

Analyse comparative humaine des modèles ASR

Bambara de RobotsMali

Auteur: Madani Amadou Tall

Institution: RobotsMali AI4D Lab

I. Introduction

La reconnaissance automatique de la parole (ASR) en bambara demeure un défi majeur, en grande partie en raison de la faible disponibilité de données, de la forte variation dialectale, du code-switching fréquent avec le français et/ou l'arabe, ainsi que de la présence notable de bruit dans les enregistrements issus de contextes réels. Au cours des dernières années, RobotsMali a développé et publié plusieurs modèles ASR open-source, entraînés sur différents corpus (Diarra et al., 2022 ; Diarra et al., 2025a ; Diarra et al., 2025b) et reposant sur diverses architectures (QuartzNet, Conformer-CTC/RNNT). Bien que leurs performances varient selon les conditions d'utilisation, aucune étude comparative approfondie n'avait jusqu'ici été menée en combinant des métriques automatiques — telles que le Word Error Rate (WER), le Character Error Rate (CER) — et des évaluations humaines.

L'objectif de ce rapport est de proposer une analyse rigoureuse permettant aux chercheurs et développeurs d'identifier le modèle le mieux adapté à leur cas d'usage : interviews de rue, dictée vocale, environnements bruyants, situations impliquant du code-switching, etc.

Cette étude vise ainsi à évaluer et comparer les performances de huit modèles ASR bambara développés par RobotsMali au moyen d'une double approche :

Approche quantitative : calcul du taux d'erreurs (WER et CER) sur un échantillon représentatif de 45 enregistrements, constitué à partir de deux jeux de données existants, jeli-asr et kunkado (Diarra et al., 2022 ; Diarra et al., 2025a) et de lectures enregistrées par 5 volontaires (2 enfants, 2 hommes, 1 femmes) sur des extraits des ouvrages du projet GAIfe de RobotsMali (Tapo et al., 2025). Ce banc de test est assez représentatif des environnements mentionnés ci dessus puisqu'il est directement tiré d'interviews de rue, d'enregistrement radio (avec assez de code-switching et de bruits de fonds) et de lecture de livre pour enfants

(contexte plus calme, bambara plus simplifié), le tout à part égale pour une évaluation équilibrée et juste.

Approche qualitative : évaluation humaine détaillée portant sur la fidélité sémantique, le traitement des noms propres, la gestion du code-switching, ainsi que la robustesse au bruit et aux chevauchements de voix.

En croisant ces deux dimensions, cette étude cherche à :

1. Identifier le modèle le plus performant de manière globale ;
2. Mettre en évidence les faiblesses systématiques liées aux particularités linguistiques et acoustiques du bambara en situation réelle ;
3. Apporter des réponses concrètes aux besoins pratiques des utilisateurs : choix du modèle pour du bambara simplifié, pour des contextes avec code-switching, pour des environnements bruyants, pour des interactions multi-locuteurs, ou encore pour des applications mobiles.

L'ensemble des modèles étudiés est accessible à l'adresse suivante:
<https://huggingface.co/RobotsMali/models>

La suite du rapport s'organise comme suit :

- II. Méthodologie (constitution du dataset, protocole d'évaluation)
- III. Résultats (WER et évaluations humaines)
- IV. Discussion et analyse croisée
- V. Conclusion et Suggestions

II. Méthodologie

Afin de mener une évaluation comparative des performances des modèles de ASR développés par RobotsMali sur un corpus diversifié et représentatif de la langue bambara, Nous présentons en détail l'approche méthodologique adoptée. Le protocole repose sur trois étapes principales :

1. la constitution du banc de test ;
2. la transcription de cet échantillon par les différents modèles ASR ;

3. l'analyse des performances, combinant le Word Error Rate (WER) et une évaluation humaine qualitative.

L'objectif est de proposer une évaluation multicritères permettant d'identifier le modèle le mieux adapté à différents contextes d'application.

1. Corpus d'évaluation (Dataset)

Le dataset utilisé pour l'évaluation est un sous-ensemble extrait d'un corpus global de 500 enregistrements (Bam ASR Eval), également publié sur [RobotsMali/Bam_ASR_Eval_500](#).

