Optimisation des performances d'un système de reconnaissance automatique de la parole pour les commentaires sportifs: fine-tuning de Whisper

Camille Lavigne, Alex Stasica, Anna Kupsc

Optimisation des performances d'un système de reconnaissance automatique de la parole pour les commentaires sportifs : fine-tuning de Whisper

Camille Lavigne Alex Stasica[1] Anna Kupsc[2]
(1) Utrecht University, Utrecht,3500-3585, Pays-Bas
(2) CLLE-ERSSàB-UMR-5263, Université Bordeaux Montaigne, 33600, Pessac, France
lavignecamille@yahoo.fr, a.stasica@uu.nl,
anna.kupsc@u-bordeaux-montaigne.fr

RÉSUMÉ

Malgré les performances élevées des systèmes automatiques de reconnaissance de la parole (Automatic Speech Recognition ; ASR) sur des corpus généraux, leur efficacité est considérablement réduite lorsqu'ils sont confrontés à des corpus spécialisés. Ces corpus peuvent notamment contenir du lexique propre à des domaines spécifiques, des accents ou du bruit de fond rendant la transcription ardue. Cette étude vise à évaluer les avantages de l'optimisation d'une transcription automatique, par opposition à manuelle, après *fine-tuning* d'un modèle d'ASR de dernière génération, Whisper (Radford *et al.*, 2023), sur un corpus spécialisé de commentaires sportifs de petite taille. Nos analyses quantitatives et qualitatives indiquent que Whisper est capable d'apprendre les particularités d'un corpus de spécialité, atteignant des performances égales où supérieures aux transcripteurs humains, avec une quantité de données limitée. Cette recherche met en lumière le rôle que l'intelligence artificielle, notamment les larges modèles de langage, peut jouer pour faciliter la création de corpus spécialisés.

ABSTRACT

Performance optimization of an automatic speech recognition system for sport commentaries : Whisper fine-tuning

Despite the great performance of automatic speech recognition (ASR) systems on general corpora, their performance is greatly impacted when confronted with specialized corpora. These corpora can include specialized lexicon, accents or background noise in the audio making the transcription harder. This study aims to evaluate the potential benefits of optimizing automatic transcription, as opposed to a manual one, by *fine-tuning* a state-of-the-art ASR model, namely Whisper (Radford *et al.*, 2023), on a small-size specialized corpus of sport commentaries. Our quantative and qualitative analyses indicate that Whisper is capable of learning the features of our specialized corpus, reaching equal or higher to human transcribers performance, with a limited quantity of data. This research highlights the role that artificial intelligence, particularly large language models, can play in facilitating the creation of specialized corpora.

MOTS-CLÉS : Whisper ; large modèle de langage ; fine-tuning ; corpus de spécialité ; reconnaissance automatique de la parole.

KEYWORDS: Whisper ; large language model ; fine-tuning, specialized corpus ; automatic speech recognition.

1 Introduction

La recherche sur les corpus oraux en France a connu un développement plus lent que celle des corpus écrits. Malgré la présence de quelques corpus oraux dès les années 70-80, comme ESLO Baude & Dugua (2016), leur exploitation a été freinée par des difficultés de transcription, d'annotation et d'automatisation de l'analyse, en raison des particularités de la langue parlée. Les corpus de langue parlée, bien qu'ils ne soient pas nécessairement des corpus de référence, soulignent la diversité des situations de parole, des locuteurs et des dialectes d'une langue. Les exigences nécessaires à l'exploitation des corpus oraux ont freiné leur développement. Par conséquent, bien que ces corpus permettent l'analyse de phénomènes linguistiques particuliers, leur utilisation pour l'étude de genres de discours spécifiques reste souvent limitée en raison de la difficulté à constituer des corpus de grande taille. De nos jours, malgré la présence de nombreux corpus oraux, les corpus de discours de spécialité restent rares.

Cet article se concentre sur le commentaire sportif en direct, un genre de discours médiatique et journalistique présentant des caractéristiques linguistiques et prosodiques particulières. Ce genre reste sous-étudié, souvent analysé sur de courts extraits et peu diversifiés en termes de sports (le plus souvent des sports d'équipes et de ballon de type football, basket, rugby). Pourtant, il présente des particularités linguistiques tout à fait intéressantes pour différents domaines de la linguistique (p.ex. syntaxe, lexicologie, prosodie ; voir Augendre *et al.* 2018; Fontagnol *et al.* 2023) et du TAL. En effet, la transcription automatique de ce type de production de parole pose un grand nombre de difficultés pour les modèles de reconnaissance automatique de la parole (ASR [1]) à cause du bruit de fond important, l'utilisation du lexique spécifique au sport ou un grand nombre d'entités nommées. Dans le même temps, les études qui s'intéressent à ce genre de discours, si elles sont fondées sur corpus, le sont très souvent sur des corpus de petite taille, car très longs à transcrire, et qui ne font pas l'objet d'une diffusion large auprès de la communauté. De ce fait, un moyen d'aligner et de transcrire automatiquement les données orales ou multimodales fournies par ce genre de corpus permettrait très certainement de développer les données prêtes pour l'analyse et de faciliter la description de ce discours de spécialité, tout en analysant un plus grand nombre de données. En effet, pour le moment la transcription de notre corpus est basé sur un système de *student sourcing* (Stasica *et al.*, 2023) : une cinquantaine d'étudiants transcrit chaque année quelques minutes d'un enregistrement de notre corpus et au vu du temps de transcription, de correction et de vérification par les chercheurs, seulement un match maximum est transcrit par année universitaire.

