Préservation du contenu médical par des systèmes de synthèse vocale ouverts

Ricardo Rodriguez[1] Stéphane Huet[1] Benoît Favre[2] Mickael Rouvier[1]
(1) LIA, Avignon Université, France
(2) LIS, Université d'Aix Marseille, France
firstname.lastname@univ-avignon.fr, lis-lab.fr

RÉSUMÉ

Le secteur des soins cliniques connaît un flux croissant de données à traiter, ce qui justifie l'introduction de nouvelles technologies mieux adaptées au domaine de la santé. Les Grands Modèles de Langue avec interfaces vocales apportent des bénéfices potentiels, mais nécessitent de grandes quantités de données audio pour l'entraînement, extrêmement difficiles à collecter vue leur haute sensibilité, introduisant des enjeux éthiques et réglementaires pour leur exploitation. Dans cet article nous étudions la capacité de quatre modèles TTS ouverts à synthètiser des textes médicaux, et nous évaluons la préservation du contenu via des méthodes automatiques : taux d'erreur de mots (WER), taux d'erreur sur des concepts médicaux (M-WER) et performance dans des tâches de TAL du modèle BioLinkBERT-large évalué sur la transcription des données synthétiques. Les textes utilisés proviennent des corpus BLURB, et l'ensemble des données générées sera rendu publiquement disponible sous licence ouverte.

ABSTRACT

Preservation of medical content with open TTS models

In clinical scenarios, the increasing volume of health data to be processed is driving a growing need for technologies adapted to the field. Speech LLMs display potential benefits in this setting but they require large amounts of audio medical data for training, which is difficult to collect and highly sensitive, raising ethics and legal questions. In this article, we study the capabilities of four highly rated open TTS models to create synthetic audio from medical text and evaluate the preservation of medical content. We synthesise text from the BLURB corpora and compare the outputs using a combination of automated tools to calculate Word Error Rate (WER), Word Error Rate over Medical terms only (M-WER), and BLURB performance of BioLinkBERT when evaluated on this synthetic data. We also make the generated audio available to the community under an open license.

MOTS-CLÉS : TTS, synthèse de la parole, médical, santé, kokoro, fish-speech, style-tts2, zipvoice, blurb.

KEYWORDS: TTS, text-to-speech, medical, health, kokoro, fish-speech, style-tts2, zipvoice, blurb.

1 Introduction

Les technologies de synthèse vocale (« text-to-speech », TTS), de reconnaissance automatique de la parole (« automatic speech recognition », ASR) et les grands modèles de langue (« large language models », LLM) ont démontré un potentiel significatif pour de nombreuses applications tant dans le domaine général que dans ceux de spécialité. Depuis quelques années nous observons un intérêt

croissant pour l'utilisation de ces technologies dans le milieu médical, que ce soit pour améliorer la prise en charge des patients, accélérer les processus administratifs, et servir de soutien du corps médical pour la réalisation de tâches qui sont souvent très chronophages, au détriment du temps disponible pour l'échange avec les patients. Les modèles ASR sont particulièrement utiles pour accélérer la transcription des consultations cliniques (Adedeji *et al.*, 2024), tandis que les LLM peuvent aider à compiler et à résumer rapidement des informations médicales en grand volume (Vasilev *et al.*, 2026) et à extraire des données du dossier médical (Chua *et al.*, 2024) dans des systèmes d'aide à la décision. Cependant, l'utilisation et l'évaluation de systèmes TTS dans le domaine médical est un champ d'étude encore largement inexploré. De plus, bien que des grands corpus textuels soient disponibles publiquement dans le domaine médical, tels que ApolloCorpus (Wang *et al.*, 2024) et MIMIC-IV (Johnson *et al.*, 2023), ce n'est pas le cas pour les données audio, nécessaires pour l'entraînement et l'amélioration de modèles ASR et de modèles LLM vocaux (« speech LLM »). De plus, l'utilisation d'enregistrements cliniques réels dans la recherche soulève des préoccupations quant à l'éthique et à la confidentialité.

Nous partons de l'hypothèse que la synthèse de grandes quantités de texte médical vers la modalité audio pourrait être d'une grande utilité pour couvrir ce manque. Dans cette étude, nous cherchons à explorer et à comparer les capacités de quatre modèles de TTS — Kokoro, Fish-Speech, Style-TTS2 et ZipVoice — à synthétiser des textes médicaux en anglais, dans des scénarios impliquant un seul locuteur ou plusieurs locuteurs. Nous évaluons leurs performances à l'aide d'outils automatiques afin de calculer le taux d'erreur de mots (« word error rate », WER) et le taux d'erreur médicale (« medical word error rate », M-WER), et nous utilisons les tâches du benchmark BLURB comme métrique pour évaluer leur capacité à conserver le contenu médical dans le document d'origine.

L'évaluation des systèmes TTS repose généralement sur deux grandes familles de métriques : les métriques subjectives, fondées sur l'écoute humaine, et les métriques objectives, calculées automatiquement à partir du signal audio ou de son contenu.