1.1. Sources et constitution du corpus global

Le corpus complet de 500 audios — chaque entrée comprenant *file_name*, *duration* et *transcription* — est issu de la combinaison de trois sous-ensembles principaux, représentant au total 36,69 minutes de parole annotée en bambara :

- **Réf. 1 :** *RobotsMali/kunkado* (Ensemble de Test) – 250 audios (~14,78 minutes)
- **Réf. 2 :** *jeli-ASR street interviews subset* – 30 audios (~1,85 minute)
- **Réf. 3 :** Extraits de lecture provenant de l'application [An Be Kalan](#) (RobotsMali) – 220 audios (~20,06 minutes)

1.2. Échantillonnage pour l'évaluation

- **Taille de l'échantillon :**
Un sous-ensemble constitué de 45 enregistrements, sélectionnés de manière équilibrée (15 par sous-ensemble) à partir du corpus global. Ce sous ensemble est également disponible à [RobotsMali/nyana-eval](#).
- **Critères de sélection :**
Les audios ont été choisis en tenant compte de la diversité des voix, de la présence de différents types de bruit, ainsi que d'une part d'aléatoire. Cette démarche vise à garantir une représentativité adéquate des défis linguistiques et acoustiques caractéristiques de l'usage réel du bambara.

2. Modèles ASR testés

Huit modèles ASR bambara développés par RobotsMali ont été évalués. Ils couvrent trois familles architecturales :

- **QuartzNet** (architecture convolutive légère),
- **Soloba** (Conformer-CTC, modèle de capacité plus élevée),
- **Soloni** (Conformer-TDT-CTC, architecture hybride).

L'ensemble des modèles est basé sur des modèles ASR anglais de NVIDIA et accepte des entrées audio mono-canal échantillonnées à 16 kHz (Kriman et al., 2019 ; Rekesh et al., 2023).

Tableau 1: Liste des modèles ASR de RobotsMali

ID Hugging Face	Architecture	Dataset d'entraînement principal	Paramètres
RobotsMali/stt-bm-quartznet15x5-v0	QuartzNet 15x5 (convolutions 1D séparables)	RobotsMali/bam-asr-early (~35h, Jeli-ASR)	~18M
RobotsMali/stt-bm-quartznet15x5-v1	QuartzNet 15x5	RobotsMali/kunkado (~40h, review humaine)	~18M
RobotsMali/stt-bm-quartznet15x5-v2	QuartzNet 15x5	RobotsMali/afvoices (98h, pre-completion subset)	~18M
RobotsMali/soloba-ctc-0.6b-v0	FastConformer + CTC	RobotsMali/kunkado (~120h) + bam-asr-early	600M
RobotsMali/soloba-ctc-0.6b-v1	FastConformer + CTC	RobotsMali/kunkado (~40h, review humaine)	600M

RobotsMali/soloni-114m-tdt-ctc-v0	FastConformer-T DT-CTC hybride	RobotsMali/bam-asr-early (~35h, Jeli-ASR)	114M
RobotsMali/soloni-114m-tdt-ctc-v1	FastConformer-T DT-CTC hybride	RobotsMali/kunkado (~40h, review humaine)	114M
RobotsMali/soloni-114m-tdt-ctc-v2	FastConformer-T DT-CTC	RobotsMali/afvoices (98h, pre-completion subset)	114M

3. Protocole d'évaluation

L'évaluation initialement envisagée devait reposer exclusivement sur une analyse humaine. Toutefois, il nous a semblé pertinent d'y ajouter le calcul du WER, afin d'examiner la corrélation entre la perception humaine de la qualité et la performance mesurée automatiquement. Cette analyse permet également d'identifier les seuils à partir desquels les variations de WER cessent d'être significatives du point de vue d'un utilisateur humain ou encore les utterances où le WER sous-estime les erreurs qui altère le sens de la communication.

3.1. WER (évaluation quantitative)

Pour assurer la fiabilité et l'équité de la mesure du WER, une étape de normalisation textuelle a été appliquée avant tout calcul. Cette normalisation vise à éviter que des divergences purement formelles (ponctuation, casse, balises, etc.) soient comptabilisées comme des erreurs. La procédure — appliquée à la fois aux transcriptions de référence et à celles produites par les modèles — inclut notamment la mise en minuscules, la suppression des signes de ponctuation et balises, ainsi que la conversion éventuelle des nombres en lettres.

