HATS : Un jeu de données intégrant la perception humaine appliquée à l'évaluation des métriques de transcription de la parole

Thibault Bañeras-Roux, Jane Wottawa, Mickael Rouvier, Teva Merlin, Richard Dufour

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

HATS : Un jeu de données intégrant la perception humaine appliquée à l'évaluation des métriques de transcription de la parole

Thibault Bañeras-Roux[1,2] Jane Wottawa[3] Michael Rouvier[2] Teva Merlin[2]
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, teva.merlin@univ-avignon.fr,
richard.dufour@univ-nantes.fr

RÉSUMÉ

Traditionnellement, les systèmes de reconnaissance automatique de la parole (RAP) sont évalués sur leur capacité à reconnaître correctement chaque mot contenu dans un signal vocal. Dans ce contexte, la mesure du taux d'erreur-mot est la référence pour évaluer les transcriptions vocales. Plusieurs études ont montré que cette mesure est trop limitée pour évaluer correctement un système de RAP, ce qui a conduit à la proposition d'autres variantes et d'autres métriques. Cependant, toutes ces métriques restent orientées "système" alors même que les transcriptions sont destinées à des humains. Dans cet article, nous proposons un jeu de données original annoté manuellement en termes de perception humaine des erreurs de transcription produites par divers systèmes de RAP. 143 humains ont été invités à choisir la meilleure transcription automatique entre deux hypothèses. Nous étudions la relation entre les préférences humaines et diverses mesures d'évaluation pour les systèmes de RAP, y compris les mesures lexicales et celles fondées sur les plongements de mots.

ABSTRACT

HATS : An open dataset integrating human perception applied to the evaluation of Automatic Speech Recognition metrics

Traditionally, Automatic Speech Recognition (ASR) systems are evaluated on their ability to correctly recognize each word contained in a speech signal. In this context, the Word Error Rate metric is the reference for evaluating speech transcripts. Several studies have shown that this measure is too limited to correctly evaluate an ASR system, which has led to the proposal of other variants and other metrics. However, all these metrics remain system-oriented, even when transcripts are intended for humans. In this paper, we describe an original manually annotated dataset in terms of human perception of transcription errors produced by various ASR systems. 143 humans were asked to choose the best automatic transcription between two hypotheses. We investigate the relationship between human preferences and various evaluation metrics for ASR systems, including lexical and embedding-based metrics.

MOTS-CLÉS : reconnaissance de la parole, jeu de données, perception, métrique, corpus.

KEYWORDS: speech recognition, dataset, perception, metric, corpus.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 10–18.

1 Introduction

La Reconnaissance Automatique de la Parole (RAP) consiste à transcrire de la parole en texte. Depuis l'utilisation des systèmes de RAP fondés sur les Modèles de Markov cachés (Juang & Rabiner, 1991), le domaine a connu d'importants progrès avec l'utilisation des réseaux de neurones profonds et des méthodes auto-supervisées telles que wav2vec (Baevski *et al.*, 2020) et HuBERT (Hsu *et al.*, 2021) qui permettent d'extraire des informations de la parole sans données étiquetées. Ces transcriptions automatiques peuvent être utilisées par les humains dans le cas par exemple du sous-titrage, de la rédaction de messages ou par des systèmes tiers tels que les assistants personnels virtuels.

Face à des erreurs dans un flux de parole ou des des textes, un humain est capable de les intégrer et éventuellement de les corriger si elles n'impactent pas fondamentalement le sens de la séquence (Cutler, 2012). Les erreurs dans les transcriptions automatiques proviennent de divers facteurs, tels que le bruit dans le signal vocal, les accents des locuteurs ou les limitations techniques. La question est de savoir quelles erreurs sont acceptables et lesquelles entraînent des difficultés de compréhension chez l'humain. Par conséquent, il apparaît souhaitable que les métriques d'évaluation des systèmes de RAP se rapprochent de la perception humaine.

