Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Un paradigme pour l'interprétation des métriques et pour mesurer la gravité des erreurs de reconnaissance automatique de la parole

Thibault Bañeras-Roux[1] Michael Rouvier[2] Jane Wottawa[3] Richard Dufour[1]
(1) Laboratoire des Sciences du Numérique de Nantes (LS2N), France
(2) Laboratoire Informatique d'Avignon (LIA), France
(3) Laboratoire d'Informatique de l'Université du Mans (LIUM), France
thibault.roux@univ-nantes.fr, jane.wottawa@univ-lemans.fr,
michael.rouvier@univ-avignon.fr, richard.dufour@univ-nantes.fr

RÉSUMÉ

Les mesures couramment employées pour l'évaluation des transcriptions automatiques de la parole, telles que le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER), ont fait l'objet d'importantes critiques en raison de leur corrélation limitée avec la perception humaine et de leur incapacité à prendre en compte les nuances linguistiques et sémantiques. Bien que des métriques fondées sur les plongements sémantiques aient été introduites pour se rapprocher de la perception humaine, leur interprétabilité reste difficile par rapport au WER et CER. Dans cet article, nous surmontons ce problème en introduisant un paradigme qui intègre une métrique choisie pour obtenir un équivalent du taux d'erreur appelé Distance d'Édition Minimale, ou Minimum Edit Distance (minED). Nous proposons également d'utiliser cette approche pour mesurer la gravité des erreurs en fonction d'une métrique, d'un point de vue intrinsèque et extrinsèque.

ABSTRACT

A Paradigm for Interpreting Metrics and Measuring Error Severity in Automatic Speech Recognition

The commonly employed metrics for the evaluation of automatic speech transcriptions, such as Word Error Rate (WER) and Character Error Rate (CER), have faced significant criticism due to their limited correlation with human perception and their inability to account for linguistic and semantic nuances. While metric-based embeddings have been introduced to approximate human perception, their interpretability remains challenging compared to WER and CER. In this article, we overcome this problem by introducing a paradigm that integrates a chosen metric to obtain an equivalent of the error rate called Minimum Edit Distance (minED). We also propose to use this approach to measure the severity of errors according to a metric, from an intrinsic and extrinsic perspective.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Métriques d'évaluation, Interprétabilité, erreurs de transcriptions.

KEYWORDS: Automatic speech recognition, Evaluation metric, Interpretability, Transcription errors..

Actes des 35èmes Journées d'Études sur la Parole, pages 191–199.

1 Introduction

Malgré les progrès considérables réalisés dans le domaine de l'apprentissage automatique et l'utilisation intensive de données pour l'entraînement des modèles, les systèmes de Reconnaissance Automatique de la Parole (RAP) présentent encore des erreurs de transcription dans des proportions variables en fonction de leurs conditions d'utilisation.

L'évaluation d'un système de RAP consiste le plus souvent à comparer les transcriptions manuelles (référence) et automatiques (hypothèse) à l'aide d'une mesure choisie, généralement le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER). Tous deux consistent à calculer une distance de Levenshtein entre la référence et l'hypothèse. Cependant, ces métriques sont critiquées pour attribuer le même poids à toutes les erreurs tout en négligeant les nuances linguistiques et sémantiques (Favre *et al.*, 2013; Ruiz & Federico, 2015; Kafle & Huenerfauth, 2017; Gordeeva *et al.*, 2021).

Pour remédier à ces limitations, des mesures fondées sur les plongements (Zhang *et al.*, 2020; Kim *et al.*, 2021; Bañeras-Roux *et al.*, 2022) ont été proposées pour intégrer les aspects sémantiques.

De même, d'un point de vue perceptif, Kafle & Huenerfauth; Kim *et al.*; Gordeeva *et al.*; Bañeras-Roux *et al.* ont utilisé des ensembles de données annotées pour évaluer rigoureusement l'alignement des métriques de reconnaissance vocale avec la perception humaine, révélant la corrélation supérieure des métriques sémantiques avec le jugement humain.