Parmi les approches subjectives, le score d'opinion moyen (« mean opinion score », MOS) est la méthode la plus couramment utilisée pour mesurer la qualité des attributs de parole tels que la fluidité, la clarté, l'intelligibilité, l'expressivité émotionnelle et la cohérence dans le clonage de voix. Les tests comparatifs paire-à-paire, tels que le score d'opinion comparatif (« comparative MOS », CMOS), demandent à des personnes évaluatrices d'attribuer leur préférence d'un enregistrement par rapport à un autre ; ils sont souvent utilisés lorsque l'on compare plusieurs TTS et ont montré des résultats significatifs dans des scénarios avec peu de personnes participantes (Cooper *et al.*, 2024). Les tests subjectifs nécessitent beaucoup de temps et de ressources pour obtenir des résultats statistiquement significatifs et sont sujets à des problèmes d'incohérence et de variabilité, principalement dus à la subjectivité des personnes impliquées.

En complément, des métriques objectives issues du traitement du signal permettent de quantifier automatiquement les écarts entre un enregistrement humain naturel (de référence) et un signal synthétique ; par exemple, la méthode connue en anglais sous le nom de « Mel-Cepstral Distortion » (MCD) mesure les différences spectrales en comparant les coefficients cepstraux (« mel-frequency cepstral coefficients », MFCCs) d'une trame synthétique et d'un enregistrement de référence. De même, le WER peut être utilisé pour évaluer la clarté et la qualité du contenu en transcrivant de l'audio synthétique à l'aide d'un système ASR et en le comparant au texte original. Néanmoins, l'utilisation d'un système intermédiaire introduit un degré de biais et d'incertitude en raison des erreurs éventuelles de transcription et des limites du modèle ASR. De plus, la métrique WER attribue autant d'importance à tous les types d'erreurs (insertions, suppressions, substitutions) et ce pour

n'importe quel mot dans la phrase (Roux *et al.*, 2022). Cette simplicité est justement la force de la métrique pour de nombreux scénarios mais ne fournit pas toute l'information nécessaire lorsque l'on évalue les TTS dans des domaines spécialisés comme la santé, où les erreurs portant sur des termes médicaux sont bien plus graves.

Nos contributions sont les suivantes :

— Nous proposons une nouvelle approche pour évaluer de manière approfondie la préservation du contenu médical dans des enregistrements audio générés utilisant des systèmes TTS en partant de texte médicaux, avec une combinaison de métriques automatiques.
— Nous proposons l'utilisation des tâches BLURB pour mesurer les capacités de modèles TTS pour synthétiser des données médicales.
— Nous allons diffuser les données générées, qui comportent plus de 20 mille heures d'audio, sous une licence ouverte.

Cet article est organisé comme suit. La Section 2, présente les modèles TTS étudiés, le modèle ASR utilisé ainsi que le corpus BLURB retenu pour nos évaluations. La Section 3 décrit le cadre expérimental de notre étude. Ensuite, la Section 4 présente les résultats comparatifs par les différents modèles TTS sur les corpus d'évaluation. Enfin, la Section 5 conclut l'article en synthétisant nos observations et en présentant les perspectives liées aux travaux en cours et futurs.

2 Méthodologie expérimentale

2.1 Sélection de modèles

Dans cette étude, nous analysons les performances de quatre systèmes TTS — Kokoro, Style-TTS2, Fish-Speech et ZipVoice — pour la synthèse de données médicales. Ces modèles ont été sélectionnés en raison de leur bon classement, en juillet 2025, sur deux plateformes publiques pour la comparaison des TTS (TTS Arena) sur Hugging Face [1], et de leurs licences gratuites et ouvertes. Notre choix était fondé sur leur capacité à générer de l'audio de haute qualité à partir de données texte, leurs capacités prouvées dans les scénarios à une seule voix et multiples voix, et la possibilité de les exécuter dans nos centres de données locaux et nationaux sans coûts d'utilisation élevés. Cette étude vise à évaluer leur aptitude à synthétiser des grandes quantités de données médicales pour finalement identifier le modèle le plus approprié pour de futurs travaux sur les grands modèles de langue appliqués au domaine de la santé.

1. Kokoro (Hexgrad, 2025) : Avec 82 millions de paramètres, Kokoro est un petit modèle avec des bons résultats dans des évaluations anglophones et a été retenu pour ce travail, malgré son impossibilité de cloner des voix, grâce au classement élevé sur les plate-formes comparatives. Basé sur Style-TTS2 avec le vocodeur ISTFTNet (Kaneko *et al.*, 2022), c'est un modèle décodeur sans couche de diffusion, utilisant l'outil qui convertit des graphèmes en phonèmes (G2P) « misaki ». Kokoro, tout comme Style-TTS2, promet le support des langues français et anglais, parmi d'autres.
2. Style-TTS2 (Li *et al.*, 2023) : Développé par une équipe de l'Université de Columbia à New York, Style-TTS2 est un modèle de diffusion de style (« style diffusion model ») entraîné avec la technique d'apprentissage par adversaire, utilisant un décodeur HifiGAN, sur des données du corpus LibriTTS (245h) (Zen *et al.*, 2019).

