 0.5903614163398743,
    "tem_whisper_transcribe": 0.09638553857803345,
    "temps_calcul_nvidia_conftrans_trancribe": 5.366302013397217,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.40101432800293,
    "temps_calcul_sb_wav2vec2_trancribe": 8.664944171905518,
    "temps_calcul_sb_whisper_trancribe": 23.13683772087097,
    "temps_calcul_whisper_transcribe": 33.7388801574707
},
"output_segments_30s/A1_12-09-17_5064_5094.wav": {
    "hypothese_nvidia_conftrans_trancribe": "comme je vous disais il avait déjà
    ↪  eu monsi le vicaire qui qui a pris sa retraite malheureusement donc oui
    ↪  ça peut être un facteur un facteur qui est facilitant mais comme je
    ↪  vous disais monsieur le commissaire le facteur important c est une
    ↪  bonne personne au bon endroit et cette personne ned doit avoir les
    ↪  capacités de développer ses connaissances et ses sensibilités au qu ils
    ↪  se retochton et est ce que dans ce cas",
    "hypothese_nvidia_fastconformer_trancribe": "comme je vous disais il y
    ↪  avait déjà eu monsieur vicaire qui a pris sa retraite malheureusement
    ↪  donc oui ça peut ça peut être un facteur un facteur qui est facilitant
    ↪  mais comme je vous disais monsieur le commissaire le facteur important
    ↪  c est une bonne personne au bon endroit et cette personne doit avoir
    ↪  les capacités de développer ses connaissances et ses sensibilités aux
    ↪  cultures autochtones",
    "hypothese_sb_wav2vec2_trancribe": "comme e vous isais il avait déj u
    ↪  monsieur vicaire uiui a pris sa rraite maeureusment donc oui sa pa peut
    ↪  êtr un facteur un fateur ui fciitant mais ce e vous isais monsiur le
    ↪  commissaire le facteur imorant c est un onne prsonne au bon nroit et
    ↪  cette prsonnes nedoit avoir les pacits de vpper ses onnaissancs et ses
    ↪  snsiiis au quilture autoctoneul s dans ce cadre",
    "hypothese_sb_whisper_trancribe": "comme je vous disais il avait déjà eu
    ↪  monsieur vicker qui a pris sa retraite malheureusement donc oui ça peut
    ↪  être un facteur un facteur qui est facilitant mais comme je vous disais
    ↪  monsieur le commissaire le facteur important c est une bonne personne
    ↪  au bon endroit et cette personne là doit avoir les capacité de
    ↪  développer ses connaissances et s",
    "hypothese_whisper_transcribe": " comme je vous disais il y avait déjà eu
    ↪  monsieur viqueur qui a pris sa retraite malheureusement donc oui ça
    ↪  peut être un facteur qui est facilitant mais comme je vous disais
    ↪  monsieur le commissaire le facteur important c est une bonne personne
    ↪  au bon endroit et cette personne là doit avoir les capacités de
    ↪  développer ses connaissances et ses sensibilités aux cultures",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.9113121628761292,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.997393786907196,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.65758216381073,
    "similarite_cosine_sb_whisper_trancribe": 0.8663078546524048,
```

```
    "similarite_cosine_whisper_transcribe": 0.9619208574295044,
    "tec_nvidia_conftrans_trancribe": 0.09176470339298248,
    "tec_nvidia_fastconformer_trancribe": 0.11764705926179886,
    "tec_sb_wav2vec2_trancribe": 0.20941177010536194,
    "tec_sb_whisper_trancribe": 0.20235294103622437,
    "tec_whisper_transcribe": 0.1411764770746231,
    "tem_nvidia_conftrans_trancribe": 0.20000000298023224,
    "tem_nvidia_fastconformer_trancribe": 0.1599999964237213,
    "tem_sb_wav2vec2_trancribe": 0.54666668176651,
    "tem_sb_whisper_trancribe": 0.25333333015441895,
    "tem_whisper_transcribe": 0.1733333319425583,
    "temps_calcul_nvidia_conftrans_trancribe": 5.40162205696106,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.327191352844238,
    "temps_calcul_sb_wav2vec2_trancribe": 7.293362379074097,
    "temps_calcul_sb_whisper_trancribe": 27.138168811798096,
    "temps_calcul_whisper_transcribe": 32.17212104797363
},
"output_segments_30s/A1_16-01-18_2083_2113.wav": {
    "hypothese_nvidia_conftrans_trancribe": "et puis la rencontre a été brève
    ↪  environ trente minutes marb st mathieu posait les questions j lui avais
    ↪  dit onlois de poser toutes les questions que tu veux pour mieux
    ↪  comprendre ta situation puis on lui répondait qu on n était pas là pour
    ↪  faire un procès on était conscient qu il avait eu des billes qu avait
    ↪  eu des choses mais que on lui expliquait que étant donné que eû",
    "hypothese_nvidia_fastconformer_trancribe": "et puis la rencontre a été
    ↪  brève lieu environ trente minutes mathieu posait les questions que je
    ↪  lui avais dit tout le droit de poser toutes les questions que tu veux
    ↪  pour mieux comprendre ta situation puis on lui répondait qu on était
    ↪  polo pour faire un procès on était conscient qu il avait eu des bris qu
    ↪  il avait vu des choses mais que on lui expliquait que étant donné qu il
    ↪  y a eu des pertes",
    "hypothese_sb_wav2vec2_trancribe": "et puis la rancontre a été brève l
    ↪  eniron trente minutes mabnatmatsuposait les usns j a avais dittuldite
    ↪  poser toutes s ustons de suf pour miux prenre la situaton puis on lui
    ↪  rponait u on était polo pour aire un procs on était consient u il avait
    ↪  eu des biris ui avait vu des coses mais ue on lui piuait ue étant onné
    ↪  u il a eut des pertes",
    "hypothese_sb_whisper_trancribe": "et puis la rencontre a été brève à
    ↪  environ trente minutes",
    "hypothese_whisper_transcribe": " et puis la rencontre a été brève là
