Diarisation multimodale : vers des modèles robustes et justes en contexte réel

Yannis Tevissen

INSTITUT POLYTECHNIQUE DE PARIS

TELECOM SudParis

IP PARIS

NNT : 2023IPPAS014

Thèse de doctorat

Diarisation multimodale : vers des modèles robustes et justes en contexte réel

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 :
École Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Signal, Images, Automatique et Robotique

Thèse présentée et soutenue à Palaiseau, le 04 décembre 2023, par

Yannis Tevissen

Composition du Jury :

Sylvain Meignier Professeur, Université du Mans - LIUM	Président & Rapporteur
Björn W. Schuller Professeur, Imperial College of London	Rapporteur
Hervé Bredin Chargé de recherche CNRS, IRIT (UMR 5505)	Examinateur
Dijana Petrovska Maître de conférence, IP Paris - Télécom SudParis, SAMOVAR	Examinatrice
Jérôme Boudy Professeur, IP Paris - Télécom SudParis, SAMOVAR	Directeur de thèse
Gérard Chollet Directeur de recherche CNRS émérite	Co-Directeur de thèse
Frédéric Petitpont Cofondateur et Directeur technique, Newsbridge	Invité

Remerciements

Je tiens tout d'abord à remercier les professeurs Björn Schuller et Sylvain Meignier pour avoir accepté de rapporter mon travail de thèse, ainsi que Dijana Petrovska et Hervé Bredin d'avoir accepté de prendre part à mon jury.

Pendant ces trois années de thèse, de nombreuses personnes m'ont également épaulé, soutenu, encouragé. Je tiens ici à les remercier. En premier lieu desquels, mes directeurs de thèse, Jérôme Boudy et Gérard Chollet.

Jérôme, un immense merci pour votre confiance, votre soutien, vos conseils éclairés et votre connaissance du milieu scientifique que vous me partagez depuis déjà avant l'entame de ma thèse. Merci également de m'avoir encouragé à faire cette thèse.

Gérard, merci énormément pour votre regard avisé et toujours pertinent sur mes travaux. Merci aussi pour vos partages réguliers qui permettent d'attiser ma curiosité scientifique.

Merci Sonia Garcia de m'avoir invité à parler de diarisation à vos étudiants. Merci à Estelle et Prateek dont j'ai eu la chance et le plaisir d'encadrer le stage.

Cette thèse ne se serait pas non plus faite sans Newsbridge qui m'a accueilli à bras ouverts et où j'ai pu me développer tant scientifiquement que dans tous les autres aspects de ma vie professionnelle. Merci beaucoup Fred et Phil pour votre confiance, votre vision et l'environnement de travail que vous avez réussi à créer. Merci aussi à Gil, Marina et Thibault qui, au quotidien, m'ont permis de devenir un meilleur ingénieur afin d'amener à la vie les projets imaginés dans cette thèse, et bien d'autres. Merci également à tous mes collègues pour leur soutien et leur convivialité : Erwan, Julien, Laura, Rachel, James, Carole, Émilie, Loïc, Gaël, Julie, Imogen, Olivier, Yvan, et tous les autres.

Merci à Aurélie, dont les livres et plus récemment les discussions me permettent de façonner la façon dont je conçois ma carrière dans les sciences.

Si ces trois années se sont aussi bien déroulées, c'est aussi en grande partie grâce à mes amis, leurs encouragements dans les bons moments et leur soutien dans les plus difficiles. Eva et Paul, avec lesquels j'ai tant partagé, galéré mais grâce à qui j'ai pu trouver ma voix. Les LGs, Carla, Farid, Maïlys, Khalil, Simon, Colombe, Marc, Judith, Julien, Raphaël, dont les conversations endiablées seraient bien difficiles à diariser. Les copains de promo que je ne vois pas assez : JB, Eloïse, Maxime, Delphine, et les autres. Et bien sûr, aux amis d'avant et de toujours, Nathan, Axelle, Cécile, Marie, Paul, Antoine, Marc, Nicolas.

Enfin je remercie ma famille et tout particulièrement mes parents, Rémi et Nadjette, qui, depuis ma plus tendre enfance m'ont accompagné et soutenu dans mes études pour finalement arriver aujourd'hui à travailler dans un domaine qui me passionne.

Table des matières

Introduction

Introduction générale

Les années 2022 et 2023 ont été particulièrement marquées par la démocratisation de solutions d'intelligence artificielle, dites conversationnelles. Ces systèmes dont la plupart fonctionnent essentiellement sur des données textuelles simulent des conversations entre un humain et une machine.

Toutefois, une large majorité des échanges entre humains se font par d'autres moyens que des échanges de messages écrits : les gestes, le regard, les expressions faciales et bien sûr la voix.

C'est notamment le cas à la télévision où les personnes qui apparaissent s'expriment, débattent, s'interpellent et essaient de transmettre des émotions au moyen de leurs voix. « Être, c'est être perçu » comme le soulignait le philosophe George Berkeley repris par Bourdieu dans son ouvrage « Sur la télévision ». Pour comprendre l'essence d'un commentaire sportif, d'une émission d'actualité, ou encore d'un reportage composé de vidéos d'archive il est donc nécessaire de pouvoir différencier les voix des différents intervenants pour percevoir ce que chacun a à dire.

Cette tâche est naturellement effectuée par le cerveau humain dès le plus jeune âge puisque même les nouveaux nés arrivent à différencier les voix de leurs parents. Pour une machine pourtant, cela reste très complexe, en particulier si on souhaite allier robustesse des systèmes et justesse du traitement pour toute la variété des locuteurs concernés.

Néanmoins, pouvoir comprendre automatiquement une conversation, même complexe, apparaît comme un enjeu majeur pour les systèmes actuels. On cherche aussi de plus en plus à faire cohabiter humains et machines mais pour ce faire, il apparait comme essentiel d'avoir des machines qui puissent analyser des échanges humains multipartites pour pouvoir apporter la valeur ajoutée qu'on en attend.

Dans le secteur de l'audiovisuel, en particulier, une vaste majorité des contenus sont structurés autour de conversations humaines : débats, interviews, commentaires sportifs, on parle même en anglais de *talk-show*, que l'on pourrait traduire littéralement par « spectacle de parole ».

I. Définitions

I.1. Diarisation du locuteur

La diarisation du locuteur, ou *speaker diarization*[1] en anglais, correspond à la tâche de déterminer automatiquement « qui parle, quand ? ». En somme, en *diarisant* on cherche à structurer un enregistrement audio ou vidéo en identifiant les différents tours de parole, puis en les attribuant aux différents locuteurs présents, comme illustré en Figure 1.

Figure 1. Représentation d'un résultat simple de diarisation

La dénomination *speaker diarization*, a été introduite par le programme EARS[2] du DARPA américain au début des années 2000 (Tranter and Reynolds, 2006), avant d'être popularisée par les évaluations RT du NIST en 2003 (NIST, 2003). Elle prend ses origines directement de la racine anglaise *diary*, le journal.

Avant ça, on retrouve dans la littérature d'autres appellations pour cette tâche telle que la *segregation of speakers* (Gish et al., 1991).

Il n'existe pas, à ce jour, de mot français équivalent. Dans la littérature scientifique francophone, on retrouve les traductions de structuration, journalisation, diarisation ou plus souvent une appellation détaillant les étapes : segmentation et regroupement (Delacourt, 2000). Dans ce document, on privilégiera le terme de diarisation qui permet facilement de faire des ponts entre les littératures scientifiques anglaise et française. On retrouve d'ailleurs en anglais les deux orthographes (*diarization* et *diarisation*).

La diarisation du locuteur est la pierre angulaire des différents systèmes de reconnaissance vocale qui visent à transcrire des conversations humaines. En effet, avant de pouvoir transformer les données audio en texte, il est nécessaire de séparer les propos prononcés par un locuteur de ceux de celui qui lui répond.

1. On trouve aussi l'écriture suivante : *speaker diarisation*.
2. https://www1.icsi.berkeley.edu/Speech/EARS/rt.html

La diarisation du locuteur trouve par exemple des applications dans l'analyse des conversations téléphoniques (Li, 2020), dans l'analyse des enregistrements multimédias diffusés à la télévision (Bendris, 2011), ou encore dans l'analyse de conversations médicales (Riad et al., 2022).

Si la diarisation est une discipline relativement récente, notamment propulsée au début des années 2000 par les recherches françaises, les approches et les systèmes qui la composent sont très divers, comme le montrait déjà Xavier Anguera en 2012 (Anguera Miro et al., 2012). Profitant du regain d'intérêt de la communauté pour l'intelligence artificielle et, en particulier avec l'apparition et la démocratisation des méthodes d'apprentissage profond (Park et al., 2021), la diarisation a vu le nombre et les performances de ses approches augmenter de façon significative au cours des cinq dernières années. Le nombre d'articles scientifiques publiés annuellement avec le mot clef *diarization* est un bon indicateur de l'intérêt croissant de la communauté scientifique pour cette tâche, avec plus de 1000 articles publiés sur Google Scholar en 2022. On constate également en comparant ces chiffres à ceux obtenus avec le mot clef *speech* un intérêt croissant au sein de la communauté de traitement de la parole (cf. Figure 2).

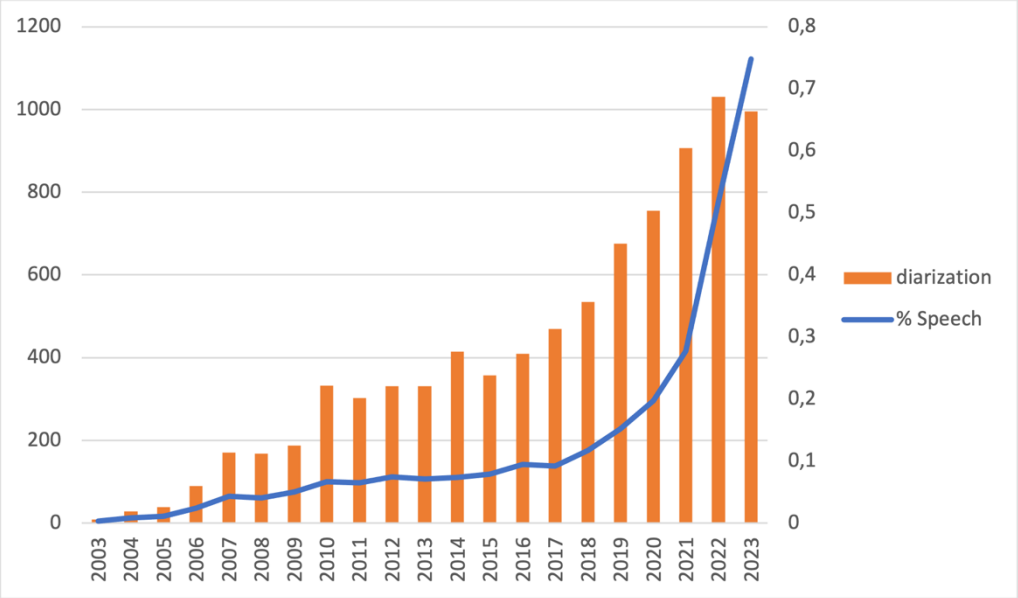

Figure 2 : Approximation du nombre d'articles scientifiques publiés au cours des dernières années avec le mot clef « diarization », comparaison avec le mot clef « speech ».
Résultats obtenus grâce à Google Scholar

I.2. Intelligence artificielle

Le terme d'intelligence artificielle est apparu en 1956 dans la bouche de John McCarthy, mathématicien et informaticien américain, lors de la conférence de Dartmouth. Le chercheur conjecture alors que tous les aspects du processus humain d'apprentissage et de l'intelligence peuvent être définis si précisément qu'une machine pourrait les reproduire à l'identique par un procédé appelé intelligence artificielle (IA)[3].

Aujourd'hui l'intelligence artificielle désigne l'ensemble des techniques avancées de calcul informatique permettant de résoudre des problèmes complexes. Cela inclut les algorithmes d'apprentissage automatique mais également d'autres méthodes telles que les algorithmes d'obtention du plus court chemin (Djikstra, A*, etc.) et les arbres de décision.

Dans ce texte, par souci de clarté, lorsque l'on fera mention d'intelligence artificielle, il s'agira exclusivement de systèmes d'apprentissage automatique dont ceux d'apprentissage profond (*deep learning*).

On parlera en particulier de systèmes tirant parti des dernières avancées en termes d'architectures de réseaux de neurones, telles que les LSTM (*long-short term memory*) (Hochreiter and Schmidhuber, 1997), les réseaux convolutifs CNN (LeCun et al., 2015) et les *transformers* (Vaswani et al., 2017) basés sur un système d'attention.

En effet, après quelques années de désintérêt[4] pour les domaines de l'intelligence artificielle dans les années 80, puis à nouveau dans les années 2000, la communauté scientifique a trouvé dans les algorithmes d'intelligence artificielle de nouvelles solutions pour résoudre des tâches de plus en plus complexes. On constate bien ces points d'inflexion dans la recherche autour de la diarisation du locuteur.

II. Résumé des contributions

Cette thèse de doctorat sur la diarisation du locuteur est l'occasion de proposer plusieurs contributions au domaine, tant sur l'aspect scientifique qu'industriel. En voici la liste :

3. « We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. », *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* (August 31st, 1955).
4. Parfois qualifiées d'« hivers de l'intelligence artificielle ».

1. Une méthode de détection d'activité vocale multi-flux avec un protocole de décision basé sur l'entropie. (Cf. Chapitre II-III, p. 51)
2. Une nouvelle méthode de détection des changements de locuteur dans une conversation textuelle grâce aux modèles de langage volumineux. (Cf. Chapitre III-III, p. 69)
3. Une première définition et mesure de la justesse des algorithmes de diarisation du locuteur. (Cf. Chapitre V : p. 89)
4. Le déploiement d'un algorithme de diarisation du locuteur pour l'analyse des données vidéos des clients de la société Newsbridge, en tenant compte des consommations énergétiques. (Cf. Chapitre VI-II, p. 103)
5. La création d'un nouveau système de diarisation du locuteur qui profite des dernières avancées de l'état de l'art pour le milieu médical. (Cf. Chapitre VI-IV, p. 109)

III. Plan du mémoire

Nous verrons tout d'abord dans une première partie le contexte qui a mené à ces travaux de recherche sur la diarisation du locuteur, en particulier en détaillant le secteur d'activité de la société Newsbridge.

Ensuite, nous détaillerons pourquoi la diarisation du locuteur, malgré ses multiples méthodes, souffre encore de ses faibles performances dans les scenarii les plus complexes. Ainsi, on discutera notamment de comment assurer la robustesse de la diarisation en contexte réel en présentant de nouvelles méthodes que nous avons introduites.

Enfin, nous nous intéresserons aux autres aspects nécessaires à l'utilisation des algorithmes diarisation à grande échelle. Nous verrons en effet qu'il est primordial d'avoir une diarisation la plus juste, la moins biaisée possible et la plus économe possible en termes de consommation énergétique.

Contexte de la recherche

Le travail présenté ici s'inscrit comme le premier partenariat de recherche entre la société Newsbridge et un acteur académique (le laboratoire SAMOVAR de Télécom SudParis). En complément du travail sur le passionnant sujet de la diarisation du locuteur, ce fut ainsi l'occasion d'explorer tous les aspects d'un partenariat de recherche, de la définition du sujet de recherche à la mise en production des premières solutions pour qu'elles soient utilisées par les clients de Newsbridge.

I. Présentation de Newsbridge

Newsbridge est une startup française du secteur des médias dont la mission est de rendre n'importe quelle archive audiovisuelle recherchable. En pratique, l'activité de la société se concentre sur l'indexation automatique de contenus photographiques et vidéos et sur la mise en place d'un moteur de recherche sémantique pour retrouver ces résultats d'indexation.

Figure 3 : Logo de Newsbridge

Grâce à la mise à disposition de ces technologies aux différents médias clients de la société, Newsbridge permet la réutilisation d'archives auparavant inutilisées et pour la plupart inutilisables. Les journalistes qui se connectent à la plateforme peuvent facilement effectuer des recherches pour illustrer et enrichir leurs reportages sur des sujets d'actualité ou sur des sujets historiques, puisque certaines chaînes de télévision disposent d'archives remontant à plus de 60 ans.

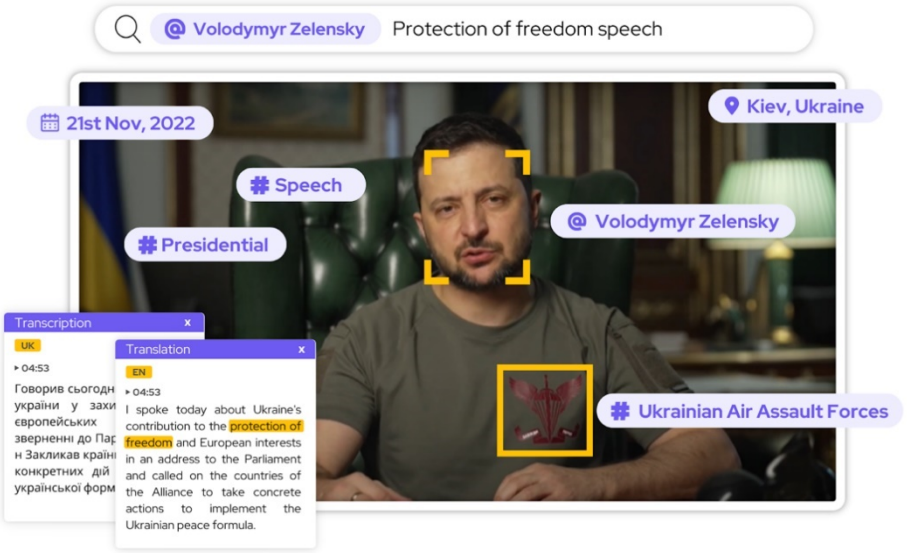

Figure 4 : Exemple du contenu des indexations effectuées automatiquement sur la plateforme Newsbridge. On retrouve une transcription, sa traduction, la personne détectée, un logo identifié ainsi qu'un certain nombre de métadonnées

La société Newsbridge travaille aujourd'hui avec des chaînes de télévision, clubs et fédérations sportives, et partis politiques principalement en Europe, aux Etats-Unis et au Moyen-Orient.

Avec plusieurs centaines de milliers d'heures archivées par Newsbridge, les besoins d'indexation automatique paraissent clairs. En effet, regarder 100 000 heures de vidéos prendrait à un humain[5] plus de 11 ans de sa vie. En parallèle, le temps de production de beaucoup des reportages diffusés à la télévision ne cesse de baisser avec parfois moins de 24h entre le tournage et la diffusion de sujets d'actualités brûlants. C'est pourquoi Newsbridge s'est tournée vers des outils d'intelligence artificielle, dont la diarisation, pour indexer automatiquement les vidéos de ses clients, et leur permettre de produire plus simplement et plus efficacement les contenus qu'ils diffusent.

5. Qui ne dormirait pas et travaillerait 24 heures sur 24.

II. Indexation automatique de contenus diffusés à la télévision

II.1. Communauté française de recherche

L'indexation automatique de contenus issus de la télévision existe depuis de nombreuses années en tant que produit commercialisé, mais également en tant que technologie issue de différents domaines de recherche.

En effet, cette thèse s'inscrit également dans un écosystème français de recherche riche qui s'est intéressé dès le début des années 2000 à la problématique de la reconnaissance des locuteurs (Meignier et al., 2001).

On note également que de nombreux travaux se sont déjà intéressés à la question de l'analyse des contenus vidéos diffusés à la télévision (Barras et al., 2006; Meignier et al., 2006; Poignant et al., 2012) dont plusieurs dans le cadre de *challenges* internationaux tels que MediaEval[6] avec en particulier les éditions 2015 (Poignant et al., 2015, 2017) et 2016 (Bredin et al., 2016) qui proposaient des tâches d'identification des personnes dans des émissions diffusées à la télévision.

Ces travaux sont riches, car la tâche est complexe notamment par la grande variété des contenus diffusés à la télévision.

II.2. Exemples de contenus diffusés à la télévision

Les contenus diffusés à la télévision sont très divers et ont souvent la marque d'une époque, d'une culture et de la volonté éditoriale de leur réalisateur. Néanmoins, on peut citer quatre larges catégories de contenus assez intemporelles et que l'on retrouve sur plusieurs chaînes de télévision à travers le monde :

- Les contenus d'actualité, c'est-à-dire les émissions dont le but premier est d'informer en relatant un fait récent ou en débattant d'un sujet de société. Ces médias sont souvent ceux qui font intervenir le plus de locuteurs différents, enregistrés dans des conditions très variables (studio, sur le terrain, en visioconférence, etc.).
- Les talk-show, émissions souvent enregistrées en studio avec un nombre d'intervenants fixe, connu à l'avance et une trame prédéfinie. La difficulté d'analyse de ces contenus réside dans le fait qu'ils sont souvent agrémentés

6. http://www.multimediaeval.org/ et https://multimediaeval.github.io/

de rires, d'applaudissements, de musiques et de bruitages, autant d'éléments pouvant perturber nos algorithmes de traitement de la parole.

- Les commentaires sportifs, fréquemment enregistrés dans des stades bruyants mais avec un nombre de locuteurs actifs restreints (2 ou 3 commentateurs le plus souvent hors-champ).
- La fiction, variété de contenu surement la plus hétérogène. On considèrera peu ce type de médias dans la suite de notre travail, d'autant que des méthodes très spécifiques existent pour les analyser (Bost et al., 2015).

II.3. L'enjeu clef de la transcription automatique pour la télévision

Pour Newsbridge, et plus généralement pour l'ensemble des acteurs du monde de la télévision, il est très important de pouvoir retranscrire automatiquement ce qui est dit à l'antenne. On pense immédiatement aux débats politiques dont l''archivage et la transcription est un véritable enjeu d'information et de démocratie, mais également aux contenus de divertissement tels que des commentaires sportifs ou des émissions de *talk-show*.

La retranscription automatique de la parole prononcée à la télévision, et sa structuration grâce à la diarisation du locuteur rendent possibles de nombreuses applications à toutes les étapes de la production d'une vidéo pour sa diffusion à la télévision.

II.3.1. Sous-titrage automatique

Le sous-titrage automatique, respectueux des tours de parole permet une meilleure accessibilité des contenus notamment pour leur publication sur les réseaux sociaux. Sans diarisation, ce sous-titrage est souvent difficile à lire et ne respecte pas les standards du secteur, tels qu'ils sont définis par la BBC[7].

7. https://www.bbc.co.uk/accessibility/forproducts/guides/subtitles/

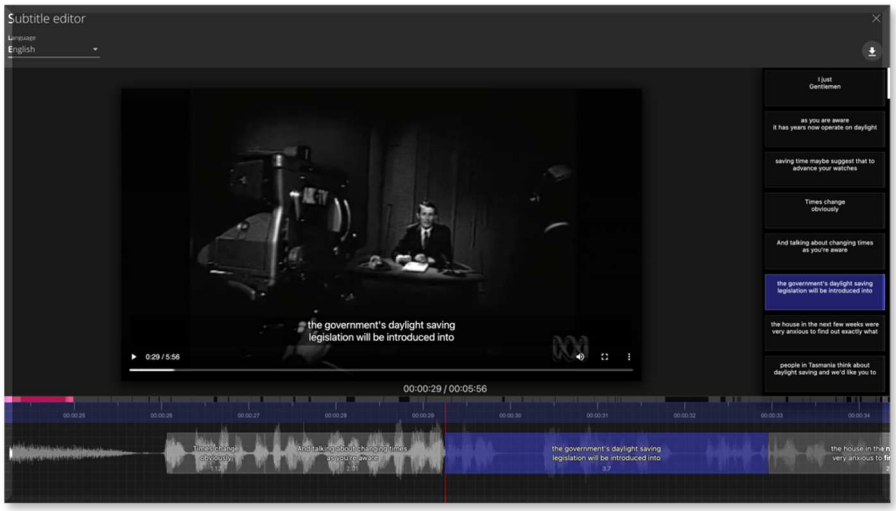

Figure 5 : Capture d'écran de l'éditeur de sous-titres développé par Newsbridge et utilisant l'algorithme de diarisation du locuteur conçu pendant cette thèse

II.3.2. Décompte du temps de parole

Le décompte automatique du temps de parole, outil éditorial fort, est également une obligation légale pour certaines émissions à certaines époques de la vie démocratique. En France, notamment, les chaînes de télévision sont tenues par l'ARCOM (Autorité de régulation de la communication audiovisuelle et numérique) de communiquer les temps de parole de chaque parti politique sur leurs antennes, en particulier lors des périodes de campagnes électorales. Celui qu'on surnomme le « gendarme de la télé » s'assure alors du respect de l'équité du temps de parole entre les partis.

Déterminer automatiquement qui parle, et pendant combien de temps, permettrait de rendre cette tâche plus simple à réaliser et de systématiser son usage pour des besoins éditoriaux ou de régulation. Plusieurs travaux ont d'ailleurs déjà montré la pertinence de l'utilisation de la diarisation du locuteur pour déterminer quel est le locuteur dominant, c'est-à-dire celui qui parle le plus, dans un enregistrement (Hung et al., 2011).

II.3.3. Outil de correction et d'extraction de transcription automatique

L'affichage ergonomique des transcriptions de vidéos incluant clairement les tours de parole permet une lecture facile et une extraction rapide des moments clefs. Pour les journalistes qui utilisent de tels outils cela représente un gain de temps certain.

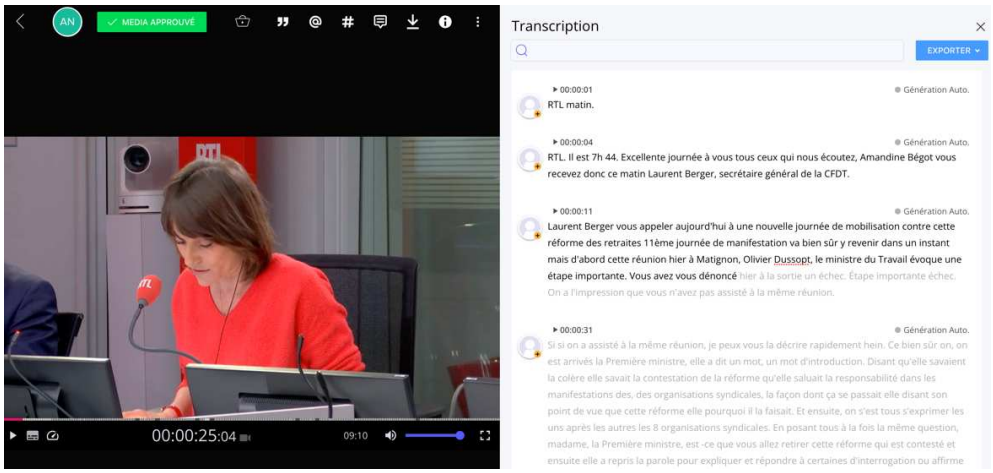

Figure 6 : Vue de la plateforme Newsbridge avec le panel « Transcription » ouvert sur un media d'actualité

II.3.4. Recherche dans de vastes archives audiovisuelles

La recherche parmi de larges archives audiovisuelles de moments clefs incluant des discours marquants est également un enjeu pour les journalistes utilisant l'IA pour les aider à archiver leurs contenus. C'est d'ailleurs là, la principale proposition de valeur de la société Newsbridge. La diarisation et la transcription automatique permettent d'indexer et de retrouver plus efficacement les phrases prononcées par un locuteur spécifique même lorsqu'elles l'ont été à l'occasion d'un débat multipartite.

En effet, couplées à des solutions d'identification des visages, ces technologies permettent de retrouver de façon multimodale tous les dires médiatiques d'une personne (Béchet et al., 2015; Bendris et al., 2013).

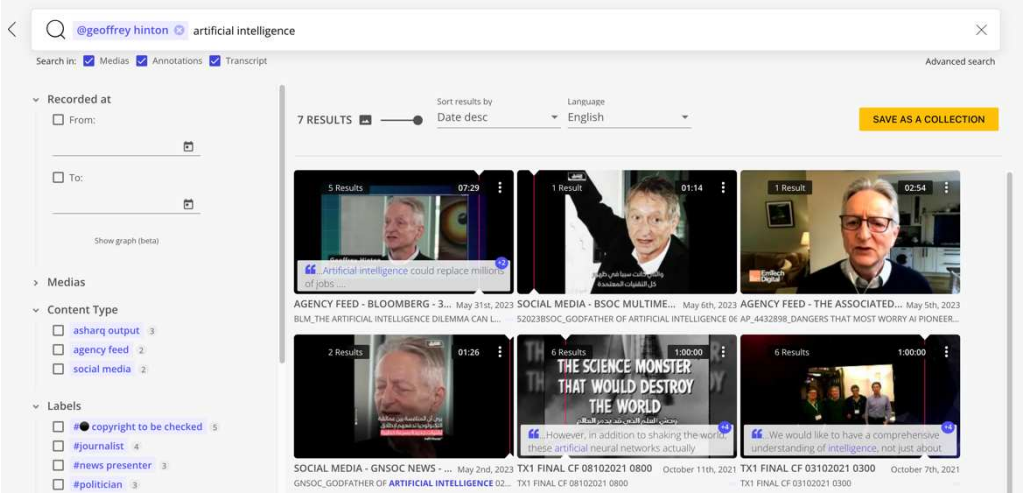

Figure 7 : Vue de la plateforme Newsbridge, la partie « Recherche » permet de retrouver dans de vastes archives les passages voulus. Ici on combine identification faciale et analyse de la parole pour retrouver toutes les fois où Geoffrey Hinton parle d'intelligence artificielle

Partie I : Robustesse de la diarisation du locuteur

La diarisation du locuteur se fait aujourd'hui essentiellement au moyen de réseaux de neurones profonds, que leurs concepteurs entrainent sur de grandes quantités de données pour essayer de garantir une certaine stabilité en termes de performance. Cette stabilité, même dans les conditions d'analyse les plus complexes, correspond à la caractéristique de robustesse du système. On dit d'un système qu'il est robuste s'il est également efficace quelques soient les variations de ses conditions d'utilisation. On s'intéressera en particulier à la robustesse des systèmes de diarisation du locuteur face aux différentes données qui peuvent leur être présentées.

Dans cette première partie, nous nous intéressons tout d'abord aux méthodes de diarisation du locuteur les plus récentes, avant de détailler les problèmes de robustesse les plus marqués de ces mêmes systèmes.

Dans le même temps, nous présenterons une nouvelle méthode centrée autour de la détection d'activité vocale.

Ensuite, nous verrons comment les approches multimodales peuvent permettre de palier certaines faiblesses des systèmes actuels, puis nous présenterons nos travaux autour de la détection de changements de locuteurs par des modèles volumineux de langage.

