Reconnaître les personnes à leur voix : définition d'un cadre scientifique pour garantir la fiabilité des résultats d'une comparaison de voix dans le cadre criminalistique

Anais Chanclu

AVIGNON UNIVERSITÉ

THÈSE DE DOCTORAT

École doctorale n° 536
Agrosciences et Sciences

Spécialité : Informatique

Laboratoire Informatique d'Avignon (UPR 4128)

Présentée par
Anaïs Chanclu

RECONNAÎTRE LES PERSONNES À LEUR VOIX

Définition d'un cadre scientifique pour garantir la fiabilité des résultats d'une comparaison de voix dans le cadre criminalistique

---

Soutenue publiquement le **11 décembre 2023** devant le jury composé de :

M^{me}	Martine Adda-Decker	Professeure	LPP	Paris	Rapporteuse
M.	Julien Pinquier	Maître de conférences	IRIT	Toulouse	Rapporteur
M^{me}	Christine Meunier	Professeure	LPL	Aix-en-Provence	Examinatrice
M.	Jean-François Bonastre	Professeur	LIA	Avignon	Directeur de thèse

Remerciements

Le travail de recherche que vous vous apprêtez à lire est le fruit de nombreuses rencontres et collaborations. Par conséquent, je tiens à remercier toutes les personnes qui m'ont accompagnée tout au long de ce travail, de près ou de loin, et qui ont fait de ces années de thèse une expérience riche et passionnante.

Je tiens tout particulièrement à remercier Jean-François Bonastre, mon directeur de thèse, de m'avoir fait confiance et m'avoir accompagnée pour mener à bien ce travail de recherche. J'ai beaucoup appris à tes côtés, et je te remercie pour ta disponibilité, ta patience et tes conseils avisés. Je te remercie également d'avoir cru en moi en toutes circonstances, et de m'avoir soutenue dans les moments difficiles. Je remercie aussi mes rapporteurs, Martine Adda-Decker et Julien Pinquier, d'avoir accepté d'évaluer ce travail de thèse. Vos remarques et vos suggestions m'ont permis d'améliorer ce manuscrit, et je vous en suis très reconnaissante. Un grand merci également à Christine Meunier pour avoir accepté de présider ce jury. Vous avez fait de ma soutenance un moment très agréable et très enrichissant. Je vous en remercie. Ces remerciements ne reflètent pas l'ensemble de ma gratitude, et je vous remercie tous, vraiment, pour votre bienveillance et vos conseils.

Je n'aurais pas pu faire cette thèse sans le consortium du projet Voxcrim, et je remercie tout particulièrement les personnes qui ont contribué à ce projet. Merci à vous tous pour votre accueil, votre bonne humeur et vos conseils : Corinne Fredouille, Emmanuel Ferragne, Cédric Gendrot, Cécile Fougeron, Agnès Delaborde, Sophie Vasseur. Je remercie plus particulièrement Alain Ghio et Olivier Galibert pour leurs suivis, conseils et expertise ainsi que Gabriele Chignoli et Benjamin O'Brien pour leur collaboration. Merci aussi à Laurianne Georgeton et Christophe Stecoli qui m'ont fait découvrir le monde de la police.

Remerciements

Bien sûr, je tiens à remercier mes collègues du LIA pour leur accueil chaleureux et leur bonne humeur en particulier Natalia Tomashenko et Titouan Parcollet ainsi que Driss Matrouf, ainsi que tous les doctorants que j'ai eu l'occasion d'y croiser ! Un énorme merci aussi à Teva Merlin qui m'a dépannée à plusieurs reprises, à la vitesse de la lumière, et à Michèle Mannen pour sa disponibilité, sa convivialité, et pour m'avoir accompagnée dans les différentes démarches administratives. Merci également à toi, Imen Ben Amor, d'avoir fait de notre collaboration un moment agréable et enrichissant.

Sans oublier les personnes sans qui je ne serai pas là aujourd'hui. Un énorme merci à Juliette Kahn et Camille Dutrey pour m'avoir donné le goût de faire de la recherche, pour votre rigueur et votre bienveillance. Je remercie aussi Moez Ajili pour son aide et sa disponibilité.

Pour votre soutien inconditionnel, votre écoute, votre bienveillance et votre bonne humeur, je remercie Émilie, Fanny, Evan, Frédérique, Marc, Laurent, Adrien, Rémi, Christopher, Jérémy. Merci à vous tous pour ces moments de partage, de rigolade, de fausses notes, et pour m'avoir aidée à garder le sourire dans les moments difficiles... et même après ! Merci également à Maureen Rivat pour son soutien et son aide depuis mon diagnostic. Bien sûr, je remercie ma famille qui m'a menée là où j'en suis, et aussi Éric et Michelle pour leur soutien.

Enfin, merci à toi, Fabien. Merci pour tout l'amour que tu me portes, merci d'avoir été à mes côtés pendant toutes ces années.

Résumé

Dans le domaine criminalistique, les pratiques de comparaison de voix ont beaucoup évolué ces dernières décennies. Cependant, elles manquent de standardisation et les résultats obtenus peuvent être décriés dans les tribunaux. L'objectif de ce travail de recherche est de définir un cadre scientifique permettant d'évaluer la fiabilité des résultats d'une comparaison de voix dans un cadre criminalistique. Dans un premier temps, nous présentons les bases de données FABIOLE 2 et PTSVOX, spécialement conçues pour répondre aux problématiques de comparaison de voix en criminalistique, ce que ne font pas les bases de données existantes telles que Voxceleb. La base de données FABIOLE 2 se concentre sur la variabilité intralocuteur, souvent négligée dans les études de comparaison de voix, tandis que la base de données PTSVOX se rapproche des conditions réelles. Dans un second temps, nous introduisons le concept de *box-rule*, qui est un cadre scientifique regroupant un ensemble de conditions dans lesquelles la fiabilité d'une comparaison de voix est connue. Pour définir ce cadre, l'influence de certains facteurs sur la performance d'une comparaison de voix est étudiée en utilisant la base de données FABIOLE 2. Nos résultats montrent que la durée des enregistrements, la différence de durée entre les enregistrements, le genre, l'âge, et l'écart temporel entre les enregistrements à comparer influent sur la performance d'un système de comparaison de voix. Ensuite, nous étudions la perception humaine des locuteurs par le biais d'une tâche de regroupement de voix en locuteurs. Cette approche permet de s'intéresser à la reconnaissance humaine des locuteurs sans passer par des tests binaires, et ainsi limiter les biais qu'ils peuvent engendrer. Nous montrons que la performance des auditeurs n'est pas homogène et qu'elle est influencée par la langue maternelle. Enfin, nous nous intéressons à la caractérisation des voix, qui peut ajouter une information supplémentaire à une comparaison de voix. Nous nous concentrons sur la détection du type de phonation et nous

proposons un nouveau système, reposant sur une architecture neuronale profonde et d'une cascade de classifieurs binaires. Ce système obtient de très bons résultats sur les voyelles prépausales de PTSVOX, en milieu fermé, ce qui nous encourage à généraliser ce système à l'ensemble des phonèmes sonores. Les résultats de la généralisation montrent une hétérogénéité entre les locuteurs, mais aussi entre les femmes et les hommes dans les tendances langagières. Pour conclure, ce travail marque une première étape dans la définition d'un cadre scientifique pour la criminalistique et explore plusieurs pistes pour garantir la fiabilité des résultats d'une comparaison de voix. Il ouvre également de nouvelles perspectives tant pour compléter la *box-rule* que pour caractériser les voix et nous espérons qu'il mènera à des pratiques de comparaison de voix standardisées pour la criminalistique.

Mots-clefs comparaison de voix, criminalistique, reconnaissance du locuteur, apprentissage automatique, apprentissage profond, fiabilité, standardisation, perception du locuteur, caractérisation des voix

Abstract

In the field of forensic science, voice comparison practices have evolved significantly in recent decades. However, they lack standardization, and the results obtained can be criticised in courtrooms. The purpose of this research is to define a scientific framework to assess the reliability of voice comparison results in a forensic context. Firstly, we introduce the FABIOLE 2 and PTSVOX databases which are specifically designed to address voice comparison issues in forensics unlike databases such as Voxceleb. The FABIOLE 2 database focuses on intraspeaker variability, often overlooked in voice comparison studies, while the PTSVOX database approaches real-life conditions. Secondly, we introduce the concept of box-rule, which is a scientific framework comprising a set of conditions under which the reliability of voice comparison is known. To define this framework, we study the influence of specific factors on voice comparison performance using the FABIOLE 2 database. Our results show that recording duration, difference of duration between recordings, gender, age, and the time gap between recordings influence the performance of a voice comparison system. Next, we study human perception of speakers through a task of speaker clustering. This approach allows us to focus on human speaker recognition without resorting to binary tests, thus limiting the biases they may introduce. We show that listener performance is not homogeneous and is influenced by the native language. Finally, we focus on voice characterization, which can provide additional information for voice comparison. We focus on detecting phonation types and offer a new system based on a deep neural architecture and a cascade of binary classifiers. This system achieves very good results on pre-pausal vowels from PTSVOX, in a closed environment, which encourages us to generalise this system to all voiced phonemes. Generalization results show heterogeneity among speakers, as well as between women and men. In conclusion, this work marks a first step in defining a scientific framework for forensics

and explores several avenues to ensure the reliability of voice comparison results. It also opens up new perspectives for both completing the box-rule and characterizing voices and we hope it will lead to standardise forensic voice comparison practices.

Keywords forensic voice comparison, speaker recognition, machine learning, deep learning, reliability, standardisation, speaker perception, voice characterization

Table des matières

Introduction

Dans le domaine de la criminalistique, il est commun de devoir effectuer une identification d'individu. Pour cela, plusieurs procédés peuvent être utilisés tels que l'identification par empreinte digitale, l'identification par empreinte génétique, l'identification par reconnaissance faciale. L'identification des personnes peut également passer par la voix à travers un procédé appelé « comparaison de voix ». La comparaison de voix peut être utilisée dans les cas d'écoutes téléphoniques, de sonorisations, d'enregistrements d'appels aux services de secours, *etc*. Les cas pouvant nécessiter une comparaison de voix comprennent les enregistrements de conversations téléphoniques, les enregistrements vidéo et audio récupérés sur les lieux de crimes, les enregistrements d'appel aux services de secours ainsi que les enregistrements provenant de dispositifs de surveillance. Sa faisabilité dépend de nombreux facteurs tels que la qualité de l'enregistrement, la durée de l'enregistrement, la présence de bruit, *etc*.

Dans l'histoire de la comparaison de voix, plusieurs méthodes ont été utilisées. Ces méthodes peuvent relever de la perception humaine, de l'expertise phonétique ou de l'utilisation de systèmes automatiques. La perception humaine est la méthode la plus ancienne, mais cette dernière a montré ses limites et n'est plus utilisée de nos jours. Aujourd'hui, la comparaison de voix peut passer par une analyse phonétique des signaux de parole souvent en complément des méthodes automatiques qui s'appuient sur des techniques de traitement du signal et d'apprentissage automatique.

Ce travail de thèse se concentre sur la comparaison de voix automatique et plus particulièrement de la garantie de la fiabilité des résultats. En effet, contrairement à d'autres méthodes d'identification, la comparaison de voix automatique n'est pas encore encadrée

par des normes pour garantir la fiabilité des résultats. L'objectif de ce travail est d'apporter des éléments pour garantir la fiabilité d'une comparaison de voix automatique. Pour cela, nous adoptons quatre approches différentes :

Utilisation de données adaptées pour la comparaison de voix Les données utilisées pour la comparaison de voix ont plusieurs objectifs. Elles servent à l'apprentissage des modèles, à la calibration des scores ou encore à la validation des systèmes de comparaison. Chacun des objectifs nécessitant un type précis de données, les jeux de données utilisés doivent être différents.

Influence de différents facteurs sur la performance de la comparaison de voix De nombreux facteurs entrent peuvent modifier les résultats d'une comparaison de voix. Ces derniers sont dépendants du ou des locuteurs, des conditions d'enregistrement, des techniques de comparaison de voix utilisées, *etc.* Nous étudions l'influence de certains de ces facteurs sur la performance de la comparaison de voix. Pour cela, nous faisons varier un ou plusieurs facteurs et nous observons l'impact de ces variations sur la performance de la comparaison de voix.

Classification humaine des enregistrements en locuteurs Les êtres humains sont en mesure d'identifier les locuteurs au quotidien notamment à chaque fois qu'ils répondent au téléphone. Nous étudions la capacité de perception de l'identité locuteur par des êtres humains et par des machines. Pour cela, nous mettons en place une expérience perceptive dans laquelle les participants doivent regrouper des enregistrements en locuteurs. Cela permet d'observer les différences de classification des enregistrements en locuteurs entre les participants. Nous comparons les résultats obtenus par les participants avec ceux obtenus par des machines.

Caractérisation des voix Il nous est possible de trouver certaines voix plus caractéristiques que d'autres. Cependant, ces caractéristiques ne sont pas toujours faciles à définir de façon factuelle. Nous nous intéressons à la caractérisation de la voix humaine en étudiant la détection automatique du type de phonation. Avec cette expérience, nous cherchons à déterminer si la voix d'un locuteur est modale, craquée ou soufflée.

Cadre de la thèse

Ce travail de recherche s'inscrit dans le projet Voxcrim [1], financé par l'Agence nationale de la recherche (ANR) et porté par un consortium composé du Service national de Police scientifique (SNPS), de l'Institut de recherche criminelle de la Gendarmerie nationale (IRCGN), du Laboratoire Informatique d'Avignon (LIA), du Laboratoire Parole et Langage (LPL), du Laboratoire Phonétique et Phonologie (LPP) et du Laboratoire nationale de métrologie et d'essais (LNE). Les objectifs du projet Voxcrim sont :

Établir un cadre de mise en œuvre d'une comparaison de voix En reprenant le concept de « *box-rule* », soit un ensemble de règles qui conditionnent la mise en œuvre de la comparaison de voix, il s'agit de définir ces conditions aussi bien au niveau perceptif grâce aux travaux des laboratoires dédiés à la parole, le LPL et le LPP, qu'au niveau automatique suivant les travaux du LIA.

Mettre en place un processus qualité Les partenaires de Voxcrim soutenus par le LNE, également partenaire, visent à mettre en place une démarche normative de type ISO/CEI 17025, avec l'ambition de diffuser le processus qualité auprès de la communauté de la parole, mais aussi après des acteurs du système judiciaire (police, magistrats).

