# **CONTEXTE**

Depuis les années 2020, la pandémie de COVID-19 a bouleversé le monde entier. En plus des effets dévastateurs du virus lui-même, la désinformation a également eu un impact négatif important. La propagation de fausses informations a conduit à des comportements dangereux, à une méfiance envers les institutions et à une politisation de la crise. Combattre la désinformation est donc devenu crucial pour protéger la santé publique et atténuer les effets de la pandémie.

Ainsi pour contrer efficacement la désinformation sur la COVID-19 et encourager des comportements responsables face à la pandémie, l'UNESCO en avril 2020 a lancé une initiative cruciale en produisant une série de messages audio destinés à être diffusés librement par les stations de radio à travers le monde. Cette initiative vise à combattre “l'infodémie", la surabondance d'informations, souvent fausses ou trompeuses, qui accompagne le virus, en fournissant des informations vérifiées et fiables. En collaboration avec l'Organisation Mondiale de la Santé (OMS), l'UNESCO a développé ces messages dans les six langues officielles des Nations Unies ainsi que dans diverses langues autochtones et locales, touchant ainsi une audience globale et diversifiée. Les contenus audio, disponibles en anglais, français, espagnol, Kazakhstan, russe et arabe, offrent des conseils précieux sur les mesures préventives, démystifient les fausses informations concernant le virus et mettent en avant l'importance de la solidarité et de la non-discrimination.

À partir de ces messages audios cruciaux, nous avons constitué notre jeu de données initial, servant de fondation à notre projet de développement d'un modèle de transcription audio multilingue. Ce corpus, méticuleusement compilé, comprend 89 fichiers audio soigneusement sélectionnés pour leur pertinence et leur diversité thématique. Ces fichiers sont répartis sur 14 thématiques essentielles, abordant divers aspects de la prévention et de la réponse à la COVID-19, reflétant ainsi la richesse et la complexité de la communication nécessaire en temps de pandémie.

Les langues choisies pour ce corpus (celles citées plus haut) ont été sélectionnées en raison de leur portée mondiale et de leur pertinence dans les régions particulièrement touchées par la désinformation autour du COVID-19. La durée totale de ce jeu de données s'élève à environ une heure et trente minutes, offrant ainsi une ressource substantielle pour l’augmentation des données, l'entraînement et l'évaluation de notre modèle de transcription.

Cependant, notre projet a rencontré des défis spécifiques, notamment liés à notre capacité à travailler avec l'ensemble du jeu de données. Malgré l'intention d'incorporer une gamme variée de langues pour refléter au mieux la diversité globale, nous avons dû faire face à des limitations significatives en ne nous servant pas des enregistrements en russe, espagnol et Kazakhstan. Cette décision découle de notre manque de compétences linguistiques dans ces langues spécifiques, ce qui a limité notre capacité à transcrire et à vérifier avec précision le contenu de ces enregistrements. Cette lacune souligne l'importance de la collaboration et de la nécessité de développer des outils de transcription et de traduction plus sophistiqués, capables de surmonter les barrières linguistiques et d'assurer une couverture plus inclusive et précise des informations.

Dans la continuité de notre engagement à lutter contre la désinformation liée à la COVID-19 par le biais de notre projet de développement d'un modèle de transcription audio multilingue, la pertinence de notre travail avec ces données demeure capitale. En effet, en intégrant ces données multilingues dans notre modèle, nous aspirons à développer un outil capable de transcrire avec précision des contenus audio variés, contribuant ainsi efficacement à la lutte contre la désinformation. De plus, notre initiative contribue significativement à renforcer la préparation de la communauté mondiale face à d'éventuelles crises futures similaires à la pandémie de COVID-19. En exploitant les leçons apprises grâce à ce projet, nous pouvons mieux comprendre comment la désinformation se propage et ainsi développer des outils plus robustes pour la contrer, garantissant ainsi une réponse plus résiliente et informée dans les moments critiques.

Alors que nous soulignons l'importance capitale du développement de capacités de reconnaissance vocale multilingues pour une communication efficace et inclusive à l'échelle mondiale, en particulier dans le contexte de la réponse aux crises sanitaires mondiales comme la pandémie de COVID-19, il devient pertinent de réfléchir à la manière dont notre jeu de données augmenté s'aligne ou diverge des efforts de recherche précédents dans ce domaine. Cette réflexion ouvre la voie à une exploration approfondie de notre contribution unique au corpus existant.

Notre projet, en intégrant des enregistrements dans une variété de langues, y compris celles moins représentées dans la recherche et les applications précédentes, représente un pas en avant vers une couverture plus globale et équitable.

De plus, en abordant directement les lacunes identifiées dans notre capacité à travailler avec certaines langues due à des contraintes de compétences linguistiques, nous soulignons l'importance de la collaboration multidisciplinaire et internationale pour surmonter ces obstacles.

Penchons-nous sur le cas d’un existant **WhisperX-(OpenAi)** qui est un système avancé pour la transcription précise de l'audio de longue durée, utilisant la détection d'activité vocale et l'alignement phonétique forcé pour générer des timestamps précis au niveau des mots. Il améliore la qualité et la vitesse de transcription en traitant par lots des segments audio pré-segmentés, et est conçu pour surmonter les défis des méthodes précédentes, tels que la dérive, les hallucinations et les répétitions dans les transcriptions de longs audios.