1. https://huggingface.co/spaces/Pendrokar/TTS-Spaces-Arena, https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2

3. Fish-Speech (Liao *et al.*, 2024) : Avec 1,2 milliards de paramètres, c'est le modèle le plus grand évalué dans ce travail. Il présente une architecture double avec deux modules auto-régressifs complémentaires, décrite comme « Serial fast-slow Dual AutoRegressive ». Il utilise une quantification GFSQ (« Grouped Finite Scalar Vector Quantization »), ne dépend pas d'une étape de conversion de graphèmes à phonèmes (G2P), et utilise un LLM en interne pour l'extraction de caractéristiques linguistiques. Le vocodeur employé appartient à la famille EVA-GAN. Le modèle promet le support des langues français et anglais, mais la taille du corpus d'entraînement en anglais est plusieurs fois plus grande que celle du français (300 mille heures vs 20 mille heures).
4. ZipVoice (Zhu *et al.*, 2025) : Ce modèle utilise une technologie de « Conditional Flow-Matching » (CFM) pour générer des enregistrements basés sur un prompt audio de quelques secondes (voix à cloner) et au départ une matrice de bruit qui est progressivement modifiée. Il est basé sur le modèle Zipformer, un encodeur initialement créé pour des système ASR. ZipVoice a été entraîné sur 555 h du corpus LibriTTS (Zen *et al.*, 2019).

Notre méthodologie d'évaluation, détaillée dans la section suivante, repose sur un modèle ASR chargé de transcrire l'audio en texte, afin de comparer la transcription obtenue au texte de référence. Dans cette étude, nous avons sélectionné le modèle Whisper, largement utilisé dans la communauté, et plus spécifiquement la version Whisper-Medium.en (Radford *et al.*, 2023). Développé par OpenAI, Whisper est un modèle de type transformeur encodeur-décodeur qui traite des entrées audio découpées en segments de 30 secondes. Il a montré d'excellentes performances dans plusieurs benchmarks pour la reconnaissance de la parole. Whisper est disponible en différentes tailles, allant du modèle « small », avec 39 millions de paramètres, aux versions « medium » et « large », qui comptent respectivement 769 et 1550 millions de paramètres. Nos tests préliminaires avec les versions Medium et Large ont montré que Medium offrait le meilleur compromis de qualité/coût, avec le coût mesuré en termes de temps d'exécution et de consommation de ressources.

Nous reconnaissons que l'utilisation d'un seul modèle ASR introduit certaines limites à cette étude. Différentes erreurs peuvent survenir lors du changement de modalité — du texte vers l'audio, puis de l'audio vers le texte — dont la cause est difficile à déterminer, pouvant provenir du modèle TTS ou ASR. Cette limite impacte la précision des métriques WER et M-WER et l'analyse des résultats. De plus, puisque Whisper-Medium.en est entraîné uniquement à partir de données en anglais, les erreurs de transcription peuvent être biaisées vers des problèmes spécifiques à la langue ou des idiosyncrasies qui ne sont pas représentatives d'autres langues ; certaines conclusions ne seront donc pas directement généralisables dans d'autres langues et dans des contextes multilingues.

Néanmoins, cette étude utilise la sortie du modèle ASR et des métriques fondées sur le taux d'erreur de mots comme moyen de comparer les performances relatives des modèles TTS. Cela nous permet d'identifier le système le plus performant dans une même configuration d'évaluation. Les résultats numériques sont donc analysés en termes de différences relatives, et non de manière absolue.

2.2 Corpus BLURB

Le benchmark BLURB (« Biomedical Language Understanding and Reasoning Benchmark ») (Gu *et al.*, 2021) est une grande collection de ressources pour le traitement du langage naturel biomédical. Il vise à évaluer la capacité des modèles de langue à comprendre et à traiter des données médicales. Développé par Microsoft, BLURB propose des ressources de haute qualité pour l'évaluation des performances sur plusieurs tâches de TAL, telles que la reconnaissance d'entités nommées (« named

entity recognition », NER), l'extraction de relations entre deux phrases (« relation extraction », RE), la similarité de phrases (« sentence similarity », SS), la classification de documents (« document classification », DC), et les questions-réponses (« question answering », Q/A).

Dans ce travail, nous nous focalisons sur 9 corpus couvrant 4 tâches, pour lesquels nous avons synthétisé la partie test. Cela représente plus de 776 000 mots et un total de 320 heures d'audio en moyenne par TTS. Les corpus utilisés sont brièvement présentés par la suite et regroupés selon les différentes catégories de tâches de TAL.

Reconnaissance d'Entités Nommées (NER) : Les données comprennent des phrases variées dans le médical avec des annotations humaines pour indiquer les début et fin des termes médicaux, sans spécifier la catégorie du terme (e.g. drogue, condition, etc.). Elles sont issues de différentes sources : BioCreative II Gene Mention Recognition (BC2GM) (Smith *et al.*, 2008), BC5CDR Drug/Chemical (BC5-Chem) (Li *et al.*, 2016), BC5CDR Disease (BC5-Disease) (Li *et al.*, 2016), JNLPBA (Collier *et al.*, 2004), NCBI Disease (Doğan *et al.*, 2014).

Similarité de Phrases (SS) : Le seul corpus dans cette catégorie, BIOSSES, est composé de 100 paires de phrases sélectionnées du jeu de données « Biomedical Summarization Track Training Dataset », de la conférence « Text Analysis Conference » (TAC), provenant d'extraits d'articles dans le domaine médical. L'ensemble de test comprend uniquement 20 paires de phrases de longueurs variées (Soğancıoğlu *et al.*, 2017).

