Choix d'une infrastructure de transcription pour les archives orales : Analyse de l'état de l'art et implémentation de WhisperX

Louis-Fiacre Franchet d'Espèrey

Mémo sur le choix de la solution WhisperX pour la tâche de transcription automatique d'archives orales

Louis-Fiacre Franchet d'Espèrey

Octobre 2025

Introduction

Dans le cadre de la thèse en cours, *« Les trajectoires professionnelles décryptées : une analyse sémantique des récits de carrière à travers les archives orales »*, une étape de transcription des archives orales est nécessaire afin d'en extraire, traiter et analyser le contenu. Cette tâche de transcription, grâce aux avancées techniques dans le domaine de la reconnaissance automatique de la parole (Automatic Speech Recognition, ASR), peut être partiellement automatisée. Il est essentiel que la transcription soit la plus fidèle possible et qu'elle prenne en compte la séparation des locuteurs, appelée diarisation, ainsi que l'horodatage des tours de parole. La tâche d'ASR transcrit ce qui est dit et la tâche de diarisation identifie qui le dit.

Sont présentés ici les systèmes à l'état de l'art permettant de produire une transcription textuelle répondant à ces exigences.

Nous évoquerons une brève histoire des systèmes d'ASR et de diarisation(I). Puis nous présenterons les métriques reconnues pour analyser la qualité d'un modèle d'ASR, le classement des solutions d'ASR de la plateforme Hugging Face et le jeu de données utilisé pour réaliser ce classement (II). Nous analyserons le résultat du classement sur les 7 meilleurs modèles d'ASR pour le français. Nous expliquerons ensuite les métriques utilisées, le jeu de données et le classement des systèmes de diarisation (III). Enfin nous justifierons le choix du logiciel WhisperX comme solution de transcription pour les archives orales dans le cadre de la thèse (IV).

Table des matières

1 Brève histoire des systèmes d'ASR et de diarisation

La tâche de transcription automatique d'entretiens sonores fait partie du domaine de la reconnaissance automatique de la parole, qui a de nombreuses applications diverses telles que la traduction automatique d'audio, la dictée en temps réel ou la transcription de témoignages oraux. Le premier système de reconnaissance automatique de la parole, appelé Audrey est développé en 1952 dans les laboratoires BELL avec une capacité de reconnaissance de quelques chiffres isolés. D'autres systèmes basés sur des règles (rules-based) voient le jours augmentant la taille reconnue du vocabulaire de ces systèmes. En 1974, le système DRAGON, basé sur des chaînes cachées de Markov, précurseur du logiciel de dictée vocale Dragon Natural Speaking, est développé par Janet et James Baker. En 1976 est développé Harpy[1], un programme basé sur l'algorithme beam search, créé au sein de l'université Carnegie Mellon par Bruce Lowerre, qui dépasse le cap des 1000 mots reconnus. En 1987 sort SPHINX-I qui est un programme combinant les deux approches de HARPY et DRAGON, des chaînes cachées de Markov et le Beam Search. L'avantage de ce système était de ne pas avoir de phase de calibrage de la voix du locuteur en amont de la reconnaissance.

L'évolution technique des approches pour réaliser une reconnaissance automatique de la parole ainsi que celles pour la diarisation a suivi l'arrivée des technologies de deep learning. De 1990 à 2010, une approche majoritaire basée sur les chaînes de Markov cachées avec des modèles de mélange gaussien (HMM GMM en anglais Hidden Markov Models + Gaussian Mixture Models) a été utilisée. Cela a donné les solutions d'IBM ViaVoice, CMU. C'est en 1997 que sort commercialement Dragon Naturally-Speaking reconnaissant jusqu'à 100 mots à la minute. Puis, l'arrivée des modèles de réseaux profonds de neurones a permis le développement de modèles de bout en bout qui ont progressivement remplacé les approches hybrides telles que l'HMM-DNN et diverses approches basées sur les RNN, les LSTM, qui ont par exemple donné la solution Deep Speech [13]. Depuis 2017 et l'arrivée de l'architecture Transformers pour les réseaux de neurones profonds, de nombreuses solutions performantes sont apparues, telles que Whisper [17] d'OpenAI. Aujourd'hui, la prédominance des solutions basées sur les transformers laisse quand même apparaître de nouvelles approches diversifiées telles que les State Space Models comme Mamba, les architectures Mixture of Experts ou encore les RWKV ainsi que les grands modèles de langues multimodaux.

