Du recueil à l'exploitation des corpus de parole " pathologique " : comment accéder à la variation physiopathologique ?

Alain Ghio, Gilles Pouchoulin, François Viallet, Antoine Giovanni, Virginie Woisard, Lise Crevier-Buchman, Fabrice Hirsch, Camille Fauth, Corinne Fredouille

▶ **To cite this version:**

Alain Ghio, Gilles Pouchoulin, François Viallet, Antoine Giovanni, Virginie Woisard, et al.. Du recueil à l'exploitation des corpus de parole " pathologique " : comment accéder à la variation physiopathologique ?. Corpus, 2021, 22, 10.4000/corpus.5677 . hal-03145102

HAL Id: hal-03145102
https://hal.science/hal-03145102

Submitted on 18 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CORPUS

Corpus

22 | 2021
Du recueil à l'outillage des corpus oraux : comment accéder à la variation ?

Du recueil à l'exploitation des corpus de parole « pathologique » : comment accéder à la variation physiopathologique ?

Alain Ghio, Gilles Pouchoulin, François Viallet, Antoine Giovanni, Virginie Woisard, Lise Crevier-Buchman, Fabrice Hirsch, Camille Fauth et Corinne Fredouille

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/5677
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Alain Ghio, Gilles Pouchoulin, François Viallet, Antoine Giovanni, Virginie Woisard, Lise Crevier-Buchman, Fabrice Hirsch, Camille Fauth et Corinne Fredouille, « Du recueil à l'exploitation des corpus de parole « pathologique » : comment accéder à la variation physiopathologique ? », *Corpus* [En ligne], 22 | 2021, mis en ligne le 08 février 2021, consulté le 16 février 2021. URL : http://journals.openedition.org/corpus/5677

Ce document a été généré automatiquement le 16 février 2021.

© Tous droits réservés

Du recueil à l'exploitation des corpus de parole « pathologique » : comment accéder à la variation physiopathologique ?

Alain Ghio, Gilles Pouchoulin, François Viallet, Antoine Giovanni, Virginie Woisard, Lise Crevier-Buchman, Fabrice Hirsch, Camille Fauth et Corinne Fredouille

1. Introduction

1.1. Une nécessité de Sciences Ouvertes

1 Les recherches sur l'évaluation des troubles de la voix et de la parole nécessitent la structuration et l'organisation d'un large ensemble de données (Schuller, 2015). En effet, le cadre « pathologique » induit une variation considérable dans ses manifestations de surface, c'est-à-dire sur les productions sonores. Aux symptômes de la maladie se superposent les effets variables des traitements ainsi que des phénomènes de compensation non uniformes des locuteurs. De ce fait, toute généralisation à une population clinique particulière nécessite l'observation d'un grand nombre de patients du fait de la très forte variation interindividuelle rencontrée.

2 De plus, la plupart des études nécessite une comparaison à un groupe contrôle qui, dans la mesure du possible, doit être similaire à celui des patients. Il est ainsi nécessaire dans le cadre de maladies neurodégénératives d'avoir des groupes contrôle de personnes âgées sans troubles de la parole, ce qui n'est pas facile à obtenir. Il est donc important de capitaliser et mutualiser les enregistrements existants.

3 En outre, pour être utilisables, ces enregistrements doivent répondre à de fortes exigences.

4 (1) des signaux de haute qualité, afin que les distorsions et le bruit ne soient pas attribués à des dysfonctionnements de la voix ou de la parole.

5 (2) des énoncés suffisamment informatifs. Les voyelles tenues sont nécessaires pour évaluer le mécanisme de phonation mais la parole continue est incontestablement plus naturelle du point de vue de la communication orale (Parsa *et al.*, 2001).

6 (3) des informations cliniques, suffisamment précises, pour gérer différents ensembles de locuteurs et différents contextes d'élocutions (avec/sans médicament, avant/après rééducation ou opération chirurgicale, durée de la maladie, durée des traitements, etc.).

7 (4) un grand nombre de locuteurs. Toute généralisation d'une population clinique spécifique nécessite la prise en compte de nombreux intervenants en raison de la très grande variabilité inter-locuteurs rencontrée (différentes évolutions de la maladie, stratégies de compensation individuelle, gravité et spécificité des maladies).

8 Si les problèmes de prise de son ou autres signaux physiologiques sont en passe de devenir anecdotiques grâce à la diffusion de matériels de qualité et à la meilleure formation des personnels en charge des enregistrements, si le stockage des signaux de parole ne constitue plus actuellement un obstacle, si le recours à du matériau linguistique suffisant se généralise, le maillon faible reste la normalisation et la structuration des données sur les locuteurs et leurs productions langagières.

1.2. La perte d'information

9 Concrètement, si les données sonores peuvent être accessibles, elles ne présentent au final aucun intérêt si les liens entre les enregistrements et les caractéristiques cliniques du locuteur sont rompus ou erronés. Or, cette information clinique doit rester consultable et pérenne de façon anonyme, ce qui est difficile à maintenir. Il ne faut surtout pas négliger les contraintes logistiques et organisationnelles qui peuvent peser sur les personnes en charge des enregistrements dans les établissements hospitaliers. Les contraintes temporelles des consultations ne permettent pas un contrôle qualité et un formatage parfait des données, ce qui nécessite un travail supplémentaire dans le cadre de la constitution de bases de données. La passation puis la saisie d'examens cliniques reste aussi difficile à rendre systématique. Nous pensons par exemple aux épreuves neuropsychologiques à garder dans le cadre de maladies neurologiques, à l'UPDRS dans le cas particulier de la maladie de Parkinson, au GRBAS des dysphoniques, au Voice Handicap Index, au Speech Handicap Index...

10 La non-connexion généralisée des ordinateurs dans les hôpitaux pour éviter le piratage rend compliqué la mise à jour d'information et rend impossible le transfert simple de données. L'expérience montre que seule l'implication de personnels clairement identifiés pour la constitution de bases de données tels qu'un attaché de recherche clinique, une orthophoniste, une psychologue, un vacataire... permet d'obtenir des données exploitables au final.

1.3. Le cadre législatif

11 Le cadre législatif s'avère comme un obstacle à la mise en œuvre de vastes bases de données de parole pathologique.

12 Alors que le mouvement d'ouverture des données de la recherche scientifique, initié en 2016 par la Loi Lemaire[1], vise à une meilleure valorisation de l'investissement public et compose un axe de travail du Comité pour la Science Ouverte[2], il n'en reste pas moins que les chercheurs sont communément confrontés à des difficultés techniques et juridiques, dues principalement à la nature des données et au contexte dans lequel celles-ci ont été produites ou collectées.

13 À cela est venue s'ajouter la promulgation en mai 2018 du Règlement Général européen sur la Protection de Données (RGPD, 2018)[3] renforçant, entre autres, le respect de la vie privée des personnes déjà exigé par la loi « Informatique et Libertés » (LIL, 1978)[4]. S'appliquant bien entendu aux données scientifiques, les chercheurs appréhendent ces obligations comme un changement majeur dans leurs activités de recherche, tout en les obligeant à s'interroger sur la gouvernance et le régime de protection des données traitées, que celles-ci soient qualifiées de « personnelles » ou « cliniques ».