    ↪  environ 30 minutes ça a été mathieu posait les questions j avais dit tu
    ↪  as le droit de poser toutes les questions tu veux pour mieux comprendre
    ↪  ta situation puis on lui répondait qu on n était pas là pour faire un
    ↪  procès on était conscients qu il y avait eu des bris qu il y avait eu
    ↪  des choses mais on lui expliquait que étant donné qu il y a eu des
    ↪  pertes ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.9306196570396423,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.927586555480957,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.6091636419296265,
    "similarite_cosine_sb_whisper_trancribe": 0.6686440706253052,
    "similarite_cosine_whisper_transcribe": 0.9970480799674988,
```

```
    "tec_nvidia_conftrans_trancribe": 0.15594059228897095,
    "tec_nvidia_fastconformer_trancribe": 0.11138613522052765,
    "tec_sb_wav2vec2_trancribe": 0.2524752616882324,
    "tec_sb_whisper_trancribe": 0.8589109182357788,
    "tec_whisper_transcribe": 0.049504950642585754,
    "tem_nvidia_conftrans_trancribe": 0.302325576543808,
    "tem_nvidia_fastconformer_trancribe": 0.22093023359775543,
    "tem_sb_wav2vec2_trancribe": 0.5930232405662537,
    "tem_sb_whisper_trancribe": 0.895348846912384,
    "tem_whisper_transcribe": 0.08139535039663315,
    "temps_calcul_nvidia_conftrans_trancribe": 5.628961801528931,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.155270338058472,
    "temps_calcul_sb_wav2vec2_trancribe": 7.546086072921753,
    "temps_calcul_sb_whisper_trancribe": 17.158891677856445,
    "temps_calcul_whisper_transcribe": 33.75832796096802
},
"output_segments_30s/A1_16-01-18_2174_2204.wav": {
    "hypothese_nvidia_conftrans_trancribe": "témoin des services de ses frères
    ↪  et sousy face était ben découragé de sou et cça a pris beaucoup de
    ↪  rencontres de quelques mots pour leur monter des soulc p si j ai
    ↪  répondu j une question peut être à mon papaty puis à vous après et mme
    ↪  chrétien vient de dire c est comme si ete pis mathieu vous l avez dit
    ↪  un peu plus tôt là c est comme si j ai",
    "hypothese_nvidia_fastconformer_trancribe": "était témoin des services de
    ↪  ses frères et sœurs aussi a que était bien découragé de ça a pris
    ↪  beaucoup de rencontres qu a de quelques mois pour leur monter de solu
    ↪  une question peut être à monsieur paplati puis à vous après et madame
    ↪  chrétien vient de dire c est comme si épiez mathieu vous l avez dit un
    ↪  peu plus tôt c est comme si",
    "hypothese_sb_wav2vec2_trancribe": "ii était témoin des ses fils de ses
    ↪  frres isont aussi faiue il était bien éuraé de étais pri baucup de
    ↪  rncontres u ell dot uus mots ur leur monter de so poussais j ai
    ↪  répondugastion peut êtr monsiur parpati piavoue ars etlmadame crétien
    ↪  viens de dire t istat comme si tpismathieu vous l ava dit un pu pustôt
    ↪  l c est o si j ",
    "hypothese_sb_whisper_trancribe": "il était témoin des services de ses
    ↪  frères saussés facques il était beaucoup découragé dessous il était s
    ↪  appris beaucoup de rencontres de quelques mois pour leur montée dessous
    ↪  facques je ne sais pas si j ai répondu",
    "hypothese_whisper_transcribe": " et il était témoin des services de ses
    ↪  frères et sœurs aussi il était bien découragé dessous ça a pris
    ↪  beaucoup de rencontres de quelques mois pour leur monter dessous je ne
    ↪  sais pas si j ai répondu je vais poser une question peut être à m
    ↪  papati puis à vous après mme chrétien vient de dire c est comme si
    ↪  mathieu vous l avez dit un peu plus tôt c est comme si j ai pas ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8380026817321777,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.875532865524292,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.671021044254303,
    "similarite_cosine_sb_whisper_trancribe": 0.7437103390693665,
    "similarite_cosine_whisper_transcribe": 0.9497861862182617,
    "tec_nvidia_conftrans_trancribe": 0.25123152136802673,
```

```
    "tec_nvidia_fastconformer_trancribe": 0.2857142984867096,
    "tec_sb_wav2vec2_trancribe": 0.38669949769973755,
    "tec_sb_whisper_trancribe": 0.6157635450363159,
    "tec_whisper_transcribe": 0.22413793206214905,
    "tem_nvidia_conftrans_trancribe": 0.4252873659133911,
    "tem_nvidia_fastconformer_trancribe": 0.4252873659133911,
    "tem_sb_wav2vec2_trancribe": 0.7471264600753784,
    "tem_sb_whisper_trancribe": 0.7471264600753784,
    "tem_whisper_transcribe": 0.3333333432674408,
    "temps_calcul_nvidia_conftrans_trancribe": 5.2904438972473145,
    "temps_calcul_nvidia_fastconformer_trancribe": 4.6708152294158936,
    "temps_calcul_sb_wav2vec2_trancribe": 7.636538982391357,
    "temps_calcul_sb_whisper_trancribe": 23.4919753074646,
    "temps_calcul_whisper_transcribe": 34.602197885513306
},
"output_segments_30s/A1_16-01-18_677_707.wav": {
    "hypothese_nvidia_conftrans_trancribe": "on dit que c était le légendaire
    ↪  tondmon l oiseau tonnain c que je me rappelle que j ai vu cet évêque
    ↪  mais je l ai vraiment vu de mes propres yeux c était un oiseau assez
    ↪  immense gros selon mes estimations l envergure de cet oiseau là
    ↪  cinquante cinquante trois mètres pila de du bec jusqu à la pointe de la
    ↪  queue dix neuf mètres de",
    "hypothese_nvidia_fastconformer_trancribe": "on dit que c était le
    ↪  légendaire tombemon l oiseau tenait c est que je me rappelle que j ai
    ↪  vu cet évêque mais je l ai vraiment vu dans mes propres yeux c était un
