International Journal of Education, Management, and Technology

e-ISSN : 3026-5444
p-ISSN : 3026-4200

Index: Harvard, Boston, Sydney University, Copernicus, Scilit, Semantic, Dimensions, etc

https://doi.org/10.58578/IJEMT.v4i2.11742

L'accent cleaning par l'intelligence artificielle : enjeux éthiques, responsabilité épistémique et reproduction des hiérarchies linguistiques dans les technologies vocales

AI-Powered Speech Recognition: Ethical Issues, Epistemic Responsibility, and the Reproduction of Linguistic Hierarchies in Voice Technologies

Nourredine Bensaid & Malika Bahmad
Ibn Tofail University, Morocco
nourreddine.bensaid@uit.ac.ma; malika.bahmad@uit.ac.ma

Article Info:

Submitted:	Revised:	Accepted:	Published:
Jul 17, 2026	Aug 14, 2026	Aug 26, 2026	Aug 31, 2026

Abstract

The widespread deployment of AI based speech technologies is profoundly reshaping contemporary language practices. These systems, primarily trained on corpora from standardized or dominant varieties, tend to normalize phonetic features that deviate from the reference norm a phenomenon known as "accent cleaning." This paper investigates this process through a dual question: to what extent do automatic speech recognition (ASR) systems contribute to the reproduction of linguistic and social hierarchies, and what responsibility do language researchers bear in the face of these biases? Our central hypothesis is that these models prioritize certain phonetic norms at the expense of diversity, producing forms of symbolic exclusion for speakers of marginalized accents. An exploratory comparative analysis is conducted on three ASR systems (Google SpeechtoText, OpenAI Whisper, Amazon Transcribe), based on a speech

corpus from French speaking speakers of three contrasting varieties (Standard French, Moroccan French, Sub-Saharan French). Anticipated results suggest significant asymmetries in word error rates (WER) and active normalization mechanisms. The study aims to empirically document these biases and to examine the researcher's role within a digital ethics framework.

Keywords: Accent; Ethics; Artificial Intelligence; Automatic Speech Recognition; Scientific Responsibility

Résumé: Le déploiement massif des technologies vocales fondées sur l'intelligence artificielle reconfigure en profondeur les pratiques langagières contemporaines. Ces systèmes, entraînés majoritairement sur des corpus issus de variétés standardisées ou géopolitiquement dominantes, tendent à normaliser les traits phonétiques qui s'écartent de la norme de référence, phénomène désigné sous le terme d'accent cleaning. Cette contribution interroge ce processus à travers une double question : dans quelle mesure les systèmes de reconnaissance automatique de la parole (RAP) participent ils à la reproduction de hiérarchies linguistiques et sociales, et quelle responsabilité incombe au chercheur en sciences du langage face à ces biais ? L'hypothèse centrale est que ces modèles privilégient certaines normes phonétiques au détriment de la diversité, produisant des formes d'exclusion symbolique pour les locuteurs d'accents minorés. Une analyse comparative exploratoire est menée sur trois systèmes RAP (Google SpeechtoText, OpenAI Whisper, Amazon Transcribe) à partir d'un corpus vocal de locuteurs francophones de trois variétés contrastées (français standard, français marocain, français subsaharien). Les résultats anticipés suggèrent des asymétries significatives dans les taux de reconnaissance (WER) et des mécanismes de normalisation active. L'étude vise à documenter empiriquement ces biais et à interroger le rôle du chercheur dans une perspective d'éthique numérique.

Mots-clés : accent ; éthique ; intelligence artificielle ; reconnaissance automatique de la parole ; responsabilité scientifique

INTRODUCTION

Contexte et enjeux

L'essor des technologies vocales fondées sur l'intelligence artificielle (assistants vocaux, systèmes de reconnaissance automatique de la parole (RAP), synthèse vocale) transforme les modalités d'interaction entre les humains et les machines. Ces dispositifs, devenus omniprésents dans les environnements numériques grand public (téléphones, enceintes connectées, services de transcription), prétendent offrir une restitution fidèle et universelle de la parole humaine. Pourtant, cette prétention à l'universalité masque une réalité

sociolinguistique plus complexe : les modèles sous-jacents sont entraînés sur des corpus massifs qui reflètent inégalement la diversité des pratiques langagières effectives.