Une fois les données normalisées, le WER moyen a été calculé pour chacun des modèles sur l'ensemble des 45 enregistrements de l'échantillon. Cette étape permet d'évaluer de manière objective la distance lexicale entre les prédictions des modèles et la transcription de référence.

3.2. Score humain (évaluation qualitative)

L'évaluation humaine a été conduite sur les 45 audios retenus, par un locuteur natif du bambara. Elle consiste en une analyse critique des transcriptions produites par chacun des huit modèles ASR, dans une perspective centrée sur l'utilité réelle, au-delà des métriques quantitatives.

Échelle de notation :

Une échelle simple et explicite, allant de 0 à 3, a été utilisée :

- 0 : transcription de qualité insuffisante ou inutilisable ;
- 3 : transcription considérée comme correcte et pleinement satisfaisante.

Tableau 2 : Échelle de pondération de l'analyse humaine

Note	Description	Interprétation
3	Excellente	Transcription sans aucune, fidèle au texte et au sens.
2	Bonne	Présence d'erreurs mineures (découpage des mots, répétition de voyelles), sens intact.
1	Acceptable	Erreurs modérées n'entravant pas la compréhension globale.
0	Médiocre	Nombreuses erreurs, compréhension difficile.

Les critères d'évaluation étaient les suivants :

- **Fidélité textuelle et sémantique** : conformité du contenu transcrit par rapport à la référence, en veillant à ce que le sens général soit préservé (critère principal déterminant le score de 0 à 3).
- **Gestion du code-switching** : capacité du modèle à traiter des segments mêlant plusieurs langues.

- **Transcription des noms propres** : aptitude à restituer correctement les noms de personnes, de lieux ou d'entités spécifiques car ces derniers représentent souvent des sujets ou objets sur lesquels porte le message.

Critère bonus – robustesse aux conditions dégradées :

Un bonus de +0,5 point pouvait être attribué lorsque le modèle parvenait à produire une transcription correcte malgré des conditions audio défavorables (bruit important, faible volume, chevauchement de voix, hésitations). Ce bonus vise à valoriser la robustesse des systèmes.

L'objectif global est donc d'identifier le « meilleur modèle selon l'humain », en se fondant sur la qualité perçue au niveau de chaque fichier et sur l'appréciation générale.

III. Résultats

1. Performance Globale WER

Le tableau ci-dessous indique les WER et CER moyen de chaque modèle pour l'ensemble des 45 audios. Notez que les modèles préfixés “soloni” apparaissent deux fois chacun car ils possèdent deux decoders à évaluer indépendamment.

Tableau 3: Performance WER des modèles sur le dataset

model_id	decoder	WER (%)	CER (%)
RobotsMali/soloni-114m-tdt-ctc-v2	ctc	36,07	20,24
RobotsMali/soloni-114m-tdt-ctc-v2	tdt	38,13	22,30
RobotsMali/soloni-114m-tdt-ctc-v1	ctc	39,44	20,50
RobotsMali/soloba-ctc-0.6b-v1	ctc	40,19	20,94
RobotsMali/soloni-114m-tdt-ctc-v1	tdt	40,19	22,30
RobotsMali/soloni-114m-tdt-ctc-v0	ctc	40,75	24,70

RobotsMali/soloba-ctc-0.6b-v0	ctc	43,36	26,72
RobotsMali/stt-bm-quartznet15x5-v2	ctc	48,97	24,22
RobotsMali/stt-bm-quartznet15x5-v1	ctc	57,20	25,71
RobotsMali/stt-bm-quartznet15x5-v0	ctc	65,42	30,66
RobotsMali/soloni-114m-tdt-ctc-v0	tdt	47,10	31,27

Nous observons donc que le modèle RobotsMali/soloni-114m-tdt-ctc-v2 est le meilleur des modèles par taux d'erreurs.

2. Performance qualitative (score humain)

Le tableau ci-après présente l'analyse humaine réalisée sur les transcriptions produites par chacun des modèles pour l'ensemble des 45 enregistrements (désignés : au1 à au45). Chaque entrée comporte trois éléments:

1. l'identifiant de l'audio évalué (dans le même ordre que sur Hugging Face);
2. le ou les modèles ayant fourni la meilleure transcription ;
3. un commentaire qualitatif détaillant les observations de l'évaluateur.