    ↪  oiseau assez immense gros selon mes estimations l envergure de cet
    ↪  oiseau là cinquante trois mètres pila de la du bec jusqu à la pointe de
    ↪  la queue dix neuf mètres de long",
    "hypothese_sb_wav2vec2_trancribe": "on it u c était le légendaire tanl ouis
    ↪  souterrain ce u e me rapplle u ai vu cet ivêtue mais je ilai vraient vu
    ↪  dans mes propres yeux c ait un oiseau assez immenses gros son mes
    ↪  estimatons l enverure e cette de cet oiseau l cinuante deu cinuante
    ↪  trois mlts pilt de la dubet jusu à la pointe de la ueue ix nuf mtres de
    ↪  long",
    "hypothese_sb_whisper_trancribe": "on dit que c était le légendaire
    ↪  tandemain l oiseau tannais c était que je me rappelle que j y ai vu c
    ↪  était vert mais j y l ai vraiment vu dans mes propres yeux c était un
    ↪  oiseau assez immense gros selon mes estimations l envergure de cet
    ↪  oiseau là cinquante trois mètres puis de la du bec jusqu à la pointe de
    ↪  la",
    "hypothese_whisper_transcribe": " on dit que c était le légendaire
    ↪  tandemann l oiseau tonnais c est ce que je me rappelle que j ai vu c
    ↪  était vert mais j ai vraiment vu dans mes propres yeux c était un
    ↪  oiseau assez immense gros selon mes estimations l envergure de cet
    ↪  oiseau là 53 mètres puis du bec jusqu à la pointe de la queue 19 mètres
    ↪  de long ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8741499185562134,
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.878187894821167,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.7517893314361572,
    "similarite_cosine_sb_whisper_trancribe": 0.8866598010063171,
    "similarite_cosine_whisper_transcribe": 0.8896396160125732,
```

```
    "tec_nvidia_conftrans_trancribe": 0.2222222238779068,
    "tec_nvidia_fastconformer_trancribe": 0.2012012004852295,
    "tec_sb_wav2vec2_trancribe": 0.30930930376052856,
    "tec_sb_whisper_trancribe": 0.23723722994327545,
    "tec_whisper_transcribe": 0.13513512909412384,
    "tem_nvidia_conftrans_trancribe": 0.2535211145877838,
    "tem_nvidia_fastconformer_trancribe": 0.23943662643432617,
    "tem_sb_wav2vec2_trancribe": 0.577464759349823,
    "tem_sb_whisper_trancribe": 0.28169015049934387,
    "tem_whisper_transcribe": 0.19718310236930847,
    "temps_calcul_nvidia_conftrans_trancribe": 5.280596733093262,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.407588958740234,
    "temps_calcul_sb_wav2vec2_trancribe": 7.4593470096588135,
    "temps_calcul_sb_whisper_trancribe": 27.49884796142578,
    "temps_calcul_whisper_transcribe": 33.56988716125488
},
"output_segments_30s/A1_23-10-18_1311_1341.wav": {
    "hypothese_nvidia_conftrans_trancribe": "ôts on nous oppose a pris à à nous
    ↪  prendre dans le broc card on était jeune les calets des parents nont
    ↪  pauvrement connu sont pour un ous c est des mal mais nous donnz nous
    ↪  trois repos par jour par exemple ils nous donnaient toutes ce qu on
    ↪  avait besoin ils nous habillaient cette même nous montraient leur amour
    ↪  mais ils nous ont jamais pris dans leurs broc il moins ça manque moi je
    ↪  suis quelqu",
    "hypothese_nvidia_fastconformer_trancribe": "nos hôtes on nous a peu appris
    ↪  à nous prendre dans nos brocas qu on était jeune les calets des parents
    ↪  n ont pas vraiment connu ce pour eux c était mal ils nous donnaient
    ↪  nous trois repas par jour par exemple ils nous donnaient tout ce qu on
    ↪  avait de besoin ils nous habillaient c est de même qu ils nous
    ↪  montraient leur amour ils nous ont jamais pris dans leurs brocs une moi
    ↪  ça manque moi je suis quelqu un",
    "hypothese_sb_wav2vec2_trancribe": "pinousot on nous a appris nous prendre
    ↪  ns nos pracr on était eune les calais des parents lont pas vraiment
    ↪  connus pour les autreses males ils nous donner notr rpas par our par ep
    ↪  is ous donna tout ce u on avait bsoin ils nous billait cet même pilrum
    ↪  ui vous montrar leur amour is n vous ont amais pris dans leurs prasils
    ↪  mois a manxe moi e suis uu un",
    "hypothese_sb_whisper_trancribe": "puis nous autres on ne nous a pas appris
    ↪  à nous prendre dans nos bras quand on était jeune les calais des
    ↪  parents n ont pas vraiment connu ça pour eux autres c était mal mais
    ↪  ils nous donnaient notre repos par jour par exemple ils nous donnaient
    ↪  tout ce qu on avait de besoin ils nous habillaient c est même qu qu ils
    ↪  nous montraient leur amour ma",
    "hypothese_whisper_transcribe": " et puis nous autres on nous a pas appris
    ↪  à nous prendre dans nos bras quand on était jeunes les câlins des
    ↪  parents on n a pas vraiment connu ça pour eux autres c était mal ils
    ↪  nous donnaient nos trois repas par jour par exemple ils nous donnaient
    ↪  tout ce qu on avait besoin ils nous habillaient c est de même qu ils
    ↪  nous montraient leur amour mais ils nous ont jamais pris dans leurs
    ↪  bras et puis de moins ça me manque ça moi je suis quelqu un qui ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.7370052337646484,
```

```
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.8088080286979675,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.6052432060241699,
    "similarite_cosine_sb_whisper_trancribe": 0.9451424479484558,
    "similarite_cosine_whisper_transcribe": 0.9954878091812134,
    "tec_nvidia_conftrans_trancribe": 0.25355449318885803,
    "tec_nvidia_fastconformer_trancribe": 0.12796208262443542,
    "tec_sb_wav2vec2_trancribe": 0.28436020016670227,