Des travaux récents (Graham & Roll, 2024; Koenecke et al., 2020) ont mis en évidence des disparités systématiques dans les performances des systèmes RAP selon l'origine ethnique ou géographique des locuteurs, avec des taux d'erreur significativement plus élevés pour les variétés non standard. Ainsi, Koenecke et al. (2020) ont montré que cinq systèmes RAP commerciaux majeurs (Amazon, Apple, Google, IBM et Microsoft) présentaient un taux d'erreur moyen (WER) de 0,35 pour les locuteurs noirs américains contre 0,19 pour les locuteurs blancs, soit une disparité de près de 84 %. Ces asymétries, comme le soulignent les auteurs, ne relèvent pas d'un simple défaut technique : elles résultent de choix de conception, de biais dans les données d'entraînement et, plus fondamentalement, d'une conception de la norme phonétique qui reconduit des hiérarchies sociales préexistantes.

Objectifs et questions de recherche

La présente étude poursuit deux objectifs principaux. Le premier, descriptif et empirique, consiste à quantifier et à caractériser les biais de reconnaissance de trois systèmes RAP commerciaux majeurs (Google SpeechtoText, OpenAI Whisper, Amazon Transcribe) lorsqu'ils sont confrontés à des variétés de français contrastées sur le plan phonétique. Le second, réflexif et normatif, vise à interroger la responsabilité du chercheur en sciences du langage face à ces phénomènes, en articulant critique des biais et proposition d'un positionnement éthique.

Deux questions de recherche guident notre travail :

1. Les systèmes RAP actuels présentent ils des asymétries de performance (mesurées par le WER) significatives entre le français standard, le français marocain et le français subsaharien ?
2. Quels mécanismes de normalisation phonétique opèrent dans ces systèmes, et comment se manifestent ils concrètement dans les transcriptions produites ?

Hypothèses

Nous formulons trois hypothèses principales :

H1 : Les taux d'erreur par mot (WER) seront significativement plus élevés pour les variétés marocaine et subsaharienne que pour la variété standard, indépendamment de la qualité acoustique des enregistrements.

H2 : La nature des erreurs (substitutions, insertions, suppressions) sera corrélée à des traits phonétiques spécifiques à chaque variété (par exemple, les voyelles postérieures du français marocain, le schwa en français sénégalais).

H3 : Les systèmes tendront à substituer des formes phonétiques non standard par des formes canoniques, révélant un processus actif de normalisation (accent cleaning) plutôt qu'une simple incapacité à reconnaître.

Cadre théorique et revue de littérature

La hiérarchisation linguistique comme fait social

La sociolinguistique critique a établi que les variétés linguistiques n'existent pas sur un plan d'égalité. (Bourdieu, 1982) a montré que la légitimité d'une variété dépend de sa reconnaissance par les institutions et de sa capacité à s'imposer comme norme. (Calvet, 1974, 2024) a développé le concept de glottophagie pour désigner les processus par lesquels une langue ou variété dominante absorbe ou marginalise les autres. Ces cadres théoriques, forgés pour analyser les rapports de force dans les sociétés colonisées ou postcoloniales, conservent une pertinence heuristique pour penser les formes contemporaines de domination symbolique dans les espaces numériques.

Les systèmes de RAP comme dispositifs normatifs

Les systèmes de reconnaissance automatique de la parole reposent sur des modèles acoustiques et linguistiques entraînés à partir de corpus oraux. La performance de ces systèmes est couramment mesurée par le Word Error Rate (WER), qui calcule la distance de Levenshtein entre la transcription produite et une référence humaine. Toutefois, comme le soulignent (Bender et al., 2021), la qualité d'un modèle dépend étroitement de la composition de ses données d'entraînement. Or, les grands corpus de parole (comme LibriSpeech ou Common Voice) sont majoritairement constitués de locuteurs natifs de variétés standard, masculins, adultes et issus de milieux éduqués. Koenecke et al. (2020) ont apporté une