Les métriques automatiques les plus couramment utilisées pour évaluer les systèmes de RAP sont le taux d'erreur-mot (WER pour *Word Error Rate* en anglais), qui mesure le nombre de mots incorrectement transcrits, et le taux d'erreurs-caractères (CER pour *Character Error Rate* en anglais) qui évalue le nombre de caractères qui diffèrent par rapport à la référence. Cependant, de nombreux travaux (Wang *et al.*, 2003; Favre *et al.*, 2013; Itoh *et al.*, 2015; Kafle & Huenerfauth, 2017) ont soulevé des problèmes liés à ces mesures tels que l'absence de pondération des erreurs ou encore le manque d'informations linguistiques et de connaissances sémantiques. En réponse à ces problèmes, le développement de nouvelles métriques a suscité un intérêt croissant dans la communauté. Des métriques alternatives se focalisant sur la qualité et l'efficacité des transcriptions automatiques ont été proposées (Nam & Fels, 2019; Gordeeva *et al.*, 2021; Kim *et al.*, 2021; Bañeras-Roux *et al.*, 2022).

Des évaluation humaines de systèmes de RAP ont par le passé été réalisées, notamment à l'aide d'une expérience *côte-à-côte* (Gordeeva *et al.*, 2021; Kafle & Huenerfauth, 2017; Kim *et al.*, 2022), consistant à demander à des sujets humains de choisir la meilleure transcription parmi deux proposées. Ces études ont également permis d'évaluer des métriques automatiques. La présente étude s'inspire de ce protocole expérimental mais, au lieu d'altérer le signal vocal ou d'utiliser différentes sorties du même système de RAP afin d'obtenir deux hypothèses différentes, notre étude met en compétition les sorties de 10 systèmes intégrant des architectures différentes. En plus, les hypothèses ont été appariées selon un ensemble de critères métriques bien définis.

Cette expérience nous permet de distribuer librement un nouveau jeu de données ouvert, nommé appelé HATS (Human Assessed Transcription Side-by-side), intégrant des préférences humaines sur des transcriptions automatiques. Comme seconde contribution, une étude originale est menée à partir de HATS sur l'évaluation des métriques automatiques de systèmes de RAP en étudiant leur accord avec les évaluations humaines. Notre objectif est de mettre en évidence les métriques qui corrèlent le mieux avec la perception humaine.

2 Systèmes de transcription et métriques

Dans la section 2.1, nous présentons les systèmes de reconnaissance automatique de la parole (RAP), y compris le protocole expérimental, utilisés pour construire le jeu de données HATS. Ensuite, nous présentons dans la section 2.2 les différentes mesures d'évaluation utilisées dans notre étude.

2.1 Systèmes de reconnaissance automatique de la parole (RAP)

Dans cette étude, nous avons implémenté huit systèmes de bout-en-bout (*end-to-end*) en utilisant la boîte à outils Speechbrain (Ravanelli *et al.*, 2021), et deux systèmes *pipeline* fondés sur des réseaux de neurones artificiels profonds et des Modèles de Markov caché (DNN-HMM) à l'état de l'art [1] en utilisant la boîte à outils Kaldi (Povey *et al.*, 2011). Concernant les systèmes de bout-en-bout, chacun a été entraîné en utilisant un modèle acoustique auto-supervisé différent (sept systèmes utilisent des variantes des modèles wav2vec2 appris sur le français et un utilise le modèle XLS-R-300m). Pour les systèmes pipeline, l'un des systèmes contient une étape supplémentaire de ré-évaluation en utilisant un modèle de langage neuronal.

Tous les systèmes ont été entraînés sur du français avec les corpus ESTER 1 et 2 (Galliano *et al.*, 2006, 2009), EPAC (Esteve *et al.*, 2010), ETAPE (Gravier *et al.*, 2012), REPERE (Giraudel *et al.*, 2012) et des données internes. Ces corpus représentent environ 940 heures d'audio composées de données de diffusion radio et télévision. Les transcriptions permettant de construire notre corpus HATS sont extraites du corpus de test de REPERE, qui représente environ 10 heures de données audio.