Classification de Documents (DC) : Le seul corpus dans cette catégorie est connu sous le nom de « Hallmarks of Cancer » (HoC) et contient des phrases extraites d'abstraits d'articles sur le site PubMed, annotées manuellement avec les classes d'appartenance dans la classification HoC (zéro classes ou plus) (Baker *et al.*, 2016a,b, 2017; Baker & Korhonen, 2017).

Questions-Réponses (Q/A) : Les données sont constituées de triplets question, contexte et réponse, où la question peut être répondue par « oui » ou « non » en s'appuyant sur le contexte donné, avec des contextes de largeur variable mais pouvant être assez longs. Elles sont de type : BioASQ (Tsatsaronis *et al.*, 2012), PubMedQA (Jin *et al.*, 2019).

Nous avons exclu de ces travaux les corpus d'extraction de relations à cause de la complexité de leur traitement. Les données de ces corpus sont composées de phrases relativement courtes (10-20 mots) où deux termes distincts ont été masqués (un mot ou plus portant sur des noms de drogues, composants chimiques ou noms de gènes), en plus d'une étiquette indiquant la relation entre les mots telle que proposée par des annotateurs humains. Ces données contiennent des masques, sous la forme `@DRUG$`, difficiles à traiter ou à ignorer par les TTS ; cela aurait nécessité un pré-traitement approprié pour les transformer avant la synthèse.

Il est important de noter que les corpus de BLURB ne sont pas directement accessibles sur le site web du benchmark [2], mais qu'ils doivent être téléchargés depuis leur source originelle et pré-traités suivant les indications de BLURB. Certains de ces corpus n'étant plus disponibles aux adresses indiquées, nous avons utilisé les fichiers disponibles sur https://github.com/michiyasunaga/LinkBERT, le dépôt git du modèle BioLinkBERT qui est accessible publiquement.

2. https://microsoft.github.io/BLURB

Corpus	Nb lignes	Nb mots	Mono-voix	Multi-voix
BC2GM	5k	143k	57h	657h
BC5-Chem	4.8k	124k	49h	588h
BC5-Disease	4.8k	124k	49h	588h
JNLPBA	3.8k	101k	45h	682h
NCBI	0.9k	24.4k	10h	640h
BIOSSES	20	1k	0.5h	288h
HoC	371	93.7k	41h	6365h
BioASQ	140	37.5k	16h	6583h
PubMedQA	500	128k	53h	6106h
Total	**20.3k**	**776.6k**	**320.5h**	**22500h**

TABLE 1 – Statistiques de la partie test des corpus BLURB. Les nombres de lignes et de mots sont exprimés en milliers. La colonne « Mono-voix » indique la durée, en heures, des audios générés en moyenne par les TTS, tandis que « Multi-voix » indique la durée pour la synthèse utilisant des voix multiples.

2.3 Métriques

Dans nos travaux, nous utilisons les métriques objectives décrites ci-dessous afin de mesurer la préservation du contenu écrit après transcription automatique, et d'analyser les différences relatives entre les quatre modèles TTS évalués.

Taux d'erreur de mots, « Word Error Rate » (WER) : le WER est une métrique largement adoptée par la communauté scientifique mesurant la similarité entre deux textes, référence et transcription automatique, généralement utilisée pour mesurer la précision des systèmes ASR. Le WER est calculé comme le rapport entre le nombre de substitutions, suppressions et insertions de mots, et le nombre total de mots dans la référence.

$$\text{WER} = \frac{S + D + I}{N}$$

où S est le nombre de substitutions, D le nombre de suppressions, I le nombre d'insertions, et N le nombre total de mots dans la référence.

Nous utilisons le WER (macro, sur l'ensemble des corpus) entre le texte de référence et la transcription automatique des audios synthétiques comme un moyen de mesurer la perte d'information à la fin du pipeline. L'utilisation d'un modèle ASR induit une marge d'erreur additionnelle : le WER final dépend à la fois des difficultés du TTS à synthétiser certains termes spécialisés et des limites d'un ASR généraliste à les reconnaître. Ainsi, notre analyse porte principalement sur les différences de taux d'erreur entre systèmes TTS, plutôt que sur les valeurs absolues obtenues.

Taux d'erreur médical, « Medical Word Error Rate » (M-WER) : le M-WER est une métrique introduite dans des travaux précédents pour l'évaluation de systèmes ASR dans le domaine médical (Afonja *et al.*, 2024). L'objectif est de porter l'attention sur les concepts médicaux pour mieux comprendre la précision de ces systèmes face à des enregistrements dans le domaine de spécialité. Le M-WER permet d'observer les systèmes sur leur capacité spécifique à traiter des termes biomédicaux, qui pourrait être négligée avec uniquement l'utilisation du WER classique. Des travaux précédents ont montré comment les modèles ASR états-de-l'art peuvent être affinés avec des données médicales

afin d'obtenir de meilleurs résultats, en particulier quand il s'agit de traiter des accents méconnus par les modèles (Afonja *et al.*, 2024). Dans notre travail, nous calculons le score de M-WER uniquement pour les jeux de données BLURB pour la reconnaissance d'entités nommées (NER) parce que ce sont les seuls annotés avec des entités médicales. La méthode de calcul est fortement inspirée de l'article cité précédemment, basée sur un algorithme d'alignement entre prédiction et référence « fuzzy string matching ».