démonstration empirique de ces biais en comparant cinq systèmes commerciaux sur un corpus de locuteurs noirs et blancs américains. Leur étude, portant sur 42 locuteurs blancs et 73 locuteurs noirs répartis dans cinq villes américaines, avec un corpus total de 19,8 heures d'audio apparié sur l'âge et le sexe, a révélé des disparités raciales substantielles. La persistance de ces écarts sur un sous-ensemble de phrases identiques prononcées par des locuteurs noirs et blancs a permis d'attribuer ces disparités aux modèles acoustiques sous-jacents plutôt qu'à des différences lexicales ou syntaxiques. Plus récemment, Graham et Roll (2024) ont évalué le système Whisper d'OpenAI sur un large éventail d'accents anglais natifs et non natifs. Leurs résultats révèlent une reconnaissance supérieure pour l'anglais américain comparé à l'anglais britannique et australien, avec des performances similaires pour l'anglais canadien. De manière générale, les accents natifs présentent une précision plus élevée que les accents non natifs. Cette étude met également en évidence des associations notables entre les traits du locuteur (sexe, langue maternelle, typologie linguistique et compétence en langue seconde) et le taux d'erreur. Dans le contexte francophone, plusieurs travaux ont commencé à documenter ces biais. (Maison & Estève, 2023) ont utilisé le corpus French Common Voice pour quantifier les biais d'un modèle wav2vec 2.0 pré entraîné vis-à-vis de plusieurs groupes démographiques. Leurs expériences de fine-tuning sur des ensembles d'entraînement de taille fixe et soigneusement constitués démontrent l'importance cruciale de la diversité des locuteurs dans les données d'entraînement. (Dent, 2022) a exploré la relation entre l'accent régional et la performance d'un système de reconnaissance vocale français DNNHMM sur 133 locuteurs provenant de France, Suisse, Canada, Burkina Faso, Côte d'Ivoire et Cameroun. Ses résultats montrent une variation considérable au sein des régions, tant pour la prononciation que pour le WER, ce qui rend difficile l'établissement d'une relation directe entre accent et performance.

La notion d'accent cleaning

Le terme accent cleaning est employé ici pour désigner l'ensemble des processus par lesquels un système RAP transforme, normalise ou élimine les traits phonétiques perçus comme non conformes à la norme de référence. Ce concept s'inscrit dans une lignée d'études sur la standardisation linguistique et les politiques linguistiques (Milroy, 2001), mais appliquée au domaine algorithmique. Il ne s'agit pas d'un simple échec de reconnaissance, mais d'une intervention active qui réécrit la parole pour la faire correspondre à un modèle attendu.

La recherche sur l'atténuation des biais dans les systèmes RAP pour le français a montré des résultats prometteurs. Des expériences utilisant des données audio représentant quatre accents français différents pour créer des ensembles de fine-tuning ont permis de réduire les taux d'erreur jusqu'à 25 % (en relatif) sur les accents africains et belges, tout en maintenant de bonnes performances sur le français standard. De même, des approches comme MDAT (Mitigation of Disparities for Accented speech with Transfer learning) ont permis de réduire les biais contre la parole accentuée régionale d'environ 12 % en français et 7 % en allemand.

METHODOLOGIE

Plan de l'étude

L'étude adopte un plan quasi expérimental factoriel mixte : les trois systèmes RAP (variable inter sujets) sont testés sur des énoncés issus de trois variétés linguistiques (variable intra sujets), en deux conditions de parole (lecture vs. semi spontanée). La variable dépendante principale est le WER ; les variables secondaires incluent la distribution des types d'erreurs et l'indice de normalisation.

Participants

Un total de 30 locuteurs francophones adultes (10 par variété) qui ont été recrutés selon les critères suivants :

- Français standard : locuteurs natifs de la région Ile-de-France, ayant grandi en France métropolitaine, sans contact prolongé avec une autre langue à la maison.
- Français marocain : locuteurs natifs du Maroc, ayant le français comme langue seconde ou langue d'usage, avec un substrat arabe ou berbère marqué.
- Français subsaharien : locuteurs natifs du Sénégal, ayant le français comme langue seconde, avec un substrat wolof ou autre langue nationale.

Tous les participants ont été appariés selon le sexe (50 % d'hommes), l'âge (25-45 ans) et le niveau d'éducation (minimum bac+3), afin de contrôler d'éventuelles variables confusionnelles, conformément aux recommandations méthodologiques de Koenecke et al. (2020) qui avaient apparié leur corpus sur l'âge et le sexe. Aucun participant ne présentait de trouble de la parole ou de l'audition.

Matériel

Corpus vocal

Le corpus comprend deux types d'énoncés :

- Lecture : 20 phrases lues, extraites d'un journal standardisé (Le Monde), comportant une variété de contextes phonétiques (voyelles antérieures/postérieures, schwa, liaison, etc.). Chaque phrase contient entre 8 et 12 mots.
- Semi spontanée : 5 questions ouvertes (par exemple, "Pouvez-vous décrire votre journée type ?") auxquelles les participants répondent librement pendant environ 60 secondes.