Chapitre I : État de l'art de la diarisation acoustique

A ce jour, l'état de l'art en diarisation est surtout focalisé autour des méthodes acoustiques. Celles-ci cherchent à déterminer « qui parle, quand ? » en se servant uniquement du signal audio.

La grande majorité des méthodes de diarisation sont aujourd'hui constituées d'une succession de modules ayant chacun des rôles distincts. On parle de diarisation modulaire ou en chaîne de composants. Ces architectures ont l'avantage d'être facilement explicables. Chaque composant a un rôle bien défini et il est facile de déterminer quelle partie du système accomplit correctement sa tâche ou non. C'est pourquoi, même s'il existe également des méthodes bout-en-bout, dans l'optique d'étudier la robustesse des systèmes de diarisation du locuteur, on s'intéressera principalement aux approches modulaires.

En effet, une architecture classique de diarisation peut être vue comme la concaténation de deux étapes principales : la segmentation et le regroupement.

Lors de la segmentation, on cherche à extraire de l'enregistrement audio des segments de parole dits homogènes, c'est-à-dire ne contenant la voix que d'un seul locuteur. Nous verrons qu'il est nécessaire de remettre parfois cette définition en cause pour s'adapter au cas de la superposition de paroles.

Une fois ces segments extraits, il s'agit alors de les regrouper (on parle de *clustering* en anglais) selon le locuteur qui les a prononcés.

A ces deux étapes, s'ajoutent fréquemment d'autres phases de post et pré-traitements mais toujours le plus généralement dans une logique linéaire et modulaire.

Ainsi la tâche de diarisation peut se schématiser facilement dans le cas général comme une succession de sous-tâches (cf. Figure 8).

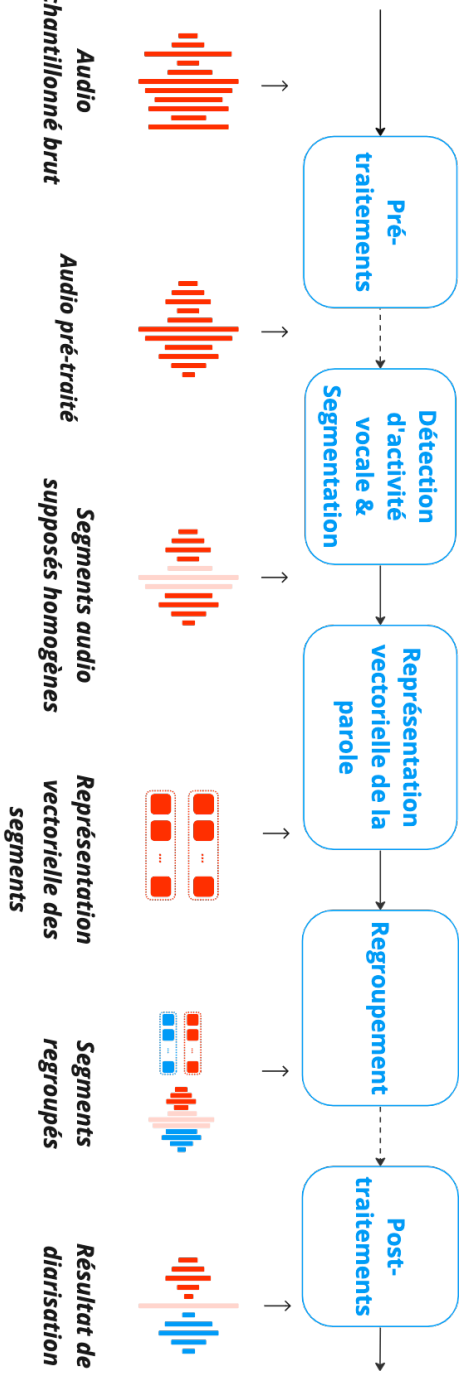

Figure 8. Schéma modulaire simplifié de la tâche de diarisation.
Ici deux segments sont représentés, chacun prononcé par un locuteur différent

I. Détection d'activité vocale et segmentation

I.1. Détection d'activité vocale

L'application d'un algorithme de diarisation du locuteur à un enregistrement, commence quasi-systématiquement par la détection du moment où quelqu'un parle. Cette détection d'activité vocale (*voice activity detection* - VAD) permet ensuite de ne travailler que sur les parties de l'audio qui contiennent une ou plusieurs voix, en éliminant notamment les silences ou les bruits parasites, non vocaux (cf. Figure 9).

De nombreuses méthodes de détection d'activité vocale existent, de la simple définition d'un seuil d'activation à des méthodes plus robustes basés sur des réseaux de neurones convolutifs et récurrents.

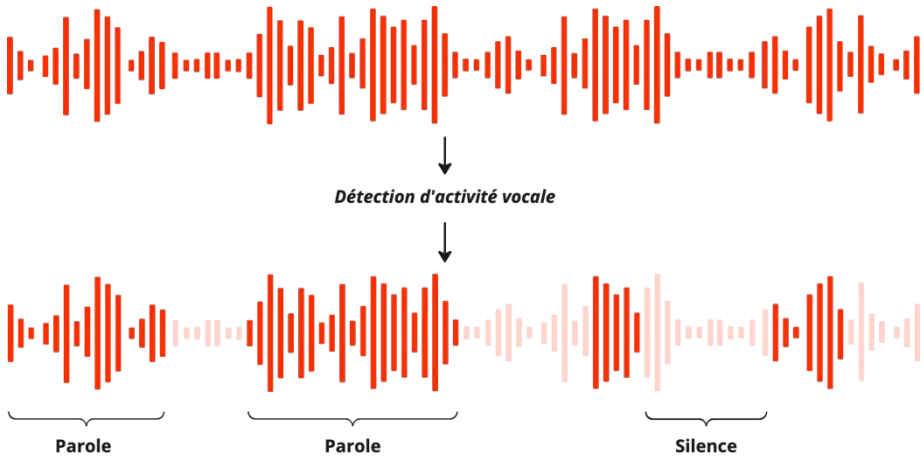

Figure 9 : Schéma représentant la tâche de détection d'activité vocale

Parmi les méthodes les plus récentes, on retrouve deux catégories de systèmes : les classificateurs binaires et les classificateurs multi-classes.

Dans le premier cas, le système de VAD cherche seulement à identifier les parties de l'enregistrement qui contiennent de la parole. Toutes les autres zones sont catégorisées comme n'étant pas de la parole. Le choix est donc binaire. C'est le cas des approches présentées dans (Bredin et al., 2019) ou (Ravanelli et al., 2021) mais aussi des systèmes les plus anciens déployés pour la téléphonie fixe et mobile (Freeman et al., 1989).

Dans le deuxième cas, celui du classificateur multi-classe, on se permet de détecter d'autres catégories sonores, dans le but d'obtenir un système plus robuste aux bruits en les caractérisant plutôt qu'en les qualifiant de parole, comme cela peut-être le cas avec

les classificateurs binaires. Ces systèmes sont souvent entraînés sur de larges bases de données audio comme Audio Set (Gemmeke et al., 2017), comme par exemple la GP-VAD introduite dans (Dinkel et al., 2020).

I.2. Segmentation

S'en suit l'étape de segmentation lors de laquelle on cherche habituellement à isoler des morceaux de l'enregistrement audio qui ne contiennent la voix que d'un seul locuteur. L'heuristique souvent admise pour l'étape de la segmentation est qu'un segment suffisamment petit est nécessairement homogène. En pratique, cette heuristique est discutable puisqu'on observe fréquemment des phénomènes de superposition des voix (*overlapping speech* en anglais), notamment lorsque les locuteurs débattent avec intensité, ce qui constitue un scénario assez fréquent en télévision. Ces chevauchements sont particulièrement complexes à traiter et sont encore source d'erreurs pour les systèmes de diarisation les plus modernes.

Dans les systèmes de diarisation du locuteur les plus récents, la notion de segment homogène se perd peu à peu. Si l'on considère toujours des segments de parole atomiques pour répondre aux contraintes intrinsèques des algorithmes de représentation de la parole, on attribue parfois à un même segment deux locuteurs ou plus, voire à un couple locuteur-musique ou locuteur-bruit.

On pourrait ainsi redéfinir l'étape de segmentation comme la partition d'un enregistrement audio en segments atomiques (et non plus homogènes), dont la caractéristique première est qu'ils ne peuvent porter chacun qu'un seul label valable pour l'entièreté du segment. Il s'agira de définir dans la suite de l'analyse le contenu porté ce label (silence, un locuteur, deux locuteurs, etc.)

II. Représentations vectorielles de la parole

Une fois les segments extraits, il s'agit désormais de trouver une représentation adaptée à leur regroupement par locuteur. De très nombreuses méthodes existent pour générer ces représentations vectorielles à partir d'un segment de parole atomique. La plupart d'entre elles tirent parti des architectures récentes de réseaux neuronaux. On peut citer par exemple les représentations issues de l'architecture wav2vec (Baevski et al., 2020) utilisées dans (Mishra et al., 2023). Beaucoup de ces méthodes ont initialement été introduites pour la tâche de vérification du locuteur. Dans les pages suivantes, nous énumérons les principales méthodes.

Pour extraire la majorité de ces représentations, un consensus a été trouvé dans le domaine de la diarisation autour de l'utilisation de coefficients cepstraux en échelle mel (MFCCs pour *mel-frequency cepstral coefficients*) présentés par Davis et Mermelstein en 1980 (Davis and Mermelstein, 1980).

L'échelle mel a pour avantage d'appliquer une transformation non linéaire aux fréquences du signal, de telle sorte qu'on obtienne davantage de précision pour les basses fréquences qui correspondent à la voix humaine. On peut passer d'une fréquence traditionnelle *f* à une valeur *m* en échelle mel grâce à la formule suivante :

$$m = 2595 log_{10}(1 + \frac{f}{10})$$

II.1. I-vecteurs

Les i-vecteurs apparus pour des problématiques de reconnaissance et de vérification du locuteur (Dehak et al., 2011; Kenny et al., 2007) permettent de représenter avec efficacité les caractéristiques de la voix d'un locuteur dans un enregistrement donné. Pour ce faire, il est tout d'abord nécessaire de passer par une représentation de la parole par des supervecteurs, que l'on nommera *M*, composés des coefficients multiplicateurs des modèles de mélange gaussiens (*gaussian mixture models* – GMM). Cette décomposition visant à approcher le signal est souvent obtenue grâce à l'algorithme EM (*expectation-maximization*).

Le i-vecteur *w*, par définition, peut se décomposer en deux termes comme suit :

$$M = m + Aw$$

Avec *m* un terme représentant l'*universal background model* (UBM) et *A* la matrice de variabilité contenant notamment les informations relatives au canal audio.

Pour la diarisation, les i-vecteurs ont souvent été combinés avec une analyse discriminante liéaire (LDA) (Sell and Garcia-Romero, 2014) et avec le critère d'information bayésien (BIC) qui sert de condition d'arrêt lors de la phase de clustering. Toutefois pour la tâche de diarisation, les i-vecteurs souffrent d'un manque de robustesse lié au fait que l'on travaille souvent à l'échelle de segments atomiques de courte durée (inférieure à 1 seconde). Si certaines limitations du critère BIC peuvent être réduites par une modification de certains hyperparamètres (Stafylakis et al., 2011), cette méthode reste limitée en contexte réel.

Toutefois, récemment, les i-vecteurs ont refait une apparition remarquée dans la communauté scientifique de la diarisation acoustique. En effet, ils sont au cœur d'une

des approches les plus performantes de l'état de l'art récent : la détection d'activité vocale centrée sur le locuteur (*target-speaker voice activity detection* – TSVAD) (Medennikov et al., 2020).

Méthode : La détection d'activité vocale centrée sur le locuteur (TS-VAD)

Cette méthode, s'inspirant de l'approche de détection d'activité vocale personnelle (Ding, S. et al., 2020), vise, après une première passe de diarisation effectuée par un algorithme *baseline*, à extraire pour chaque locuteur identifié un modèle de sa voix basé sur l'extraction de i-vecteurs. S'en suit un nombre variable d'itérations lors desquelles on vient affiner les résultats de diarisation grâce aux modèles extraits, avant d'améliorer à nouveau, de façon cyclique, les modèles avec ces nouveaux résultats (cf. Figure 10 pour un exemple à 4 locuteurs, les étapes encadrées en pointillés sont répétés N fois).

Cette approche, bien qu'elle soit longue et coûteuse, a montré des performances quasiment inégalées lors des dernières compétitions de diarisation (Wang, W. et al., 2021, 2022a).

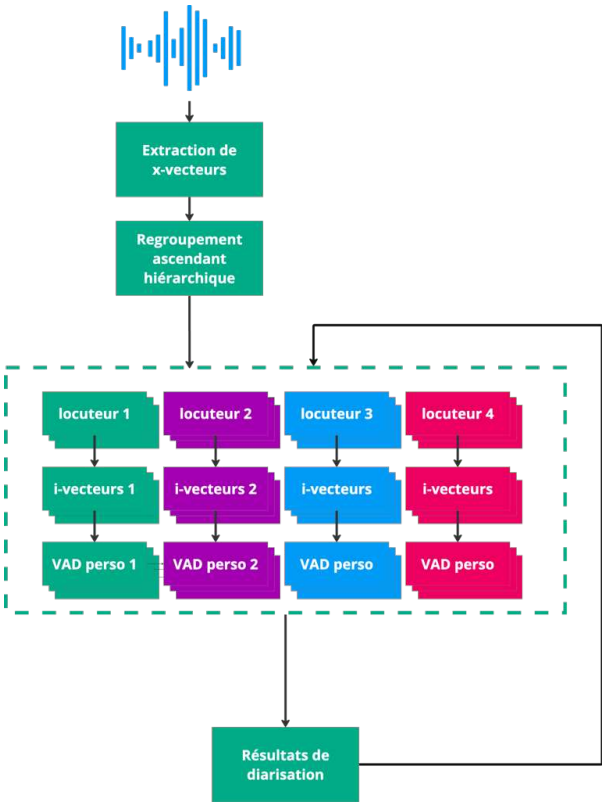

Figure 10 : Schéma représentant la méthode itérative de détection d'activité vocale centrée sur le locuteur

II.2. D-vecteurs

Après de nombreuses itérations avec les méthodes de représentation par GMMs (Imseng and Friedland, 2010), et, avec l'essor des réseaux de neurones profonds, de nouvelles approches de représentation vectorielle de la parole sont nées (Cyrta et al., 2018; Garcia-Romero et al., 2017). Parmi elles, on retient en particulier l'approche dite des d-vecteurs[8] (Variani et al., 2014).

Comme les i-vecteurs, ces représentations vectorielles abstraites ont été créées initialement pour la tache de vérification du locuteur. Des réseaux de neurones profonds sont donc entraînés à générer des vecteurs proches pour deux segments de parole dont les voix sont proches. Ce qui, *a fortiori*, permet de vérifier l'identité d'un locuteur.

L'architecture du réseau est la suivante : en entrée du réseau on utilise 40 filtres énergétiques calculés pour chaque échantillon audio (souvent choisis d'une durée d'environ 40 ms). Suivent ensuite plusieurs couches de neurones entièrement connectés avec un taux de désactivation (*dropout rate*) souvent égal à 0,5. Ce qui permet, en particulier, de se prévenir de phénomènes de surapprentissage des caractéristiques d'un locuteur particulier.

De nombreuses approches récentes de diarisation s'appuient sur les d-vecteurs. On peut citer par exemple les systèmes de diarisation par LSTM (Wang, Q. et al., 2018), ou encore le système UIS-RNN (*unbounded interleaved state recurrent neural network*) (Zhang, A. et al., 2019), tous deux présentés par les équipes de Google en 2018.

Méthodes : Diarisation par LSTM & UIS-RNN

Pour ces deux méthodes, le principe d'extraction de représentations est le même et s'appuie sur les d-vecteurs. Un réseau LSTM extracteur de d-vecteurs est entraîné sur les bases VoxCeleb 1 et 2 (Nagrani et al., 2017, 2020) à extraire des représentations vectorielles de fenêtre de 25ms toutes les 10ms à partir de 40 filtres énergétiques en échelle mel.

S'en suivent ensuite respectivement pour la méthode dite de diarisation par LSTM et le système UIS-RNN, un regroupement par la méthode des K-moyennes et un regroupement obtenu en quasi temps réel et construit à partir d'un RNN et du procédé de calcul de distances dit du restaurant chinois (Blei and Frazier, 2011).

8. Appelés ainsi en écho aux ***d****eep neural networks* qui servent à les calculer

II.3. C-vecteurs

On peut également mentionner les c-vecteurs (Sun et al., 2021), moins répandus, mais qui permettent une amélioration des performances de reconnaissance du locuteur et de diarisation en fusionnant habilement plusieurs d-vecteurs, calculés avec des paramètres différents. En particulier, on vient fusionner des représentations vectorielles de segments issus de deux systèmes de segmentation différents. Le premier, plutôt classique, inclut une détection d'activité vocale alors que le second se dote également d'un réseau de neurones, dont le but est de détecter les signes d'un changement de locuteur (*change point detection* – CPD).

Les auteurs de cette méthode proposent ensuite plusieurs façons de combiner les deux représentations calculées en parallèle jusqu'à obtenir 4 % de gain de DER (métrique principale utilisée pour évaluer la diarisation, cf. Partie II, Chapitre IV-I.1) sur le corpus AMI.

II.4. X-vecteurs

En appliquant toujours le même principe d'apprentissage des caractéristiques du locuteur, et quelques années après l'introduction des d-vecteurs, sont apparus les x-vecteurs, décrits notamment comme plus robustes au bruit.

De la même façon que pour les d-vecteurs, on va générer une représentation vectorielle au niveau de l'échantillon audio mais, dans le cas du x-vecteur, une couche dite de sous-échantillonnage (*pooling layer*) temporel est ajoutée au réseau extracteur permettant de créer cette fois un vecteur de représentation de la parole sur un segment et non plus à l'échelle de l'échantillon.

Ainsi, par rapport aux d-vecteurs on gagne en robustesse grâce à la moyenne faite sur un segment pour l'extraction des x-vecteurs, mais on perd en précision puisqu'il n'est plus possible de détecter un changement de locuteur à l'échelle de l'échantillon (Snyder et al., 2018).

II.4.1. pyannote

D'autres méthodes de représentations vectorielles robustes de la parole fonctionnent grâce à des représentations vectorielles des locuteurs calculées sur de plus longs segments.

Méthode : pyannote

C'est le cas de la méthode pyannote[9], introduite dans (Bredin et al., 2020) et améliorée dans (Bredin, 2023). Celle-ci calcule des représentations vectorielles sur des segments longs, et garantis homogènes par une détection de paroles superposés. L'architecture du réseau extracteur, PyanNet est décrite dans (Bredin et al., 2020). Ces représentations sont ensuite regroupées au moyen d'un algorithme de regroupement ascendant avec un critère basé sur les centroïdes.

II.5. ECAPA TDNN

En plus des x-vecteurs, d'autres méthodes de représentations robustes de la parole sont récemment apparues. On peut en particulier citer les ECAPA (*Emphasized Channel Attention, Propagation, and Aggregation*) TDNN. En effet, celles-ci sont également générées par un TDNN (*time-delay neural network*), que les auteurs de (Desplanques et al., 2020) ont combiné à différents blocs SE (*Squeeze Excitation*), dont l'architecture est illustrée en Figure 11, pour extraire des données audio des éléments de contexte sensiblement utiles à la génération de représentations vectorielles discriminantes.

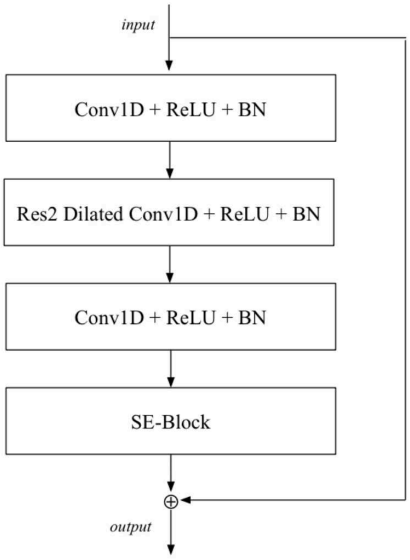

Figure 11 : Bloc SE-Res2 présenté dans (Desplanques et al., 2020) pour l'architecture ECAPA-TDNN

9. Outil *open-source* : https://pyannote.github.io/

Appliquées à la diarisation du locuteur (Dawalatabad et al., 2021), ces nouvelles représentations vectorielles permettent d'atteindre 3,01 % de DER sur le jeu d'évaluation du corpus AMI (Cf. Chapitre II-II.2, p. 49).

III. Méthodes de regroupement pour la diarisation

Une fois la bonne représentation trouvée et générée pour chaque segment de parole, il s'agit de regrouper ces segments selon le locuteur qui les a prononcés, afin de former des groupes, ou *clusters*, dans lesquels chaque segment a été prononcé par la même personne.

III.1. Regroupement agglomératif hiérarchique

Le regroupement agglomératif hiérarchique (*agglomerative hierarchical clustering* - AHC) est une méthode de regroupement utilisée dans de nombreuses méthodes de diarisation du locuteur. C'est une méthode de regroupement ascendante (dite *bottom-up*) introduite en 2003 dans (Ajmera and Wooters, 2003).

En pratique, la phase de regroupement est initialisée avec un groupe par segment, puis à chaque étape de l'algorithme, les segments proches sont rassemblés de façon itérative selon un critère de distance.

On peut notamment citer le critère BIC (*Bayesian Information Criterion*) (Zhu et al., 2005) les métriques Kullback-Leibler (Rougui et al., 2006) ou encore le ratio GLR (*Generaled Likelihood Ratio*) (Tsai et al., 2004) fréquemment utilisés comme critères d'arrêt du regroupement des segments.

III.2. Regroupement par modèles de Markov bayésiens

On retrouve également des méthodes de regroupement par modèles de Markov cachés (*Hidden Markov Models* - HMM) dans certains systèmes de diarisation du locuteur (Diez et al., 2018; Fox et al., 2011; Meignier et al., 2000). La principale limite de ces méthodes réside dans le fait qu'il est nécessaire d'initialiser le HMM avec un certain nombre d'états, en fonction du nombre de locuteurs.

Méthode : VBx

L'approche VBx, introduite en 2019 (Diez et al., 2019) puis détaillée en 2022 (Landini et al., 2022), utilise des x-vecteurs dont l'extraction est suivie par un regroupement ascendant hiérarchique. Celui-ci sert d'initialisation à la méthode de

regroupement par HMM et permet ; entre autres, d'inférer le nombre de locuteurs actifs. Cette dernière méthode se fait au moyen d'un modèle de Markov caché ergodique[10] avec un état par locuteur un état initial non émetteur (cf. Figure 12, exemple de ce HMM pour un enregistrement avec 3 locuteurs).

Pour construire ce HMM bayésien, on établit pour chaque locuteur *k* une probabilité π_k qu'il prenne la parole, ainsi qu'une probabilité commune P_{loop} que la parole reste au locuteur actif. A titre d'exemple, on constatera fréquemment lors de l'enregistrement d'une émission en plateau, des valeurs π_p relatives aux présentateurs plus élevées que les π_i relatives aux autres intervenants puisqu'entre les différentes interventions, la parole revient fréquemment aux présentateurs comme expliqué dans (Canseco et al., 2005).

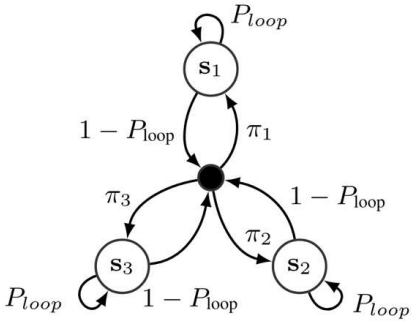

Figure 12 : Schéma du HMM utilisé lors de la phase de regroupement.
L'exemple pris dans (Diez et al., 2020) est celui d'un enregistrement avec trois locuteurs (s1, s2, s3)

Les excellents résultats de cette méthode, telle que présentée dans (Diez et al., 2020) nous poussent à la choisir comme méthode de référence dans plusieurs des travaux de cette thèse.

III.3. Regroupement spectral

Le regroupement spectral (Ng et al., 2001), non spécifique à la tâche de diarisation du locuteur, est une méthode classique de regroupement, prisée pour sa simplicité et sa rapidité. Sa particularité consiste en la réduction de la dimension des données à regrouper par le calcul des vecteurs propres de la matrice laplacienne associée. Un regroupement par l'algorithme des K-moyennes est ensuite appliqué aux lignes de la matrice nouvellement composée par ces vecteurs.

10. Dont tous les états sont liés entre eux.

La complexité temporelle du regroupement spectral est relativement faible (kn) avec n le nombre de segments de parole, et k le nombre de vecteurs propres extraits. Cela fait de cette méthode une des solutions de regroupement les plus prisées pour des diarisations *online*, c'est-à-dire pouvant être exécutées en quasi temps-réel.

III.4. Regroupement en quasi temps réel

En effet, parmi toutes les méthodes de regroupement, certaines présentent l'avantage notable de pouvoir être exécutées en quasi-temps réel à mesure que l'audio est décodé. C'est notamment le cas de la méthode UIS-RNN présentée précédemment, du regroupement spectral ou encore du regroupement incrémental présenté dans (Coria et al., 2021).

Méthode : Diart

La méthode Diart introduite en 2021, permet une diarisation en « live » ou plutôt une diarisation dite en-ligne avec une latence réglable jusqu'à 5 secondes. Cette méthode s'appuie sur les modèles de segmentation et d'extraction des représentations vectorielles de l'outil pyannote (Bredin, 2023; Bredin et al., 2020).

IV. Resegmentation

Après l'étape de regroupement, certains systèmes de diarisation du locuteur opèrent une nouvelle étape de segmentation qui vise à raffiner les segments de paroles identifiés en début de chaîne, cette fois en tenant compte des groupes obtenus dans la phase de regroupement.

Cette phase dite de resegmentation se fait par exemple au moyen de l'algorithme d'optimisation de Viterbi.

L'étape de resegmentation permet dans certains cas de tenir compte d'informations issues d'une détection de paroles superposées (Bredin and Laurent, 2021; Bullock et al., 2019) et d'améliorer ainsi drastiquement les performances de systèmes de diarisation qui ne traitaient originellement pas ce cas.

V. Diarisation bout-en-bout

Comme expliqué précédemment, beaucoup d'approches de diarisation du locuteur adoptent des architectures modulaires, dîtes en chaîne de composants. Toutefois, il est important de noter que des méthodes bout-en-bout (*end-to-end*) existent également et gagnent en popularité notamment grâce à la simplicité de leur implémentation (Dissen et al., 2022; Horiguchi et al., 2023; Izquierdo Del Alamo et al., 2022; Lai et al., 2021).

Leur fonctionnement est simple, ce sont souvent des réseaux de neurones profonds, d'architectures différentes (RNN, CNN, Transformers) qui prennent en entrée des caractéristiques audio telles que les MFCCs, et qui produisent en sortie les résultats de diarisation robustes aux instances de paroles superposées et aux audios multi-canaux (Horiguchi et al., 2022).

Comme de nombreuses approches bout-en-bout, entraînées avec de larges bases de données, elles produisent d'excellents résultats. Cependant elles restent assez peu explicables et ne permettent pas d'avoir accès aux résultats intermédiaires.

V.1. End-to-end neural diarization

Une des premières méthode unifiant les différents modules dans un unique réseau de neurones est celle présentée dans (Fujita et al., 2019).

Méthode : EEND

La méthode EEND a résolu un problème inhérent à la diarisation du locuteur dans le contexte des méthodes d'apprentissage automatique : la permutabilité des labels associés aux différents locuteurs d'un enregistrement. En effet, s'agissant d'une identification seulement relative du locuteur, il est tout à fait correct d'appeler « Locuteur 1 » le locuteur numéro 2 et vice-versa. Ainsi la proposition nouvelle de l'article (Fujita et al., 2019) est de proposer deux nouvelles fonctions d'objectifs insensibles aux permutations des labels des locuteurs, et adaptées à l'entraînement d'un système de diarisation.

V.2. Transformers

Avec l'arrivée récente des *transformers* dans différents domaines à commencer par le traitement du langage naturel, la vision par ordinateur et plus récemment le traitement de l'audio et de la parole, la diarisation du locuteur a également vu apparaître son lot de méthodes s'appuyant sur des réseaux basés sur le principe d'attention.

Méthode : DiFormer

L'une d'elle, la méthode dite des DiFormer (Lai et al., 2021), propose une architecture dans laquelle chaque module de la diarisation est réalisé par une tête d'attention qui prend pour valeurs d'entrée les caractéristiques temporelles des locuteurs extraites par un réseau encodeur.

Cette méthode permet d'obtenir des résultats très convaincants sur la base de données VoxConverse allant jusqu'à 8.21 % de DER.

On observe ainsi que les méthodes de diarisation essentiellement acoustiques sont très diverses (cf. Tableau 1). Cela s'explique par la diversité des approches : diarisation modulaire, ou bout-en-bout et par la multiplicité des méthodes adoptées pour représenter vectoriellement la parole puis pour regrouper ces représentations. Enfin, nous pouvons noter que chacune de ces méthodes ont leurs forces par rapport aux différents problèmes de robustesse rencontrés lorsqu'on construit un algorithme de diarisation du locuteur.

Tableau 1 : Récapitulatif des principales méthodes de diarisation acoustique

Méthode	Représentation vectorielle utilisée	Particularité
UIS-RNN	d-vecteurs	Regroupement en-ligne via RNN
TS-VAD	i-vecteurs	Système itératif basés sur les VAD personelles
Pyannote 2.0	x-vecteurs modifiés (PyanNet)	Système modulaire (1 NN par module)
VBx	x-vecteurs	Double regroupement (AHC + HMM bayésien)
EEND	Aucune explicite	Regroupement multi-classe dans une approche bout-en-bout
Diarisation auto-supervisée	Audio brut	Matrice de de similarité + AHC
LIUM	GMMs / i-vecteurs	Deux regroupements (AHC)
Diart	x-vecteurs modifiés (PyanNet)	Regroupement *en-ligne*
FlexSTB (EEND-EDA)	Aucune explicite	Approche bout-en-bout, *en-ligne*, pour un nombre ∞ de locuteurs
DiFormer	Audio brut	Une tête d'attention par module de diarisation

Chapitre II : Les problèmes de robustesse de la diarisation classique

I. Principales limites des algorithmes de diarisation

I.1. Robustesse au bruit

La robustesse au bruit est certainement la catégorie la plus étudiée en ce qui concerne la robustesse des algorithmes de traitement de la parole. En effet, qu'il s'agisse de la reconnaissance de la parole ou de la diarisation, on cherche souvent à obtenir les systèmes les plus robustes possibles au bruit. A la télévision en particulier, on s'expose à des niveaux de bruit très variables : entre une émission enregistrée en studio avec un microphone par locuteur, et une interview réalisée au bord d'un terrain de football, les conditions d'acquisition de l'audio sont radicalement différentes.