2 La tâche de reconnaissance automatique de la parole

2.1 Les unités de mesure en reconnaissance de la parole

Pour mesurer la qualité et la rapidité d'un modèle de reconnaissance automatique de la parole, deux métriques très populaires sont employées : le WER (Word Error Rate ou, en français, le taux d'erreur par mot) et le RTF (Real Time Factor ou, en français, facteur)et ses variantes (inversed Real Time Factor ou Speed Factor).

2.1.1 Word Error Rate (WER)

Le taux d'erreurs de mots (WER en anglais) est utilisé pour mesurer la qualité d'un système d'ASR. C'est un dérivé de la distance de Levenshtein à l'échelle des mots. Une valeur basse indique une meilleure qualité qu'une valeur plus élevée.

$$WER = \frac{\text{Substitutions} + \text{Suppressions} + \text{Insertions}}{\text{Total des mots}} \tag{1}$$

2.1.2 Real Time Factor (RTF)

Le RTF et le RTFx rendent compte de l'efficience d'un modèle d'ASR ou de diarisation. Le Real-Time Factor correspond à la division de la durée d'inférence du modèle sur la durée de l'audio traité, une valeur plus basse est la meilleure.

$$RTF = \frac{\text{Durée de l'inférence}}{\text{Durée de l'audio}} \tag{2}$$

Par exemple, si un système prend 8 heures à traiter 2 heures d'audio son RTF et de 4. À l'inverse, si un système prend 1 heure pour traiter 2 heures d'audio son RTF et de 0,5. Sur les modèles à l'état de l'art depuis des services API tels que Google, AWS, Azure, les valeurs de RTF sont souvent inférieures à 1. Sur des systèmes embarqués tournant hors-ligne sur du matériel local, ce sont les ressources matérielles (processeurs et mémoire vive) qui ont le plus grand impact sur le RTF.

2.1.3 Inverse Real-Time Factor (RTFx) ou Speed Factor

À l'inverse, l'unité de mesure Inverse Real-Time Factor (RTFx) ou aussi appelée Speed Factor est la durée de l'audio divisée par la durée d'inférence.

$$RTFx = \frac{1}{RTF} = \frac{\text{Durée de l'audio}}{\text{Durée de l'inférence}} \tag{3}$$

Une valeur RTFx plus élevée indique un temps de latence plus faible. Cette métrique est considérée comme plus lisible que le RTF car le facteur d'efficacité du modèle est directement corrélé à un ratio. Pour un système qui prend 8 heures à traiter 2 heures d'audio son RTFx est de 4, et pour un système qui traite en 1 heure 2 heures d'audio son RTFx et de 2.

2.2 Comparer les solutions d'ASR

Pour comparer les divers systèmes d'ASR, de nombreux classements (Leaderboard, benchmark en anglais) mettant en concurrence les systèmes d'ASR publics existent. Actuellement, le plus reconnu est l'open_asr_leaderboard publié par Hugging Face [3]. L'en-

semble des modèles et des données sont présentés dans l'optique de fournir des résultats transparents et reproductibles suivant un niveau d'exigence académique.

2.2.1 Hugging Face

Hugging Face est une entreprise franco-américaine créée en 2016 qui propose des bibliothèques de programmation, des ensembles de données et des modèles de réseaux de neurones en accès open source. La plateforme Hugging Face Hub permet l'hébergement de dépôts Git, d'applications web et de modèles par ses utilisateurs. Actuellement, plus de 2 millions de modèles différents de réseaux de neurones et plus de 530 000 ensembles de données en accès open source y sont recensés. Le fondateur Clément Delangue affirme que la plateforme compte plus de 164 000 organisations, dont des entreprises, des instituts de recherche et des collectifs de contributeurs, ainsi que près de 10 millions d'utilisateurs. Par ailleurs, la majorité des grands modèles de langues open source sont publiés directement sur Hugging Face, tels que ceux de Meta, de Mistral ou encore DeepSeek. En résumé, cette plateforme est un hub leader et reconnu par le monde professionnel et le monde de la recherche pour le partage autour du ML/DNN. [2]