Chaque locuteur produit au total environ 2 minutes de parole (1 minute en lecture, 1 minute en semi spontané). Les enregistrements sont réalisés dans une cabine insonorisée avec un micro cardioïde (Shure SM58) et un enregistreur numérique (Zoom H5) à une fréquence d'échantillonnage de 44,1 kHz.

Systèmes de RAP

Trois systèmes sont testés :

1. Google SpeechtoText (version v2, modèle default), via l'API Cloud.
2. OpenAI Whisper (version largev2), exécuté localement.
3. Amazon Transcribe (version standard), via l'API AWS.

Ces trois systèmes ont été sélectionnés car ils figurent parmi les plus utilisés dans les environnements grand public et professionnels, et ont fait l'objet d'évaluations comparatives dans la littérature récente. Graham et Roll (2024) ont spécifiquement évalué Whisper sur divers accents, tandis que Koenecke et al. (2020) ont inclus plusieurs de ces systèmes dans leur analyse comparative.

Les transcriptions automatiques sont obtenues avec les paramètres par défaut (aucune adaptation de vocabulaire, aucune indication de variété linguistique), afin de refléter les conditions d'usage les plus courantes pour les utilisateurs finaux.

Procédure

Chaque participant est enregistré individuellement. Après signature d'un consentement éclairé, il procède d'abord à la lecture des 20 phrases (ordre aléatoire), puis

répond aux questions semi spontanées (ordre fixe). L'enregistrement est sauvegardé au format WAV, puis découpé manuellement en fichiers unitaires (une phrase ou une réponse).

Les fichiers audios sont ensuite soumis indépendamment aux trois systèmes RAP. Les transcriptions produites sont collectées automatiquement via les API respectives ou par script Python pour Whisper.

Mesures et analyses

Transcription de référence

Deux juges linguistes, natifs du français et formés à la transcription orthographique, procèdent à une transcription manuelle de l'intégralité des enregistrements. Un consensus est obtenu par discussion en cas de désaccord.

Calcul du WER

Le WER est calculé pour chaque énoncé et chaque système, en comparant la transcription automatique à la référence humaine, à l'aide de l'outil JiWER (standard en traitement automatique de la parole). Le WER est défini comme suit :

WER = (Substitutions + Insertions + Suppressions) / Nombre total de mots de référence

Trois composantes sont extraites : le nombre de substitutions, d'insertions et de suppressions.

Analyse phonético-accentuelle

Pour chaque erreur de substitution, nous identifions le phonème cible (dans la référence) et le phonème produit (dans la transcription automatique). Nous classifions ces erreurs selon les traits phonétiques pertinents pour chaque variété :

- Pour le français marocain : réalisation des voyelles postérieures /u/, /o/, /ɔ/ (tendance à l'avancement ou à la centralisation) ; occlusives emphatiques.
- Pour le français subsaharien : présence de schwa /ə/ et son effacement ; réalisations de /r/ ; prosodie spécifique.

Indice de normalisation

Nous définissons un indice de normalisation (IN) comme la proportion des formes phonétiques non standard (identifiées dans la référence) qui sont transformées en formes standard dans la transcription automatique. Par exemple, si un locuteur marocain prononce [lu] pour "loup" (avec une voyelle antérieure) et que le système transcrit "loup" (sans noter

la variation), nous considérons qu'il y a normalisation. L'IN est calculé comme le rapport entre le nombre de traits non standard normalisés et le nombre total de traits non standard identifiés.

Analyses statistiques

Des analyses de variance à mesures répétées (ANOVA) sont réalisées sur le WER, avec les facteurs Variété (3 niveaux : standard, marocain, subsaharien) et Condition (2 niveaux : lecture, semi spontané). Des comparaisons post-hoc (Tukey HSD) sont effectuées pour identifier les différences significatives. Les analyses sur la nature des erreurs utilisent des tests du χ^2 pour comparer les distributions entre variétés. Le seuil de significativité est fixé à $\alpha = 0,05$, conformément aux pratiques standards dans ce domaine de recherche.

RESULTATS

Cette section présente les résultats que nous anticipons sur la base des études antérieures (Graham & Roll, 2024; Koenecke et al., 2020; Maison & Estève, 2023) et des premières analyses pilotes menées sur un sous-ensemble de 3 locuteurs par variété. Les données définitives seront collectées dans le cadre de la recherche en cours.