Notre projet se distingue de WhisperX principalement par son application spécifique à l'analyse des messages de l'UNESCO sur la COVID-19 et sa concentration sur l'identification de tendances et mots clés à travers une approche enrichie d'augmentation de données. Contrairement à WhisperX, qui optimise la transcription précise de l'audio de longue durée pour une gamme variée de contenus, notre projet vise une compréhension plus profonde du contenu, utilisant des réseaux neuronaux avancés pour extraire des insights spécifiques des enregistrements.   
La valeur ajoutée réside dans l'application ciblée et la capacité à générer des transcriptions enrichies d'analyses contextuelles, offrant une perspective unique sur les communications liées à la santé publique.

Il invite à une discussion plus large sur les directions potentielles de la recherche et du développement technologique, en mettant en lumière la nécessité d'approches holistiques qui prennent en compte à la fois les avancées technologiques et les implications sociétales de leur déploiement.

# **DESCRIPTION DES DONNEES**

**INVENTAIRE DU JEU DE DONNEES**

Notre inventaire détaillé du jeu de données se présente comme suit :

**Langues et Dialectes** :

* *Anglais,*
* *Français,*
* *Arabe*.   
  Les enregistrements en espagnol, kazakh et russe ont été exclus cause du manque de compétences linguistiques dans ces langues, réduisant le corpus à ces trois langues principalement utilisées.

**Thématiques (14)** :

*1. Éducation à la Médiation Informationnelle et Linguistique (MIL) pour enfants*

*2. Identification de sources d'information fiables*

*3. Écoles et COVID-19*

*4. Effets secondaires du vaccin contre le COVID-19*

*5. Marchés alimentaires traditionnels*

*6. Utilisation des masques*

*7. Prévention et protection contre le COVID-19*

*8. Gestion des réactions émotionnelles*

*9. Actions post-vaccination*

*10. Rassemblements publics*

*11. Recherche de contacts*

*12. Vaccination contre le COVID-19*

*13. Méfiance envers les faux experts*

*14. Transmission du virus*

**Format et Durée** :

*Les 42 enregistrements,* d'une durée totale d'environ *35 minutes,* sont répartis équitablement entre les trois langues et sont disponibles en formats *WAV et MP3,* avec des durées variantes entre *29 et 52 secondes.*

**Interprètes :**

Les audios sont interprétés par des *voix masculines et féminines,* enrichissant la diversité du corpus.

**DOCUMENTATION ET METADONNEES**

Dans le cadre de notre projet de recherche, nous avons entrepris un processus d'organisation et d'annotation méticuleux de notre jeu de données audio. Cette démarche visait à faciliter l'accès et l'analyse des fichiers audio pour leur exploitation dans des travaux de recherche ultérieurs. Pour ce faire, une réorganisation préliminaire des fichiers a été jugée nécessaire. Elle s'est articulée autour de deux étapes principales : le regroupement thématique et le renommage des fichiers. Les fichiers audios ont été soigneusement examinés et classés en fonction de leur contenu thématique, permettant ainsi de regrouper tous les enregistrements traitant d'un même sujet dans un dossier dédié. Chaque enregistrement a ensuite été renommé selon le format "***Msg\_sexe\_langue",*** où "Msg" représente un identifiant unique de l'enregistrement, "sexe" le sexe du locuteur, et "langue" la langue de l'enregistrement. Cette structure vise à offrir une vue d'ensemble claire des caractéristiques principales de chaque fichier, simplifiant les recherches et analyses ultérieures.

La phase suivante de notre travail a porté sur l'annotation des fichiers audio avec des métadonnées pertinentes. Un ensemble de métadonnées clés a été défini pour accompagner chaque fichier audio, incluant le titre de l'enregistrement, la durée, le signal audio, les caractéristiques chromatiques ainsi que d’autres caractéristiques relatives au contenu”. Les métadonnées identifiées ont été soigneusement ajoutées à chaque fichier audio à l'aide d'outils logiciels adaptés, permettant une saisie précise et uniforme des informations pour l'ensemble du jeu de données. Avant l'annotation proprement dite, une série de préanalyses a été effectuée sur le jeu de données, permettant d'identifier des tendances et des caractéristiques clés au sein des enregistrements. Ces analyses préliminaires ont orienté le choix des métadonnées à annoter, enrichissant chaque fichier d'informations supplémentaires facilitant son identification et son analyse.

La réorganisation et l'annotation des fichiers audio constituent ainsi une étape fondamentale dans la préparation de notre jeu de données pour des recherches futures. Elles garantissent une accessibilité et une gestion efficace des données audios, posant les bases d'une exploitation réussie dans le cadre de nos travaux de recherche.

## Dans les lignes qui suivront, nous nous pencherons sur les métadonnées, ces données sur les données, ces les clés qui déverrouillent le potentiel de nos informations. En les analysant, nous exploitions une mine d'informations pour améliorer la gestion, la sécurité et la prise de décision basée sur les données.

1. Le signal audio

Une image contenant Tracé, diagramme, capture d’écran, ligne

Description générée automatiquement

L'image représente la forme d'onde d'un signal audio sur un graphique où l'axe horizontal illustre le temps et l'axe vertical l'amplitude, variant entre -1 et 1, indiquant le silence à zéro et la plus grande amplitude des signaux positifs et négatifs. Cette visualisation est cruciale pour analyser le contenu sonore, les variations denses suggérant des séquences sonores intenses ou de haute fréquence et les parties plates pouvant signifier des moments plus calmes. Les métadonnées déduites de cette image, comme le *pic d'amplitude*, *la gamme dynamique*, et *la puissance moyenne*, requièrent des connaissances sur le taux d'échantillonnage pour préciser la durée.