    "tec_sb_whisper_trancribe": 0.26777252554893494,
    "tec_whisper_transcribe": 0.06635071337223053,
    "tem_nvidia_conftrans_trancribe": 0.4651162922382355,
    "tem_nvidia_fastconformer_trancribe": 0.23255814611911774,
    "tem_sb_wav2vec2_trancribe": 0.6162790656089783,
    "tem_sb_whisper_trancribe": 0.3604651093482971,
    "tem_whisper_transcribe": 0.09302325546741486,
    "temps_calcul_nvidia_conftrans_trancribe": 5.414780855178833,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.404529571533203,
    "temps_calcul_sb_wav2vec2_trancribe": 7.542479991912842,
    "temps_calcul_sb_whisper_trancribe": 27.49142861366272,
    "temps_calcul_whisper_transcribe": 37.31102800369263
},
"output_segments_30s/A1_23-10-18_1540_1570.wav": {
    "hypothese_nvidia_conftrans_trancribe": "c est un doucier qui s excida des
    ↪  trou trop longtemps mais pourquoi qui me l ont faite faire d abord st
    ↪  ils été hores là je comprends un peor mais c est quoi qui se pose
    ↪  madame hervieux si je vous posais la question aujourd hui si c était à
    ↪  refaire vous revoyez l émission en quête ou des stes porter plainte si
    ↪  c était à refaire est ce que vous leur referiez oui oui",
    "hypothese_nvidia_fastconformer_trancribe": "c était un dossier qui s
    ↪  excédât trop longtemps mais pourquoi ils me l ont fait faire d abord ce
    ↪  qu ils étaient des horloges je comprends paulissez quoi qui se passe
    ↪  madame hervieux si je vous posais la question aujourd hui si c était à
    ↪  refaire on vous revoyait l émission enquête ou décidez de porter
    ↪  plainte si c était à refaire est ce que vous le referiez oui oui ",
    "hypothese_sb_wav2vec2_trancribe": "faites le docier ui ecidatdtroup trop
    ↪  longtemps mais pouruoi guilmel ont faits faire d abord tstes orlo e
    ↪  omprends moc est quoi qui se pose maame mervieux si e ous posais la
    ↪  uesion auourdui si c tait à rfaire o vous renvoit l émission enquête
    ↪  pous cider de porter painte si c it à rfaire est ce ue vous le
    ↪  refrezoui oui",
    "hypothese_sb_whisper_trancribe": "c était un dossier qui s est daté trop
    ↪  trop longtemps mais pourquoi qu ils me l ont fait faire d abord cette
    ↪  des heures là je comprends pas moi c est quoi qui se passe madame
    ↪  mervieux si je vous posais la question aujourd hui si c était à refaire
    ↪  vous revoyer l émission enquête vous décider porter plainte si",
    "hypothese_whisper_transcribe": " c était un dossier qui s est daté trop
    ↪  longtemps mais pourquoi qu ils me l ont fait faire d abord ces heures
    ↪  là je comprends pas moi c est quoi qui se passe madame merveilleux si
    ↪  je vous posais la question aujourd hui si c était à refaire vous
    ↪  revoyez l émission enquête vous décidez de porter plainte si c était à
    ↪  refaire est ce que vous le referiez oui ",
    "reference": "n avaient pas une seule nation ils faisaient comme celui de
    ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
    ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
    ↪  deux dans le même souffle la communauté qui faisait une demande à la
    ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
    ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
    ↪  cette nation là a un grand conseil de la nation donc un moment donné
    ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
    "similarite_cosine_nvidia_conftrans_trancribe": 0.8791674375534058,
```

```
    "similarite_cosine_nvidia_fastconformer_trancribe": 0.9598582983016968,
    "similarite_cosine_sb_wav2vec2_trancribe": 0.6171836853027344,
    "similarite_cosine_sb_whisper_trancribe": 0.9784035682678223,
    "similarite_cosine_whisper_transcribe": 0.9900859594345093,
    "tec_nvidia_conftrans_trancribe": 0.17977528274059296,
    "tec_nvidia_fastconformer_trancribe": 0.12921348214149475,
    "tec_sb_wav2vec2_trancribe": 0.2247191071510315,
    "tec_sb_whisper_trancribe": 0.24438202381134033,
    "tec_whisper_transcribe": 0.07303370535373688,
    "tem_nvidia_conftrans_trancribe": 0.3333333432674408,
    "tem_nvidia_fastconformer_trancribe": 0.1944444477558136,
    "tem_sb_wav2vec2_trancribe": 0.5694444179534912,
    "tem_sb_whisper_trancribe": 0.3472222089767456,
    "tem_whisper_transcribe": 0.1111111119389534,
    "temps_calcul_nvidia_conftrans_trancribe": 5.367574691772461,
    "temps_calcul_nvidia_fastconformer_trancribe": 5.226211309432983,
    "temps_calcul_sb_wav2vec2_trancribe": 7.511625289916992,
    "temps_calcul_sb_whisper_trancribe": 27.303059816360474,
    "temps_calcul_whisper_transcribe": 34.138951539993286
},
"output_segments_30s/A1_23-10-18_174_204.wav": {
    "hypothese_nvidia_conftrans_trancribe": "la salle fera sa demande de vuis
    ↪  clos lorsqu elle sera en présence tout à l heure oui où sonnez vous lor
    ↪  son lisère oui osi déjà telle c est déjà cté si voudrait mentionner à
    ↪  maître la salle que son témoin pourra témoigner à vuiscloses pesce
    ↪  beaucoup alors aujourd hui monsieur le commissaire nous avons madme
    ↪  marie louise hervieux qui est de pessamite elle est accompagnée de son
    ↪  conjoint m robert lapointe et",
    "hypothese_nvidia_fastconformer_trancribe": "la salle fera sa demande de
    ↪  huis clos lorsqu elle sera en présence tout à l heure où sommez vous
    ↪  oui c est déjà parfait je ferai mentionner à mettre à la salle son
    ↪  témoin pourra témoigner à huis clos faites passe beaucoup alors aujourd
    ↪  hui monsieur le commissaire nous avons madame marie louise hervieux qui
