Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Pratiques d'évaluation en ASR et biais de performance

Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Nous proposons une réflexion sur les pratiques d'évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d'un point de vue légal et la notion d'équité dans les systèmes d'intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d'évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l'individu n'est pas prise en compte dans les protocoles d'évaluation actuels rendant impossible l'étude de biais potentiels dans les systèmes.

ABSTRACT

Evaluation methodology in ASR and performance bias.

We propose a reflection on the evaluation practices of automatic speech recognition (ASR) systems. After defining the notion of discrimination from a legal point of view and the notion of equity in artificial intelligence systems, we look at the practices in large evaluation campaigns. Current protocols do not yet take into account the variability of speech, especially speaker variability, rendering the study of potential bias in systems impossible.

MOTS-CLÉS : reconnaissance automatique de la parole, évaluation, éthique.

KEYWORDS: automatic speech recognition, evaluation, ethics.

1 Introduction

Suite aux progrès amenés par l'essor combiné du big data et de l'apprentissage machine, les systèmes de TAL sont capables aujourd'hui d'atteindre des performances impressionnantes. Mais passé l'effervescence des premières réussites, un discours parallèle s'est construit sur l'impact de ces technologies sur nos sociétés (Boyd & Crawford, 2012; Barocas & Selbst, 2016; Hovy & Spruit, 2016). Une des études les plus médiatisées s'intéressant aux biais présents dans les systèmes issus d'apprentissages supervisés est celle rendue publique par Pro Publica dénonçant le système COMPAS (Angwin *et al.*, 2016). Ce système était utilisé par les cours de justice pour évaluer le taux de récidive d'une personne inculpée et présentait des résultats biaisés selon l'origine des individus. Par la suite, des biais ont également été découverts dans des systèmes de reconnaissance faciale (Buolamwini & Gebru, 2018) et de génération automatique de légendes d'images [1] ou encore de tri de CV. [2] Dans le domaine du

1. https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-p
Dernière consultation le 03/03/2020.
2. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
Dernière consultation le 13/03/2020

2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL), pages 1–9.
HAL : hal-02750220.

TAL, les articles sur les biais de genre, notamment concernant les représentations vectorielles de mots (plongement de mots ou *word-embeddings*) et les systèmes de traduction automatique rappellent à la communauté le caractère hautement social et situé des données langagières (Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017; Garg *et al.*, 2018; Vanmassenhove *et al.*, 2018). De manière assez surprenante en revanche, la littérature concernant l'existence possible de tels biais dans les systèmes de traitement automatique de parole reste pauvre. Cet article est une réflexion générale sur la notion d'équité dans la reconnaissance automatique de la parole [3] et sur l'utilisation du WER comme métrique d'évaluation. Il est organisé comme suit : une première partie présente la notion d'équité dans les performances de systèmes d'apprentissage automatique. Dans un second temps, nous présentons les pratiques actuelles d'évaluation des systèmes de reconnaissance automatique de la parole, en nous appuyant sur des grandes campagnes d'évaluation. Dans une troisième partie, nous questionnons les pratiques d'évaluation face à la variabilité des résultats.

2 Equité et systèmes d'apprentissage automatique

Les systèmes d'apprentissage automatique peuvent être résumés comme étant une modélisation algorithmique d'un processus décisionnel. D'un point de vue légal, Berendt & Preibusch (2014) distinguent la notion de différenciation de la notion de discrimination. Là où la différenciation se définit comme une distinction de traitement, et donc une prise de décision différente, selon un ensemble de caractéristiques ou de paramètres, la discrimination est une différenciation faite sur des caractéristiques considérées comme non-acceptables par le contrat social. En France, il existe 25 critères non-acceptés dont le sexe, l'identité de genre, les origines, la religion, la situation économique ou encore la situation familiale (LOI n° 2008-496 du 27 mai 2008 portant sur diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations [4]). Il est donc important de rappeler que les notions de discrimination et de biais restent fortement culturelles, la législation variant selon les pays. Un article de Sánchez-Monedero *et al.* (2020) s'intéressant aux systèmes automatiques pour l'embauche soulignait d'ailleurs que les travaux sur les biais de ces systèmes sont majoritairement faits en considérant le cadre socio-légal des Etats-Unis.