Si les mesures sémantiques offrent une perspective d'évaluation différente, leurs scores, calculés par similarité cosinus, manquent d'interprétabilité, contrairement au WER qui s'appuie simplement sur les mots. Dans cet article, nous proposons d'intégrer une métrique dans un nouveau paradigme, appelé Distance d'Édition Minimale, ou en anglais Minimum Edit Distance (minED), afin de rendre interprétables les scores des métriques basés sur les plongements. Ce paradigme est également appliqué pour mesurer la gravité des erreurs, ce qui peut être utilisé pour l'analyse des métriques.

Le document est organisé comme suit. La section 2 présente les métriques de RAP et un ensemble de données avec des annotations de perception humaine. La section 3 décrit le paradigme minED proposé pour l'interprétabilité des mesures, tandis que la section 4 examine la capacité du paradigme à mesurer la gravité des erreurs. Enfin, nous concluons le travail et donnons des perspectives dans la section 5.

2 Méthodologie

Dans la section 2.1, nous fournissons des détails sur les métriques de RAP utilisées dans cette étude. Ensuite, dans la section 2.2, nous présentons le jeu de données HATS, utilisé pour évaluer les métriques ainsi que le paradigme proposé.

2.1 Métriques

Comme indiqué précédemment, la communauté a développé diverses métriques s'appuyant sur les plongements. En utilisant BERT (Devlin *et al.*, 2019), nous pouvons extraire des représentations sémantiques des phrases. L'une de ces mesures, **SemDist** (Kim *et al.*, 2021) calcule la similarité cosinus entre la référence et l'hypothèse à l'aide des plongements obtenus au niveau de la phrase.

Une autre mesure, **BERTScore** (Zhang *et al.*, 2020), appliquée dans diverses tâches de traitement automatique du langage (TAL) (Yilmaz *et al.*, 2019; Hanna & Bojar, 2021), calcule un score de similarité pour chaque token de la phrase candidate avec chaque token de la phrase de référence à l'aide de plongements contextuels.

Dans cette étude, SemDist intègre la version Sentence-BERT (Reimers & Gurevych, 2019) de CamemBERT (Martin *et al.*, 2020) [1], une version française de BERT, et BERTScore utilise un BERT (Devlin *et al.*, 2019) multilingue. Pour nos expériences, nous avons normalisé toutes les mesures sur une échelle de [0, 1] en appliquant la règle "le plus faible le meilleur".

2.2 Jeu de données HATS

L'un des moyens d'évaluer correctement les métriques consiste à utiliser un ensemble de données d'annotations humaines. L'ensemble de données HATS est un corpus en libre accès, pour le français, conçu pour évaluer la corrélation entre les mesures d'évaluation de RAP et la perception humaine du point de vue du lecteur. L'ensemble de données HATS a été développé à l'aide d'une expérience côte-à-côte (Gordeeva *et al.*, 2021; Kafle & Huenerfauth, 2017; Kim *et al.*, 2022). Une référence textuelle, ainsi que deux hypothèses erronées produites par des systèmes de RAP (8 systèmes de bout-en-bout (Ravanelli *et al.*, 2021) et deux systèmes basés sur une architecture DNN-HMM [2] (Povey *et al.*, 2011)), ont été présentées à au moins 7 sujets qui ont sélectionné la meilleure hypothèse. L'ensemble des données comprend 1 000 triplets : une référence, chacune accompagnée de deux hypothèses et du nombre de votes associé.

En calculant le nombre de fois qu'une métrique est en accord avec les annotations humaines (la métrique indique le meilleur score pour l'hypothèse choisie par les humains), nous pouvons calculer un ratio correspondant à la corrélation entre cette métrique et l'évaluation humaine.

La métrique SemDist obtient la corrélation la plus forte avec la perception humaine sur HATS. L'ensemble des données sera utilisé pour déterminer si l'utilisation du paradigme minED entraîne une réduction de la corrélation avec la perception humaine par rapport à l'utilisation de la métrique seule.