    ↪  est de pessamit elle est accompagnée de son conjoint monsieur robert
    ↪  lapointe et",
    "hypothese_sb_wav2vec2_trancribe": "la salle fera sa ane de huisclos orsu
    ↪  ele se ra en présente tout à l heure vousrdonnesnatui asou tojatre so
    ↪  toujoursrvrairement sormu a metraasal ue son témoin prouate monioviclo
    ↪  fat marci beaucoup lauourhui monsieur l issaire nus avons mae marie
    ↪  louise ervieux ui est de pissamits elle est acne son onoint monsieur
    ↪  rabert lapointeet",
    "hypothese_sb_whisper_trancribe": "elle a salle fera sa demande de huitlot
    ↪  lorsqu elle sera en présence tout à l heure ou soignez vous l ordonner
    ↪  immédiatement oui oui c est déjà accordé faurait mentionner à mettre
    ↪  elle açalle que son témoin peura témoigner à huitlot perfect
    ↪  personneur",
    "hypothese_whisper_transcribe": " la salle fera sa demande de huitlots
    ↪  lorsqu elle sera en présence tout à l heure ou soit il vous l ordonner
    ↪  immédiatement oui c est déjà accorder faudrait mentionner à maître
    ↪  alassalle que son témoin pourra témoigner à huitlots parfait merci
    ↪  beaucoup alors aujourd hui monsieur le commissaire nous avons madame
    ↪  marie louise hervieux qui est de pessamites elle est accompagnée de son
    ↪  conjoint monsieur robert lapointe ",
```

```
        "reference": "n avaient pas une seule nation ils faisaient comme celui de
        ↪  manawan s occupait également de kanesatake donc ils ne devenaient pas
        ↪  bons avec juste une nation ils étaient tout le temps à cheval entre les
        ↪  deux dans le même souffle la communauté qui faisait une demande à la
        ↪  sûreté du québec pouvait recevoir une réponse x de montréal une réponse
        ↪  y de trois rivières puis une réponse z du saguenay lac saint jean et
        ↪  cette nation là a un grand conseil de la nation donc un moment donné
        ↪  ils se parlaient puis ils disaient bien voyons j ai 3 réponses",
        "similarite_cosine_nvidia_conftrans_trancribe": 0.8602554798126221,
        "similarite_cosine_nvidia_fastconformer_trancribe": 0.8706654906272888,
        "similarite_cosine_sb_wav2vec2_trancribe": 0.6735278367996216,
        "similarite_cosine_sb_whisper_trancribe": 0.772281289100647,
        "similarite_cosine_whisper_transcribe": 0.8801697492599487,
        "tec_nvidia_conftrans_trancribe": 0.2868369221687317,
        "tec_nvidia_fastconformer_trancribe": 0.29273083806037903,
        "tec_sb_wav2vec2_trancribe": 0.4381139576435089,
        "tec_sb_whisper_trancribe": 0.5697445869445801,
        "tec_whisper_transcribe": 0.23968565464019775,
        "tem_nvidia_conftrans_trancribe": 0.4204545319080353,
        "tem_nvidia_fastconformer_trancribe": 0.3636363744735718,
        "tem_sb_wav2vec2_trancribe": 0.7613636255264282,
        "tem_sb_whisper_trancribe": 0.6590909361839294,
        "tem_whisper_transcribe": 0.35227271914482117,
        "temps_calcul_nvidia_conftrans_trancribe": 5.510642051696777,
        "temps_calcul_nvidia_fastconformer_trancribe": 4.819784879684448,
        "temps_calcul_sb_wav2vec2_trancribe": 7.441242218017578,
        "temps_calcul_sb_whisper_trancribe": 24.463451862335205,
        "temps_calcul_whisper_transcribe": 35.33129072189331
    },
    "similarite_cosine_moyen_nvidia_conftrans_trancribe": 0.881527821222941,
    "similarite_cosine_moyen_nvidia_fastconformer_trancribe": 0.9096531669298807,
    "similarite_cosine_moyen_sb_wav2vec2_trancribe": 0.6414618558353848,
    "similarite_cosine_moyen_sb_whisper_trancribe": 0.8446069094869826,
    "similarite_cosine_moyen_whisper_transcribe": 0.9488603274027506,
    "tec_moyen_nvidia_conftrans_trancribe": 0.1834719330072403,
    "tec_moyen_nvidia_fastconformer_trancribe": 0.1624220311641693,
    "tec_moyen_sb_wav2vec2_trancribe": 0.2954781651496887,
    "tec_moyen_sb_whisper_trancribe": 0.4495841860771179,
    "tec_moyen_whisper_transcribe": 0.12032224237918854,
    "tem_moyen_nvidia_conftrans_trancribe": 0.3177570104598999,
    "tem_moyen_nvidia_fastconformer_trancribe": 0.24566088616847992,
    "tem_moyen_sb_wav2vec2_trancribe": 0.6395193338394165,
    "tem_moyen_sb_whisper_trancribe": 0.5313751697540283,
    "tem_moyen_whisper_transcribe": 0.18291054666042328,
    "temps_moyen_nvidia_conftrans_trancribe": 5.610402743021647,
    "temps_moyen_nvidia_fastconformer_trancribe": 5.429712083604601,
    "temps_moyen_sb_wav2vec2_trancribe": 8.061181757185194,
    "temps_moyen_sb_whisper_trancribe": 25.253914965523613,
    "temps_moyen_whisper_transcribe": 35.02862779299418,
    "temps_total_nvidia_conftrans_trancribe": 50.49362397193909,
    "temps_total_nvidia_fastconformer_trancribe": 48.86740708351135,
    "temps_total_sb_wav2vec2_trancribe": 72.5506341457367,
    "temps_total_sb_whisper_trancribe": 227.28523325920105,
    "temps_total_whisper_transcribe": 315.2576484680176
}
```


Listing C.2 – Résultats des modèles sur les segments de 30 secondes

# ANNEXE D
# TRANSCRIPTION AVEC LA PREUVE DE CONCEPT

Ce fichier est un extrait d'un début d'audience de la commission Viens. Sa durée est de 3 minutes et 58 secondes. Le prototype prend environ 8 mins à le transcrire avec un modèle de haute qualité (modèle whisper « medium »). 4 minutes de transcriptions, 4 mins de diarisation.

```
0.00 6.00 SPEAKER_03  de la commission.
6.00 8.00 SPEAKER_03  La commission d'enquête sur les relations entre les
8.00 14.00 SPEAKER_02  Autochtones et certains services publics au Québec,  prise
↪  d'éparele honorable de Jacques Viens, est maintenant  ouverte.