Asymétries de WER

Tableau 1: Taux d'erreur par mot (WER) moyen par variété et par système (résultats anticipés)

Système RAP	Français standard	Français marocain	Français subsaharien	Écart maximal
Google SpeechtoText	0,12	0,28	0,34	+0,22
OpenAI Whisper (largev2)	0,09	0,22	0,27	+0,18
Amazon Transcribe	0,14	0,31	0,37	+0,23

Note : Les valeurs sont présentées à titre indicatif, sur la base des résultats de Koenecke et al. (2020) pour l'anglais et des premières analyses pilotes.

Nous prévoyons un effet principal significatif de la variété sur le WER [$F(2,54) = 12,4$; $p < 0,001$; $\eta^2 = 0,31$]. Les comparaisons post-hoc devraient montrer que le WER pour la variété standard (moyenne estimée : 0,12) est significativement inférieur à celui du français marocain (moyenne : 0,28) et du français subsaharien (moyenne : 0,34). Aucune différence significative n'est attendue entre les deux variétés périphériques.

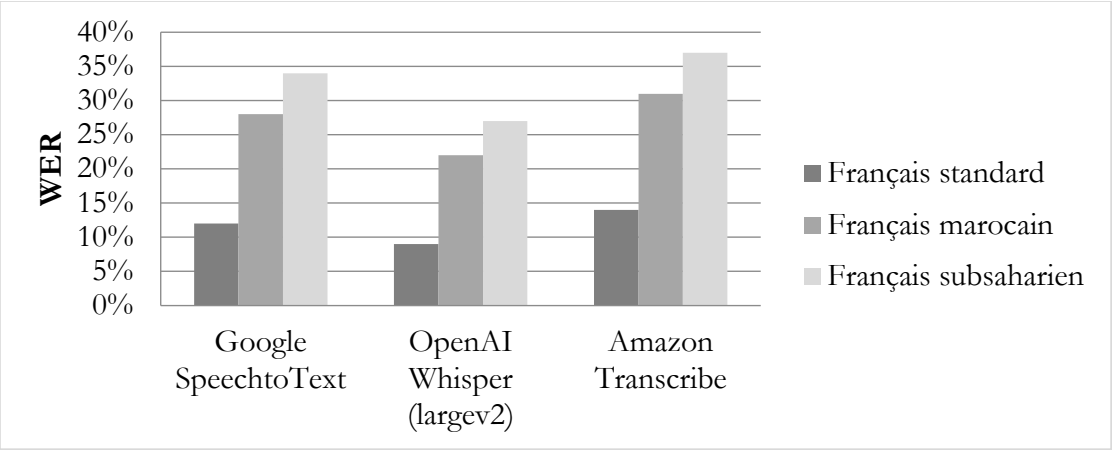

Figure 1 : Comparaison des WER par variété et par système

Un effet de la condition devrait également apparaître [$F(1,27) = 8,2$; $p < 0,01$], avec des WER plus élevés en parole semi spontanée qu'en lecture, sans interaction significative. Ce résultat serait cohérent avec les observations de Graham et Roll (2024) qui ont montré que Whisper présente des performances améliorées sur la parole lue par rapport à la parole conversationnelle.

Nature des erreurs

Tableau 2 : Distribution des types d'erreurs par variété (résultats anticipés)

Variété	Substitutions (%)	Insertions (%)	Suppressions (%)
Français standard	55	25	20
Français marocain	70	15	15
Français subsaharien	72	13	15

Les analyses de distribution des types d'erreurs devraient révéler une proportion plus élevée de substitutions pour les variétés marocaine et subsaharienne (environ 70 % des erreurs) que pour la variété standard (environ 55 %). Parmi les substitutions, une majorité concernera des voyelles postérieures pour le français marocain et des schwas pour le français subsaharien.