Pour simplifier la lecture et réduire la longueur des étiquettes, les modèles ont été renommés selon la convention suivante :

- soloni-114m-tdt-ctc-v2 → soloni-v2
- soloni-114m-tdt-ctc-v1 → soloni-v1
- soloba-ctc-0.6b-v1 → soloba-v1
- soloba-ctc-0.6b-v0 → soloba-v0
- soloni-114m-tdt-ctc-v0 → soloni-v0
- stt-bm-quartznet15x5-v2 → quartznet-v2
- stt-bm-quartznet15x5-v1 → quartznet-v1
- stt-bm-quartznet15x5-v0 → quartznet-v0

Ces abréviations permettent de présenter les résultats de manière plus compacte et de faciliter la comparaison entre modèles dans l'analyse qualitative. Pour les modèles soloni, ayant deux decoders, nous utilisons le décodeur par défaut "TDT" pour produire les transcriptions utilisées lors de l'analyse ci-dessous.

Tableau 4: Perception humaine des modèles par audio

ID	Meilleur Modèle / Modèles	Commentaire
au1	soloni-v0	soloni-v2 pénalisé par une substitution à la fin qui change partiellement le sens. Soloni-v0 a une transcription qui correspond à une contraction de la phrase (omission). soloba-v0, soloni-v0, soloni-v2 transcrivent bien le nom Adama.
au2	soloni-v1	soloni-v1 est le meilleur. Les modèles soloni ont globalement les meilleures transcriptions ici, suivi des modèles soloba.
au3	aucun	Aucune bonne transcription, aucun modèle n'est acceptable. Problème avec 'notiw' (notes en Bambara).
au4	soloni-v2	soloni-v2 est le meilleur. soloni-v0 acceptable mais omet des mots avec quelques soucis de découpage.
au5	soloba-v0, soloba-v1; soloni-v0; soloni-v1; quartznet-v2; soloni-v2	Ces modèles ont excellé dans la transcription.
au6	soloni-v0	soloni-v0 est le seul qui se rapproche de la lecture sonore, mais problèmes de découpage/substitutions.
au7	soloni-v2	Transcription excellente. soloba-v1 est le second à avoir une transcription presque parfaite.
au8	soloba-v0; soloba-v1; soloni-v0; soloni-v1; quartznet-v2; soloni-v2	quartznet-v0 a une transcription correcte mais un mauvais découpage à la fin. quartznet-v1 quant à lui présente une substitution qui ne change pas le sens.
au9	soloba-v0; soloni-v0; soloni-v1; quartznet-v2; soloni-v2	Ces modèles ont été excellents.
au10	aucun (soloni-v2 passable)	Aucun modèle n'a réussi à transcrire correctement. soloni-v2 est passable avec plusieurs omissions et substitutions.

au11	soloni-v2 le plus acceptable	soloni-v2 se positionne comme le plus acceptable.
au12	soloba-v1, soloni-v1	soloni-v1 présente une erreur mineure de découpage dans la transcription,néanmoins cela reste presque parfait.
au13	quartznet-v1, soloni-v2, quartznet-v2	ces modèles ont -v1 mieux géré la transcription et gardent le sens de la phrase.
au14	soloni-v2	transcription parfaite de soloni-v2. quartznet-v2 suit avec une transcription presque parfaite
au15	quartznet-v0, quartznet-v2, soloba-v0, soloba-v1, soloni-v0, soloni-v2	quartznet-v0, quartznet-v2, soloba-v0, soloba-v1, soloni-v0, soloni-v2 sont acceptable, soloni-v1 passable
au16	quartznet-v1; soloba-v0; soloba-v1; soloni-v0; soloni-v1; quartznet-v2; soloni-v2	Tous gèrent parfaitement la transcription.
au17	aucun	Aucun
au18	soloni-v1	Transcription correcte. soloba-v1 passable.
au19	quartznet-v1	Transcription excellente, soloni-v1 à du mal avec “mais” et soloni-v2 omet des mots mais garde le sens.
au20	aucun	Aucun modèle n'a gardé le sens (difficulté avec les hésitations dans la voix et le mot francais “parce que”).
au21	soloni-v1	soloni-v1 a une transcription acceptable.
au22	soloba-v1, soloni-v1, soloni-v2	soloba-v1, soloni-v1, soloni-v2 sont les meilleurs. soloni-v0 vient après.
au23	soloni-v2	Le modèle a parfaitement transcrit.
au24	soloba-v0; soloba-v1; soloni-v1	Ces modèles ont parfaitement transcrit.
au25	soloba-v1	soloba-v1 est le meilleur, puis soloba-v0 et soloni-v0 sont acceptables.
au26	tous les modèles soloba, et soloni	Mention spéciale à soloni-v1 qui a même transcrit parfaitement les un chevauchement dans la discussion.
au27	soloni-v1	Transcription non totalement correcte (omission) mais sens fidèlement gardé.
au28	soloni-v1	Transcription excellente.