Nous proposons dans cet article de discuter de l'apport que constituent les ASR pour le traitement de corpus oraux de spécialité. Plus spécifiquement, nous détaillons une première évaluation de la performance de la famille de modèles Whisper (Radford *et al.*, 2023) sur la transcription orthographique d'un corpus multimodal de petite taille (9h30 heures d'enregistrement) de commentaires télévisuels en direct de matchs de rugby. Nous étudions également l'apport de modèles *fine-tuned*(FM [2]) sur notre corpus, permettant de combler les limites des modèles pré-entraînés (PM [3]) sur une des particularités de notre corpus : le lexique de spécialité. Nous explorons ainsi la possibilité qu'après un *fine-tuning* sur un petit ensemble de données, ces modèles atteignent des performances comparables à celles des transcripteurs humains.

Cet article est structuré comme suit : dans la section 2 nous décrivons l'état de l'art des modèles

1. Automatic Speech Recognition
2. *fine-tuned* Models
3. *Pre-trained Models*

d'ASR. Dans la section 3, nous décrivons notre méthodologie et notre corpus. Dans la section 4, nous présentons les résultats de notre recherche et enfin dans la section 5, nous discutons des perspectives possibles pour de futures recherches.

2 Etat de l'art

Depuis l'introduction du premier ASR il y a plusieurs décennies (Davis *et al.*, 1952), de nombreuses méthodologies ont été élaborées pour améliorer la précision de ces modèles, convertissant les ondes sonores en impulsions électriques. Plusieurs revues de la littérature existantes ont été publiées pour examiner l'évolution des différentes approches utilisées dans le développement des ASR (Ghai & Singh, 2012; Karpagavalli & Chandra, 2016). Pendant longtemps, le modèle de Markov caché (HMM [4]) associé au modèle de mélange gaussien (GMM [5]) est resté dominant et le plus performant dans le domaine de la reconnaissance vocale. Cependant, l'avènement des techniques du *deep learning* a ouvert la voie à des modèles surpassant les performances des HMM-GMM, notamment les modèles HMM-*Deep Neural Network* (HMM-DNN) et les modèles *end-to-end* (Wang *et al.*, 2019).

Les récents modèles d'ASR ont abandonné l'utilisation des HMM, mais ont continué à exploiter les techniques de *deep learning*. Parmi les modèles ayant révolutionné la reconnaissance vocale, on trouve le modèle Wav2Vec (Schneider *et al.*, 2019). Il s'agit du premier modèle à adopter une approche de pré-entraînement non supervisé, apprenant à partir de données audio non transcrites. Ce pré-entraînement permet à Wav2Vec d'apprendre des représentations générales de la parole, exploitables pour améliorer les performances sur des tâches ultérieures où les données annotées sont limitées. C'est le cas pour des tâches telles que la reconnaissance vocale, où l'annotation de données est fastidieuse et rare.

L'architecture de Wav2Vec se compose d'un réseau neuronal convolutif (CNN) suivi d'un réseau de neurones basé sur des *transformers*. Le CNN encode les formes d'ondes audio brutes pour extraire des représentations de haut niveau, puis les transmet à l'encodeur du *transformer* pour un traitement ultérieur, permettant la capture d'informations contextuelles.

Après le pré-entraînement, Wav2Vec peut être *fine-tuned* sur des données annotées spécifiques à des tâches d'ASR. Pendant cette phase de *fine-tuning*, le modèle apprend de manière supervisée, adaptant les représentations pré-entraînées aux caractéristiques spécifiques des données cibles, ce qui améliore les performances sur les tâches de reconnaissance vocale. L'utilisation d'un pré-entraînement non supervisé permet l'exploitation d'une plus grande quantité de données, car il n'est pas nécessaire de compléter les données sources avec des transcriptions. Wav2Vec maintient actuellement des performances de pointe sur divers ensembles de tests ASR, même lorsqu'il est formé sur des données annotées limitées.

Cependant, les encodeurs pré-entraînés entièrement non supervisés, tels que Wav2vec, présentent des limites. Selon Radford *et al.* (2023), bien qu'ils apprennent des représentations vocales de haute qualité, ils souffrent d'un manque de décodeur aussi performant pour associer ces représentations à des sorties utilisables. Cette lacune nécessite une étape de *fine-tuning* pour des tâches telles que la reconnaissance vocale, ce qui peut être complexe et limiter leur utilité.

Les auteurs notent également que des études ont démontré que les ASR pré-entraînés sur plusieurs en-

4. Hidden Markov Model
5. Gaussian Mixture Model

sembles de données ou domaines sont plus robustes et généralisent plus efficacement. Par conséquent, des recherches ont mis en avant l'utilisation de données audio avec des transcriptions de référence où l'exigence de validation manuelle a été assouplie, permettant l'utilisation d'un volume accru de données (p.ex. de 10 à 30 000 heures d'audio) pour l'entraînement des ASR. Cela permet d'établir un équilibre entre qualité et quantité de données.

Ainsi, Radford *et al.* (2023) proposent les modèles Whisper pour pallier les problèmes cités ci-dessus, étendant la reconnaissance vocale faiblement supervisée à un vaste ensemble de données comprenant 680,000 heures de données audio transcrites. Ces modèles incluent un entraînement multilingue et multitâche, et produisent des résultats comparables aux modèles de pointe actuellement disponibles. Les auteurs choisissent une architecture *transformer* d'encodeur-décodeur (Vaswani *et al.*, 2017), car cette architecture a montré son efficacité pour la génération de texte.