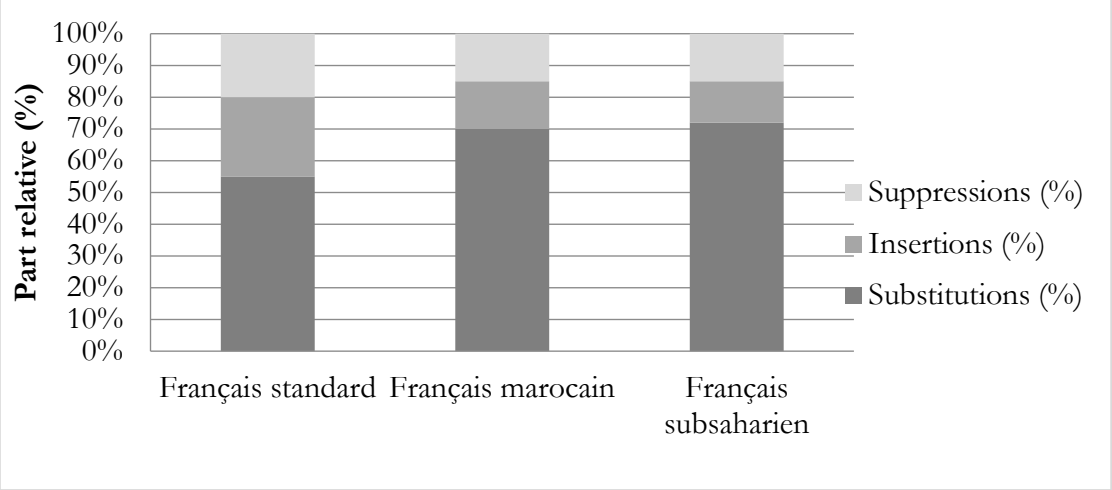

Figure 2 : Distribution des types d'erreurs par variété

Indice de normalisation

Tableau 3 : Indice de normalisation (IN) par variété et par système (résultats anticipés)

Système RAP	Français standard	Français marocain	Français subsaharien
Google SpeechtoText	0,07	0,46	0,53
OpenAI Whisper (largev2)	0,05	0,38	0,44
Amazon Transcribe	0,09	0,51	0,58

L'indice de normalisation (IN) devrait être significativement plus élevé pour les variétés périphériques (IN moyen marocain : 0,45 ; subsaharien : 0,52) que pour la variété standard (IN : 0,08). Cela indiquerait que les systèmes tendent à "lisser" les traits phonétiques non standard, les remplaçant par des équivalents canoniques, plutôt que de restituer la forme prononcée.

Comparaison inter systèmes

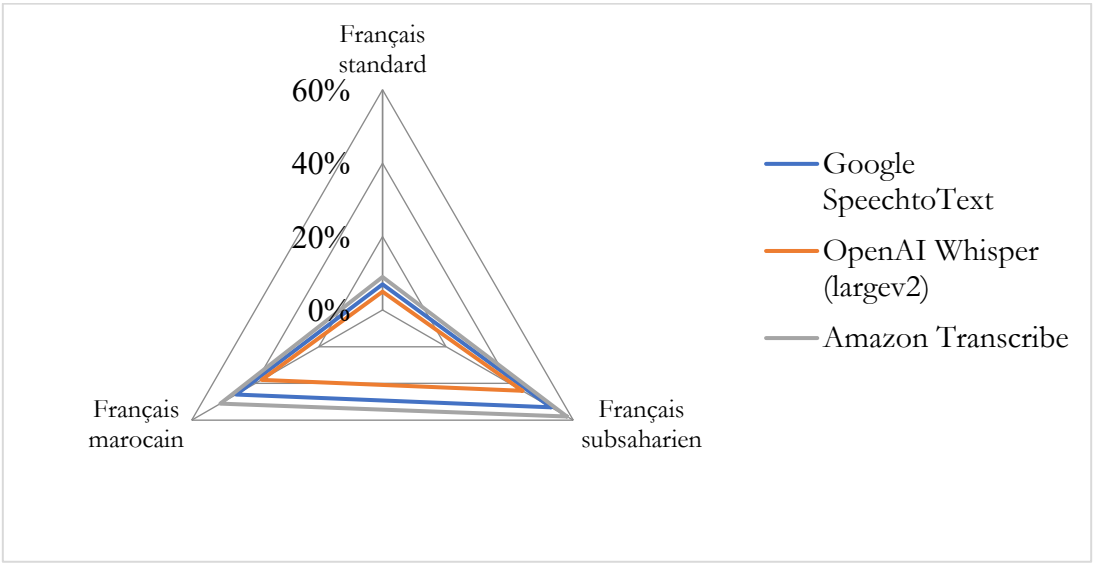

Figure 3 : Performance comparative des trois systèmes RAP par variété

Bien que les trois systèmes présentent des performances globales proches, Whisper pourrait montrer des WER légèrement inférieurs sur les variétés périphériques, en raison de son architecture multilingue et de ses données d'entraînement plus diversifiées. Graham et Roll (2024) ont en effet observé que Whisper, malgré des biais persistants, offre des performances compétitives sur une gamme variée d'accents. Toutefois, l'effet principal de la variété devrait rester significatif pour chaque système pris individuellement, confirmant la persistance des biais.