au29	soloni-v2	soloba-v1; soloni-v1 sont aussi acceptables.
au30	aucun	Mauvaise transcription (difficultés avec les noms 'Kuyate', 'jabate').
au31	aucun	Le son est extrêmement difficile à entendre (voix basse).
au32	aucun	chevauchement des paroles.
au33	aucun	aucun n'a su transcrire. soloni-v1 seul a pu transcrire 'Allahu Akbar'.
au34	aucun	Problème de transcription avec 'Segou', 'souvenir' (Problème de Code-Switching).
au35	aucun	Difficultés avec le changement de langue bambara-français ('Ségou'; 'ville'; 'région').
au36	soloni-v1	Excellent ('Traoré' transcrit en 'Tarawele'). Les deux orthographes sont acceptable
au37	soloni-v1	Excellent pour les noms ('Fatou=Fatu'; 'Coulibaly=Kulibali') et le mot journaliste.
au38	soloni-v0; soloni-v2	Sont les meilleurs dans la transcription.
au39	soloni-v1; soloni-v2	Sont les meilleurs dans la transcription.
au40	quartznet-v1; soloni-v1; soloni-v2	Sont les meilleurs.
au41	tous les soloni, soloba, quartznet-v2 et quartznet-v1	Tous très bons.
au42	aucun	tous mauvais
au43	soloni-v1, soloni-v2	Meilleurs modèles pour cette transcription.
au44	aucun	Difficultés avec 'fête' et 'donc'.
au45	aucun	Difficultés avec les mots 'meilleur' et 'souvenir'.

De ce tableau nous avons donc tiré les pondérations suivantes pour les modèles

Tableau 5: Pondération des modèles par audio après Perception humaine

ID	soloni-v2	soloni-v1	soloni-v0	soloba-v1	soloba-v0	quartznet-v2	quartznet-v1	quartznet-v0
au1	1	0	2	0	0	0	0	0

au2	1	3	1	1	1	0	0	0
au3	0	0	0	0	0	0	0	0
au4	3	0	2	0	0	0	0	0
au5	3	3	3	3	3	3	0	0
au6	0	0	2	0	0	0	0	0
au7	3	0	0	2	0	0	0	0
au8	3	3	3	3	3	3	2	2
au9	3	3	3	0	3	3	0	0
au10	1	0	0	0	0	0	0	0
au11	2	0	0	0	0	0	0	0
au12	0	3	0	3	0	0	0	0
au13	2	0	0	0	0	2	2	0
au14	3	0	0	0	0	2	0	0
au15	2	1	2	2	2	2	0	2
au16	3	3	3	3	3	3	3	0
au17	0	0	0	0	0	0	0	0
au18	0	3	0	0	1	0	0	0
au19	1	1	0	0	0	0	3	0
au20	0	0	0	0	0	0	0	0
au21	0	2	0	0	0	0	0	0
au22	3	3	2	3	0	0	0	0
au23	3	0	0	0	0	0	0	0
au24	0	3	0	3	3	0	0	0
au25	0	0	2	3	2	0	0	0
au26	3	3,5	3	3	3	0	0	0
au27	0	2	0	0	0	0	0	0
au28	0	3	0	0	0	0	0	0
au29	3	2	0	2	0	0	0	0
au30	0	0	0	0	0	0	0	0
au31	0	0	0	0	0	0	0	0
au32	0	0	0	0	0	0	0	0
au33	0	0	0	0	0	0	0	0
au34	0	0	0	0	0	0	0	0
au35	0	0	0	0	0	0	0	0
au36	0	3	0	0	0	0	0	0
au37	0	3	0	0	0	0	0	0
au38	3	0	3	0	0	0	0	0
au39	3	3	0	0	0	0	0	0
au40	3	3	0	0	0	0	3	0

au41	3	3	3	3	3	3	3	0
au42	0	0	0	0	0	0	0	0
au43	3	3	0	0	0	0	0	0
au44	0	0	0	0	0	0	0	0
au45	0	0	0	0	0	0	0	0