Ainsi, Whisper se distingue des modèles précédents. Contrairement à ses prédécesseurs, en tant que modèle *end-to-end*, Whisper ne requiert pas de *fine-tuning* avant son application à un ensemble de données, ce qui lui confère une robustesse surpassant les autres modèles, même si, ses performances sont plus faibles sur des données moins représentées lors de son entraînement (p.ex. Jain *et al.* (2023a)). En effet, ayant été entraîné sur une grande variété de données audio et évalué dans un cadre *zero-shot*, c'est-à-dire sur des ensembles de données différents de ceux de l'entraînement, il est capable de généraliser à travers divers domaines et tâches. Les auteurs visent ainsi à développer un système de traitement vocal fonctionnel adaptable à différents domaines, tâches et langues, sans nécessiter de réglages supplémentaires, ce qui le différencie des modèles antérieurs, souvent limités dans leur capacité à faire des généralisations. En effet, même de légères divergences entre les ensembles d'entraînement et de test altèrent significativement les performances des modèles précédents (Radford *et al.*, 2023).

Malgré les atouts de Whisper par rapport aux autres modèles, il convient de souligner que, dans le cadre de notre recherche sur les corpus de spécialité, certaines fonctionnalités requises ne sont pas intégrées dans les modèles Whisper de base. En particulier, ces modèles ne sont pas capables de diariser, c'est-à-dire de reconnaître les différents locuteurs, et leur alignement au signal sonore est mal adapté, fonctionnant davantage comme un générateur de sous-titres, pas assez précis pour notre objectif de recherche. Malgré la robustesse du modèle, Radford *et al.* (2023) admettent la nécessité d'étudier le *fine-tuning* de Whisper pour améliorer ses performances dans des contextes spécifiques. Par exemple, des adaptations ont été nécessaires pour décrire la parole enfantine (Jain *et al.*, 2023a), pour des langues peu représentées dans les données d'entraînement (comme le roumain (Păis *et al.*, 2023) ou le turc (Oyucu, 2023)), ainsi que pour des langues minoritaires ou mixtes (Xie *et al.*, 2023).

3 Corpus et méthodologie

3.1 Données

3.1.1 Données brutes

Notre corpus est un corpus oral de spécialité consistant en plusieurs matchs de rugby de deux coupes du monde différentes : celle de 2007 (Lortal & Mathon, 2008) et celle de 2015. Pour la première nous avons 36 heures d'enregistrement audio pour certains matchs et vidéos pour d'autres et pour la seconde nous possédons 25 heures d'enregistrement vidéo. Ce corpus est considéré de spécialité

de par ces nombreuses particularités linguistiques, notamment lexicales (Fontagnol *et al.*, 2023) et syntaxiques (Augendre *et al.*, 2018) avec une présence moindre de structures verbales et une proéminence de structures nominales spécifiques (p.ex. entités nommées) et de lexique spécifique au rugby et au commentaire sportif.

L'omniprésence de ce lexique spécifique dans notre corpus pourrait potentiellement affecter les performances des modèles d'ASR. C'est pourquoi le *fine-tuning* serait nécessaire pour permettre au modèle d'acquérir ces spécificités lexicales et ainsi augmenter sa performance jusqu'à atteindre celle d'un transcripteur humain.

La production des transcriptions se fait en 2 phases. La première phase repose sur le *student sourcing* pour obtenir une première transcription de l'audio, alignement compris. Ces transcriptions sont ensuite validées par un transcripteur expert (ie. enseignant chercheur ou doctorant). Toutes les transcriptions sont réalisées sur Transcriber (Barras *et al.*, 2001), et l'annotation d'actions de jeu sur les images des vidéos de deux matchs sont réalisées avec Aegissub [6]. Néanmoins, ces annotations ne sont pas incluses dans la présente étude.

Tous les matchs ne sont pas totalement transcrits et annotés. Pour cette raison, nous n'utilisons comme corpus de travail pour cette recherche que six matchs pour lesquels nous possédons une transcription «gold», c'est-à-dire vérifiée par un transcripteur expert. Les six matchs cumulent 9h30 d'audio. Ces 9h30 d'audio transcrits contiennent 94514 mots soit 138751 tokens (mots et la ponctuation) après tokenization. Pour chaque match, la transcription est orthographique, alignée au signal sonore, notant les pauses silencieuses d'au moins 200 ms comme requis dans l'état de l'art (Candea, 2000) et distingue les différents locuteurs présents dans l'audio.

Le tableau 1 (Annexe A) présente les deux commentateurs principaux, un journaliste et un expert commentant chaque match, le nom de chaque match du corpus de travail ainsi que l'année où il a été joué.

3.1.2 Pré-traitement des données

Le corpus de travail est préalablement divisé en un ensemble d'entraînement/validation et un ensemble de test. Pour les matchs utilisés dans l'ensemble d'entraînement, un échantillon sur six a été réservé pour la validation soit 1h20 d'audio (14% du corpus de travail). Les échantillons restant, 6h30 d'audio (68% du corpus de travail) constituent l'ensemble d'entraînement. Pour s'assurer de la capacité de généralisation après *fine-tuning*, nous réservons un match entier à l'ensemble de test soit 1h40 d'audio (18% du corpus de travail). Ce match, «Japon-Fidji», est donc *out-of-sample*, c'est-à-dire qu'aucun passage de ce match n'est donné au modèle durant l'entraînement. De plus, les commentateurs de ce match ne sont pas représentés dans l'ensemble d'entraînement. Pour chaque match l'audio et la transcription correspondante ont été découpés en tranches de 30 secondes, soit la taille maximale de la fenêtre de contexte de Whisper. Le découpage est fait sur les unités inter-pausales, définies par Nguyen *et al.* (2022) comme une détection d'activité vocale continue par un locuteur, délimitée par un silence de plus de 200 ms des deux côtés.