2.2 Métriques pour la RAP

En plus des métriques lexicales classiques telles que le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER), nous étudierons trois métriques sémantiques fondées sur les plongements de mots. La première, le **taux d'erreur-plongement** ou **EmBER** (pour *Embedding Error Rate*) (Bañeras-Roux *et al.*, 2022), est un WER où les erreurs de substitution sont pondérées en fonction de la distance cosinus entre le plongement lexical d'un mot de référence et le plongement du mot substituant. Les plongements lexicaux sont obtenus à partir de fastText (Grave *et al.*, 2018; Bojanowski *et al.*, 2017). La deuxième métrique, **SemDist** (Kim *et al.*, 2021), consiste à calculer une similarité cosinus entre la référence et l'hypothèse en utilisant des plongements obtenus au niveau de la phrase. Différentes méthodes sont utilisées afin d'évaluer leur impact sur la métrique : utiliser le plongement du premier token des modèles CamemBERT (Martin *et al.*, 2020) ou FlauBERT (Le *et al.*, 2020) ou d'un modèle de plongement de phrase (SentenceBERT (Reimers & Gurevych, 2019)). La dernière métrique sémantique est le **BERTScore** (Zhang* *et al.*, 2020), qui calcule un score de similarité pour chaque mot de la phrase candidate avec chaque mot de la phrase de référence en utilisant des plongements contextuels. Dans notre étude, nous utilisons un modèle multilingue BERT et le modèle CamemBERT. Alors que les transcriptions textuelles sont issues de la parole, nous considérons finalement une métrique **taux d'erreur-phonème** ou **PER** (pour *Phoneme Error Rate*) qui consiste à calculer une distance de Levenshtein entre les séquences de phonèmes de référence et d'hypothèse obtenues avec l'aide d'un convertisseur texte-à-phonème [2].

3 Évaluation humaine

La collecte du corpus HATS est décrite dans cette partie. La section 3.1 résume la mise en place de l'expérience perceptive tandis que la section 3.2 décrit le protocole permettant la sélection des transcriptions automatiques en vue de leur évaluation humaine.

1. https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/
2. https://github.com/Remiphilius/PoemesProfonds

3.1 Expérience perceptive

Dans notre étude, l'expérience côte-à-côte consiste à présenter d'une part, au sujet, une référence textuelle transcrite manuellement à partir d'un court extrait de parole, et d'autre part deux transcriptions automatiques, chacune réalisée par un système de RAP différent. Les transcriptions automatiques présentaient toujours des déviations par rapport à la référence (*i.e.* contenaient des erreurs de transcription). À l'aide de la souris, les participants devaient choisir, selon eux, la meilleure hypothèse en fonction de la référence. L'étude a utilisé un protocole d'instruction minimal, permettant aux participants de déterminer eux-mêmes les critères qui étaient importants pour déterminer la qualité d'une transcription. La référence était uniquement sous forme écrite afin que les sujets et méthodes automatique aient accès aux mêmes informations (Vasilescu *et al.*, 2012).

Pour l'étude, 143 participants se sont portés volontaires en ligne. Avant de débuter l'évaluation, les participants ont rempli un questionnaire afin de renseigner leur âge, le nombre de langues parlées, leur langue maternelle, ainsi que leur niveau d'éducation. Chaque participant a évalué 50 triplets de transcription dans un ordre aléatoire, avec un temps total moyen de 15 minutes par participant.

3.2 Protocole de sélection des transcriptions automatiques

Les triplets de transcriptions, issues du corpus de test REPERE, ont été sélectionnées en respectant les 3 critères suivants : (1) les deux hypothèses sont différentes et doivent avoir au moins un caractère de différence avec la référence, (2) chaque système doit avoir au moins quelques hypothèses face à des hypothèses de chaque systèmes et (3) la sélection des paires d'hypothèses s'appuie sur le respect de critères basées sur les scores des métriques.

Categorie	**Critères de métriques**	**Référence**	**Hypothèse A**	**Hypothèse B**
(A)	WER =	et on découvre les spectateurs	ϵ on découvre les spectateurs	et on découvre les spectacles
(A)	WER =	et on découvre les spectateurs	ϵ on découvre les spectateurs	et on découvre les spectacles
(A)	CER >	sur la vie politique	ϵ la vie politique	c' la vie politique
(A)	SemDist >>	c' est à paris	ϵ est à paris	c' est appau ϵ
(B)	WER = ; SemDist >	encore du rock	corps du rock	encore du rok
(C)	WER $\neq$ BERTscore	où les passions sont si vives	ϵ les patients sont si vive	où les patients sont si vifs

TABLE 1 – Détails de quelques critères de choix des stimuli avec des exemples.

Le point (3) peut être divisé en trois catégories différentes : (A) chaque métrique a été comparée à elle-même en présentant soit le même score, soit un score légèrement différent, soit un score très différent entre les deux hypothèses, (B) dans les deux hypothèses, le WER ou le CER étaient égaux mais le WER ou le CER, EmbER, SemDist, BERTScore étaient différents, (C) les métriques indiquaient des prédictions opposées sur quelle est la meilleure hypothèse (*e.g.* $\text{WER}_{(\text{hypA})} > \text{WER}_{(\text{hypB})}$ mais $\text{CER}_{(\text{hypA})} < \text{CER}_{(\text{hypB})}$). Le tableau 1 illustre la manière dont les hypothèses ont été confrontés avec des exemples concrets utilisés dans la tâche d'évaluation humaine.