Comme souligné par Kate Crawford (2017) dans son intervention à NeurIPS, le terme biais, largement utilisé pour parler de systèmes discriminatoires, est polysémique et complique donc parfois les échanges entre les communautés de l'apprentissage machine et d'autres domaines comme le droit ou la linguistique. Si historiquement la notion de biais a un sens technique en statistiques, où il décrit les différences systématiques entre un échantillon et une population, il est aujourd'hui largement utilisé pour parler de discriminations, dont il est pratiquement devenu synonyme.

Face aux définitions légales, a émergé le concept d'équité (*fairness*) dans les systèmes automatiques. Ntoutsi *et al.* (2020) dénombrent plus de 20 définitions mathématiques différentes. Chen *et al.* (2018) distinguent deux types d'équité à savoir l'équité de groupe, et l'équité individuelle. L'équité individuelle pose l'hypothèse que pour des individus équivalents ne différant que par la valeur de la variable protégée, le résultat sera équivalent et se testera donc avec des modèles linéaires mixtes. L'équité de groupe suppose que les performances suivent des distributions similaires dans les sous-groupes créés par les différentes valeurs de la variable protégée et se mesure avec des tests statistiques comme le test U de Mann-Whitney.

3. Par la suite nous utiliserons l'acronyme anglais ASR qui signifie *Automatic Speech Recognition*
4. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000018877783

Afin de déterminer si le système ne donne pas les mêmes opportunités aux individus (*opportunity-based bias*) ou présente des différences significatives dans les résultats entre groupes (*outcome-based bias*), il est nécessaire d'avoir accès aux informations concernant la variable protégée pour quantifier ces biais. Ces informations sont cependant rarement prises en compte dans les pratiques d'évaluation de systèmes d'ASR. Si la récolte ou l'accès aux méta-données est un premier obstacle, d'une manière générale, la notion de variabilité intrinsèque à la parole semble disparaître dans les procédures d'évaluation.

3 Évaluation en ASR

3.1 Bref rappel de la genèse de la tâche

Historiquement, la première tâche de reconnaissance de la parole consistait à reconnaître les dix chiffres isolément pour un locuteur donné à l'aide d'un dispositif cablé (Davis *et al.*, 1952). À partir des années 1960, les méthodes numériques ont été introduites, améliorant les performances sur les mots isolés, la parole continue restant une tâche particulièrement complexe (Haton *et al.*, 2006). Les tâches ont donc d'abord été simplifiées en supprimant des facteurs de difficulté : reconnaissance de mots isolés, puis de mots enchaînés, souvent sur des configurations mono-locuteur. Les phénomènes de coarticulation présents sur de la parole continue et la variabilité inter-locuteurs ont été traités par la suite, grâce aux progrès en informatique et en électronique, pour permettre maintenant de traiter des situations de communication écologiques.

La reconnaissance de la parole ayant pour but principal d'accéder au message et donc au contenu lexical, tout ce qui relevait de la variation phonostylistique a été considéré comme du bruit. L'objectif était d'augmenter "la robustesse des systèmes à l'environnement (bruit, locuteurs...)" (Calliope, 1989). Ont donc été proposées des techniques comme la normalisation des paramètres acoustiques, permettant de gérer différents environnements (téléphone, radio, variabilité des microphones, etc.). Dans ce contexte, l'individu est considéré comme une source de variabilité de même que son sexe, son âge, son accent, sa catégorie sociale ou encore son état physique et émotionnel, chaque critère impactant la production de la parole (Kreiman & Sidtis, 2011). En évoluant vers des systèmes indépendants du locuteur, la variabilité due à l'individu a été prise en compte, d'abord à l'aide de modèles en fonction du genre, puis ensuite par les différentes techniques d'adaptation. Mais cette prise en compte de la variabilité des locuteurs et locutrices dans le développement des systèmes ne se retrouve pas dans les pratiques d'évaluation.