Pour illustrer ce problème, prenons le jeu de test de la base de données VoxConverse et ajoutons artificiellement à chaque audio du bruit avec des niveaux de rapport signal à bruit (*signal to noise ratio* – SNR) variables. Pour ce faire on utilise un bruit blanc issu de la base de données MUSAN (Snyder et al., 2015).

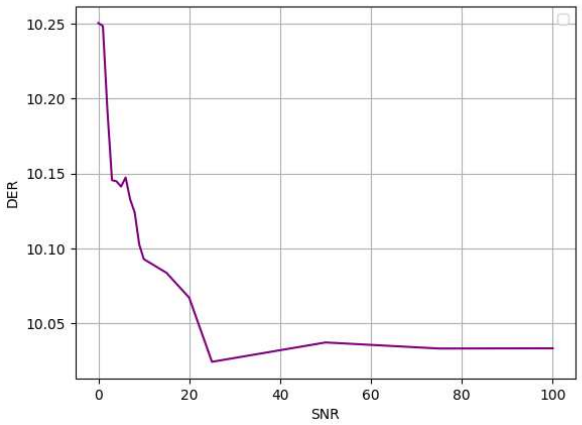

Figure 13 : Résultats de diarisation du locuteur (DER) en fonction du niveau de bruit ajouté artificiellement sur la base VoxConverse

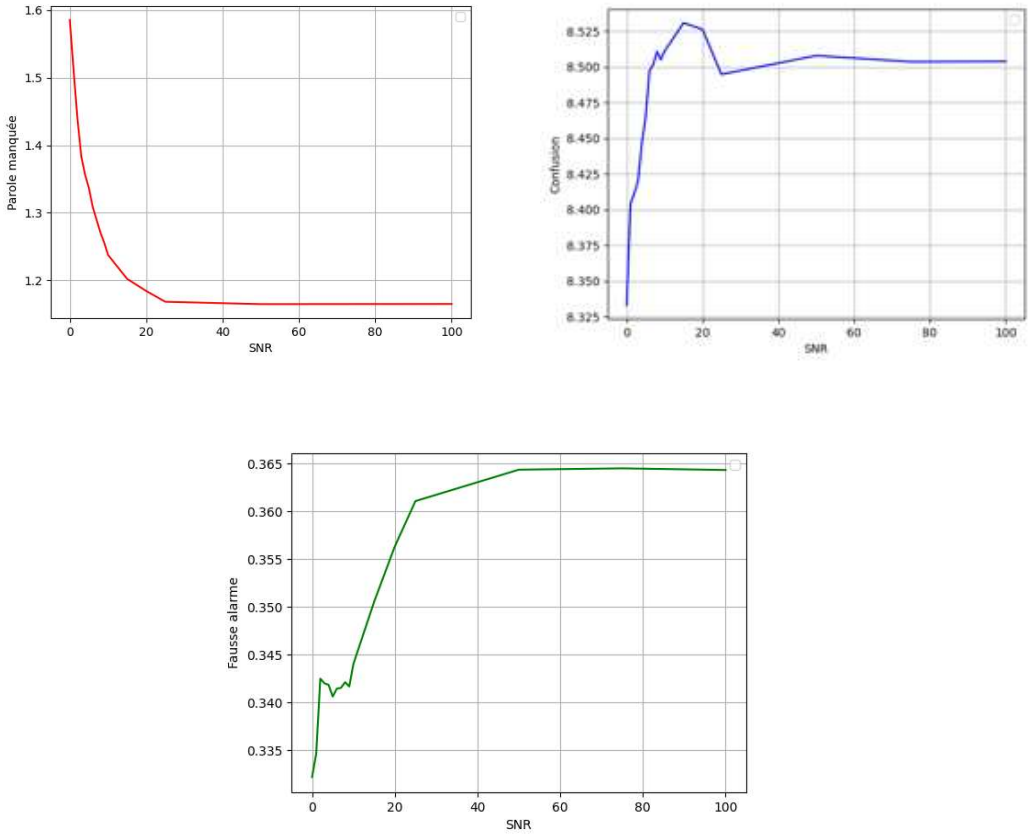

Figure 14 : Détails des résultats de diarisation en fonction du SNR

On constate logiquement qu'à mesure que l'on augmente le SNR, les résultats de diarisation s'améliorent (cf. Figure 13), avec en particulier la nette diminution de la quantité de parole manquée par la diarisation. Inversement, le taux de fausses alarmes et la quantité de confusion entre les locuteurs augmentent (cf. Figure 14), cela est vraisemblablement dû aux données d'entraînement de la méthode utilisée, elles-mêmes augmentées par des données bruitées (Bredin et al., 2020), ne permettant pas d'obtenir des performances idéales lorsque le SNR est trop élevé.

Néanmoins, les résultats observés montrent que la méthode de diarisation employée (Diart cf. Chapitre I-III.4, p. 38) est plutôt robuste aux bruits blancs ajoutés artificiellement. Cela ne permet en revanche pas de conclure sur la robustesse de cette méthode à des bruits moins réguliers, comme ceux que l'on rencontre à la télévision.

I.2. Paroles superposées

Un des principaux problèmes qui se posent lorsque l'on cherche à effectuer une diarisation du locuteur sur des données réelles est qu'il n'y a souvent pas de tours de parole distincts. Les locuteurs d'une conversation s'interrompent fréquemment les uns les autres, qu'ils s'agissent de brefs « oui » ou « non », ou de phrases plus construites. C'est particulièrement le cas à la télévision (Lebourdais et al., 2022) lors de débats politiques notamment.

Par exemple, lors du dernier débat de l'entre-deux tours des élections présidentielles de 2022 en France, on comptabilise, avec un algorithme de l'état de l'art de détection de parole superposée, environ 10 minutes pendant lesquelles candidats ou journalistes ont parlé simultanément (cf. Figure 15). Cela représente environ 5 % de la durée totale du débat.

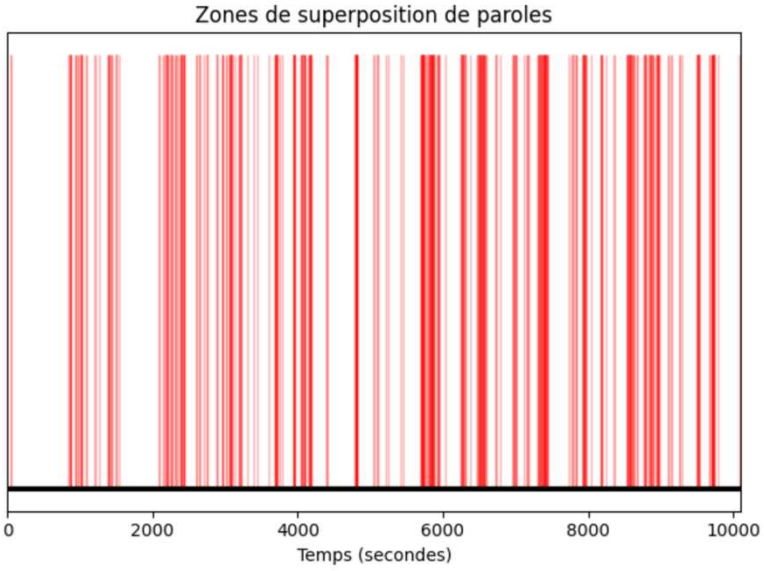

Figure 15 : Représentation des zones (en rouge) où l'algorithme de l'outil pyannote détecte des paroles superposées lors du débat télévisé entre Emmanuel Macron et Marine Le Pen du 20 avril 2022

Si les oreilles et les cerveaux humains sont bien entraînés à traiter ce type d'informations, ce n'est pas le cas des algorithmes de diarisation.

Les premiers algorithmes de diarisation du locuteur, dont certains sont encore utilisés aujourd'hui, font tout simplement comme si le problème n'existait pas. En effet, fidèles au modèle du segment homogène, ils considèrent qu'une seule personne peut parler à un moment donné, ce qui introduit un pourcentage incompressible d'erreurs de diarisation.

Par exemple, lorsque l'on considère le jeu d'entraînement du corpus vocal AMI *MixHeadset* et que l'on part de l'hypothèse que deux locuteurs ne peuvent pas parler en même temps, il devient mécaniquement impossible d'obtenir un score inférieur à 8 % de DER. Toutefois, de nombreuses études ont aussi montré qu'avec des systèmes adaptés à la gestion de l'*overlap*, ces scores d'erreurs peuvent être drastiquement réduits (Boakye et al., 2008).

C'est pourquoi, les méthodes plus récentes de diarisation tiennent compte de cette problématique en incluant dans leur chaîne de traitement des modules de détection de paroles superposées (Bullock et al., 2019; Otterson and Ostendorf, 2007). Ces derniers sont de plus en plus nombreux (Cornell et al., 2020; Geiger et al., 2012) à mesure que leur rôle devient essentiel pour améliorer les performances des systèmes récents de diarisation.

Méthode : DOVER-lap

La méthode DOVER-lap (Raj et al., 2021) vise à combiner différents résultats de diarisation grâce à une méthode de vote qui tient compte des éventuels passages où plusieurs locuteurs ont été détectés simultanément. L'approche fait suite à une précédente méthode de fusion intitulée DOVER (*diarization output voting error reduction*), elle-même inspirée de l'approche ROVER pour l'ASR. DOVER, dont le principe repose sur l'algorithme d'optimisation dit algorithme hongrois (Kuhn, 1955), ne tenait pas compte des passages de paroles superposées.

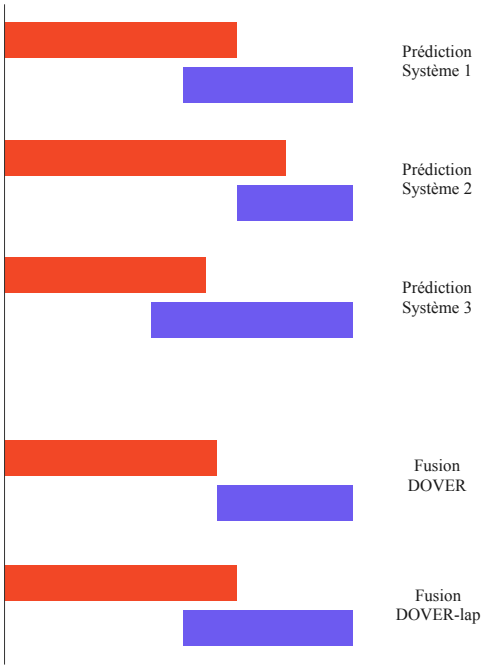

Figure 16 : Schéma des résultats obtenus grâce aux fusions DOVER et DOVER-lap de trois systèmes de diarisation du locuteur

Cette méthode, dont on présente un exemple simple en Figure 16, s'est illustrée notamment en permettant de fusionner plusieurs résultats de diarisation et d'obtenir la première place du challenge VoxSRC 2022 (Wang, W. et al., 2022a).

I.3. Nombre de locuteurs

Le but de la diarisation du locuteur étant de différencier au sein d'un même enregistrement les voix des personnes s'y exprimant, on comprend que la difficulté de la tâche croit avec le nombre de locuteurs actifs. A la télévision, on distingue deux types d'émissions.

Premièrement, celles où le nombre de locuteurs est fixe, souvent connu à l'avance et relativement faible. C'est le cas des débats politiques, des commentaires sportifs, ou encore des jeux télévisés mais aussi souvent des « rushs », c'est-à-dire des morceaux de vidéos non montés provenant directement de la caméra.

Deuxièmement, celles dont le nombre de locuteurs est variable d'une émission à l'autre. C'est le cas des fictions mais surtout des émissions d'actualité telles que les journaux télévisés quotidiens. Par exemple, lors du journal télévisé de 13h sur France 2 le 13 juin 2023, 81 locuteurs différents ont pris la parole dont :

- 6 traducteurs, qui traduisaient en français les paroles de personnes s'exprimant dans une autre langue. Les deux voix superposées étant souvent audibles, cela accentue encore la difficulté de la diarisation d'un journal télévisé ;
- 1 présentatrice en plateau ;
- 10 journalistes sur le terrain ;
- 6 voix off ;
- 8 voix d'ambiances distinctes et compréhensibles ;
- 50 personnes interviewées, ou s'exprimant au sujet d'une actualité, ne prononçant souvent que quelques mots.

Ces derniers cas sont parmi les plus complexes à traiter pour la diarisation du locuteur. En effet, une grande majorité des méthodes de l'état de l'art doivent leurs performances à une restriction des algorithmes de regroupement dont on majore le nombre de groupes qu'ils doivent trouver. On trouve souvent une majoration autour de 20 locuteurs comme c'est le cas dans les implémentations *open-source* de pyannote[11] et VBx[12].

C'est pourquoi est apparue la nécessité d'avoir des méthodes adaptées à l'analyse d'enregistrements composés d'un grand nombre de locuteurs (Horiguchi et al., 2020, 2021, 2023).

Méthode : EEND-EDA

La méthode dite EEND-EDA (*end-to-end neural diarization with encoder-decoder based attractors*) est une méthode basée sur le calcul d'attracteurs (Horiguchi et al., 2020) intervenant après le calcul des représentations vectorielles de la parole (apparentées à des d-vecteurs de dimension *D*). En pratique, pour N segments de parole, les attracteurs $a_n \in \{0; 1\}^D$ sont calculés grâce à un réseau encodeur-décodeur et font ensuite office de détecteurs d'activité vocale.

Le problème de cette méthode réside dans le fait qu'elle ne permet pas de détecter plus de locuteurs que le maximum de locuteurs présents dans les enregistrements du jeu d'entraînement. Si cette limite est virtuellement modifiable, il est en réalité très compliqué d'obtenir et de labelliser un jeu de données pour la diarisation dans lequel chaque enregistrement contient un nombre important de locuteurs actifs. Cette limitation est résolue par (Horiguchi et al., 2021) avec l'introduction d'attracteurs

11. https://github.com/pyannote/pyannote-audio
12. https://github.com/BUTSpeechFIT/VBx

locaux. Finalement dans (Horiguchi et al., 2023) cette méthode est adaptée à un fonctionnement *online*, c'est-à-dire en quasi temps réel.

La particularité du système de diarisation que nous cherchons à construire dans cette thèse réside dans le fait qu'on le souhaite robuste et adapté aux cas où le nombre de locuteurs est grand, mais également lorsque seules deux ou trois personnes débattent.

II. Jeux de données adaptés à la diarisation

Il existe de nombreux jeux de données adaptés à la diarisation du locuteur, dont beaucoup sont associés à des challenges annuels, dont le but est d'évaluer les différents systèmes sur une typologie de données spécifique à ce dernier. En plus des quatre bases de données citées ci-après, on peut citer les challenges DIHARD (Ryant et al., 2019, 2021), FearlessStep (Joglekar and Hansen, 2019) ou Ego4D (Min, 2023), chacun associé à un jeu de données destiné à l'évaluation de la diarisation du locuteur.

II.1. CALLHOME

Le jeu de données CALLHOME[13] développé en 1997 par le LDC (*Linguistic Data Consortium*) est un des plus anciens *datasets* permettant l'évaluation de la diarisation du locuteur. En effet, celui-ci se compose de 120 conversations téléphoniques avec des tours de parole clairement identifiés.

Au total, 60 heures de discussions sont annotées, mais chaque enregistrement ne contient que deux locuteurs actifs.

Ainsi, si ce jeu de données est l'un des plus présents dans les différentes évaluations de diarisation publiées, ses enregistrements sont très éloignés des conditions présentes dans les contenus diffusés à la télévision.

II.2. AMI

Le corpus AMI (McCowan et al., 2005) développé par l'université d'Édimbourg contient 100 heures d'enregistrements de réunions jouées par des acteurs et enregistrées au moyen de plusieurs caméras et microphones.

13. https://catalog.ldc.upenn.edu/LDC97S42

Pour son utilisation en diarisation du locuteur on se sert souvent de la sous partie de cette base intitulée *MixHeadset,* qui correspond à un mixage de tous les microphones des différents acteurs, aligné avec la retranscription des discussions.

II.3. VoxConverse

Créée par l'université d'Oxford, VoxConverse (Chung et al., 2020) est une base de données de plus de 70 heures d'audio multi-locuteurs extraites de vidéos YouTube et annotées de façon semi-automatique. Les enregistrements audio composant ce jeu de données contiennent en particulier des débats politiques, des émissions enregistrées en plateau et des captations de grand discours.

Tableau 2 : Caractéristiques des deux sous-ensembles du jeu de données VoxConverse

Sous-ensembles	Nombre de vidéos	Nombre d'heures	Nombre de locuteurs moyen	Pourcentage moyen de parole superposée
Développement	216	20,3	4,5	3,8
Test	310	53,5	6,3	3,0

VoxConverse 0.3[14] est un jeu de données particulièrement difficile en raison de la présence de nombreux passages de paroles superposées (cf. Tableau 2), et de conditions de bruit variables et parfois difficiles (applaudissements, bruits de foule, musique, etc.). En outre, s'il est majoritairement en langue anglaise, on y retrouve certains enregistrements en allemand, français ou encore en arabe.

C'est par ailleurs le jeu de données qui, par sa diversité, s'avère le plus proche des cas d'usage qui intéressent Newsbridge, raison pour laquelle l'on choisira d'évaluer nos systèmes sur ces enregistrements.

II.4. CHiME

La base de données CHiME issue du challenge du même nom est un jeu d'enregistrements audio capturés par plusieurs microphones dans le contexte particulier du « dinner party scenario ». En pratique, il s'agit de plusieurs personnes qui jouent une scène de diner entre amis avec tous les détails associés : conversations naturelles, bruits de vaisselle, rires, chuchotements, cris, etc. Cela fait de cette base l'une des bases les plus difficiles pour la tâche de diarisation du locuteur. L'analyse de ce corpus requiert d'ailleurs fréquemment des méthodes de prétraitement spécifiques (Boeddecker et al., 2018).

14. https://github.com/joonson/voxconverse/commit/48312391280c284a82a616899892b2efa948eadb

Pour chaque scène jouée, on dispose de plusieurs enregistrements effectués au moyen de différents microphones dont le placement dans l'appartement est également labellisé.

Cette base existe en plusieurs versions, correspondant aux différentes itérations du challenge. Dans notre cas, lorsque l'on se réfère à cette base de données il s'agit de sa version 5.

II.5. Corpus en langue française

La recherche sur la diarisation du locuteur a également été marquée par différents corpus en langue française tels que les bases de données ESTER1 (Galliano et al., 2005 ; Gravier, G et al., 2004) et ESTER2 (Galliano et al., 2009), popularisées par les campagnes d'évaluation « Évaluation de Systèmes de Transcription enrichie d'Émissions Radiophoniques » (ESTER).

Initiées en 2004, elles visaient à évaluer les systèmes de transcription automatique pour les actualités radiophoniques en français. Les tâches d'évaluation comprenaient la transcription orthographique, la détection d'événements particuliers (musique, applaudissements, etc.) et bien sûr la reconnaissance des locuteurs.

Plus récemment, le projet ANR REPERE (REconnaissance de PERsonnes dans des Emissions audiovisuelles) lancé en 2010 a lui aussi permis plusieurs campagnes d'évaluation sur un corpus éponyme. Ces travaux ont ensuite mené à la publication du corpus ETAPE en 2012 (Gravier, Guillaume et al., 2012), fournissant 30 heures de vidéos diffusées à la télévision française, annotées notamment pour la diarisation.

Si certains de ces corpus semblent appropriés dans le cadre de nos travaux, nous leur avons néanmoins préféré leurs équivalents en langue anglaise, voire dans le cas de VoxConverse un corpus multilingue, plus alignés avec les besoins stratégiques de Newsbridge.

III. Contribution : Détection d'activité vocale multi-flux pour la diarisation

Se situant en début de chaîne de traitement, la détection d'activité vocale joue nécessairement un rôle prédominant dans l'efficacité de toute diarisation. Forts de cette réflexion, nous avons souhaité combiner certaines des méthodes de regroupement les plus performantes avec des méthodes récentes de VAD, afin d'étudier l'impact de ce

prétraitement nécessaire sur les systèmes les plus performants. La méthode résultant de la combinaison des VAD pour diarisation sera désignée comme la détection d'activité vocale multi-flux (MSVAD – *multi-stream voice activity detection*). Cette méthode, utilisée pour la première lors d'un challenge VoxSRC (Tevissen, Boudy, and Petitpont, 2022) a été décrite dans (Tevissen, Boudy, et al., 2023). Nous détaillons dans les pages suivantes les différents système qui la composent et leur combinaison.

On constate par ailleurs dans la littérature (Desnous et al., 2018), ainsi que dans nos résultats empiriques, que la meilleure méthode de détection d'activité vocale ne produit pas systématiquement les meilleurs résultats de diarisation. Voyons désormais avec la MSVAD comment tirer parti de ce constat empirique et tentons de vérifier sa véracité sur des données issues de la télévision.

III.1. Choix des méthodes de détection d'activité vocale

III.1.1. VAD par seuil d'énergie

La détection d'activité vocale basée sur l'énergie avec un seuil d'activation est une des plus anciennes méthodes de VAD. Elle consiste à calculer pour chaque fenêtre (de taille fixée N_w) w l'énergie e_w du signal sonore s_w selon la formule suivante :

$$e_w = \sum_{t}^{N_w} |s_w(t)|^2$$

Cette méthode a été choisie comme système de référence en premier lieu car elle était recommandée par les auteurs de la méthode VBx dans (Landini et al., 2022). C'est d'ailleurs la méthode incluse dans l'implémentation *open-source*[15] de cet algorithme de diarisation.

III.1.2. pyannote VAD

Cette approche est exclusivement basée sur des réseaux de neurones, car elle repose sur l'architecture PyanNet décrite dans (Bredin et al., 2020) permettant une classification en K classes (ici, K=2). Cette architecture se compose de plusieurs couches convolutives pour l'extraction des *features* (Ravanelli and Bengio, 2018), puis de couches récurrentes. Le signal audio échantillonné est directement utilisé comme caractéristiques d'entrée pour un classificateur binaire. Pour cette expérience, le modèle pyannote pré-entraîné sur le corpus AMI a été utilisé.

15. https://github.com/BUTSpeechFIT/VBx

Parmi les 4 systèmes de VAD que nous considérons, celui-ci est celui qui, utilisé seul, produit les meilleurs résultats.

III.1.3. Speechbrain VAD

On utilisera également la solution de détection d'activité vocale proposée par l'outil *open-source* speechbrain (Ravanelli et al., 2021). A l'image de pyannote, cette méthode utilise un modèle, issu d'un réseau CRDNN, pré-entraîné[16] pour détecter la présence ou non de voix. Le modèle choisi a été entraîné sur la base de données Libriparty[17], dérivée de Librispeech (Panayotov et al., 2015).

III.1.4. GP-VAD

Pour obtenir une détection d'activité vocale plus robuste, les auteurs de (Dinkel et al., 2020) ont proposé une approche basée sur un réseau de neurones conçu avec une combinaison de couches de convolution et de couches récurrentes. Cette méthode, nommée GP-VAD (*global purpose voice activity detection*), est basée sur un schéma d'entraînement faiblement supervisé (*weakly supervised sound event detection* – WSSED). Entraînée sur le jeu de données Audioset (Gemmeke et al., 2017), cette approche apprend à détecter 527 classes de sons différents parmi lesquelles la classe parole. La grande variété de classes prédites est ensuite réduite virtuellement pour obtenir un classificateur binaire *Parole/Non-parole*.

III.2. Méthode de détection d'activité vocale multi-flux

Comme évoqué précédemment, les méthodes de détection d'activité vocale sont nombreuses dans l'état de l'art et il s'est avéré complexe de déterminer laquelle serait la plus adaptée à la diarisation.

En effet, on constate de façon empirique qu'une VAD jugée meilleure, selon les critères d'évaluation de la VAD, ne produit pas forcément de meilleurs résultats de diarisation. Il vaut parfois mieux accepter de se tromper de quelques centaines de millisecondes, notamment en début et en fin de tour de parole, si l'on veut obtenir une meilleure représentation vectorielle du segment de parole. Cette étape étant en début de chaîne de traitement, le gain de performances que l'on souhaite obtenir sera naturellement propagé au regroupement et permettra finalement d'obtenir une meilleure diarisation.

16. https://huggingface.co/speechbrain/vad-crdnn-libriparty
17. https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriParty/generate_dataset

C'est dans cet esprit qu'une partie de notre travail s'est orientée vers la recherche d'une VAD optimale pour la diarisation du locuteur.

Une fois plusieurs VAD testées, nous est venue l'idée de fusionner les résultats de plusieurs VAD dans l'espoir de tirer parti des résultats de chacune dans les situations où elles sont les meilleures.

Pour ce faire, nous nous sommes inspirés des travaux d'Hervé Bourlard (Misra et al., 2003; Misra and Bourlard, 2005) en reconnaissance automatique de la parole qui s'était servi d'un système de décision basée sur l'entropie de plusieurs systèmes classificateurs fonctionnant en parallèle. Nous détaillons ces travaux dans (Tevissen, Boudy, et al., 2023).

Appliquée à la VAD, voici, en Figure 17, ce à quoi ressemble notre système de décision basé sur l'entropie.

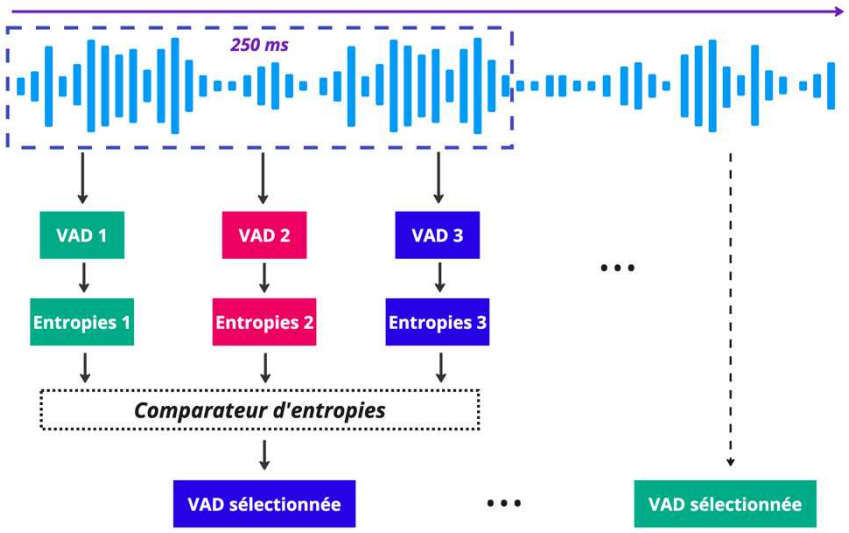

Figure 17 : Schéma décrivant le processus de choix dynamique de la méthode de détection d'activité vocale basé sur l'entropie

Chacun des trois algorithmes de détection d'activité vocale sélectionnés fonctionnent grâce à un réseau de neurones profond. Après les avoir fait fonctionner en inférence, on obtient pour chaque système les probabilités de présence de parole qu'ils ont prédites (cf. Figure 18).

On sélectionne ensuite comme taille de fenêtre temporelle de traitement minimale le plus petit commun multiple des trois niveaux de précision des systèmes (cf. Tableau 3).

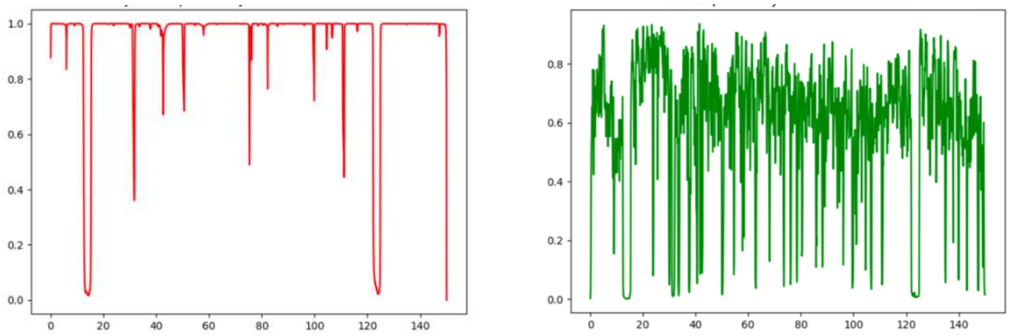

Figure 18 : Résultat de VAD obtenus grâce aux systèmes GP-VAD (vert) et pyannote (rouge) sur un enregistrement de VoxConverse

En pratique on travaille avec les tailles de fenêtre suivantes :

Tableau 3 : Tailles des fenêres temporelles par chaque système de VAD utilisé

Pyannote VAD	250 ms
Speechbrain VAD	250 ms
GP-VAD	20 ms
PPCM utilisé	250 ms

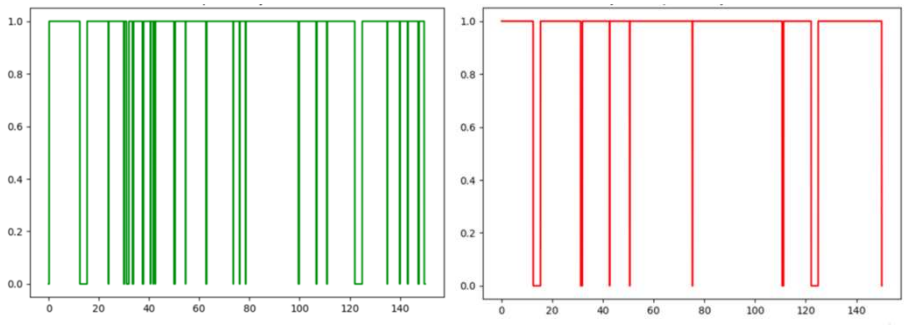

Figure 19 : Résultat de VAD obtenus grâce aux systèmes GP-VAD (vert) et pyannote (rouge) après application de leurs seuils respectifs.

Sur chaque fenêtre, on prend la moyenne des probabilités de chaque classificateur et on calcule l'entropie locale $h_{k,i}$ associée grâce à la formule suivante :

$$h_{k,i} = -P\left(Speech\middle|o_{k,i}\right) log_2\, P\left(Speech\middle|o_{k,i}\right) - P\left(Non\ Speech\middle|o_{k,i}\right) log_2\, P\left(Non\ Speech\middle|o_{k,i}\right) \quad (1)$$

Avec i indice de la fenêtre temporelle, et k l'indice du classifieur pour l'observation notée $o_{k,i}$.