3 Intégrer des métriques pour l'interprétabilité

Le paradigme minED est conçu pour améliorer l'interprétabilité des mesures produisant des scores difficiles à comprendre. Pour ce faire, nous intégrons une métrique non interprétable telle que SemDist dans minED. Cela consiste à calculer le nombre minimum de modifications à appliquer à l'hypothèse pour qu'elle soit suffisamment proche de la référence en ce qui concerne sa perception humaine. Suivant cette idée, nous appliquons cette méthode aux mots (minWED) et aux caractères (minCED).

Le paradigme est décrit dans la section 3.1, tandis que la section 3.2 traite du paramétrage de la méthode. Nous discutons ensuite de deux types de mesures (cohérentes, incohérentes) influençant le coût de calcul (section 3.3), et explorons la corrélation entre minED et la perception humaine (section 3.4).

1. https://huggingface.co/dangvantuan/sentence-camembert-large
2. https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/

FIGURE 1 – Graphe de chaque modification possible pour obtenir une hypothèse sans erreur avec le paradigme minWED. Chaque arête correspond à une erreur corrigée. Étant donné la référence, nous avons trois erreurs de mots, chacune d'un type différent : 1 substitution, 1 insertion, 1 suppression. La métrique est basée sur la règle "le plus faible le meilleur". Le symbole ϵ correspond à une suppression.

3.1 Distance d'Édition Minimum (minED)

La correction de mots, ou de caractères, consiste à éditer l'hypothèse afin qu'il n'y ait plus de substitutions, d'insertions ou de suppressions. Le paradigme minED calcule le nombre minimum de corrections (mots ou caractères) nécessaires pour rendre une hypothèse "acceptable" sur la base d'une métrique non interprétable. Pour ce faire, nous générons un graphique qui représente toutes les modifications qui peuvent être apportées à l'hypothèse pour qu'elle devienne la référence (voir figure 1). Pour chaque élément corrigé, nous calculons un score entre la référence et la nouvelle hypothèse à l'aide de la métrique incorporée. Si le score est inférieur à un seuil prédéfini, l'hypothèse est jugée "acceptable". Il n'est donc pas nécessaire de calculer le reste du graphique. Le score minED est le nombre de niveau minimum à calculer pour obtenir un score en dessous du seuil.

La définition du seuil est cruciale, et la section 3.2 détaille la manière de l'établir.

La Figure 1 présente le graphe des possibilités pour la référence « *I will book them an appointment* » et l'hypothèse « *will book them an appointment and* ». Dans ce scénario, nous avons trois erreurs : une suppression, une substitution et une insertion. L'erreur de suppression est représentée par un symbole ϵ.

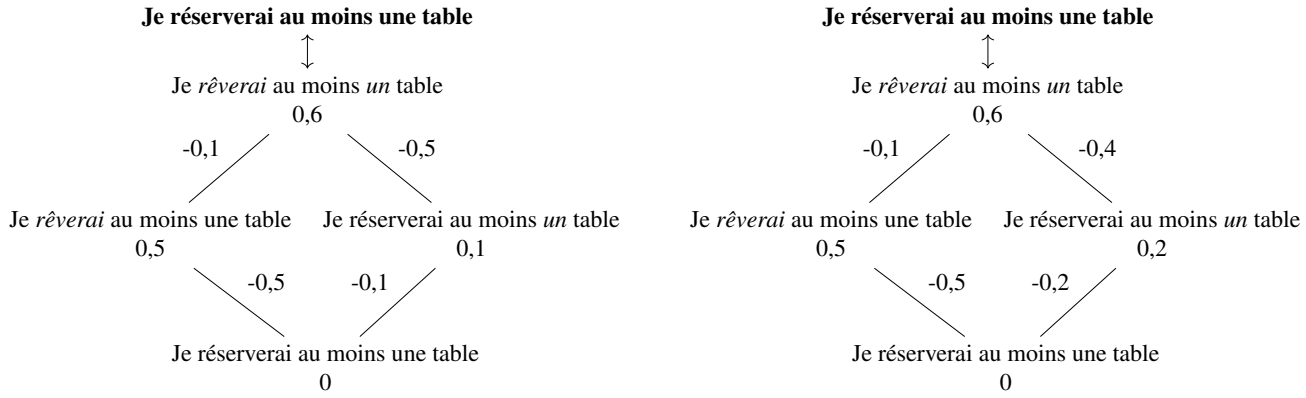
3.2 Fixation du seuil d'acceptabilité

Comme indiqué dans la section 3.1, minED signifie les modifications nécessaires pour une hypothèse acceptable. Ce concept repose sur le fait qu'une métrique puisse donner un score considérée comme acceptable par les humains. Par exemple, lorsqu'un humain lit une hypothèse erronée, si une métrique sémantique indique un score inférieur au seuil (dans un contexte où la valeur la plus basse est la meilleure), le sens de la phrase originale est censé être compris.