Conformément à la méthode utilisée par les auteurs de Whisper pour leur propre corpus, tous les enregistrements audio ont été rééchantillonnés à 16 000 Hz, et une représentation du Mel-spectrogramme en magnitude logarithmique à 80 canaux a été calculée sur des fenêtres de 25 millisecondes avec un pas de 10 millisecondes (Radford *et al.*, 2023).

6. https://aegisub.org/

3.2 Fine-tuning de Whisper

Comme mentionné précédemment, en raison de la fréquence élevée d'entités nommées et du lexique spécifique au rugby, les PM présentent généralement des performances limitées sur notre corpus. Ainsi, un processus de *fine-tuning* sur nos propres données s'avère nécessaire pour améliorer la reconnaissance de ce lexique spécifique. De plus, Whisper, comme tout modèle d'ASR, est sensible au bruit. Il est probable que les puissants bruits de fond particuliers à l'environnement des grands évènements sportifs dégradent la qualité de la transcription, dégradation que pourrait compenser un fine-tuning du modèle. Différentes méthodes de *fine-tuning* ont été proposées ces dernières années, parmi lesquelles le *Low-Rank Adaptation* (LoRA) (Hu *et al.*, 2021), que nous avons retenu pour sa sobriété computationnelle et sa facilité d'implémentation. L'utilisation de cette méthode nous a permis d'effectuer le *fine-tuning* de la famille de modèles Whisper du Tiny (33 millions de paramètres) au Large (1600 millions de paramètres) avec une unique NVDIA RTX 6000 24GB VRAM [7]. La *quantization* (Li *et al.*, 2023) couplée à l'utilisation d'un float half-précision (float 16bit) dégradaient fortement les performances des modèles lors de l'inférence. Nous avons donc entraîné nos modèles en full précision (float 32bit).

Concernant les hyper-paramètres de LoRA, le rang r=1, le *dropout* des matrices de poids a été fixé à 0.3 et alpha=64. Ce paramétrage permet de réduire l'*overfitting*, très problématique sur les modèles les plus larges. Le *learning rate* est fixé à 1e-3 avec *warm-up*. Le nombre d'epochs par défaut est fixé à 15 avec *early stopping*. L'entraînement d'un modèle sur les 6h30 heures d'audio des données d'entraînement prend entre 1h et 3h selon la taille du modèle.

3.3 Évaluation

L'enjeu du *fine-tuning* revêt une double importance : améliorer la qualité globale de la transcription (p.ex. transcription plus précise malgré le bruit de fond, meilleure performance en français), et affiner la transcription des éléments spécifiques à notre corpus, notamment le lexique propre aux commentaires sportifs pour les matchs de rugby.

Pour évaluer les améliorations apportées par nos modèles *fine-tuned*, nous avons adopté une approche double, à la fois quantitative et qualitative. Tout d'abord, l'évaluation quantitative suit le même protocole que celui utilisé dans les publications majeures en ASR ces dernières années, à savoir le calcul de la *Word Error Rate* (WER) entre la référence et une prédiction. Cette métrique présente néanmoins un défaut majeur dans notre cas. Elle pénalise de manière disproportionnée les hallucinations de Whisper [8], qui sont rares, facilement détectables et corrigeables en remplaçant la génération *greedy search* par du *sampling* avec *beam search* et température *schedduling* (Radford *et al.*, 2023). Puisque les transcriptions ayant une WER supérieure à 100 sont systématiquement des hallucinations, pour chaque segment de transcription évalué, nous définissons une borne supérieure telle que : WER $\in$ [0, 100].

Ensuite, nous avons procédé à des évaluations qualitatives afin de déterminer (i) si les performances des FM augmentent par rapport aux PM sur les éléments spécifiques à notre corpus de spécialité, et (ii) pour établir un lien entre son éventuelle amélioration quantitative et une meilleure compréhension

7. Les expériences présentées dans cet article ont été réalisées par l'intermédiaire de Gilles Boyé et de Catherine Mathon en utilisant la plateforme OSIRIM qui est administrée par l'IRIT et soutenue par CNRS, la région Midi-Pyrénées, le gouvernement français, et le FEDER (voir http://osirim.irit.fr/site/fr).

8. Un exemple d'hallucination qui peut être trouvé dans notre corpus : "alors on du mal au au au au au au au au au"

et transcription des caractéristiques de notre corpus. Plusieurs éléments peuvent être représentatifs de notre corpus, tels que les noms propres (p.ex. noms des joueurs, de l'arbitre, des commentateurs), dans la mesure où nous manquons de moyens efficaces pour permettre au modèle d'apprendre des noms propres qu'il n'aurait pas rencontrés lors de l'entraînement, nous nous concentrons sur le lexique spécialisé du rugby dans notre analyse qualitative.