14.00 22.00 SPEAKER_02  Alors, bonjour. Bienvenue dans cette autre semaine de nos
↪  audiences à Val-d'Or en territoire à Nuchinabé. Je vais  demander au procureur
↪  de s'identifier tout d'abord  pour les fins de l'enregistrement.
22.00 24.00 SPEAKER_03  Donald Bourget pour la commission.
24.00 26.00 SPEAKER_02  Bonjour, jargon.
26.00 28.00 SPEAKER_03  Bonjour, monsieur le commissaire.
28.00 30.00 SPEAKER_03  Bonjour, maître Miller pour la Ferme autochtone du Québec.
30.00 32.00 SPEAKER_03  Bonjour, maître Miller.
32.00 36.00 SPEAKER_03  Bonjour, maître Marie-Paul Boucher pour la procureur
↪  générale du Québec.
36.00 38.00 SPEAKER_03  Bonjour, maître Boucher.
38.00 42.00 SPEAKER_02  Bonjour, Donnise Robillard pour la procureur générale du
↪  Québec.
42.00 44.00 SPEAKER_01  Bonjour, maître Robillard.
44.00 48.00 SPEAKER_01  Alors, maître Bourget, maître Richard, vous allez nous
↪  présenter le programme de la journée?
48.00 54.00 SPEAKER_01  Oui, en fait, cette après-midi, nous entendrons  Madame
↪  Vianney et Madame Charon du CISAT  au niveau de la santé et des services
↪  sociaux.
54.00 62.00 SPEAKER_01  Mais pour l'heure, ce matin, nous  comprenons que la
↪  compréhension de la santé est à
62.00 66.00 SPEAKER_02  l'écran de la province.  Nous avons aussi le directeur de
↪  la protection des
66.00 70.00 SPEAKER_01  jeunes ici en région de  Bitté-Bitté-Misskemenque.
70.00 74.00 SPEAKER_01  Alors, je comprends que vous allez être prêt après que la
↪  graffière aura assormenté M. Gagné?
```

FIGURE D.1 – Transcription d'un extrait audio d'un début d'audience de la commission Viens.

```
74.00 78.00 SPEAKER_03  Oui, tout à fait. Je ferai une  brève présentation du témoin.
78.00 82.00 SPEAKER_02  Il pourra compléter en connaissance personnelle  de cause.
82.00 84.00 SPEAKER_02  Vous affirmez salariellement de dire la vérité?
84.00 88.00 SPEAKER_02  Oui, c'est très agréable de vous  être reçu.
88.00 100.00 SPEAKER_01  Alors, on comprend que  je pense que c'est à l'automne  que vous
↪ étiez venu en panel  avec deux de vos collègues des PSJ d'Outaouais  et de Côte-Nord.
↪ Maintenant, vous êtes  seul au pretoire.
100.00 116.00 SPEAKER_01  Alors, je comprends que vous êtes des PSJ  depuis 2011 en
↪ région de Bitté-Bitté-Misskemenque  et que vous avez fait du kilométrage avant  dans
↪ les services de protection  de la jeunesse et particulièrement  des communautés
↪ autochtones. Pouvez-vous  nous faire le parcours de votre  carrière
116.00 126.00 SPEAKER_00  ici en Abitibi?  Avec plaisir. Donc, effectivement,  j'ai
↪ commencé ma carrière professionnelle à la direction  de la protection de la jeunesse
↪ en  1994.
126.00 144.00 SPEAKER_00  Et puis, la première affectation  au secteur autochtone a été
↪ en  1995. Donc, j'ai été embauché  à l'époque. Il y avait une liste d'attentes
↪ importantes d'enfants en attente  d'évaluation. Donc, mon premier défi  professionnel
↪ d'importance,  on va dire comme ça, a été  d'oeuvrer auprès de la communauté de
↪ L'Aximont.
144.00 158.00 SPEAKER_00  J'ai fait ça  un an à peu près pour ensuite  repartir pour le
↪ Témiscamingue avec  des responsabilités multiprogrammes,  c'est-à-dire que je
↪ m'occupais de l'ensemble  de la trajectoire en protection  de la jeunesse, adoption
↪ et le GPA.
158.00 164.00 SPEAKER_00  Par la suite, je suis  revenu dans  l'Abitibi.
164.00 168.00 SPEAKER_00  J'ai finalisé mon  parcours d'éducation universitaire.
168.00 172.00 SPEAKER_00  Donc, j'ai fini mon bac à l'OREA  à l'Université du Québec en
↪ Abitibi-Témiscamingue.
172.00 176.00 SPEAKER_00  Donc, je suis bachelier en travail  social.
176.00 190.00 SPEAKER_00  J'ai  vers les années 2001-2002  travaillé pour  l'agence
↪ Minogun. Donc, j'ai été  embauché comme superviseur clinique  en remplacement d'un
↪ congé de maternité.
190.00 200.00 SPEAKER_00  Par la suite, quand la fermeture  des services sociaux Minogun
↪ a eu lieu  en 2002, j'ai été nommé  président du directeur de la protection de la
↪ jeunesse  pour les communautés autochtones.
200.00 212.00 SPEAKER_00  Donc, mon travail était  de m'assurer  la viabilité des
↪ ententes  qu'on avait avec les communautés,  notamment sur la vigie en regard  des
↪ responsabilités 32.
212.00 220.00 SPEAKER_00  Donc, je faisais la révision  des situations et je m'occupais
↪ de la bonne collaboration avec  l'agence Minogun.
220.00 228.00 SPEAKER_00  Par la suite, en 2002-2003,  j'ai été nommé chef de service au
↪ programme  évaluation et orientation dans la MRC Val-et-L'Art.
228.00 254.00 SPEAKER_00  Et par la suite,  en 2011, comme vous l'avez mentionné,  j'ai
↪ été nommé à la direction  de la protection de la jeunesse,  poste que j'occupe depuis
↪ 2011.
```

FIGURE D.2 – Transcription d'un extrait audio d'un début d'audience de la commission Viens.