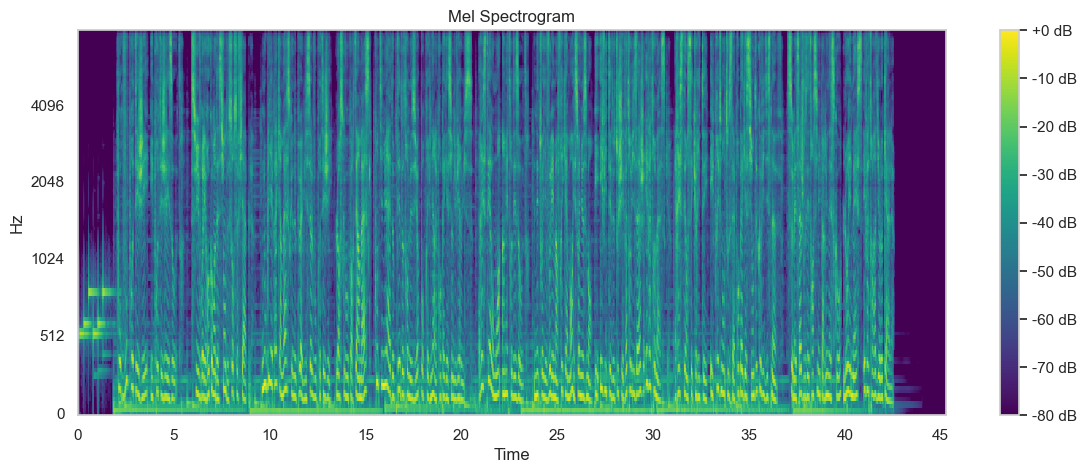
1. Les amplitudes

Une image contenant texte, Tracé, capture d’écran, diagramme

Description générée automatiquement

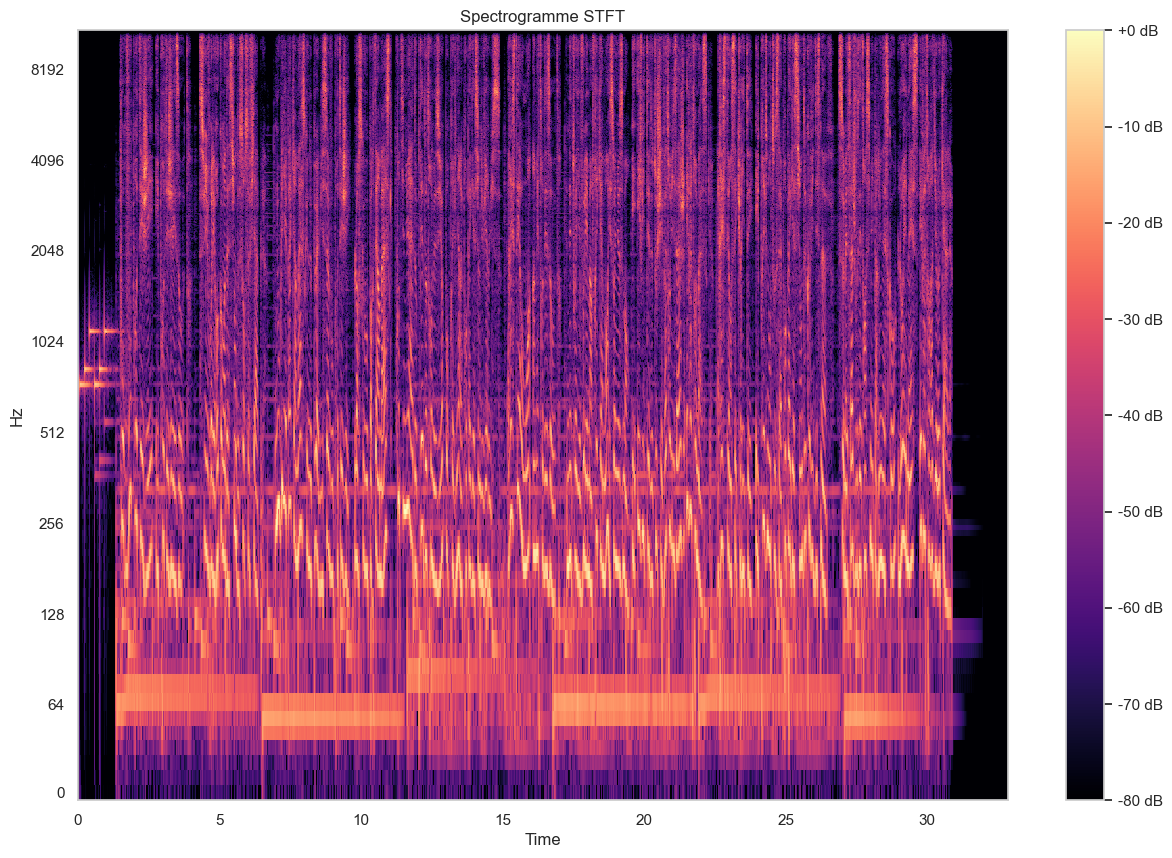
L'image illustre un signal audio sous forme d'onde où les marqueurs rouges identifient les points de plus haute amplitude. Ces points d'intensité sonore maximale peuvent être des indicateurs utiles pour diverses analyses, comme la détection de pics de volume ou la présence de sons transitoires qui sont souvent associés à des événements sonores comme des claps, des coups de cymbale ou des syllabes accentuées dans la parole. La représentation graphique fournit une visualisation directe de la dynamique du signal, ce qui est essentiel pour la reconnaissance de la parole où les variations d'amplitude sont particulièrement significatives.