4 Le jeu de données HATS

4.1 Description du corpus

Le corpus HATS comprend 1 000 références avec respectivement deux transcription automatiques provenant de systèmes de RAP différents. 143 humains ont chacun évalué 50 triplets référence-hypothèses, ce qui a conduit au final à 7 150 annotations. Notons que tous les triplets de transcription ont été évaluées par au moins 7 participants.

4.2 Méthodologie d'évaluation des métriques

Nous ne conservons que les annotations de transcription ayant obtenu un niveau de consensus suffisant entre les annotateurs. Nous calculons leur accord de la façon suivante : Soit A le nombre de sujets choisissant la première transcription automatique et B ceux ayant choisi l'autre, alors l'accord humain se calcule en prenant le maximum entre A et B, divisé par la somme de A et B. Les choix humains sont alors considérés selon trois pourcentages d'accord : **100%** (seulement les triplets où tous les sujets ont choisi la même hypothèse), **70%**, ou **0%** (pas de filtre) ; ce qui correspond respectivement à 371, 819 et 1000 triplets. Le seuil de 70 % a été choisi afin d'avoir un accord cohérent des annotateurs même si tous les participants ne répondent pas de la même manière (Nowak & Rüger, 2010).

Pour l'évaluation des métriques automatiques, nous cherchons à savoir si celles-ci corrèlent avec la perception humaine (*i.e.* la métrique automatique désigne la même que celle choisit par les humains). Nous obtenons donc au final, pour chaque métrique, son taux de couverture par rapport à l'annotation humaine. Le code ainsi que les données de HATS sont mises disponibles publiquement [3].

5 Évaluation des métriques

Le tableau 2 résume les résultats obtenus par chaque métrique automatique selon leur accord avec la perception humaine. Sans surprise, plus l'accord humain est élévée, plus les performances des métriques sont élevées. En contradiction avec les résultats d'études antérieures (Kim *et al.*, 2022), notre étude montre que le CER corrèle mieux avec la perception humaine que le WER. Cette divergence peut être attribuée à l'utilisation d'un texte écrit comme référence dans notre expérience perceptive, plutôt qu'un texte audio, ou à des variations linguistiques intrinsèques entre le français et l'anglais (l'orthographe du français comporte beaucoup de lettres muettes).

Il est intéressant de noter qu'au niveau des phonèmes, le PER donne de bons résultats, meilleurs que ceux du WER et du CER, malgré le fait que les humains aient fait leurs choix sur la base du texte uniquement. Cela montre que les humains semblent tenir compte de la façon dont les phrases sont phonétisées, même pendant la lecture. Ceci est particulièrement vrai si les phrases sont contrastées avec une référence.

Bien que les hypothèses choisies selon BERTScore avec les plongements BERT-base-multilingue soient 8 % meilleures que celles choisies selon SemDist utilisant des plongements de phrases multilingues, il serait précipité de conclure que la stratégie BERTScore est meilleure pour déterminer la qualité des transcriptions car les deux métriques utilisent différents plongements. En comparant ces métriques avec les mêmes plongements, SemDist est meilleure que BERTScore quand il s'agit des

3. https://github.com/thibault-roux/metric-evaluator

Accord	100%	70%	0% (Full)
Taux d'erreur mot	63% (23%)	53% (28%)	49% (28%)
Taux d'erreur caractère	77% (17%)	64% (21%)	60% (22%)
Taux d'erreur plongement	73% (12%)	62% (16%)	57% (17%)
BERTScore BERT-base multilingue	84% (0%)	75% (1%)	70% (1%)
BERTScore CamemBERT-base	81% (0%)	72% (0%)	68% (0%)
BERTScore CamemBERT-large	80% (0%)	68% (0%)	65% (0%)
SemDist CamemBERT-base	86% (0%)	74% (0%)	70% (0%)
SemDist CamemBERT-large	80% (0%)	71% (0%)	67% (0%)
SemDist Phrases CamemBERT-base	86% (0%)	75% (0%)	71% (0%)
SemDist Phrases CamemBERT-large	90% (0%)	78% (0%)	73% (0%)
SemDist Phrases multilingual	76% (0%)	66% (0%)	62% (0%)
SemDist FlauBERT-base	65% (0%)	62% (0%)	59% (0%)
Taux d'erreur phoneme	80% (14%)	69% (16%)	64% (17%)

TABLE 2 – Performance de chaque métrique en fonction de l'accord hymain. **Full** signifie qu'aucun filtre sur l'accord n'a été appliqué à l'ensemble des données. Le nombre entre parenthèses indique le pourcentage de fois où la mesure a donné la même note aux deux hypothèses.

plongements de CamemBERT-base, et SemDist a des performances similaires à BERTScore quand il s'agit des plongements de CamemBERT-large.