On observe alors notamment, comme illustré en Figure 20, que certains réseaux ont des tendances à l'incertitude plus marquées que d'autres. En particulier le système GP-VAD mais ceci s'explique par le fait qu'initialement il s'agit d'un classificateur à 527 classes et non à 2 classes comme le sont la plupart des VADs.

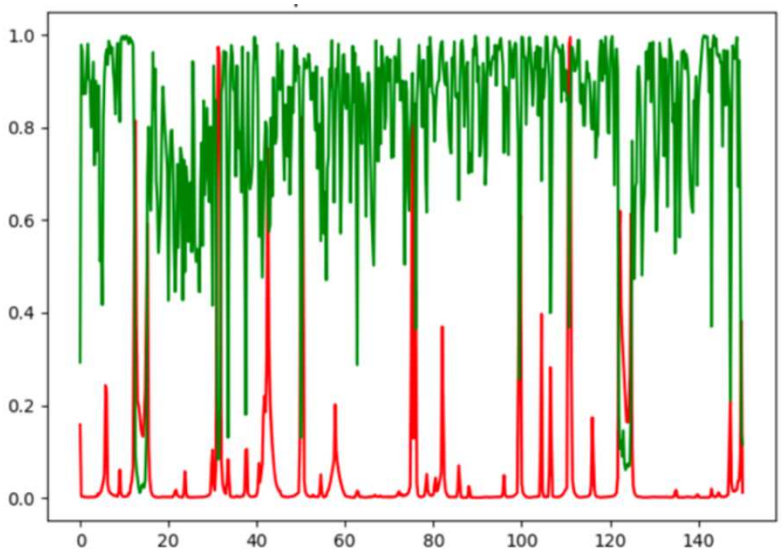

Figure 20 : Entropies des deux systèmes de VAD

Une fois les différentes entropies calculées, pour chaque fenêtre on sélectionne le classificateur k_{opt} avec l'entropie la plus faible, i.e. le plus sûr de lui. On choisit alors pour cette fenêtre i la valeur $o_{k_{opt},i}$ qui servira de VAD pour la suite du système.

Chacun des classificateurs ayant son fonctionnement propre, il faut ensuite déduire des probabilités obtenues un résultat binaire (parole / non-parole) en appliquant un seuil propre à chaque classificateur (cf. Figure 19 et Tableau 4).

Tableau 4 : Seuils utilisés pour chaque système de détection d'activité vocale. Lorsque x remplit les conditions ci-dessous on considère que le segment étudié contient de la parole

Pyannote VAD	Activation : $x > 0,767$; Désactivation : $x < 0,713$
Speechbrain VAD	Activation : $x > 0,500$; Désactivation : $x < 0,250$
GP-VAD	Activation : $x > 0,150$; Désactivation : $x < 0,100$

Pour chacun de ces seuils, des valeurs optimales ont été trouvées sur le jeu de données de développement de la base VoxConverse.

Ainsi pour obtenir le résultat final de VAD (cf. Figure 21) qui sera utilisé pour la suite de l'analyse, on sélectionne parmi les résultats des différents systèmes, celui dont l'entropie est la plus faible.

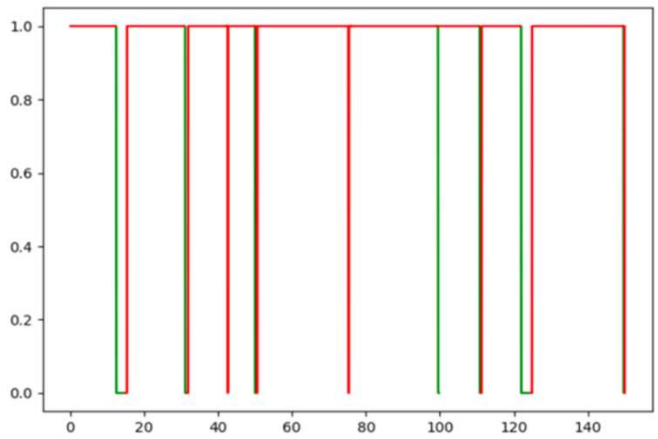

Figure 21 : Résultats fusionnés de VAD (pyannote en rouge et GP-VAD en vert) après appliction des seuils

Ensuite, on applique la suite de l'algorithme de diarisation du locuteur en extrayant les x-vecteurs, puis en appliquant le regroupement VBx (décrit en Partie I, Chapitre 1-III.2). On obtient donc finalement les résultats de diarisation comme présenté en Figure 22.

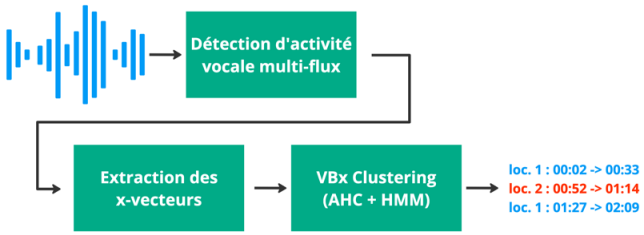

Figure 22 : Schéma global du système de diarisation du locuteur mis en place

III.3. Performances de la méthode

On choisit ici de rapporter les résultats de cette méthode en DER et JER, métriques principales de la diarisation du locuteur, afin de mesurer l'impact des détections d'activité vocale sur la tâche de diarisation du locuteur, et ce indépendamment de leurs performances propres en tant que VAD. Nous présentons toutefois ces derniers en fin de section afin d'amener une discussion.

En bref, le DER correspond à la somme de toutes les erreurs possibles de diarisation (parole manquée, fausse alarme, et confusion entre locuteurs). Le JER, quant à lui, vient normaliser les scores de parole manquée et de fausse alarme pour chaque locuteur indépendamment. Ces deux métriques sont présentées en détails en page 83.

III.3.1. VoxCeleb Speaker Recognition Challenge

Cette méthode de VAD multi-flux a été utilisée dans le cadre du VoxCeleb Speaker Recognition Challenge (Huh et al., 2023) à l'occasion duquel nous avons présenté notre système (Tevissen, Boudy, and Petitpont, 2022).

Si nous n'avons pas pu rivaliser avec les meilleurs systèmes, qui utilisaient plusieurs algorithmes de diarisation en parallèle (dont la TS-VAD) avant d'en fusionner les résultats avec l'approche DOVER-lap, nous avons cependant obtenu de bons résultats avec notre méthode de VAD multi-flux (cf. Tableau 5) notamment avec un score de JER (cf. Partie 2, Chapitre IV - I. p. 83) inférieur à la 3ème meilleure méthode du challenge.

Tableau 5 : Résultats de diarisation obtenus lors du challenge VoxSRC 2022

	DER	JER
DKU-SMIIP (Wang, W. et al., 2022b)	4.745	27.847
Kriston AI (Cai et al., 2022)	4.866	25.488
GIST-AiTeR (Park, D. et al., 2022)	5.120	30.815
MS VAD (Tevissen, Boudy, et al., 2023)	6.622	29.006
VGG *Baseline*	19.602	41.428

III.3.2. Cas général

Malgré les résultats encourageants obtenus lors du challenge VoxSRC, on constate que lorsque l'on applique notre méthode de VAD multi-flux à l'ensemble de la base de données de test de la base VoxConverse, les résultats ne conduisent pas à une

amélioration des performances, comparativement à la même analyse effectuée en n'utilisant qu'une seule détection d'activité vocale (cf. Tableau 6).

En effet, la méthode consistant à associer la détection d'activité vocale de l'outil pyannote à un regroupement VBx reste meilleure dans le cas général, que ne l'est notre système commençant par une VAD multi-flux avec 1,1 point de DER en moins.

Tableau 6 : Résultats détaillés de la diarisation sur le jeu de test de la base VoxConverse

VAD utilisée	DER	PM	FA	CF
Énergie	22.58	10.34	7.30	4.94
GP-VAD	9.76	3.78	2.30	3.68
speechbrain	9.94	2.43	3.74	3.76
pyannote	6.66	3.09	0.79	2.78
MSVAD	7.76	2.51	1.88	3.36

En analysant les résultats, on constate assez clairement un phénomène de disparité des résultats avec environ 20 % des enregistrements pour lesquels la MSVAD conduit à l'obtention d'un DER supérieur à 20 (cf. Figure 23). Ces quelques enregistrements sont donc responsables de la majorité des erreurs rencontrées. Inversement, environ 80 % des enregistrements ne sont responsables que de 20 % des erreurs de diarisation.

On constate bien ces résultats dans le diagramme de Pareto ci-après qui nous dit qu'en résolvant les DER des quelques médias représentés à gauche de la courbe, on réduirait les scores de DER de la MSVAD d'environ 80 %.

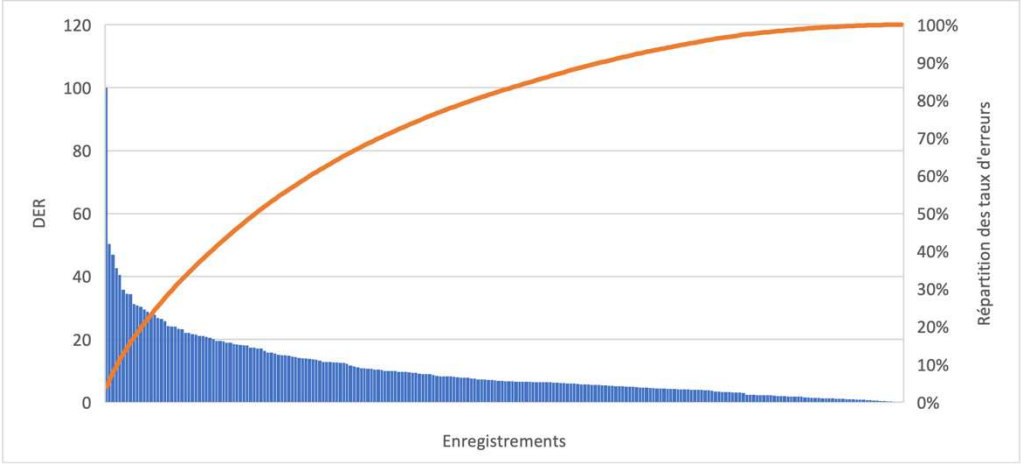

Figure 23 : Diagramme de Pareto des résultats en DER de la méthode MSVAD sur le jeu de test de VoxConverse

III.3.3. Sous ensemble de VoxConverse

On constate néanmoins que pour une partie des fichiers sur lesquels l'étude a été réalisée, 30 enregistrements, la méthode MSVAD a permis d'améliorer assez nettement les scores de diarisation avec environ 2 points d'amélioration du DER (cf. Tableau 7). On appellera ce sous-ensemble *V+*.

Par ailleurs, sur ces mêmes enregistrements, on constate que la MSVAD permet un décompte plus précis du nombre de locuteurs, comme illustré par les résultats du Tableau 8. Cela permet de montrer que cette méthode n'est pas systématiquement moins performante. On constate en effet qu'elle est dans certains cas meilleure, bien qu'elle reste en moyenne moins performante.

Tableau 7 : Résultats détaillés de la diarisation sur un sous-ensemble du jeu de test de la base VoxConverse

VAD utilisée	DER	PM	FA	CF
pyannote	8.21	3.92	0.66	3.63
MSVAD	6.20	2.30	1.50	2.40

Tableau 8 : Pourcentages de bonne détection du nombre de locuteurs présents sur un sous-ensemble du jeu de test de la base VoxConverse

Nombre de locuteurs	1	2	3	4	5	>5
Pyannote	66.7	88.9	42.9	66.7	33.3	89.5
MSVAD	66.7	100.0	71.4	100.0	66.7	94.7

Un des premiers constats que nous avons fait sur ce sous ensemble est qu'en moyenne la VAD multi-flux présente de meilleurs résultats que pyannote sur la tâche de détection de parole seule, malgré des résultats non systématiquement meilleurs sur la diarisation. Pour ce faire, on a calculé sur le corpus de test de VoxConverse les performances des quatre VAD qui nous intéressent (cf. Tableau 9). Pour ce faire, on utilise le F1-score, qui s'exprime pour une classification binaire comme :

$$F1 - score = \frac{VP}{VP + \frac{1}{2}(FN + FP)}$$

Avec, respectivement VP, FN et FP les valeurs de vrais-positifs, faux-négatifs et faux positifs.

On observe également que la VAD issue de l'outil Speechbrain propose systématiquement des meilleurs résultats pour la première étape mais des résultats moins bons en terme de diarisation en fin de chaîne de traitement. Cela nous permet de confirmer avec certitude que la meilleure VAD n'est pas toujours la plus adaptée pour la tâche de diarisation du locuteur.

Tableau 9 : Résultats des quatre méthodes utilisées en F1-score pour la tâche de détection d'activité vocale

F1-score (%)	MSVAD	pyannote	speechbrain	GP-VAD
VoxConverse test set	96.16	95.96	**97.63**	95.48
V+	67.36	66.74	**67.62**	67.61

Il s'agira dans de futurs travaux de déterminer dans quels cas ce comportement se produit, afin de toujours choisir la méthode qui produit les meilleurs résultats. En particulier, on cherchera un estimateur qui permettrait de généraliser ces résultats encourageants. Nous envisageons notamment de mettre en place un système basé sur une entropie pondérée ou lissée.

Chapitre III : Diarisation multimodale

Pour pallier les problèmes de robustesse décrits précédemment, les recherches dans le domaine de la diarisation du locuteur s'orientent également vers des approches multimodales. Lorsque cela est possible, au lieu de ne considérer que le flux audio, il s'agit alors de profiter des informations issues d'autres modalités pour obtenir les résultats de diarisation, c'est-à-dire les tours de parole.

De nombreuses équipes de recherche ont tenté d'ajouter d'autres modalités aux systèmes effectuant la diarisation du locuteur. On peut citer (Kumar et al., 2022) qui s'intéresse au contexte de l'audio pour choisir les hyperparamètres du système, (Flemotomos et al., 2020) qui cherche à déterminer le rôle du locuteur dans la conversation, ou encore (Friedland et al., 2009; Zhang, Y. et al., 2017) qui démontrent la possibilité de se servir de caractéristiques prosodiques et paralinguistiques inférées à l'échelle d'un segment de parole pour effectuer la diarisation. Enfin, plus spécifiques à certains cas d'usage, des méthodes comme celle présentée (Kang et al., 2020) se servent des informations de positionnement des microphones dans l'espace.

Dans le cadre de notre travail sur des données audiovisuelles issues de la télévision, on s'intéressera en particulier à ce que peuvent apporter deux modalités : l'image et le texte.

I. État de l'art des approches audio-visuelles

Parmi les approches de l'état de l'art en diarisation, on retrouve également des approches multimodales dont le but est de pallier les limites des approches acoustiques en utilisant les informations fournies par d'autres modalités. Plusieurs méthodes aux fonctionnements divers existent (Bendris, 2011; Le, D. H. N., 2019). On peut citer celles qui tentent de classifier individuellement les paramètres audios et visuels au moyen d'un SVM (Vallet et al., 2013) ou d'une PLDA (Mingote et al., 2022), ou encore celles qui visent à créer des représentations vectorielles multimodales comme DyViSE (Wuerkaixi et al., 2022).

I.1. Détection visuelle de la parole

Les principales approches de diarisation qui font intervenir la modalité visuelle (Chung et al., 2019; Ding, Y. et al., 2020) cherchent à détecter les mouvements des lèvres des personnes apparaissant à l'image afin de les faire correspondre avec un flux de parole actif.

I.1.1. Détection de locuteur actif

Afin d'enrichir la diarisation d'informations visuelles, on peut adopter une stratégie de détection multimodale du locuteur actif (*active speaker detection* – ASD) comme présenté dans (Chung et al., 2020) ou (Gebru et al., 2015).

Le systèmes d'ASD récents reposent sur des réseaux de neurones, souvent un transformer, dont la partie *encoder* prend comme entrée les paramètres audio (MFCC) et des caractéristiques visuelles extraites des visages préalablement détectés et *trackés* (Auguste, Martinet, et al., 2015; Auguste, Tirilly, et al., 2015) sur l'image. Des *features* visuelles sont ensuite extraites au moyen d'un réseau composé de différentes couches convolutives (3D Conv, ResNet18, V-TCN) ainsi que des représentations audios calculées à partir des MFCCs par un réseau ResNet34. Ce double réseau encodeur permet de mettre en place un mécanisme d'attention croisée et d'identifier des mouvements caractéristiques des lèvres qui donnent un indice sur le fait que le visage détecté est parlant ou non (cf. résultats en Figure 24).

Figure 24 : Exemple de résultat de détection du locuteur actif lors d'un débat politique. Le visage parlant est encadré en vert et l'autre visage détecté en rouge

La détection du locuteur actif est un champ de recherche riche par les nombreuses méthodes, dont l'une des plus performantes, *TalkNet* introduite dans (Tao et al., 2021).

I.1.2. Méthode de diarisation associée

A partir de cette information sur les visages parlants fournie par l'ASD, on peut construire un système de diarisation multimodale. Pour ce faire il est également nécessaire d'ajouter des règles pour la gestion des locuteurs hors-champ, qui par définition ne peuvent pas être gérés par la détection (en partie visuelle) du locuteur actif.

Toutefois ces règles peuvent s'avérer très différentes selon le type de contenu considéré (journaux télévisés, *talk-show*, commentaires sportifs, débats, etc.), rendant ainsi ces méthodes audio-visuelles difficilement généralisables.

Un système complet et modulaire de diarisation du locuteur, basé sur une détection multimodale du locuteur actif, devrait ainsi être composé d'un bloc de regroupement des visages parlants et d'un système expert s'appuyant par exemple sur une heuristique pour la gestion des locuteurs hors-champ (cf. Figure 25).

Ainsi, malgré les avantages évidents de cette approche lorsque les locuteurs actifs sont à l'image, on se rend rapidement compte que sur du contenu d'actualité (journaux télévisés notamment), le locuteur actif n'est présent que lors d'environ 30 % de la durée de la vidéo. En effet, on retrouve beaucoup de commentaires qui, prononcés par des voix-off ou extraits de discours, sont illustrés par des images ne correspondant pas aux moments où les paroles sont délivrées. Sur les contenus relatifs au sport, le constat est encore plus flagrant puisque souvent, la quasi-totalité des paroles sont prononcées par des commentateurs hors-champ.

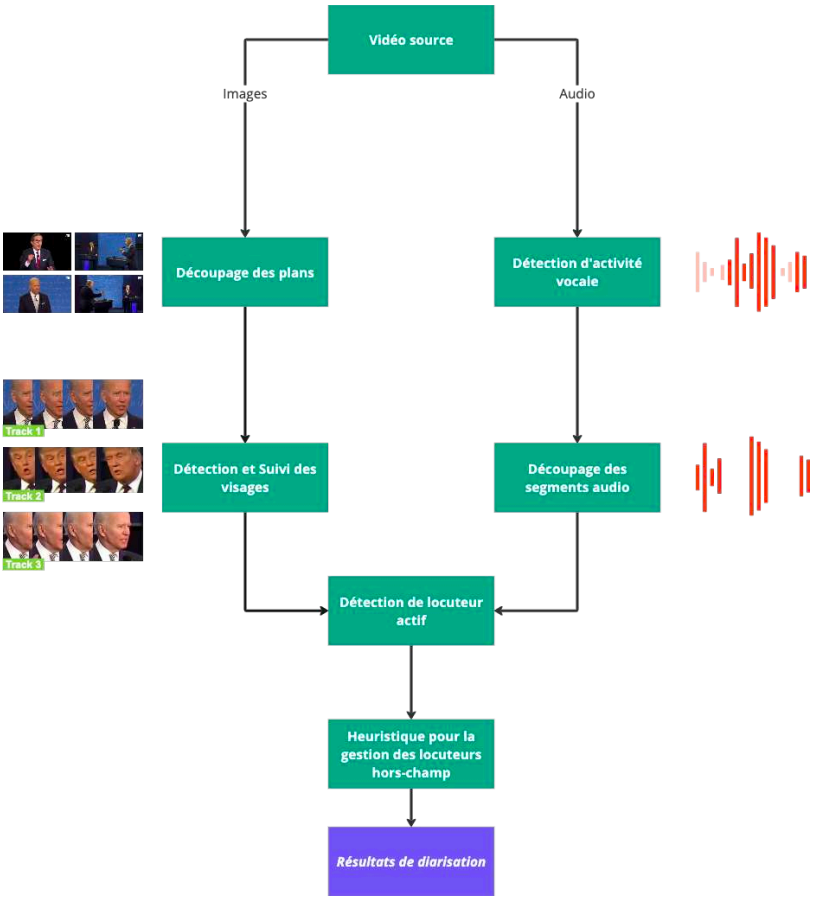

Figure 25 : Schéma d'un possible système de diarisation multimodale du locuteur intégrant une détection visuelle du locuteur actif

De plus, les approches multimodales audio-visuelles amènent de nouvelles problématiques de synchronisation (Garau et al., 2010) qui ne sont encore aujourd'hui que trop partiellement résolues pour être ignorées en contexte réel.

Raisons pour lesquelles, ces problématiques s'ajoutant à une nouvelle complexité temporelle d'analyse liée à l'extraction de chaque image de la vidéo, ces méthodes ne seront pas retenues pour la suite de nos travaux.

I.2. Vers une identification multimodale du locuteur

Plus généralement, l'état de l'art en analyse multimodale des vidéos issues de la télévision a montré, notamment lors des campagnes d'évaluation REPERE ou MediaEval, que pour regrouper efficacement les locuteurs, il peut être intéressant d'utiliser d'autres informations relatives au nom de la personne. En effet lorsqu'il est mentionné à l'oral (Jousse, Meignier, et al., 2009; Jousse, Petit-Renaud, et al., 2009),

ou lorsqu'il est écrit à l'écran (Poignant et al., 2014), on peut tenter de lier le nom de la personne à une des voix entendues et à des visages pour faire une vérification croisée de l'identité du locuteur (Le, N. et al., 2017). Cette fusion se fait fréquemment à l'aide de fonctions de croyance (El Khoury et al., 2012; Petitrenaud et al., 2010).

Cette recherche de l'identité du locuteur permet de passer d'une tâche de diarisation, qui a pour objet une identification relative (locuteur 1, locuteur 2, etc.), à une identification nommée du locuteur (Gay et al., 2014; Tranter, 2006).

Si ces approches semblent prometteuses sur des corpus normés, nous les avons testées sur les cas d'usages de Newsbridge et sommes arrivés à la conclusion qu'elles génèrent dans le cas général trop de faux positifs. Ce constat est confirmé par (Mauclair et al., 2006) qui observe jusqu'à 18,2% de données mal catégorisées sur la base ESTER.

Cette limitation se fait encore plus ressentir sur les données internationales traitées en raison des différences de domaines. En effet, entre une vidéo de *news* française, américaine, récente ou d'archive, les normes sont si différentes, que les règles et modèles adaptés à un cas ne produisent pas de bons résultats dès que l'on s'en éloigne trop.

II. État de l'art des approches audio-sémantiques

Le traitement du langage naturel est depuis longtemps déjà utilisé dans l'analyse de contenus multimédias. On peut ainsi en résumer le contenu, en extraire les principaux thèmes ou même prédire si une personne mentionnée apparaît où non dans une vidéo (Bechet et al., 2012).

II.1. Détection automatique des changements de locuteurs

Une approche assez évidente de la diarisation est la détection automatique des changements de locuteurs (*speaker change point detection - SCPD*). Ce qui peut se faire sur l'audio grâce à des modèles entraînés sur le signal brut, peut aussi se faire sur du texte, éventuellement après une étape de reconnaissance vocale. C'est ce que présentent notamment Anidjar et son équipe dans (Anidjar et al., 2021).

L'approche est la suivante : un jeu de données de détection de changements de locuteurs est créé à partir de la transcription automatique d'un jeu données audio. On extrait ensuite via la méthode de la fenêtre glissante des séquences de six mots. Pour chacune des séquences, on attribue un label de SCPD si et seulement si un changement

de locuteur intervient entre le 3ème et le 4ème mot de la séquence, soit au milieu de la fenêtre glissante (cf. Figure 26).

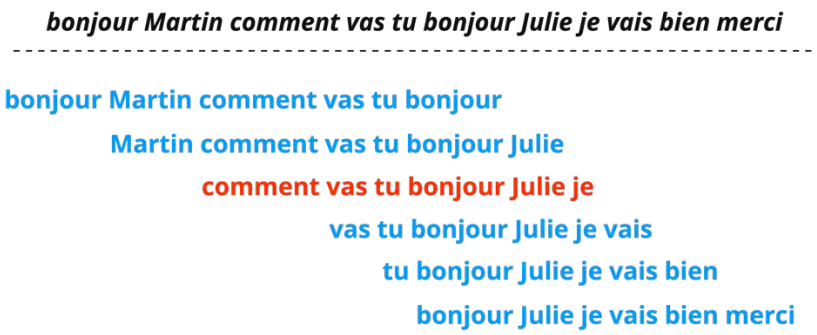

Figure 26 : : Exemple d'une annotation utilisée par l'algorithme de SCPD décrit. Des segments de six mots sont extraits. Seule la séquence (en rouge) pour laquelle le changement de locuteur est situé entre le 3ème et le 4ème mot sera annotée comme contenant un SCP

A titre d'exemple, sur la base de données AMI, environ 4 % des séquences ont été répertoriées comme contenant un changement de locuteur.

Ensuite, un réseau de neurones profond (DNN) est entraîné à effectuer une classification binaire entre les séquences. Pour ce faire, on utilise en entrée de réseau la représentation vectorielle des différentes séquences obtenues en faisant la moyenne des vecteurs associés à chaque mots via l'algorithme word2vec (Mikolov et al., 2013).

Cette méthode de détection de changement de locuteurs dans un texte, comme d'autres, permettent ensuite d'envisager une approche multimodale de la diarisation du locuteur. On peut citer (Park, T. J. et al., 2019) et (Park, T. J. and Georgiou, 2018) qui utilisent de tels classificateurs.

Méthode : Turn-to-diarize

Cette méthode introduite par les équipes de Google Research (Xia et al., 2022) ajoute à une diarisation acoustique standard, basée sur l'extraction de d-vecteurs et sur un regroupement spectral, une dimension sémantique. En effet, l'audio est transcrit par un algorithme de reconnaissance automatique de la parole, puis un algorithme de détection des changements de locuteur est appliqué.

La fusion se produit ensuite lors de la phase de regroupement puisque l'on vient contraindre le regroupement spectral avec les informations fournies par le détecteur de changements de locuteur.

Avec *i* et *j* les indices de deux segments de parole, on définit deux ensembles CL et ML tels que :

- CL correspond aux paires d'indices de segments qui ne peuvent pas avoir été prononcés par le même locuteur selon le système dédié à la détection de ces changements.
- ML correspond aux paires d'indices de segments entre lesquels le système n'a détecté aucun changement de locuteur.

Ainsi, en pratique, une matrice *C* de contrainte est créée telle que :

$$c_{i,j} = \begin{cases} -1, si\ (i,j)\ \epsilon\ CL\ et\ scpd(chgt_{loc}) > \theta; \\ 1, si\ (i,j)\ \epsilon\ ML; \\ 0, sinon. \end{cases}$$

Cette dernière méthode permet, selon ses auteurs, de gagner plus de 2 % de DER sur le jeu de données CALLHOME.

III. Contribution : Détection de changement du locuteur grâce aux modèles de langage volumineux

III.1. Enthousiasme autour des modèles de langage volumineux

Récemment, le secteur de l'intelligence artificielle a fortement été mis en lumière par le déploiement et la mise à disposition au grand public de ChatGPT, une application conversationnelle créée par la société OpenAI. Cette application utilise des modèles de langage volumineux (GPT-3.5 dans sa première version puis GPT-4) afin de donner l'impression à l'utilisateur qu'il converse avec un humain qui répond à ses questions et réalise pour lui certaines tâches.

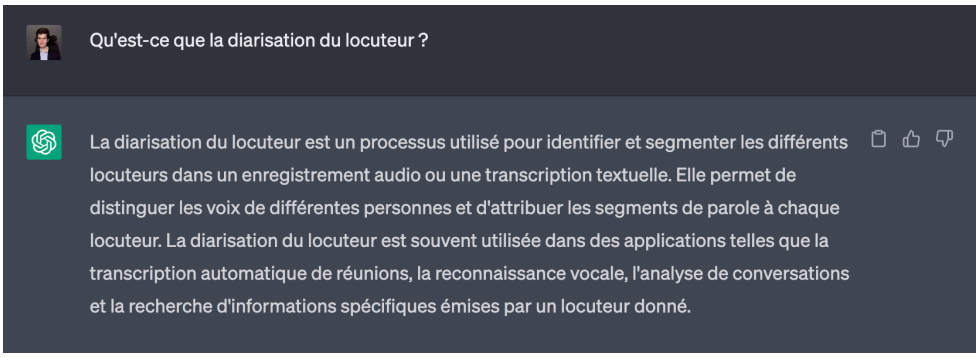

Figure 27 : Exemple d'une requête simple soumise à l'outil ChatGPT et sa réponse

Si l'on sait depuis longtemps que les modèles de langage volumineux (*large language models* - LLMs) sont particulièrement efficaces pour certaines tâches telles que la traduction, la reformulation, la génération de résumés ou de code, les modèles les plus récents ont également démontré l'émergence de nouvelles capacités.

On parle plus particulièrement de capacités émergentes d'un modèle d'intelligence artificielle lorsque celui-ci produit des résultats qualitatifs sur une certaine tâche pour laquelle il n'a pas particulièrement été entraîné. On retrouve de tels comportements dans plusieurs systèmes basés sur les réseaux convolutifs (CNN) et *transformers*.

Si l'on peut débattre d'une véritable émergence de capacités dans ces modèles d'IAs génératives, on constate néanmoins d'intéressantes performances sur des tâches impliquant du traitement du langage naturel et une analyse de la structure syntaxique d'un texte.

Nous avons cherché à évaluer si ces modèles autorégressifs, c'est-à-dire qui sont entraînés à toujours prédire le prochain mot ou groupe de mots, peuvent nous aider à comprendre automatiquement une conversation. En particulier nous les avons testés sur la tâche de détection automatique de changement de locuteurs.

III.2. Détection de changements de locuteur par des modèles de langage volumineux

III.2.1. Données utilisées pour l'étude

Afin d'établir les capacités des LLMs pour cette tâche, nous avons choisi de les évaluer sur les retranscriptions automatiques de la base données VoxConverse. Ainsi nous avons utilisé la solution par API VoxSigma de Vocapia[18] pour transcrire automatiquement le jeu de développement VoxConverse. A celui-ci, on appose un marqueur de changement de locuteur <*spk*> dès que la vérité terrain de la diarisation l'indique. Il est intéressant de noter que le service de Vocapia effectue déjà une diarisation et une ponctuation, dont on ne se sert pas pour cette étude.