Pour l'évaluation qualitative de la reconnaissance du lexique spécialisé, nous avons utilisé AnaText [9] afin de sélectionner dans notre corpus d'entraînement 20 lemmes qui sont peu fréquents dans un corpus de référence standard, mais qui présentent une spécificité élevée dans notre corpus. Nous étudions également la transcription de la disfluence "euh", de l'interjection "hein" et de la répétition d'amorces, qui sont fréquentes dans tout discours oral mais non transcrits par Whisper. Nous avons exclu de notre analyse les emprunts à l'anglais (p.ex. «drop») afin d'éviter que le large pré-entraînement dont Whisper a bénéficié sur l'anglais, n'influe sur nos analyses : si ces lemmes sont rares en français et spécifiques à notre corpus, ils sont possiblement plus fréquents en anglais. Nous avons également exclu les lemmes spécifiques à certains matchs, tels que la nationalité des joueurs (p.ex. roumaine, argentine), ainsi que des termes plus généraux (p.ex. «introduction»). Bien que ces termes puissent être analysés dans une étude future afin de graduer les performances de FM sur différents types d'éléments lexicaux.

Le premier tableau (Tableau 2) de l'annexe 6.2 résume les lemmes utilisés, leur fréquence dans notre corpus, leur fréquence dans le corpus de référence, et leur spécificité (LogLike). Nous avons ensuite vérifié que tous ces lemmes apparaissent dans la transcription du match utilisée pour le test, et les lemmes ainsi que leur fréquence et leur spécificité sont présentés dans le deuxième tableau (Tableau 3) de l'annexe 6.2. Enfin, nous avons examiné, pour chaque taille de modèles (PM et FM), la différence d'occurrence de ces lemmes avec la transcription de référence.

4 Résultats

4.1 Quantitatif

Les valeurs de WER sont reportées pour chacun des modèles Whisper par ordre croissant de nombre de paramètres en Figure 1. Pour tous les modèles, quel que soit leur taille, le *fine-tuning* avec la méthode LoRA a permis de diminuer significativement la métrique WER sur l'ensemble de test.

On remarque que plus le modèle est large, plus l'écart de performance entre le modèle *Pre-trained* et sa version *Fine-tuned* est important. Les modèles les plus larges sont donc ceux qui ont le plus bénéficié du fine-tuning. Par exemple le modèle Tiny a vu sa WER diminuer de 14% tandis que cette diminution est de 54% pour whisper Large-v2. Pour évaluer la qualité de la transcription d'un ASR, il est également pertinent de la comparer avec les performances humaines sur cette même tâche. La transcription de notre corpus reposant en partie sur du *student sourcing*, il nous est possible de calculer la WER entre les transcriptions des étudiants et ces mêmes transcriptions après correction par un expert (étudiants vs gold sur la Figure 1). Il est notable que les modèles medium et large pré-entraînés obtiennent des performances similaires à celles des étudiants, confirmant une fois de plus que Whisper produit des transcriptions décentes en *zero-shot learning*. L'écart se creuse après *fine-tuning*, Whisper Small dépassant déjà largement les étudiants en terme de WER. D'un point de vue quantitatif, il est certain que le *fine-tuning* est un excellent choix pour qui dispose de la puissance

9. http://phraseotext.univ-grenoble-alpes.fr/anaText/index.php

de calcul nécessaire. Néanmoins on ne peut déduire de ce simple chiffre ce qui a été amélioré par le *fine-tuning* dans la transcription.

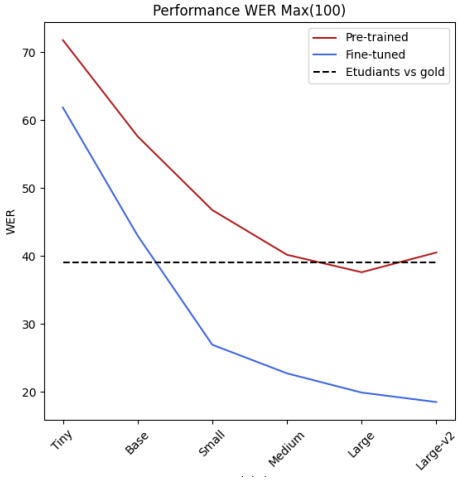

FIGURE 1 – Performance des modèles Whisper *Pre-trained* et *Fine-tuned* par ordre croissant de taille

4.2 Qualitatif

Les résultats de l'analyse qualitative sont présentés en Figure 2. Pour chaque PM et FM nous avons calculé l'erreur absolue moyenne (MAE [10]) du nombre total d'occurrences entre la transcription prédite et le gold pour 20 lemmes de notre ensemble de test appartenant au vocabulaire spécifique du sport. Les résultats montrent deux choses : (i) plus le modèle est large, plus la MAE est faible [11], (ii) les FM ont une MAE systématiquement plus faible que les PM, preuve que le modèle, avec seulement 6h30 d'audio et de transcriptions est capable d'apprendre du lexique spécifique.

Le détail des lemmes pour chaque modèle est exposé en annexe 6.3 (Tableau 4). Certains lemmes, tels que «pénalité», «essai» et «touche», sont déjà bien reconnus, même par les modèles de petites tailles PM et FM, suggérant leur présence fréquente dans l'ensemble d'apprentissage du pré-entraînement. Le Tiny-FM présente dans sa transcription une hallucination faisant apparaître ces lemmes plus de fois que dans la transcription de référence, comme observé avec «essai» qui apparaît 96 fois dans les transcriptions du Tiny-FM contre seulement 67 fois dans le gold.