Les notes obtenues par les modèles sont :

1. soloni-v1: 59.5

2. soloni-v2: 58

3. soloba-v1: 34

4. soloni-v0: 34

5. soloba-v0: 27

6. quartznet-v1: 21

7. quartznet-v2: 16

8. quartznet-v0: 4

Le modèle soloni-v1 apparaît comme le meilleur selon l'évaluation humaine. De façon générale, lors de l'analyse humaine, les modèles soloni sont remarquablement cités sur presque chaque audio, constituant le top 3. De plus, soloni-v1 se distingue pour sa forte performance sur les noms propres et le code-switching (ex. Adama, Kulibali, Allahu Akbar, journaliste, mais).

Cependant, l'ensemble des modèles présente encore des limites face à plusieurs défis récurrents : chevauchement de locuteurs, code-switching, bégaiements, variations de qualité audio ou encore conditions d'enregistrement dégradées.

Par ailleurs, plusieurs modèles semblent avoir été entraînés sur des variantes orthographiques mixtes pour certains noms propres (ex. *Ségou/Segu*, *Kouyaté/Kuyate*, *Allah/Ala*, *brousse/burusi*). Ces variantes sont acceptables linguistiquement, mais elles reflètent une absence de standardisation dans les données d'entraînement.

En revanche, tous les modèles traitent de manière satisfaisante les expressions et interjections fréquentes du bambara, telles que *nka* (mais), *baara* (travail), *waati* (moment/temps) entre autres.

IV. Analyse des résultats

1. Croisement WER / évaluation humaine

Nous avons croisé les résultats issus des évaluations qualitative et quantitative (WER “etc” pour soloni) afin d’examiner visuellement la relation entre le WER et le score humain, et d’identifier les éventuelles corrélations entre ces deux dimensions. Nous observons clairement que :

- soloni-v2 a le meilleur WER. En revanche sa version v1, entraîné sur 58 h de moins (kunkado) est le meilleur de l’évaluation humaine. De cela nous pouvons déduire que les métriques de distance d’édition comme le WER, bien que d’excellents indicateurs de différence textuelle/lexicale n’incorpore pas la notion de sémantique.
- Une différence de moins de 1% de WER seulement entre soloba-v1 et soloni-v1 a donné un écart de 25,5 points dans la préférence humaine en cumulé. Cela confirme que le WER sous-estime certaines erreurs sémantiques notables et que pour un locuteur natif les erreurs sur certains mots clés ont logiquement plus de poids que d’autres dans la compréhension d’une communication, cette remarque appuie encore le point précédent.
- La faible performance des modèles QuartzNet sur les deux axes, confirmation d’une architecture devenue obsolète, montre que ces petits modèles convolutifs requièrent plus de données et des environnement simplifié pour performer. Quartznet-v1 s’est démarquée avec quelques transcriptions intéressantes sur les données venant des lectures de GAIfE et sur kunkado. Sa version v0 n’a obtenu que 4 points lors de l’évaluation faisant de lui le modèle le moins apprécié de l’analyse et celui avec le pire Word Error Rate.
- Les modèles **soloba**, plus volumineux au sein de la suite RobotsMali, affichent des performances décevantes. Cela pourrait indiquer que l’augmentation du nombre de paramètres d’un modèle améliore les performances jusqu’à un certain plafond à moins que le corpus d’entraînement soit proportionnellement étendue, un phénomène qui corrobore les **lois d’échelle des modèles neuronaux** qui montrent que les deux facteurs doivent être augmentés de concert pour obtenir des gains de performance optimaux (Kaplan et al., 2020).

La figure ci-dessous montre un croisement des métriques sur deux axes.