Organisation du manuscrit

Ce manuscrit de thèse comprend 8 chapitres organisés comme suit :

Reconnaître les locuteurs et comparer les voix

De la voix aux locuteurs Reconnaître les personnes à leur voix peut sembler trivial, mais il s'agit d'un processus complexe. Ce chapitre présente les différentes informations que le locuteur laisse transparaître sur lui-même à travers sa voix, telles que son genre, son âge ou

1. VoxCrim, ANR-17-CE39-0016 - https://voxcrim.univ-avignon.fr

encore son orientation sexuelle. Les processus cognitifs dans la reconnaissance humaine des locuteurs sont également abordés. Enfin, nous présentons les différentes méthodes utilisées pour la reconnaissance automatique des locuteurs.

Contexte de la comparaison de voix en criminalistique La comparaison de voix consiste à comparer deux enregistrements de voix afin de déterminer s'ils ont été produits par la même personne. Cependant, de nombreux facteurs de variabilité peuvent être présents dans les pièces à comparer, ce qui peut influer le résultat d'une comparaison de voix. Ce chapitre présente le statut de la voix en criminalistique ainsi que les cadres scientifiques et légaux de la comparaison de voix.

Contributions

Quelles données pour la comparaison de voix en criminalistique ? Les bases de données VoxCeleb sont aujourd'hui très largement utilisées dans l'état de l'art de la reconnaissance automatique du locuteur pour l'apprentissage des modèles et leur test. VoxCeleb contient au total plus de 7000 locuteurs et plus de 2000 heures de parole. Néanmoins, cette base de données connaît quelques limites, notamment dans la représentation de certains sous-groupes de locuteurs. Les bases de données FABIOLE 2 et PTSVOX, introduites dans le cadre du projet Voxcrim, sont également présentées spécifiquement pour la comparaison de voix. FABIOLE 2 est une base de données de parole spontanée de près de 400 locuteurs rencontrés dans plusieurs contextes différents dans l'objectif de représenter au mieux la variabilité intra-locuteur. La base de données PTSVOX, quant à elle, reprend les protocoles de prélèvements de voix du SNPS afin d'avoir une base de données plus proche des conditions réelles de comparaison de voix.

Premiers éléments de la *box-rule* Ce chapitre présente les premiers éléments de la *box-rule*. Pour cela, nous utilisons la base de données FABIOLE 2. En faisant varier des paramètres définis, nous observons l'influence de ces variations sur la performance d'un système de comparaison de voix basé sur l'architecture *Emphasized Channel Attention, Propagation and Aggregation* (ECAPA)-*Time Delay Neural Network* (TDNN) et entraîné sur les données VoxCeleb Ces paramètres sont la durée des enregistrements, la différence de durée entre

les enregistrements, le genre, l'âge, le locuteur cible et l'écart temporel entre les enregistrements à comparer.

Évaluer la capacité humaine à regrouper les locuteurs Dans le cadre du challenge VoicePrivacy 2020, nous avons mis en place une expérience perceptive dans laquelle les participants doivent regrouper des enregistrements en locuteurs. L'objectif est, d'une part, d'évaluer la tâche de regroupement en locuteurs comparativement à une comparaison binaire et, d'autre part, d'évaluer la capacité humaine à regrouper des enregistrements. Les données utilisées proviennent de la base de données VCTK qui contient des phrases lues en langue anglaise. Les participants sont invités à regrouper les enregistrements en locuteurs. Leur performance est comparée à celle d'un système de comparaison de voix basé sur l'architecture *x-vector*.

Vers la caractérisation de la voix : le cas du type de phonation Dans l'objectif de caractériser les locuteurs, une expérience de détection du type de phonation est mise en place. Cette expérience est focalisée sur les voyelles prépausales d'un jeu de données de PTSVOX avant d'être généralisée sur l'ensemble des phonèmes voisés de ce jeu de données. Le système présenté repose sur *Problem-Agnostic Speech Encoder* (PASE) pour l'extraction des paramètres et un Perceptron multi-couches (MLP) pour la classification. Ce système est comparé à un système de référence basé sur les coefficients *Mel-frequency cepstral coefficient* (MFCC) et un machine à support de vecteur (SVM).

Conclusions et perspectives

Apports de la thèse et Perspectives Cette partie reprend et conclut les conclusions de ce travail de thèse en mettant l'accent sur les apports et en présentant les perspectives envisagées.

Première partie

Reconnaître les locuteurs et comparer les voix

1 | De la voix aux locuteurs

Résumé : *La voix est un signal produit par la vibration des plis vocaux, modulé par les cavités crâniennes, nasales et buccales. Outre l'aspect linguistique, certaines informations sur l'identité de la personne qui parle, son état émotionnel, son état de santé, son âge ou encore son sexe peuvent être déduites par la voix. Ce chapitre explore les différentes informations qui peuvent en être extraites ainsi que les processus cognitifs mis en place dans la reconnaissance des locuteurs.*

Sommaire

1.1 Reconnaître les personnes à leur voix

La parole est un moyen de communication. Elle permet de délivrer des informations d'une personne à l'autre par la voie acoustique. Ce message est porté par la voix qui, outre l'information linguistique, peut également véhiculer des informations sur la personne qui parle. Ces informations peuvent être perçues de manière consciente ou inconsciente et donnent parfois lieu à des jugements de valeur. Il nous est également possible d'identifier des personnes à partir de leur voix, bien qu'il nous arrive de nous tromper. Dans ce chapitre, nous abordons la parole et la voix humaine, puis nous nous intéressons aux processus cognitifs impliqués dans la perception humaine des locuteurs.

1.1.1 Langage, langue, parole et voix

Les termes langage, langue, parole et voix sont utilisés dans le langage courant. Parfois, les acceptions se recoupent entre les termes. Par exemple, le « traitement automatique du langage naturel » et le « traitement automatique des langues » désignent la discipline qui consiste à créer des outils pour le traitement de données linguistiques (traduction automatique, recherche et extraction d'informations, *etc.*). Une confusion entre « parole » et « voix » s'observe dans certaines locutions telles que la « reconnaissance vocale », utilisée abusivement pour les techniques utilisées pour la transcription automatique de la parole [1].

Langage En 1916, Saussure, considéré comme le père de la linguistique, définit le langage comme étant un système de signes où chaque signe est composé d'un concept, d'une part, que Saussure nomme le signifié et d'une forme physique, d'autre part, qui prend le nom de signifiant (Saussure, 1916). La notion de signes est reprise par Martinet lorsqu'il évoque la double articulation du langage : tout énoncé linguistique peut être segmenté à deux niveaux. Le premier niveau correspond à une suite de signes selon la définition saussurienne ; le second niveau permet de former les signes en utilisant des unités plus fines : les phonèmes (Martinet, 1960).

1. Le terme exact est « reconnaissance automatique de la parole ».

Langue Les signes du langage sont conditionnés par la langue, que Saussure définit comme étant un « produit social de la faculté de langage et un ensemble de conventions nécessaires, adoptées par le corps social » (Saussure, 1916). Martinet conçoit la langue comme un « instrument de communication selon lequel l'expérience humaine s'analyse, différemment dans chaque communauté » (Martinet, 1960).

Parole Le *Cours de linguistique générale* de Saussure introduit plusieurs dichotomies dont l'une d'entre elle est la distinction entre langue et parole. Selon Saussure, la parole est un « acte individuel » du langage (Saussure, 1916). Dans ce manuscrit, le terme parole est utilisé pour désigner le contenu linguistique délivré sous forme acoustique.

Voix La voix est « le support acoustique de la parole » (Cornut, 2019). Il s'agit de la combinaison du son produit par la vibration des plis vocaux[2] situés au niveau du larynx et des résonances de ce son dans les cavités buccales, nasales et crâniennes. Dans ce travail de recherche, le terme « voix » correspond au transport de la communication parlée.

1.1.2 Que dit notre voix lorsque nous parlons ?

1.1.2.1 La voix : un phénomène issu de la phonation et de l'articulation

La production de la voix est très fortement liée à l'activité respiratoire. En effet, l'essentiel des productions linguistiques orales survient durant l'expiration. Avec l'expulsion de l'air des poumons, deux étapes permettent la production de la parole : la phonation et l'articulation.

Phonation Après avoir été expulsé des poumons, l'air circule dans le larynx où sont situés les plis vocaux. L'anatomie du larynx est présentée dans la figure 1.1. Les plis vocaux peuvent être amenés à se resserrer et à vibrer, provoquant ainsi la voix laryngée. Ce phénomène intervient dans la production des phonèmes voisés. *A contrario*, lors de la production des phonèmes non voisés, les plis vocaux sont relâchés et l'air y circule librement. La liste des phonèmes voisés et non voisés en français est présentée dans le tableau 1.1. La voix laryngée

2. Le langage courant utilise la locution « cordes vocales ». situés au niveau du larynx et des résonances de ce son dans les cavités buccales, nasales et crâniennes.

constitue la source des phonèmes voisés. Elle est composée d'une fréquence fondamentale, notée F_0, qui correspond à la fréquence de vibrations des plis vocaux et d'harmoniques (multiples entiers de la F_0). D'un point de vue perceptif, la F_0 correspond à la hauteur tonale de voix : plus la F_0 est élevée, plus la voix est perçue comme étant aiguë. Les variations de la F_0 forment l'intonation d'un énoncé. L'amplitude des vibrations des plis vocaux coïncide avec l'intensité, soit la puissance sonore, mesurée en décibels. Plus l'amplitude des vibrations des plis vocaux est élevée, plus la voix sera perçue comme étant forte. Le mode de vibration des plis vocaux permet de définir le type de phonation. Il s'agit d'un descripteur de qualité de voix qui se traduit par un continuum articulatoire basé sur l'aperture des plis vocaux, allant d'une glotte fermée pour la voix craquée à une glotte ouverte pour la voix soufflée (Gordon et Ladefoged, 2001). Le type de phonation normal, dit modal, se trouve entre la voix craquée et la voix soufflée. Les types de phonation non-modaux entraînent des perturbations de la F_0.

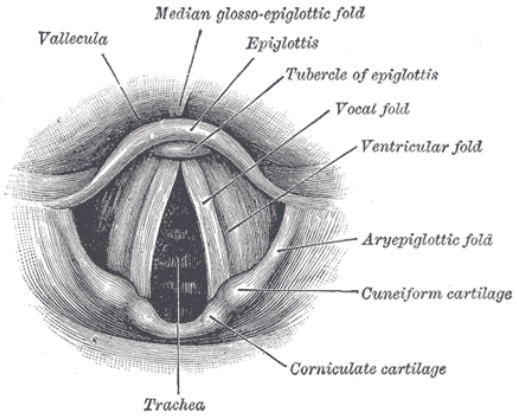


Figure 1.1 – *Anatomie du larynx, faisant apparaître les plis vocaux* (vocal fold) *(Gray, 1918)*

Articulation Après la phonation, la voix laryngée est filtrée par les résonateurs (caisses de résonance) que sont les cavités pharyngales, buccales, nasales et crâniennes. Les résonateurs amplifient certaines harmoniques et en diminuent d'autres. Les harmoniques amplifiées constituent les fréquences de résonance. Leur amplitude décroît à mesure que les fréquences augmentent. Ces dernières sont à l'origine des formants (également nommés F_1, F_2, F_3 et F_4). L'organisation des formants les uns par rapport aux autres permet de distinguer les phonèmes voisés. Les phonèmes non voisés sont caractérisés par l'absence de for-

	Phonèmes voisés	Phonèmes non voisés
Voyelles	/i/, /u/, /e/, /ø/, /ɛ/, /ɛː/, /œ/, /a/, /ə/, /ʉ/, /o/, /ɔ/, /ɑ/, /ɛ̃/, /œ̃/, /ɑ̃/, /ɔ̃/	
Consonnes	/b/, /d/, /g/, /v/, /z/, /ʒ/, /ʁ/, /m/, /n/, /ɲ/, /ŋ/, /l/, /ɥ/, /w/, /j/	/p/, /t/, /k/, /f/, /s/, /ʃ/

Tableau 1.1 – *Phonèmes voisés et non voisés en français*

mants et peuvent être distingués par leur mode d'articulation et leur lieu d'articulation. Les consonnes occlusives [p], [t], [k] sont caractérisées par une fermeture complète du conduit vocal à des lieux différents (respectivement au niveau des lèvres, de la langue contre les dents et de la langue contre le voile du palais). Les consonnes [f], [s], [ʃ] sont caractérisées par une fermeture incomplète du conduit vocal au niveau des lèvres, des alvéoles et entre les alvéoles et le palais dur. Il s'agit de consonnes fricatives.

1.1.2.2 La dimension sociale de la parole

Saussure indique que la parole est un acte individuel qui intervient dans un contexte social (Saussure, 1916). Ce contexte social conditionne les productions linguistiques, ce qui peut être à l'origine de schibboleths ou de sociolectes tels que les argots. Les schibboleths définissent les usages linguistiques exclusifs à un groupe social, et s'observent particulièrement à l'oral, permettant ainsi de distinguer les personnes qui font partie du groupe social des personnes qui n'en font pas partie. Les sociolectes sont également des marqueurs sociaux qui peuvent être oraux ou écrits, mais qui ne possèdent pas la même exclusivité que les schibboleths. Les sociolectes peuvent évoluer vers d'autres sociolectes.

En français, l'un des argots les plus connus est le largonji. Attesté à partir de la fin du XVIII^{e} siècle, cet argot consiste à remplacer la consonne initiale par la lettre « l » et d'ajouter un suffixe au choix. Le terme « largonji » est une déformation en largonji du terme « argot ». Le largonji a permis aux bouchers de mettre en place le loucherbem [3], argot qui est toujours utilisé dans le milieu boucher (Plénat, 1985). Le largonji et le loucherbem sont à l'origine

3. Également orthographié « loucherbem », « louchebem » ou encore « louch'bem ».

de termes et de locutions qui sont aujourd'hui employées dans le langage courant telles que « loufoque » (loucherbem de « fou ») ou encore « en loucedé » qui signifie « en douce ». Le largonji et le loucherbem ne sont pas les seuls argots du français. Depuis la fin du XX^{e} siècle, le verlan s'est propagé des banlieues au reste de la France. Le principe du verlan est d'inverser les syllabes ou les phonèmes d'un mot. Certains termes verlan sont restés dans le langage courant tels que « loubard », « ripou ». Le verlan a également donné naissance au veul qui se veut une reverlanisation, c'est-à-dire être le « verlan du verlan ». Alors que le terme « femme » donne « meuf » en verlan, en veul, « meuf » devient « feumeu ».

Le contexte social conditionne également le système phonologique des locuteurs, et par conséquent leur prononciation. Cela est particulièrement flagrant chez les locuteurs de langues différentes, mais cela est également observé chez les locuteurs d'une même langue. À titre d'exemple, [la.pɔm] dans la partie nord de la France correspond au fruit du pommier. Une personne du sud-ouest peut se représenter [la.pɔm] comme étant le fruit du pommier (la pomme) ou bien comme étant la face intérieure de la main (la paume). Seul le contexte permet d'indiquer s'il s'agit du fruit ou de la face intérieure de la main. Le système phonologique représente l'ensemble des phonèmes qu'un individu emploie dans ses productions langagières orales. Les phonèmes représentent l'unité minimale significative qui permet de faire la distinction entre les mots du lexique.