2.2.2 Jeu de données de l'open_asr_leaderboard

Sur le classement open_asr_leaderboard, chaque modèle soumis est évalué sur un ensemble de corpus publics et multilingues couvrant diverses situations. Par exemple ce classement inclus LibriSpeech qui est un corpus de 980 heures de livres audio aux conditions d'enregistrement optimales mais aussi le corpus SPGISpeech constitué d'audio liés à des activités financières et économiques dans des conditions d'enregistrement moins qualitatives. La diversité des corpus de données utilisées permet d'évaluer les modèles de reconnaissance de la parole dans des conditions très variées, reflétant à la fois des scénarios optimaux (comme LibriSpeech pour l'anglais avec des audio lus et sans bruits parasites) et des contextes plus complexes ou réalistes (comme SPGISpeech pour l'anglais des affaires dans des conditions d'enregistrement parfois bruitées). Cette diversité permet de mesurer non seulement la précision des transcriptions (via le Word Error Rate – WER), mais aussi la robustesse des modèles face à la longueur, aux accents, aux types d'interférences, et d'en évaluer leur rapidité (facteur RTFx). Néanmoins, ces corpus restent anglo-centrés, l'anglais représentant 57% avec 980 heures d'audio et le français représente 11% avec 190 heures d'audio.

Table 1: Corpus employés pour le classement open_asr de HuggingFace

Corpus	Domaine	Style	Train (h)	Dev (h)	Test (h)	Transcriptions
LibriSpeech	Audiobook	Narrated	960	11	11	Normalised
VoxPopuli	EU Parliament	Oratory	523	5	5	Punctuated
TED-LIUM	TED talks	Oratory	454	2	3	Normalised
GigaSpeech	Audiobook, podcast, YouTube	Spontaneous	2500	12	40	Punctuated
SPGISpeech	Financial meetings	Oratory, spontaneous	4900	100	100	Punctuated & Cased

Corpus	Domaine	Style	Train (h)	Dev (h)	Test (h)	Transcriptions
Earnings-22	Financial meetings	Oratory, spontaneous	105	5	5	Punctuated & Cased
AMI	Meetings	Spontaneous	78	9	9	Punctuated & Cased

En résumé, la méthodologie du classement, suivant les recommandations de l'ESB [12] en matière de constitution du jeu de donnée, assure la pertinence et la comparabilité des résultats sur un ensemble de scénarii réels d'application en reconnaissance automatique de la parole. Les corpus comportant un style spontané semblent plus proches de nos données audio issues d'archives orales. Dans le cadre d'un usage d'ASR sur des témoignages oraux réalisés en tant qu'archives orales, il reste difficile de savoir si ces corpus sont représentatifs des archives orales que nous traitons dans le cadre de la thèse.

2.3 Analyse du classement

Le classement permet de filtrer par qualité des modèles sur les données en français, c'est-à-dire les taux d'erreur par mots (Word Error Rate) triés par ordre décroissant sur le jeu de données en français. Ici est présenté pour le 15 août 2025 un extrait du classement prenant les métriques de 7 modèles au WER le plus bas sur les corpus français.

Table 2: Classement des modèles ASR pour le français (WER et RTFx) issu du Leaderboard open_asr de HuggingFace

Modèle	WER moyenne multilangue	WER French	RTFx	Licence	Nb de Paramètres (en milliard)	Date de sortie
nvidia/canary-1b-v2	4.89	4.86	630.22	CC BY 4.0	1	2025/08/18
microsoft/Phi-4-multimodal-instruct	4.6	5.13	25.12	MIT	6	2025/02/26
nvidia/parakeet-tdt-0.6b-v3	5.05	5.38	2154.22	CC BY 4.0	0.6	2025/08/14
mistralai/Voxtral-Mini-3B-2507	5.18	5.96	58.93	Apache 2.0	3	2025/08/03
openai/whisper-large-v3	4.91	6.59	126.16	MIT	1.55	2023/11/06
openai/whisper-large-v3-turbo	5.44	7.01	188.6	MIT	0.8	2024/10/17
elevenlabs/scribe_v1	12.57	15.11	NA	Propriétaire	NA	2025/02/26

On peut constater qu'à l'opposé du modèle scribe_v1 d'ElevenLabs, les modèles les plus performants sont des modèles dont l'architecture et les poids sont dits ouverts, c'est-à-dire open-source. Cela signifie qu'ils sont disponibles au téléchargement pour fonctionner sur n'importe quel matériel et non uniquement accessibles via une API. Cela est particulièrement adapté pour traiter des fichiers audio tout en garantissant la confidentialité des données, sans devoir recourir à des serveurs tiers.