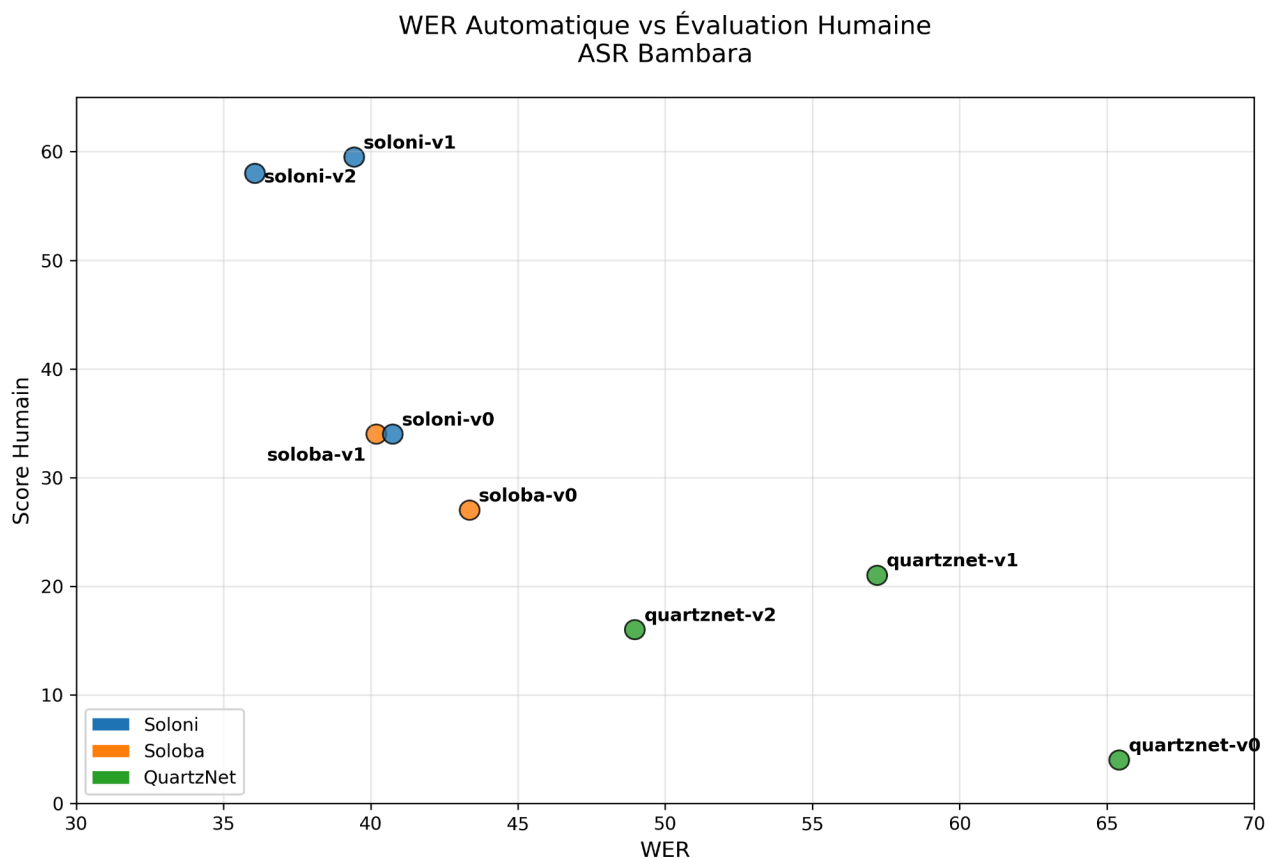


Figure 1: Croisement des métriques WER (“etc” pour soloni) et de la perception Humain.

2. Analyse des erreurs systématiques des modèles

La plupart des modèles ont échoué sur des audios qui comportaient des noms propres ou du code-switching et ont eu également du mal face à des défis comme le bégaiement et des voix basses.

Tableau 6 : Erreurs systématiques des modèles ASR

Thème	Observation
Noms propres	Très mal gérés globalement (Adama, Kouyaté, Diabaté, Traoré, Fatou, Coulibaly). Souvent substitués ou découpés. Exception soloni-v1

Mots français en Bambara et Code switching	<p>parce que → souvent transcrit "paseke" (surtout soloni-v2).</p> <p>Ordinateur, journaliste, souvenir, ville, région → mal compris (tous les modèles). Souvent gros problème de segmentation et perte de sens.</p> <p>"Traoré" → "Tarawele" soloni-v1 semble entraîné sur cette forme (meilleur modèle sur les noms propres)</p>
Hésitations vocales	Les modèles perdent souvent le sens dès qu'il y a des hésitations ou bégaiements.
Voix basse / bruit	au31 → inaudible → tous échouent.
Chevauchement vocal/bruits de fond	Fin d'audio avec deux voix → découpage erroné.
Contractions / liaisons	soloni-v0 se rapproche souvent d'une prononciation orale, mais souffre de mauvais découpage.
Omission de mots	soloni-v1 et soloni-v2 omettent souvent des mots mais gardent le sens.