14 D'un point de vue pragmatique, le contexte législatif et politique en matière de protection et d'ouverture des données, ne facilite pas l'implémentation et l'utilisation des bases de données cliniques, sonores et physiologiques. En effet, le respect des obligations légales et réglementaires soulève de nombreuses interrogations juridiques et techniques concernant le droit d'auteur et de propriété, l'anonymisation des données, l'accès et la diffusion des données, etc. (Lalain *et al.*, 2020)

1.4. Un contexte clinique réticent à la Science Ouverte

15 Un obstacle sérieux à la constitution de bases de données de parole pathologique est l'appréhension liée au partage de données dans les établissements hospitaliers, indépendamment de la question du secret médical qui peut être maîtrisée. En effet, la culture de la Science Ouverte y reste peu répandue essentiellement pour des raisons historiques et de culture scientifique plus compétitive qu'en Sciences Humaines. Chaque équipe clinique a tendance à exploiter son groupe de patients de façon exclusive sauf dans le cadre de vastes projets multicentriques (Schuepbach *et al.*, 2013). Or, du fait que certaines pathologies sont rares et qu'il n'est pas toujours facile d'enregistrer certains patients, l'acquisition de données vocales pathologiques dans plusieurs centres reste une condition indispensable à la constitution de cohortes suffisantes pour en tirer des conclusions généralisables. En revanche, le recours à divers centres d'enregistrements peut engendrer des variabilités non désirées liées à la spécificité du service. Il est donc important d'adopter un certain nombre de pratiques communes et partagées de façon à limiter ces biais contextuels.

16 Dans tous les cas, concernant le partage possible de données, il est indispensable de gérer de façon fine les privilèges/rôles accordés aux demandeurs de données et ce, en fonction des desiderata des producteurs de données (les hôpitaux). En effet, seule la mise en place de tels contrôles précis permettra de lever les réticences légitimes des partenaires hospitaliers.

1.5. Une reconnaissance interdisciplinaire des contributions

17 Dans la difficile interdisciplinarité à mettre en place dans les recherches sur l'évaluation des troubles de la voix et de la parole, il est fréquent d'assister à une forme de dénis de la contribution des partenaires hospitaliers qui peuvent être considérés,

par les chercheurs en sciences du langage ou en traitement automatique, comme de simples fournisseurs de données exploitées ensuite par les disciplines non cliniques. On assiste alors à des travaux dans lesquels les producteurs de données hospitaliers n'apparaissent pas dans la liste des auteurs, minimisant leur contribution pourtant essentielle dans ces travaux.

18 Il faut donc très clairement se mettre d'accord sur un modèle de licence d'utilisation des données de façon à ne pas considérer les cliniciens comme de simples fournisseurs de patients mais au contraire, en les plaçant de façon active dans le processus de recherche. Il peut être ainsi proposé que l'investigateur clinique principal apparaisse systématiquement sur tous les travaux qui découlent de la collection de données enregistrée sous son égide. Cela revêtira divers avantages : une reconnaissance de sa contribution (et des membres du service), la mise au courant de l'avancée des travaux, le regard du clinicien, la crédibilité du résultat final liée à la présence du spécialiste médical.

2. Les corpus de parole pathologique en français

19 Au niveau francophone, il existe un certain nombre d'initiatives locales. En préambule, nous tenons à rappeler le rôle du « Groupe Francophone d'Étude de la Dysarthrie » initié en 2004, regroupant des neurologues, ORL, phoniatres, orthophonistes, ingénieurs et chercheurs, issus des centres de recherche d'Aix-en-Provence, Boulogne-sur-Mer, Lille, Marseille, Paris, Rouen et Toulouse. Cette initiative précurseur, décrite en détail dans Jan (2007), mériterait une remise à jour contemporaine pour permettre de créer une initiative fédératrice au niveau national. En attente, nous présentons différentes initiatives régionales de façon non exhaustive.

2.1. Le corpus MTO (Marseille Timone ORL) de voix dysphoniques[5]

20 Pendant plus de vingt ans, le service ORL du CHU de la Timone à Marseille (à présent localisé sur l'hôpital de la Conception) a enregistré des patients dysphoniques qui venaient en consultation médicale (Ghio *et al.*, 2012). Pour des raisons logistiques, les informations sur les patients étaient stockées sur des cahiers dans lesquels sont indiqués l'identité des locuteurs, leur pathologie, la date de l'examen, le contexte pré/post-opératoire, etc. Un important travail de numérisation, d'indexation et de saisie d'informations a permis de constituer une collection de 1530 patients dysphoniques produisant des voyelles tenues, lisant un texte, chantant une chanson pour un total de 1953 sessions d'enregistrements (certains locuteurs sont enregistrés plusieurs fois). Cette collection comprend des données provenant de 504 hommes et 1026 femmes. Les principales pathologies sont les nodules, les paralysies laryngées, les polypes, les œdèmes de Reinke et les dysphonies dysfonctionnelles à larynx normal. Parmi ces locuteurs dysphoniques, 332 d'entre eux ont été enregistrés plusieurs fois (ex : avant et après chirurgie).

21 La plupart des productions vocales (1766 sessions) ont été évaluées de manière perceptive à l'aide de l'échelle GRBAS (Hirano, 1981). Cette évaluation réalisée par une unique orthophoniste lors de la session d'enregistrement doit être considérée comme un niveau approximatif de la dysphonie.

22 Ces données ont notamment contribué à la réalisation de l'*International consensus on basic voice assessment for unilateral vocal fold* (Mattei *et al.*, 2018).

2.2. Le corpus AHN (Aix Hôpital Neurologie) de dysarthries

23 Pendant plus de vingt ans, le service de neurologie du CH du Pays d'Aix à Aix-en-Provence a enregistré des patients dysarthriques qui venaient en consultation médicale. Un formulaire informatisé a été utilisé pour stocker les données cliniques. Nous avons actuellement collecté les enregistrements sonores et aérodynamiques de 990 patients et 160 sujets témoins plutôt âgés. La population pathologique est composée de divers troubles neuromoteurs : AVC, sclérose latérale amyotrophique (SLA), maladie de Friedreich, maladie de Huntington... La maladie de Parkinson (601) et les syndromes parkinsoniens (98) représentent l'essentiel de ce corpus car une attention importante a été portée aux études sur cette maladie (Pinto *et al.*, 2010).

24 L'originalité de ce corpus réside dans :

25 (1) La présence de signaux complémentaires aux signaux sonores, tels que l'intensité SPL, le débit d'air oral, la pression de l'air sous-glottique, etc. (Ghio *et al.*, 2012)

26 (2) Les différents contextes pour les enregistrements de 601 patients atteints de la maladie de Parkinson (avec/sans médicament, avec/sans stimulation subthalamique...), qui représentent 1616 séances d'enregistrement

27 (3) La collecte d'informations précises sur les locuteurs (date et lieu de naissance, langue maternelle...) et les conditions cliniques (date de détection de la maladie, localisation des symptômes, traitement régulier et traitement réel lors de l'enregistrement, résultats des examens cliniques...)