Définition · Phonème

Si deux sons apparaissent exactement dans la même position phonique et ne peuvent se substituer l'un à l'autre sans modifier la signification des mots, ou sans que le mot ne devienne méconnaissable, alors, ces sons sont des réalisations de deux phonèmes.

– Nikolaï Sergueïevitch Troubetzkoï (Troubetzkoï, 1949)

Le système phonologique est variable entre les locuteurs d'une même langue. Les locuteurs du français québécois possèdent un système phonologique plus fourni que les locuteurs du français dit « standard [4] ». Cela s'observe surtout au niveau des voyelles où les

4. Il s'agit du français de Paris, celui des médias nationaux.

distinctions entre les phonèmes tendent à se perdre en français standard, mais perdurent dans d'autres pays et régions de la francophonie. Les phonèmes qui tendent à être confondus par une partie de la population francophone sont présentés dans le tableau 1.2. À titre d'exemple, dans certaines variétés du français, les locuteurs ne font plus la distinction entre les sons ɛ̃ et œ̃, ce qui a pour effet que la prononciation des termes « brin » et « brun » est strictement identiques à leurs oreilles. Dans d'autres variétés du français, à l'instar du français québécois, cette distinction est maintenue.

[bʁɛ̃] brin	**[bʁœ̃]** brun
[pat] patte	**[pɑt]** pâte
[koʃe] cocher	**[koʃɛ]** cochet
[ʒø] jeu	**[ʒə]** je
[mɛtʁ] mètre	**[mɛːtʁ]** maître
[manje] manier	**[maɲe]** magner

Tableau 1.2 – *Distinctions phonémiques qui tendent à se perdre dans certaines régions de la francophonie. La perte de ces distinctions amène à prononcer les mots de la colonne de droite comme celle de la colonne de gauche.*

Depuis les travaux de Labov, la sociolinguistique s'applique à étudier l'influence du contexte social sur les productions langagières des individus (Labov, 1972). Dans la parole, le contexte social est particulièrement flagrant dans l'accent et les termes employés. En effet, ces derniers donnent des informations sur la classe sociale ou le lieu d'origine d'un individu. Les différences sociolinguistiques peuvent également être la source de « débats ». L'exemple qui marque les jeunes générations de Françaises et Français est l'appellation d'une viennoiserie composée de pâte feuilletée et de barres de chocolat. Dans le sud-ouest de la France, cette viennoiserie se nomme « chocolatine » alors que dans le reste de la France, elle est majoritairement appelée « pain au chocolat ». Bien que le nom de la viennoiserie, sur

fond de chauvinisme régional et de second degré, « divise la France en deux camps » selon Jeannot Nymouce (Mello et Nymouce, 2021), il n'en reste pas moins que l'affrontement est généralement de l'ordre de la taquinerie et très générateur de mèmes Internet. Le nom de la viennoiserie a cependant été source de violences. En effet, le 9 février 2020, à Mimizan, dans les Landes, où il est de coutume de dire « chocolatine », une bagarre a éclaté après que des personnes du Nord ont appelé la viennoiserie « pain au chocolat ». Une des personnes en cause a écopé de 4 mois de prison ferme [5].

Figure 1.2 – *Exemple de photomontage humoristique sur la dénomination « pain au chocolat »/« chocolatine »*

1.1.3 Qui entendons-nous lorsqu'une personne nous parle ?

Outre le contenu linguistique délivré par la parole, les êtres humains sont en mesure de percevoir des informations supplémentaires sur leur interlocuteur. La perception de la voix est problématisée dès les années 1930 à partir des voix entendues à la radio. En effet, Pear,

5. Article de 20 Minutes daté du 17 février 2022 : https://www.20minutes.fr/faits_divers/3237227-20220217-landes-defenseur-chocolatine-frappe-violemment-adeptes-pain-chocolat

Herzog ainsi que Cantril et Allport se sont penchés sur la « personnalité » des voix entendues à la radio (Cantril et Allport, 1935 ; Herzog, 1933 ; Pear, 1931). Pear aborde non seulement les « stéréotypes vocaux » et l'influence que ces derniers ont sur les auditeurs, mais également sur la perception du « sexe » et de l'âge.

1.1.3.1 Accent et origines géographiques et sociales

> **Définition · Accent**
>
> 1. Augmentation de l'intensité ou élévation de la hauteur de la voix, qui met en relief telle syllabe ou telle articulation d'un mot ou d'un groupe de mots.
> 2. Manière particulière de placer l'accent, et par extension, ensemble des traits de prononciation qui s'écartent de la prononciation considérée comme normale et révèlent l'appartenance d'une personne à un pays, une province, un milieu déterminés.
>
> – Trésor de la langue française informatisé (TLFi) [6]

La voix peut laisser transparaître l'origine d'un individu par le biais de l'accent. En français, l'accent de référence est l'accent « standard » que nous retrouvons particulièrement dans la moitié nord de la France avec une implantation plus ou moins forte en fonction des régions (Boughton et Armstrong, 1998). La perception de l'accent est le principal indice pour indiquer une origine géographique. Ainsi l'accent permet d'indiquer si un individu francophone vient du sud-ouest de la France, des Antilles, de Polynésie, de Suisse, de Belgique, du Canada, d'Haïti, de Madagascar ou encore du Cambodge. Certains accents sont mieux reconnus que d'autres. En effet, Woehrling et Boula de Mareüil publient en 2006 une étude sur la perception et la classification des accents régionaux du français dans laquelle deux groupes de personnes – un groupe originaire d'Île-de-France et un groupe originaire de Marseille –, sont soumis à la reconnaissance auditive d'accents provenant de Vendée, de Normandie, du canton de Vaud en Suisse, du Pays basque, du Languedoc et de Provence (Woehrling et

6. https://www.cnrtl.fr/definition/accent

Boula de Mareüil, 2006). Les résultats montrent que l'accent suisse obtient 72 % de bonne reconnaissance. En revanche, il existe une confusion entre l'accent languedocien, l'accent provençal et, dans une bien moindre mesure, l'accent basque. Une autre confusion est observée entre l'accent vendéen et l'accent normand. Selon Ikeno et Hansen, notre accent a un effet sur la perception que nous avons des accents (Ikeno et Hansen, 2007). L'étude conclut que bien que l'accent permette d'indiquer l'origine d'un individu, nous ne sommes pas tous égaux dans la reconnaissance des différents accents.

L'accent peut également être un marqueur social. Les accents « bourgeois » et les accents dits « de banlieue » en sont des termes qui, bien que courant, sont assez difficiles à définir. Bien que ces deux accents fassent directement référence au statut social d'un individu, leur dénomination dans le langage courant laisse penser que l'accent « de banlieue » réfère à une aire géographique et non à la condition sociale des individus qui utilisent cet accent. La perception des accents en tant que marqueurs sociaux et des biais implicites que ces derniers peuvent induire. Dans leur étude, Pélissier et Ferragne ont procédé à une expérience dans laquelle ils ont enregistré les données électroencéphalographiques d'auditeurs soumis à des enregistrements prononcés par des personnes ayant un accent « bourgeois » parisien et des personnes ayant un accent dit « de banlieue » (Pélissier et Ferragne, 2022). Les enregistrements comprennent des phrases correspondant à des stéréotypes liés à l'un ou l'autre accent. Lorsqu'il y a incongruence entre l'accent entendu et la phrase prononcée, les chercheurs ont observé une réponse cérébrale négative environ 400 millisecondes après présentation des stimuli (N400). La N400 est notamment activée lorsque la perception du contenu langagier (écrit ou oral) est considérée comme « inattendu » par l'interlocuteur, ce qui met en évidence des biais implicites dans la perception de l'accent.

Cas réel · L'Éventreur du Yorkshire - Peter Sutcliffe

En Angleterre, la publication d'une enquête sur les dialectes anglais (Orton *et al.*, 1962–1971) a permis de déterminer précisément l'origine d'un homme par son accent. Entre 1975 et 1980, Peter Sutcliffe, surnommé « L'Éventreur du Yorkshire [7] », assassine 13 femmes et tente d'en tuer 9 autres dans le Yorkshire, au nord de l'Angleterre.

Entre 1978 et 1979, la police reçoit trois lettres et un enregistrement audio. L'expertise de l'enregistrement est effectuée par Jack Windsor Lewis, un phonéticien, et Stanley Ellis, un linguiste spécialiste des accents régionaux anglais. Ce dernier était l'un des principaux chercheurs lors de l'enquête sur les dialectes anglais et, par ailleurs, la première personne à fournir une expertise d'identification de voix dans les tribunaux anglais en 1967. Il leur est demandé d'indiquer l'origine géographique du locuteur. Lewis indique qu'il s'agit d'un accent du Tyneside, au nord-est de l'Angleterre. Fort de ses compétences et travaux, Ellis est plus précis et indique que l'accent du locuteur est typique du quartier Castletown à Sunderland, soit à environ 150 kilomètres de la zone des meurtres. Pour cette raison, Lewis et Ellis pensent que l'enregistrement audio est un canular, et recommandent de l'ignorer. La police fait fi de leur recommandation pour se concentrer sur cette nouvelle piste qui ne donnera aucun résultat.

Pendant ce temps, Sutcliffe tuera trois autres femmes. Le 2 janvier 1981, Sutcliffe est arrêté par la police à cause des fausses plaques d'immatriculation de sa voiture. Il est transféré dans un commissariat où il est interrogé sur l'affaire de « L'Éventreur du Yorkshire » avec lequel il partage beaucoup de similitudes physiques. Il avouera, malgré son accent du Yorkshire. L'enregistrement reçu par les policiers était un canular de John Samuel Humble, originaire de Sunderland, domicilié à quelques minutes à pied du quartier de Castletown. Ce dernier sera condamné en 2006 pour obstruction à la justice.

7. *The Yorkshire Ripper* en anglais

1.1.3.2 Discrimination à l'accent : la glottophobie

Définition · Glottophobie

Le terme « glottophobie » désigne les discriminations à prétexte linguistique et inclut le processus de stigmatisation qui conduit à ces discriminations.

– Philippe Blanchet (Blanchet, 2021)

Selon Blanchet, la glottophobie tient sa source de la normalisation du français, qui débute dès le XVI^{e} siècle et qui perdure depuis la création de l'Académie française. La volonté des grammairiens de l'Académie française était d'élaborer une norme du français qui soit proche du français de la Cour, mais également pour fixer l'orthographe et la grammaire pour « distinguer les lettrés des ignorants et des simples femmes » (Blanchet, 2016). La normalisation de la langue française a eu pour effet que le français dit « standard » soit la norme et que tout individu qui s'en écarte est susceptible d'être dévalorisé. Le 3 juillet 2020, le président de la République française, Emmanuel Macron, nomme son nouveau Premier ministre : Jean Castex. Originaire de Gascogne, le nouveau Premier ministre s'exprime oralement avec un accent méridional marqué, ce qui lui vaudra des remarques péjoratives. Dans les médias, les accents dits « régionaux » (par opposition à l'accent standard) sont très minoritaires. Les journalistes Aphatie et Feltin-Palas ont mené une enquête et concluent que 50 % de la population française indique « avoir un accent » et que 16 % de la population française estime avoir été victime de discrimination pour l'obtention d'un emploi ou d'un diplôme du fait de leur accent (Aphatie et Feltin-Palas, 2020). Le 3 décembre 2019, un projet de loi reconnaissant l'accent comment un critère de discrimination est déposé au parlement. Le 26 novembre 2020, l'Assemblée nationale adopte le texte en première lecture [8].

La glottophobie est un concept qui se retrouve au Canada sous le terme linguicisme ainsi qu'au Cameroun sous la locution hygiène verbale.

8. https://www.vie-publique.fr/loi/277433-proposition-de-loi-glottophobie-promouvoir-la-france-des-accents

1.1.3.3 La voix, une manifestation du genre ?

La voix est souvent considérée comme étant un caractère sexuel secondaire. Selon Childers et Wu, la F_0 constitue la principale différence entre une « voix de femme » et une « voix d'homme » (Childers et Wu, 1991). Coleman indique que cette différence de F_0 entre les femmes et les hommes s'explique par la physiologie des plis vocaux (Coleman, 1983). À l'adolescence, sous l'impulsion des hormones sexuelles, l'anatomie des femmes et des hommes se modifie, y compris le système phonatoire. À l'âge adulte, les plis vocaux des hommes sont plus larges et situés plus bas que chez les femmes. Il est admis que les femmes ont une voix plus aiguë que les hommes : elles possèdent donc une F_0 plus élevée. Lorsque Cartei *et al.* ont demandé à des personnes d'imiter des « voix de femme » et des « voix d'homme ». Celles-ci ont adopté une voix plus aiguë dans l'imitation des « voix de femme » et une voix plus grave pour imiter une « voix d'homme » (Cartei *et al.*, 2012). La perception du genre s'observe aussi dans la transidentité. La transition et le *passing* [9] passent également par la voix et plus particulièrement par des modifications de la F_0 (Spencer, 1988 ; Wolfe *et al.*, 1990).

En 1995, Traunmüller et Eriksson ont comparé des études sur les valeurs de fréquences fondamentales moyennes entre locuteurs, reprenant ainsi les valeurs de 10 études sur le sujet pour l'allemand, l'anglais (lecture et conversation), le français, le mandarin, le suédois et le wú [10]. Le résultat de leur comparaison est présenté dans le tableau 1.3. Dans chacune des études citées, et bien que certaines d'entre elles ne disposent pas d'un nombre représentatif de locuteurs, la fréquence fondamentale moyenne des femmes est plus élevée que la fréquence fondamentale moyenne des hommes, et ce, quels que soient la langue et le style de parole (parole spontanée, lecture ou théâtre). Selon Simpson, cette comparaison entre différentes langues parlées par des populations qui présentent des différences anatomiques notables montre que la différence entre une « voix de femme » et une « voix d'homme » est également le fruit de comportements sociaux acquis (Simpson, 2009), reprenant ainsi les conclusions de Boë *et al.* et Dolson (Boë *et al.*, 1975 ; Dolson, 1994).