D'autres lemmes, tels que «mêlée», voient leur reconnaissance s'améliorer avec l'augmentation de la taille du modèle et bénéficient d'une meilleure détection suite au *fine-tuning*. Les PM éprouvent des difficultés à détecter certains lemmes spécifiques au rugby, tels que «chandelle» et «rebond», détectés moitié de fois moins que par les FM en faisant la somme du total de détection pour ces lemmes pour toutes les tailles de PM comparés à la somme de leur détection par toutes les tailles de FM. Les deux lemmes comprenant un tiret («en-avant», «en-but») ne sont pas reconnus par les PM, exception faite du modèle Base-PM qui identifie presque toutes les occurrences de «en-but». Les tentatives de

10. Mean Absolute Error
11. Excepté pour le modèle large, faisant plus d'erreur que le modèle medium

recherche de ces lemmes sans tiret ou avec seulement une partie du lemme (avant, but) n'ont pas donné de résultats concluants non plus, les modèles ne parvenant pas à les détecter.

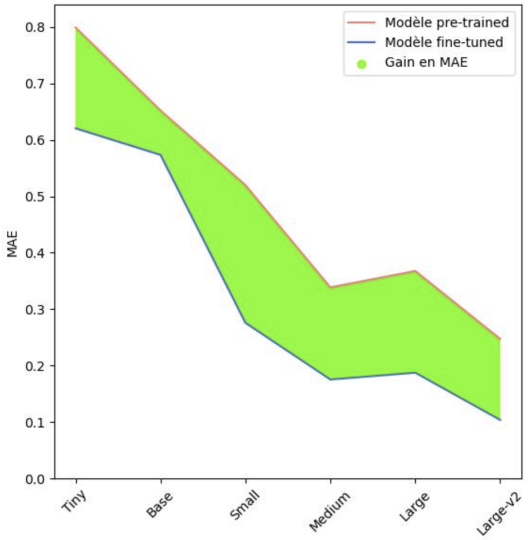

FIGURE 2 – MAE du nombre d'occurrences des PM et FM par rapport à la référence

Certaines disfluences et interjections apparaissent dans notre corpus. Elles présentent l'intérêt de ne pas être spécifiques à un corpus de spécialité, car présentes dans tout type de discours oral et plutôt absentes de notre type de corpus, contrairement au lexique, tout en étant absentes des transcriptions des modèles PM. On constate une détection de 0% sur toute la famille de modèles PM. Après *fine-tuning*, Whisper reconnaît entre 38% (base) à 92% (large-v2) la disfluence "euh" et l'interjection "hein", les plus courantes. D'autres types de disfluences, comme les amorces, ne sont jamais identifiés par Whisper. Cette lacune peut être attribuée à notre convention de transcription (ICOR, 2013), mais aussi au fait que Whisper n'a jamais été exposé à ces disfluences lors de son entraînement. En effet, Whisper est principalement basé sur du sous-titrage de films/séries, où la spontanéité du langage peut être discutée, et où l'annotation des disfluences n'est ni nécessaire ni souhaitée.

5 Conclusion et Perspectives

La présente étude vise à évaluer l'impact du *fine-tuning* des différentes tailles de modèles de Whisper sur la transcription de corpus spécialisés, en particulier dans le contexte des commentaires sportifs. Les résultats obtenus, tant qualitatifs que quantitatifs, indiquent une corrélation positive entre la performance du modèle et sa taille, soulignant ainsi l'efficacité accrue des modèles plus larges. Par ailleurs, plus les modèles sont larges plus l'écart entre leur performance et celle de leur équivalent PM est important. Cependant, en raison de la taille limitée de notre corpus et du nombre de modèles, il est difficile de tirer des conclusions définitives, et des recherches supplémentaires avec plus de données sont nécessaires pour confirmer ou infirmer cette corrélation taille du modèle /gain après *fine-tuning*. De plus, il serait intéressant de mettre en relation la taille des modèles avec la quantité de données annotées nécessaire pour atteindre un certain niveau de performance. Jain *et al.* (2023b) ont

amorcé une telle démarche mais de façon limitée.

Ensuite, alors même qu'une analyse lexicale a été effectuée dans cette étude, la spécificité de notre corpus est multidimensionnelle, justifiant ainsi la réalisation d'autres analyses pour évaluer l'impact, par exemple, des accents des commentateurs, de la syntaxe particulière des commentaires sportifs ou du bruit de fond sur la performance des ASR. Dans ce contexte, il convient d'explorer dans quelle mesure le *fine-tuning* de modèles tels que Whisper peut renforcer leur robustesse face à ces défis en les comparant aux PM.

En outre, pour mener des analyses grammaticales, l'utilisation d'analyseurs syntaxiques est courante, mais ces outils nécessitent de la ponctuation, que Whisper génère. Des études sont donc nécessaires pour évaluer la capacité de Whisper à insérer une ponctuation cohérente aux endroits appropriés.

Enfin, des recherches supplémentaires pourraient être entreprises pour mieux appréhender les types de périodes discursives les plus difficiles à traiter pour les PM et pour évaluer le potentiel des FM à surmonter ces difficultés. Dans le contexte des commentaires sportifs, qui se divisent généralement en deux types, à savoir les *colour commentaries* (Hartmann, 2013) et les narrations d'action en temps réel (*play-by-play*), une analyse séquentielle pourrait permettre de déterminer si les modèles performent mieux sur l'un ou l'autre type, et si le *fine-tuning* améliore leur performance dans ces deux catégories.

Cette étude s'est principalement concentrée sur l'amélioration de la transcription à partir d'un large modèle de langage affiné, visant à réduire le besoin de corrections humaines par rapport à une transcription manuelle ou avec un PM. Toutefois, pour optimiser la création de corpus spécialisés, notamment dans le cadre des études en linguistique de l'oral, et particulièrement dans notre contexte où des analyses prosodiques sont réalisées, il est essentiel de garantir la précision de l'alignement des unités inter-pausales. Malgré les capacités de Whisper dans cette tâche, son niveau de précision reste comparable à celui d'un générateur de sous-titres, ce qui n'est pas suffisant pour nos besoins de recherche. De plus, Whisper ne prend pas en charge la diarisation, c'est-à-dire la distinction entre les différents locuteurs.