Lorsque le seuil est trop bas, les mesures minWED et minCED tendent à se rapprocher des valeurs WER ou CER. Inversement, des valeurs de seuil trop élevées font converger ces mesures vers des

scores nuls, ce qui signifie qu'aucune correction n'est nécessaire.

Une approche pourrait consister à sélectionner un seuil qui maximise la corrélation avec la perception humaine, mais il n'est pas exclu de réfléchir à d'autres méthodes.



(a) Métrique cohérente.

(b) Métrique incohérente.

FIGURE 2 – Comparaison de l'impact des corrections sur des métriques cohérentes et incohérentes. Les métriques respectent une règle de "le plus faible le meilleur".

3.3 Cohérence des métriques

Lors de la correction d'une hypothèse pour la rapprocher de la référence, nous pouvons observer une amélioration du score selon la métrique incorporée.

La correction peut avoir deux effets connus : soit elle améliore le score indépendamment des modifications précédentes (voir figure 2a), soit elle améliore le score en fonction des modifications précédentes (voir figure 2b). Par exemple, dans la figure 2a, la correction de la substitution *réserverai*/*rêverai* améliorera la performance de la métrique de *0,5*, que *une*/*une* ait été corrigé ou non. Dans la figure 2b, la correction de *réserverai*/*rêverai* améliorera la performance de la métrique de *0,5* ou *0,4*, selon que *une*/*un* ait été corrigé ou non.

La propriété de cohérence permet de calculer plus rapidement le nombre minimum de modifications, car il n'est plus nécessaire de calculer l'ensemble du graphique. Une approche pratique consiste plutôt à calculer le deuxième niveau où une seule erreur dans l'hypothèse est corrigée. Ensuite, on soustrait le score de l'hypothèse initiale du nombre minimum d'améliorations rédactionnelles requises pour que le score obtenu soit inférieur au seuil.

WER et CER sont des exemples de métriques cohérentes, tandis que BERTScore et SemDist sont des exemples de métriques incohérentes.

3.4 Corrélation avec la perception humaine

La figure 3 montre la corrélation entre la perception humaine et minED pour différentes valeurs de seuil (θ). Les seuils les plus bas donnent des corrélations plus proches de la métrique associée à l'édition (WER ou CER), tandis que les valeurs trop élevées entraînent une baisse des performances.

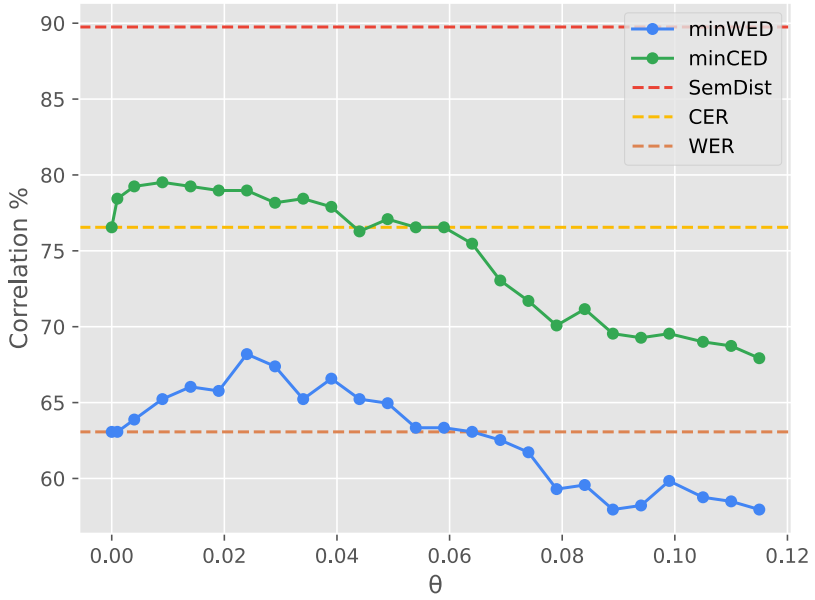


FIGURE 3 – Corrélation de MinED avec le jeu de données HATS en fonction de différentes valeurs de seuil (θ).