La F_0 n'est pas le seul critère déterminant qui permet la distinction du genre dans la voix. Les études de Klatt montrent que les locuteurs adoptent un type de phonation différent. Les

9. Reconnaissance et identification de l'identité de genre par la société
10. Langue parlée dans l'est de la Chine

Étude	Langue	Locutrices et locuteurs	Tranche d'âge	F_0 μ	F_0 δ
(Rappaport, 1958)	Allemand	108 F 190 H		239 129	1.9 2.3
(Chevrie-Muller *et al.*, 1971)	Français	21 F 21 H	19-72 20-61	226 145	2.3 2.5
(Takefuta *et al.*, 2017)	Anglais	24 F 24 H		186 127	5.4 3.8
(Chen, 1974)	Mandarin	2 F 2 H	30-50 30-50	184 108	4.1 3.8
(Boë *et al.*, 1975)	Français	30 F 30 H		207 118	3 18
(Kitzing, 1979)	Suédois	141 F 51 H	21-70 21-70	193 110	2.7 3.0
(Johns-Lewis, 1986)	Anglais (Conversation) Anglais (Lecture) Anglais (Théâtre)	5 F 5 H 5 F 5 H 5 F 5 H	24-49 24-49 24-49 24-49 24-49 24-49	182 101 213 128 239 142	2.7 3.4 4.5 4.35 5.3 4.85
(Graddol, 1986)	Anglais (Lecture A) Anglais (Lecture B)	15 F 12 H 15 F 12 H	25-40 25-40 25-40 25-40	207 119 219 131	3.05 3.6 3.9 4.55
(Pegoraro Krook, 1988)	Suédois	467 F 198 H	20-89 20-79	188 113	2.55 2.65
(Rose, 1991)	Wú	3 F 4 H	30-64 25-62	187 170	3.8 4.1

Tableau 1.3 – *Fréquences fondamentales moyennes des locuteurs à partir des données de 10 études. D'après (Traunmüller et Eriksson, 1995)*

femmes tendent à avoir une voix plus soufflée que les hommes (Klatt, 1986 ; Klatt, 1987 ; Klatt et Klatt, 1990). Cette différence est observée en anglais américain (Podesva, 2011) et en fran-

çais (Pépiot, 2015). Néanmoins, des études plus récentes montrent que certaines locutrices adoptent plus volontiers une voix craquée comme marqueur socioculturel (Dallaston et Docherty, 2020 ; Hornibrook *et al.*, 2018 ; Yuasa, 2010). Les études sur la reconnaissance du genre à partir de la voix montrent que le genre est très bien reconnu par les auditeurs, et ce, aussi bien dans le flux de parole (Mullennix *et al.*, 1995 ; Pernet et Belin, 2012 ; Whiteside, 1998) que dans des vocalisations extralinguistiques telles que le rire ou les soupirs (Childers et Wu, 1991 ; Wu et Childers, 1991).

1.1.3.4 Quel âge parlez-vous ?

La voix varie avec l'âge. Ce phénomène est particulièrement observé lors de puberté, notamment chez les locuteurs, où des changements physiologiques entraînent des variations de la F_0. Cependant, tout au long de la vie adulte, des variations de la voix sont observées (Beck, 2010 ; Linville, 1995). Ces changements physiologiques peuvent s'expliquer par la perte d'élasticité des tissus pulmonaires (Mahler, 1983), de la raideur du thorax (Kahane, 1981) et de l'affaiblissement des muscles respiratoires (Kahane, 1981 ; Rossi *et al.*, 1996). Des modifications du larynx (Linville, 1995) peuvent également intervenir, et se traduisent par une ossification et une calcification des cartilages (Malinowski, 1967 ; Roncallo, 1948), une atrophie musculaire (Bach *et al.*, 1941 ; Rodeno *et al.*, 1993) et une érosion articulaire (Kahane, 1990). Plusieurs études indiquent que ces changements physiologiques provoquent des modifications acoustiques dans la voix (Ramig et Ringel, 1983 ; Torre et Barlow, 2009), aussi bien au niveau de la F_0 (Hollien et Shipp, 1972 ; Linville, 1981 ; McGlone et Hollien, 1963 ; Morris et Brown Jr, 1988 ; Mysak et Hanley, 1958 ; Ptacek *et al.*, 1966), que du type de phonation (Linville et Fisher, 1985 ; Linville *et al.*, 1990 ; Linville *et al.*, 1989 ; Ramig et Ringel, 1983 ; Wilcox et Horii, 1980).

En 1931, Pear a mené une expérience dans laquelle entre 3710 et 4285 auditeurs devaient estimer l'âge de 3 locutrices et 6 locuteurs âgés de 11 à 57 ans (μ = 35.89 et δ = 13.97) à partir d'enregistrements de lecture diffusés à la radio sur trois soirées consécutives (Pear, 1931). Pear indique une tendance des auditeurs à indiquer un multiple de 10 pour estimer l'âge des locuteurs. Pour chaque locuteur, la médiane est calculée et indique un écart entre l'âge réel et l'âge supposé de 0 à 12 ans (μ = 6.22 et δ = 4.02).

Tout au long du XX^{e} siècle, d'autres études ont été menées sur la perception de l'âge dans la voix sur tout type de conditions de parole. Ptacek et Sander ont étudié la perception de l'âge pour 2 groupes de locuteurs (moins de 35 ans et plus de 65 ans) à partir d'enregistrements de voyelles prolongées, d'enregistrements de lecture joués à l'endroit et à l'envers (Ptacek et Sander, 1966). Les 10 auditeurs ont pu identifier les groupes d'âge des locuteurs avec en moyenne 78 % d'identification correcte pour les voyelles prolongées, 87 % pour les enregistrements joués à l'envers et 99 % pour ceux joués à l'endroit.

Shipp et Hollien se sont intéressés à la perception de l'âge dans la voix masculine. Pour cela, ils ont réuni 175 locuteurs répartis équitablement dans 7 tranches d'âge (allant de 20-29 ans à 80-89 ans) (Shipp et Hollien, 1969). Les locuteurs ont été enregistrés au microphone en prononçant la phrase « *These take the shape of a long, round arch with its path high above, and its two ends apparently beyond the horizon* ». Les auditeurs sont répartis en 3 groupes, chacun ayant une tâche précise dans l'estimation de l'âge. Le premier groupe (30 auditeurs) devait indiquer si la voix entendue était une « voix jeune », une « voix âgée » ou une « voix ni jeune ni âgée ». Le second groupe (40 auditeurs) devait indiquer la décennie des locuteurs. Le dernier groupe (25 auditeurs) devait estimer l'âge des locuteurs. Les résultats du premier groupe montrent que le score augmente au fur et à mesure que l'âge des locuteurs augmente. Ceux du second groupe indiquent que, passé 50 ans, l'âge perçu par les auditeurs s'écarte, à la baisse, de l'âge réel. Enfin, l'estimation directe de l'âge des locuteurs démontre un coefficient de corrélation $r = 0.88$ entre l'âge réel et l'âge perçu. Les auteurs observent toutefois le même phénomène que pour le deuxième groupe : après 50 ans, les auditeurs tendent à sous-estimer l'âge réel des locuteurs. Les résultats de Shipp et Hollien sont confirmés par différentes études perceptives (Cerrato *et al.*, 2000 ; Huntley *et al.*, 1987 ; Ryan et Burk, 1974).

Linville et Fisher ont étudié l'influence de la phonation dans la perception de l'âge chez les locutrices (Linville et Fisher, 1985). Pour cela, 75 locutrices réparties de façon équilibrée dans 3 groupes d'âge (25-35 ans, 45-55 ans et 70-80 ans) sont entraînées pour maintenir la voyelle /æ/ durant 3 secondes minimum avec une fréquence fondamentale à 210 Hz avec une voix modale et avec une voix chuchotée. L'âge des locutrices est estimé par 23 auditrices, âgées de 20 à 28 ans, qui devaient indiquer si la voix entendue était « jeune », « d'âge mûr » ou « âgée ». Pour chacune des conditions, les auditrices ont fait mieux que le hasard, avec 51 % d'identification correcte pour les enregistrements de voix modale et 43 % pour les enre-

gistrements de voix chuchotée. Les chercheuses concluent que le type de phonation donne des informations sur l'âge des locutrices.

Mulac et Giles ont mené une expérience sur la perception de la voix âgée aussi bien chez les locutrices que chez les locuteurs (Mulac et Giles, 1996). Leur étude montre que 214 jeunes auditeurs (μ = 19.5 et δ = 2.41) ont décrit la voix de locuteurs âgés de 59 à 92 ans, en utilisant 24 échelles binaires différentes (par exemple : compétent/incompétent ou fort/faible). Les résultats indiquent qu'une voix sourde, tendue, voilée et où une élongation des voyelles est observée, est perçue comme plus âgée. Les résultats montrent également que les voix perçues comme étant plus âgées sont associées à certains stéréotypes tels que la fragilité, les mauvaises intentions, la soumission, l'incompétence et la dépendance.

1.1.3.5 Y a-t-il une voix homosexuelle ?

À l'instar du genre, la sexualité est d'abord vue de façon binaire : soit une personne est hétérosexuelle, soit elle est homosexuelle. Au niveau linguistique, la définition d'un langage gay a essuyé les mêmes revers que la définition d'un langage féminin (Baker, 2008). La perception de l'orientation sexuelle repose essentiellement sur les stéréotypes de genre. Selon si les femmes répondent aux stéréotypes de féminité et les hommes répondent aux stéréotypes de masculinité, la perception de leur orientation sexuelle change (stéréotypes de « la camionneuse » ou du « gay efféminé »). À l'instar du genre ou de l'âge, l'orientation sexuelle transparaît-elle dans la voix ?

Smyth *et al.* se sont penchés sur la question et indiquent qu'il existe une « voix gay » et une « voix hétérosexuelle » (Smyth *et al.*, 2003). Leur étude porte sur la perception de 25 locuteurs, dont 17 sont homosexuels. Les locuteurs ont été choisis en vue d'avoir « une distribution homogène des voix sur un continuum allant d'une voix à consonance gay à une voix à consonance hétérosexuelle ». Ce détail indique que l'étude comporte un biais qui induit les auditeurs vers une distinction entre « voix gay » et « voix hétérosexuelle ». Ces 25 locuteurs sont enregistrés dans 3 conditions différentes : en lisant un paragraphe scientifique, en lisant un texte dramatique et en répondant à une question ouverte. Des extraits des enregistrements sont évalués par 46 auditeurs, dont 14 hommes gays. Le reste des auditeurs sont supposés hétérosexuels (leur orientation sexuelle ne leur a pas été demandée).

Pour chaque enregistrement, les auditeurs se trouvaient face à un choix binaire et devaient indiquer si la voix entendue « semblait gay ou hétérosexuelle ». Ce choix était accompagné d'un score de confiance de 0 à 6 où 0 est une réponse au hasard et 6 une certitude absolue. Les auditeurs ont d'abord évalué les locuteurs sur le passage scientifique, puis le texte dramatique et enfin sur la réponse à la question ouverte. Ainsi, ils ont dû émettre pour chaque locuteur 3 évaluations de leur voix. Les résultats globaux montrent que seulement 10 locuteurs sont perçus par la majorité des auditeurs comme étant homosexuels avec un indice de confiance variant de 3.6 à 5. Parmi ces 10 locuteurs, un seul est hétérosexuel. Les résultats ne sont pas détaillés pour chaque tâche ; il y a cependant un « score moyen » (non accompagné de l'indice de confiance) qui semble indiquer que les gays ont une voix qui semble « plus gay » que les hétérosexuels.

En 2020, Sulpizio *et al.* mettent en place une expérience similaire sur des lesbiennes en vue de montrer si le « *gaydar*[11] auditif » existe effectivement (Sulpizio *et al.*, 2020). Leur étude porte sur 30 femmes issues de 3 nationalités différentes. Sur les 10 femmes de chaque nationalité, 5 ont déclaré être hétérosexuelles et 5 ont déclaré être saphiques. Les résultats montrent que la voix ne permet pas d'indiquer l'orientation sexuelle et ce, que ce soit en italien, portugais ou allemand.

1.1.3.6 Petites et grosses voix : estimer la taille et le poids d'un individu à sa voix

La question de la perception de la taille et du poids à l'écoute de la voix s'est posée. Entre 1976 et 1980, Lass *et al.* se sont penchés sur la perception de la taille et du poids à partir de la voix (Lass *et al.*, 1979a ; Lass *et al.*, 1978 ; Lass *et al.*, 1980a ; Lass et Colt, 1980 ; Lass et Davis, 1976 ; Lass *et al.*, 1979b ; Lass *et al.*, 1980b ; Lass *et al.*, 1980c). Chacune des études se présente sous un prisme différent.

Dans (Lass et Davis, 1976), 15 locutrices et 15 locuteurs âgés de 18 à 25 ans sont enregistrés en lecture. Sur les 15 locutrices, 7 mesurent entre 152 centimètres et 165 centimètres et 8 mesurent entre 168 centimètres et 183 centimètres. Au niveau du poids, 1 pèse moins de 45 kilogrammes, 13 pèsent entre 45 et 68 kilogrammes et 1 pèse entre 68 et 91 kilogrammes.

11. Mot valise entre « gay » et « radar » qui désigne une capacité intuitive à deviner l'orientation sexuelle d'une personne

Du côté des locuteurs, 11 mesurent entre 168 centimètres et 183 centimètres et 4 mesurent plus de 183 centimètres. En termes de poids, 3 pèsent entre 46 et 68 kilogrammes et 11 pèse entre 68 et 91 kilogrammes et 1 pèse plus de 91 kilogrammes. Les locuteurs doivent être répartis par intervalles de tailles et de poids par les 30 auditeurs. Les auditeurs ont obtenu une meilleure performance que le hasard. L'étude indique que les indices utilisés sont la F_0, le volume sonore, le débit, l'intonation et la résonance.

Dans (Lass *et al.*, 1978), l'estimation de la taille et du poids est cette fois-ci directe. Les locuteurs sont les mêmes que l'étude précédente, publiée en 1976. 40 auditeurs ont participé à l'expérience durant laquelle ils devaient indiquer la taille et le poids de façon directe. L'étude conclut que les auditeurs ont eu moins de difficulté dans l'estimation de la taille que dans l'estimation du poids. Les études suivantes s'intéressent à des indices tels que le débit (Lass *et al.*, 1979a), la « complexité phonétique » (Lass *et al.*, 1979b), la parole filtrée (Lass *et al.*, 1980c), la phonation (Lass *et al.*, 1980b) et la combinaison d'indices acoustiques et visuels (Lass et Colt, 1980).

Chacune de ces études indique les paramètres étudiés intervenant dans la perception de la taille et du poids des locuteurs. Cohen *et al.* reprend l'ensemble des expériences de Lass *et al.* précédemment citées et indique les raccourcis empruntés par les auteurs (Cohen *et al.*, 1980). Les auteurs se concentrent sur le poids. Elles mettent en évidence que la « complexité phonétique » (Lass *et al.*, 1979b) et la parole filtrée (Lass *et al.*, 1980c) n'ont aucun effet sur les résultats issus des études précédentes (Lass *et al.*, 1978 ; Lass et Davis, 1976) puisque la différence entre les résultats n'est pas significative. Elles indiquent également que l'estimation du poids est directement liée à la perception du genre des locuteurs, et non de leur voix, ce qui explique le faible écart-type observé dans les études. La conclusion de l'étude indique que les indices acoustiques étudiés n'ont pas donné de résultats probants et que les résultats relèvent davantage de la corrélation que d'un lien de cause à effet.