2.3. Le corpus CCM de parole dysarthrique (Paris)

28 Pendant plus de 30 ans (1965-1997), le laboratoire de la voix, INSERM U3 à l'hôpital de la Salpêtrière, puis à hôpital Laennec et l'hôpital HEGP ont enregistré plus de 700 patients présentant des dysarthries. Ces enregistrements des voix et parole dysarthriques chez l'adulte ont constitué un corpus appelé CCM (Claude Chevrie-Muller, directrice du laboratoire).

29 Les patients étaient adressés par les différents services de neurologie pour un diagnostic de dysarthrie basé sur les troubles de la voix et de la parole. Les dossiers patients comprenaient les informations personnelles : sexe, date de naissance, lieu de naissance, langue maternelle et les langues parlées, l'activité professionnelle, ainsi que le dossier médical partagé avec le service de neurologie. Les caractéristiques de la pathologie étaient consignées dans le dossier comme le mode d'apparition, la durée d'évolution, les prises en charge thérapeutiques (médicamenteuses, chirurgicales, physiothérapies, orthophonies...). Un certain nombre de patients ont été enregistrés plusieurs fois permettant d'avoir un suivi longitudinal de leur dysarthrie.

30 Les enregistrements acoustiques étaient réalisés sur un enregistreur à deux pistes Revox permettant d'acquérir le son et l'EGG (électroglottographie). Ces enregistrements se faisaient en chambre sourde, de façon systématique avec le même protocole pour tous les patients. Les différentes tâches comprenaient le comptage et séries automatiques (1 à 10 et les mois de l'année), la lecture d'une phrase intonative, la

lecture d'une liste de mots explorant les différentes situations de co-articulation, la tenue des 5 voyelles cardinales, la lecture des syllabes avec toutes les consonnes du français (CV - VCV), la lecture d'un texte (conte pour enfant), la description d'une histoire en image et de la parole spontanée. Ces données ont été numérisées dans le cadre de l'ANR DesPhoAPaDy (08-Blan-0125) en 2009 (Fougeron *et al.*, 2010). On a ainsi pu constituer une collection de plus de 1000 enregistrements avec dossier médical associé, lui aussi ayant été numérisé. Les principales pathologies étaient la Sclérose Latérale Amyotrophique (SLA), la maladie de Parkinson et les pathologies extra-pyramidales (Huntington), les ataxies cérébelleuses, la maladie de Friedreich, les dysarthries vasculaires (accidents vasculaires cérébraux).

31 L'accès aux données acoustiques dans leur contexte clinique permet de développer les connaissances pour la caractérisation perceptive et acoustique des dysphonies et dysarthries (Crevier-Buchman, 2005 ; Crevier-Buchman, 2019).

2.4. Le corpus C2SI (Carcinologic Speech Severity Index) de patients post cancer des VADS (Toulouse)

32 Dans le cadre du projet C2SI (Carcinologic Speech Severity Index) financé par l'INCA, le service d'oncoréhabillitation de l'Oncopole à Toulouse a collecté une série d'enregistrements de la parole de patients post cancer des VADS. Un tel corpus est utilisé pour mesurer l'impact du cancer de la cavité buccale et pharyngée sur la production de la parole (Woisard *et al.*, 2020). Il permettra à terme d'évaluer la qualité de vie des patients après le traitement. Le corpus est composé d'enregistrements audio de 134 sessions avec les métadonnées associées (taille et localisation de la tumeur, traitement...). Plusieurs niveaux d'intelligibilité et de compréhensibilité des fonctions langagières ont été évalués : pseudomots (Ghio *et al.*, 2018), phrases, fonctions prosodiques (Nocaudie *et al.*, 2018), lecture de texte. Des taux d'évaluation perceptive de jurys naïfs et d'experts sont en cours d'élaboration ainsi que des analyses automatiques (Laaridh *et al.*, 2018). Il est destiné à fournir aux orthophonistes et aux médecins des outils objectifs, qui prennent en compte l'intelligibilité des patients ayant reçu un traitement anticancéreux (chirurgie et/ou radiothérapie et/ou chimiothérapie). Ce corpus C2SI sera mis à la disposition de la communauté scientifique par le biais du groupe d'intérêt scientifique Parolothèque[6].

2.5. Le corpus Paroles disfluentes du laboratoire Praxiling (Montpellier)

33 Si le bégaiement fait l'objet d'un grand nombre d'études dans les pays anglo-saxons, cela est moins vrai dans le monde francophone. Cette situation peut s'expliquer par le fait qu'il s'agit d'un trouble ne touchant qu'environ 1% de la population (Didirkova, 2016) et que, contrairement aux autres altérations de la parole, aucune structure ne centralise sa prise en charge. C'est donc pour favoriser la recherche sur le bégaiement que le laboratoire Praxiling, aidé par un financement du consortium CORLI, a proposé le corpus intitulé *Paroles disfluentes* (Didirkova *et al.*, 2017).

34 *Paroles disfluentes* se compose de 38 fichiers audio au format *.wav*, chacun des fichiers étant accompagné d'une transcription au format *Textgrid*. Ces enregistrements portent

sur 17 locuteurs adultes qui bégaient, autrement dit 13 hommes et 4 femmes âgés en moyenne de 32 ans (écarts-type : 11 ans).

35 Les données proviennent de plusieurs études qui portaient sur les situations de double tâche en parole bègue ou encore sur la description articulatoire et acoustique du bégaiement. En conséquence, les tâches enregistrées consistaient en de la lecture, de la parole spontanée et des résumés de contes pour enfants.

36 Le corpus est actuellement disponible, après demande, sur la plateforme Ortolang, à l'adresse suivante : https://www.ortolang.fr/market/corpora/paroles-disfluentes. Il sera complété par les données actuellement recueillies dans le cadre du programme ANR BENEPHIDIRE (ANR-18-CE36-0008, responsable : Fabrice Hirsch), une fois celui-ci arrivé à son terme.

2.6. Le corpus de l'Institut de Phonétique de Strasbourg

37 En 2014, suite à un projet financé (initiative d'excellence « projets attractivité » - porteuse : Béatrice Vaxelaire), l'équipe Parole et Cognition (Institut de Phonétique de Strasbourg) a procédé à l'inventaire et l'archivage systématique de ses corpus en parole pathologique pour les verser sur la Plateforme Unistra de Linguistique et de Phonétique Clinique. Ce projet au long cours repose sur des investigations comparatives entre des productions linguistiques normales, produites par des sujets sains, et des productions déviantes, émises par des patients atteints de diverses pathologies de la parole et du langage. Plusieurs corpus enregistrés pour des travaux de thèse ont ainsi pu être sauvegardés en procédant à un travail d'anonymisation et d'archivage systématique des métadonnées pour permettre leur exploitation en recherche. À ce jour, les données acoustiques annotées (à l'aide de textgrid) versées dans cette collection en parole pathologique concernent les personnes qui bégaient (Hirsch, 2007), les productions d'enfants porteurs de fentes labio-palatines (Béchet, 2011), la voix après thyroidectomie (Fauth, 2012 et Xiu, 2018), et les productions de patients glossectomisés (Zaouali, 2019).