Ensuite, on vient découper le texte en segments contenant 20 mots, soit plus de deux fois plus de contexte que la méthode précédemment évoquée (Anidjar et al., 2021). Au total, on obtient 13 100 segments avec la répartition suivante :

18. http://www.voxsigma.com/speech-recognition-software.html

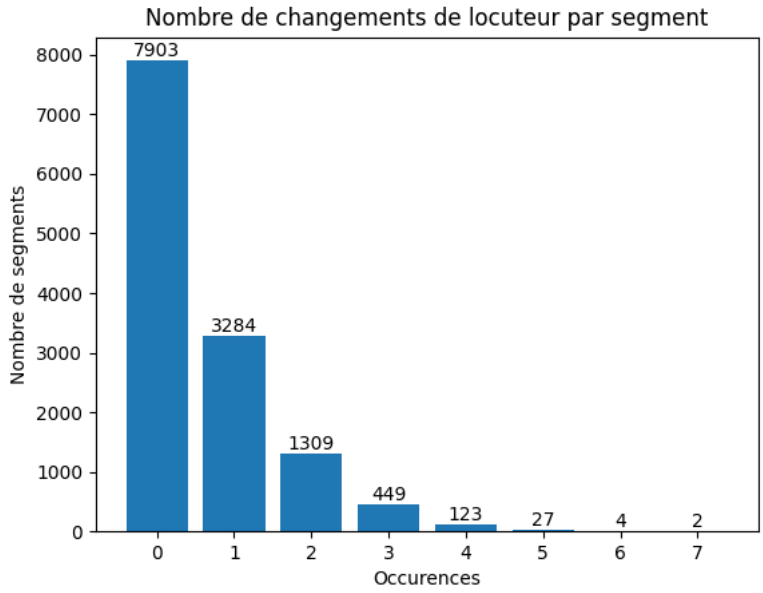

Figure 28 : Répartition du nombre de changements de locuteur dans les segments de texte utilisés pour notre étude

III.2.2. Méthodologie

Une fois la vérité terrain construite grâce aux annotations de diarisation et à la transcription automatique obtenue, nous cherchons à interroger chaque modèle de langage de façon optimale pour obtenir des résultats de détection des points de changement de locuteurs.

Ainsi on fait en sorte que lorsque nous envoyons nos segments de texte issus des transcriptions de VoxConverse, le LLM nous réponde avec le même segment de texte auquel il aura ajouté des occurrences de <spk> là où il prédit un changement de locuteur. On évalue les résultats en les classant selon quatre catégories :

- Les détections parfaites lorsque le système prédit un changement de locuteur exactement là où il a été placé par les annotations de diarisation.
- Les détections proches lorsque le système prédit un changement dans le segment mais le place mal.
- Les décisions correctes, somme des décisions proches et parfaites.
- Les décisions manquantes dans le cas où le système ne parvient pas à prédire un changement de locuteur présent dans la vérité terrain.

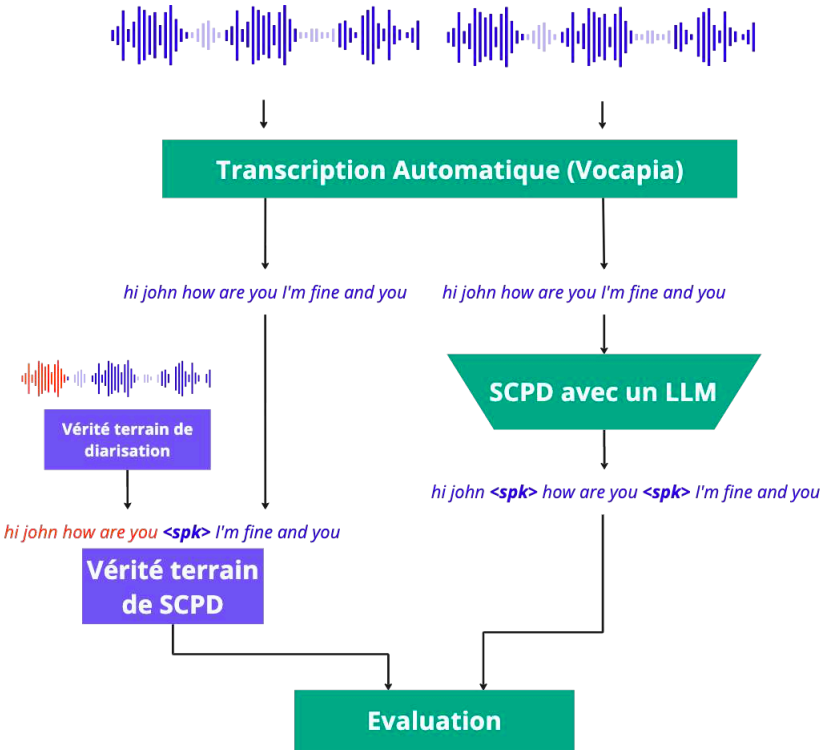

Figure 29 : Schéma descriptif de la méthodologie employée pour évaluer les performances des LLMs sur la tâche de détection des changements de locuteurs

III.2.3. Échec avec les LLMs d'ancienne génération

Avant le succès récent des modèles de langages volumineux, d'autres systèmes similaires avaient prouvé leur efficacité sur des tâches telles que la génération de résumé ou encore la traduction automatique. On peut citer les modèles BERT ou les architectures T5 (*Text-to-text transfer transformer* – T5) (Raffel et al., 2020) que l'on peut facilement spécialiser pour différentes tâches. Nos premiers tests se sont donc orientés vers un modèle T5 pour lequel nous avons essayé d'adapter une méthode de *prompt tuning*[19], normalement destiné à améliorer les performances de la tâche de génération automatique de résumé.

Ainsi nous avons construit un jeu de données composés de paires segment / segment + SCPD (cf. Figure 30). Puis nous avons tenté de spécialiser le modèle T5 large tel que décrit dans (Raffel et al., 2020). Malheureusement, ceux-ci se sont avérés infructueux

19. https://research.ibm.com/blog/what-is-ai-prompt-tuning

puisque le modèle T5 que nous avons tenté de spécialiser, a produit de trop nombreuses hallucinations[20] rendant son évaluation impossible.

- `diarize: I know what you said about sport I did it this year You said we should just talk about sport`
- `I know what you said about sport <spk> I did it this year <spk> You said we should just talk about sport`

Figure 30 : Exemple d'une paire utilisée pour spécialiser le modèle T5. Le suffixe diarize est ajouté pour désigner au modèle qu'une nouvelle tâche lui est demandée

III.2.4. Premiers succès avec GPT-4

Suite à cela, et au regain d'enthousiasme pour les modèles de langages volumineux avec la mise à disposition de ChatGPT et de GPT-4, nous avons décidé de tenter d'utiliser ces solutions pour accomplir notre tâche de détection de changements du locuteur, cette fois sans aucun apprentissage supervisé. Après quelques itérations pour trouver la bonne requête (*prompt* en anglais). Nous avons finalement réussi à trouver une *system prompt* qui permet de réaliser la tâche de TSCPD avec GPT-4 :

```
You are a system designed to detect speaker change
in an automatic transcript. Copy all the text I give
you without correcting it and add <spk> when you
believe a speaker change happened.
```

Figure 31 : Prompt utilisée pour effectuer la tâche de TSCPD avec GPT-4 sans spécialisation du modèle

Tableau 10 : Résultats de détection de changements locuteur obtenus sans supervision avec un modèle T5 et les modèles de langage GPT

	% Détections parfaites	% Détections proches	% Détections correctes	% Détections manquantes
T5	N/A	N/A	N/A	100.0
GPT-3.5	N/A	N/A	N/A	100.0
GPT-4	11.3	22.6	34.0	66.0

20. Terme utilisé pour désigner les générations erronées des modèles génératifs comme les LLMs.

Ces premiers résultats montrent que seul GPT-4 permet d'obtenir des résultats concluants avec environ un tiers de détections correctes des changements de locuteur (cf. Tableau 10).

III.2.5. Vers des LLMs spécialisés dans la compréhension des conversations

Suite aux premiers succès obtenus avec GPT-4, nous avons cherché à reproduire les résultats obtenus avec un LLM *open-source,* dont nous connaissons les données ayant servi à l'entraîner. Après plusieurs échecs avec les versions disponibles de Falcon-13B, Falcon-40B (Almazrouei et al., 2023), LLaMa-7B, LLaMa-30B (Touvron et al., 2023), qui produisaient plus d'hallucinations que de résultats exploitables, nous avons finalement obtenu des résultats intéressants en spécialisant un modèle Falcon-40B.

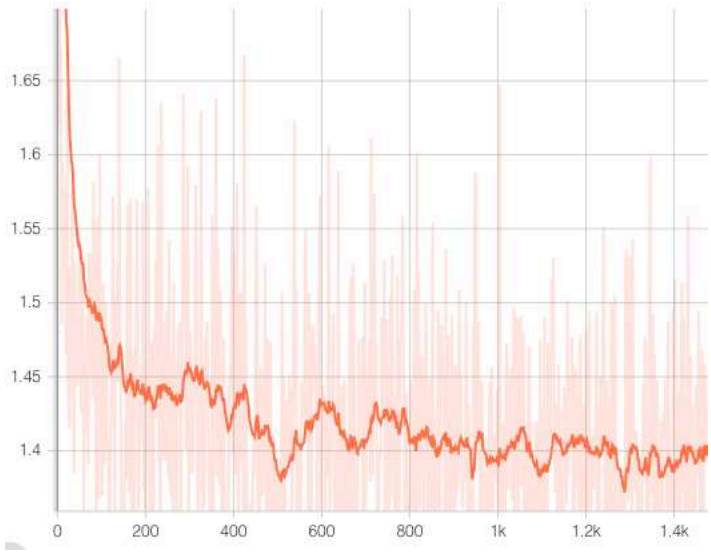

Figure 32 : Évolution de la fonction de coût au fur et à mesure des étapes de la spécialisation du LLM

Pour ce faire, nous avons employé une méthode PEFT (*parameter efficient finetuning*) appelée QLoRA (*quantized low-rank adapter*) (Dettmers et al., 2023). Celle-ci permet, au lieu de modifier l'ensemble des paramètres du réseau de neurones, de n'entraîner qu'un plus petit réseau avec peu de couches, que l'on vient finalement ajouter aux dernières couches du réseau initial, dont on gèle les paramètres en en réduisant la complexité numérique. Lors de cette dernière phase, dite de quantification, au lieu de travailler avec des nombres à virgule flottante sur 16 bits, on travaille avec des entiers sur 4 bits.

La spécialisation du modèle a été caractérisée par une convergence rapide de la fonction de coût (cf. Figure 32), laissant espérer de bons résultats avec le LLM spécialisé.

Tableau 11 : Résultats de détection de changements locuteur obtenus avec les modèles de langage Falcon-40B

	% Détections parfaites	% Détections proches	% Détections correctes	% Détections manquantes
Falcon-40B Instruct	N/A	N/A	N/A	100,0 %
Falcon-40B spécialisé	17,7 %	64,5 %	82,2 %	17,8 %

Conformément à nos attentes, les résultats du Tableau 11 montrent que le modèle Falcon spécialisé permet de remplir la tâche de détection de locuteur actif, de façon même supérieure aux modèles GPT testés précédemment puisqu'on obtient ici jusqu'à 82,2 % de détections correctes.

Ces résultats encourageants, obtenus grâce à un LLM de l'état de l'art spécialisé sur relativement peu de données annotées, suggèrent que les modèles de langage volumineux peuvent effectivement apprendre à identifier les structures sous-jacentes aux conversations humaines.

III.2.6. Étude de la consistance des résultats

Une des propriétés inhérentes aux modèles génératifs que sont les modèles de langage utilisés, est la variabilité des résultats générés si l'on reproduit plusieurs requêtes avec la même instruction. Dans notre cas, on souhaite minimiser cet effet et ne générer que le segment de parole augmenté des éventuels changements de locuteur détectés. Afin d'appréhender au mieux les performances évoquées précédemment, on souhaite vérifier leur consistance lorsque l'on reproduit la même requête plusieurs fois.

Après avoir passé 100 fois un des segments de notre base de données dans le modèle évalué, nous référençons les différentes réponses qu'il produit. Notre segment candidat est choisi tel qu'il contient deux changements de locuteur. Par ailleurs, nous choisissons délibérément un segment pour lequel la phase de transcription n'a pas parfaitement fonctionné. Ce qui permet d'attester la consistance de nos systèmes dans les cas les plus complexes.

Pour GPT-4, sur 100 itérations, nous obtenons 17 réponses différentes qui varient souvent d'un caractère (souvent des espaces). Parmi ces alternatives, nous rapportons

- 1 alternative où aucun des deux points de changement du locuteur n'a été détecté,
- 1 alternative avec une détection parfaite et une manquante,
- 15 alternatives différentes avec une détection proche et un point de changement manquant. Pour ces dernières la distance entre le point de changement de locuteur de la vérité de terrain et celui qui a été détecté est d'au plus 6 mots.

En termes de pourcentage d'apparition des occurrences, l'alternative avec la détection parfaite est la plus récurrente, avec 52 % des expériences ayant permis de la générer (cf. Tableau 12).

Tableau 12 : Résultats de consistance de la détection de changements locuteur obtenus sans supervision avec le modèle de langage GPT-4

1 détection parfaite + 1 manquée	52 % (1 alternative)
1 détection proche + 1 manquée	42 % (15 alternatives)
2 détections manquées	6 % (1 alternative)

Concernant notre modèle spécialisé Falcon-40B, malgré de meilleurs résultats, nous constatons une beaucoup plus forte variabilité dans les réponses générées. Voici pour illustration les résultats de consistance obtenus :

Tableau 13 : Résultats de consistance de la détection de changements locuteur obtenus avec le modèle Falcon-40B spécialisé pour cette tâche.

2 détections parfaites	21 % (2 alternatives)
1 détection parfaite + 1 proche	34 % (7 alternatives)
1 détection parfaite + 1 manquée	17 % (9 alternatives)
2 détections proches	2 % (2 alternatives)
1 détection proche + 1 manquée	5 % (5 alternatives)
2 détections manquées	21 % (19 alternatives)

Il est également intéressant de noter que ce modèle est plus sujet aux hallucinations que celui précédemment testé. En effet, parmi les 19 alternatives pour lesquelles aucun des deux changements de locuteur n'a été détecté, beaucoup étaient même loin de reproduire la phrase initialement fournie.

III.3. Mise en perspective

Les performances illustrées sont très encourageantes et tendent à conforter l'hypothèse de l'apparition d'une nouvelle capacité émergente de certains modèles de langage volumineux.

Cette notion d'émergence a été démocratisée par les travaux du prix Nobel de physique Philip Warren Anderson dans un article intitulé « *More is different* ». Plus récemment elle a été largement reprise pour qualifier les nouvelles tâches que les LLMs parviennent à réaliser sans avoir appris spécifiquement à les réaliser.

En effet, on parle de capacités émergentes (Wei et al., 2022) lorsqu'un système, souvent un réseau de neurones, montre de bonnes performances sur une tâche pour laquelle il n'a pas été entraîné et ce sans aucune adaptation spécifique.

Dans notre cas GPT-4 semble être capable de détecter environ un tiers des changements de locuteurs dans la transcription d'une conversation. Pourtant, sans en avoir une absolue certitude, il ne semble pas que la société OpenAI ait entraîné son modèle sur ce type de données très spécifiques (OpenAI, 2023).

Néanmoins, certains travaux sur les capacités émergentes des LLMs sont largement débattus en particulier ceux qui visent à démontrer que les modèles de langage sont un premier pas vers une intelligence artificielle forte (*artificial general intelligence* – AGI)[21] (Bubeck et al., 2023).

Par ailleurs, nos résultats avec le modèle Falcon spécialisé confirment que cette catégorie de modèles peut être adaptés à la tâche de détection des changements de locuteur dans du texte. Ce qui va dans le sens d'une compréhension générale acquise par les LLMs de la structure des conversations.

Ces travaux de détection des changements de locuteurs dans du texte doivent permettre deux choses. Premièrement ils doivent permettre de rendre plus accessibles les sous titres. En effet, aujourd'hui lorsqu'un média est sous-titré, il est rare que le fichier de sous titres compagnon contienne des informations sur les locuteurs. Or pour les personnes sourdes et malentendantes, il est recommandé par les normes d'accessibilité d'afficher de couleurs différentes deux phrases consécutives prononcées par deux locuteurs différents[22]. La deuxième utilisation de cette méthode pourrait être une correction a posteriori des résultats obtenus par une approche audio de diarisation ou de SCPD. C'est dans ce sens que se poursuivront nos travaux.

21. https://fr.wikipedia.org/wiki/Intelligence_artificielle_g%C3%A9n%C3%A9rale
22. https://www.bbc.co.uk/accessibility/forproducts/guides/subtitles/#PRESENTATION

Pour poursuivre cette étude il faudra également évaluer dans quel contexte et quel type de conversations sont les mieux analysées (questions / réponses, discussions naturelles, débats, etc.), afin de déterminer si d'autres facteurs jouent implicitement sur ces résultats, comme par exemple le ton de la phrase prononcée, la rhétorique employée, la ponctuation, etc.

Finalement, parmi les nombreuses méthodes de diarisation multimodale citées (cf. Tableau 14), et les approches tendant vers l'identification multimodale, les méthodes tenant compte de l'audio et dans un second temps du texte semblent les plus adaptées au traitement des médias télévisés dans toute leur diversité.

Tableau 14 : Récapitulatif des principales méthodes de diarisation multimodale

Méthode	Modalité(s)	Représentation vectorielle utilisée	Particularité
Speaker diarization with lexical information	Audio, Sémantique	x-vecteurs	Matrice d'adjacence + regroupement spectral
Turn-to-diarize	Audio, Sémantique	d-vecteurs	Détection textuelle de changements de locuteur pour contraindre le regroupement spectral
Diarisation avec ASD	Audio, Image	Obtenus via un CNN	Détection audiovisuelle du locuteur actif
Approche paralinguistique	Audio, paralinguistique	Eléments paralinguistiques inférés	Ulisation d'éléments paralinguistiques

Conclusion sur la robustesse de la diarisation

Si les méthodes de diarisation du locuteur sont de plus en plus diverses, performantes et complexes, on a pu voir également dans cette partie qu'il existe encore de nombreux problèmes de robustesse.

Il reste de nombreux obstacles à franchir pour obtenir une diarisation du locuteur également performante selon les cas, en particulier dans les conditions très changeantes qu'offre le cas l'analyse de vidéos issues de la télévision.

Néanmoins, nous avons vu que différentes méthodes de fusion, multimodale ou non, peuvent permettre de régler certains des problèmes des récents systèmes de diarisation du locuteur.

En particulier, on a souligné l'importance qui doit être accordée à l'étape initiale de détection d'activité vocale. En introduisant une approche multi-flux, qui tire sa force des différents points forts des approches existantes, nous avons démontré les possibles gains de performance en termes de DER, en faisant uniquement varier ce module.

Par ailleurs, nous observons que la diarisation du locuteur ne se résume plus seulement à un problème de traitement du signal audio mais que de nombreuses autres modalités peuvent être considérées. Pour illustrer ce point, nous avons, entre autres, montré que les modèles de langage volumineux peuvent servir à effectuer une détection de changement de locuteur dans une conversation retranscrite automatiquement.

Voyons désormais comment il devient nécessaire, au-delà des performances de la diarisation du locuteur, d'en considérer la justesse algorithmique et la consommation énergétique afin de garantir une diarisation responsable, adaptée à un large panel de cas d'utilisation.

Partie II : Vers une diarisation responsable

Afin de permettre une utilisation de la diarisation du locuteur sur de grandes quantités de données et dans des secteurs où l'apparition de biais peut avoir un fort impact, il est nécessaire de travailler à produire une diarisation responsable. Celle-ci se doit de rester performante tout en étant robuste à certains biais et énergétiquement efficace.

Ces contraintes s'appliquent en particulier aux solutions technologiques développées et commercialisée dans l'Union Européenne puisque la future législation sur l'utilisation de l'intelligence artificielle[23] prévoit notamment une obligation de contrôle des biais pour certains des acteurs qui développent ces technologies. Si la diarisation du locuteur ne tombe pas directement sous le coup des règles qui s'imposeront aux systèmes d'identification biométrique, on peut souligner que les technologies employées sont souvent similaires voire parfaitement identiques (Mtibaa et al., 2021).

A la lumière de ceci, il nous a semblé nécessaire de considérer la diarisation du locuteur sous l'angle nouveau de sa justesse algorithmique. En effet, tout algorithme, d'autant plus s'il s'appuie sur des principes probabilistes, comme c'est le cas pour l'ensemble des solutions d'apprentissage automatique, peut présenter des biais qu'il s'agit d'identifier puis de corriger.

Dans cette seconde partie, nous tentons d'instaurer un premier état des lieux de la justesse algorithmique de la diarisation du locuteur, en étudiant les biais d'une des méthodes de l'état de l'art. Nous nous intéressons également à la consommation énergétique de ces algorithmes et explorons comment leurs performances peuvent permettre de nouveaux usages notamment dans le milieu médical.

23. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Chapitre IV : État de l'art de l'évaluation de la diarisation

I. État de l'art : Métriques existantes

I.1. DER

Le DER ou *Diarization Error Rate* est la métrique de référence pour l'étude d'un système de diarisation. Pour le calculer, on évalue les durées pendant lesquelles on n'a pas su détecter la parole (PM), les durées où de la parole a été détectée à tort (FA) et enfin les durées où le système s'est trompé de locuteur actif (CF), toutes trois illustrées en Figure 33. On applique ensuite la formule suivante pour trouver le DER :

$$\text{DER} = \frac{PM+FA+CF}{Durée\ totale\ de\ parole}$$

Il est intéressant de noter qu'à l'image du WER *(Word Error Rate*) pour la reconnaissance de la parole, le DER peut être supérieur à 1.

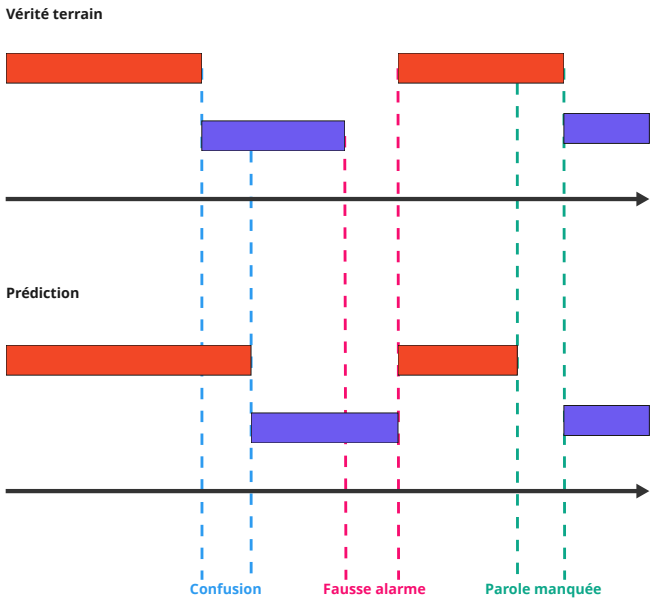

Figure 33 : Schéma illustrant les trois cas d'erreur de diarisation utilisés pour calculer le DER

Les premières implémentations de cette métrique introduite par le *National Institute of Standards and Technology* (NIST) américain (Galibert, 2013) ne tenaient initialement pas compte des passages avec des paroles superposées. Ce n'est que depuis quelques années que désormais la quasi-totalité des performances publiées tiennent compte de ces passages particulièrement difficiles à traiter.

Pour le calcul du DER, on s'accorde parfois une marge, ou *collar*, en deçà de laquelle on ne décompte pas les erreurs. Une valeur fréquemment utilisée pour cette marge est de 250 millisecondes. Toutefois, l'utilisation d'une telle marge peut entraîner des biais, en particulier si les tours de paroles sont courts. En effet, 250 ms de marge n'a, en pourcentage pas le même impact sur des segments de 5 secondes et sur d'autres de 20 secondes. C'est pourquoi, compte tenu de la grande diversité des contenus que nous souhaitons pouvoir traiter, avec de longs tours de parole et d'autres plus brefs, les résultats présentés dans cette thèse sont tous calculés sans marge.

I.2. JER

Si le DER reste la métrique phare pour l'évaluation de la tâche de diarisation du locuteur, quelques autres métriques sont parfois mentionnées. Parmi elle, la principale est le JER ou *Jaccard Error Rate* a été introduite lors du challenge DIHARD.

Pour chaque locuteur l d'un enregistrement dont l'ensemble des locuteurs actifs est de taille L, on peut calculer le JER spécifique au locuteur :

$$JER_l = \frac{PM_l + FA_l}{Total_l}$$

Puis, pour l'ensemble de l'enregistrement,

$$JER = \frac{1}{L}\sum_{l=1}^{L} JER_l$$

Cette métrique a pour intérêt de faire peser un poids similaire aux erreurs venant des différents locuteurs, et ce, indépendamment du fait qu'ils aient ou non parlé plus que leurs co-locuteurs.

Ces métriques, JER et DER, peuvent être calculées grâce à des outils *open-source* tels que *dscore* ou *pyannote.metrics*.

II. Protocoles d'annotation et d'évaluation de la diarisation

II.1. Difficulté de l'annotation et de l'évaluation de la diarisation

Malgré l'existence des métriques précédemment évoquées, il reste très compliqué d'évaluer la diarisation du locuteur en termes de performances et de robustesse. En effet, selon que l'on souhaite une diarisation indépendante optimale ou une diarisation optimisée pour la reconnaissance automatique de la parole, le choix de la méthode ou des hyperparamètres peut varier significativement.

Enfin, on peut également imaginer une diarisation dont le seul but serait de compter avec précision le nombre de locuteurs actifs auquel cas des réglages spécifiques pourraient également être nécessaires.

En outre, annoter un segment audio avec la vérité terrain de diarisation du locuteur peut s'avérer complexe (Broux, 2020) et il est fréquemment difficile d'arriver à un consensus entre les annotateurs (McKnight et al., 2022). En effet, il n'est pas toujours facile de définir les frontières d'un tour de parole. Commence-t-il lorsque la personne précédente finit de parler, lorsque le nouveau locuteur prend son inspiration pour entamer une phrase ou à l'instant où le premier son sort de la bouche de ce dernier ? Tant de questions qui n'ont jamais été formellement tranchées dans le contexte de l'évaluation de la diarisation.

Une conversation entre deux personnes, ou plus, est infiniment complexe et dépend de très nombreux paramètres qu'il est difficile de reproduire précisément. En effet, la nature d'une conversation peut varier en fonction de l'identité des deux locuteurs bien sûr mais aussi en fonction de leurs éventuels accents, de leurs âges, de leurs prosodies, etc.

Autant de paramètres dont il faudrait tenir compte pour effectuer une analyse précise des biais de la diarisation dans une démarche scientifique. Or, il est très compliqué, voire impossible de ne faire varier qu'un seul de ces paramètres et de reproduire plusieurs fois la même conversation avec une stabilité paramétrique de tous les autres aspects. Par exemple, imaginons que nous disposions d'un enregistrement d'une discussion entre un locuteur A de sexe masculin, qui a 35 ans et un accent provençal et un second locuteur B féminin de 60 ans sans accent. Il sera alors très difficile de reproduire exactement la même conversation en ne faisant varier qu'un seul de ces paramètres, pourtant tous constitutifs de la voix des locuteurs étudiés.

Ainsi, si plusieurs bases de données existent pour l'analyse des biais de la reconnaissance vocale (Ardila et al., 2020; Hazirbas et al., 2022), aucune n'est vraiment adaptée à l'étude de la tache de diarisation du locuteur.

III. Besoin de justesse dans le traitement des contenus destinés à une large diffusion

Les technologies d'IA développées à l'occasion de cette thèse, ainsi que de nombreuses autres, ont pour finalité d'être utilisées dans de multiples rédactions par des journalistes souhaitant couvrir des sujets d'actualité.

Le métier de journaliste demande une certaine rigueur dans la sélection des contenus à présenter, afin qu'ils soient les moins biaisés possible. Lorsque ce métier est aidé par des outils d'intelligence artificielle comme ceux développés dans le cadre de cette thèse, il devient nécessaire de prendre garde aux éventuels nouveaux biais que ces outils induisent.

Les solutions d'intelligence artificielle déployées dans le secteur des médias servent aujourd'hui à accélérer la production et la diffusion du contenu créé pour de toujours plus larges audiences. Si cela permet d'informer, de sensibiliser, et de divertir de plus en plus de personnes, cela participe aussi à forger des opinions sur certains sujets. En ce sens, il est primordial que les informations diffusées et produites pour être représentatives de la réalité d'une société le soient, avec des outils qui eux même ne contiennent pas de biais sur cette même réalité.

On sait que certaines voix sont plus agréables à écouter que d'autres, du fait notamment des fréquences qui les composent (Gallardo et al., 2018; Weiss and Burkhardt, 2010). Victor Hugo parlait même de « voix d'or » à propos de l'actrice Sarah Bernhardt.

Il en va de même pour les algorithmes de traitement automatique de la parole. Selon la façon dont ils ont été conçus, ceux-ci peuvent être plus ou moins sensibles à certaines catégories de voix. Néanmoins, il ne faut pas que ces voix soient particulièrement effacées des sujets d'actualité. On pense en particulier aux voix d'enfants, de personnes âgées, de personnes avec un accent prononcé, qui risquent souvent d'être invisibilisées.

En effet, il devient primordial que les outils utilisés lors de la production de contenus audiovisuel restent les moins biaisés possible.

Dans un contexte de couverture journalistique, on ne peut pas se permettre de n'indexer que les « voix d'or ». C'est pourquoi la justesse de la diarisation du locuteur est essentielle en complément de ses bonnes performances.

III.1. Exemple de biais sur des contenus diffusés à la télévision

Dans le vaste domaine de la recherche en intelligence artificielle, une attention croissante est portée aux biais des méthodes et modèles aujourd'hui utilisés par des millions de personnes (Chhabra et al., 2021 ; Mehrabi et al., 2022). Cette tendance s'est notamment vue accélérée par plusieurs scandales qu'ont subis les géants américains[24,25]. Ceux-ci ont été souvent précédés ou suivis par des publications scientifiques alarmantes sur les biais des modèles d'IA (Raji and Buolamwini, 2019). Si ces recherches se sont fortement concentrées sur les algorithmes de vision par ordinateur, on observe également la nécessité de telles études concernant les systèmes de traitement automatique de la parole (Jin et al., 2022; Rajan et al., 2022; Shen et al., 2022).

En parallèle, des études montrent d'importants biais dans les émissions diffusées à la télévision (Doukhan et al., 2018), notamment en ce qui concerne le genre des personnes s'y exprimant.