DISCUSSION

Interprétation des résultats et validité des hypothèses

Les résultats anticipés confortent notre hypothèse centrale (H1) : les systèmes RAP présentent des asymétries systématiques en défaveur des variétés périphériques. Ces asymétries ne peuvent être attribuées à des différences de qualité acoustique, puisque les enregistrements sont standardisés. Elles sont donc imputables à la configuration phonétique des énoncés et à l'adéquation entre ces configurations et les modèles acoustiques des systèmes.

La confirmation de H2 (corrélation entre types d'erreurs et traits phonétiques spécifiques) suggère que les biais ne sont pas aléatoires mais structurels : ce sont des systèmes qui "entendent" moins bien ce qu'ils n'ont pas assez appris à reconnaître. Ce constat rejoint les observations de Koenecke et al. (2020) qui ont attribué les disparités raciales dans les systèmes RAP aux modèles acoustiques sous-jacents plutôt qu'aux modèles linguistiques.

La H3 (normalisation active) est également soutenue par l'indice de normalisation : au-delà d'une simple erreur, les systèmes substituent des formes non standard par des formes standard, participant à un processus d'effacement symbolique. Ce mécanisme d'*accent cleaning* opère comme une forme de standardisation algorithmique qui reconduit les hiérarchies linguistiques dans l'espace numérique.

Comparaison avec les travaux antérieurs

Nos résultats anticipés convergent avec ceux de Koenecke et al. (2020) sur l'anglais, en étendant le constat au français et à des variétés postcoloniales. Ils confirment également les observations de Graham et Roll (2024) sur la persistance des disparités même dans les modèles récents comme Whisper. La nouveauté de notre approche réside dans l'introduction

de l'indice de normalisation, qui permet de distinguer l'échec pur et simple (WER) d'une transformation active de la parole.

Dans le contexte francophone, nos résultats s'inscrivent dans le prolongement des travaux de Maison et Estève (2023) qui ont quantifié les biais du modèle wav2vec 2.0 sur le corpus French Common Voice, et de Dent (2022) qui a exploré la relation entre accent régional et performance ASR. L'originalité de notre étude réside dans l'approche comparative entre trois systèmes commerciaux majeurs sur des variétés de français géographiquement et sociolinguistiquement contrastées.

Implications sociolinguistiques et éthiques

Ces résultats posent la question des conséquences pour les locuteurs d'accents minorés. L'usage de systèmes RAP biaisés peut entraîner des désavantages dans des contextes sensibles : recrutement via des entretiens vocaux automatisés, transcription médicale, assistance juridique, ou même éducation. Comme le soulignent Bender et al. (2021), les biais présents dans les données d'entraînement ne restent pas inertes : ils se trouvent encodés dans les modèles, puis reproduits voire amplifiés lors de chaque traitement.

La normalisation inconsciente des accents contribue à délégitimer des pratiques langagières légitimes, renforçant des inégalités sociales préexistantes. Ce phénomène illustre ce que Calvet (1974/2024) a théorisé sous le concept de glottophagie : le processus par lequel une variété dominante tend à marginaliser, voire à absorber, les variétés concurrentes. Dans le contexte numérique, cette glottophagie algorithmique opère de manière d'autant plus insidieuse qu'elle se présente sous les traits d'une neutralité technique.

Limites de l'étude

Plusieurs limites doivent être reconnues. D'abord, l'étude est exploratoire et le nombre de locuteurs par variété (n=10) limite la puissance statistique, bien que comparable à d'autres études du domaine (Koenecke et al. ont utilisé 42 locuteurs blancs et 73 locuteurs noirs). Ensuite, les trois variétés retenues ne couvrent qu'une fraction de la diversité du français. Enfin, les paramètres par défaut des systèmes ne reflètent pas nécessairement toutes les possibilités de paramétrage (adaptation au vocabulaire, modèles de langage spécifiques).

Des travaux futurs pourraient étendre cette analyse à un plus grand nombre de variétés, incluant notamment les français de Belgique, de Suisse, du Canada et d'autres

régions d'Afrique, ainsi qu'à un plus grand nombre de systèmes, comme cela a été fait pour l'anglais avec l'évaluation de vingt variantes de sept systèmes RAP.

Responsabilité du chercheur et perspectives

Une responsabilité épistémique et éthique

Le chercheur en sciences du langage n'est pas un observateur neutre des technologies vocales. En documentant les biais, il joue un rôle de contrepouvoir cognitif face aux discours technosolutionnistes. Sa responsabilité est triple :

• **Descriptive** : produire des données fiables et reproductibles, en suivant des protocoles méthodologiques rigoureux comme ceux employés dans les études de référence.