3. Préconisation d'organisation en base de données

3.1. Les concepts de base de données

38 L'intérêt principal de la construction d'une base de données (BD) regroupant différentes ressources est de pérenniser les informations et de permettre à un groupe de travail d'échanger et d'améliorer progressivement la base de données via un serveur de données. Le modèle de la BD s'est appuyé sur une analyse fonctionnelle. Il a été réalisé dans un environnement clinique, basé sur des corpus empiriques, tels que ceux présentés précédemment.

39 Si les concepts autour des bases de données sont familiers aux informaticiens, ce n'est pas le cas pour les non-spécialistes[7] (Ghio *et al.*, 2012). Il est fréquent de lire qu'une collection d'enregistrements sonores est une BD. Pourtant, une BD se distingue d'un corpus ou d'une collection par une structuration et une organisation cohérente qui est régie par un modèle qui peut être partagé par un groupe de personnes et qui peut être stocké dans un support informatique. Une telle architecture organisée facilite la sélection des données, en utilisant des critères précis. Cela nous amène à aborder la

notion de système de gestion de base de données (SGBD) qui doit gérer ces concepts selon le modèle de données. Ce système a pour objet de (1) clarifier le partage des données entre les différents utilisateurs, (2) protéger la confidentialité des données si nécessaire, (3) répondre aux requêtes et (4) fournir différentes langues d'accès en fonction du profil de l'utilisateur.

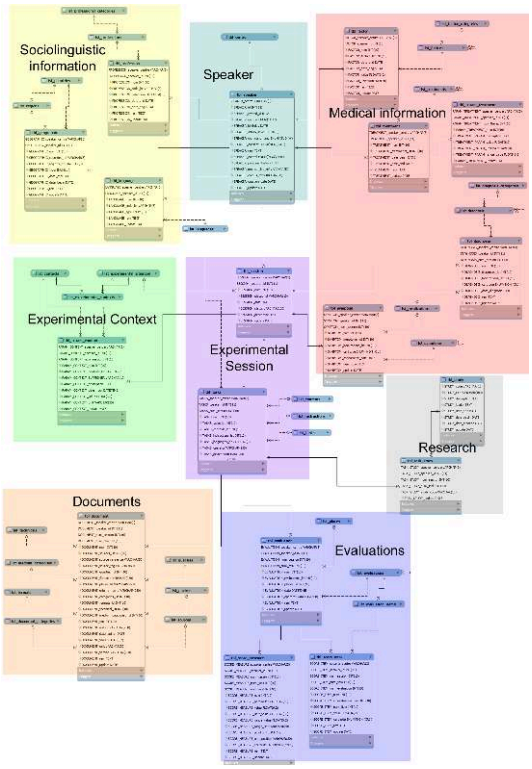
40 Dans notre cas, nous avons opté pour un modèle relationnel, considéré comme le modèle de base de données le plus simple et le plus élégant. Sa simplicité vient de l'organisation tabulaire des données, atomistique et minimaliste, rendant l'architecture des données intuitive, les éléments de chaque table étant liés par des relations.

41 Le modèle conceptuel adopté et le choix des tables ont émergé par l'usage en concertation avec les cliniciens. Le choix des métas données sélectionnées est le résultat de l'informatisation des usages en dossier papier et des données de gestion. En effet, la plupart des études cliniques impose la tenue d'un cahier d'observation (Case Report Form, CRF[8]) qui rassemble les données individuelles de chaque patient. Traditionnellement, ce CRF est sous la forme de fiches au format papier remplies au moment de l'examen. L'exploitation ensuite de ces données nécessite une saisie informatisée manuelle des informations. Cet usage est, d'une part, chronophage mais peut aussi être source d'erreurs ou peut engendrer des pertes de données. L'informatisation d'un CRF en e-CRF (format électronique) est donc une bonne solution et passe par une organisation en base de données suffisamment généraliste pour s'adapter à des contextes différents (usages variables selon les centres hospitaliers, les services, les objectifs des études). Un autre exemple d'adaptation aux nécessités cliniques réside dans les relations que nous avons mises en place entre les tables de la base de données. Si par exemple, en neurologie, un diagnostic peut être directement mis en lien avec un locuteur (ce qui se traduit en termes de base de données par une jonction des tables 'tbl_speaker' et 'tbl_diagnoses', Figure 1), cette propriété est directement liée au fait qu'une maladie neurodégénérative telle que la maladie de Parkinson une fois diagnostiquée ne disparaitra pas et restera « attachée » à la personne. En revanche, un diagnostic ORL tel que des nodules ou des polypes pourra être vrai au temps *T* mais ce diagnostic pourra ensuite disparaitre si ces nodules se résorbent suite à un repos vocal, après une rééducation orthophonique ou une chirurgie (voir corpus MTO). Attacher un diagnostic directement à un locuteur n'est donc pas adapté. C'est ainsi qu'est apparue la nécessité d'introduire la notion de 'session' (Figure 1) qui représente l'état du patient à un temps *T* (celui de la visite médicale durant laquelle il est enregistré). Cette session est alors associée au locuteur. Les diagnostics et les symptômes ne sont pas directement associés au locuteur mais ils le sont à la session. Lors des requêtes qui permettent de sélectionner des enregistrements répondant à un critère, on ne cherchera pas directement les locuteurs qui répondent à un choix (ex : locutrices porteuses de nodules) mais on s'attachera à rechercher les sessions attachées à des locutrices durant lesquelles un diagnostic de nodules a été porté.

42 Comme le résume la figure 1, la BD est composée d'une cinquantaine de tables qui donnent les informations civiles (date et lieu de naissance, lieu de résidence...), sociolinguistiques (langue maternelle, professions...), médicales (symptômes, diagnostic, traitements habituels), sessions d'enregistrement (date, lieu, opérateur...), contexte d'enregistrement (avec/sans traitement), protocole expérimental (tâche, instructions au participant, contenu linguistique, dispositifs utilisés...), documents

associés (données sonores) et éventuellement des notes d'évaluation (perceptive, instrumentale...).

Figure 1. Modèle conceptuel de données préconisé pour la gestion de base de données de parole pathologique issu de la mise en conformité RGPD du modèle proposé par (Ghio *et al.*, 2012). Afin de garantir la sécurité et la protection des données personnelles, cette mise en conformité a nécessité la suppression de la table 'tbl_medical_history' qui contenait des informations trop personnelles ainsi que la table 'tbl_civil' qui permettait de lever l'anonymat.