Si les algorithmes que nous développons permettent de détecter certains de ces biais (Lebourdais et al., 2022), il devient nécessaire que ceux-ci, lorsqu'ils sont utilisés pour produire de nouveaux contenus, ne perpétuent pas ces biais, ni n'en créent de nouveaux.

24. https://www.lemonde.fr/pixels/article/2017/04/15/quand-l-intelligence-artificielle-reproduit-le-sexisme-et-le-racisme-des-humains_5111646_4408996.html
25. https://www.rtl.fr/actu/sciences-tech/l-intelligence-artificielle-de-facebook-a-confondu-des-personnes-noires-avec-des-primates-7900069176

Chapitre V : Contribution : étude de la justesse de la diarisation

A l'heure actuelle, les recherches en diarisation du locuteur s'orientent principalement vers la mesure des performances et notamment la réduction du DER dans les scenarii les plus complexes.

Ceci s'explique notamment par le fait que les métriques principales d'évaluation de la diarisation du locuteur sont le DER et dans une moindre mesure le JER, qui ne permettent pas d'expliciter les éventuels biais de la diarisation.

I. Introduction du taux de justesse de la diarisation

I.1. Restriction du domaine de l'évaluation

Afin de pallier ce manque de métrique dans le contexte difficile de l'évaluation de la justesse de la diarisation du locuteur, nous proposons d'introduire le taux de justesse de la diarisation, ou DFR (*diarization fairness rate*). Cette métrique et son protocole associé doivent permettre une première évaluation scientifique de la justesse de la diarisation selon plusieurs critères (Tevissen, Boudy, Chollet, et al., 2022).

Pour ce faire, on choisit de se restreindre au cas où une seule personne parle. Ce cas, supposé simple pour la diarisation, nous permet de mener des tests unitaires sur des bases de données existantes. On pourra alors produire des résultats dans un nombre suffisant pour qu'ils soient statistiquement représentatifs.

I.1.1. Définition du DFR

Soient δ un critère étudié tel que l'âge, le sexe ou l'accent du locuteur, i l'indice de l'observation $o_{\delta i}$ et P la loi de probabilité qui représente les différentes valeurs prises par les $o_{\delta i}$, c'est-à-dire les différents résultats possibles de diarisation. Alors on définit le DFR comme suit :

$$DFR(\delta) = P(o_{\delta i} = 1)$$

En d'autres termes, le DFR représente la proportion des expériences pour lesquelles la diarisation évaluée a produit le résultat attendu, à savoir la détection d'un unique locuteur.

Dans notre étude, on s'intéressera aussi aux cas où $P(o_{\delta,i} = 0)$ et $P(o_{\delta,i} > 1)$, notés p_0 et p_+, c'est-à-dire la proportion des cas dans lesquels la diarisation a respectivement failli à détecter le locuteur actif et ceux dans lesquels elle a à tort caractérisé l'enregistrement comme ayant été prononcé par plusieurs locuteurs.

I.2. Protocole mis en place

Ainsi, afin d'évaluer le DFR, on choisit d'appliquer notre algorithme de diarisation de référence sur un grand nombre de phrases prononcées par une grande variété de locuteurs différents.

Pour ce faire on choisit la base de données Mozilla Commonvoice[26]. Normalement destinée à l'entraînement ou à l'évaluation de méthodes d'ASR, cette base est l'une des seules à proposer un très grand nombre de locuteurs et plusieurs attributs renseignés pour chacun d'entre eux. En se basant sur ces attributs, on choisit d'évaluer la justesse de la diarisation du locuteur selon quatre critères :

- l'âge du locuteur, considéré par tranche de dix ans de 10 à 100 ans ;
- le sexe du locuteur ;
- l'accent du locuteur ;
- la longueur de la phrase prononcée.

Pour chacun des critères étudiés et chacune des mesures faites, on s'intéressera tout particulièrement à la composition du jeu de données qui a servi à l'étude. En particulier, on calculera toujours les intervalles de confiance à 99 % définis tels que le résultat réel $\tilde{p}$ puisse être calculé en fonction de l'observation p des N expériences indépendantes réalisées, comme suit :

$$p - \varepsilon(p) \leq \tilde{p} \leq p + \varepsilon(p)$$

Avec,

$$\varepsilon(p) = 2{,}58\sqrt{\frac{p(1-p)}{N}}$$

26. https://commonvoice.mozilla.org/fr/datasets

Nous choisissons de mener cette étude en évaluant un algorithme de diarisation du locuteur dont les performances rivalisent encore avec l'état de l'art : VBx (Cf. Chapitre I-III.2, p. 36). L'implémentation choisie est celle décrite dans (Landini et al., 2022).

II. Jeu de données Mozilla Commonvoice

Le jeu de données Mozilla Commonvoice est une collection collaborative et *open-source* d'enregistrements audio prononcés par chacun des contributeurs via le site de la fondation Mozilla[27].

Chaque contributeur s'enregistre en prononçant une phrase et complète les attributs demandés : âge, sexe, accent. Cette phase d'auto-labellisation des enregistrements peut certes être source d'un biais dans notre étude, mais la très grande quantité d'enregistrement permet d'en minimiser le risque.

En effet, pour cette étude nous utilisons la version anglaise du jeu de données Commonvoice dans sa neuvième version qui contient 81 085 locuteurs différents, chacun prononçant une phrase.

S'il n'est pas initialement destiné à évaluer la diarisation du locuteur mais plutôt la reconnaissance de la parole, ce jeu de données est, au jour de notre étude, le seul à permettre une étude à grande échelle des biais d'une solution de traitement de la parole.

III. Résultats et biais identifiés

Cette première étude de la justesse de la diarisation du locuteur a permis d'identifier un certain nombre de biais, ou inversement un comportement robuste de l'algorithme VBx sur les enregistrements de la base Commonvoice.

27. https://commonvoice.mozilla.org/

III.1. Age du locuteur

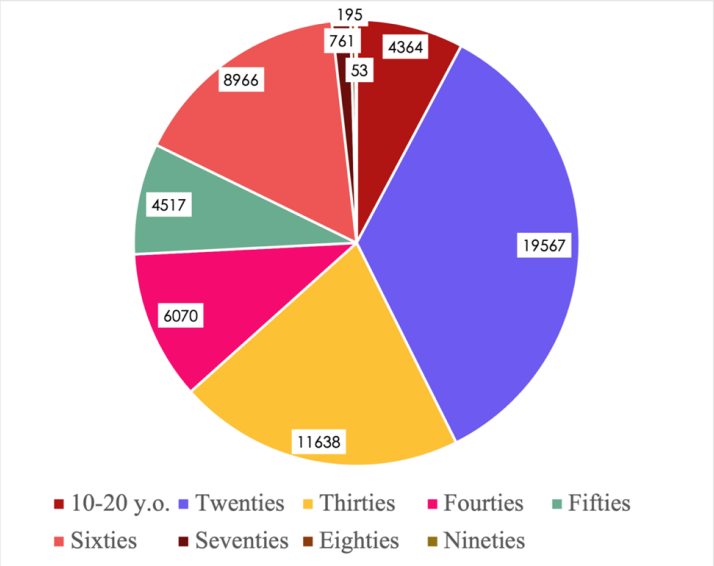

Figure 34 : Répartition des différentes classes de l'attribut « Age » de la base de données Commonvoice

Concernant l'âge du locuteur, on constate une certaine stabilité, à quelques pourcents près des performances de la diarisation (cf. Figure 35, Tableau 15). On note cependant deux comportements intéressants :

- Les enregistrements de personnes plus jeunes semblent être plus sujets à la non détection de parole. Ceci est sûrement lié à la méthode de détection d'activité vocale utilisée (seuil d'énergie) et peut-être aux prosodies différentes. Les personnes plus âgées ont par exemple la caractéristique de parler un peu plus lentement que les personnes jeunes.
- Le DFR pour les personnes de plus de 90 ans semble nettement plus bas que la moyenne. Cela est sans doute multifactoriel. On peut supposer que les personnes très âgées sont moins présentes dans les jeux de données ayant servi à entraîner la méthode de diarisation utilisée, en particulier l'extraction des x-vecteurs (VoxCeleb 1 et 2). De plus, c'est pour cette classe que nous avons le moins d'enregistrements de test (seulement 53, cf. Figure 34), c'est pourquoi l'incertitude sur ces résultats est plus élevée (cf .Tableau 16).

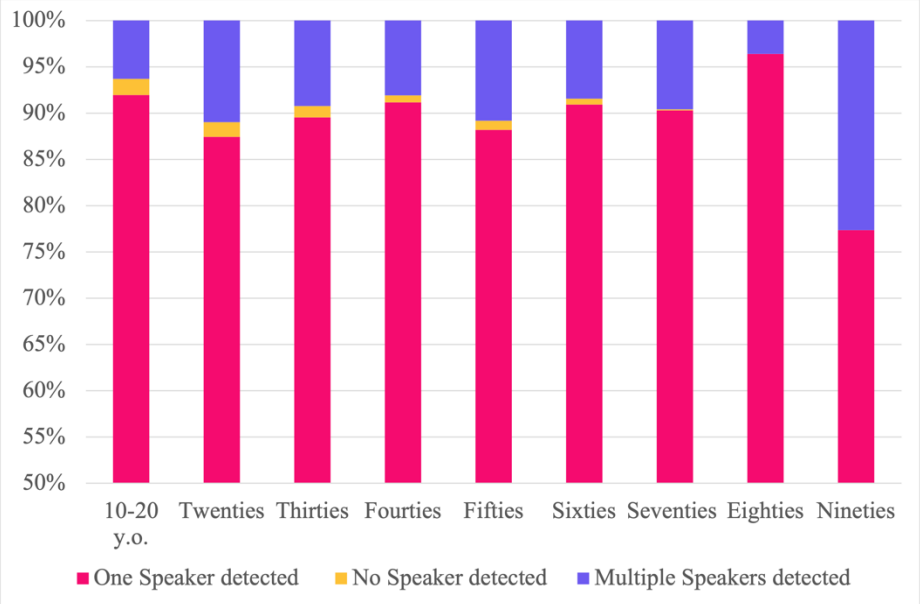

Figure 35 : Résultats obtenus par la diarisation selon l'attribut « Age »

Tableau 15 : Résultat de l'analyse de biais selon le critère d'âge des différents locuteurs

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
p_0	1,76	1,55	1,23	0,72	1,00	0,62	0,13	0,00	0,00
DFR	**91,93**	**87,45**	**89,53**	**91,17**	**88,19**	**90,92**	**90,28**	**96,41**	**77,36**
p_+	6,30	11,00	9,25	8,11	10,81	8,46	9,59	3,59	22,64

Tableau 16 : Incertitudes liées à l'intervale de confiance à 99 %
pour les résultats concernant l'âge des locuteurs

	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
$\varepsilon(p_0)$	1,06	0,61	0,73	0,94	1,24	0,78	2,77	3,44	14,83
$\varepsilon(DFR)$	0,51	0,23	0,26	0,28	0,38	0,21	0,34	0,00	0,00
$\varepsilon(p_+)$	0,95	0,58	0,69	0,90	1,19	0,76	2,75	3,44	14,83

III.2. Sexe du locuteur

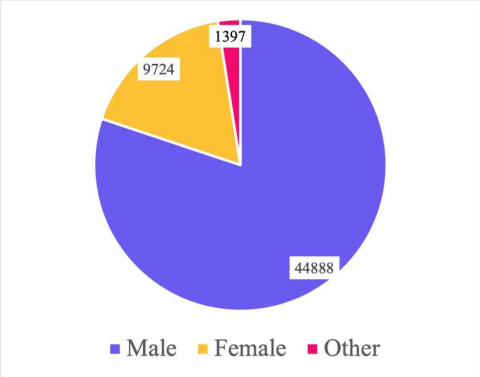

Figure 36 : Répartition des différentes classes de l'attribut « Sexe » de la base de données Commonvoice

C'est sûrement pour le sexe du locuteur que l'on obtient un résultat des plus intéressants de cette étude puisque l'on constate que sur l'expérience unitaire proposée, la diarisation fournit des meilleurs résultats d'environ 4 % pour les femmes en comparaison à ceux des hommes (cf. Figure 37, Tableau 17). En particulier, les voix des hommes sont plus souvent détectées comme émanant de plusieurs locuteurs différents, et ce malgré le grand nombre d'échantillons considérés permettant une faible marge d'erreurs (cf. Figure 36, Tableau 18).

Une des explications possibles est la répartition des fréquences dans les voix masculines et féminines. En pratique, il semble en effet que les voix féminines couvrent un plus large spectre de fréquences que les voix masculines. On peut alors conjecturer qu'elles sont plus facilement différentiables lors de la phase de regroupement.

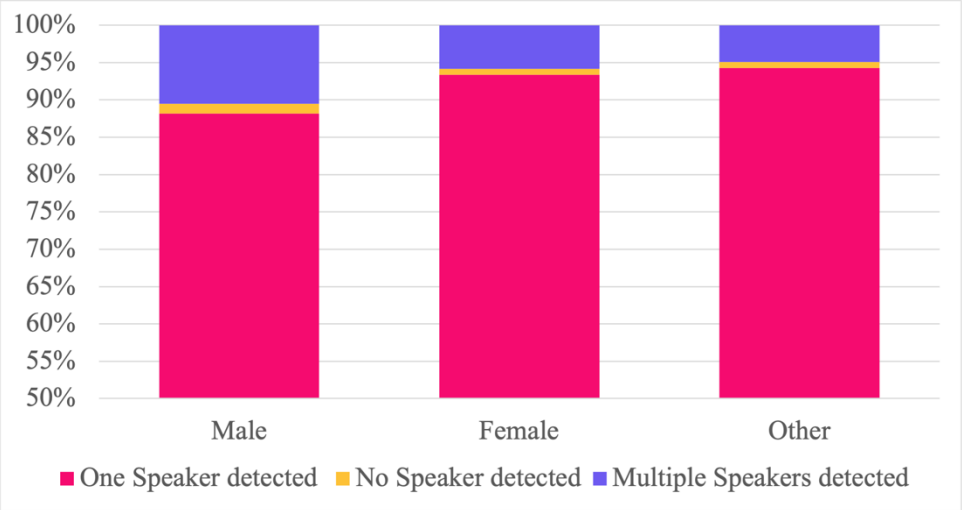

Figure 37 : Résultats obtenus par la diarisation selon l'attribut « Sexe »

Tableau 17 : Résultat de l'analyse de biais selon le critère de sexe déclaré des différents locuteurs

	Homme	Femme	Autre
p_0	1.31	0.77	0.79
DFR	**88.20**	**93.39**	**94.27**
p_+	10.49	5.84	4.94

Tableau 18 : Incertitudes liées à l'intervale de confiance à 99 % pour les résultats concernant le sexe des locuteurs

	Homme	Femme	Autre
$\varepsilon(p_0)$	0,39	0,65	1,60
$\varepsilon(DFR)$	0,14	0,23	0,61
$\varepsilon(p_+)$	0,37	0,61	1,50

III.3. Accent du locuteur

Si la base de données Commonvoice n'est pas très équilibrée en ce qui concerne les accents des locuteurs prononçant des phrases (cf. Figure 38), on note néanmoins un certain nombre de résultats intéressants.

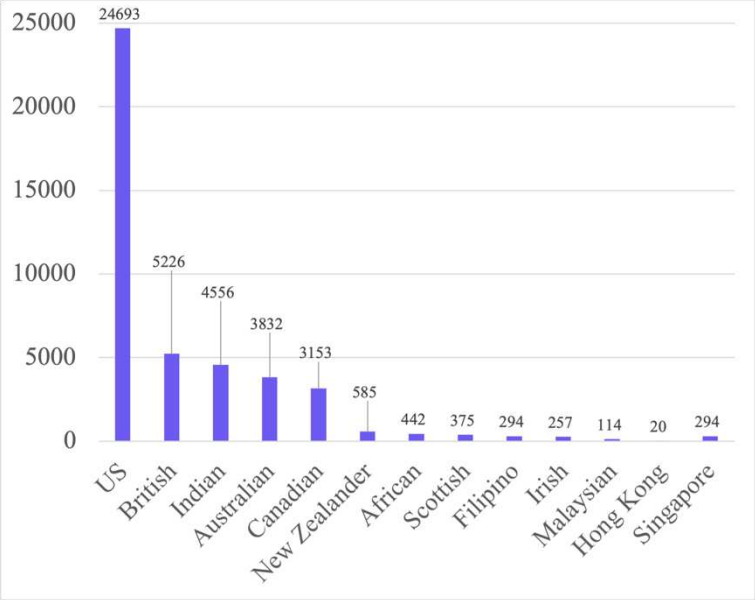

Figure 38 : Répartition des différentes classes de l'attribut « Accent » de la base de données Commonvoice

On remarque par exemple que si les accents américains et britanniques sont sensiblement traités de la même façon par la diarisation testée, les accents indiens, africains, irlandais sont moins bien traités par cet algorithme. Ceci se produit en particulier au niveau de la phase de regroupement, puisque de nombreuses fois les enregistrements sont catégorisés comme étant prononcés par plusieurs locuteurs.

On peut proposer plusieurs explications à ces résultats qui devront être testées :

- Les erreurs proviennent peut-être de l'entraînement du réseau extracteur des x-vecteurs. Celui-ci a été entrainé sur VoxCeleb 1 et 2 qui ne propose pas de labels concernant l'accent des locuteurs. Il est possible que lors de l'apprentissage, le réseau ait sur-appris des caractéristiques liées aux accents des locuteurs pour les discriminer.
- Il est également possible que ces erreurs viennent des prosodies qui diffèrent entre les différents accents. Une prosodie plus lente, par exemple d'une personne parlant une langue autre que sa langue maternelle, peut laisser penser que deux personnes s'expriment.

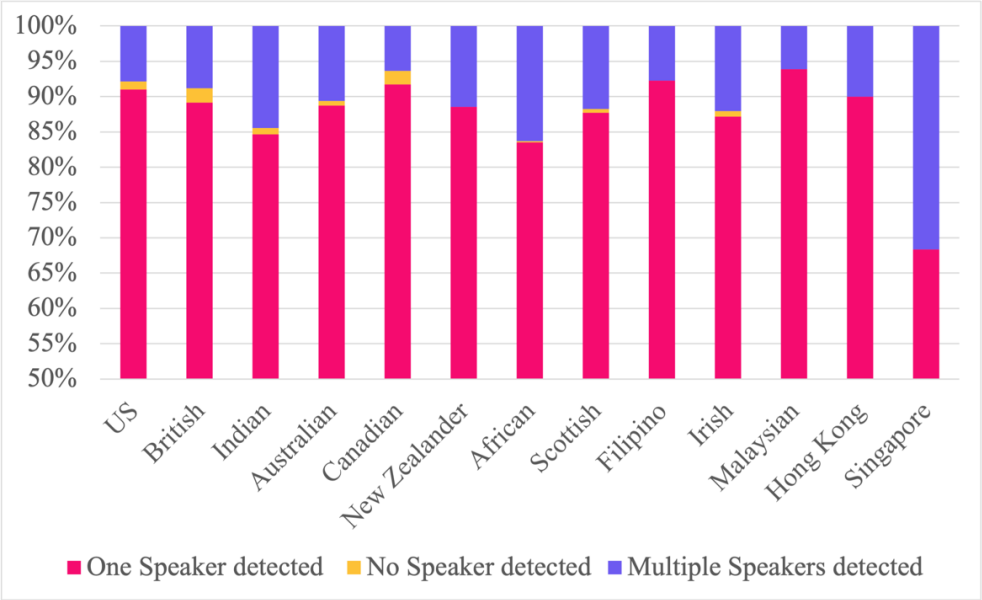

Figure 39 : Résultats obtenus par la diarisation selon l'attribut « Accent »

Tableau 19 : Résultat de l'analyse de biais selon le critère d'accent des différents locuteurs

	Américain	Britannique	Indien	Australien	Canadien	Neo-Zelandais	Africain	Ecossais	Philippin	Irlandai	Malaisien	Hong-Kongais	Singapourien
p_0	1,17	2,05	0,92	0,63	1,93	0,00	0,23	0,53	0,00	0,78	0,00	0,00	0,00
DFR	**90,98**	**89,13**	**84,64**	**88,75**	**91,69**	**88,55**	**83,48**	**87,73**	**92,24**	**87,16**	**93,86**	**90,00**	**68,37**
p_+	7,85	8,82	14,44	10,62	6,37	11,45	16,29	11,73	7,76	12,06	6,14	10,00	31,63

Tableau 20 : Incertitudes liées à l'intervale de confiance à 99 %
pour les résultats concernant l'accent des locuteurs

	Américain	Britannique	Indien	Australien	Canadien	Neo-Zelandais	Africain	Ecossais	Philippin	Irlandai	Malaisien	Hong-Kongais	Singapourien
$\varepsilon(p_0)$	0,47	1,11	1,38	1,32	1,27	3,40	4,56	4,37	4,03	5,38	5,80	17,31	7,00
$\varepsilon(DFR)$	0,18	0,51	0,37	0,33	0,63	0,00	0,58	0,97	0,00	1,41	0,00	0,00	0,00
$\varepsilon(p_+)$	0,44	1,01	1,34	1,28	1,12	3,40	4,53	4,29	4,03	5,24	5,80	17,31	7,00

III.4. Longueur de phrase

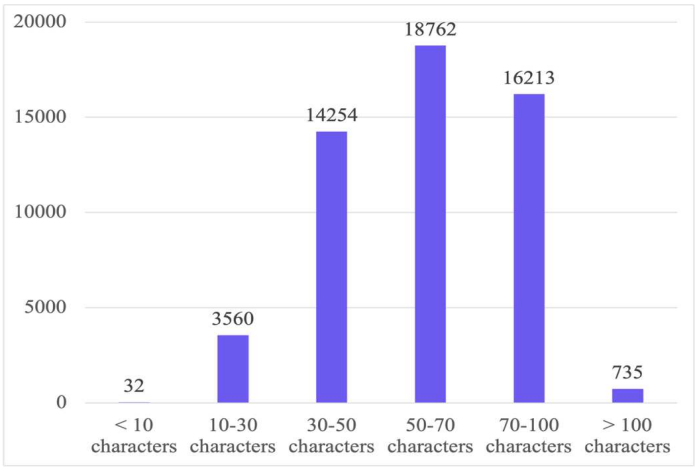

Figure 40 : Répartition des différentes classes de l'attribut « Longueur de phrase » généré à partir des données de la base Commonvoice

Les résultats concernant les longueurs des phrases prononcées étaient surement les plus attendus. En effet, on obtient logiquement que plus la phrase prononcée est courte, plus la diarisation, et plus spécifiquement la détection d'activité vocale, ne réussit pas à détecter la présence de parole dans l'enregistrement. De même la littérature montre que l'identification d'un locuteur par sa voix est plus complexe s'il parle peu (Charlet et al., 2015). Pour les phrases plus longues en revanche le DFR reste stable (cf. Tableau 21 et Figure 41).

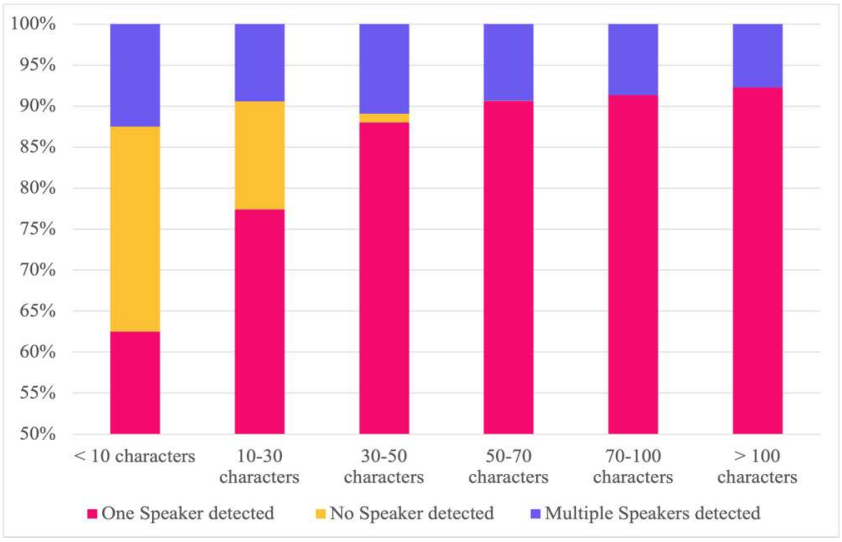

Figure 41 : Résultats obtenus par la diarisation selon l'attribut « Longueur de phrase »

Ce résultat, même s'il paraît anodin, peut en réalité remettre en cause l'une des propriétés fondamentales de la diarisation : l'indépendance des résultats selon la langue parlée. En effet plusieurs études ont montré (Futrell et al., 2015) que le nombre de mots par phrase dépend de la langue parlée, ainsi un biais de la diarisation concernant la durée des phrases prononcées peut se traduire par un biais sur les langues compatibles avec la diarisation. Des langues comme l'hébreu ou l'italien, dont les locuteurs prononcent en moyenne moins de mots qu'en anglais, pourraient ainsi poser davantage de problèmes aux algorithmes de diarisation du locuteur.

Tableau 21 : Résultat de l'analyse de biais selon le critère de longueur des phrases prononcées par les différents locuteurs

	< 10 c	10-30 c	30-50 c	50-70 c	70-100 c	> 100 c
p_0	25,00	13,20	1,08	0,05	0,00	0,00
DFR	**62,50**	**77,39**	**87,99**	**90,60**	**91,32**	**92,24**
p_+	12,50	9,41	10,93	9,34	8,68	7,76

Tableau 22 : Incertitudes liées à l'intervale de confiance à 99 % pour les résultats concernant la longueur des phrases prononcées par les différents locuteurs

	< 10 c	10-30 c	30-50 c	50-70 c	70-100 c	> 100 c
$\varepsilon(p_0)$	22,08	1,81	0,70	0,55	0,57	2,55
$\varepsilon(DFR)$	19,75	1,46	0,22	0,04	0,00	0,00
$\varepsilon(p_+)$	15,08	1,26	0,67	0,55	0,57	2,55

IV. Limites de cette approche

Cette approche nouvelle de détection des biais de la diarisation du locuteur présente toutefois un certain nombre de limites.

En particulier, elle ne tient pas compte du caractère multi-locuteurs des conversations habituellement analysées par la diarisation. En effet, afin de pouvoir travailler sur une base de données qui contenait suffisamment d'information pour l'étude des biais présentée, nous avons dû faire le choix de nous restreindre au seul cas où une unique personne prend la parole. Si une telle analyse ne permet pas d'identifier de façon exhaustive l'ensemble des biais de la diarisation du locuteur, elle permet néanmoins de tester une forme de spécificité de la diarisation, c'est-à-dire sa capacité à ne pas identifier deux locuteurs lorsqu'il n'y en aurait qu'un d'actif.

C'est d'ailleurs cette spécificité qui sera privilégiée dans une majorité d'applications industrielles, dont celle de Newsbridge, puisqu'on préfère regrouper à tort deux locuteurs différents, plutôt que d'avoir une multitude de segments dont l'affichage et la transcription seraient complexes.

IV.1. Protocole envisagé pour étendre l'étude

IV.1.1. Base de données idéale

Afin d'avoir une évaluation encore plus précise de la justesse de la diarisation du locuteur on peut imaginer étendre le protocole présenté précédemment au cas où plusieurs locuteurs conversent. Pour ce faire et afin de conserver l'unicité des changements de paramètre entre les expériences, il sera nécessaire de construire les conversations étudiées.

Prenons un exemple, soit une liste de phrases composant un jeu de données :

Tableau 23 : Jeu de données fictif permettant une étude minimale des biais dans une conversation multi-locuteurs

Numéro de phrase	Phrase prononcée	Locuteur	Age	Sexe
1	A	1	20-30 ans	M
2	B	2	20-30 ans	F
3	A	2	20-30 ans	F
4	C	3	40-50 ans	M
5	B	3	40-50 ans	M
6	B	4	10-20 ans	F

Si l'on dispose d'un tel jeu de données (Tableau 23), pour l'analyse des biais de la diarisation dans le contexte d'une courte conversation entre deux locuteurs seules les comparaisons suivantes sont pertinentes :

- Phrases 1 et 2 comparées avec les Phrases 1 et 4 : Ce qui nous donnerait une indication sur l'impact de l'âge du locuteur.
- Phrases 1 et 2 comparées avec les Phrase 1 et 3 : Ce qui nous donnerait une indication sur l'impact du contenu de la phrase prononcée (longueur par exemple).
- Toute autre combinaison ne permettrait en revanche pas de conclure car, en ce cas, plusieurs paramètres varieraient.

En pratique, une analyse précise des biais dans un contexte multi-locuteurs nécessiterait un jeu de données beaucoup plus conséquent, que l'on pourrait construire en généralisant l'exemple ci-dessus avec des plans en carrés latins[28].

IV.1.2. Nouveau protocole

Une fois en possession d'un tel jeu de données, il deviendra plus simple de mener une étude des biais pour la diarisation dans le contexte d'une conversation entre plusieurs locuteurs.

On pourra ainsi procéder à deux analyses :

- De façon similaire à l'analyse présentée précédemment on pourrait vérifier que la diarisation parvient bien à identifier le nombre de locuteurs de chacun de nos enregistrements de test. La réponse attendue ne serait plus systématiquement 1 mais on pourrait la déterminer à partir de la vérité terrain.
- On pourrait également calculer sur l'ensemble des médias la somme des JER selon chaque attribut étudié. En effet le JER permet de calculer de façon non pondérée le taux d'erreur de diarisation dans un enregistrement pour chaque locuteur.

28. https://fr.wikipedia.org/wiki/Carr%C3%A9_latin

Chapitre VI : Consommation énergétique de la diarisation appliquée à grande échelle

I. Diarisation économe en énergie

Avoir une diarisation économe en énergie permet d'analyser la structure d'un discours dans de nombreuses situations où cela n'était pas envisageable précédemment pour des raisons pratiques, énergétiques et financières.

Dans notre cas de figure, si on souhaite traiter plusieurs centaines de milliers d'heures de vidéos d'actualité archivées[29], la limitation est triple. Tout d'abord il serait très compliqué de disposer de suffisamment de processeurs graphiques pour paralléliser les calculs et obtenir les résultats d'analyse après un temps acceptable.

De plus, la consommation énergétique de telles analyses serait énorme et leurs coûts trop importants pour faire de telles analyses une réalité économique.

C'est pourquoi, il apparaît nécessaire de travailler à la conception d'architectures basses consommation de diarisation du locuteur.