## BIBLIOGRAPHIE


An, K., Xiang, H. et Ou, Z. (2020). CAT : A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency. *arXiv*. https://doi.org/10.48550/arXiv.2005.13326

Anjos, A., El-Shafey, L., Wallace, R., Günther, M., McCool, C. et Marcel, S. (2012). Bob : A free signal processing and machine learning toolbox for researchers. Dans *20th ACM International Conference on Multimedia*, 1449–1452. ACM

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. et Weber, G. (2020). Common voice : A massively-multilingual speech corpus. Dans *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222. ELRA

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, *5*(9), 341–345. https://cir.nii.ac.jp/crid/1572261550900588928

Bonastre, J.-F., Wils, F. et Meignier, S. (2005). ALIZE, a free toolkit for speaker recognition. Dans *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 737–740. IEEE

Bredin, H. (2023). pyannote. audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. Dans *Interspeech 2023*, 1983–1987. ISCA

Bredin, H. et Laurent, A. (2021). End-to-End Speaker segmentation for overlap-aware resegmentation. Dans *Interspeech 2021*, 3111–3115. ISCA

Chadha, H. S., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R. et Raghavan, V. (2022). Vakyansh : ASR toolkit for low resource indic languages. *arXiv*. https://doi.org/10.48550/arXiv.2203.16512

Chiu, C.-C., Kannan, A., Prabhavalkar, R., Chen, Z., Sainath, T., Han, W., Zhang, Y., Pang, R., Kishchenko, S., Nguyen, P., Narayanan, A., Liao, H. et Zhang, S. (2019). A comparison of end-to-end models for long-form speech recognition. Dans *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 889–896. IEEE

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C. et Bapna, A. (2023). Fleurs : Few-shot learning evaluation of universal representations of speech. Dans *2022 IEEE Spoken Language Technology Workshop (SLT)*, 798–805. IEEE

Davis, K. H., Biddulph, R. et Balashek, S. (1952). Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America*, *24*(6), 637–642. https://doi.org/10.1121/1.1906946

Del Rio, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Żelasko, P. et Jetté, M. (2021). Earnings-21 : A Practical Benchmark for ASR in the Wild. Dans *Interspeech 2021*, 3465–3469. ISCA

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

Ferreira, E., Nocera, P., Goudi, M. et Do Thi, N. D. (2012). YAST : A scalable ASR toolkit especially designed for under-resourced languages. Dans *2012 International Conference on Asian Language Processing*, 141–144. IEEE

Giannakopoulos, T. (2015). pyaudioanalysis : An open-source python library for audio signal analysis. *PLOS ONE*, *10*, 1–17. https://doi.org/10.1371/journal.pone.0144610

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F. et Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, *10*, 522–538. https://doi.org/10.1162/tacl_a_00474

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. *et al.* (2020). Conformer : Convolution-augmented transformer for speech recognition. *arXiv*. https://doi.org/10.48550/arXiv.2005.08100

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. et Ng, A. Y. (2014). Deep Speech : Scaling up end-to-end speech recognition. *arXiv*. https://doi.org/10.48550/arXiv.1412.5567

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hayashi, T., Yamamoto, R., Inoue, K., Yoshimura, T., Watanabe, S., Toda, T., Takeda, K., Zhang, Y. et Tan, X. (2020). ESPnet-TTS : Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit. *arXiv*. https://doi.org/10.48550/arXiv.1910.10909

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N. et Esteve, Y. (2018). Ted-lium 3 : Twice as much data and corpus repartition for experiments on speaker adaptation. Dans *Speech and Computer : 20th International Conference*, 198–208. Springer

Huang, W. R., Chang, S.-y., Rybach, D., Prabhavalkar, R., Sainath, T. N., Allauzen, C., Peyser, C. et Lu, Z. (2022). E2E Segmenter : Joint Segmenting and Decoding for Long-Form ASR. *arXiv*. https://doi.org/10.48550/arXiv.2204.10749

Inaguma, H., Kiyono, S., Duh, K., Karita, S., Soplin, N. E. Y., Hayashi, T. et Watanabe, S. (2020). ESPnet-ST : All-in-One Speech Translation Toolkit. *arXiv*. https://doi.org/10.48550/arXiv.2004.10234

Kang, T. G., Kim, H.-G., Lee, M.-J., Lee, J. et Lee, H. (2021). Partially Overlapped Inference for Long-Form Speech Recognition. Dans *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5989–5993. IEEE

Kendall, T. et Farrington, C. (2023). The corpus of regional african american language

Khoury, E., Shafey, L. E. et Marcel, S. (2014). Spear : An open source toolbox for speaker recognition based on Bob. Dans *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1655–1659. IEEE

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J. *et al.* (2019). NeMo : A toolkit for building AI applications using Neural Modules. *arXiv*. https://doi.org/10.48550/arXiv.1909.09577

Larcher, A., Lee, K. A. et Meignier, S. (2016). An extensible speaker identification sidekit in Python. Dans *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5095–5099. IEEE

Levenshtein, V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, *10*(8), 707–710.