On peut également constater que deux des modèles, Phi-4-multimodal-instruct et Voxtral-Mini-3B-2507, sont de grands modèles de langage multimodaux. Quant aux autres, ce sont des implémentations de réseaux de neurones encodeur-décodeur Transformer, sauf pour parakeet-tdt-0.6b-v3, qui utilise une variante de cette architecture avec un décodeur Transducteur.

Il est intéressant de noter que, bien que les modèles multimodaux figurent dans le top 7 des meilleurs modèles ASR pour le français, leur latence d'exécution, indiquée par le RTFx, les place parmi les modèles les plus lents, et de loin. À l'inverse, ce sont les modèles produits par NVIDIA qui sont les plus rapides. Le modèle parakeet-tdt-0.6b, avec son architecture spécifique, présente un WER légèrement supérieur à celui de canary-1b-v2, mais il est quasiment quatre fois plus rapide.

Ce classement permet notamment de comprendre l'évolution des modèles : pendant presque deux ans, les modèles de la famille Whisper d'OpenAI ont dominé le classement et continuent d'être compétitifs, avec leur place dans le top 7 des meilleures solutions ASR, tant du point de vue de la qualité que de la rapidité d'exécution.

D'autres classements connus tels que celui d'Artificial Analysis [7], Voice Writer [6] ou encore PicoVoice [4] sont parfois utilisés mais leur méthodologie ou leurs résultats sont soit pas reproductibles, soit ne cherchent pas à suivre un niveau d'exigence académique. Aussi, des classements par corpus spécifique existent comme le LibriSpeech-PC Benchmark [15], mais n'ont pas la dimension ou la diversité de l'open_asr_leaderboard.

3 La séparation des locuteurs

Pour produire une transcription des témoignages oraux pertinente dans le cadre de la thèse, la tâche de séparation des locuteurs appelée diarisation est une étape nécéssaire et complémentaire de la tâche de transcription. Elle permet de répondre à la question "qui parle, et quand ?". Les modèles présentés dans le classement open_asr_leaderboard de Hugging face ne gèrent pas cette tâche, hormis le modèle propriétaire d'Elevenlabs qui fonctionne comme un service complet avec ASR + diarisation pour produire des transcriptions fonctionnelles mais qui est surtout un service commercial fonctionnant uniquement par API. La diarisation est une tâche de classification multi-label avec l'objectif d'identifier des clusters pour les tours de parole d'un même locuteur. Généralement, cette tâche se réalise sans information préalable sur le nombre de locuteurs présents dans un audio ni n'identifie les locuteurs eux-mêmes (c'est-à-dire qu'il différencie un locuteur A d'un locuteur B, pas qu'il identifie une personne en tant que tel). Cette tâche est une étape complémentaire à la reconnaissance automatique de la parole permettant la production d'une transcription pleinement exploitable. Nous présentons les unités de mesures pertinentes pour la tâche de diarisation, et ferons le commentaire de deux papiers proposant des classements de systèmes de diarisation.

BENCHMARKING DIARIZATION MODELS[14] SDBench : A Comprehensive Benchmark Suite for Speaker Diarization[16]

3.1 Métriques pertinentes pour la diarisation

3.1.1 Diarization Error Rate (DER)

La DER (Diarization Error Rate) est calculée comme la somme du temps de confusion des locuteurs, des faux positifs (durée taguée comme parole alors qu'il n'y en a pas), et

des segments manqués, divisée par la durée totale annotée. Cet indicateur est le standard du domaine pour comparer la qualité des systèmes de diarisation.

$$DER = \frac{\text{Faux positifs} + \text{Faux négatifs} + \text{Confusions de locuteur}}{\text{Durée totale de parole}} \tag{4}$$

Une plus faible valeur de DER est meilleure. Il existe plusieurs variantes de cette unité de mesure telles que la Jaccard Error Rate qui harmonise la contribution de chaque locuteur sans prendre en compte la durée de leur temps de parole pour lisser le poids entre les locuteurs, nous n'avons pas à faire à cette unité dans les articles que nous présentons.

3.1.2 RTF, RTFx et Speed Factor

Le Real-Time Factor (RTF) et le Speed Factor (aussi appelé l'inverse RTFx) sont utilisés pour analyser l'efficience d'un modèle de diarisation ; ce sont les mêmes mesures de l'ASR appliquées à la diarisation.