II. Contribution : Mise en production d'un algorithme de diarisation du locuteur pour l'analyse multimédia à grande échelle

Afin de pouvoir intégrer l'algorithme de diarisation du locuteur choisi dans l'infrastructure de calcul de Newsbridge, il a été nécessaire de comparer différents algorithmes selon plusieurs critères, autres que les performances. En particulier, pour

29. Pour donner un ordre de grandeur, une chaîne de télévision locale d'un état américain comme la Caroline du Nord possède à elle seule dans ses archives plus de 500 000 heures de contenus filmés.

une utilisation réaliste en production, le système déployé doit pouvoir fonctionner relativement rapidement sur une machine peu consommatrice d'énergie.

L'ensemble des tests effectués l'ont été sur des machines du fournisseur *cloud* Amazon Web Services (AWS) avec lequel Newsbridge travaille.

Concernant la consommation énergétique, les calculs ont été réalisés avec la formule suivante :

$$CE = Pe * \frac{t_u}{3600}$$

Avec *Pe*, la quantité d'énergie primaire utilisée pour une heure de fonctionnement de la machine considérée[30] et t_u le temps d'utilisation pour l'inférence de la diarisation sur un média d'une heure (journal télévisé).

Tableau 24 : Détails des tests effectués pour comparer la rapidité et la consommation énergétique des algorithmes de diarisation du locuteur les plus adaptés à l'analyse multimédia à grande échelle

Méthode de diarisation utilisée	Machine utilisée	Processeur principal de calcul	Temps d'inférence (en secondes)	Consommation énergétique (en kilojoules)
Pyannote 2.1	AWS c5.xlarge	CPU	14 088,01	1 174,00
MSVAD	AWS p3.2xlarge	GPU	2 804,77	545,37
MSVAD	AWS g4dn.xlarge	GPU	2 509,35	209,11
MSVAD	AWS c5.xlarge	CPU	2 249,71	187,48
Diart	AWS p3.2xlarge	CPU	935,14	181,83
Diart	AWS p3.2xlarge	GPU	123,86	24,08
Pyannote 2.1	AWS p3.2xlarge	GPU	98,04	19,06

30. Obtenue grâce à un calculateur en ligne : https://doc.api.boavizta.org/getting_started/single_cloud_instance/

Diart	AWS g4dn.xlarge	CPU	867,34	7,23
Diart	AWS c5.xlarge	CPU	838,25	6,99
Pyannote 2.1	AWS g4dn.xlarge	GPU	171,77	1,43
Diart	AWS g4dn.xlarge	GPU	120,51	1,00

Il est intéressant de noter que dans une situation de *cloud computing*, c'est-à-dire lorsque les traitements se font sur des serveurs distants, il devient également primordial de considérer le mix énergétique[31] local pour mesurer l'impact d'une solution technologique sur l'environnement. Pour cette thèse, la quasi-totalité des machines utilisées pour l'entraînement, la conception ou les tests de nos modèles et algorithmes étaient situées dans l'Union Européenne, au sein de centres de données en Irlande ou en France dont voici en Figure 42 les mix énergétiques respectifs.

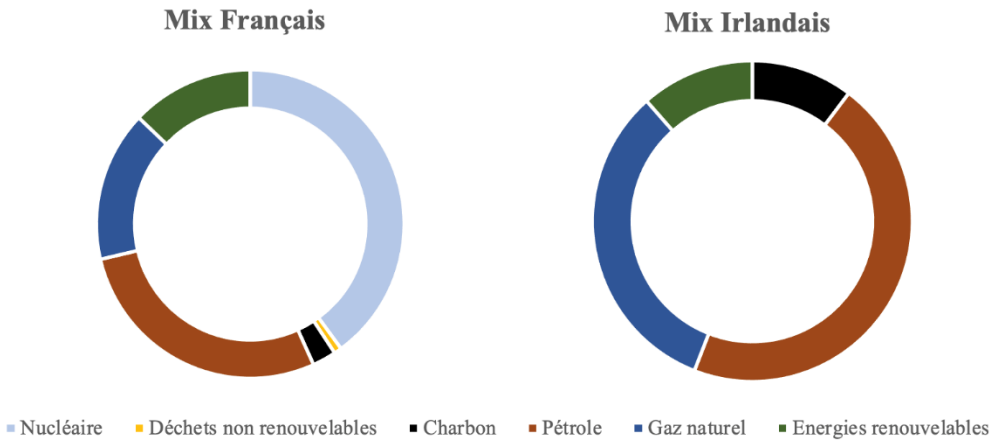

Figure 42 : Mix énergétiques français et irlandais respectivement en 2020 et 2018. Sources : Eurostat et Agence internationale de l'énergie

Finalement notre choix s'est porté vers la solution la moins consommatrice d'énergie (cf. Tableau 24) qui sert désormais à analyser plusieurs centaines d'heures arrivant quotidiennement sur la plate-forme Newsbridge. Cette solution inspirée de Diart (Coria et al., 2021) a pour avantage de proposer des performances à l'état de l'art, tout en

31. https://fr.wikipedia.org/wiki/Mix_%C3%A9nerg%C3%A9tique

permettant d'analyser le contenu au fur et à mesure de telle sorte qu'il devient possible d'en commencer la transcription avant la fin de la diarisation.

III. Applications médicales de la diarisation

Les technologies de diarisation du locuteur ayant beaucoup évolué ces dernières années, elles trouvent de nouveaux usages et ce notamment dans le secteur de la santé.

On peut citer (Robin et al., 2017), (Riad et al., 2022) et (Shafey et al., 2019) mais on va ici surtout s'intéresser à l'usage de la diarisation du locuteur pour permettre une amélioration des technologies qui permettent un suivi et un maintien à domicile des personnes âgées.

En effet, d'après l'INSEE la population française âgée de plus de 65 ans est passée de 19,7 % en 2018 à 20,5 % en 2020. On retrouve de tels phénomènes plus largement en Europe, et dans le reste du monde. De plus en plus de personnes sont donc sujettes à vouloir rester vieillir chez elles et requièrent de plus en plus de solutions technologiques permettant d'adapter leurs habitats à leur mode de vie.

Beaucoup de ces solutions font intervenir de l'analyse du signal audio que ce soit pour détecter certains sons synonymes de détresse ou alors pour reconnaître les paroles d'un utilisateur et lui permettre d'utiliser un assistant vocal[32] (cf. Figure 43).

Grâce à la diarisation du locuteur, il est également devenu possible de compter le nombre de personnes présentes dans une pièce en y surveillant les interactions parlées. Le décompte du nombre de locuteur permet plusieurs avancées pour le maintien à domicile des personnes âgées parmi lesquelles :

- En termes de domotique, la possibilité de connaître le nombre de personnes présentes dans un lieu de vie permet de réguler la température par exemple.
- Un suivi de l'isolement précis, en captant si la personne a de la visite ou si elle reste seule trop longtemps.
- En cas d'interaction avec un assistant vocal, en permettant à cet assistant de cibler la voix de son utilisateur principal.
- L'amélioration des systèmes d'urgence en adaptant les protocoles d'urgence selon que la personne soit seule chez elle ou non.

32. Google Assistant, Amazon Alexa ou une autre solution plus spécialisée comme un robot compagnon mobile (cf. III.1).

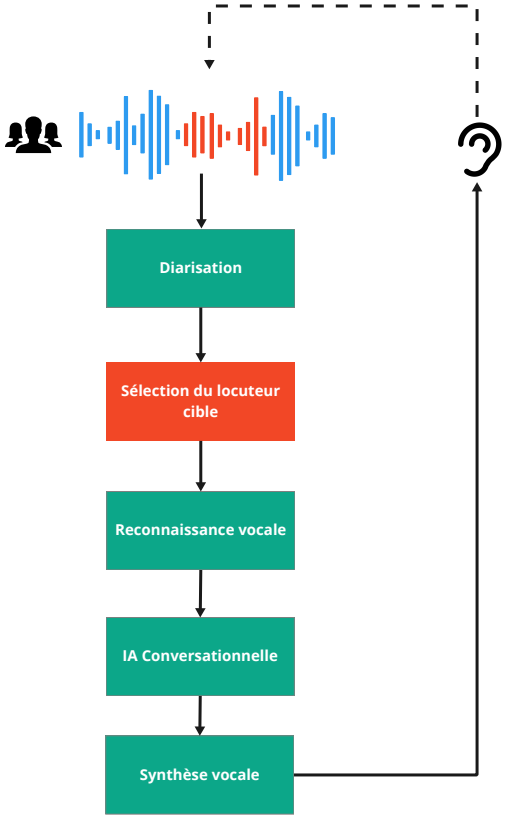

Figure 43 : Schéma d'un possible système d'assistant vocal utilisant la diarisation pour cibler son interlocuteur dans un environnement multi-locuteurs

En supposant que la personne discute avec les personnes qui lui rendent visite, on peut détecter via des microphones placés intelligemment dans l'habitation que plusieurs locuteurs parlent, et combiner cette information aux autres systèmes déjà présents dans l'habitat adapté (cf. Figure 44).

De plus, de la sorte, on parvient à identifier les moments où la personne parle au téléphone, seule chez elle.

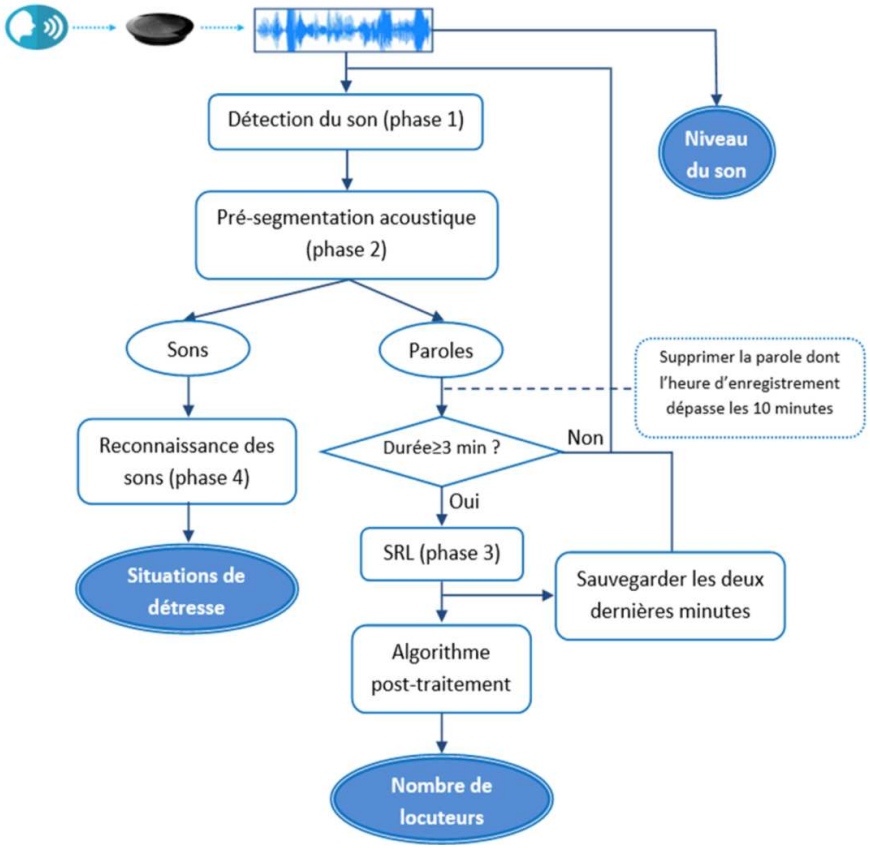

Figure 44 : Schéma descriptif d'un système de diarisation du locuteur (Boutamine et al., 2019) intégré dans une « Smart home »

III.1. Projet E-vita

Le projet européen E-vita[33], réalisé en partenariat avec le Japon vise à regrouper de nombreux partenaires (12 en Europe, 10 au Japon) académiques, industriels et institutionnels pour la création d'un coach virtuel à destination des personnes âgées dépendantes.

Figure 45 : Logo du projet E-vita

33. https://www.e-vita.coach/

Ce projet mêle des expertises en science de l'information, en ingénierie, psychologie, sociologie et robotique pour proposer un coach virtuel qui soit le plus utile possible, et pour assurer sa bonne adoption par les utilisateurs finaux, et ce quelque soient leurs différences culturelles.

IV. Contribution : application des méthodes récentes de diarisation au domaine médical

Dans le cadre de ce projet, mené à l'Institut Mines Telecom par le professeur Jérôme Boudy, nous avons souhaité réactualiser les travaux sur la diarisation du locuteur appliquée au maintien à domicile des personnes âgées dépendantes.

C'est pourquoi, en nous appuyant les précédents travaux ayant été réalisés sur ce sujet (cf. Figure 44 & Figure 46), nous avons cherché à améliorer le système d'interactions entre le « robot » et la personne âgée utilisatrice grâce aux méthodes récentes de diarisation du locuteur.

Le système présenté précédemment présente deux limites majeures :

- D'une part, celle de ses performances qui s'avèrent insuffisantes pour une utilisation en conditions réelles. Sur notre base de données de test, nous avons établi que seuls 33,5 % des enregistrements audio sont caractérisés par la diarisation *LIUM_SpkDiarization* avec le nombre correct de locuteurs actifs.
- D'autre part sa relativement lente vitesse d'exécution, puisque qu'on doit attendre d'avoir regroupé au minimum 3 minutes de parole consécutive pour lancer une analyse. Dans le cas de l'amélioration des systèmes de détection de situation de détresse grâce à la diarisation, ce gain de temps peut s'avérer précieux.

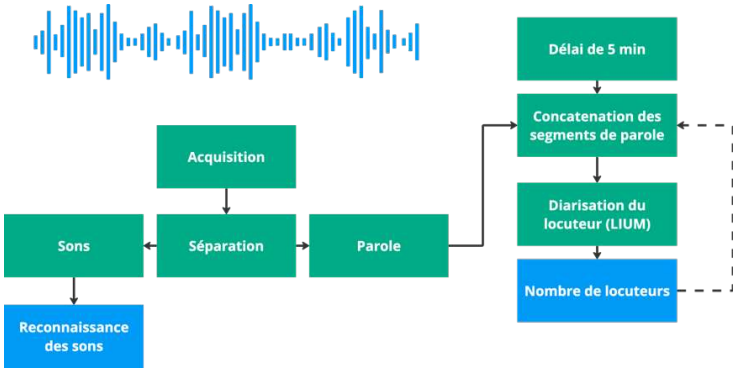

Figure 46 : Schéma du système tel que présenté dans (Boutamine et al., 2019) avec une diarisation du locuteur effectuée sur des segments de parole de 3 à 5 min extraits régulièrement

Méthode LIUM_Spkdiarization

Cette méthode de diarisation du locuteur, introduite et rendue disponible dans un outil *open-source*[34] en 2010 (Meignier and Merlin, 2010; Rouvier et al., 2013), s'appuie sur un regroupement basé sur le critère d'information bayésien (BIC) de représentations vectorielles de la parole, elles-mêmes obtenues grâce à 8 GMMs. S'en suivent une resegmentation grâce à l'algorithme de Viterbi, et éventuellement un second regroupement avec un critère d'arrêt différent, comme par exemple le *cross likelihood ratio* (CLR).

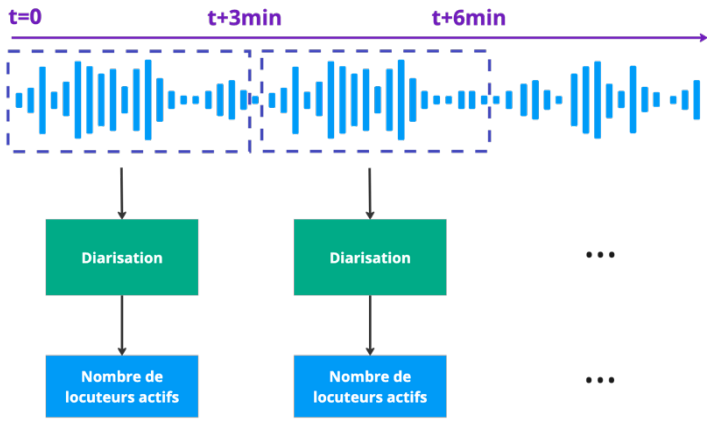

Figure 47 : Schéma simplifié du système que l'on se propose d'améliorer

34. https://projets-lium.univ-lemans.fr/spkdiarization/

Notre contribution, présentée dans (Tevissen, Istrate, et al., 2023), a consisté en l'amélioration du système de diarisation du locuteur utilisé dans le contexte du maintien à domicile des personnes âgées. Pour ce faire, nous avons décidé d'intégrer au système précédent deux algorithmes récents de l'état de l'art pour évaluer leurs performances et déterminer s'ils rendent cette utilisation de la diarisation plus réaliste.

Nous avons choisi d'utiliser l'algorithme de détection d'activité vocale multi-flux MSVAD présenté précédemment ainsi que Diart qui présente l'avantage de fonctionner en quasi temps-réel. Ces deux solutions seront évaluées selon leur capacité à déterminer avec précision le nombre de locuteurs s'exprimant dans une cinquantaine d'enregistrements effectués dans les conditions réelles d'une pièce à vivre, avec un microphone *Jabra Speak 510*.

Tableau 25 : Résultats du pourcentage d'enregistrements pour lesquels le nombre de locuteur a été correctement prédit

Méthode	Pourcentage d'enregistrement correctement analysés
LIUM_SpkDiarization	(32,5 ± 2,4) %
Diart (en-ligne)	(57,5 ± 3,1) %
MSVAD	**(77,5 ± 3,6) %**

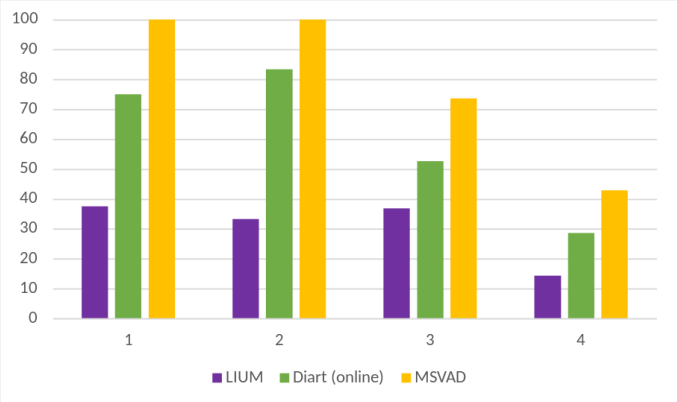

Figure 48 : Détails des résultats selon le nombre de locuteurs actifs dans l'enregistrement

Tableau 26 : Résultats de l'écart relatif entre le nombre de locuteurs prédits et le nombre de locuteurs réels

Méthode	Distance relative avec la verité terrain
LIUM_SpkDiarization	0,9 locuteurs
Diart (en-ligne)	0,3 locuteurs
MSVAD	**0,1 locuteurs**

On constate alors assez logiquement que nos deux systèmes se sont systématiquement avérés meilleurs que l'approche précédemment utilisée pour cette tâche. La MSVAD affiche des scores plus de deux fois meilleurs et Diart reste systématiquement entre les deux solutions. Ces résultats, biens que très positifs pour notre application, sont logiques puisque la méthode précédemment utilisé s'appuie sur des GMMs, peu robustes aux changements de domaines, et sujettes aux erreurs de segmentation (Mathieu Ben et al., 2004), en particulier dans les conditions difficiles de la diarisation à domicile.

La dernière étape de notre étude a consisté à évaluer la faisabilité du déploiement de tels algorithmes à domicile.

IV.1. Utilisation des algorithmes de diarisation en environnement embarqué

Dans cette optique d'une évaluation de l'utilisation de nos algorithmes dans le contexte du maintien des personnes âgées à domicile, il est essentiel d'envisager leur déploiement sur du matériel embarqué, c'est-à-dire sur des appareils souvent moins puissants que ceux que nous pouvons utiliser habituellement. C'est pourquoi, nous avons déployé la MSVAD et Diart sur une Raspberry Pi 4[35], un ordinateur miniature, doté pour seule puissance de calcul d'un processeur quatre cœurs Cortex-A72 (ARMv8).

Ce qui a nécessité plusieurs adaptations, notamment en ce qui concerne la détection de l'activité vocale multi-flux, qui s'est avérée complexe à faire fonctionner sur une telle architecture embarquée. En effet, si nous faisions habituellement tourner ce système en exécutant trois algorithmes de détection d'activité vocale en parallèle, dans cette situation, il nous a été impossible de faire tourner la GP-VAD, par manque de

35. https://www.raspberrypi.com/products/raspberry-pi-4-model-b/

mémoire. Nous avons donc dû adapter notre système pour qu'il ne fonctionne qu'avec deux détections d'activité vocale en parallèle.

Pour la méthode Diart nous avons également dû procéder à plusieurs adaptations minimes, notamment en ce qui concerne la durée minimale des segments audio considérés, afin de garantir une rapidité d'analyse suffisante. Partant du principe que les résultats précis de diarisation ne nous intéressent pas, mais que nous sommes plus concernés par le nombre de locuteurs présents dans la pièce, nous avons choisi de considérer des segments de parole de cinq secondes comme unité atomique de diarisation.

Tableau 27 : Temps d'inférence des différentes méthodes de diarisation du locuteur utilisées sur une raspberry pi pour analyser un segment audio de 3 minutes

	Temps d'inférence
LIUM_SpkDiarization[36]	113,4 s
MSVAD	639,4 s
Diart (0,5 s *step*)	293,7 s
Diart (5 s *step*)	**30,1 s**

Après toutes ces adaptations, nous avons finalement pu obtenir des résultats très encourageants, notamment avec Diart qui analyse un fichier audio environ 6 fois plus rapidement que sa durée initiale (cf. Tableau 27), rendant ainsi possible l'utilisation de cet algorithme à domicile sur du matériel embarqué. La méthode du LIUM reste intéressante pour sa rapidité d'exécution mais nous ne la choisissons pas en raison de ses moindres performances.

36. Temps obtenu sur un CPU Intel i5 à 2,4 GHz

Conclusion sur la diarisation responsable

Dans ce chapitre, nous avons pu montrer l'importance de considérer d'autres aspects que les performances pour choisir un système de diarisation du locuteur. Nommément nous avons vu qu'il est primordial de garantir la justesse de tels systèmes ainsi que leur faible consommation énergétique.

On a pu voir les difficultés de l'évaluation de la justesse des algorithmes de diarisation compte tenu de la complexité inhérente aux conversations humaines. Néanmoins, on a pu proposer un protocole pour son évaluation et obtenir des premiers résultats dans un cas simple.

On a pu établir la faisabilité de l'usage de la diarisation du locuteur sur des machines peu consommatrices d'énergie, ce qui rend possible de nouveaux usages de la diarisation pour les secteurs des médias et de la santé. Dans ce dernier cas, on a d'ailleurs pu proposer un nouveau système qui profite pleinement des différentes avancées des méthodes récentes de diarisation du locuteur.

Ainsi, on a pu voir les multiples prérequis à considérer afin de rendre possible l'usage des algorithmes de diarisation du locuteur, à grande échelle, et de façon responsable.

Conclusion

I. Conclusion

La diarisation du locuteur joue un rôle clef dans les applications modernes d'indexation audiovisuelle des contenus diffusés à la télévision. Son utilisation permet de mieux structurer, transcrire et, finalement, comprendre les conversations humaines archivées.

Nous avons vu que l'état de l'art de la diarisation du locuteur est riche par les différentes approches existantes pour traiter la tâche. En particulier on peut souligner la coexistence des approches bout-en-bout et modulaires. Nous avons pu présenter deux nouveaux systèmes ayant pour but d'améliorer les performances des algorithmes actuels de diarisation en contexte réel.

Premièrement, avec la détection d'activité vocale multi-flux, nous avons démontré l'importance et l'impact des premières phases d'un algorithme de diarisation. En particulier en concentrant notre travail sur la détection d'activité vocale, dont les approches sont aussi nombreuses, nous avons montré qu'il était possible de battre les méthodes de l'état de l'art sur certains enregistrements.

Ensuite, nous avons exploré les approches multimodales de la diarisation du locuteur, en commençant par les approches qui intègrent des éléments visuels, sans succès pour nos cas d'usages. Nous avons dans un second temps pu voir comment les récentes avancées en traitement du langage naturel et, notamment de quelle façon les modèles de langage volumineux, pourront améliorer la façon dont nous comprenons automatiquement les conversations humaines.

Enfin, dans l'optique d'utiliser ces algorithmes à grande échelle, nous avons pu souligner l'importance de développer des méthodes algorithmiquement justes. Pour ce faire, nous avons posé les premières pierres d'une méthode qui permet d'identifier les éventuels biais d'un système de diarisation du locuteur.

Ces travaux nous ont également amené à réfléchir à la consommation énergétique des algorithmes de diarisation. Nous proposons sur ce sujet des premières pistes de réflexion et avons pu déployer un des algorithmes de l'état de l'art sur une architecture embarquée, adaptée à un usage à domicile.

Si la diarisation a connu ces dernières années d'importants progrès, nous souhaitons conclure sur le fait qu'il reste encore de nombreux défis pour atteindre des modèles de diarisation robustes et justes en contexte réel. L'adage humoristique et populaire dans le domaine de la recherche en apprentissage automatique « *no such thing as a free*

lunch »[37] nous semble rester encore aujourd'hui très vrai dans le domaine de la diarisation du locuteur.

En effet, malgré les importantes avancées permises par les nouveaux algorithmes de diarisation et les méthodes de fusion multimodale, les passages où plusieurs locuteurs se coupent mutuellement la parole restent très difficiles à traiter. De la même façon, détecter avec précision les passages à diariser, c'est-à-dire contenant de la parole, peut s'avérer particulièrement complexe en contexte réel.

II. Perspectives

Nous pensons que pour aller vers des systèmes de diarisation du locuteur plus robustes et encore plus utiles, le principal défi réside dans la gestion de la parole superposée. Il est probable qu'à l'avenir les recherches en diarisation du locuteur se concentrent encore davantage sur ce problème de la détection et de la gestion des passages contenant des paroles prononcées simultanément par plusieurs locuteurs. En effet ce problème reste encore largement irrésolu et il devient probablement nécessaire de combiner certaines des approches décrites précédemment avec des algorithmes de séparation de source, afin de pouvoir ensuite transcrire indépendamment ces passages dans les médias analysés.

Il faudra également poursuivre les travaux sur la justesse de la diarisation du locuteur, par la création d'un jeu de données adapté à son évaluation afin de pouvoir s'assurer que les modèles entrainés sur de plus en plus de données n'intègrent pas de nouveaux biais.

Enfin, nous envisageons de poursuivre nos recherches en utilisant la diarisation du locuteur comme le point d'entrée d'autres algorithmes d'analyse des contenus multimédias. Une telle approche de *diarisation-as-a-modality* s'apparente d'ailleurs à d'autres travaux menés par Newsbridge sur l'identification multimodale de personnes dans des contenus télévisés.

Plus généralement, nous poursuivrons certainement nos recherches vers des méthodes de fusion multimodale asynchrone (Bloch, 2003) permettant notamment d'intégrer davantage de modalités, en considérant par exemple des éléments de contexte sur la nature du média analysé ou des informations plus larges sur la sémantique employée par chaque locuteur.

37. https://en.wikipedia.org/wiki/No_free_lunch_theorem

Bibliographie

Ajmera, J., and Wooters, C. (2003) A robust speaker clustering algorithm. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. (2023) Falcon-40B: an open large language model with state-of-the-art performance.

Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012) Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing* 20(2): 356–370.

Anidjar, O. H., Hajaj, C., Dvir, A., and Gilad, I. (2021) A Thousand Words are Worth More Than One Recording: NLP Based Speaker Change Point Detection. In *Interspeech 2021*.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020) Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Auguste, R., Martinet, J., and Tirilly, P. (2015) Space-time Histograms And Their Application To Person Re-identification In TV Shows. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. Presented at the ICMR '15: International Conference on Multimedia Retrieval Shanghai China: ACM.

Auguste, R., Tirilly, P., and Martinet, J. (2015) Introducing FoxPersonTracks: A benchmark for person re-identification from TV broadcast shows. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. Presented at the 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI) Prague, Czech Republic: IEEE.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020, October 22) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv.

Barras, C., Xuan Zhu, Meignier, S., and Gauvain, J.-L. (2006) Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing* 14(5): 1505–1512.

Béchet, F., Bendris, M., Charlet, D., Damnati, G., Favre, B., Auguste, R., Bigot, B., Dufour, R., Fredouille, C., Senay, G., Tirilly, P., and Martinet, J. (2015) Identification de personnes dans des flux multimédia. In *CORIA 2015*.

Bechet, F., Favre, B., and Damnati, G. (2012) Detecting person presence in TV shows with linguistic and structural features. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Presented at the ICASSP 2012 - 2012 IEEE International Conference on Acoustics, Speech and Signal Processing Kyoto, Japan: IEEE.

Bendris, M. (2011) *Indexation audio-visuelle des personnes dans un contexte de télévision*. Télécom ParisTech.

Bendris, M., Favre, B., Charlet, D., Damnati, G., Senay, G., Auguste, R., and Martinet, J. (2013) Unsupervised face identification in TV content using audio-visual sources. In *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*. Presented at the 2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI) Veszpr m , Hungary: IEEE.

Blei, D. M., and Frazier, P. I. (2011) Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research* 12: 2461–2488.

Bloch, I. (2003) *Fusion d'informations en traitement du signal et des images*. (Hermes Science Publications.).

Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G. (2008) Overlapped speech detection for improved speaker diarization in multiparty meetings. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Boeddecker, C., Heitkaemper, J., Schmalenstroeer, J., Drude, L., Heymann, J., and Haeb-Umbach, R. (2018) Front-end processing for the CHiME-5 dinner party scenario. In *5th International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*.

Bost, X., Linarès, G., and Gueye, S. (2015) Audiovisual speaker diarization of TV series. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Boutamine, S., Istrate, D., Boudy, J., and Tannous, H. (2019) Smart Sound Sensor to Detect the Number of People in a Room. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Bredin, H. (2023) pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*. Presented at the INTERSPEECH 2023 ISCA.

Bredin, H., Barras, C., and Guinaudeau, C. (2016) Multimodal Person Discovery in Broadcast TV at MediaEval 2016. In *MediaEval 2016*.

Bredin, H., and Laurent, A. (2021) End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech 2021*.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020) pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Broux, P.-A. (2020) *Segmentation et regroupement en locuteurs dans des documents audiovisuels, en interaction avec des annotateurs humains*. Université du Maine.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023, April 13) Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/abs/2303.12712.

Bullock, L., Bredin, H., and Garcia-Perera, L. P. (2019) Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Cai, Q., Hong, G., Ye, Z., Li, X., and Li, H. (2022) The Kriston AI System for the VoxCeleb Speaker Recognition Challenge 2022. In *VoxCeleb Speaker Recognition Challenge 2022*.