43 Pour standardiser certaines informations et pour suivre les bonnes pratiques de la constitution d'un CRF[8], un ensemble de listes permet de collecter des informations normalisées telles que professions, langues, pays/régions, symptômes, thérapies, diagnostics, facteurs de risque, localisation des pathologies, contextes expérimentaux, méthodes d'évaluation... L'intérêt de ces listes fermées vise à éviter l'augmentation des dénominations pour une même terminologie. Par exemple, un diagnostic de « maladie de Parkinson » peut être noté comme PD, maladie de Parkinson, Parkinson, Park... Il est plus efficace de proposer une liste fermée où « maladie de Parkinson » est assigné comme diagnostic n° 11 (valeur arbitraire fixe). Tous les patients parkinsoniens seront alors référés à cet identifiant. Un avantage important d'un tel codage est la compatibilité internationale. En effet, si tous les éléments des listes sont traduits, l'ensemble du contenu de la base de données est opérationnel et adapté à la nouvelle langue. Une liste des diagnostics habituels relatifs aux troubles de la voix et de la parole est proposée mais cette liste peut être augmentée en fonction des besoins. Des détails sont disponibles dans (Ghio *et al.*, 2012).

3.2. Le stockage des informations cliniques

44 Comme mentionné ci-dessus, l'étude de la parole pathologique nécessite spécifiquement la collecte et le stockage d'informations précises – personnelles et

médicales – relatives aux locuteurs et aux contextes médicaux dans lesquels elles ont été enregistrées. Ces informations sont essentielles pour pouvoir espérer appréhender correctement les multiples sources de variation à la fois linguistique et clinique que l'on retrouve dans la parole pathologique. Par exemple, étudier la variation induite par la maladie de Parkinson n'est pas possible si le chercheur n'a accès qu'aux enregistrements sonores. Il aura besoin, en plus des informations socio-démographiques traditionnelles, de connaitre l'ancienneté de la maladie de chaque locuteur, l'évaluation motrice effectuée par le neurologue (UPDRS), la sévérité de la dysarthrie, le traitement médicamenteux usuel, l'état médicamenteux au moment de l'enregistrement (délai de la dernière prise de médicament)... En effet, seule la connaissance de ces informations permettra de comparer ce qui est comparable (voir corpus AHN ou CCM). À l'inverse, essayer de dégager de l'information linguistique sur un corpus de locuteurs parkinsoniens dont on ignore l'ancienneté de la maladie, le traitement thérapeutique, l'état moteur... ne permettra en aucune façon d'expliquer la/les variation(s) observées dans ce type de parole. Il en est de même pour l'étude des productions langagières de patients ayant un handicap de parole post cancer de la cavité buccale et de l'oro-pharynx (voir corpus C2SI ou Strasbourg). Les variations impactant l'intelligibilité de ces patients doivent être mises en perspective avec la localisation précise de la tumeur, la taille de la tumeur, le geste chirurgical pratiqué, la possible reconstruction, la dose de radio et/ou chimiothérapie, le délai depuis la chirurgie... Bref, le stockage des informations cliniques sous une forme organisée en base de données est incontournable pour l'étude de la variation physiopathologique dans la parole. Par conséquent, il est recommandé d'obtenir un maximum d'informations sur les aspects suivants :

3.2.1. Informations sociolinguistiques

- Sexe, année de naissance
- Lieux de naissance et de résidences successives
- Langue maternelle et langues parlées
- Statut professionnel ou niveau d'études
- Main dominante
- Remarques générales (par ex. Difficulté de lecture, analphabétisme, surdité, port de lunettes, bégaiement, pratique du chant, niveau de sport...)

45 Pour illustrer l'importance de ce type d'information : nous avons été confrontés dans certains cas de dysarthrie, au phénomène d'élision du /r/ qui peut être similaire à celui que l'on retrouve dans les accents « créoles » ; seule la connaissance des lieux de naissance et de résidence du locuteur nous a permis de savoir si ce phénomène était pathologique ou sociolinguistique.

3.2.2. Informations médicales générales

46 Il est conseillé de compléter les informations des locuteurs par des commentaires sur l'état du patient.

- Suivi médical (ex : état psychologique, syndrome dépressif, hallucinations, troubles du comportement et/ou cognitifs, autres troubles)
- Traitements thérapeutiques (ex : chirurgie, médecine, orthophonie, électrophysiologie...)

- Facteurs pouvant provoquer ou favoriser la maladie (ex : Alcool et tabac, pollution sonore et atmosphérique, allergie respiratoire, abus vocal, stress, intubation...).

47 Ces informations permettent d'inclure ou d'exclure des patients en fonction des finalités de l'étude.

3.2.3. Informations symptomatiques

48 Les symptômes du patient et les signes observés par le médecin doivent également être indiqués (ex : dysphonie, dysarthrie, tremblements, fuite glottique, trouble cognitif, trouble du traitement auditif), ainsi que la date à laquelle ils ont été observés, donnant éventuellement une indication de certitude et si nécessaire, la localisation anatomique (par exemple mâchoires, membre supérieur droit / gauche-supérieur, membre droit / gauche-inférieur...).

3.2.4. Informations pathologiques

49 Les diagnostics posés par le médecin (ex : nodule, polype, maladie de Parkinson, maladie de Charcot, traumatisme crânien...), la date de leur établissement, avec une indication possible de certitude, et si nécessaire, leur anatomie la localisation (par exemple, à gauche/droite, lobe frontal, lobe pariétal...) doit également être indiquée.

3.2.5. Informations contextuelles

50 Le contexte clinique dans lequel le patient est enregistré représente une information importante à collecter afin d'effectuer des analyses rigoureuses et significatives. Voici quelques-uns des contextes expérimentaux à collecter :

- Statut pharmacologique (par exemple, la date et l'heure du dernier médicament, la nature et la quantité habituelle du médicament et la médication pendant l'enregistrement du patient...)
- État de neurostimulation activé et désactivé
- Situation pré/post-opératoire (par exemple la date de l'opération...)
- Informations complémentaires (par exemple « le patient a une bronchite, porte un corset, a eu son médicament il y a 4 heures, a oublié ses lunettes... »)

3.2.6. Protocole

51 En raison de la diversité des caractéristiques acoustiques liées aux troubles de la voix et de la parole, nous proposons de distinguer d'une part les tâches d'élocution vocale produites par les locuteurs (ex : chant, voyelle soutenue, lecture d'un texte, répétition, description d'image, discours spontané...) et d'autre part le contenu linguistique (ex : voyelle /a/, jours de la semaine, Rainbow Passage...). De plus, il est intéressant et pertinent de stocker les instructions données pour les différentes tâches (ex : rapide, lent, cadence habituelle...). Si l'utilisation d'un système de gestion de base de données est recommandée pour la traçabilité et l'exploitation des métadonnées, la standardisation du protocole qui vise à collecter des données sonores ou physiologiques est difficilement compatible avec le contexte clinique. En fait, un protocole complet comprenant la production de voyelles tenues, d'efforts vocaux, de phrases, de répétitions, de textes lus, de parole spontanée, est difficilement réalisable en raison de la fatigabilité causée par de trop longs efforts. Il est donc préférable d'adapter les

tâches d'élocution à l'état de dysfonctionnement du locuteur. Par exemple, une étude sur la nasalité est particulièrement intéressante dans le cas de la dysarthrie paralytique en raison de l'immobilité du voile du palais mais moins importante dans la maladie de Parkinson pour laquelle les exercices phonatoires peuvent être préférés en raison de l'hypophonie.