3.2 Évaluation

Lorsque sont reportés des résultats de systèmes de reconnaissance automatique de la parole, la métrique utilisée est le taux d'erreur-mots ou WER (*word-error rate*), basé sur la distance de Levenshtein et se calculant comme la somme des erreurs (insertion, délétion et substitution) de l'hypothèse divisée par le nombre de mots total dans la référence. En pratique, le WER est calculé à l'échelle du corpus de test, lissant ainsi les variations dues à la longueur des énoncés. Le développement des systèmes d'ASR ayant souvent donné lieu à des campagnes d'évaluation (campagnes NIST de 2002 à 2009 [5]

5. https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

pour l'anglais, campagnes ESTER [6] et ETAPE [7] pour le français), le report d'une mesure unique permettait la comparaison directe des systèmes entre eux, les données de test étant communes. En 2017, IBM reportait un WER de 5,5% sur SwitchBoard et 10,3% sur CallHome (Saon *et al.*, 2017), et Microsoft des WER de 5,1% et 9,8% sur ces mêmes corpus, considérant avoir atteint un niveau de performance similaire voire supérieur à l'humain (Xiong *et al.*, 2018). Mais ce mode d'évaluation de la parole, à l'aide d'une mesure unique, va de pair avec une conception complètement désincarnée du langage. On évalue le système en décorrélant complètement le fait que cette parole est produite de manière située, par un individu en contexte.

Comme expliqué dans la Section 3.1, la variabilité de la parole, venant des individus ou des environnements a été vue comme un bruit à gommer dans la conception des systèmes. Des campagnes ont donc vu le jour pour prendre en compte ces défis. Les différentes campagnes de NIST 2002 à 2009 ont notamment largement travaillé sur différents enjeux techniques dûs aux environnements sonores : bande-passante du téléphone, enregistrements bruités, séparation parole/musique, etc., et ces problématiques sont d'ailleurs toujours d'actualité avec des campagnes telles que CHiME. [8] La notion de variation stylistique est quant à elle prise en compte dans les campagnes MGB Challenge [9] où les performances sont reportées en fonction des différentes émissions avec des variations de WER pouvant aller jusqu'à plus de 30 points (Bell *et al.*, 2015). L'âge n'a pas encore été pris en compte dans des campagnes d'évaluation de grande envergure mais il est abordé dans certains travaux portant sur la reconnaissance de la parole des enfants (Kennedy *et al.*, 2017) ou la parole des personnes agées (Aman *et al.*, 2013) qui montrent des différences de performances liées à l'âge.

Notre travail de recherche s'intéresse plus spécifiquement à la variation de genre. Nous parlons de genre, car nous nous intéressons aux caractéristiques présentes dans la parole des individus et relevant d'une instanciation individuelle de la représentation sociale sexuée de l'individu (Ochs, 1992). En ASR, la plupart des études et des données langagières ne font référence qu'à deux catégories femme/homme. À notre connaissance, seules trois études se sont réellement intéressées aux différences de performances entre hommes et femmes dans des tâches d'ASR durant les deux dernières décennies. Adda-Decker & Lamel (2005) observaient de meilleures performances sur les voix de femmes. Les travaux plus récents de (Tatman, 2017), au contraire, mettaient en avant une performance moins bonne dans le sous-titrage automatique de YouTube sur les voix de femmes. Cependant, cette tendance n'était plus significative dans son étude de la même année avec Kasten, contrairement aux variations de performances en fonction de l'accent ou de l'origine éthnique (Tatman, 2017; Tatman & Kasten, 2017).

On observe donc que s'il est de coutume de reporter des résultats en fonction de certains types de variations considérées comme posant des problèmes techniques (environnement sonore, phonostyle), dans le cadre des campagnes, peu d'études se consacrent à la variation des performances en fonction des caractéristiques des individus. Le sexe, l'âge, l'appartenance ethnique, qui sont pourtant des variables protégées aux yeux de la loi française, ne sont pas prises en compte explicitement comme facteurs de variation des performances des systèmes. Or, s'assurer de l'équité de groupe pour ces facteurs constitue une étape nécessaire pour une diffusion éthique de ces technologies dans la société.