Canseco, L., Lamel, L., and Gauvain, J.-L. (2005) A comparative study using manual and automatic transcriptions for diarization. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*

Charlet, D., Poignant, J., Bredin, H., Fredouille, C., and Meignier, S. (2015) What Makes a Speaker Recognizable in TV Broadcast? Going Beyond Speaker Identification Error Rate. In *ERRARE Workshop, a Satellite Event of Interspeech 2015.* Sinaia, Romania.

Chhabra, A., Masalkovaite, K., and Mohapatra, P. (2021) An Overview of Fairness in Clustering. *IEEE Access* 9: 130698–130720.

Chung, J. S., Huh, J., Nagrani, A., Afouras, T., and Zisserman, A. (2020) Spot the conversation: speaker diarisation in the wild. In *Interspeech 2020*.

Chung, J. S., Lee, B.-J., and Han, I. (2019) Who said that?: Audio-visual speaker diarisation of real-world meetings. In *Interspeech 2019*.

Coria, J. M., Bredin, H., Ghannay, S., and Rosset, S. (2021) Overlap-Aware Low-Latency Online Speaker Diarization Based on End-to-End Local Segmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Cornell, S., Omologo, M., Squartini, S., and Vincent, E. (2020) Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. In *Interspeech 2020*.

Cyrta, P., Trzciński, T., and Stokowiec, W. (2018) Speaker Diarization Using Deep Recurrent Convolutional Neural Networks for Speaker Embeddings. In Borzemski, L., Świątek, J., and Wilimowska, Z. (Eds.), *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology – ISAT 2017* (Vol. 655). Cham.

Davis, S., and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28(4): 357–366. Presented at the IEEE Transactions on Acoustics, Speech, and Signal Processing.

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., and Na, H. (2021) ECAPA-TDNN Embeddings for Speaker Diarization. In *Interspeech 2021*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011) Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19(4): 788–798.

Delacourt, P. (2000) *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio*. EURECOM.

Desnous, F., Larcher, A., and Meignier, S. (2018) Impact de la détection de la parole pour différentes tâches de traitement automatique de la parole. In *XXXIIe Journées d'Études sur la Parole*. Presented at the XXXIIe Journées d'Études sur la Parole ISCA.

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023, May 23) QLoRA: Efficient Finetuning of Quantized LLMs. arXiv.

Diez, M., Burget, L., Landini, F., Wang, S., and Cernocky, H. (2020) Optimizing Bayesian Hmm Based X-Vector Clustering for the Second Dihard Speech Diarization Challenge. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Diez, M., Burget, L., and Matejka, P. (2018) Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. In *The Speaker and Language Recognition Workshop (Odyssey 2018)*.

Diez, M., Burget, L., Wang, S., Rohdin, J., and Černocký, J. (2019) Bayesian HMM Based x-Vector Clustering for Speaker Diarization. In *Interspeech 2019*.

Ding, S., Wang, Q., Chang, S., Wan, L., and Moreno, I. L. (2020) Personal VAD: Speaker-Conditioned Voice Activity Detection. In *Odyssey 2020 The Speaker and Language Recognition Workshop*.

Ding, Y., Xu, Y., Zhang, S.-X., Cong, Y., and Wang, L. (2020) Self-supervised learning for audio-visual speaker diarization. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Dinkel, H., Chen, Y., Wu, M., and Yu, K. (2020) Voice activity detection in the wild via weakly supervised sound event detection. In *Interspeech 2020*.

Dissen, Y., Kreuk, F., and Keshet, J. (2022) Self-supervised Speaker Diarization. In *Interspeech 2022*.

Doukhan, D., Poels, G., Rezgui, Z., and Carrive, J. (2018) Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. *Journal of European Television History and Culture* 7(14): 103.

El Khoury, E., Laurent, A., Meignier, S., and Petitrenaud, S. (2012) Combining transcription-based and acoustic-based speaker identifications for broadcast news. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Presented at the ICASSP 2012 - 2012 IEEE International Conference on Acoustics, Speech and Signal Processing Kyoto, Japan: IEEE.

Flemotomos, N., Georgiou, P., and Narayanan, S. (2020) Linguistically Aided Speaker Diarization Using Speaker Role Information. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011) A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics* 5(2A).

Freeman, D. K., Cosier, G., Southcott, C. B., and Boyd, I. (1989) The voice activity detector for the Pan-European digital cellular mobile telephone service. In *International Conference on Acoustics, Speech, and Signal Processing*.

Friedland, G., Vinyals, O., Huang, Y., and Muller, C. (2009) Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* 17(5): 985–993. Presented at the IEEE Transactions on Audio, Speech, and Language Processing.

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., and Watanabe, S. (2019) End-to-End Neural Speaker Diarization with Permutation-Free Objectives. In *Interspeech 2019*.

Futrell, R., Mahowald, K., and Gibson, E. (2015) Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33): 10336–10341.

Galibert, O. (2013) Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech. In *Interspeech 2013*.

Gallardo, L. F., Mittag, G., Möller, S., and Beerends, J. (2018) Variable Voice Likability Affecting Subjective Speech Quality Assessments. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, G. (2005) The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Interspeech 2005*. Presented at the Interspeech 2005 ISCA.

Galliano, S., Gravier, G., and Chaubard, L. (2009) The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts. In *Interspeech 2009*.

Garau, G., Dielmann, A., and Bourlard, H. (2010) Audio-visual synchronisation for speaker diarisation. In *Interspeech 2010*.

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., and McCree, A. (2017) Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gay, P., Dupuy, G., Lailler, C., Odobez, J.-M., Meignier, S., and Deleglise, P. (2014) Comparison of two methods for unsupervised person identification in TV shows. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. Presented at the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI) Klagenfurt, Austria: IEEE.

Gebru, I. D., Ba, S., Evangelidis, G., and Horaud, R. (2015) Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*.

Geiger, J. T., Vipperla, R., Bozonnet, S., Evans, N., Schuller, B., and Rigoll, G. (2012) Convolutive Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization. In *Interspeech 2012*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017) Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gish, H., Siu, M.-H., and Rohlicek, R. (1991) Segregation of speakers for speech recognition and speaker identification. In *ICASSP 91: International Conference on Acoustics, Speech, and Signal Processing*. Toronto, Ont., Canada: IEEE.

Gravier, G, Bonastre, J.-F., Geoffrois, E., Galliano, S., Tait, K. M., and Choukri, K. (2004) The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC*.

Gravier, Guillaume, Adda, G., Paulsson, N., Carre, M., Giraudel, A., and Galibert, O. (2012) The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC - Eighth International Conference on Language Resources and Evaluation*.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., and Ferrer, C. C. (2022) Towards Measuring Fairness in AI: the Casual Conversations Dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4(3): 324–332.

Hochreiter, S., and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation* 9(8): 1735–1780.

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., and Nagamatsu, K. (2020) End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors. In *Interspeech 2020*.

Horiguchi, S., Takashima, Y., Watanabe, S., and Garcia, P. (2022) Mutual Learning of Single- and Multi-Channel End-to-End Neural Diarization. In *IEEE SLT 2022*.

Horiguchi, S., Watanabe, S., García, P., Takashima, Y., and Kawaguchi, Y. (2023) Online Neural Diarization of Unlimited Numbers of Speakers Using Global and Local Attractors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31: 706–720.

Horiguchi, S., Watanabe, S., Garcia, P., Xue, Y., Takashima, Y., and Kawaguchi, Y. (2021) Towards Neural Diarization for Unlimited Numbers of Speakers Using Global and Local Attractors. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2021*.

Huh, J., Brown, A., Jung, J., Chung, J. S., Nagrani, A., Garcia-Romero, D., and Zisserman, A. (2023, March 6) VoxSRC 2022: The Fourth VoxCeleb Speaker Recognition Challenge. https://arxiv.org/abs/2302.10248.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011) Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* 19(4): 847–860. Presented at the IEEE Transactions on Audio, Speech, and Language Processing.

Imseng, D., and Friedland, G. (2010) Tuning-Robust Initialization Methods for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* 18(8): 2028–2037. Presented at the IEEE Transactions on Audio, Speech, and Language Processing.

Izquierdo Del Alamo, S., Labrador, B., Lozano-Diez, A., and Toledano, D. T. (2022) Efficient Transformers for End-to-End Neural Speaker Diarization. In *IberSPEECH 2022*.

Jin, M., Ju, C., Chen, Z., Liu, Y. C., Droppo, J., and Stolcke, A. (2022) Adversarial Reweighting for Speaker Verification Fairness. In *Interspeech 2022*.

Joglekar, A., and Hansen, J. H. L. (2019) Fearless Steps Challenge Phase-1 Evaluation Plan. In *Interspeech 2019 Special Session*.

Jousse, V., Meignier, S., Jacquin, C., Petitrenaud, S., Estève, Y., and Daille, B. (2009) Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteurs. *Traitement Automatique des Langues* 50(1): 201–225.

Jousse, V., Petit-Renaud, S., Meignier, S., Esteve, Y., and Jacquin, C. (2009) Automatic named identification of speakers using diarization and ASR systems. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Presented at the ICASSP 2009 - 2009 IEEE International Conference on Acoustics, Speech and Signal Processing Taipei, Taiwan: IEEE.

Kang, W., Roy, B. C., and Chow, W. (2020) Multimodal Speaker Diarization of Real-World Meetings Using D-Vectors With Spatial Features. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007) Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing* 15(4): 1435–1447.

Kuhn, H. W. (1955) The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2(1–2): 83–97.

Kumar, A. K., Waldekar, S., Sahidullah, M., and Saha, G. (2022) Robust Acoustic Domain Identification with its Application to Speaker Diarization. *International Journal of Speech Technology* 25(4): 933–945.

Lai, Y., Tang, X., Fu, Y., and Fang, R. (2021, December 14) End-to-end speaker diarization with transformer. https://arxiv.org/abs/2112.07463.

Landini, F., Profant, J., Diez, M., and Burget, L. (2022) Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language* 71.

Le, D. H. N. (2019) *Multimodal person recognition in audio-visual streams*. Ecole Polytechnique Fédérale de Lausanne.

Le, N., Bredin, H., Sargent, G., India, M., Lopez-Otero, P., Barras, C., Guinaudeau, C., Gravier, G., Da Fonseca, G. B., Freire, I. L., Patrocínio, Z., Guimarães, S. J. F., Martí, G., Morros, J. R., Hernando, J., Docio-Fernandez, L., Garcia-Mateo, C., Meignier, S., and Odobez, J.-M. (2017) Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. Presented at the CBMI '17: International Workshop on Content-Based Multimedia Indexing Florence Italy: ACM.

Lebourdais, M., Tahon, M., Laurent, A., Meignier, S., and Larcher, A. (2022) Overlaps and Gender Analysis in the Context of Broadcast Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference - LREC 2022*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep learning. *Nature* 521(7553): 436–444.

Li, Y. (2020) *Speaker Diarization System for Call-center data*. Master Thesis, KTH Royal Institute of Technology.

Mathieu Ben, Michaël Bester, Fréderic Bimbot, and Guillaume Gravier (2004) Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *Interspeech 2004*. Presented at the Interspeech 2004 ISCA.

Mauclair, J., Meignier, S., and Esteve, Y. (2006) Speaker Diarization: About whom the Speaker is Talking ? In *2006 IEEE Odyssey - The Speaker and Language Recognition*

Workshop. Presented at the 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop San Juan, PR: IEEE.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemo, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005) The AMI Meeting Corpus. https://www.researchgate.net/publication/228341280_The_AMI_meeting_corpus.

McKnight, S. W., Hogg, A. O. T., Neo, V. W., and Naylor, P. A. (2022) Studying Human-Based Speaker Diarization and Comparing to State-of-the-Art Systems. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., Laptev, A., and Romanenko, A. (2020) Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario. In *Interspeech 2020*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021) A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54(6): 1–35.

Meignier, S., Bonastre, J.-F., Fredouille, C., and Merlin, T. (2000) Evolutive HMM for multi-speaker tracking system. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Vol. 2).

Meignier, S., Bonastre, J.-F., and Igounet, S. (2001) E-HMM approach for learning and adapting sound models for speaker indexing. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2001)*.

Meignier, S., and Merlin, T. (2010) LIUM SpkDiarization: An Open Source Toolkit For Diarization. In *CMU SPUD Workshop*.

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., and Besacier, L. (2006) Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language* 20(2–3): 303–330.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations (ICLR)*.

Mingote, V., Viñals, I., Gimeno, P., Miguel, A., Ortega, A., and Lleida, E. (2022) Multimodal Diarization Systems by Training Enrollment Models as Identity Representations. *Applied Sciences* 12(3): 1141.

Mishra, J., Patil, J. N., Chowdhury, A., and Prasanna, M. (2023) End to End Spoken Language Diarization with Wav2vec Embeddings. In *INTERSPEECH 2023*. Presented at the INTERSPEECH 2023 ISCA.

Misra, H., Bourlard, H., and Tyagi, V. (2003) New entropy based combination rules in HMM/ANN multi-stream ASR. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. (Vol. 2). Presented at the International Conference on Acoustics, Speech and Signal Processing (ICASSP'03) Hong Kong, China: IEEE.

Misra, Hemant, and Bourlard, H. (2005) Spectral entropy feature in full-combination multi-stream for robust ASR. In *Interspeech 2005*. Presented at the Interspeech 2005 ISCA.

Mtibaa, A., Petrovska-Delacrétaz, D., Boudy, J., and Ben Hamida, A. (2021) Privacy-preserving speaker verification system based on binary I-vectors. *IET Biometrics* 10(3): 233–245.

Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020) Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017) VoxCeleb: a large-scale speaker identification dataset. In *Interspeech 2017*.

Ng, A., Jordan, M., and Weiss, Y. (2001) On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (Vol. 14).

NIST (2003) The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan.

OpenAI (2023, March 27) GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Otterson, S., and Ostendorf, M. (2007) Efficient use of overlap information in speaker diarization. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015) Librispeech: An ASR corpus based on public domain audio books. In *ICASSP 2015 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Park, D., Yu, Y., Park, K. W., Kim, J. W., and Kim, H. K. (2022) GIST-AiTeR System for the Diarization Task of the 2022 VoxCeleb Speaker Recognition Challenge. In *VoxCeleb Speaker Recognition Challenge 2022*.

Park, T. J., and Georgiou, P. (2018) Multimodal Speaker Segmentation and Diarization using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks. In *Interspeech 2018*.

Park, T. J., Han, K. J., Huang, J., He, X., Zhou, B., Georgiou, P., and Narayanan, S. (2019) Speaker Diarization with Lexical Information. In *Interspeech 2019*.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022) A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language* 72(C).

Petitrenaud, S., Jousse, V., Meignier, S., and Estève, Y. (2010) Reconnaissance Automatique de Locuteurs à l'aide de Fonctions de Croyance. In *17e Congrès Francophone Reconnaissance Des Formes et Intelligence Artificielle (RFIA'10)*. Caen, France.

Poignant, J., Besacier, L., and Quenot, G. (2014) Unsupervised Speaker Identification in TV Broadcast Based on Written Names. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* : 1–1.

Poignant, J., Bredin, H., and Barras, C. (2015) Multimodal Person Discovery in Broadcast TV at MediaEval 2015. In *MediaEval 2015*.

Poignant, J., Bredin, H., and Barras, C. (2017) Multimodal person discovery in broadcast TV: lessons learned from MediaEval 2015. *Multimedia Tools and Applications* 76(21): 22547–22567.

Poignant, J., Bredin, H., Le, V.-B., Besacier, L., Barras, C., and Quénot, G. (2012) Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast. In *Interspeech 2012*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21: 1–67.

Raj, D., Paola Garcia-Perera, L., Huang, Z., Watanabe, S., Povey, D., Stolcke, A., and Khudanpur, S. (2021) DOVER-Lap: A Method for Combining Overlap-Aware Diarization Outputs. In *2021 IEEE Spoken Language Technology Workshop (SLT)*.

Rajan, S. S., Udeshi, S., and Chattopadhyay, S. (2022) AequeVox: Automated Fairness Testing of Speech Recognition Systems. In *Fundamental Approaches to Software Engineering*. Springer.

Raji, I. D., and Buolamwini, J. (2019) Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Ravanelli, M., and Bengio, Y. (2018) Speaker Recognition from Raw Waveform with SincNet. In *SLT 2018*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., et al. (2021, June 8) SpeechBrain: A General-Purpose Speech Toolkit. https://arxiv.org/abs/2106.04624.

Riad, R., Titeux, H., Lemoine, L., Montillot, J., Sliwinski, A., Bagnou, J., Cao, X., Bachoud-Levi, A.-C., and Dupoux, E. (2022) A comparison study on patient-psychologist voice diarization. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*.

Robin, M., Istrate, D., and Boudy, J. (2017) Remote monitoring, distress detection by slightest invasive systems: Sound recognition based on hierarchical i-vectors. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.

Rougui, J. E., Rziza, M., Aboutajdine, D., Gelgon, M., and Martinez, J. (2006) Fast Incremental Clustering of Gaussian Mixture Speaker Models for Scaling up Retrieval In On-Line Broadcast. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013) An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech 2013*.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019) The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In *Interspeech 2019*.

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. (2021) The Third DIHARD Diarization Challenge. In *Interspeech 2021*.

Sell, G., and Garcia-Romero, D. (2014) Speaker diarization with plda i-vector scoring and unsupervised calibration. In *2014 IEEE Spoken Language Technology Workshop (SLT)*.

Shafey, L. E., Soltau, H., and Shafran, I. (2019) Joint Speech Recognition and Speaker Diarization via Sequence Transduction. In *Interspeech 2019*.

Shen, H., Yang, Y., Sun, G., Langman, R., Han, E., Droppo, J., and Stolcke, A. (2022) Improving fairness in speaker verification via Group-adapted Fusion Network. In *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Snyder, D., Chen, G., and Povey, D. (2015, October 28) MUSAN: A Music, Speech, and Noise Corpus. https://arxiv.org/abs/1510.08484.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018) X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Stafylakis, T., Anguera, X., Katsouros, V., and Carayannis, G. (2011) Closed-form expressions vs. BIC: A comparison for speaker clustering. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Presented at the ICASSP 2011 - 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Prague, Czech Republic: IEEE.

Sun, G., Zhang, C., and Woodland, P. (2021) Combination of Deep Speaker Embeddings for Diarisation. *Neural Networks* 141: 372–384.

Tao, R., Pan, Z., Das, R. K., Qian, X., Shou, M. Z., and Li, H. (2021) Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F. (2022) Towards Measuring and Scoring Speaker Diarization Fairness. https://arxiv.org/abs/2302.09991.

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F. (2023) Détection d'activité vocale Multi-flux pour la Diarisation du locuteur. In *GRETSI 2023*.

Tevissen, Y., Boudy, J., and Petitpont, F. (2022) The Newsbridge - Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description. In *VoxCeleb Speaker Recognition Challenge 2022*.

Tevissen, Y., Istrate, D., Zalc, V., Boudy, J., Chollet, G., Petitpont, F., and Boutamine, S. (2023) Home monitoring for frailty detection through sound and speaker diarization analysis. In *JETSAN 2023*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023, February 27) LLaMA: Open and Efficient Foundation Language Models. arXiv.

Tranter, S. E. (2006) Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings* (Vol. 1). Presented at the 2006 IEEE International Conference on Acoustics Speed and Signal Processing Toulouse, France: IEEE.

Tranter, S. E., and Reynolds, D. A. (2006) An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing* 14(5): 1557–1565.

Tsai, W.-H., Cheng, S.-S., and Wang, H.-M. (2004) Speaker clustering of speech utterances using a voice characteristic reference space. In *Interspeech 2004*.

Vallet, F., Essid, S., and Carrive, J. (2013) A Multimodal Approach to Speaker Diarization on TV Talk-Shows. *IEEE Transactions on Multimedia* 15(3): 509–520.

Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014) Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP 2014 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017) Attention Is All You Need. In *International Conference on Neural Information Processing Systems*.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., and Moreno, I. L. (2018) Speaker Diarization with LSTM. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Wang, W., Cai, D., Lin, Q., Yang, L., Wang, J., Wang, J., and Li, M. (2021) The DKU-DukeECE-Lenovo System for the Diarization Task of the 2021 VoxCeleb Speaker Recognition Challenge. In *VoxCeleb Speaker Recognition Challenge 2021*.

Wang, W., Qin, X., Cheng, M., Zhang, Y., Wang, K., and Li, M. (2022a) The DKU-DukeECE Diarization System for the VoxCeleb Speaker Recognition Challenge 2022. In *VoxCeleb Speaker Recognition Challenge 2022*.

Wang, W., Qin, X., Cheng, M., Zhang, Y., Wang, K., and Li, M. (2022b) The DKU-SMIIP Diarization System for the VoxCeleb Speaker Recognition Challenge 2022. In *VoxCeleb Speaker Recognition Challenge 2022*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022) Emergent Abilities of Large Language Models. *Transactions of Machine Learning Researchs*.

Weiss, B., and Burkhardt, F. (2010) Voice attributes affecting likability perception. In *Interspeech 2010*.

Wuerkaixi, A., Yan, K., Zhang, Y., Duan, Z., and Zhang, C. (2022) DyViSE: Dynamic Vision-Guided Speaker Embedding for Audio-Visual Speaker Diarization. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*.

Xia, W., Lu, H., Wang, Q., Tripathi, A., Huang, Y., Moreno, I. L., and Sak, H. (2022) Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection. In *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Zhang, A., Wang, Q., Zhu, Z., Paisley, J., and Wang, C. (2019) Fully Supervised Speaker Diarization. In *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Zhang, Y., Weninger, F., Liu, B., Schmitt, M., Eyben, F., and Schuller, B. (2017) A Paralinguistic Approach To Speaker Diarisation: Using Age, Gender, Voice Likability and Personality Traits. In *Proceedings of the 25th ACM International Conference on Multimedia*.

Zhu, X., Barras, C., Meignier, S., and Gauvain, J.-L. (2005) Combining speaker identification and bic for speaker diarization. In *Interspeech 2005*.

Annexes

Liste des communications

Articles

Publiés :

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F., *Détection d'activité vocale Multi-flux pour la Diarisation du locuteur*, GRETSI 2023.

Tevissen, Y., Istrate D., Zalc V., Boudy, J. Chollet G., Petitpont F., and Boutamine S., *Home Monitoring for Fraitly Detection through Sound and Speaker Diarization Analysis*, JETSAN 2023.

Tevissen, Y., Boudy, J., & Petitpont, F. *The Newsbridge-Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description*, VoxSRC 2022.

Autres :

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F., *Textual Speaker Change Point Detection using Large Language Models*, en cours de revue ICNLSP 2023.

Tevissen, Y., Istrate D., Zalc V., Boudy, J. Chollet G., Petitpont F., *Robust Home Monitoring with Online and Embedded Speaker Diarization*, accepté IRBM 2023.

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F., *Towards Measuring and Scoring Speaker Diarization Fairness*, pré-publication arXiv, https://arxiv.org/pdf/2302.09991.pdf, 2023.

Présentations orales

Tevissen, Yannis, *Home Monitoring for Fraitly Detection through Sound and Speaker Diarization Analysis*, JETSAN 2023, présentée le 2 juin 2023, **Prix de la meilleure présentation.**

Tevissen Yannis, *Des médias à la santé : nouveaux usages de la diarisation du locuteur*, Webinaire interne de l'Institut Mines Télécom, présentée le 25 mai 2023.

Tevissen Yannis, *Archivages audiovisuels à grande échelle grâce à une diarisation robuste*, Journée des jeunes chercheurs en parole (JCEP), présentée le 9 mai 2023.

Posters

Tevissen Y., Boudy J., Chollet G., *Multi-Stream Voice Activity Detection for Robust Speaker Diarization*, GDR ISIS 2022 : Information, Signal, Image et ViSion: Traitement du signal pour la voix, 20 octobre 2022, IRCAM Paris.

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F., *Détection d'activité vocale multi-flux pour la diarisation du locuteur*, GRETSI, 29 août 2023, Grenoble.

Tevissen, Y., Boudy, J., Chollet, G., and Petitpont, F., *Zero-Shot Speaker Change Point Detection using Large Language Models*, Journée des doctorants Paris Saclay, 20 juin 2023, Palaiseau.

Résumé des principales méthodes de diarisation

Méthode	Modalité(s)	Représentation vectorielle utilisée	Particularité
UIS-RNN	Audio	d-vecteurs	Regroupement en-ligne via RNN
TS-VAD	Audio	i-vecteurs	Système itératif basés sur lesVAD personelles
Pyannote 2.0	Audio	x-vecteurs modifiés (PyanNet)	Système modulaire (1 NN par module)
VBx	Audio	x-vecteurs	Double regroupement (AHC + HMM bayésien)
Speaker diarization with lexical information	Audio, Sémantique	x-vecteurs	Matrce d'adjacence + regroupement spectral
EEND	Audio	Aucune explicite	Regroupement multi-classe dans une approche bout-en-bout
Turn-to-diarize	Audio, Sémantique	d-vecteurs	Détection textuelle de changements de locuteur pour contraindre le regroupement spectral

Diarisation auto-supervisée	Audio	Audio brut	Matrice de de similarité + AHC
Diarisation avec ASD	Audio, Image	Obtenus via un CNN	Détection audiovisuelle du locuteur actif
LIUM	Audio	GMMs / i-vecteurs	Deux regroupements (AHC)
MSVAD	Audio	x-vecteurs	3 VADs en parallèle
Diart	Audio	x-vecteurs modifiés (PyanNet)	Regroupement *en-ligne*
FlexSTB (EEND-EDA)	Audio	Aucune explicite	Approche bout-en-bout, *en-ligne*, pour un nombre ∞ de locuteurs
Approche paralinguistique	Audio, paralinguistique	Eléments paralinguistiques inférés	Ulisation d'éléments paralinguistiques
DiFormer	Audio	Audio brut	Une tête d'attention par module de diarisation

Table des abréviations

- *ASD : Active Speaker Detection*
- *BIC : Bayesian Information Criterion*
- *CF : Confusion entre locuteurs*
- *CLR : Cross Likelihood Ratio*
- *CNN : Convolutional Neural Network*
- *CRNN : Convolutional and Recurrent Neural Network*
- *DER : Diarization Error Rate*
- *DFR : Diarization Fairness Rate*
- *DNN : Deep Neural Network*
- *EEND : End to end neural diarization*
- *FA : Fausse Alarme*
- *GMM : Gaussian Mixture Model*
- *GPT : Generative Pretrained Transformer*
- *GP-VAD : Global Purpose Voice Activity Detection*
- *JER : Jaccard Error Rate*
- *LLM : Large Language Model*
- *MFCC : Mel Frequency Cepstral Coefficient*
- *MSVAD : Multi-Stream Voice Activity Detection*
- *PEFT : Parameter Efficient Finetuning*
- *PM : Parole Manquée*
- *QLoRA: Quantized Low Rank Adapter*
- *RNN : Recurrent Neural Network*
- *TSVAD : Target Speaker Voice Activity Detection*
- *UIS-RNN : Unbounded Interleaved State Recurrent Neural Network*
- *VAD : Voice Activity Detection*
- *VBx : Variational Bayes x-vector diarization*
- *WSSED : Weakly Supervised Sound Event Detection*

Liste des figures

Liste des tableaux

Liste des projets *open-sources* utilisés

Méthodes de détection d'activité vocale

- https://superkogito.github.io/blog/2020/02/09/naive_vad.html
- https://github.com/RicherMans/GPV
- https://github.com/RicherMans/Datadriven-GPVAD
- https://huggingface.co/speechbrain/vad-crdnn-libriparty

Algorithme de diarisation du locuteur

- https://github.com/BUTSpeechFIT/VBx
- https://github.com/pyannote/pyannote-audio
- https://github.com/juanmc2005/diart
- https://github.com/wq2012/SpectralCluster
- https://github.com/hitachi-speech/EEND
- https://github.com/kaldi-asr/kaldi/tree/master/egs/chime6/s5b_track2
- https://projets-lium.univ-lemans.fr/spkdiarization/

Modèles de langage volumineux

- https://huggingface.co/docs/transformers/main/model_doc/falcon
- https://github.com/rmihaylov/falcontune
- https://github.com/langchain-ai/langchain

Divers

- https://github.com/wq2012/awesome-diarization
- https://github.com/nryant/dscore
- https://github.com/TaoRuijie/TalkNet-ASD

ECOLE DOCTORALE

Titre : Diarisation multimodale : vers des modèles robustes et justes en contexte réel

Mots clés : diarisation, robustesse, justesse, multimodalité

Résumé : La diarisation du locuteur, c'est à dire la tache de déterminer automatiquement « qui parle, quand ? » dans un enregistrement audio ou vidéo, est un des piliers des systèmes modernes d'analyse des conversations.
A la télévision, les contenus diffusés sont divers et couvrent à peu près tous les types de conversations, de la discussion calme entre deux personnes, aux débats passionnés, en passant par les interviews en terrain de guerre.
L'analyse de ces contenus, réalisée par la société Newsbridge, requiert, en vue de leur archivage et de leur indexation, des méthodes de traitement robustes et justes.
Dans ce travail, nous présentons deux nouvelles méthodes permettant d'améliorer la robustesse des systèmes via des approches de fusion.
La première se concentre sur la détection d'activité vocale, prétraitement nécessaire à tout système de diarisation. La seconde est une approche multimodale qui tire notamment parti des dernières avancées en traitement du langage naturel.
Nous voyons également que les récentes avancées des systèmes de diarisation rendent l'utilisation de la diarisation du locuteur réaliste y compris dans des secteurs critiques tels que l'analyse de larges archives audiovisuelles ou le maintien à domicile de personnes âgées.
Enfin ce travail présente une nouvelle méthode d'évaluation de la justesse algorithmique de la diarisation du locuteur en vue de rendre son utilisation plus responsable.

Title: Multimodal diarization: towards robustness and fairness in the wild

Keywords: diarization, robustness, fairness, multimodality

Abstract: Speaker diarization, or the task of automatically determining "who spoke, when?" in an audio or video recording, is one of the pillars of modern conversation analysis systems.
On television, the content broadcasted are very diverse and covers about every type of conversation, from calm discussions between two people to impassioned debates and wartime interviews.
The archiving and indexing of this content, carried out by the Newsbridge company, requires robust and fair processing methods.
In this work, we present two new methods for improving systems robustness via fusion approaches.
The first method focuses on voice activity detection, a necessary pre-processing step for every diarization system. The second is a multimodal approach that takes advantage of the latest advances in natural language processing.
We also show that recent advances in diarization systems make the use of speaker diarization realistic, even in critical sectors such as the analysis of large audiovisual archives or the home care of the elderly.
Finally, this work shows a new method for evaluating the algorithmic fairness of speaker diarization, with the objective to make its use more responsible.