Plus récemment, Krauss *et al.* indiquent qu'il est possible d'estimer l'âge d'une personne, sa taille et son poids à partir d'un enregistrement de deux phrases (*Joe took father's shoe bench out* et *She is waiting at my lawn*) (Krauss *et al.*, 2002). Ces deux phrases ont été sélectionnées, car elles « couvrent l'ensemble de l'espace vocalique de l'anglais américain » et que les « différences physiques entre les locuteurs sont surtout rencontrées au niveau des voyelles, donnant ainsi des informations sur le système phonatoire ». Alors que la conclusion

de l'étude indique que ces deux phrases permettent de reconnaître les attributs physiques des individus, les informations sur les locuteurs montrent peu de variations pour les tailles, que ce soit pour les femmes (μ = 165.86 et δ = 8.25) ou pour les hommes (μ = 179.32 et δ = 8.25). En termes de poids, des différences d'écart-type s'observent entre femmes (μ = 58.01 et δ = 6.38) et hommes (μ = 79.97 et δ = 18.56). En 2018, la taille moyenne des États-Uniens était plus élevée que dans les années 2001-2002. Des variations de poids sont également observées puisque le poids moyen est réduit (Fryar *et al.*, 2018). Outre qu'il n'existe pas une grande disparité de la taille et du poids des locuteurs, les résultats montrent également qu'il existe peu de significativité pour l'ensemble des auditeurs ($p < 0.1$ pour les résultats de la moitié des auditeurs). Les mêmes critiques que pour les études de Lass *et al.* peuvent être apportées à cette étude.

Smith *et al.* ont également tenté de répondre à cette question en indiquant que la taille de l'appareil phonatoire est dépendant de la taille du locuteur, ce qui a été démontré (Laver et Trudgill, 1979 ; van Bezooijen, 1987), et que par conséquent, la valeur de la F_0 en découle (Smith *et al.*, 2005). Cependant, plusieurs études admettent qu'il n'existe pas de lien de cause à effet entre les attributs physiques d'une personne (taille, poids) et la valeur de la F_0 (Cohen *et al.*, 1980 ; Künzel, 1989).

1.1.4 La perception de la voix en psychologie évolutionniste

En 1859, Darwin publie *De l'origine des espèces* (Darwin, 1859). Dans cet ouvrage, il propose une théorie selon laquelle les espèces vivantes sont issues d'espèces éteintes par le biais de la sélection naturelle. La psychologie évolutionniste reprend la théorie de l'évolution de Darwin dans le but d'expliquer les mécanismes de pensées et comportementaux des êtres humains. La psychologie évolutionniste prétend décrire « l'architecture de l'esprit humain » qui s'est fixé il y a 100 000 ans (au plus tôt) durant le Pléistocène, la période géologique précédant l'ère actuelle qui perdure depuis près de 12 000 ans.

La recherche en psychologie évolutionniste s'est intéressée à la voix, notamment pour y déceler des traits tels la force physique ou encore le pouvoir ou l'infidélité. Certaines études portent sur la question de la perception de la beauté [12] dans la voix.

1.1.4.1 Une affaire de beaux parleurs ?

Zuckerman et Driver ont travaillé sur l'expression « ce qui est beau est bon » pour évaluer si ce stéréotype transparaît dans la voix (Zuckerman et Driver, 1989). Dans leur étude, 200 locuteurs sont filmés en lecture. 10 sujets (5 femmes et 5 hommes) sont soumis à l'écoute des voix seules d'une part, et de la vidéo seule d'autre part (les locuteurs diffèrent entre les sessions). Après chaque stimulus, un score de 1 à 7 est donné (1 signifiant « très laid » et 7 signifiant « très beau »). Les auteurs mettent en évidence une corrélation entre les résultats de l'écoute de la voix seule et ceux de la visualisation de la vidéo seule (r = 0.84 pour les femmes et r = 0.93 pour les hommes). Les chercheurs concluent que le stéréotype « ce qui est beau est bon » est confirmé dans la perception de la voix. Ce stéréotype est repris par Bruckert *et al.* qui ont évalué des voix de femmes et d'hommes et les ont moyenné de manière acoustique afin de déterminer la beauté du rendu. Leurs conclusions indiquent que plus le nombre de voix moyennées est élevé, plus la voix générée est considérée comme étant « belle ».

Hughes *et al.* reprennent le stéréotype « ce qui est beau est bon » et tentent de déterminer si la « beauté de la voix » peut être corrélée avec des traits physiques (rapport épaules/hanches, rapport poitrine/hanches, indice de masse corporelle) ou des comportements sexuels (âge de la première masturbation, du premier rapport sexuel, nombre de partenaires sexuels et de partenaires sexuels en couple avec une autre personne). Les auteurs concluent que les femmes qui possèdent un faible rapport poitrine/hanches et les hommes qui ont un rapport élevé épaules/hanches ont une voix plus « belle » que les autres.

12. Traduction du terme *attractiveness*

1.1.4.2 Perception de la force physique dans la voix

Plusieurs expériences menées sur différentes langues à travers le monde indiquent qu'il est possible de déterminer la force physique d'un locuteur masculin en écoutant sa voix (Sell *et al.*, 2010). Ils mesurent la force physique par la circonférence du biceps en flexion, la force exercée en serrant le poing, la force exercée en contractant la poitrine ou les épaules et/ou l'évaluation de photos. L'étude porte sur des locuteurs de la tribu amérindienne des Tsimane en Bolivie, d'une tribu de chasseurs-cueilleurs des Andes et des étudiants des États-Unis et de Roumanie. Un total de 8 expériences ont été menées avec des protocoles expérimentaux différents décrits dans le tableau 1.4.

ID	Locuteurs	Pays	Âge	Indicateurs de force	Stimuli
1	63 H	États-Unis	18-22 (μ = 18.7, δ = 0.88)	Poing, biceps, photo	Phrase
2	49 H	Bolivie	19-68 (μ = 35.8, δ = 13.5)	Poitrine, biceps	Phrase
3	20 H	Andes	15-71 (μ = 34.8, δ = 19.1)	Poitrine, épaules, poing, biceps	Phrase
4	50 H	États-Unis	18-31 (μ = 20.2, δ = 2.24)	Poitrine, épaules, photo	Phrase
5	45 H	Roumanie	20-38 (μ = 21.7, δ = 3.48)	Poitrine, poing, biceps	Phrase
6	50 F	États-Unis	18-22 (μ = 18.8, δ = 0.95)	Poitrine, épaules, photo	Phrase
7	30 F	Roumanie	20-29 (μ = 21.1, δ = 1.89)	Poitrine, poing, biceps	Phrase
8	54 H	États-Unis	18-23 (μ = 19.9, δ = 2.0)	Poitrine, biceps	Voyelles

Tableau 1.4 – *Protocoles des expériences menées dans (Sell* et al.*, 2010)*

Les résultats des différentes expériences sont exprimés avec une corrélation de Pearson entre la perception de la force physique et les mesures prises. Les résultats moyennés sont compris entre 0.26 (p = 0.07) et 0.51 (p = 0.001). Les auteurs indiquent que la forte

significativité du (faible) coefficient de corrélation permet d'indiquer que la force physique transparaît dans la voix et ce que les auditeurs soient familiers de la langue parlée ou non. Les auteurs indiquent également que les auditeurs « extraient des informations additionnelles de « formidabilité [13] » à partir de la voix qui ne transparaissent pas dans les indices visuels. »

1.1.4.3 Parler comme un chef

Dans la lignée des travaux de Sell *et al.*, Klofstad *et al.* ont mené une expérience sur la perception de la capacité à diriger d'une personne en fonction de sa voix (Klofstad *et al.*, 2012). Pour cela, 17 femmes et 10 hommes sont enregistrés en prononçant « *I urge you to vote for me this November* ». Les locutrices sont âgées de 21 à 60 ans (μ = 31) et les locuteurs de 20 à 55 ans (μ = 33). La F_0 des locuteurs est également calculée en utilisant le logiciel Praat (Boersma et Weenink, 2001). La F_0 moyenne des femmes se situe entre 162 Hz et 207 Hz (μ = 187) et celle des hommes entre 91 Hz et 116 Hz (μ = 107). Un premier panel d'auditeurs, composé de 83 étudiants (46 femmes et 37 hommes) de l'université de Miami, évalue les voix féminines et un second panel de 40 femmes et 49 hommes de Duke University évaluent les voix masculines. Après écoute de deux stimuli de F_0 différentes, les auditeurs devaient répondre à la question « *If they were running against each other in an election, which voice would you vote for?* ». La seconde partie de l'expérience requiert trois panels d'auditeurs composés chacun de 35 femmes et 35 hommes de Duke University pour répondre à chacun à une question différente.

1. *Which voice is more competent (e.g. capable, experienced, knowledgeable, effective)?*
2. *Which voice is stronger (e.g. confident, determined, resolute, self-assured)?*
3. *Which voice is more trustworthy (e.g. honest, straightforward, reliable, believable)?*

Lors du choix binaire entre deux voix, les résultats montrent que les auditeurs tendent à sélectionner la voix la plus grave (≈60 % du temps). Les auteurs indiquent qu'il existe possiblement un biais induit par le contexte électoral de l'énoncé et que dans un autre contexte (professionnel par exemple), d'autres résultats pourraient être observés. Au niveau de la

13. *awesomeness* en anglais.

définition des voix, les résultats montrent que les voix féminines les plus graves sont considérées comme plus « compétentes », « fortes » et « dignes de confiance ». Cette tendance ne s'observe que très peu chez les hommes où seuls les auditeurs considèrent les voix les plus graves comme étant « compétentes » et « fortes ». Malheureusement, ces résultats ne sont pas mis en corrélation avec des capacités réelles à diriger, ce qui indique avant tout des attendus plus que des véritables compétences.

1.1.4.4 Fréquence fondamentale et infidélité

Hughes et Harrison ont également travaillé sur la perception de la voix et plus particulièrement sur la perception de l'infidélité dans la voix (Hughes et Harrison, 2017). Reprenant le travail de O'Connor *et al.*, elles ont rassemblé 10 locutrices et 10 locuteurs dont la moitié (5 par genre) a indiqué avoir eu une relation sexuelle avec une personne extérieure à leur relation « romantique, sérieuse et exclusive » actuelle ou passée (O'Connor *et al.*, 2011). Les locuteurs sont hétérosexuels, caucasiens, non mariés et sont en couple. Les stimuli vocaux sont composés de l'énumération de chiffres de 1 à 10 à un rythme de 1 chiffre par seconde.

Les stimuli sont évalués par 54 étudiants de différentes ethnicités et tous se déclarent hétérosexuels. 45 % indiquent être dans une « relation romantique sérieuse et exclusive » et 55 % indiquent ne pas l'être. Les auditeurs devaient répondre à la question « *Using the following scale, please rate how likely you think the person speaking has 'cheated' on their romantic partner with whom they are in an exclusive, committed relationship.* » sur une échelle à 10 paliers où 1 correspondait à « impossible » et 10 est « très probable ».

Les résultats de l'expérience montrent que sur les voix naturelles, les locuteurs infidèles sont mieux reconnus que leurs homologues fidèles (4.24 pour les locutrices fidèles, 4.73 pour les locutrices infidèles, 4.99 pour les locuteurs fidèles et 5.35 pour les locuteurs infidèles). Cependant, l'effet n'est pas significatif pour chacun des groupes. En effet, les résultats indiquent une faible significativité pour les hommes fidèles et une forte significativité pour les femmes infidèles.

1.1.4.5 Critiques de la psychologie évolutionniste

Outre les critiques que les précédentes études citées ont pu recueillir, la psychologie évolutionniste est également la cible de critiques. En effet, les sujets de controverse s'orientent autour de la modularité de l'esprit (Gibbs et Van Orden, 2010 ; Grossi *et al.*, 2014 ; Hamilton, 2008), de l'environnement, de l'adaptation évolutionniste (Plotkin, 2008), de l'ethnocentrisme (Davies, 2012 ; Lancaster, 2003 ; Paulson, 2018), de réductionnisme et de déterminisme génétique (Ehrlich et Feldman, 2003 ; Hamilton, 2008 ; Lancaster, 2003 ; Lickliter et Honeycutt, 2003 ; Maiers, 2001 ; Plotkin, 2008) et de sexisme (Huteau, 2021 ; Ruti, 2015).

Rose et Rose sont à l'origine d'un ouvrage très critique envers la psychologie évolutionniste et indiquent que les fondements de cette dernière sont erronés (Rose et Rose, 2010). Les auteurs indiquent que les études de la discipline essentialisent certains traits, tels que le bagou des hommes ou encore la fausse pudeur des femmes. Ces études sont également critiquées pour leur manque de généralité et de généralisabilité. En effet, les études sont souvent réalisées sur des échantillons de sujets issus de populations occidentales et de classes sociales élevées et sont souvent réalisées sur des sujets de sexe masculin. Les auteurs indiquent que les résultats de ces études ne sont pas généralisables aux femmes et aux classes sociales inférieures.

1.1.5 Processus cognitifs mis en œuvre dans la reconnaissance des voix

1.1.5.1 Reconnaître les voix

Les êtres humains sont en mesure de discriminer les voix. Cette capacité arrive très tôt dans le développement humain. Dès le troisième trimestre de grossesse, les fœtus possèdent un appareil auditif fonctionnel et commencent à réagir aux différents sons, dont les voix. *In utero*, les fœtus, d'en moyenne 38 semaines, montrent qu'ils sont plus sensibles à la voix de leur mère qu'à celle d'autres personnes (Kisilevsky *et al.*, 2003). Cela se traduit par une augmentation du rythme cardiaque lorsqu'ils sont stimulés avec la voix de leur mère et une diminution du rythme cardiaque après stimulation avec la voix d'une autre personne. Après la naissance, le même phénomène est observé. Les nouveau-nés et nourrissons sont

reliés à une tétine qui permet de mesurer le nombre de succions non nutritives. À l'écoute de la voix de leur mère, les nouveau-nés et nourrissons augmentent la fréquence des succions alors que cette dernière diminue à l'écoute de la voix d'une autre personne, et ce, qu'ils soient âgés de trois jours seulement (DeCasper et Fifer, 1980) ou de quatre semaines (Mills et Melhuish, 1974).