Métriques BLURB : Finalement, nous utilisons les métriques proposées par le benchmark BLURB afin de quantifier l'impact de la perte de données à la fin du pipeline TTS+ASR. Nous avons sélectionné le modèle avec les meilleurs scores dans le classement BLURB, BioLinkBERT-large (Yasunaga *et al.*, 2022), que nous avons entraîné sur les tâches BLURB à partir des ensembles d'entraînement et de validation de chaque corpus, puis évalué sur l'ensemble de test issu de l'ASR. Ainsi, les ensembles d'entraînement et de validation des corpus sont utilisés sans modification aucune. Notre hypothèse est que la perte d'information dans les donnés synthétiques peut être mesurée en fonction de la perte de précision du modèle BERT ; des termes médicaux manquant devraient impacter négativement la capacité du modèle à suivre une tâche donnée. Là encore, nous considérons les résultats mesurés en différences relatives entre les TTS et non en valeurs absolues.

2.4 Pipeline d'évaluation

Notre pipeline de synthèse de textes médicaux et d'évaluation de la préservation du contenu comprend plusieurs étapes : (1) chaque corpus est prétraité afin de produire une chaîne textuelle unique, ensuite segmentée en sous-phrases pour obtenir des enregistrements d'une durée maximale de 30 secondes ; (2) chaque système TTS est utilisé pour synthétiser le jeu de test de chaque corpus BLURB sélectionné ; (3) le modèle ASR transcrit les fichiers audio afin d'obtenir un texte comparable aux données initiales, puis les métriques WER et M-WER sont calculées, en micro et en macro ; (4) enfin, nous affinons BioLinkBERT sur chaque jeu d'entraînement et exécutons les différentes tâches BLURB avec les hyperparamètres définis par les développeurs de BioLinkBERT.

Toutes les étapes sont d'abord réalisées en utilisant une voix unique, commune à tous les systèmes TTS. Étant donné que Kokoro n'est pas capable de cloner des voix, nous utilisons l'une des voix connues de ce modèle pour la génération avec tous les autres ; pour ce faire nous avons synthétisé avec Kokoro une phrase simple en anglais que nous avons utilisé par la suite en tant que donnée d'entrée pour les autres systèmes : `This is just a simple test with a sample voice, reading letters A B C and numbers 1 2 3`. Pendant cette phase, la totalité des jeux de données de test est synthétisée. Chaque modèle est configuré avec les paramètres par défaut ou recommandés par ses créateurs.

Une deuxième phase vise à évaluer les modèles dans un scénario de voix multiples à partir d'une sélection de 576 voix différentes et uniques provenant du jeu de données « CommonVoice v24 Scripted » (anglais) (Ardila *et al.*, 2020). Cette sélection permet de couvrir une diversité de voix et d'accents, tant féminines que masculines. Pour cette deuxième phase, nous avons sélectionné un petit échantillon de chaque corpus composé de 100 exemples pris au hasard, afin de limiter le coût d'exécution. Nous avons exclu Kokoro de cette analyse car il ne permet pas de cloner des voix en zéro-shot.

CommonVoice est une plate-forme gratuite et libre soutenue par la Fondation Mozilla, où les utilisateurs peuvent créer et partager des jeux de données texte et vocaux à des fins multiples et sous différentes formats, y compris des enregistrements vocaux scénarisés et spontanés. Les jeux de

données sont accompagnés de données démographiques auto-déclarées telles que l'accent, l'âge et le genre, ou encore le nombre de votes positifs et négatifs apportés par la communauté mesurant la qualité des échantillons. Pour ce travail nous avons filtré et sélectionné un échantillon de 576 voix uniques, en équilibrant les accents et les genres auto-déclarés femme-homme. Nous n'avons conservé que des enregistrements ayant reçu des votes positifs, associés à un texte de référence d'au moins trois lettres, et dont la durée était comprise entre 3 et 5 secondes. L'âge n'a pas été un facteur discriminant. Nous avons en tout 24 enregistrements uniques pour chaque combinaison de genre et d'accent, avec les 12 accents anglais suivants : australien, écossais, états-unien, indien et sud-asiatique (Inde, Pakistan, Sri Lanka), des Philippines, canadien, de Hong Kong, de Malaisie, sud-africain (Afrique du Sud, Zimbabwe, Namibie), néo-zélandais, d'Irlande, ou encore d'Angleterre.

3 Résultats et discussion

La Table 2 présente les métriques WER et M-WER macro calculées sur l'ensemble des corpus. Nous pouvons constater l'importance d'inclure une métrique de taux d'erreur sur les termes spécialisés, le M-WER, qui permet de déterminer les meilleurs TTS en fonction de leur performance sur le vocabulaire spécialisé. Le tableau présente les résultats pour les scénarios de voix unique et multiple. Dans le cas de voix unique, nous voyons que les meilleurs modèles en termes de WER ne sont pas les mêmes que pour M-WER. Kokoro et Style-TTS2 présentent des comportements similaires qui pourraient s'expliquer par le fait qu'ils partagent la même architecture. Comme nous le verrons par la suite dans l'analyse des résultats BLURB, le score M-WER semble avoir une corrélation avec les performances du modèle BERT dans les tâches de TAL. En ce qui concerne les scores multi-voix, le classement des modèles reste le même (à l'exception de Kokoro qui est exclu de cette phase) et l'écart-type montre une infime différence des résultats en faisant varier l'accent du locuteur. Nous pouvons donc conclure que les accents évalués sont bien supportés par les modèles et qu'ils n'ont pas un impact significatif dans l'intelligibilité de l'audio généré.