V. Conclusion et perspectives

L'évaluation de ces modèles ASR a mis en évidence une distinction intéressante dans les performances de soloni-v2 et soloni-v1 selon les axes d'analyse.

Synthèse des Résultats Clés

Tableau 7: Synthèse des résultats

Modèle	Métrique	Performance	Interprétation

soloni-v2	WER (Word Error Rate)	36,07 % (le plus faible)	Meilleure précision textuelle brute.
soloni-v1	Score Humain	59,5/135 (le plus élevé)	Meilleure qualité perçue et fidélité sémantique.

Bien que le modèle soloni-v2 se soit distingué par le WER le plus faible, traduisant une supériorité en termes de précision textuelle stricte sur Nyana Eval, c'est le modèle soloni-v1 qui a obtenu la meilleure évaluation humaine.

Nyana Eval, compilé pour refléter les conditions réelles d'utilisation au Mali — où le bruit de fond et le mélange linguistique (code-switching) sont omniprésents — a ainsi révélé une faiblesse sémantique du modèle soloni-v2 dans ces environnement par rapport à soloni-v1, résultant en une petite différence de 1,5 points au général. Ainsi, soloni-v1 bénéficie d'avoir été préparé pour ce type de conditions lors de son entraînement plus que de la taille du corpus utilisé.

Travaux futurs

Les travaux futurs devraient se concentrer sur l'intégration des forces des deux modèles. Le WER de soloni-v2 indique qu'il possède un meilleur lexique Bambara mais échoue à s'adapter au bruit de fond et au code-switching (largement absent de son ensemble d'entraînement). Nous proposons donc :

1. Extension du corpus :

- Ajouter des enregistrements bruités, des noms propres de lieux et de personnes maliens, et des segments comportant du code-switching au données d'entraînement de soloni-v2 afin de mieux représenter les conditions rencontrées dans des applications réelles. Éventuellement l'entraîner sur le même ensemble que sa version v1 (Diarra et al., 2025a)

2. Post-traitement :

- Normalisation des noms et des variantes orthographiques (ex. *Traoré* → *Tarawele*, *Paseke*, *Segu*, etc.) pour assurer une cohérence dans les transcriptions.

3. Robustesse acoustique :

- Entraîner les modèles avec des audios contenant du bruit ambiant, des chevauchements de voix ou des effets de réverbération pour améliorer leur résilience. Encore une fois, RobotsMali/kunkado semble être le bon ensemble pour cela.

4. Version mobile :

- Préférer la quantification des modèles soloni (v1 ou v2 en fonction de l'environnement de production) pour les déploiements sur mobile plutôt que l'utilisation de QuartzNet, pour de meilleures performances.

Références

Sébastien Diarra, Michael Leventhal, et Allahsera Auguste Tapo. 2022. Robotsmali griots speech dataset, and asr. <https://github.com/robotismali-ai/jeli-asr/>.

Yacouba Diarra, Nouhoum Coulibaly, Panga Azazia Kamaté, et Michael Leventhal. 2025a. [kunafonidilaw ka cadeau: an ASR dataset to power the development of models that understand present-day bambara](#). Hugging Face Datasets. Arxiv coming soon

Yacouba Diarra, Nouhoum Coulibaly, Panga Azazia Kamaté, Madani Amadou Tall, Emmanuel Élisé Koné et Michael Leventhal. 2025b. [African Next Voices Bambara: A 600h open ASR dataset for Bambara](#). Hugging Face Datasets. Arxiv coming soon

Samuel Krman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, et Yang Zhang. 2019. [Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions](#). Preprint, arXiv:1910.10261.

Dima Rekish, Nithin Rao Koluguri, Samuel Krman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et Boris Ginsburg. 2023. [Fast conformer with linearly scalable attention for efficient speech recognition](#). Preprint, arXiv:2305.05084.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). Preprint, arXiv:2001.08361.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, et Michael Leventhal. 2025. [GAIfE: Using GenAI to improve](#)

[literacy in low-resourced settings](#). In Findings of the Association for Computational Linguistics: NAACL 2025, pages 7914–7929, Albuquerque, New Mexico. Association for Computational Linguistics.