La reconnaissance humaine des locuteurs s'appuie sur les informations fournies par la voix ainsi que leur perception par les auditeurs. Plusieurs études se sont intéressées à la reconnaissance des voix « connues » en opposition à la reconnaissance des voix « inconnues ». Ces études indiquent que la reconnaissance des voix est particulièrement bonne lorsqu'il s'agit de personnes proches ou d'amis (Ladefoged, 1980), de collègues (Hollien *et al.*, 1982) ou encore de personnalités publiques (Van Lancker *et al.*, 1985a ; Van Lancker *et al.*, 1985b). Pour chacune de ces études, les voix « connues » sont correctement identifiées en moyenne entre 67 % et 98 % du temps. Du côté des voix « inconnues », les résultats chutent en moyenne entre 27 % et 50 % (Hollien *et al.*, 1982 ; Ladefoged, 1980).

En 1937, McGehee montre qu'après une première exposition à une voix inconnue, les auditeurs la reconnaissent mieux (McGehee, 1937). Néanmoins, leur capacité à reconnaître la voix entendue s'estompe au fil du temps sans autre exposition à la voix que l'exposition initiale. Ainsi, dans l'étude citée, alors que le taux de bonne reconnaissance après 2 jours est de 83 %, le taux de bonne reconnaissance après 5 mois est de 13 %.

1.1.5.2 Paramètres acoustiques et prototypage

Au niveau cognitif, la reconnaissance des voix est abordée en utilisant la dichotomie « voix connue »/« voix inconnue ». Van Lancker et Kreiman indiquent que la reconnaissance des voix « connues » et celle des voix « inconnues » sont deux processus cognitifs distincts et indépendants (Van Lancker et Kreiman, 1987). En effet, la reconnaissance des voix « connues » s'effectuerait dans l'hémisphère droit alors que celle des voix « inconnues » s'effectue dans les deux hémisphères.

En 1976, Bricker et Pruzansky indiquent que la reconnaissance des voix s'effectue par extraction de paramètres acoustiques (Bricker et Pruzansky, 1976). À l'écoute d'une voix, les

auditeurs extraient des indices acoustiques et comparent leur similarité avec les paramètres acoustiques déjà connus. Cela impliquerait que la fréquence d'exposition à la voix d'une personne donnée permet de définir sa voix plus précisément, ce qui permet d'expliquer la différence de performance dans la reconnaissance des voix « connues » et voix « inconnues ».

Lavner *et al.* suggèrent plutôt que les auditeurs construisent un prototype générique qui contient une idée globale de ce qu'est un locuteur (Lavner *et al.*, 2001). Ainsi, lorsque l'auditeur entend une voix nouvelle, cette dernière est comparée au prototype et seules les différences entre le prototype et la voix sont « mémorisées ». Le prototype est construit à partir des voix déjà rencontrées, ce qui peut expliquer pourquoi il est plus facile de reconnaître les voix de personnes qui parlent notre langue (Lavner *et al.*, 2001). Lavner *et al.* indiquent également que toutes les voix « inconnues » ne sont pas égales dans leur reconnaissance. En fonction du prototype construit par l'auditeur, une même voix « inconnue » peut être plus ou moins facile à reconnaître qu'une autre (Sullivan et Schlichting, 2000).

1.1.5.3 Le cas particulier de la phonagnosie

En 1982, Van Lancker et Canter introduisent le terme « phonagnosie » (du grec *phônê* (« voix ») et *gnôsia* (« connaissance ») précédé du préfixe privatif *a*-) pour désigner une capacité lacunaire dans la perception des personnes à la voix (Van Lancker et Canter, 1982). La phonagnosie peut être innée (Garrido *et al.*, 2009) ou acquise (Assal *et al.*, 1976 ; Neuner et Schweinberger, 2000 ; Van Lancker et Canter, 1982), et se manifeste sous deux formes différentes : la phonagnosie aperceptive et la phonagnosie associative (Buchtel et Stewart, 1989).

Les personnes atteintes de phonagnosie aperceptive peinent à percevoir les caractéristiques acoustiques qui permettent l'identification d'une personne alors que les personnes atteintes de phonagnosie associative ne parviennent pas à « mettre un nom ou un visage » sur une voix entendue (Hailstone *et al.*, 2011 ; Roswandowitz *et al.*, 2014 ; Van Lancker et Kreiman, 1987). Les différentes études sur la phonagnosie montrent que seule la capacité à percevoir l'identité par la voix est touchée. En effet, l'acuité auditive, la reconnaissance des visages, la perception de la parole, la perception des sons et la perception de la musique ne diffèrent pas significativement entre une personne phonagnosique et une personne non

phonagnosique (Garrido *et al.*, 2009 ; Roswandowitz *et al.*, 2014). Les recherches de Roswandowitz *et al.*, dans la lignée de celles de Garrido *et al.*, montrent que des lésions dans la zone temporale et dans le lobe pariétal inférieur droit entraînent une difficulté à reconnaître les personnes à leur voix (Roswandowitz *et al.*, 2018). Plus précisément, Roswandowitz *et al.* constatent que la reconnaissance des voix « familières » et celles des voix « nouvelles » ne se trouvent pas dans les mêmes zones cérébrales. La reconnaissance des voix « familières » est rendue difficile lorsque le lobe temporal postérieur gauche a subi des lésions alors que la reconnaissance des voix « nouvelles » est rendue difficile après des lésions dans la zone temporale droite et dans le lobe pariétal inférieur droit.

Pour résumer, la phonagnosie nous indique que la perception de l'identité par la voix passe par deux étapes : la première se penche sur l'analyse de la voix afin d'en extraire des caractéristiques qui puissent déterminer l'identité du locuteur ; la deuxième étape consiste à pouvoir associer un nom ou un visage à la voix perçue. La phonagnosie confirme également que la reconnaissance des voix « connues » et celle des voix « nouvelles » demandent des processus bien distincts, faisant appel à l'hémisphère droit du cerveau pour l'identification des voix « nouvelles » et à l'hémisphère gauche pour l'identification des voix « connues ». Roswandowitz *et al.* estiment que la phonagnosie touche 1 % de la population (Roswandowitz *et al.*, 2014).

1.2 Reconnaître les locuteurs avec des systèmes automatiques

De nos jours, de plus en plus de services nécessitent de reconnaître les personnes à leur voix pour diverses raisons. C'est le cas des services bancaires en ligne, des assistants personnels, des systèmes de sécurité, *etc.* La Reconnaissance automatique du locuteur (RAL) est une discipline qui permet de déterminer si une personne donnée est la source d'un enregistrement de voix.

1.2.1 Qu'est-ce que reconnaître les locuteurs en informatique ?

La RAL est une discipline qui contient plusieurs tâches liées à la discrimination des voix en locuteurs :

Identification du locuteur L'objectif de l'identification du locuteur est de déterminer si une personne donnée est la source d'un enregistrement de voix. Pour cela, nous construisons un modèle à partir d'enregistrements d'un ensemble d'individus connus. La comparaison entre le modèle et l'enregistrement à tester permet de déterminer si le locuteur est connu ou non.

Vérification du locuteur La tâche de vérification du locuteur a pour objectif de déterminer si un locuteur donné est la source d'un enregistrement de voix. Pour cela, il est nécessaire d'utiliser un ou plusieurs enregistrements dudit locuteur afin d'en former un modèle. L'enregistrement est comparé au modèle du locuteur pour déterminer si le locuteur est à l'origine de l'enregistrement.

Diarization La diarization consiste à distinguer les différents locuteurs d'un signal de parole. Pour cela, le signal de parole est segmenté puis les segments sont regroupés par locuteur.

1.2.2 Histoire de la reconnaissance automatique du locuteur

1.2.2.1 Premiers systèmes de reconnaissance du locuteur : paramètres acoustiques et chaînes de Markov cachées

Le premier système de RAL est apparu en 1963 dans les Laboratoires Bell (Pruzansky, 1963). L'objectif du système est d'identifier 10 locuteurs en milieu fermé à partir des enregistrements de mots isolés extraits de signaux de parole. Ce système s'appuie sur les banques de filtres de Mel pour l'analyse du signal de parole. Pruzansky calcule une corrélation temporelle et fréquentielle des enregistrements de mots isolés avec les signaux de parole des locuteurs. Le locuteur est identifié par le modèle qui possède la plus grande corrélation avec l'enregistrement à tester. Ce système obtient 89% d'identification correcte. Trois ans plus

tard, en 1966, Li *et al.* proposent un système de vérification du locuteur reposant également sur les banques de filtres, introduisant également des imposteurs (locuteurs qui ne sont auteurs d'aucun enregistrement) (Li *et al.*, 1966). Doddington, travaillant alors pour Texas Instruments, propose un système de vérification du locuteur s'appuyant sur les formants du signal de parole (Doddington, 1971).

À partir des années 1980, les systèmes de RAL s'appuient sur des modèles statistiques pour représenter les locuteurs. Ainsi apparaissent des systèmes utilisant des *Hidden Markhov Model* (HMM), des automates à état fini qui permettent d'estimer la probabilité d'une séquence d'observations (Ferguson, 1980). Dans un contexte de RAL, les HMM estiment la probabilité qu'une séquence de vecteurs acoustiques soit prononcée par un locuteur donné. Les HMM permettent également de reconnaître les locuteurs dans un environnement bruité (Matsui *et al.*, 1996 ; Rose *et al.*, 1994).

1.2.2.2 Introduction du rapport de vraisemblance : UBM-GMM et i-vector

Plusieurs facteurs influent sur les résultats d'un système de RAL : l'environnement, le canal de transmission, le matériel d'enregistrement, *etc.* Ces facteurs sont présentés dans Comparer les voix de façon empirique. Pour prendre en compte ces facteurs, l'adoption du *LR* provoque un bouleversement dans la conception des systèmes de RAL. Il est nécessaire de pouvoir donner le résultat d'une tâche de RAL tout en prenant en compte la variabilité mais aussi les erreurs de décision.

Dans le cadre d'une vérification du locuteur, soit e un enregistrement de parole, l un locuteur, le *LR* est défini par l'équation 1.1

- H_0 est l'hypothèse que e a été prononcé par l ;
- H_1 est l'hypothèse que e a été prononcé par une personne différente de l.

$$LR = \frac{P(e|H_0)}{P(e|H_1)} \tag{1.1}$$

Pour répondre à cette contrainte, les systèmes de RAL doivent s'adapter. Pour cela, l'approche UBM-GMM est utilisée. Cette méthode consiste à construire un modèle de locuteurs universel (*Universal Background Model* (UBM)) à partir d'un ensemble de locuteurs (Reynolds

et al., 2000). D'autres approches sont également utilisées telles que les machines à support de vecteur SVM (Campbell *et al.*, 2006 ; Mclaren *et al.*, 2007). Durant les années 1990, les systèmes de RAL connaissent des améliorations, notamment en utilisant des modèles paramétriques et en quantifiant les vecteurs acoustiques : les *Gaussian Mixture Model* (GMM) (Reynolds et Rose, 1995). Les GMM permettent d'estimer la distribution des paramètres acoustiques en les modélisant sous forme de gaussiennes. Ils sont par ailleurs le résultat de la somme de ces gaussiennes. Ces dernières sont caractérisées par leurs moyenne, variance et amplitude. Celles-ci peuvent être optimisées par l'algorithme d'espérance-maximisation (Bilmes *et al.*, 1998 ; Dempster *et al.*, 1977). Ainsi, les GMM permettent un premier pas dans la modélisation de la variabilité intralocuteur. Soit un GMM λ, une séquence de vecteurs de paramètres acoustiques i, la probabilité de la séquence correspond à la somme des probabilités de chaque vecteur i_n en fonction du modèle, ainsi que décrit dans l'équation 1.2, qui correspond au numérateur de l'équation 1.1.

$$P(i|\lambda) = \sum_{n=1}^{T} P(x_n|\lambda) \tag{1.2}$$

Les UBM sont des « super GMM » qui modélisent la distribution des paramètres acoustiques d'un ensemble de locuteurs pour former un « modèle du monde » (Reynolds *et al.*, 2000). Autant que possible, l'UBM doit être représentatif de la population de locuteurs. Typiquement, les UBM sont conçus pour représenter l'hypothèse H_1 de l'équation 1.1.

Les systèmes UBM-GMM ont connu plusieurs améliorations au cours des années 2000. Les UBM se sont adaptés aux modèles de locuteurs, notamment grâce au *Maximum A Posteriori* (Gauvain et Lee, 1994). Les supervecteurs ont également été introduits pour pallier les variations des conditions d'enregistrement (Campbell *et al.*, 2006). Il s'agit d'un vecteur global pour l'ensemble des locuteurs. C'est le cas des *Factor Analysis* et des améliorations qui en ont été faites telles que la Joint Factor Analysis (JFA) (Kenny et Dumouchel, 2004). Les *Factor Analysis* se sont simplifiés, devenant ainsi la matrice de variabilité totale, ce qui donne naissance aux i-vector (Dehak *et al.*, 2011). La matrice de variabilité totale est un espace dans lequel l'ensemble de la variabilité de la parole est représentée. Soit un locuteur l, des supervecteurs GMM m et une matrice de variabilité totale T et un i-vector i, le modèle du locuteur est présenté dans l'équation 1.3.

$$m_l = m_{ubm} + Ti_l \tag{1.3}$$

Après extraction des i-vector, ces derniers sont normalisés avant d'être comparés.

1.2.2.3 Réseaux de neurones profonds et *x-vector*

Vers le milieu des années 2010 apparaissent des systèmes de RAL basés sur des réseaux de neurones profonds (Heigold *et al.*, 2016 ; Matějka *et al.*, 2016 ; Rouvier *et al.*, 2015 ; Variani *et al.*, 2014).

En 2017, Snyder *et al.* introduit les réseaux de neurones profonds pour la RAL (Snyder *et al.*, 2017). L'année suivante, l'architecture *x*-vector apparaît (Snyder *et al.*, 2018) et s'est imposée dans les campagnes d'évaluation NIST-SRE, notamment grâce à ses performances inégalées (Snyder *et al.*, 2018 ; Villalba *et al.*, 2020a). L'architecture *x-vector* repose sur un réseau de neurones TDNN, spécifiquement conçu pour les données séquentielles, à l'origine prévu pour la classification de phonèmes (Snyder *et al.*, 2018 ; Waibel *et al.*, 1989).

Couche	Contexte	Taille du contexte	Entrée×sortie
frame1	$[t-2, t+2]$	5	120×512
frame2	$t-2, t, t+2$	9	1536×512
frame3	$t-3, t, t+3$	15	1536×512
frame4	$\{t\}$	15	512×512
frame5	$\{t\}$	15	512×1500
statistiques groupées	$[0, T)$	T	1500T×3000
segment6	$\{0\}$	T	3000×512
segment7	$\{0\}$	T	512×512
softmax	$\{0\}$	T	512×N

.