3.2 Le classement du BENCHMARKING DIARIZATION MODELS

Le preprint [14] publié le 30 septembre 2025 propose de comparer 6 modèles de diarisation sur 4 jeux de données en utilisant le DER. 2 modèles présentés sont issus de la famille de modèles de l'entreprise PyannoteAI. Le 29 septembre 2025, PyannoteAI a sorti de nouveaux modèles (community-1 remplaçant speaker-diarization-3.1 et precision-2 remplaçant precision-1) entraînés sur le super-calculateur JeanZay, ils n'ont donc pas pu être pris en compte dans le classement proposé par ce preprint. Le preprint propose des comparaisons de DER par corpus, par langue, par RTFx (RTF écrit dans le preprint) et par DER agrégé par nombre de locuteurs présents.

3.2.1 Le jeu de données

Le jeu de données est composé de 4 corpus distincts qui sont diversifiés et composés au total de 5 langues dont le français ne fait pas partie. La présence de l'anglais, de l'allemand et de l'espagnol permet par la proximité de ces langues avec le français de supposer que les résultats seraient sensiblement similaires pour le français. L'anglais reste sur-représenté et 2 langues asiatiques sont présentes, le japonais et le mandarin. Les corpus représentent des conversations téléphoniques, des conversations multi-locuteurs en anglais, et des conversations issues d'un cadre professionnel. Tant par le nombre de langues que par la durée totale, le jeu de données de diarisation est plus petit que le jeu de données présenté par l'ESB. Il semble que la tâche manuelle d'annotation des segments de prise de parole par locuteur soit une tâche plus complexe et plus chronophage que la transcription manuelle de référence, cela pourrait expliquer l'écart de taille entre ces 2 jeux de données de référence, l'un pour l'ASR et l'autre pour la diarisation.

3.2.2 Analyse des résultats

Table 3: Comparaison des modèles de diarisation vocale — DER moyen, RTFx, licence et date de sortie issu du preprint.

Modèle	Ali (DER)	Ami (DER)	Callhome (DER)	Voxconverse (DER)	Moyenne (DER)	Moyenne RTFx	Licence	Date de sortie
pyannoteAI/precision-1	15.5	13.2	10.7	5.4	9.1	N/A	Propriétaire	2024/04/..
DiariZen	11.5	23.7	12.8	5.2	13.3	20.2	CC BY-NC 4.0	2025/05/30
pyannote/speaker-diarization-3.1	17.1	18.5	22.5	8.9	14.5	45	MIT	2023/06/..
nvidia/sortformer-v2	7.9	23.7	12.6	15.1	17.1	214.3	CC BY-NC 4.0	2024/09/10
nvidia/sortformer-v2-stream	7.0	22.2	12.6	15.1	16.4	209.5	CC BY-NC 4.0	2024/09/10
nvidia/sortformer	15.1	27.7	14.2	13.8	17.7	164.7	CC BY-NC 4.0	2024/09/10

Le projet Pyannote est à l'origine issu de la thèse de Hervé Bredin de 2008 qui, depuis, a créé l'entreprise PyAnnoteAI qui propose des modèles de diarisation open-source et des modèles plus performants propriétaires par accès API. Une très grande majorité des solutions commerciales de transcription automatique (tel que Gladia) emploie les modèles de PyannoteAI pour la tâche de diarisation. PyannoteAI produit depuis plusieurs années les modèles les plus performants pour la diarisation. À titre d'information, durant le mois d'octobre 2025, HuggingFace recense 12,8 millions de téléchargements du modèle speaker-diarisation-3.1 de Pyannote. Il semble que DiariZen et NVIDIA commencent depuis peu à bien concurrencer Pyannote. DiariZen semble plus précis mais plus lent, tandis que NVIDIA produit des modèles très rapides mais moins précis.

3.3 Le classement du SDBench

Le second et dernier classement sur la tâche de diarisation présentée ici est mis en complément du précédent classement. Il a été publié dans la revue INTERSPEECH d'août 2025 et semble donc avoir plus de poids scientifiquement parlant. Ce classement utilise les unités de mesures de DER pour la qualité et de Speed Factor pour l'efficience des modèles. Le Speed Factor est une unité de mesure similaire au RTFx. Cet article s'intéresse particulièrement à décomposer la métrique de DER ainsi que les différents étapes internes à la tâche de diarisation pour comprendre quelles sont les étapes les plus difficile à résoudre et les moins performantes. L'objectif est d'orienter le travail des producteurs de système de diarisation vers ces étapes précises pour identifier et améliorer les solutions. Il est aussi présenté une solution SpeakerKit, développé dans le cadre de la recherche de cet article, optimisée pour l'efficience basée sur le modèle de pyannote open-source Ce classement compare des solutions API fermées tel que AWS Transcribe, Deegram, PicoVoice et Pyannote-AI face à deux solutions open-source Pyannote (speaker-diarization-3.1) et une implémentation optimisée développée par les auteurs de l'article SpeakerKit.