Levy, G., Sitman, R., Amir, I., Golshtein, E., Mochary, R., Reshef, E., Reichart et Allouche, O. (2019). Gecko - a tool for effective annotation of human conversations. Dans *Interspeech 2019*. ISCA

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y. et Wang, Y. (2022). Neufa : Neural network based end-to-end forced alignment with bidirectional attention mechanism. Dans *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8007–8011. IEEE

Liu, A. T., Li, S.-W. et yi Lee, H. (2020a). Tera : Self-supervised learning of transformer encoder representation for speech. *arXiv*. https://doi.org/10.48550/arXiv.2007.06028

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c. et Lee, H.-y. (2020b). Mockingjay : Unsupervised speech representation learning with deep bidirectional transformer encoders. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp40776.2020.9054458

Loakes, D. (2024). Automatic speech recognition and the transcription of indistinct forensic audio : how do the new generation of systems fare ? *Frontiers in Communication*, *9*, 1–9. https://doi.org/10.3389/fcomm.2024.1281407/full

Louw, S. (2021). Automated transcription software in qualitative research. Dans *Proceedings of the International Conference*, 1–12. ACM

Lu, L., Xiao, X., Chen, Z. et Gong, Y. (2019). Pykaldi2 : Yet another speech toolkit based on kaldi and pytorch. *arXiv*. https://doi.org/10.48550/arXiv.1907.05955

Maupomé, D., Armstrong, M. D., Rancourt, F., Soulas, T. et Meurs, M.-J. (2021). Early detection of signs of pathological gambling, self-harm and depression on social media. Dans *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CLEF-WN 2021)*, volume 2936 de *CEUR-WS.org*, 1031–1045. CEUR Workshop Proceedings

Maupomé, D., Soulas, T., Rancourt, F., Cantin-Savoie, G., Winterstein, G., Mosser, S. et Meurs, M.-J. (2023). Lightweight methods for early risk detection. Dans *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF-WN 2023)*, volume 3497 de *CEUR-WS.org*, 718–726. CEUR Workshop Proceedings

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. et Sonderegger, M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. Dans *Interspeech 2017*, 498–502. ISCA

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://doi.org/10.48550/arXiv.1301.3781

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. et Auli, M. (2019). Fairseq : A Fast, Extensible Toolkit for Sequence Modeling. *arXiv*. https://doi.org/10.48550/arXiv.1904.01038

Panayotov, V., Chen, G., Povey, D. et Khudanpur, S. (2015). Librispeech : an asr corpus based on public domain audio books. Dans *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. IEEE

Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., Heitkaemper, J., Olvera, M., Stöter, F.-R., Hu, M., Martín-Doñas, J. M., Ditter, D., Frank, A., Deleforge, A. et Vincent, E. (2020). Asteroid : The PyTorch-based audio source separation toolkit for researchers. *arXiv*. https://doi.org/10.48550/arXiv.2005.04132

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. *et al.* (2011). The kaldi speech recognition toolkit. Dans *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. et Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv :2212.04356*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., ... et Bengio, Y. (2021). SpeechBrain : A General-Purpose Speech Toolkit. *arXiv*. https://doi.org/10.48550/arXiv.2106.04624

Reimers, N. et Gurevych, I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. Dans K. Inui, J. Jiang, V. Ng, et X. Wan (dir.). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992., Hong Kong, China. ACL

Rio, M. D., Ha, P., McNamara, Q., Miller, C. et Chandra, S. (2022). Earnings-22 : A practical benchmark for accents in the wild. *arXiv*. https://doi.org/10.48550/arXiv.2203.15591

Robertson, S. (2004). Understanding inverse document frequency : On theoretical arguments for idf. *Journal of Documentation*, *60*, 503–520. https://doi.org/10.1108/00220410410560582

Roux, T. B., Rouvier, M., Wottawa, J. et Dufour, R. (2022). Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. Dans *Interspeech 2022*, 3968–3972. ISCA

Salaün, Y., Vincent, E., Bertin, N., Souviraà-Labastie, N., Jaureguiberry, X., Tran, D. T. et Bimbot, F. (2014). The Flexible Audio Source Separation Toolbox Version 2.0. ICASSP

Saravani, S. H. H., Normand, L., Maupomé, D., Rancourt, F., Soulas, T., Besharati, S., Normand, A., Mosser, S. et Meurs, M.-J. (2022). Measuring the severity of the signs of eating disorders using similarity-based approaches. Dans *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (CLEF-WN 2022)*, volume 3180 de *CEUR-WS.org*, 936–946. CEUR Workshop Proceedings

Schuller, B., Lehmann, A., Weninger, F., Eyben, F. et Rigoll, G. (2009). Blind enhancement of the rhythmic and harmonic sections by nmf : Does it help ? Dans *Proceedings of International Conference on Acoustics (NAG/DAGA 2009)*, 361–364. IEEE

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Srivastav, V., Majumdar, S., Koluguri, N., Moumen, A., Gandhi, S., Hugging Face Team, Nvidia NeMo Team et SpeechBrain Team (2023). Open automatic speech recognition leaderboard. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

Tkachenko, M., Malyuk, M., Holmanyuk, A. et Liubimov, N. (2020-2022). Label Studio : Data labeling software. Open source software available from https ://github.com/heartexlabs/label-studio

Wang, C., Hsu, W.-N., Adi, Y., Polyak, A., Lee, A., Chen, P.-J., Gu, J. et Pino, J. (2021a). fairseq sˆ 2 : A scalable and integrable speech synthesis toolkit. *arXiv*. https://doi.org/10.48550/arXiv.2109.06912

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J. et Dupoux, E. (2021b). VoxPopuli : A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. Dans *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 993–1003. ACL

Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D. et Pino, J. (2022a). fairseq S2T : Fast speech-to-text modeling with fairseq. *arXiv*. https://doi.org/10.48550/arXiv.2010.05171

Wang, J., Tong, X., Guo, J., He, D. et Maas, R. (2022b). VADOI : Voice-activity-detection overlapping inference for end-to-end long-form speech recognition. Dans *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8367–8371. IEEE

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. et Ochiai, T. (2018). ESPnet : End-to-End Speech Processing Toolkit. *arXiv*. https://doi.org/10.48550/arXiv.1804.00015

Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C. *et al.* (2021). C2si corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, *55*(1), 173–190. https://doi.org/10.1007/s10579-020-09496-3

Wollin-Giering, S., Hoffmann, M., Höfting, J. et Ventzke, C. (2024). Automatic transcription of english and german qualitative interviews. *Forum : Qualitative Social Research*, *25*(1). https://doi.org/10.17169/fqs-25.1.4129

Xiang, H. et Ou, Z. (2019). CRF-based single-stage acoustic modeling with CTC topology. Dans *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5676–5680. IEEE

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., ... et Lee, H.-y. (2021). SUPERB : Speech processing Universal PERformance Benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2105.01051

Zhang, H., Yuan, T., Chen, J., Li, X., Zheng, R., Huang, Y., Chen, X., Gong, E., Chen, Z., Hu, X., Yu, D., Ma, Y. et Huang, L. (2022). Paddlespeech : An easy-to-use all-in-one speech toolkit. *arXiv*. https://doi.org/10.48550/arXiv.2205.12007

Zhao, C., Wang, M., Dong, Q., Ye, R. et Li, L. (2021). NeurST : Neural Speech Translation Toolkit. Dans *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations*, 55–62. ACL