	Mono-voix		**Multi-voix**	
TTS	**WER**	**M-WER**	**WER** (σ)	**M-WER** (σ)
fish-speech	0,379	<u>0,228</u>	**0,287** (0,0085)	<u>0,249</u> (0,0095)
kokoro	**0,261**	0,239	–	–
style-tts2	<u>0,323</u>	0,302	<u>0,317</u> (0,0024)	0,264 (0,0056)
zipvoice	0,428	**0,183**	0,385 (0,0821)	**0,190** (0,0152)

TABLE 2 – Scores macro de WER et de M-WER obtenus par les différents systèmes TTS, en scénarios mono-voix et multi-voix. Le M-WER est calculé uniquement sur les corpus de NER. Pour le scénario multi-voix, l'écart-type σ indique la variabilité des scores entre accents. Les meilleurs scores, les valeurs les plus faibles, sont indiqués en **gras**, et les deuxièmes meilleurs sont <u>soulignés</u>.

Ensuite, l'analyse par groupe de corpus selon la tâche de TAL, Table 3, montre certaines différences au sein d'un même système TTS. Celles-ci peuvent s'expliquer par la complexité du contenu textuel et l'impact des termes médicaux de spécialité. En particulier nous pouvons observer que les performances de ZipVoice diminuent de manière drastique pour les corpus de classification de documents (DC) et de questions-réponses (Q/A). Pour le premier, certaines phrases courtes mais complexes ont un impact fort sur le score WER, telles que : `The 3-UTRs of p27 and connexin-43 were`

PCR amplified from genomic DNA and cloned into pGL3 (Promega). Pour le deuxième, le résultat pourrait s'expliquer en partie par la longueur des textes à traiter (contexte de chaque question).

	Mono-voix				Multi-voix			
TTS	**NER**	**SS**	**DC**	**Q/A**	**NER**	**SS**	**DC**	**Q/A**
fish-speech	0,440	0,202	0,316	**0,221**	0,302	**0,219**	**0,343**	**0,249**
kokoro	**0,251**	**0,174**	**0,315**	0,256	–	–	–	–
style-tts2	0,305	0,276	0,365	0,355	0,297	0,276	0,362	0,296
zipvoice	0,257	0,202	0,796	0,748	**0,267**	0,229	0,454	0,382

TABLE 3 – Scores macro de WER obtenus par les différents systèmes TTS sur les groupes de tâches BLURB, en scénarios mono-voix et multi-voix.

Nous observons aussi qu'il ne semble pas y avoir de différence significative entre les résultats du scénario mono et multi-voix, impliquant que les TTS restent stables dans la génération de parole pour les accents testés.

En ce qui concerne les performances du modèle BioLinkBERT évalué sur les transcriptions des données synthétiques, nous pouvons voir dans la Table 4 une baisse des scores pour toutes les tâches, notamment pour la reconnaissance d'entités, signifiants que la perte d'information à la fin du pipeline à un impact non négligeable sur un modèle spécialisé. Les performances de ZipVoice en termes de M-WER semblent se reflèter dans les bons résultats pour la tâche BLURB de NER.

Malgré sa taille réduite, Kokoro obtient les résultats les plus élevés en moyenne pour toutes les tâches, ce qui pourrait être expliqué par les bonnes performances en taux d'erreur de mots (WER) et ce en dépit du taux M-WER relativement faible. Nous observons donc que la parte d'information nettement clinique ne jouerait pas un rôle déterminant dans l'exécution des tâches BLURB par BioLinkBERT, ce qui soulève un problème important dans l'évaluation de modèles spécialisés dans la santé : les tâches de TAL sont suffisamment générales au point qu'un modèle peut obtenir des bons résultats même quand l'information clinique du texte traité n'est pas exacte. Une observation similaire concerne le très mauvais WER de ZipVoice sur le corpus de Q/A qui contraste avec le très bon résultat dans la tâche respective ; le WER montre alors une perte importante d'information qui ne se reflète pas dans la capacité du modèle BioLinkBERT à compléter la tâche avec succès.

4 Perspectives

Nous travaillons actuellement sur l'évaluation des données synthétiques par des professionnels de la santé, afin de classer les systèmes TTS au moyen d'un score d'optinion moyen « Mean Opinion Score » (MOS). L'objectif étant de mesurer la préservation du contenu médical par des professionnels anglophones, nous prévoyons d'extraire un échantillon de 20 à 25 phrases par système TTS avec des WER et M-WER variables, qui seront évaluées par des experts du domaine. Un échantillon de contrôle de 10% d'enregistrements humains sera inclus. Afin de diminuer le biais introduit par la variation dans les accents, nous nous restreindrons à un petit nombre d'accents de haute qualité (e.g. Angleterre, États-Unis). Les questionnaires seront présentés sans le texte de référence et incluront une partie de compréhension orale qui permettra d'évaluer si une question médicale peut être répondue uniquement