3.3.1 Le jeu de données

Le jeu de donnée utilisé dans l'article est composé de 13 corpus distincts dont 4 sont les mêmes que présentés dans le preprint précédent. Il est composé de corpus multilingues, avec des typologies d'audio variés. Une partie des corpus n'est pas open-source, l'autre partie est mise à disposition sur Hugging Face [10]. Le français n'est pas mis en avant dans cet article, ce sont les langues anglaise, chinoise, arabe, tagalog et espagnole qui sont explicitement présentées, les autres langues dont le français sont inclus dans un groupe Other.

3.3.2 Analyse du classement

Ce qui nous intéresse dans ce classement ce sont notamment les figures 1 et 5 qui présentent les taux de DER des modèles de Pyannote et de l'implémentation otpimisée SpeakerKit basée sur Pyannote speaker-diarization-3.1 face aux solutions propriétaires par API. Il apparaît clairement que PyannoteAI, depuis la figure 1, est le modèle avec le DER le plus faible, suivi du modèle Pyannote en second place puis de SpeakerKit basé sur Pyannote en troisième position. SpeakerKit est le modèle le plus efficient avec le Speed Factor le plus élevé tandis que les modèles de Pyannote sont en 4 et 5ème position.

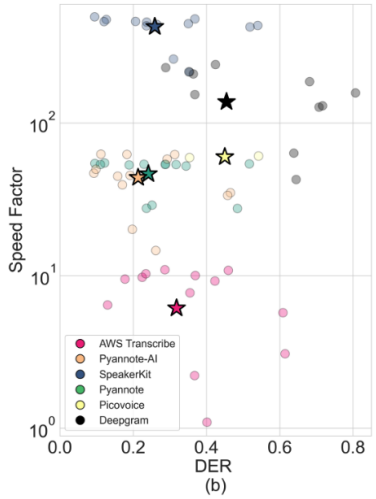

FIGURE 1 – Figure 1 : DER vs Speed Factor : (b) 6 speaker diarization systems across 13 datasets. Circle markers represent per-dataset results whereas star markers represent cross-dataset aggregation. (Extrait du SDBench)

La seconde figure reproduite ici intitulée dans SDBench Figure 5 : "DER breakdown across multiple systems, categorized by language." montre que le modèle speaker-diarization-3.1 et le modèle fermé precision-1 de Pyannote ainsi que l'implémentation efficiente développée pour le SDBench appelée SpeakerKit issu du modèle speaker-diarization-3.1 sont les modèles avec le DER le plus faible, tous modèles confondus et toutes langues analysées. Ce classement, comparant donc les modèles Pyannote à des solutions différentes du précédent classement, met encore les modèles Pyannote en premières positions.

3.4 Pyannote : promesse des nouveaux modèles

Les nouveaux modèles publiés fin septembre par Pyannote sur Hugging Face [5] tels que community-1 (open-source) et precision-2 (propriétaire par API) obtiennent des scores

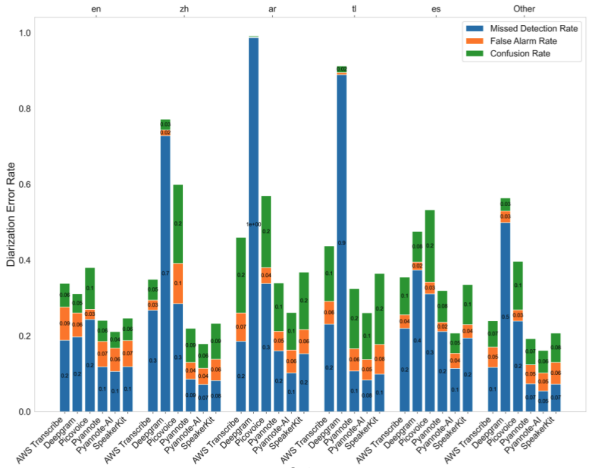

FIGURE 2 – Figure 5 : DER breakdown across multiple systems, categorized by language. (Extrait du SDBench)

DER encore plus faibles que les modèles précédents speaker-diarization-3.1 et precision-1 (propriétaire par API).