1. Le spectogramme de Mel



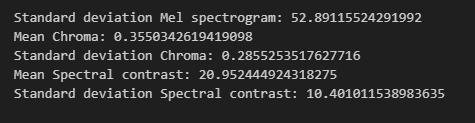
L'image illustre un spectrogramme de Mel, qui est une représentation visuelle de la fréquence des sons dans un fichier audio sur l'échelle de Mel, en fonction du temps. L'échelle de Mel est une échelle de fréquence perceptive qui tient compte de la façon dont les humains perçoivent les sons, en particulier pour les différenciations de fréquences basses. Les couleurs du spectrogramme représentent l'intensité (ou la puissance) des fréquences à différents moments : les jaunes et les verts indiquent des fréquences plus fortes, tandis que les bleus et les violets représentent des fréquences plus faibles. L'axe horizontal montre le temps en secondes, et l'axe vertical montre les fréquences, du plus bas au plus haut. Cet outil est particulièrement utile dans les domaines de la reconnaissance vocale et de la musique, car il permet de visualiser et d'analyser le contenu fréquentiel d'un signal audio de manière plus intuitive.

1. Le spectogramme STFT (Short-Time Fourier Transform ou la Transformée de Fourier à court terme)



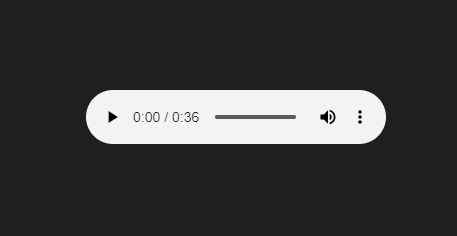
L'image illustre un spectrogramme réalisé à partir d'une transformation de Fourier à court terme (STFT), qui décompose un signal audio en composantes fréquentielles au fil du temps. L'axe horizontal montre le temps en secondes, l'axe vertical représente les fréquences en hertz, et les couleurs indiquent l'intensité (ou la puissance) des fréquences à différents moments, avec le rouge et le jaune indiquant les intensités les plus élevées et le violet les plus faibles. Cette analyse fréquentielle détaillée est cruciale pour diverses applications, telles que l'identification des caractéristiques du signal audio, la reconnaissance de la parole, la classification de la musique, ou encore l'analyse acoustique environnementale. Elle permet aux ingénieurs et chercheurs de détecter les motifs, les harmoniques et les bruits de fond d'un enregistrement audio.

1. Les statistiques chromatiques (Ton chromatique, contraste spectral



L'image affiche un ensemble de statistiques décrivant les caractéristiques d'un signal audio:

1. *Écart type du spectrogramme de Mel (Standard deviation Mel spectrogram)*:   
   Il indique la variabilité de l'intensité des fréquences dans le spectrogramme de Mel. Un écart type plus élevé signifie que l'intensité des fréquences varie grandement tout au long de l'enregistrement, ce qui pourrait indiquer une complexité accrue dans le contenu sonore ou des variations dynamiques importantes.
2. *Moyenne de Chroma (Mean Chroma):*Le chroma, souvent utilisé en analyse musicale, se réfère à une représentation où le spectre est projeté en 12 régions qui correspondent aux 12 notes de la gamme tempérée occidentale. La moyenne ici suggère la présence prédominante d'une ou de plusieurs notes au sein du signal analysé.
3. *Écart type de Chroma (Standard deviation Chroma):* Cet écart type mesure la variabilité des intensités des notes chromatiques. Un écart type faible pourrait indiquer une régularité ou une homogénéité dans les notes utilisées, tandis qu'un écart type élevé révèlerait une diversité plus grande dans la composition chromatique du signal.
4. *Contraste spectral moyen (Mean Spectral contrast)*:  
   Le contraste spectral reflète la différence en décibels (dB) entre les pics et les vallées dans le spectre sonore. Un contraste élevé signifie généralement une clarté plus grande entre les composants sonores, pouvant indiquer une séparation nette entre les sources sonores dans l'enregistrement.
5. *Écart type du contraste spectral (Standard deviation Spectral contrast):*Cela montre la variabilité du contraste spectral tout au long de l'audio. Un grand écart type dans le contraste spectral indique que le degré de séparation entre les pics et les vallées du spectre varie considérablement dans le temps, ce qui pourrait être dû à des changements dans les types de son ou dans la dynamique de l'enregistrement.
6. La durée



La durée d'un fichier audio, telle que 36 secondes, est précieuse pour les analyses qui nécessitent une connaissance contextuelle du contenu, permettant par exemple d'identifier la nature d'un enregistrement (comme un jingle ou une réponse d'interview), d'évaluer la quantité de parole ou de musique présente, ou encore de structurer des segments pour la synchronisation avec d'autres médias. Elle est aussi cruciale en recherche linguistique pour segmenter la parole et étudier l'intonation, en musicologie pour l'analyse des formes musicales, et en sciences comportementales pour évaluer l'attention des auditeurs. En somme, la durée est un indice fondamental dans la gestion, le traitement et l'interprétation des données audio, ainsi que dans l'optimisation des processus d'analyse pour de larges collections audio ou des études comparatives.

# **CONCEPTION EXPERIMENTALE, MATERIAUX ET METHODES**

Dans le cadre du processus d'acquisition et d'augmentation des données audio, différents logiciels et modules Python ont été utilisés, chacun apportant des fonctionnalités spécifiques essentielles au traitement et à l'analyse du son.

1. Python :

Une image contenant Police, texte, Graphique, logo

Description générée automatiquement

Langage de programmation de haut niveau très prisé pour sa simplicité et sa puissance. Python sert de socle aux différents modules et bibliothèques utilisés pour l'analyse de données.

1. Matplotlib (matplotlib) :



Il s'agit d'une bibliothèque de visualisation de données en Python, largement utilisée pour créer des graphiques et des figures de qualité publication.