	NER			SS	DC			Q/A
Système	**Préc.**	**Rappel**	**F1**	**Pearson**	**Préc.**	**Rappel**	**F1**	**Acc.**
baseline	82,79	89,34	85,76	96,85	79,93	90,50	84,88	81,44
fish-speech	65,31	67,62	66,31	73,65	79,31	85,54	82,31	**79,84**
kokoro	73,98	61,82	66,95	**93,54**	79,74	90,29	**84,69**	77,94
style-tts2	57,18	59,26	58,11	86,99	82,74	81,20	81,96	78,23
zipvoice	66,39	69,21	**67,65**	82,36	78,56	85,54	81,90	78,61

TABLE 4 – Performances du modèle BioLinkBERT-large sur les tâches BLURB à partir de données synthétiques générées en scénario mono-voix. Les scores sont reportés pour la reconnaissance d'entités nommées (NER), la similarité de phrases (SS), la classification de documents (DC) et les questions-réponses (Q/A).

avec le contenu audio. Ils incluront également une évaluation subjective, sur une échelle de 1 à 5, visant à déterminer dans quelle mesure les voix sont perçues comme proches d'une voix humaine et à quantifier l'effort requis par les experts pour comprendre le texte énoncé (e.g. les difficultés de locution ou d'intelligibilité).

À l'issue de ces travaux, les données seront mises à disposition de la communauté sous une licence ouverte permettant à d'autres équipes de les exploiter pour l'évaluation de modèles audio. Nous allons continuer nos travaux dans l'évaluation de modèles de langue vocaux pour la santé. Les conclusions de nos expériences nous permettront de choisir le TTS le plus approprié pour réaliser la synthèse d'un grand corpus textuel médical et l'exploiter sur plusieurs tâches, via l'affinage de grands modèles de langue généralistes et spécialisés, avec et sans modalité audio pré-entraînée.

Les observations de cette étude se limitent à une sélection de quatre TTS en langue anglaise et les scores sont impactés par le biais du système ASR choisi. D'autres évaluations sur des modèles plus grands et ayant une variété plus importante en termes d'architecture et de données d'entraînement serait nécessaire pour généraliser les conclusions. Nous espérons que notre approche éclaire les défis auxquels font face les systèmes TTS ouverts pour produire de l'audio de qualité dans un domaine de spécialité, ouvrant la voie à des développements futurs de systèmes plus robustes pour le domaine de la santé.

Les limitations de nos évaluations basées sur la modalité textuelle pourraient être réduites en recourant à des méthodes utilisant directement la modalité audio, ce qui éliminerait la dépendance à un système ASR. Il est ansi envisageable de recourir à de grands modèles de langue en tant que juges (« LLM-as-a-Judge ») pour évaluer la qualité de plus petits modèles. D'autres approches sémantiques pourraient inclure la projection du texte de référence et de l'audio généré vers le même espace de représentation pour ensuite évaluer leur similarité.

Références

ADEDEJI A., JOSHI S. & DOOHAN B. (2024). The sound of healthcare : Improving medical transcription ASR accuracy with large language models. *arXiv preprint arXiv :2402.07658*. DOI : 10.48550/arXiv.2402.07658.

AFONJA T., OLATUNJI T., OGUN S., ETORI N. A., OWODUNNI A. & YEKINI M. (2024). Performant ASR models for medical entities in accented speech. *arXiv preprint arXiv :2406.12387*. DOI : 10.48550/arXiv.2406.12387.

ARDILA R., BRANSON M., DAVIS K., KOHLER M., MEYER J., HENRETTY M., MORAIS R., SAUNDERS L., TYERS F. & WEBER G. (2020). Common Voice : A massively-multilingual speech corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4218–4222, Marseille, France : European Language Resources Association.

BAKER S., ALI I., SILINS I., PYYSALO S., GUO Y., HÖGBERG J., STENIUS U. & KORHONEN A. (2017). Cancer Hallmarks Analytics Tool (CHAT) : a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics*, **33**(24), 3973–3981. DOI : 10.1093/bioinformatics/btx454.

BAKER S. & KORHONEN A. (2017). Initializing neural networks for hierarchical multi-label text classification. In K. B. COHEN, D. DEMNER-FUSHMAN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 16th BioNLP Workshop*, p. 307–315, Vancouver, Canada, : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2339.

BAKER S., KORHONEN A. & PYYSALO S. (2016a). Cancer hallmark text classification using convolutional neural networks. In S. ANANIADOU, R. BATISTA-NAVARRO, K. B. COHEN, D. DEMNER-FUSHMAN & P. THOMPSON, Éds., *Proceedings of the fifth workshop on building and evaluating resources for biomedical text mining (BioTxtM2016)*, p. 1–9, Osaka, Japan : The COLING 2016 Organizing Committee.

BAKER S., SILINS I., GUO Y., ALI I., HÖGBERG J., STENIUS U. & KORHONEN A. (2016b). Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, **32**(3), 432–440. DOI : 10.1093/bioinformatics/btv585.

CHUA C. E., CLARA N. L. Y., FURQAN M. S., KIT J. L. W., MAKMUR A., THAM Y. C., SANTOSA A. & NGIAM K. Y. (2024). Integration of customised LLM for discharge summary generation in real-world clinical settings : a pilot study on RUSSELL GPT. *The Lancet Regional Health–Western Pacific*, **51**, 101211. DOI : 10.1016/j.lanwpc.2024.101211.