Alors que minWED gagne 5,12 % par rapport au WER et améliore l'interprétabilité par rapport à SemDist, il perd 21,56 % de corrélation par rapport à SemDist. De même, minCED est mieux corrélé avec la perception humaine que CER, mais présente une proportion significative de perte par rapport à SemDist. Ces résultats montrent les limites de l'utilisation des taux d'erreur pour évaluer les transcriptions de RAP d'un point de vue humain.

4 Mesurer la gravité des erreurs

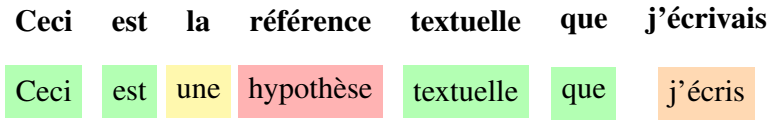


FIGURE 4 – Visualisation de la gravité des erreurs selon notre paradigme MinED intégrant une métrique sémantique.

Dans cette section, nous étudions la capacité de notre paradigme à identifier les erreurs et à mesurer leur gravité, comme l'illustre la figure 4. La section 4.1 présente notre méthode d'évaluation de la gravité des erreurs, tandis que la section 4.2 se penche sur les résultats et l'analyse.

4.1 Protocole d'évaluation

Pour évaluer correctement la capacité de notre paradigme à identifier les erreurs critiques, nous partons du principe que la correction d'une erreur grave devrait avoir un impact plus important sur une tâche en aval que la correction d'une erreur mineure.

Dans notre étude, nous avons choisi une tâche de traduction du français vers l'anglais à partir de données vocales. Cette tâche comprend d'abord une transcription automatique, considérée comme l'évaluation intrinsèque des sorties de RAP à l'aide des métriques SemDist et CER. La transcription résultante est ensuite transmise à un traducteur automatique pour générer l'hypothèse finale, qui est considérée comme l'évaluation extrinsèque du système de reconnaissance de la parole à l'aide des métriques BLEU et BERTScore.

	Transcriptions	**Traductions**	**SemDist**	**BERTScore**
Référence	à nos résultats	to our results		
Hypothèses	un non résultat	a no result	57,8	28,1
Hypothèses Corrigées	à non résultat	to no result	50,1 (+7,7)	23,6 (+4,5)
	un nos résultat	a our result	20,2 (+37,6)	21,4 (+6,7)
	un non résultats	a no results	52,7 (+5,1)	28,0 (+0,1)

TABLE 1 – Exemple des améliorations de SemDist et de BERTScore obtenues par la correction de l'hypothèse « *à nos résultats* ». Les scores sont projetés dans une règle "le plus faible le meilleur" et une échelle de [0, 100] pour une meilleure lisibilité.

Comme le montre le tableau 1, nous générons autant de corrections à une hypothèse erronée qu'il y a d'erreurs de transcription. Cette approche nous permet d'obtenir, pour chaque correction, le score d'amélioration pour nos mesures intrinsèques et extrinsèques. Si nous observons une corrélation entre ces deux valeurs, cela signifie que le paradigme est effectivement capable de mesurer la sévérité des erreurs.

Notre dispositif expérimental utilise le jeu de données HATS pour obtenir les références et les hypothèses erronées associées, les traductions étant générées à l'aide de Google Traduction.

4.2 Résultats et analyse

Le tableau 2 présente la corrélation de Spearman entre l'amélioration intrinsèque et extrinsèque de la transcription automatique pour la tâche de traduction. Une corrélation notable entre SemDist et BERTScore est observée, démontrant la capacité du paradigme à mesurer la gravité des erreurs. Des corrélations différentes apparaissent pour les mesures intrinsèques et extrinsèques, suggérant des variations potentielles dans les résultats pour des tâches autres que la traduction.