1. NumPy (numpy) :

Une image contenant cube, carré, Casse-tête mécanique, Rubik's Cube

Description générée automatiquement

Ce module est fondamental pour le calcul scientifique en Python. Il offre des fonctionnalités de manipulation de tableaux et matrices efficaces et est souvent utilisé pour les opérations numériques complexes.

1. Pandas (pandas) :

Une image contenant capture d’écran, Graphique, graphisme, Police

Description générée automatiquement

Bibliothèque de manipulation et d'analyse de données offrant des structures de données et des opérations pour manipuler des tableaux numériques et des séries temporelles.

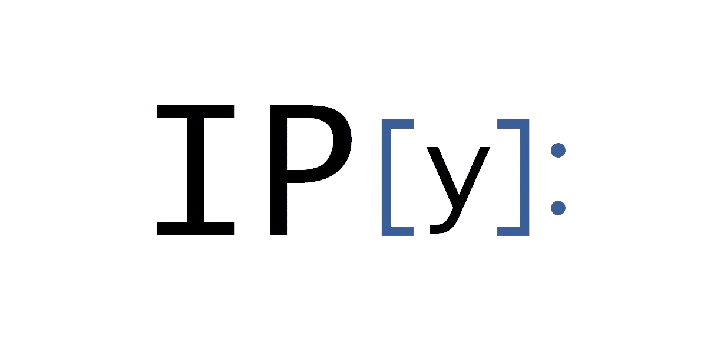
1. Librosa (librosa):

Une image contenant Graphique, graphisme, art, conception

Description générée automatiquement

C'est une bibliothèque spécialisée dans l'analyse audio et musique en Python. Elle permet d'effectuer des opérations complexes comme la transformation de Fourier, la création de spectrogrammes et la détection des battements.

1. IPython.display (IPython.display) :



Fournit des outils pour l'affichage interactif de contenus tels que des images, du son, et des vidéos dans les notebooks IPython.

Chacun de ces outils et modules contribue à une étape spécifique du processus d'acquisition d'analyse et d’augmentation des données audio, allant de la manipulation de base des fichiers à des opérations plus complexes de traitement du signal audio et d’augmentation des données.

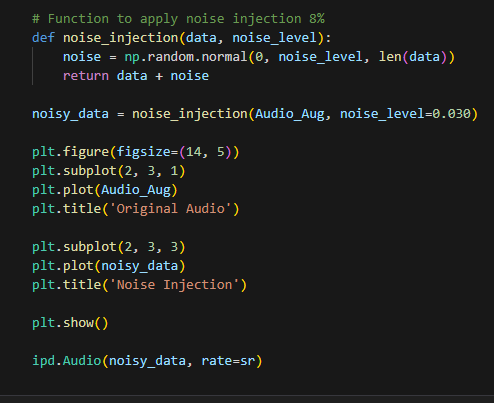
**L’AUGMENTATION DES DONNEES**

Dans le cadre de l'enrichissement de notre base de données initiale, constituée de 42 fichiers audio représentant un total de 30 minutes, nous avons procédé à une série de manipulations visant à augmenter substantiellement le volume de nos données. Cette démarche nous a permis de constituer un corpus enrichi comprenant 505 fichiers audio, pour une durée globale de 6 heures et 20 minutes. Cette expansion significative a été rendue possible grâce à l'application de diverses techniques d'augmentation de données sur chacun des fichiers du corpus original. La présente section détaillera les méthodes et les scripts spécifiquement mis en œuvre pour réaliser cette augmentation de notre jeu de données, soulignant ainsi les étapes clés de notre démarche d'enrichissement.

1. Injection de bruits

L'injection de bruit dans le traitement des données audio est une technique d'augmentation visant à améliorer la robustesse des modèles en simulant divers environnements acoustiques. Elle consiste à ajouter artificiellement différents types et niveaux de bruit de fond (tel que le bruit blanc, le bruit de trafic ou les conversations de foule) aux enregistrements audios originaux. Ce processus a pour but d'enrichir le jeu de données avec des conditions sonores variées, sans masquer les informations importantes du signal initial. L'objectif est de préparer le modèle à être efficace dans des situations réelles, où les conditions acoustiques peuvent grandement varier. Entraîner un modèle sur des données ainsi diversifiées lui permet de mieux généraliser et d'extraire l'information pertinente même en présence de perturbations, augmentant ainsi sa performance et sa fiabilité pour des applications telles que la reconnaissance vocale ou la détection d'événements sonores dans des environnements bruyants.

Dans le cadre de notre recherche, nous avons implémenté une stratégie d'augmentation de données par injection de bruit contrôlée réalisée à l'aide d'un script Python personnalisé. Ce script implémente une méthode de type "bruit gaussien", largement plébiscitée dans la communauté scientifique pour sa simplicité et son efficacité. Le code du script est le suivant :



L'image ci-dessous montre une comparaison entre l'audio original et l'injection de bruit. Le bruit rend la forme d'onde plus rugueuse et le spectre de fréquences montre des pics à certaines fréquences. Le type de bruit injecté n'est pas clair, mais gaussien à un niveau élevé.

Une image contenant capture d’écran, diagramme, Tracé, ligne

Description générée automatiquement

1. Suppression de bruits

La suppression de bruit est une technique d'augmentation de données qui vise à améliorer la clarté des enregistrements audio en réduisant le bruit de fond indésirable. Cette méthode consiste à appliquer des algorithmes de traitement du signal pour identifier et éliminer les composantes sonores non pertinentes, tout en préservant autant que possible la qualité du signal original. Le but de la suppression de bruit est de générer un corpus de données audio plus propre, ce qui est particulièrement utile pour entraîner des modèles de reconnaissance vocale, de transcription automatique ou toute autre application sensible au bruit de fond. En s'entraînant sur des données nettoyées, les modèles peuvent apprendre à se concentrer sur les caractéristiques importantes du signal, comme la parole, sans être distraits par des éléments extérieurs. Cela peut conduire à une amélioration significative de la précision et de l'efficacité des modèles dans des environnements contrôlés ou lorsque le bruit de fond peut être prédit et éliminé à la source.