Une précédente étude (Stasica *et al.*, 2023) a souligné la similarité en termes de temps nécessaire entre la transcription, l'alignement et la diarisation entièrement manuelles des commentaires sportifs, et le temps requis pour corriger la transcription de Whisper *pre-trained*, réaligner et diariser. Les résultats ont montré que la majeure partie du temps n'est pas consacrée à la correction de la transcription, mais à l'alignement et à la diarisation.

Actuellement, nos recherches se concentrent sur l'automatisation de l'alignement et de la diarisation afin d'optimiser la création de nos corpus. Whisper dispose déjà de *timestamps* sous forme de *token* spéciaux ajoutés durant la génération de texte. Néanmoins ceux-ci souffrent d'une précision instable, due à la nature de l'ensemble d'entraînement faiblement supervisé. Pour remédier à ce problème, des solutions ont été proposées offrant une précision à plus ou moins 100ms, suffisant pour notre usage, notamment le *Dynamic Time Warping* (Giorgino, 2009).

Aucune solution directe n'existe pour la diarisation avec Whisper. La solution la plus couramment utilisée est d'ajouter un second modèle indépendant de Whisper pour effectuer la diarisation puis fusionner diarisation / transcription / alignement, comme proposé par la librairie `pyannote` (Bredin *et al.*, 2020).

Au vue de la difficulté que pose ce triple problème, nous travaillons en parallèle à la création d'un modèle généraliste qui rendrait l'implémentation plus simple tout en limitant le risque de propagation d'erreurs entre les différentes étapes.

Références

AUGENDRE S., KUPŚĆ A., BOYÉ G. & MATHON C. (2018). Live TV sports commentaries : specific syntactic structures and general constraints. In D. LEGALLOIS, T. CHARNOIS & M. LARJAVAARA, Éds., *The Grammar of Genres and Styles : From Discrete to Non-Discrete Units*, p. 194–218. De Gruyter Mouton.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1-2), 5–22.

BAUDE O. & DUGUA C. (2016). Les ESLO, du portrait sonore au paysage digital. *Corpus*, **15**.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). Pyannote. audio : neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7124–7128 : IEEE.

CANDEA M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. *Etude sur un corpus de récits en classe de français*.

DAVIS K. H., BIDDULPH R. & BALASHEK S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, **24**(6), 637–642.

FONTAGNOL C., HANOTE S. & MATHON C. (2023). Sélection lexicale et réalisations prosodiques : impact des contraintes d'un genre discursif spécifique, le commentaire sportif télévisuel en direct. *Lexique et frontières de genres*, p. 97–116.

GHAI W. & SINGH N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications*, **41**(8).

GIORGINO T. (2009). Computing and visualizing dynamic time warping alignments in R : the dtw package. *Journal of statistical Software*, **31**, 1–24.

HARTMANN C. (2013). *Pre-fabricated speech formulas as long-term memory solutions to working memory overload in routine language*. Thèse de doctorat, University of Zurich.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). Lora : Low-rank adaptation of large language models. *arXiv preprint arXiv :2106.09685*.

ICOR (2013). Convention ICOR. *Lyon : université de Lyon. URL : http ://icar. cnrs. fr/projets/corinte/documents/2013_Conv_ICOR_250313. pdf*.

JAIN R., BARCOVSCHI A., YIWERE M., CORCORAN P. & CUCU H. (2023a). Adaptation of Whisper models to child speech recognition. *arXiv preprint arXiv :2307.13008*.

JAIN S., KIRK R., LUBANA E. S., DICK R. P., TANAKA H., GREFENSTETTE E., ROCKTÄSCHEL T. & KRUEGER D. S. (2023b). Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv :2311.12786*.

KARPAGAVALLI S. & CHANDRA E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, **9**(4), 393–404.

LI Y., YU Y., LIANG C., HE P., KARAMPATZIAKIS N., CHEN W. & ZHAO T. (2023). Loftq : Lora-fine-tuning-aware quantization for large language models. *arXiv preprint arXiv :2310.08659*.

LORTAL G. & MATHON C. (2008). Motion and Emotion or how to align emotional cues with game actions. In *WORKSHOP PROGRAMME &*, p.̃79.

NGUYEN T. A., KHARITONOV E., COPET J., ADI Y., HSU W.-N., ELKAHKY A., TOMASELLO P., ALGAYRES R., SAGOT B., MOHAMED A. & DUPOUX E. (2022). Generative spoken dialogue language modeling.

OYUCU S. (2023). Comparing the Fine-Tuning and Performance of Whisper Pre-Trained Models for Turkish Speech Recognition Task. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, p. 1–4 : IEEE.

PĂIS V., MITITELU V. B., ION R. & IRIMIA E. (2023). Evaluating a Fine-Tuned Whisper Model on Underrepresented Romanian Speech. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, p. 141–145 : IEEE.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, p. 28492–28518 : PMLR.

SCHNEIDER S., BAEVSKI A., COLLOBERT R. & AULI M. (2019). wav2vec : Unsupervised pre-training for speech recognition. *arXiv preprint arXiv :1904.05862*.

STASICA A., BOYÉ G., KUPŚĆ A. & MATHON C. (2023). Chuchoter à l'oreille des corpus : Whisper, pour augmenter la production de corpus oraux de spécialité. COSEDI.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WANG D., WANG X. & LV S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry*, **11**(8), 1018.