3.2.7. Document

52 Dans la table « document », les noms de fichiers d'enregistrement, les caractéristiques (par exemple la fréquence d'échantillonnage, le format, la qualité d'un fichier de signal...), ainsi que le nom de l'expérimentateur sont stockés. Un document peut être un fichier signal mais également être composé de transcriptions orthographiques, d'annotations ou d'images associées à la tâche. Les questions de format de fichiers de ce type de données sont détaillées dans Ghio *et al.* (2012). Concernant les noms de fichiers, il n'est pas pertinent de coder toutes les informations lors de la dénomination d'un fichier car il peut générer des noms extrêmement complexes. Il faut cependant normaliser ces noms et obtenir une dénomination unique, non ambiguë et si possible universelle non dépendante de la spécificité du corpus. Le principe que nous proposons est le suivant :

53 (FRA-)MTO-000052-03-L02.wav dont le nom est suffisamment informatif pour déduire que les données proviennent du corpus MTO, locuteur n° 52. Le document est le fichier wave relatif à la troisième session d'enregistrement de ce locuteur, exécutant la deuxième tâche de lecture (L) pendant la session. Pour obtenir des informations sur le contexte, la pathologie, l'âge, l'origine géographique, la catégorie socioprofessionnelle, les traitements, le contexte pharmacologique, il est nécessaire d'interroger la base de données créée à cet effet.

3.2.8. Évaluations

54 Les évaluations perceptives ou instrumentales sont des ressources informatives qui doivent être stockées.

3.3. Aspects juridiques

55 Nous ne nous intéressons pas ici à la protection des bases de données c.-à-d. ni au droit d'auteur reposant sur la structure originale de la base de données, ni au droit *sui generis* destiné à protéger l'investissement financier, matériel et humain entrepris par le producteur de la base de données.

56 Comme décrit plus haut, une base de données dédiée aux troubles de la voix et de la parole est amenée à être alimentée par différentes sources[9] de données de natures diverses (clinique, sonore, physiologique...) collectées auprès de patients et de sujets contrôles. Cette approche multicentrique oblige le producteur de base de données à s'assurer au respect des conditions de collecte et de cessation des différents corpus qui la composent. Cette obligation est d'autant plus essentielle que la plupart des données des corpus sont dites « sensibles »[10] car elles informent sur l'état de santé des patients enregistrés durant leur parcours de soins courants.

3.3.1. Quelles obligations à l'égard du producteur de données ?

57 La livraison d'un corpus en vue d'être migré dans une base de données, doit être formalisée sous une forme contractuelle ou conventionnelle, entre le service hospitalier collecteur des données et le producteur de la base de données. Cet acte juridique bilatéral permet d'organiser la cession des droits de propriété intellectuelle, et plus particulièrement des droits d'auteur, dans le respect des exigences légales. Il permet également de définir la gestion des aspects relatifs à la confidentialité et à la gouvernance des données.

58 Tel que décrit au paragraphe « Une reconnaissance interdisciplinaire des contributions » et même si cela n'est pas une obligation à respecter, nous préconisons fortement de conditionner la diffusion d'un jeu de données pour un requérant avec l'accord du producteur des données et la délivrance d'une licence d'utilisation. Cette licence permet de fixer les modalités spécifiques de la mise à disposition des données, ainsi que l'obligation de citer l'investigateur clinique dans tous les travaux publiés et fondés sur les données collectées sous son égide.

3.3.2. Peut-on anonymiser une BD dédiée aux troubles de la voix et de la parole ?

59 De par le caractère « sensible » des données collectées dans les services hospitaliers, l'anonymisation des données doit s'imposer avant leur migration dans la base de données et ce, même si le consentement éclairé est recueilli auprès des patients en préambule de la passation hospitalière. L'objectif est que le producteur de la base de données n'ait aucune possibilité de pouvoir identifier nominativement les personnes enregistrées dans celle-ci. Le renforcement de la protection des données et des personnes (RGPD, loi Jardé) nous impose d'adopter des solutions limitant l'usage de texte libre pour préférer l'utilisation de listes à choix forcé, de cases à cocher, ce qui impacte directement le modèle conceptuel de la base de données.

60 Afin de ne pas diffuser d'informations permettant d'identifier (in)directement les personnes, différentes techniques d'anonymisation peuvent être appliquées sur la base de données :

- Hachage du nom et du prénom (algorithme SHA-2 *i.e.* Secure Hash Algorithm)[11]
- Minimisation des données (suppression de l'anamnèse, de l'histoire personnelle...)
- Généralisation des lieux de résidence au département, de la date de naissance à l'année, de la profession à la catégorie socio-professionnelle...

61 Cependant certaines de ces techniques présentent des limites ne permettant pas d'atteindre une anonymisation complète des données. Un risque résiduel pour les personnes concernées peut encore exister. Tout d'abord, l'anonymat par hachage ne peut être garanti de façon absolue en raison de risques d'attaque par « force brute » consistant à tester toutes les solutions possibles pour établir une table de correspondance. Ensuite, l'anonymisation des données sonores ne peut être envisagée dans le cadre de la recherche scientifique. En effet, même si la CNIL définit la voix[12] comme une donnée personnelle permettant d'identifier indirectement une personne physique, le bruitage ou la déformation des enregistrements sonores entraverait considérablement toute recherche en linguistique et plus particulièrement en phonétique clinique. C'est la raison pour laquelle le terme pseudonymisation[13] est plus approprié dans ce cadre (Lalain *et al.*, 2020).

3.3.3. Comment gérer l'accès et la diffusion des données ?

62 La mise en œuvre d'une base de données scientifiques répond aux besoins de la recherche en offrant une meilleure mutualisation et partage des connaissances destinées aux chercheurs pour réaliser leurs travaux. Un chercheur doit donc pouvoir accéder à la base de données, la consulter et requérir la mise à disposition de collections de données extraites par des interrogations multicritères. Pour cela, l'implémentation d'une stratégie d'accès sécurisé et de confidentialité est indispensable pour

63 [1] garantir le contrôle par modération des utilisateurs autorisés à accéder à la base de données et

64 [2] définir les conditions d'accès aux données qui dépendent de leur caractère « sensible » et des objectifs des utilisateurs.

65 Nous préconisons de soumettre la gestion des comptes de connexion à la base de données, à une modération « scientifique » conduite par le producteur de la base de données c.-à-d. un contrôle préalable visant à s'assurer de la validité de l'identité déclarée et des finalités de recherche. Cette modération pourra éventuellement associer le producteur des données.