Tableau 1.5 – *Architecture neuronale d'un TDNN. Les* x-vector *sont extraits en sortie du segemnt6. D'après (Snyder* et al.*, 2018)*

Depuis, les architectures *x-vector*s n'ont cessé d'être améliorées pour atteindre une meilleure performance, particulièrement avec les TDNN étendus (Snyder *et al.*, 2019) ou les

TDNN factorisés (Povey *et al.*, 2018). En 2020, Desplanques *et al.* mettent au point l'architecture ECAPA-TDNN qui obtient une très bonne performance (Desplanques *et al.*, 2020).

1.2.3 Score et seuil de décision

1.2.3.1 Faux rejets et fausses acceptations

Dans le cadre d'une vérification du locuteur, soit e un enregistrement de parole, l un locuteur. Alors, deux hypothèses sont possibles :

- H_0 est l'hypothèse que e a été prononcé par l ;
- H_1 est l'hypothèse que e a été prononcé par une personne différente de l.

S'il est connu *a priori* que le locuteur a prononcé l'enregistrement, nous parlons de comparaison cible. Dans le cas de deux locuteurs différents, nous parlons de comparaison imposteur. Deux types d'erreur peuvent être rencontrés :

Faux rejet Si lors d'une comparaison cible, l'hypothèse H_1 a été retenue, alors le résultat est un faux rejet (FR).

Fausse acceptation Si lors d'une comparaison imposteur, l'hypothèse H_0 a été retenue, alors le résultat est une fausse acceptation (FA).

Ces deux types d'erreurs permettent de calculer les métriques suivantes :

Taux de faux rejets Pour calculer le taux de faux rejets, seules les comparaisons cibles sont évaluées. Le *FRR* est défini par l'Équation 1.4.

$$FAR = \frac{\text{nombre de faux rejets}}{\text{nombre de comparaisons cibles}} \tag{1.4}$$

Taux de fausses acceptations Pour calculer le taux de fausses acceptations, seules les comparaisons imposteurs sont évaluées. Le *FAR* est défini par l'équation 1.5.

$$FRR = \frac{\text{nombre de fausses acceptations}}{\text{nombre de comparaisons imposteurs}} \tag{1.5}$$

1.2.3.2 Prise de décision

La prise de décision d'un système de RAL est basée sur le *LR* défini par l'équation 1.6. Soit :

- e un enregistrement de parole ;
- l un locuteur ;
- H_0 l'hypothèse selon laquelle e a été prononcé par l ;
- H_1 l'hypothèse selon laquelle e a été prononcé par une personne différente de l ;
- τ le seuil de décision.

$$LR = \frac{P(e|H_0)}{P(e|H_1)} \begin{cases} \geq \tau & \text{L'hypothèse } H_0 \text{ est retenue.} \\ \leq \tau & \text{L'hypothèse } H_1 \text{ est retenue.} \end{cases} \tag{1.6}$$

Le score seul ne permet pas de déterminer l'hypothèse la plus probable. Il est nécessaire de le coupler avec le seuil de décision pour considérer l'hypothèse la plus probable. Au-delà du seuil de décision, l'hypothèse H_0 est retenue, et en deçà, l'hypothèse H_1 est retenue.

Le seuil de décision peut être calculé *a priori*, en fonction de la distribution des scores des comparaisons cibles et imposteurs. Le seuil de décision est calculé de sorte de minimiser le taux d'erreur égale (*EER*) qui correspond au score pour lequel le taux de faux rejets (*FRR*) est égal au taux de fausses acceptations (*FAR*). L'*EER* est un indicateur du pouvoir discriminant du système.

Les courbes *Detection Error Trade-off* (DET) permettent de visualiser les variations du *FRR* et du *FAR* en fonction du seuil de décision. La figure 1.3 permet de comparer la performance de plusieurs systèmes. Plus la courbe est proche de l'origine, plus le système est performant. La figure 1.3 la courbe DET montre les variations de *FAR* et *FRR* en fonction du seuil de décision pour trois systèmes/conditions.

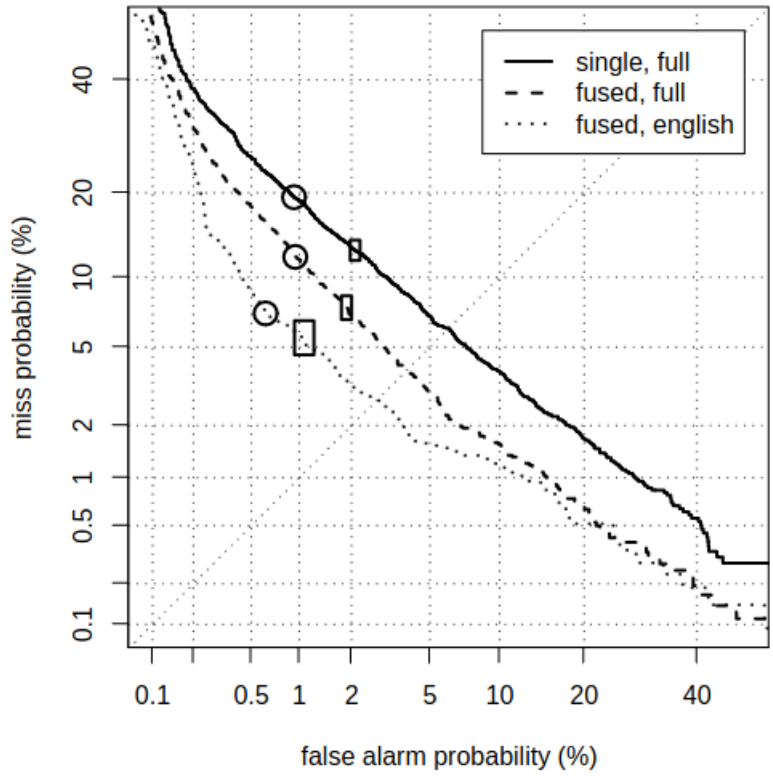


Figure 1.3 – *Courbes DET pour 3 systèmes/conditions (Brummer et Preez, 2006)*

1.2.4 Campagnes d'évaluation NIST-SRE

Entre 1996 et 2020, le *National Institute of Standards and Technology* (NIST) conduit 18 campagnes d'évaluation des systèmes de reconnaissance du locuteur sous le nom de *Speaker Recognition Evaluation* (SRE) (Greenberg *et al.*, 2020). Les objectifs de ces campagnes sont multiples. Non seulement elles permettent d'orienter la recherche et les progrès technologiques dans le domaine mais, en plus, elles permettent de rendre compte de l'état de l'art des systèmes utilisés. Les campagnes d'évaluation ont été adaptées au fur et à mesure des progrès des systèmes de RAL.

Les compagnes d'évaluation comprennent plusieurs tâches issues de la RAL. Les participants ont à disposition des millions de comparaisons qui peuvent être cibles ou imposteurs

(Greenberg *et al.*, 2020). Ces données doivent être aussi variées que possible, aussi bien au niveau des langues parlées, que du matériel d'enregistrement utilisé ou encore la distance entre le locuteur et le microphone. Les enregistrements montrent également des variations en termes de volume vocal, d'environnement bruité ou encore l'utilisation de différents réseaux de télécommunications (téléphone, VoIP) et canaux audiovisuels (vidéos issues d'Internet).

La métrique utilisée lors des campagnes NIST-SRE est la fonction du coût de détection (Detection Cost Function (DCF)) qui correspond à la somme des coûts des deux types d'erreur (*FR* et *FA*) en prenant en compte la probabilité a priori de rencontrer une comparaison cible. Son calcul est défini par l'équation 1.7.

$$DCF = C_{FR} P_{PR} P_{cible} + C_{FA} P_{FA} (1 - P_{cible}) \tag{1.7}$$

Synthèse du chapitre

La reconnaissance des locuteurs, qu'elle soit automatique ou humaine, est une tâche complexe. Pour les êtres humains, la reconnaissance des locuteurs vise à créer une représentation mentale de la voix d'un individu. En effet, ils construisent un prototype générique à partir des voix déjà rencontrées et, lorsqu'ils entendent une voix nouvelle, ne mémorisent que les caractéristiques qui s'écartent de ce prototype. Cependant, les êtres humains sont sujets à des erreurs de reconnaissance des locuteurs. La phonagnosie nous apprend que la reconnaissance des locuteurs repose sur la perception de caractéristiques acoustiques indépendamment du message linguistique délivré.

En informatique, les débuts de la reconnaissance des locuteurs se font dans les années 1960, dans des expériences en milieu fermé à partir de mots isolés. Un premier essor est observé dans les années 2000 avec l'approche UBM-GMM ainsi que l'adoption du *LR* pour la prise de décision. Depuis quelques années maintenant, les systèmes de reconnaissance des locuteurs sont basés sur des réseaux de neurones profonds, notamment avec l'architecture *x-vector* qui est aujourd'hui l'état de l'art. Les progrès technologiques des systèmes de RAL sont également observés dans les campagnes d'évaluation NIST-SRE.

2 | Contexte de la comparaison de voix en criminalistique

Résumé : *La comparaison de voix est un processus qui vise à comparer deux enregistrements de voix afin de déterminer s'ils ont été produits par la même personne. Toutefois, de nombreux facteurs de variabilité peuvent être présents dans les pièces à comparer. Ces facteurs peuvent être liés au locuteur, au contenu linguistique, aux conditions d'enregistrement ou encore au support numérique. Ils peuvent influer sur le résultat d'une comparaison de voix. Il est donc nécessaire de les prendre en compte afin d'en garantir la fiabilité. Ce chapitre passe en revue le statut de la voix en criminalistique ainsi que les cadres scientifiques et légaux de la comparaison de voix.*

Sommaire

2.1 La comparaison de voix en criminalistique : cadre légal et scientifique

Depuis 2000, l'Autorité de régulation des communications électroniques, des postes et de la distribution de la presse (ARCEP) lance une enquête annuelle intitulée « Baromètre du numérique » sur l'usage numérique de la population française. Ces enquêtes montrent que depuis 2011, la part des personnes de plus de 12 ans résidant en France et possédant un téléphone mobile ne cesse d'augmenter et jusqu'à atteindre 94 % en 2020 (Berhuet *et al.*, 2021). Cette démocratisation des téléphones mobiles conduit à un accroissement du nombre d'affaires judiciaires dans lesquelles une interception téléphonique est impliquée. Ces interceptions téléphoniques peuvent donner lieu à une comparaison de voix afin d'identifier l'auteur d'un propos. Le recours à la comparaison de voix est également très demandé dans le cadre de sonorisation de domicile ou de véhicule. Le traitement des interceptions téléphoniques et sonorisations d'espaces sont régies par la Plate-forme nationale des interceptions judiciaires (PNIJ) qui elle-même est encadrée par le Code de procédure pénale (CPP) et la Commission nationale de l'informatique et des libertés (CNIL).

Dans ce chapitre, nous nous intéressons à la comparaison de voix dans le cadre judiciaire. Dans un premier temps, les définitions relatives à la criminalistique, l'identification et l'individualisation sont présentées. Puis, la question de la voix en tant qu'indice biométrique est posée. Pour cela, nous définissons les propriétés d'un trait biométrique et estimons si la voix entre dans cette définition. Puis, nous présentons le cadre légal dans lequel une comparaison de voix peut être effectuée. Enfin, nous en définissons le cadre scientifique en évoquant les problématiques liées aux différentes techniques utilisées et la nécessité de suivre des méthodes scientifiques.

2.1.1 Définitions

2.1.1.1 Criminalistique

La criminalistique s'intéresse à l'étude des traces témoignant d'une activité délictueuse ou criminelle en appliquant des méthodes et des techniques scientifiques. Il s'agit avant tout d'un processus de comparaison entre un élément, appelé trace, et une pièce de référence appelée pièce de comparaison. Elle permet notamment – mais pas uniquement – la reconnaissance, l'identification ou l'individualisation de personnes.

La criminalistique voit le jour au début du XX^{e} siècle sous l'égide d'Edmond Locard, professeur de médecine légale, par ailleurs fondateur du premier laboratoire de police scientifique suivant les méthodes d'anthropométrie dédiées à l'identification criminelle, mises en place par Alphonse Bertillon.

> **Principe de Locard**
>
> Nul ne peut agir avec l'intensité que suppose l'action criminelle sans laisser de multiples marques de son passage.
>
> – Edmond Locard

Locard indique qu'il existe un lien direct entre le responsable d'une action criminelle et un objet physique présent sur la scène de crime, cet objet permettant de confondre le coupable au cours de l'investigation. Les marques laissées par le coupable, les traces, peuvent être physiques, analogiques ou numériques, et peuvent prendre différentes formes : traces papillaires (dont font partie les empreintes digitales), empreintes de pas, enregistrements de vidéosurveillance, ou encore enregistrements sonores.

2.1.1.2 Identification et individualisation

Les œuvres de fiction montrent souvent que l'objectif de la criminalistique est de relier une trace à une identité. Or, la notion d'identité comme entendu en criminalistique n'est pas aussi tranchée. En effet, Locard admet le concept d'identité comme étant « l'ensemble

des caractères par lesquels un homme définit sa personnalité propre et se distingue de tout autre » (Locard, 1931). Cependant, pour Kwan et Meuwly, le terme « identité » est ambigu du fait des définitions ambivalentes qui lui sont données (Kwan, 1977; Meuwly, 2006). Ils distinguent deux types d'identité :

Identité qualitative La trace contient des caractéristiques spécifiques à des classes d'individus. Cela implique que plusieurs sources peuvent être à l'origine de la trace, et qu'il n'est donc pas possible de la relier à une seule personne. Il s'agit d'un processus de classification.

Identité numérique La trace contient des caractéristiques spécifiques à un individu en particulier. L'analyse permet de mettre en évidence un unique individu comme étant la source de la trace. Il s'agit d'un processus d'individualisation.

La figure 2.1 présente la différence entre les processus de classification et d'individualisation. Alors que la classification permet de déterminer que le véhicule est une DeLorean DMC-12, la plaque d'immatriculation permet de l'individualiser et d'en connaître le possesseur.

(a) *Classification* (b) *Individualisation*

Figure 2.1 – *Classification et individualisation : la plaque d'immatriculation permet d'individualiser le véhicule*

2.1.2 La voix, un trait biométrique ?

L'identification des individus repose sur des méthodes variées :

1. Connaissance : seul l'individu détient l'information nécessaire à son identification. C'est le cas de l'authentification par mot de passe ou par code, *etc.*

2. Possession : seul l'individu détient l'objet nécessaire à son identification. C'est le cas des badges d'accès, des clefs, *etc.*

3. Biométrie : l'individu est identifié à partir de ses caractéristiques biologiques et/ou comportementales. C'est le cas des empreintes digitales, de l'iris, de la signature, *etc.*

Certaines d'entre elles sont utilisées au quotidien. Ces moyens peuvent reposer sur une information que *théoriquement* seule la personne détient. C'est notamment le cas de l'authentification par mot de passe (avec ou sans identifiant), nécessaire pour accéder à des contenus dédiés. Les paiements par carte bleue nécessitent de connaître le code associé à la carte. Dans certaines entreprises, un badge est nécessaire pour entrer dans certains espaces. Durant un voyage, dans les aéroports, une identification par empreinte digitale est parfois nécessaire afin de vérifier que la personne détentrice du passeport est bel et bien celle qui l'utilise.