XIE P., LIU X., CHEN Z., CHEN K. & WANG Y. (2023). Whisper-MCE : Whisper Model Finetuned for Better Performance with Mixed Languages. *arXiv preprint arXiv :2310.17953*.

6 Annexes

6.1 Annexe A

TABLE 1 – Noms des matchs, des experts et des journalistes et année des match

Match	Année	Journaliste	Expert
France/Argentine	2007	Thierry Gilardi	Thierry Lacroix
France/Canada	2015	Christian Jeanpierre	Bernard Laporte
France/Roumanie	2015	Christian Jeanpierre	Bernard Laporte
France/Nouvelle-Zélande	2015	Christian Jeanpierre	Bernard Laporte
Afrique du Sud/Argentine	2015	Christian Jeanpierre	Bernard Laporte
Japon/Fidji	2007	Nicolas Delage	Jérôme Papin

6.2 Annexe B : détail lexique spécifique

TABLE 2 – Fréquence et degré de spécificité de vingt lemmes dans la transcription gold des matchs servant à l'entraînement de Whisper classé par Anatext

Lemme	Fréquence	CorpusRef (par million)	LogLike (spécificité)
en-avant	126	0.035	2684.069
pénalité	132	1.09	1916.386
essai	208	20.895	1601.322
touche	184	16.28	1467.165
mêlée	121	4.875	1185.473
plaquage	45	0.4	646.637
talonneur	29	0.035	532.528
relance	29	0.595	341.643
pénaliser	26	0.47	319.679
buteur	20	0.15	294.184
pilier	44	10.305	260.996
hors-jeu	24	1.47	211.049
envoi	32	6.1	203.258
sélectionneur	12	0.035	199.177
rebond	18	0.91	166.349
plaqueur	10	0.035	162.334
en-but	11	0.12	153.557
chandelle	27	8.725	142.668
renvoi	20	4.36	121.545
ailier	14	1.065	116.263

TABLE 3 – Fréquence et degré de spécificité de vingt lemmes dans la transcription gold du match servant au test du *fine-tuning* de Whisper classé par Anatext

Lemme	**Fréquence**	**CorpusRef (par million)**	**LogLike (spécificité)**
pénalité	41	1.09	575.298
essai	67	20.895	564.366
touche	61	16.28	533.237
mêlée	30	4.875	293.029
en-avant	13	0.035	258.228
ailier	18	1.065	215.516
plaquage	12	0.4	160.688
hors-jeu	11	1.47	111.966
buteur	7	0.15	103.093
talonneur	5	0.035	89.761
en-but	5	0.12	71.732
pénaliser	6	0.47	68.017
pilier	9	10.305	52.310
plaqueur	3	0.035	50.791
renvoi	7	4.36	49.157
relance	4	0.595	39.808
sélectionneur	2	0.035	31.103
rebond	3	0.91	25.429
envoi	4	6.1	20.997
chandelle	4	8.725	18.216

6.3 Annexe C : Détails des résultats de l'analyse qualitative

TABLE 4 – Nombre d'occurrences des lemmes spécifiques pour chaque taille de modèle Whisper *Fine-tuned* (FM) et Pre-trained (PM)

	Gold	Tiny		Base		Small		Medium		Large		Largev2	
		PM	FM	PM	FM	PM	FM	PM	FM	PM	FM	PM	FM
pénalité	41	26	***27***	33	***36***	39	***39***	40	***36***	39	***41***	39	***40***
essai	67	14	***96***	38	***33***	42	***58***	55	***58***	57	***62***	57	***63***
touche	61	30	***75***	41	***69***	49	***50***	61	***48***	62	***62***	63	***63***
mêlée	30	0	***5***	5	***13***	7	***21***	12	***19***	15	***27***	20	***23***
en-avant	13	0	***0***	0	***3***	0	***6***	0	***4***	0	***7***	0	***8***
ailier	18	0	***0***	0	***0***	0	***0***	4	***16***	0	***9***	7	***13***
plaquage	12	0	***0***	0	***2***	0	***8***	8	***9***	5	***10***	7	***9***
hors-jeu	11	0	***0***	0	***3***	2	***0***	3	***9***	5	***5***	6	***7***
buteur	7	5	***5***	5	***4***	6	***7***	7	***7***	7	***6***	6	***7***
talonneur	5	1	***2***	1	***1***	2	***5***	5	***5***	4	***4***	5	***4***
en-but	5	0	***0***	4	***0***	0	***0***	0	***0***	0	***0***	0	***3***
pénaliser	6	3	***4***	6	***4***	5	***5***	4	***5***	6	***5***	6	***6***
pilier	9	4	***0***	6	***6***	9	***9***	9	***9***	8	***9***	7	***9***
plaqueur	3	0	***0***	0	***1***	0	***2***	2	***3***	3	***2***	3	***2***
renvoi	7	0	***0***	0	***4***	5	***7***	5	***7***	7	***7***	6	***7***
relance	4	3	***4***	5	***4***	4	***3***	4	***4***	4	***4***	4	***4***
sélectionneu	2	0	***2***	0	***0***	1	***2***	2	***2***	0	***2***	2	***2***
rebond	3	0	***0***	0	***0***	2	***3***	0	***3***	2	***3***	3	***3***
envoi	4	0	***0***	0	***3***	3	***4***	3	***3***	4	***4***	3	***5***
chandelle	4	0	***0***	0	***0***	0	***3***	4	***4***	0	***4***	3	***4***