66 Concernant les demandes de mise à disposition de collections de données, celles-ci doivent être encadrées par un contrat ou une licence qui, en raison du caractère « sensible » des données, doit être adapté aux risques d'utilisations des données non conformes à la loi et à l'éthique. Concrètement, pour établir ce type de licence, nous nous sommes appuyés sur la licence du Speech and Language Data Repository (Figure 2) adaptée au contexte clinique en intégrant notamment la reconnaissance des producteurs hospitaliers. Hormis le fait de garantir l'intégrité, la sécurité et la confidentialité des données, le demandeur devra aussi s'engager à ne pas « dés-anonymiser » les données transmises, ni à les diffuser.

Figure 2. Licence du Speech and Language Data Repository (SLDR/Ortolang ; www.sldr.fr/)

Licence SLDR

Préambule

Cette licence est destinée à garantir :

1. l'intégrité des données diffusées par le SLDR ;
2. le suivi des utilisations de ces données, au bénéfice de leurs créateurs.

Termes de la licence

En téléchargeant du site du SLDR un corpus, une ressource linguistique ou un outil (ci-après désignés comme « la ressource »), l'utilisateur s'engage sur les points suivants :

1. ne pas distribuer la ressource à de tierces personnes ; la diffusion se fait uniquement, de manière nominative, par téléchargement du site SLDR, après accord sur la présente licence ;
2. signaler clairement l'origine (identifiant unique "sldrxxxxxx") dans toute publication ou utilisation (même non commerciale) de la ressource, et reproduire les références bibliographiques des articles mentionnés sur la fiche descriptive, le cas échéant ;
3. informer le SLDR de cette publication ou utilisation. Cette information sera saisie dans l'espace de l'utilisateur (lien "Déposer -> Publication") ;
4. informer le SLDR de tout enrichissement de la ressource et mettre cet enrichissement au service de la communauté scientifique, via le SLDR, en accord avec les auteurs de la ressource.

67 De plus, en cas d'un transfert de données hors de l'Union Européenne[14], il faudra prévoir un encadrement contractuel spécifique si le pays de destination n'offre pas « *un niveau de protection adéquate reconnu par l'UE* »[15].

4. Conclusion

68 Bien que l'état de l'art fasse apparaître d'importantes avancées dans la compréhension des mécanismes de production de la voix et de la parole, il existe un besoin continu d'améliorer l'analyse des locuteurs sains et pathologiques. Une collecte de données à grande échelle est nécessaire pour prendre en compte la variabilité « normale » et « pathologique » de la parole. Une base de données structurée de la parole pathologique représente un jalon dans la progression vers ces objectifs.

69 Une telle base de données peut fournir aux développeurs et aux utilisateurs de logiciels cliniques des données de référence pour former la base sur laquelle différentes méthodes peuvent être comparées. Les bases de données ont été au cœur du développement des dispositifs automatiques de reconnaissance de la parole et des locuteurs. Une base de données des troubles de la parole peut permettre de fournir un élan similaire pour les applications cliniques.

70 À ce jour, il existe une réalisation technique développée au Laboratoire Parole et Langage à Aix-en-Provence baptisée Speedi DB[16] (speech disorders database). La genèse de ce projet est détaillée dans Ghio *et al.* (2006). On y trouve notamment les difficultés rencontrées dans le rassemblement des données et les arbitrages qui ont dû être fait. Ce serveur de base de données intègre pour le moment les corpus français AHN, MTO et CCM décrits précédemment. Une interface utilisateur permet de faire des requêtes complexes telles que « je cherche les extraits de lecture de la chèvre de monsieur Seguin de locuteurs masculins de plus de 60 ans, francophones natifs, droitiers, atteint de la maladie de Parkinson ». Si la vocation de ce serveur de base de données de parole pathologique n'a pas vocation à accueillir toutes les données, il peut servir de modèle de référence pour des initiatives pouvant revêtir une couverture nationale. À ce propos, cet outil a été utilisé pour les projets ANR DESPHO-APADY (2009-2012), TYPALOC (2012-2015) et RUGBI (2019-2023).

71 À l'image de ce qui s'est fait dans divers autres pays, il serait important que la communauté française, voire francophone, se mobilise de façon fédératrice pour se doter de bases de données de parole pathologique permettant aux neurologues, ORL, phoniatres, orthophonistes, phonéticiens et informaticiens de la parole de faire progresser les connaissances, les procédures d'évaluations ou les technologies vocales adaptées au handicap. L'implication de la communauté française dans une dynamique européenne telle que DELAD[17] (« Database Enterprise for Language And speech Disorders ») serait aussi la bienvenue.

BIBLIOGRAPHIE

Bechet M. (2011). *Perturbation de la production des occlusives chez des locuteurs présentant une division palatine ou labio-palatine*, Thèse de doctorat, Univ. Strasbourg.

Crevier-Buchman L. (2005). « La modélisation de la parole normale ». In Ozsancak C., Auzou P. (éd.), *Les troubles de la parole et de la déglutition dans la maladie de Parkinson*, Solal, 63-93.

Crevier-Buchman L. (2019). « Clinical Illustrations of Voice Quality ». In Esling J.H., Moisik S.R. (éd.), *Voice Quality The Laryngeal Articulator Model*, Cambridge University Press.

Didirkova I. (2016). *Parole, langues et disfluences : une étude linguistique et phonétique du bégaiement.* Thèse de Doctorat, Univ. Montpellier.

Didirkova I., Hirsch F. & Luxardo G. (2017). « Paroles disfluentes : corpus de parole produite par des personnes qui bégaient », *Colloque Corpus oraux, corpus écrits : pratiques croisées*. Montpellier.

Fauth C. (2012). *Perturbation de la production de la parole suite à une opération de la glande thyroïde,* Thèse de doctorat, Univ. Strasbourg.

Fougeron C., Crevier-Buchman L., Fredouille C., Ghio A., Meunier C., Chevrie-Muller C. *et al.* (2010). « Developping an acoustic-phonetic characterisation of dysarthric speech in French ». *Proceed. LREC*, 2831-2838.

Ghio A., Teston B., Viallet F., Jankowski L., Purson A. *et al.* (2006). « Corpus de parole pathologique, état d'avancement et enjeux méthodologiques », *TIPA*, Laboratoire Parole et Langage, 25 : 109-126.

Ghio A., Pouchoulin G., Teston B., Pinto S., Fredouille C., De Looze C., Robert D., Viallet F. & Giovanni A. (2012). « How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers ? », *Speech Communication*, 54(5) : 664-679.

Ghio A., Lalain M., Giusti L., Pouchoulin G., Robert D. *et al.* (2018). « Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique », *JEP*, Aix-en-Provence, France, 285-293.

Hirano M. (1981). *Clinical Examination of Voice*. Springer Verlag.

Hirsch F. (2007). *Le bégaiement : Perturbation de l'organisation temporelle de la parole et conséquences spectrales*, Thèse de doctorat, Univ. Strasbourg.