6. http://www.afcp-parole.org/camp_eval_systemes_transcription/present.html
7. http://www.afcp-parole.org/etape.html
8. https://chimechallenge.github.io/chime6/
9. http://www.mgb-challenge.org/

4 Une évaluation différenciée : première analyse sur les données de Librispeech

Dès les premiers systèmes d'ASR, les enjeux différenciés des évaluations en fonction des objectifs de chaques acteurs : chercheurs, industriels, etc., ont été problématisés (Pallett, 1985). En effet, évaluer les performances d'un système n'a de sens qu'au regard de l'usage qu'il en est fait. Cependant, les systèmes étant de plus en plus utilisés dans nos sociétés, il est nécessaire de penser l'impact qu'ils peuvent avoir sur la population. Dans le cadre de la reconnaissance automatique de la parole, l'évaluation en fonction des caractéristiques individuelles des locuteurs et locutrices n'est jamais abordée, alors que des outils d'évaluation permettent de prendre en compte ces préoccupations. Le National Institute of Standards and Technology [10] (NIST) a développé un outil d'évaluation des systèmes de reconnaissance automatique de la parole : le Speech Recognition Scoring Toolkit [11] (SCTK), largement utilisé et intégré dans la boîte à outils KALDI (Povey *et al.*, 2011). Cet outil permet des évaluations par locuteur ou locutrice, mais ces options sont rarement utilisées dans les reports de résultats.

Nous présentons ici un rapport d'évaluation d'un système d'ASR prenant en compte la variabilité propre à l'individu sur les données de Librispeech (Panayotov *et al.*, 2015). Le système utilisé a été développé en utilisant ESPnet (Watanabe *et al.*, 2018) et la recette fournie pour le corpus. Nous atteignons un WER de 4.2% sur le jeu de données nommé *test-clean*, validant ainsi notre système comme état de l'art.

La Figure 1 présente la répartition des WER par individu en fonction du genre, obtenue par notre système. Nous avons regroupé ici les deux partitions de test (clean et other) en raison du faible nombre de locuteurs et locutrices distinct·es (l'échantillon total contient 37 locutrices et 36 locuteurs). On semble observer une répartition différente en fonction du genre, mais un nombre plus conséquent d'observations nous permettrait d'avoir une distribution plus fine des WER pour attester ou non de l'équité de notre système. En l'état actuel des mesures, le test U de Mann-Whitney n'est pas significatif (U=715, p-valeur=0.59). Bien qu'indicatifs, ces résultats semblent montrer que les performances varient selon le genre des individus, ce que nous avions aussi observé dans une précédente étude sur le français (Garnerin *et al.*, 2019) et il serait intéressant d'étudier ces variations en fonction des différentes caractéristiques des individus, notamment dans les corpus largemement utilisés par la communauté.

5 Discussion

Dans une précédente étude (Garnerin *et al.*, 2019) nous avons montré que la faible présence des femmes dans les données médiatiques conduisait à un bias de performance genré. Face à ce constat, une solution serait de rééquilibrer les données, sacrifiant les performances du système pour assurer une équité de groupe. Nous ne cherchons pas à niveller par le bas, mais à alerter sur les pratiques actuelles d'évaluation des systèmes de reconnaissance de parole, qui ne permettent pas de connaître la variabilité de performance des systèmes sur différents groupes de locuteurs et locutrices. Cette évaluation lacunaire peut conduire à des problèmes éthiques dans la mesure où cela ne permet pas de définir