Dans le cadre de notre étude, nous avons procédé à une suppression des bruits parasites en début et en fin d'enregistrements audio, ainsi qu'à l'élimination des silences non significatifs. Cette étape de prétraitement a été réalisée à l'aide d'un script Python personnalisé, dont le code est présenté ci-après :

Une image contenant texte, capture d’écran, affichage, logiciel

Description générée automatiquement

L'image ci-dessous met en exergue une comparaison entre l'audio original et celui dont on a supprimé les bruits indésirables.

En observant le graphique "Original Audio", on note une amplitude qui varie constamment entre -1 et 1, ce qui est caractéristique d'un signal audio qui contient à la fois le son souhaité (comme de la parole ou de la musique) et du bruit de fond. L'étendue complète de l'amplitude sur tout le graphique suggère que le bruit de fond est omniprésent et potentiellement de volume variable.

Le graphique "Cut Audio" montre un signal qui reste également entre -1 et 1 en amplitude, mais on peut remarquer une légère diminution de la densité des oscillations, en particulier vers le début et la fin de la piste. Cela indique une réduction du bruit de fond, suggérant que l'algorithme ou la technique utilisée pour nettoyer l'audio a effectivement diminué certaines des composantes du bruit.

Une image contenant capture d’écran, Tracé, ligne, Police

Description générée automatiquement

1. Décalage temporel

Le décalage temporel est une méthode d'augmentation des données audio qui consiste à modifier légèrement le minutage des signaux audio dans le but d'augmenter la diversité temporelle du corpus de formation. Cette technique implique de décaler les enregistrements audios dans le temps, soit en les avançant, soit en les retardant, créant ainsi de nouvelles versions du même signal qui semblent survenir plus tôt ou plus tard qu'elles ne le faisaient initialement. L'objectif de cette approche est de rendre notre modèle plus flexible et résilient face à des variations temporelles dans les données d'entrée. En s'entraînant sur un jeu de données enrichi par des décalages temporels, les modèles peuvent mieux généraliser et maintenir une performance élevée même lorsque les événements audios ne se produisent pas exactement au même moment dans différents échantillons. Cela est particulièrement utile dans des scénarios réels où les délais dans la signalisation ou la parole peuvent varier d'un enregistrement à l'autre, aidant ainsi à améliorer la précision globale et la robustesse du modèle.

Pour ce qui est de notre travail, nous avons procédé à un décalage audio de 20 secondes à partir du début. Cette étape a été réalisée à l'aide d'un script Python personnalisé, dont le code est présenté ci-après :

Une image contenant texte, capture d’écran, Police, affichage

Description générée automatiquement

signal qui a subi un décalage temporel. Sur ces graphiques, l'amplitude est représentée sur l'axe vertical et l'axe horizontal représente les échantillons de temps.

Le graphique intitulé "Original Audio" montre une onde sonore continue avec des variations d'amplitude qui se propagent sur toute la durée de l'échantillon. Cela indique un enregistrement audio avec les variations naturelles de l'intensité sonore au fil du temps.

Le second graphique, intitulé "Time décalage audio", présente ce qui semble être le même signal audio, mais avec un décalage notable au milieu. Ce décalage crée un espace vide dans l'onde sonore, qui représente un silence ou une absence de son. Après cet espace vide, le signal semble reprendre, mais avec une correspondance légèrement décalée par rapport au signal original.

Une image contenant capture d’écran, Tracé, diagramme, ligne

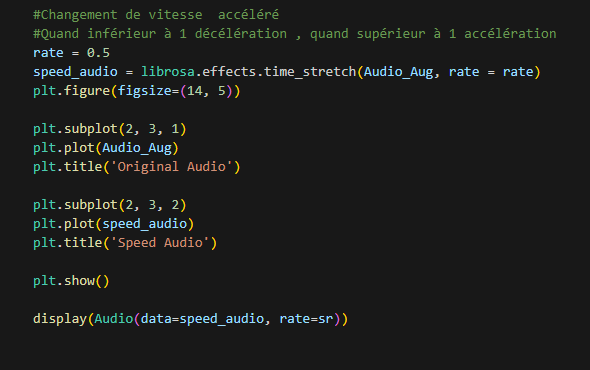
Description générée automatiquement

1. Variation (Augmentation et diminution) de la vitesse

La variation de la vitesse, englobant à la fois l'augmentation et la diminution de la vitesse d'un enregistrement audio, est une méthode d'augmentation de données visant à enrichir le corpus d'entraînement par l'ajout de diversité dans la temporalité et le pitch des signaux sonores. Cette technique consiste à modifier la vitesse de lecture des fichiers audio, accélérant ou ralentissant le signal sans altérer sa tonalité originale de manière significative. L'objectif de cette pratique est de préparer les modèles de machine learning, notamment ceux dédiés à la reconnaissance vocale ou à la détection d'événements sonores, à mieux gérer les variations naturelles de débit de parole ou de mouvement dans les enregistrements audio. En intégrant des versions accélérées et ralenties des mêmes pistes audio dans le jeu de données, les modèles peuvent apprendre à reconnaître et à comprendre le contenu audio indépendamment de la vitesse à laquelle il est présenté. Cela est particulièrement utile pour améliorer la flexibilité et la robustesse des systèmes face à des variations de vitesse de parole dans le langage naturel ou à des changements de tempo dans les environnements sonores, contribuant ainsi à une meilleure performance et à une plus grande précision des modèles dans des conditions variées.