Intrins./Extrins.	**BERTScore**	**BLEU**
SemDist	0,39	0,26
CER	0,22	0,23

TABLE 2 – Moyenne de la corrélation de Spearman entre les améliorations intrinsèques et extrinsèques en fonction de différentes métriques pour la tâche de traduction.

5 Conclusions et perspectives

Nous avons proposé un paradigme qui non seulement rend les mesures de RAP interprétables, mais qui permet également de mesurer la gravité des erreurs. L'approche minED fournit un cadre plus transparent pour l'évaluation des systèmes de RAP. Alors que notre étude a révélé une diminution notable de la corrélation avec la perception humaine lors de l'intégration d'une métrique dans minWED (*i.e.* sur les mots), nos résultats démontrent que minCED (sur les caractères) maintient une performance relativement forte dans la capture de la perception de l'erreur comparé à d'autres métriques évalué par Bañeras-Roux *et al.*.

L'étude montre également, par la perte significative de corrélation avec l'interprétabilité, qu'une mesure du nombre d'erreurs ne correspond pas à la manière dont les humains se comportent. Il semble que les humains donnent la priorité à la prise en compte de la gravité des erreurs plutôt qu'à la simple proportion d'erreurs graves.

Une autre stratégie pour développer des mesures interprétables plus étroitement liées à la perception humaine consisterait à développer des métriques qualitatives plutôt que quantitatives. Par exemple, des ensembles de données comme HypRatings (Kim *et al.*, 2022) intègrent des annotations qualitatives telles que "`exact`", "`hyp utile`", "`hyp fausse`", et "`hyp incohérente`". L'étude du développement de métriques prédisant ces caractéristiques qualitatives pourrait constituer une perspective intéressante pour de futures recherches.

6 Considérations éthiques

Lors de la mise en production d'un système de reconnaissance de la parole, si minED est utilisé pour évaluer le système, il convient d'être prudent dans le choix du seuil : l'acceptabilité des erreurs est subjective et peut ne pas s'appliquer à tous les humains.

Tous les modèles et données utilisés dans cet article sont publics et librement accessibles à des fins de reproductibilité. Notre code est disponible sur un dépôt GitHub public [3].

Références

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2022). Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In *Interspeech 2022*.

BAÑERAS-ROUX T., WOTTAWA J., ROUVIER M., MERLIN T. & DUFOUR R. (2023). Hats : An open data set integrating human perception applied to the evaluation of automatic speech recognition metrics. In *Text, Speech and Dialogue 2023 - Interspeech Satellite*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

3. https://anonymous.4open.science/r/mined

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of ASR systems : Does WER really predict performance ? In *INTERSPEECH*, p. 3463–3467.

GORDEEVA L., ERSHOV V., GULYAEV O. & KURALENOK I. (2021). Meaning Error Rate : ASR domain-specific metric framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, p. 458–466.

HANNA M. & BOJAR O. (2021). A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, p. 507–517.

KAFLE S. & HUENERFAUTH M. (2017). Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 165–174.

KIM S., ARORA A., LE D., YEH C.-F., FUEGEN C., KALINLI O. & SELTZER M. L. (2021). Semantic Distance : A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In *Proc. Interspeech 2021*, p. 1977–1981. DOI : 10.21437/Interspeech.2021-1929.

KIM S., LE D., ZHENG W., SINGH T., ARORA A., ZHAI X., FUEGEN C., KALINLI O. & SELTZER M. (2022). Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In *Proc. Interspeech 2022*, p. 3978–3982. DOI : 10.21437/Interspeech.2022-11144.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, volume CONF : IEEE Signal Processing Society.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In *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. 3982–3992.

RUIZ N. & FEDERICO M. (2015). Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, p. 296–302 : IEEE.

YILMAZ Z. A., WANG S., YANG W., ZHANG H. & LIN J. (2019). Applying bert to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) : System Demonstrations*, p. 19–24.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.