TABLE 4: Classement par DER modèles Pyannote (mis à jour en 2025-09)

Benchmark	legacy (3.1)	community-1	precision-2
AISHELL-4	12.2	11.7	11.4
AliMeeting (channel 1)	24.5	20.3	15.2
AMI (IHM)	18.8	17.0	12.9
AMI (SDM)	22.7	19.9	15.6
AVA-AVD	49.7	44.6	37.1
CALLHOME (part 2)	28.5	26.7	16.6
DIHARD 3 (full)	21.4	20.2	14.7
Ego4D (dev.)	51.2	46.8	39.0
MSDWild	25.4	22.8	17.3
RAMC	22.2	20.8	10.5
REPERE (phase2)	7.9	8.9	7.4
VoxConverse (v0.3)	11.2	11.2	8.5

Sur les pages de présentation des nouveaux modèles, les scores de DER des anciens modèles sont sensiblement différents des scores obtenus dans les deux classements présentés ici. Il faut souligner que, bien que ce soient les mêmes corpus utilisés dans la présentation des nouveaux modèles sur Hugging Face et dans les deux classements de diarisation présentés ici, ce sont majoritairement des versions différentes de ces corpus qui sont utilisées par Pyannote pour présenter ses nouveaux modèles. Cela permet d'expliquer en partie les résultats différents de DER entre les deux tableaux pour le même modèle speaker-diarization-3.1. Néanmoins, les résultats montrent que proportionnellement les nouveaux modèles de Pyannote sont plus performants (environ 6% pour le community-1 face à

speaker-diarization-3.1). Renforçant encore la certitude que l'entreprise PyannoteAI reste leader et produit des modèles à l'état de l'art pour la tâche de diarisation.

4 WhisperX, une solution complète et modulaire

Le logiciel WhisperX, développé initialement par Max Bain dans le cadre de ses recherches doctorales et présenté à INTERSPEECH 2023[11], se présente comme une solution complète, modulaire et à l'état de l'art pour la transcription automatique et la diarisation. Il étend le modèle Whisper d'OpenAI, déjà reconnu pour sa robustesse multilingue, en y ajoutant des améliorations significatives qui le rend particulièrement adapté aux besoins de la recherche et des corpus d'archives orales. WhisperX intègre une étape d'alignement phonétique post-traitement (via le modèle Wav2Vec2 d'OpenAI ou d'autres modèles forçant l'alignement), permettant de corriger les imprécisions temporelles du modèle Whisper original. Les transcriptions sont synchronisées mot à mot avec l'audio, ce qui est crucial pour l'analyse fine des prises de parole, la segmentation ou l'annotation qualitative et facilite l'alignement avec l'étape de diarisation. Contrairement au modèle Whisper d'origine, optimisé pour des segments d'environ 30 secondes, WhisperX gère efficacement des audios longs (plusieurs heures) sans dégradation majeure des performances. Cette solution se distingue aussi par sa capacité à connecter des modules externes de diarisation, notamment les modèles Pyannote open-source, via une interface simple et reproductible. WhisperX repose sur les poids ouverts des modèles Whisper (MIT License) et publie son code sous licence BSD-2-Clause, garantissant la transparence, la réplicabilité et le contrôle des données. Cela permet de répondre aux exigences académiques et éthiques de la recherche en SHS. Il est possible d'importer d'autres modèles que ceux d'origine tels que des modèles raffinés sur des données en français [8] ou des modèles optimisés pour du matériel moins puissant qui présentent des taux d'erreur de mots (WER) plus faibles que Whisper de base. Le pipeline WhisperX améliore le temps d'exécution (RTFx) grâce à un découpage dynamique des segments audio, une gestion optimisée du GPU et des buffers, et une parallélisation automatique. WhisperX peut être exécuté sur des machines locales (CPU/GPU) ou sur des serveurs distants, et il supporte plusieurs formats audio et sorties structurées (JSON, TXT, SRT, VTT). Cela facilite l'intégration dans des pipelines de recherche ou de traitement documentaire existants (annotation, indexation, analyse sémantique).