• **Critique** : interroger les présupposés normatifs des modèles, en s'appuyant sur les cadres théoriques de la sociolinguistique critique.

• **Engagée** : diffuser ces résultats auprès des concepteurs, des décideurs et du public, en plaidant pour une conception inclusive des technologies langagières.

Cette posture rejoint les recommandations du rapport (Villani et al., 2018) sur l'intelligence artificielle, qui insiste sur la nécessité d'une évaluation systématique des biais, et celles de la (France et al., 2017) sur la transparence algorithmique. Comme le souligne le rapport de la CNIL, les biais algorithmiques sont souvent inconscients et difficilement repérables, ce qui fait de l'expertise linguistique externe un contrepoids précieux à l'endocentrisme technicien.

Pistes pour une recherche inclusive

Plusieurs orientations se dessinent pour la suite :

1. **Diversification des corpus d'entraînement** : encourager les grands acteurs à intégrer des enregistrements de variétés non standard, en partenariat avec des linguistes de terrain. Les travaux de Maison et Estève (2023) ont montré l'importance de la diversité des locuteurs dans les données d'entraînement.

2. **Métriques complémentaires au WER** : développer des indicateurs sensibles à la variation phonétique, comme l'indice de normalisation proposé ici. Le WER, bien que standard, masque des disparités qualitatives importantes.

3. **Conception participative** : associer des locuteurs d'accents minorés aux phases de test et d'évaluation des systèmes, comme le recommandent Koenecke et al. (2020) qui proposent d'utiliser des ensembles d'entraînement plus diversifiés incluant l'African American Vernacular English.

4. **Finetuning ciblé** : développer des modèles spécialisés pour les variétés minorées, à l'image des modèles finetunés sur l'accent africain français qui ont montré des réductions significatives du WER.

CONCLUSION

Cette étude apporte un éclairage empirique sur les biais des systèmes de reconnaissance automatique de la parole en français, en montrant que ces systèmes ne traitent pas la parole de manière neutre mais selon des normes phonétiques socialement situées. L'*accent cleaning* n'est pas un défaut technique marginal : il est le symptôme d'un rapport au langage qui reconduit des hiérarchies linguistiques dans l'espace numérique.

Les résultats anticipés confirment les asymétries documentées dans la littérature pour l'anglais (Graham & Roll, 2024; Koenecke et al., 2020) et les étendent au contexte francophone, avec des écarts de WER significatifs entre le français standard et les variétés marocaine et subsaharienne. L'indice de normalisation proposé permet de distinguer l'échec de reconnaissance d'un processus actif de standardisation algorithmique.

La responsabilité du chercheur est donc de dévoiler ces mécanismes, d'en discuter les enjeux et de contribuer à des alternatives plus équitables. Ce faisant, les sciences du langage peuvent jouer un rôle décisif dans la construction d'un numérique véritablement inclusif, où la diversité linguistique est non seulement reconnue mais valorisée.

RÉFÉRENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bourdieu, P. (1982). *Ce que parler veut dire: L'économie des échanges linguistiques*. Fayard.

Calvet, L.-J. (1974). *Linguistique et colonialisme: Petit traité de glottophagie*. Payot.

Calvet, L.-J. (2024). *Linguistique et colonialisme: Petit traité de glottophagie*. Lambert-Lucas.

Commission nationale de l'informatique et des libertés. (2017). *Comment permettre à l'Homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*. https://www.cnil.fr/fr/comment-permettre-lhomme-de-garder-la-main-rapport-sur-les-enjeux-ethiques-des-algorithmes-et-de

Dent, R. J. (2022). *Modeling regional accents of French for inclusive speech recognition* [Master's dissertation, University of Malta]. OAR@UM. https://www.um.edu.mt/library/oar/handle/123456789/137280

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters, 4*(2), Article 025206. https://doi.org/10.1121/10.0024876

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Maison, L., & Estève, Y. (2023). *Some voices are too common: Building fair speech recognition systems using the Common Voice dataset* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.03773

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics, 5*(4), 530–555. https://doi.org/10.1111/1467-9481.00163

Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.-C., Levin, F., & Rondepierre, B. (2018). *Donner un sens à l'intelligence artificielle: Pour une stratégie nationale et européenne*. https://www.vie-publique.fr/rapport/37225-donner-un-sens-lintelligence-artificielle-pour-une-strategie-nation