COLLIER N., OHTA T., TSURUOKA Y., TATEISI Y. & KIM J.-D. (2004). Introduction to the bio-entity recognition task at JNLPBA. In N. COLLIER, P. RUCH & A. NAZARENKO, Éds., *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, p. 73–78, Geneva, Switzerland : COLING.

COOPER E., HUANG W.-C., TSAO Y., WANG H.-M., TODA T. & YAMAGISHI J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, **45**(4), 161–183. DOI : 10.1250/ast.e24.12.

DOĞAN R. I., LEAMAN R. & LU Z. (2014). NCBI disease corpus : a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, **47**, 1–10. DOI : 10.1016/j.jbi.2013.12.006.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. DOI : 10.1145/3458754.

HEXGRAD (2025). Kokoro-82m (revision d8b4fc7). DOI : 10.57967/hf/4329.

JIN Q., DHINGRA B., LIU Z., COHEN W. & LU X. (2019). PubMedQA : A dataset for biomedical research question answering. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), p. 2567–2577, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1259.

JOHNSON A. E. W., BULGARELLI L., SHEN L., GAYLES A., SHAMMOUT A., HORNG S., POLLARD T. J., HAO S., MOODY B., GOW B., LEHMAN L.-W. H., CELI L. A. & MARK R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, **10**(1), 1. DOI : 10.1038/s41597-022-01899-x.

KANEKO T., TANAKA K., KAMEOKA H. & SEKI S. (2022). iSTFTNet : Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6207–6211 : IEEE. DOI : 10.1109/ICASSP43922.2022.9746713.

LI J., SUN Y., JOHNSON R. J., SCIAKY D., WEI C.-H., LEAMAN R., DAVIS A. P., MATTINGLY C. J., WIEGERS T. C. & LU Z. (2016). BioCreative V CDR task corpus : a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, **2016**. DOI : 10.1093/database/baw068.

LI Y. A., HAN C., RAGHAVAN V. S., MISCHLER G. & MESGARANI N. (2023). StyleTTS 2 : Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models.

LIAO S., WANG Y., LI T., CHENG Y., ZHANG R., ZHOU R. & XING Y. (2024). Fish-speech : Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv :2411.01156*. DOI : 10.48550/arXiv.2411.01156.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, p. 28492–28518 : PMLR. DOI : 10.48550/arXiv.2212.04356.

ROUX T. B., ROUVIER M., WOTTAWA J. & DUFOUR R. (2022). Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In *Interspeech*, Incheon, South Korea. HAL : hal-03712735.

SMITH L., TANABE L. K., ANDO R. J. N., KUO C.-J., CHUNG I.-F., HSU C.-N., LIN Y.-S., KLINGER R., FRIEDRICH C. M., GANCHEV K., TORII M., LIU H., HADDOW B., STRUBLE C. A., POVINELLI R. J., VLACHOS A., BAUMGARTNER W. A., HUNTER L., CARPENTER B., TSAI R. T.-H., DAI H.-J., LIU F., CHEN Y., SUN C., KATRENKO S., ADRIAANS P., BLASCHKE C., TORRES R., NEVES M., NAKOV P., DIVOLI A., MAÑA-LÓPEZ M., MATA J. & WILBUR W. J. (2008). Overview of BioCreative II gene mention recognition. *Genome Biology*, **9**(2), S2. DOI : 10.1186/gb-2008-9-s2-s2.

SOĞANCIOĞLU G., ÖZTÜRK H. & ÖZGÜR A. (2017). BIOSSES : a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, **33**(14), i49–i58. DOI : 10.1093/bioinformatics/btx238.

TSATSARONIS G., SCHROEDER M., PALIOURAS G., ALMIRANTIS Y., ANDROUTSOPOULOS I., GAUSSIER E., GALLINARI P., ARTIERES T., ALVERS M. R., ZSCHUNKE M. *et al.* (2012). BioASQ : A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium : Information retrieval and knowledge discovery in biomedical text* : Arlington, VA : Citeseer.

VASILEV Y., RAZNITSYNA I., PAMOVA A., BURTSEV T., BOBROVSKAYA T., KOSOV P., VLADZYMYRSKYY A., OMELYANSKAYA O. & ARZAMASOV K. (2026). Evaluating medical text summaries using automatic evaluation metrics and LLM-as-a-judge approach : A pilot study. *Diagnostics*, **16**(1). DOI : 10.3390/diagnostics16010003.

WANG X., CHEN N., CHEN J., HU Y., WANG Y., WU X., GAO A., WAN X., LI H. & WANG B. (2024). Apollo : Lightweight multilingual medical LLMs towards democratizing medical ai to 6b people. DOI : 10.48550/arXiv.2403.03640.

YASUNAGA M., LESKOVEC J. & LIANG P. (2022). LinkBERT : Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*. DOI : 10.18653/v1/2022.acl-long.551.

ZEN H., DANG V., CLARK R., ZHANG Y., WEISS R. J., JIA Y., CHEN Z. & WU Y. (2019). LibriTTS : A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv :1904.02882*. DOI : 10.48550/arXiv.1904.02882.

ZHU H., KANG W., YAO Z., GUO L., KUANG F., LI Z., ZHUANG W., LIN L. & POVEY D. (2025). ZipVoice : Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv :2506.13053*. DOI : 10.48550/arXiv.2506.13053.