Sur les accords à 70 % et 0 %, le taux d'erreur mot a des performances proche d'un choix aléatoire. Cela est dû au fait que dans notre jeu de données, de nombreux cas présentent des hypothèses avec le même WER, des prédictions égales étant considérées comme un échec de la métrique puisque les humains sont capables de sélectionner une hypothèse. De plus, nous pouvons observer que SemDist utilisant les plongements de FlauBERT a de moins bonnes performances que le CER. Cela met en évidence la nécessité de choisir soigneusement les plongements et de les évaluer sur un jeu de données tel que HATS avant de tirer des conclusions sur les systèmes au niveau sémantique. Enfin, selon notre corpus orienté vers l'humain, la meilleure métrique est SemDist utilisant les plongements de phrase de CamemBERT-large, ce qui peut s'expliquer par le fait que cette métrique s'appuie sur des plongements spécifiquement entraînés pour maximiser la similarité entre des phrases ayant un sens similaire. Il est important de noter qu'une grande quantité de données annotées est nécessaire pour utiliser ces métriques fondées sur les plongements.

6 Conclusion et perspectives

Dans cette étude, des métriques automatiques appliquées à différents systèmes de RAP ont été comparées à l'évaluation humaine de différentes hypothèses erronées selon une référence écrite.

Nos résultats montrent que SemDist avec les plongements de phrases de BERT évaluent les transcriptions d'une manière qui semble acceptable pour les évaluateurs humains. Dans le cas de plongements de phrases, BERTScore semble être la deuxième meilleure option. Cette métrique est plus stable que SemDist sur les plongements de BERT. Néanmoins, si possible, les métriques devraient être évaluées sur des ensembles de données comprenant également des annotations humaines, comme HATS.

Bien que ces nouvelles méthodes d'évaluation soient intéressantes en RAP, l'avantage des métriques WER et CER est leur faible coût de calcul et l'interprétabilité du score. Par conséquent, la prochaine

étape pourrait consister à développer des métriques en corrélation avec la perception humaine tout en restant interprétables, ce qui n'est pour l'instant pas le cas de la métrique SemDist par exemple.

Dans le cadre d'un travail futur, une étude supplémentaire pourrait être menée en reproduisant l'expérience actuelle en utilisant une référence audio au lieu d'une référence textuelle, de sorte que les sujets ne disposent pas d'informations sur les caractères. Cette approche nous permettrait d'examiner les variations éventuelles et de déterminer si la métrique CER est toujours considérée comme meilleure que le WER dans un contexte multimodal.

Références

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, **33**, 12449–12460.

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

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, **5**, 135–146.

CUTLER A. (2012). *Native listening : Language experience and the recognition of spoken words*. Mit Press.

ESTEVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. In *International Conference on Language Resources and Evaluation (LREC)*.

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.

GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J.-F., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *International Conference on Language Resources and Evaluation (LREC)*, p. 139–142.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *International Conference on Language Resources and Evaluation (LREC)*, p. 1102–1107.

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.

GRAVE É., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *International Conference on Language Resources and Evaluation (LREC)*, p. 114–118.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460.

ITOH N., KURATA G., TACHIBANA R. & NISHIMURA M. (2015). A metric for evaluating speech recognizer output based on human-perception model. In *Sixteenth Annual Conference of the International Speech Communication Association*.

JUANG B. H. & RABINER L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**(3), 251–272.

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.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 2479–2490.

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.

NAM S. & FELS D. (2019). Simulation of Subjective Closed Captioning Quality Assessment Using Prediction Models. *International Journal of Semantic Computing*, **13**(01), 45–65.

NOWAK S. & RÜGER S. (2010). How reliable are annotations via crowdsourcing : a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, p. 557–566.

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.

VASILESCU I., ADDA-DECKER M. & LAMEL L. (2012). Cross-lingual studies of ASR errors : paradigms for perceptual evaluations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 3511–3518.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, p. 577–582 : IEEE.

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