10. https://www.nist.gov/itl
11. https://github.com/usnistgov/SCTK

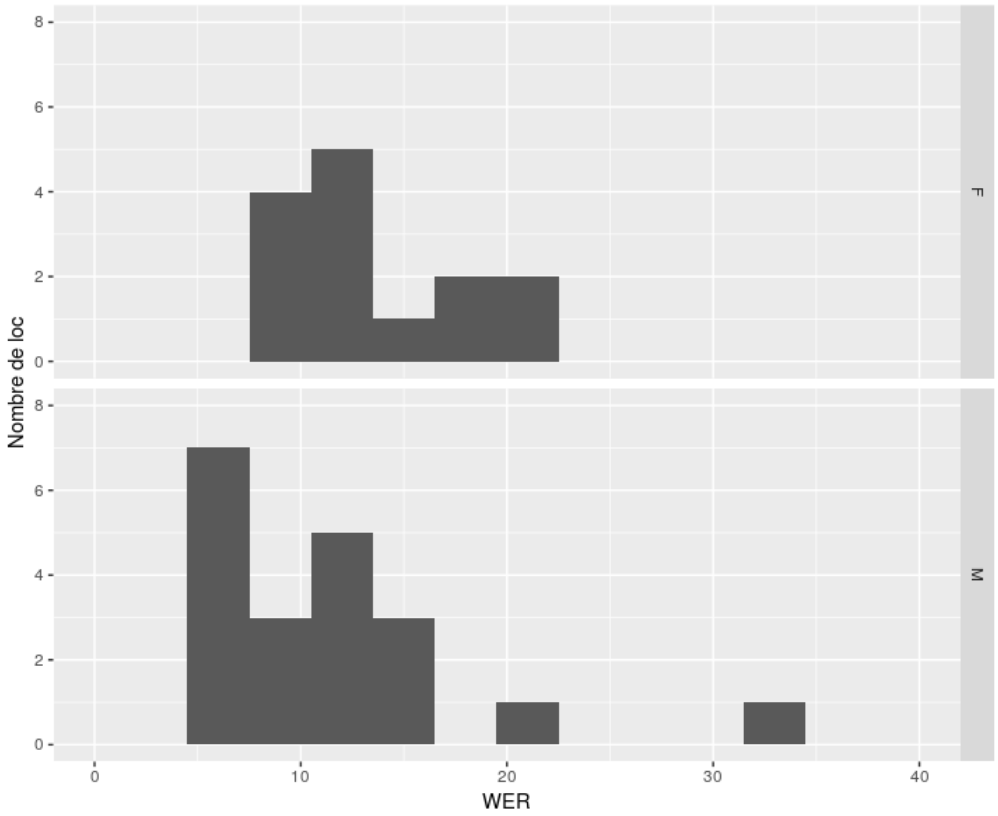

FIGURE 1 – Distribution des WER (en %) en fonction du genre (en haut : femmes ; en bas : hommes) sur le regroupement des jeux de données *test-clean* et *test-other* du corpus Librispeech

clairement les usages qu'il peut être fait des systèmes. En effet, la question éthique ne peut pas être envisagée indépendamment des usages. Actuellement, la reconnaissance de parole est principalement utilisée dans l'industrie pour un ensemble de tâches telles que le compte rendu de réunion ou la saisie de documents, mais également pour le sous-titrage et l'indexation de contenus audio-visuels ainsi que pour le maintien des personnes agées à domicile à l'aide de systèmes de domotique. La dictée vocale ou le compte-rendu ne posent pas en eux-même des problèmes particuliers, étant donné que ces systèmes fonctionnent en local, avec une étape d'adaptation à l'utilisateur ou à l'utilisatrice. En revanche, en ce qui concerne l'indexation et le sous-titrage, on peut se demander si cela ne va pas contribuer à l'invisibilisation de certaines catégories de personnes dans les médias : on peut penser aux femmes (Doukhan & Carrive, 2018), mais également à l'impact sur la parole accentuée. En ce qui concerne le service à la personne, la prise en compte du genre est incontournable étant donné que les hommes ne représentent que 38,9% de la population des 75 ans et plus, comme rapporté par l'édition 2019 du portrait social de l'INSEE. [12] Le faible nombre d'études sur les différences de performances entre hommes et femmes dans les systèmes de reconnaissance automatique de la parole s'explique donc peut-être par le faible impact sociétal des écarts de performances lors de l'utilisation actuelle de ces systèmes. En revanche, avec l'émergence des assistants vocaux, qui sont des services fonctionnant sur serveurs, sans adaptation directe au locuteur ou à la locutrice, on observe une volonté de faire de la voix la nouvelle interface de nombreux produits de service. On peut également se questionner sur l'effet que pourraient avoir des performances différenciées dans le cas de systèmes de traductions speech to text ou speech to speech.