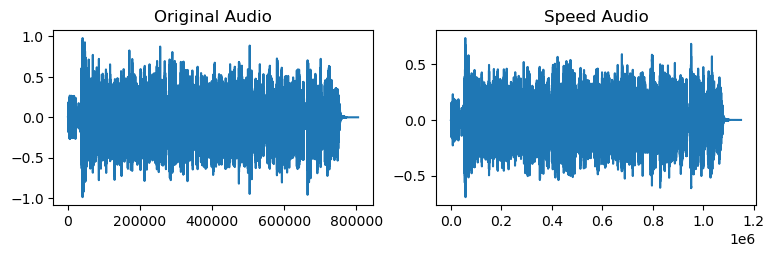
**Dans le cadre de notre étude, nous avons enrichi notre corpus d'enregistrements audio en appliquant à chaque fichier une variation de la vitesse de lecture de 10%, 30% et 50%.** Cette stratégie d'augmentation de données vise à explorer un large éventail de variations temporelles susceptibles d'influencer la performance des modèles d'apprentissage automatique ultérieurement entraînés. L'augmentation de la vitesse de lecture simule une accélération du locuteur, utile pour la reconnaissance vocale dans des environnements bruyants ou l'analyse de discours émotionnels. La diminution de la vitesse de lecture simule un ralentissement du locuteur, utile pour la transcription automatique ou l'analyse de discours complexes. Cette approche s'est avérée efficace pour améliorer la robustesse des modèles et élargir le champ d'application de notre recherche.

Elle a été réalisée grâce au script python suivant :



Les graphiques comparatifs affichés ci-suit démontrent par exemple des effets d'une diminution de la vitesse de lecture audio de 30%. Dans le graphique "Original Audio", on observe un signal compact avec des variations d'amplitude qui se propagent sur l'intégralité de l'axe horizontal, représentatif d'un fichier audio classique. L'amplitude varie de manière dynamique entre -1 et 1, illustrant la pleine intensité des variations sonores capturées dans l'enregistrement original.

Le second graphique, intitulé "Speed Audio", présente le même signal audio après la réduction de vitesse. L'échelle horizontale est étendue, reflétant une durée accrue due au ralentissement. L'amplitude semble être réduite, avec des pics ne dépassant pas 0.5, ce qui peut indiquer une baisse de volume associée ou une conséquence de la méthode de ralentissement appliquée. Cette modification étire non seulement la durée de l'audio, mais peut aussi altérer la hauteur tonale des sons enregistrés, une conséquence naturelle du ralentissement de la lecture. L'analyse technique devrait inclure une évaluation auditive pour détecter les distorsions ou artefacts potentiels introduits par le traitement et pour assurer que la qualité et la clarté du son restent acceptables malgré la réduction de vitesse.



1. Extraction d’un extrait audio

L'extraction d'un audio à partir d'un audio original est une méthode d'augmentation de données qui consiste à sélectionner et isoler des segments spécifiques au sein d'un enregistrement audio plus long pour créer de nouveaux échantillons. Cette technique implique d'identifier des portions d'intérêt telles que des phrases, des mots, ou des sons distincts et de les extraire comme des fichiers audios indépendants. L'objectif est de diversifier le jeu de données en introduisant une variété de contextes et de conditions acoustiques, tout en concentrant l'attention du modèle sur des éléments sonores clés. En s'entraînant sur ces extraits, notre modèle, peut devenir plus efficaces à identifier et à traiter des caractéristiques audios spécifiques. Cette approche permet non seulement d'augmenter quantitativement le jeu de données mais aussi d'enrichir sa qualité en offrant des perspectives variées sur le matériel sonore. En résultat, elle contribue à améliorer la généralisation du modèle et sa capacité à opérer avec précision dans divers scénarios audio, rendant les systèmes plus robustes et adaptables à un large éventail de situations réelles.

Afin d'enrichir notre corpus et d'alimenter notre modèle d'apprentissage automatique, nous avons extrait des segments de 20 secondes de chacun de nos enregistrements audios à partir de ce script Python :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Les graphiques présentés ci après illustrent la comparaison entre un enregistrement audio original et une version extraite ou "coupée" de cet enregistrement. Le graphique "Original Audio" montre une onde sonore avec une amplitude fluctuant fortement entre -1 et 1 sur toute la durée, ce qui indique un signal riche et potentiellement complexe avec de nombreuses variations sonores.

En contraste, le graphique "Cut Audio" présente un signal dont la durée est considérablement réduite, suggérant l'extraction d'une portion spécifique de l'audio original. Cette section semble conserver la même plage d'amplitude que l'audio original, indiquant que le niveau de volume est resté relativement constant. Toutefois, la réduction de la longueur du signal suggère que seul un segment de l'enregistrement original a été conservé, probablement en supprimant les parties non désirées au début et à la fin.