4.1 Usage de WhisperX sur les archives orales

Dans le cadre de la thèse en cours, WhisperX a été déployé sur le corpus des archives orales Michelin, une mission réalisée en partenariat avec l'Observatoire B2V des mémoires et l'agence Perles d'Histoire. La phase de transcription automatique m'a été confiée et a permis de tester sur un corpus réel d'archives orales ce logiciel. Le modèle de transcription choisi est le deepdml/faster-whisper-large-v3-turbo-ct2, un dérivé de Whisper Large v3 Turbo qui est quantifié et au format Ctranslate2. La quantification et le changement de format permettent d'augmenter l'efficacité du modèle, notamment au niveau de son temps d'inférence, tout en conservant une qualité similaire au modèle de base. Le modèle de diarisation choisi a été speaker-diarization-3.1 de Pyannote [9]. Par ailleurs la sortie récente d'un nouveau modèle open-source community-1 de Pyannote pourra être facilement intégré au logiciel WhisperX. Plusieurs paramètres pour améliorer les prédictions de transcription et de diarisation ont été spécifiés. Les paramètres donnant un intervalle pour

le nombre minimal et maximal de locuteurs à identifier étaient indiqués pour chaque entretien. Le paramètre du prompt initial a été rempli d'un vocabulaire métier et technique propre à l'entreprise Michelin pour améliorer l'orthographe prédite de termes spécifiques.

Conclusion

En conclusion, WhisperX s'impose comme la solution la plus adaptée pour la transcription des archives orales dans le cadre de cette thèse. Sa combinaison d'un modèle de reconnaissance de la parole à l'état de l'art (Whisper), d'un alignement temporel précis et d'une intégration directe de la diarisation en fait un outil complet répondant aux exigences scientifiques de fidélité, de traçabilité et de reproductibilité. Son fonctionnement open-source et local garantit par ailleurs le respect de la confidentialité des données sensibles contenues dans les témoignages, tout en offrant une flexibilité technique et une performance suffisante pour traiter de grands volumes d'audio. WhisperX constitue ainsi une solution robuste et transparente, parfaitement alignée avec les objectifs méthodologiques et éthiques de cette recherche sans compromis avec la qualité. Néanmoins nous avons conscience que le domaine de l'ASR évolue de plus en plus rapidement et qu'il reste probable que de nouveaux modèles surpassent rapidement les solutions actuellement à l'état de l'art. Comme dirait Károly Zsolnai-Fehér de la chaîne Two Minute Papers : "What a time to be alive ! ".

Références

[1] The HARPY Speech Recognition System. https ://apps.dtic.mil/sti/citations/ADA035146.

[2] Interview [Redif] – Hugging Face : Le poids-lourd de l'IA open source (Clément Delangue, Hugging Face) | Monde Numérique - Podcast | Actualité des Technologies. https ://mondenumerique.info/episode/interview-redif-hugging-face-le-poids-lourd-de-lia-open-source-clement-delangue-hugging-face.

[3] Open ASR Leaderboard - a Hugging Face Space by hf-audio. https ://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

[4] Open-Source Speech-to-Text Benchmark - Picovoice Docs. https ://picovoice.ai/docs/benchmark/stt/.

[5] pyannoteAI Speaker Intelligence and Diarization. https ://www.pyannote.ai/.

[6] Real-World Speech-to-text API Leaderboard | Voice Writer. https ://voicewriter.io/speech-recognition-leaderboard.

[7] Speech to Text (ASR) Providers Leaderboard & Comparison | Artificial Analysis. https ://artificialanalysis.ai/speech-to-text.

[8] Bofenghuang/whisper-large-v3-french · Hugging Face. https ://huggingface.co/bofenghuang/whisper-large-v3-french, February 2023.

[9] Pyannote/speaker-diarization-3.1 · Hugging Face. https ://huggingface.co/pyannote/speaker-diarization-3.1, September 2025.

[10] Speaker Diarization Datasets - a argmaxinc Collection. https ://huggingface.co/collections/argmaxinc/speaker-diarization-datasets, June 2025.

[11] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX : Time-Accurate Speech Transcription of Long-Form Audio, July 2023.

[12] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. ESB : A Benchmark For Multi-Domain End-to-End Speech Recognition, October 2022.

[13] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep Speech : Scaling up end-to-end speech recognition, December 2014.

[14] Luca A. Lanzendörfer, Florian Grötschla, Cesare Blaser, and Roger Wattenhofer. Benchmarking Diarization Models, September 2025.

[15] Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. LibriSpeech-PC : Benchmark for Evaluation of Punctuation and Capitalization Capabilities of end-to-end ASR Models, October 2023.

[16] Eduardo Pacheco, Atila Orhon, Berkin Durmus, Blaise Munyampirwa, and Andrey Leonov. SDBench : A Comprehensive Benchmark Suite for Speaker Diarization, August 2025.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, December 2022.