12. https://www.insee.fr/fr/statistiques/4238375?sommaire=4238781#consulter

6 Conclusion

La notion d'éthique en traitement automatique des langues est un enjeu majeur dont la communauté de parole doit s'emparer. Dans cet article, nous proposons de repenser l'évaluation des systèmes de reconnaissance automatique de la parole en termes d'équité. Il est clair que la parole est un domaine dans lequel le sexe, l'identité de genre, l'âge, l'appartenance ethnique et la classe sociale sont des sources importantes de variabilité. Les possibilités d'analyse des performances des systèmes en fonction de ces variables protégées par la loi française existent. Nous ne pouvons que regretter qu'elles ne fassent pas partie des pratiques d'évaluation de la communauté. Les grandes campagnes d'évaluation ont pour objectif de comparer les systèmes des différentes équipes afin d'améliorer les architectures et ne comparent donc que les systèmes entre eux. Mais avec la diffusion de ces systèmes dans la société, il est nécessaire de penser une évaluation différente, dans laquelle il ne s'agit non pas que de trouver le meilleur système mais aussi de s'assurer de la conformité des systèmes au cadre légal, pour en limiter d'éventuels impacts sociétaux négatifs.

Références

ADDA-DECKER M. & LAMEL L. (2005). Do speech recognizers prefer female speakers ? In *Actes de INTERSPEECH 2005 (International Speech Communication Association)*, p. 2205–2208, Lisbon, Portugal : ISCA.

AMAN F., VACHER M., ROSSATO S. & PORTET F. (2013). Speech recognition of aged voice in the AAL context : Detection of distress sentences. In *Actes de SPED13 (Conference on Speech Technology and Human-Computer Dialogue)*, p. 1–8 : IEEE.

ANGWIN J., LARSON J., MATTU S. & KIRCHNER L. (2016). Machine bias. *ProPublica*, **23**.

BAROCAS S. & SELBST A. D. (2016). Big data's disparate impact. *California Law Review*, **104**, 671.

BELL P., GALES M. J., HAIN T., KILGOUR J., LANCHANTIN P., LIU X., MCPARLAND A., RENALS S., SAZ O., WESTER M. *et al.* (2015). The MGB challenge : Evaluating multi-genre broadcast media recognition. In *Actes de ASRU 2015 (Workshop on Automatic Speech Recognition and Understanding)*, p. 687–693 : IEEE.

BERENDT B. & PREIBUSCH S. (2014). Better decision support through exploratory discrimination-aware data mining : foundations and empirical evidence. *Artificial Intelligence and Law*, **22**(2), 175–209.

BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? Debiasing word embeddings. In *Actes de NeurIPS 2016 (Neural Information Processing Systems)*, p. 4349–4357.

BOYD D. & CRAWFORD K. (2012). Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, **15**(5), 662–679.

BUOLAMWINI J. & GEBRU T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. In *Actes de FAT 2018 (Fairness, Accountability and Transparency)*, p. 77–91, New-York City, USA : ACM.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

CALLIOPE (1989). *Ergonomie et évaluation du traitement de la parole par ordinateur*, In J. TUBACH, Éd., *La parole et son traitement automatique*, chapitre 26, p. 689–705. Paris : Masson.

CHEN L., MA R., HANNÁK A. & WILSON C. (2018). Investigating the impact of gender on rank in resume search engines. In *Actes de CHI 2018 (Conference on Human Factors in Computing Systems)*, p. 1–14, Montréal, QC, Canada.

CRAWFORD K. (2017). The trouble with bias. NIPS 2017 Keynote.

DAVIS K. H., BIDDULPH R. & BALASHEK S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, **24**(6), 637–642.

DOUKHAN D. & CARRIVE J. (2018). Description automatique du taux d'expression des femmes dans les flux télévisuels français. In *Actes de JEP 2018 (Journées d'Études sur la Parole)*, p. 496–504, Aix-en-Provence, France.

GARG N., SCHIEBINGER L., JURAFSKY D. & ZOU J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, **115**(16), E3635–E3644.

GARNERIN M., ROSSATO S. & BESACIER L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *Actes de AI4TV 2019 (Workshop on AI for Smart TV Content Production, Access and Delivery)*, p. 3–9, Nice, France : ACM. DOI : 10.1145/3347449.3357480.

HATON J.-P., CERISARA C., FOHR D., LAPRIE Y. & SMAÏLI K. (2006). *Introduction à la reconnaissance automatique de la parole*, In *Reconnaissance automatique de la parole : Du Signal à son Interprétation*, chapitre 1, p. 1–15. Paris : Dunod.