Jan M. (2007). « L'évaluation instrumentale de la dysarthrie en France », In *Les dysarthries,* Auzou P., Rolland-Monnoury V., Pinto S., Ozsancak C. (éd.), Solal, 119- 122.

Laaridh I., Fredouille C., Ghio A., Lalain M., Woisard V. (2018). « Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers », *Interspeech* : 2943-2947.

Lalain M., Pouchoulin G. (2020). « De la protection des données à la protection de la personne : Réflexions sur l'impact des nouvelles réglementations sur la collecte des corpus », *Revue Corpus*, 22.

Mattei A., Desuter G., Roux M., Lee B.-J., Louges M.-A., ... A. Giovanni, (2018). « International consensus (ICON) on basic voice assessment for unilateral vocal fold paralysis », *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 135(1S).

Nocaudie O., Astésano C., Ghio A., Lalain M., Woisard V. (2018). « Évaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx », *JEP*, Aix-en-Provence, 196-204.

Parsa V., Donald G.J. (2001). « Acoustic Discrimination of Pathological Voice : Sustained Vowels Versus Continuous Speech », *J Speech Hear Res*. 44(2): 327-339.

Pinto S., Ghio A., Teston B., Viallet F. (2010). « La dysarthrie au cours de la maladie de Parkinson. Histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie », *Revue Neurologique*, 166(10) : 800-810.

Schuepbach W.M.M., Rau J., Knudsen K., Volkmann J., Krack P., Timmermann L., Hälbig, ... Deuschl G. (2013). « Neurostimulation for Parkinson's Disease with Early Motor Complications », *New England Journal of Medicine*, 368(7) : 610-622.

Schuller B.W. (2015). « Speech Analysis in the Big Data Era ». In : Král P., Matoušek V. (éd.), *Text, Speech, and Dialogue*. TSD 2015. Lecture Notes in Computer Science, vol. 9302. Springer.

Woisard V., Astésano C., Balaguer M., Farinas J., Fredouille C. *et al.* (2020). « C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers », *Language Resources and Evaluation*, Springer Verlag.

Xiu N. (2018). *Perturbation de la production de la parole chez le patient atteint d'une paralysie laryngée : Données acoustiques et aérodynamiques*, Thèse de doctorat, Univ. Strasbourg.

Zaouali H. (2019). *Etude acoustique de la production de la parole chez des patients glossectomisés*, Thèse de doctorat, Univ. Strasbourg.

NOTES

1. Loi n° 2016-1321 du 7 octobre 2016 Pour une République numérique. [En ligne]
2. Axe 2 du CoSO : structuration et ouverture « autant que possible » des données de la recherche. [En ligne]
3. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016. [En ligne]
4. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. [En ligne]
5. Partie extraite de notre article en anglais (Ghio *et al*., 2012).
6. https://www.irit.fr/parolotheque/
7. Partie extraite de notre article en anglais (Ghio *et al*., 2012).
8. https://www.rechercheclinique pariscentre.fr/wp-content/uploads/2016/12/DIU-CP-CRF-09-12-2016-partie-1-et-2-S.-Makhlouf.pdf
9. Une source de données correspond à un corpus produit par un service hospitalier appelé « producteur de données » ; chaque corpus est identifié à un centre au sein de la base de données.
10. Catégorie particulière des données personnelles. [En ligne]
11. Le hachage n'est pas réversible c.-à-d. la reconstitution de l'entrée hachée n'est plus possible. Néanmoins, il est utilisé pour l'appariement de données entre une nouvelle source et la base de données, et ainsi éviter les doublons qui seraient susceptibles de constituer un biais scientifique.
12. Définition de la donnée personnelle. [En ligne]
13. « *La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces.* », https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles. Par exemple, il est aisé de comprendre que l'identification d'un notaire de village dont on connait l'âge et dont on a un extrait de parole est possible en regroupant ces sources d'information.

14. Attention, une simple consultation des données à distance constitue un transfert !
15. La liste des pays offrant une protection adéquate figure sur le site internet de la CNIL. [En ligne]
16. https://speedi-db.lpl-aix.fr/physio
17. https://delad.ruhosting.nl

RÉSUMÉS

L'étude des troubles de la voix et de la parole est sortie du cadre de la recherche clinique. Par l'observation des dysfonctionnements, les chercheurs non cliniciens confrontent les résultats de leur recherche établis sur des corpus de parole « normale » à des situations de dysfonctionnement. Le défi est immense car le cadre « pathologique » induit une variation considérable dans ses manifestations de surface. Toute généralisation à une population clinique particulière nécessite l'observation d'un grand nombre de patients du fait de la très forte variation interindividuelle. Il est donc important de capitaliser et mutualiser les enregistrements existants. Or pour être utilisables, ces enregistrements doivent répondre à de fortes exigences. Le maillon faible reste la normalisation et la structuration des données sur les locuteurs et leurs productions langagières. Concrètement, si les données sonores sont souvent accessibles, elles ne présentent au final aucun intérêt si les liens entre les enregistrements et les caractéristiques cliniques du locuteur sont rompus ou erronés. L'objectif de ce travail est de présenter différentes actions de terrain et de proposer des recommandations pour la structuration des données sonores, physiologiques et cliniques dans le cas de corpus de parole issue de patients atteints de troubles de la voix et de la parole.

Voice and speech disorders are now studied beyond the framework of clinical research. By observing dysfunctions, non-clinical researchers compare the results of their research established on "normal" speech with dysfunctional situations. The challenge is important because the "pathological" framework induces a great variation in its audible manifestations. Any generalization to a particular clinical population requires the observation of a large number of patients due to the very strong interindividual variation. It is therefore important to capitalize and share existing records. However, to be usable, these recordings require a high level of quality. The main problem remains the standardization and structuring of data on speakers and speech productions. Concretely, if the audio data is accessible, it is useless if the links between the recordings and the speaker's clinical characteristics are broken or erroneous. The objective of this work is to present various actions in the field and to propose recommendations for the structuring of sound, physiological and clinical data in the case of speech corpus from patients with voice and speech disorders.

INDEX

Keywords : clinical phonetics, voice speech disorders, database
Mots-clés : phonétique clinique, troubles de la voix, troubles de la parole, base de données

AUTEURS

ALAIN GHIO

Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France

GILLES POUCHOULIN

Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France

FRANÇOIS VIALLET

Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
Service de neurologie, Centre Hospitalier du Pays d'Aix, France

ANTOINE GIOVANNI

Aix-Marseille Univ, CNRS, LPL, UMR 7309, Aix-en-Provence, France
CHU Timone-Conception, APHM, Marseille, France

VIRGINIE WOISARD

CHU Toulouse, Oncopole Toulouse, France

LISE CREVIER-BUCHMAN

Laboratoire de Phonétique et Phonologie, UMR7018, Hôpital Foch, Paris, France

FABRICE HIRSCH

Praxiling, Université de Montpellier 3, France

CAMILLE FAUTH

LILPA, Université de Strasbourg, France

CORINNE FREDOUILLE

LIA, Université d'Avignon, France