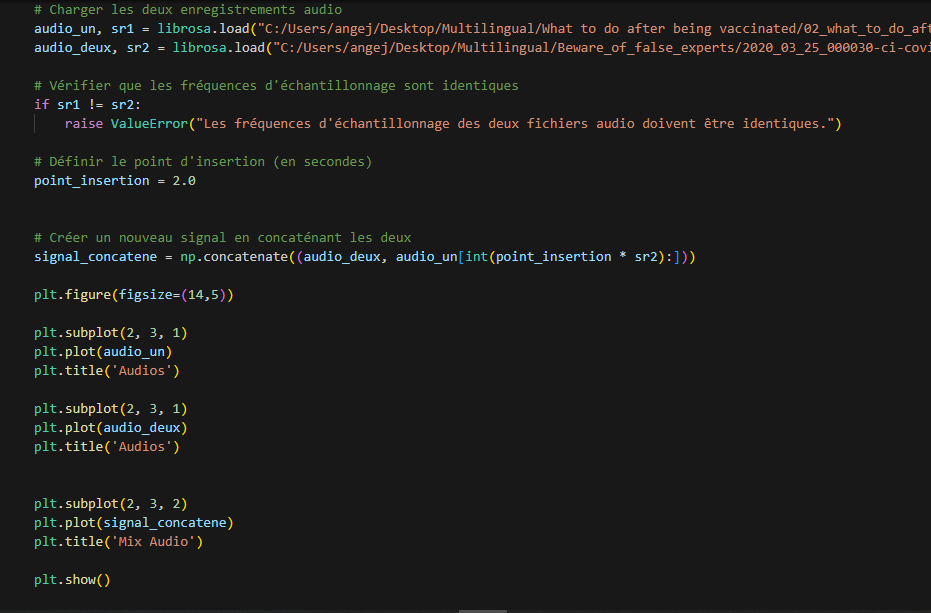
Une image contenant capture d’écran, Tracé, Police, ligne

Description générée automatiquement

1. Mix de deux audios

Le mix ou la concaténation de deux audios est une technique d'augmentation des données qui consiste à fusionner deux fichiers audios pour créer un nouvel enregistrement. Cette méthode implique de combiner harmonieusement deux signaux audio distincts, soit en les superposant pour créer un effet de mixage, soit en les joignant bout à bout pour former un enregistrement plus long. L'objectif de cette pratique est d'enrichir le jeu de données avec des échantillons audio qui présentent une complexité accrue ou une durée étendue, permettant ainsi aux modèles d'apprentissage automatique d'être exposés à des contextes sonores plus variés et dynamiques. En s'entraînant sur des données issues de mixes audio, notre modèle peut mieux apprendre à distinguer et à traiter simultanément plusieurs sources sonores, améliorant ainsi leur capacité à généraliser à partir d'expériences auditives diversifiées. En fin de compte, cette méthode vise à accroître la robustesse et l'efficacité des systèmes en les préparant à naviguer dans un monde sonore riche et en constante évolution.

La mise en œuvre de la concaténation des fichiers audio a été réalisée à l'aide d'un script Python personnalisé, dont le détail sera présenté ultérieurement. Pour orchestrer cette opération, nous avons adopté une approche méthodique basée sur une règle spécifique : chaque thématique de notre corpus initial était représentée par trois fichiers audio, chacun dans une langue différente (français, anglais, arabe). En exploitant le principe mathématique de la permutation (3!), nous avons généré, pour chaque thématique, six fichiers audio mixés. Cette démarche systématique nous a permis d'assurer une diversification maximale au sein de notre corpus enrichi, en veillant à ce que chaque combinaison linguistique possible soit représentée, enrichissant ainsi le jeu de données avec une variété de contextes multilingues et interculturels.



Dans les graphiques présentés dans la suite, nous observons deux visualisations différentes liées à l'audio. Le premier graphique, étiqueté " Mix Audio ", semble représenter deux signaux audio distincts, probablement superposés : l'un est affiché en bleu et l'autre en orange. Chaque signal montre des variations d'amplitude entre -1 et 1, indiquant deux pistes audios qui ont été combinées visuellement mais pas encore mélangées acoustiquement, car on peut distinguer clairement les deux couches superposées.

L’autre graphique, "Audio ", montre le résultat de ce qui semble être une concaténation des originaux de ces deux fichiers audios. La durée du signal est allongée, suggérant l'ajout séquentiel d'un fichier audio à la fin de l'autre. Cela est mis en évidence par la forme du signal qui change au fil du temps, ce qui est typique lors de la fusion de deux sources sonores distinctes. L'ampleur des variations reste constante, avec des pics atteignant les limites supérieure et inférieure, indiquant que le volume global n'a pas été significativement modifié lors de la concaténation.

Une image contenant capture d’écran, Tracé, ligne, diagramme

Description générée automatiquement

# **LIMITATIONS**

L'objectif de cette section est de présenter une analyse critique des limitations rencontrées dans le traitement de données audio pour la transcription automatique. Elle résume les observations pour donner suite à la révision approfondie du jeu de données audio utilisé.

**Analyse critique des données**

Une évaluation minutieuse des fichiers audio a révélé plusieurs limitations significatives affectant la précision de la transcription :

*Ajout de bruit :*

Lors de la phase d'ajout de bruit, il a été constaté que les fichiers audios avec un niveau de bruit élevé ne sont pas correctement transcrits. Le bruit de fond perturbe l'algorithme de reconnaissance vocale, menant à des erreurs de transcription ou à l'omission de segments audio importants.

*Fichiers mixés :*

Les fichiers audio résultant d'une concaténation de deux sources distinctes présentent également une baisse de précision dans la transcription. Il semble que la fusion de plusieurs flux audio dégrade certains attributs du signal nécessaires à une identification claire par les systèmes de reconnaissance vocale.

*Format MP3 et vitesse de lecture :*

Il a également été observé que les fichiers au format MP3, ainsi que les fichiers dont la vitesse de lecture est considérablement réduite, posent des problèmes de transcription. Le format de compression MP3 peut introduire des artefacts auditifs qui obscurcissent la parole, tandis que la réduction de la vitesse de lecture peut modifier la fréquence et la clarté des mots prononcés.