HOVY D. & SPRUIT S. L. (2016). The social impact of Natural Language Processing. In *Actes de ACL 2016 (Volume 2 : Short Papers)*, p. 591–598, Berlin, Allemagne : Association for Computational Linguistics. DOI : 10.18653/v1/P16-2096.

KENNEDY J., LEMAIGNAN S., MONTASSIER C., LAVALADE P., IRFAN B., PAPADOPOULOS F., SENFT E. & BELPAEME T. (2017). Child speech recognition in human-robot interaction : evaluations and recommendations. In *Actes de HRI 2017 (International Conference on Human-Robot Interaction)*, p. 82–90 : ACM/IEEE.

KREIMAN J. & SIDTIS D. (2011). *Physical Characteristics and the Voice : Can We Hear What a Speaker Looks Like ?*, In *Foundations of Voice Studies*, chapitre 4. Wiley-Blackwell.

NTOUTSI E., FAFALIOS P., GADIRAJU U., IOSIFIDIS V., NEJDL W., VIDAL M.-E., RUGGIERI S., TURINI F., PAPADOPOULOS S., KRASANAKIS E., KOMPATSIARIS I., KINDER-KURLANDA K., WAGNER C., KARIMI F., FERNANDEZ M., ALANI H., BERENDT B., KRUEGEL T., HEINZE C., BROELEMANN K., KASNECI G., TIROPANIS T. & STAAB S. (2020). Bias in Data-driven AI Systems - An Introductory Survey. *arXiv preprint 2001.09762*.

OCHS E. (1992). *Indexing gender*, In A. DURANTI & C. GOODWIN, Éds., *Rethinking Context : Language as an interactive phenomenon*, p. 335—350. Cambridge University Press.

PALLETT D. S. (1985). Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards*, **90**, 371–387.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an ASR corpus based on public domain audio books. In *Actes de ICASSP 2015 (Acoustics, Speech and Signal Processing*, p. 5206–5210, Brisbane, Australie : IEEE.

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 *Actes de IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society.

SÁNCHEZ-MONEDERO J., DENCIK L. & EDWARDS L. (2020). What does it mean to solve the problem of discrimination in hiring ? social, technical and legal perspectives from the uk on automated hiring systems. In *Actes de FAT 2020 (Fairness, Accountability and Transparency)*, Barcelona, Spain : ACM.

SAON G., KURATA G., SERCU T., AUDHKHASI K., THOMAS S., DIMITRIADIS D., CUI X., RAMABHADRAN B., PICHENY M., LIM L.-L., ROOMI B. & HALL P. (2017). English conversational telephone speech recognition by humans and machines. In *Actes de INTERSPEECH 2017 (International Speech Communication Association)*, p. 132–136 : ISCA. DOI : 10.21437/Interspeech.2017-405.

TATMAN R. (2017). Gender and dialect bias in youtube's automatic captions. In *Actes de ACL Workshop on Ethics in Natural Language Processing*, p. 53–59.

TATMAN R. & KASTEN C. (2017). Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Actes de INTERSPEECH 2017 (International Speech Communication Association)*, p. 934–938 : ISCA.

VANMASSENHOVE E., HARMEIER C. & WAY A. (2018). Getting gender right in neural machine translation. In *Actes de EMNLP 2018 (Empirical Methods in Natural Language Processing)*, p. 3003–3008.

WATANABE S., HORI T., KARITA S., HAYASHI T., NISHITOBA J., UNNO Y., ENRIQUE YALTA SOPLIN N., HEYMANN J., WIESNER M., CHEN N., RENDUCHINTALA A. & OCHIAI T. (2018). Espnet : End-to-end speech processing toolkit. In *Actes de INTERSPEECH 2018 (International Speech Communication Association)*, p. 2207–2211, Hyderabad, India : ISCA. DOI : 10.21437/Interspeech.2018-1456.

XIONG W., WU L., ALLEVA F., DROPPO J., HUANG X. & STOLCKE A. (2018). The Microsoft 2017 conversational speech recognition system. In *Actes de ICASSP 2018 (Acoustics, Speech and Signal Processing*, p. 5934–5938, Calgary, Alberta, Canada : IEEE.