Cette dernière caractéristique repose sur un trait biométrique. Les traits biométriques comportent des caractéristiques variées telles que les empreintes papillaires, l'Acide désoxyribonucléique (ADN), l'iris, la démarche, la forme des mains et des doigts, la répartition de la chaleur dans le corps, la signature, *etc.* (Jain *et al.*, 2007). La biométrie s'intéresse à l'unicité d'un individu, ce qui le rend unique par rapport aux autres individus. En France, le Fichier automatisé des empreintes digitales (FAED) et le Fichier national automatisé des empreintes génétiques (FNAEG) sont deux bases de données biométriques qui permettent de faciliter la recherche et l'identification des individus. Le FAED contient les empreintes digitales des personnes condamnées pour des infractions pénales. Le FNAEG contient les informations génétiques utilisées dans la résolution de crimes et délits. Ces deux fichiers sont consultables par les services de police et de gendarmerie, mais aussi par les services de renseignement et de sécurité.

Les traits biométriques peuvent être purement biologiques (telles les empreintes papillaires et génétiques), mais elles peuvent également être comportementales comme la démarche ou encore l'écriture (Jain *et al.*, 2007).

2.1.2.1 Définition d'un trait biométrique

Selon Jain *et al.* et Maltoni *et al.*, pour être un trait biométrique, toute caractéristique physique et/ou comportementale doit remplir les critères suivants (Jain *et al.*, 1999 ; Maltoni *et al.*, 2009) :

1. Universalité : tout individu doit posséder cette caractéristique ;
2. Spécificité : la caractéristique doit être suffisamment différentes entre les individus afin de les distinguer ;
3. Permanence : la caractéristique est stable au cours du temps ;
4. Quantification : la caractéristique peut être mesurée quantitativement.

Dans les cas pratiques d'utilisation biométrique, de nouvelles contraintes s'ajoutent :

1. Performance : le système d'identification doit avoir un taux de reconnaissance satisfaisant tout en étant rapide et en utilisant aussi peu de ressources que nécessaires ;
2. Acceptabilité : l'utilisation au quotidien du trait biométrique doit être acceptée de toutes et tous ;
3. Contournement : les systèmes biométriques doivent pouvoir faire face aux tentatives de tromperie.

Selon Jain *et al.*, tous les traits biométriques remplissent les conditions, mais pas forcément dans les mêmes proportions comme indiqué dans la figure 2.2. Jain *et al.* considèrent la voix comme étant un trait biométrique dont la permanence et la spécificité sont plutôt basses, aisément contournables, et dont les systèmes de reconnaissance délivrent un taux de reconnaissance qui n'est pas aussi satisfaisant que pour d'autres traits tels que les empreintes papillaires.

Tous les traits biométriques suivent le même processus de comparaison. Dans un premier temps, des paramètres sont extraits du trait biométrique et la pièce de comparaison, puis ces paramètres sont comparés à ceux d'une base de données. Enfin, un score de similarité est calculé entre les paramètres extraits et ceux de la base de données. La figure 2.3 illustre le processus de comparaison entre deux traits biométriques. Les traits biométriques sont comparés par paramétrisation et comparaison. La paramétrisation consiste à extraire des informations (paramètres ou *features*). La comparaison consiste à comparer les paramètres extraits entre les deux traits biométriques tout en les comparant à une large base

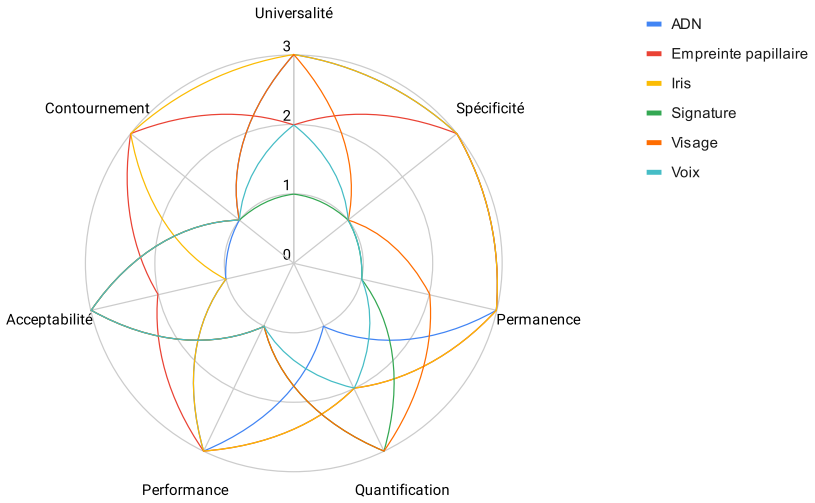


Figure 2.2 – *Différences entre quelques traits biométriques, dont la voix. D'après (Jain et al., 2007).*

de données. La base de données permet d'estimer les traits qu'ils partagent avec une plus grande population ainsi que les traits qui leur sont propres. Le résultat de cette comparaison indique la similarité entre les traits biométriques comparés (Li et Jain, 2015).

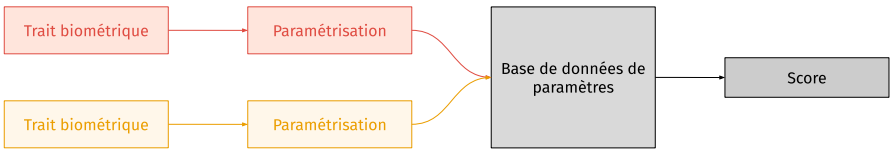


Figure 2.3 – *Schéma illustrant la comparaison entre deux traits biométriques*

2.1.3 Le procès pénal

Lorsqu'une infraction a été commise, le CPP prévoit un cadre juridique qui régit la découverte de l'auteur de l'infraction, sa poursuite et son jugement. De la commission de l'infraction au jugement, la procédure pénale en France reconnaît quatre phases :

1. la phase d'enquête ;

2. la phase de poursuites ;
3. l'instruction préparatoire ;
4. le jugement.

2.1.3.1 La phase d'enquête

La phase d'enquête débute après constatation d'une infraction. La police judiciaire est alors saisie après une plainte, une dénonciation, d'initiative ou sur instruction d'un magistrat. La phase d'enquête est menée par un Officier de police judiciaire (OPJ), un Agent de police judiciaire (APJ) ou un Agent de police judiciaire adjoint (APJA) dont, conformément à l'article 14 du CPP, la mission est de :

- constater les infractions ;
- rassembler les preuves ;
- rechercher les auteurs.

Deux cadres juridiques régissent la phase d'enquête : l'enquête de flagrance et l'enquête préliminaire.

Enquête de flagrance Lorsque l'infraction (crime ou délit punissable d'une peine d'emprisonnement) est en train de se commettre ou vient juste de se commettre, ou si l'auteur présumé est poursuivi par la clameur publique, peu de temps après la commission de l'infraction, s'il est en possession d'indices ou d'objets qui l'accusent, une enquête de flagrance est ouverte pour une durée de huit jours renouvelable une fois. L'enquête de flagrance permet la coercition.

Enquête préliminaire Contrairement à l'enquête de flagrance, l'enquête préliminaire n'est pas coercitive. Elle n'est pas non plus limitée dans le temps et peut être déclenchée pour des contraventions ainsi que des crimes et délits que ne sont pas flagrants.

Quel que soit le cadre juridique, la police judiciaire est alors sous la direction du ministère public ou du procureur de la République.

Parmi les actes et les compétences conférés à l'OPJ, l'APJ ou l'APJA se trouvent entre autres la saisine, l'interpellation, la garde à vue, la conservation des traces et indices et les réquisitions.

2.1.3.2 La phase de poursuites

Après la phrase d'enquête, le procureur de la République peut décider ou non de poursuites. Si poursuites il y a, une information judiciaire est ouverte et est confiée à un juge d'instruction qui dispose désormais de tous les pouvoirs pour mener les investigations. Dans le cadre d'une commission rogatoire, le juge d'instruction peut demander à un OPJ d'accomplir des actes d'instruction.

Si le procureur de la République décide de ne pas engager de poursuites, il peut classer l'affaire sans suite ou proposer une alternative aux poursuites (rappel à la loi, participation à une médiation, *etc.*).

2.1.3.3 L'instruction préparatoire

L'instruction préparatoire vise à récolter des indices en vue d'identifier la ou les personne(s) susceptible(s) d'avoir commis l'infraction, et ainsi déterminer si elles l'ont commise. Si les indices récoltés constituent des charges suffisantes, ces personnes sont alors renvoyées devant une juridiction de jugement. L'instruction préparatoire est menée à charge et à décharge.

L'article 79 du CPP indique que l'instruction préparatoire est obligatoire lorsque l'infraction est un crime. Pour les délits et les contraventions, elle est facultative.

Afin d'ouvrir l'instruction, le juge d'instruction doit être saisi par le procureur de la République ou par une plainte avec constitution partie civile. Le juge d'instruction dirige alors les investigations. Une partie des actes d'investigations peuvent être directement effectués par le juge d'instruction. C'est le cas de l'interrogatoire des personnes mises en examen, des auditions, des confrontations, des perquisitions et des saisies. En revanche, pour les

actes spécifiques qu'il n'est pas en mesure d'assurer, il peut procéder à une délégation des pouvoirs à un autre magistrat ou à un OPJ : c'est la commission rogatoire.

2.1.3.4 La phase de jugement

Au terme de l'instruction préparatoire, si le juge d'instruction estime que les charges sont suffisantes, il peut émettre une ordonnance de renvoi (pour les contraventions et délits) ou de mise en accusation (pour les crimes).

Dans le cas d'une ordonnance de renvoi, au sujet des contraventions, la phase de jugement s'opère dans un tribunal judiciaire où siègent un juge unique ainsi que le ministère public. Pour les délits, le tribunal correctionnel est saisi et présidé par un juge d'instruction, deux assesseurs et le procureur de la République.

Dans le cas d'une ordonnance de mise en accusation, la phase de jugement a lieu en cour d'assises, présidée par un juge d'instruction accompagné de deux assesseurs, six jurés et un procureur général ou un avocat général.

Pour chacun des cas, après un premier verdict, il est possible de faire appel auprès de la cour d'appel pour les contraventions et les délits, ou de la cour d'assises en appel pour les crimes. En cas de désaccord avec le verdict de l'appel, l'affaire est pourvue en cassation. La Cour de cassation étudiera la procédure pénale employée et peut annuler les décisions de justice ou bien rejeter le pourvoi.

2.1.4 La comparaison de voix dans le procès pénal

Au cours du procès pénal, un expert peut être saisi pour effectuer des actes sous le contrôle de l'OPJ, l'APJ ou l'l'APJA lors de la phase d'enquête, de l'OPJ lors de l'instruction préparatoire en commission rogatoire ou bien du juge d'instruction lors de l'instruction préparatoire.

Tout au long de l'instruction, des enregistrements contenant une ou plusieurs voix peuvent être prélevés. Les canaux peuvent être multiples : écoutes téléphoniques, messages vocaux

ou piste audio d'une vidéo. Il peut être alors nécessaire de procéder à une comparaison de voix en vue de permettre l'identification ou non de la ou des personnes qui parlent dans ces enregistrements. Une comparaison de voix peut être demandée sous différentes formes régies par le CPP : la réquisition à personne qualifiée et l'ordonnance de commission d'expertise.

2.1.4.1 Réquisition à personne qualifiée

Les réquisitions à personne qualifiée permettent à un OPJ de faire appel à des personnes physiques ou morales inscrites près des tribunaux pour effectuer des actes spécifiques. Lors d'une réquisition à personne qualifiée [1], pour le cas de la police scientifique, le chef du SNPS [2] est saisi par l'enquêteur, le procureur de la République ou le magistrat instructeur dans le cadre d'une commission rogatoire. Les réquisitions peuvent être demandées dans le cadre :

- de l'enquête de flagrance, et sont alors régies par l'article 60 du CPP ;
- de l'enquête préliminaire, et sont alors régies par l'article 77-1 du CPP ;
- de la commission rogatoire, et sont alors régies par les articles 81 alinéas 4, 151 et 152 du CPP.

La réquisition comprend une mission qui indique à la personne qualifiée les actions à effectuer. Une comparaison de voix peut être accompagnée de l'exploitation d'un enregistrement en vue de déterminer si ce dernier est exploitable. Dans certains cas, un prélèvement de voix peut s'avérer nécessaire afin d'obtenir une pièce de comparaison.

Le chef du SNPS désigne une personne qualifiée pour effectuer la comparaison de voix. Cette personne qualifiée doit rédiger un rapport technique mentionnant de façon exhaustive la méthode employée pour déterminer la faisabilité d'une comparaison de voix d'une part, et, si cette dernière est possible, procéder au prélèvement de voix si nécessaire ainsi qu'à ladite comparaison. Le rapport présente le résultat de la comparaison de voix et est accompagné d'une prestation de serment.

1. La personne qualifiée peut être travailler dans une institution telles que le SNPS ou l'IRCGN, ou être
2. Suivant les cas, le chef d'un Service régional de police technique et scientifique (SRPTS) ou d'une Antenne régionale de police technique et scientifique (ARPTS) peut être saisi.

2.1.4.2 Ordonnance de commission d'expertise

Dans le cadre d'une ordonnance de commission d'expertise, le chef du SNPS est saisi par un magistrat instructeur. L'ordonnance comprend une mission qui indique à l'expert les actions à effectuer. À l'instar d'une réquisition, une comparaison de voix peut être accompagnée de l'exploitation d'un enregistrement en vue de déterminer si une comparaison de voix est possible, et d'un prélèvement de voix pour constituer une pièce de comparaison. L'expert accomplit le travail décrit dans la mission et rédige un rapport d'expertise qui indique de façon exhaustive la méthode employée pour déterminer la faisabilité d'une comparaison de voix d'une part, et, si cette dernière est possible, procéder au prélèvement de voix si nécessaire ainsi qu'à ladite comparaison.

En France, les conclusions de la comparaison de voix reposent sur une interprétation verbale du rapport de vraisemblance. Cette interprétation permet de rendre accessible la valeur du rapport de vraisemblance, *renforçant très fortement*, *renforçant fortement*, *renforçant*, *renforçant légèrement* soit l'hypothèse que les enregistrements proviennent de la même personne, soit l'hypothèse selon laquelle les enregistrements proviennent de deux personnes différentes, voire ne renforçant aucune des deux hypothèses.

Le rapport d'expertise est accompagné d'une prestation de serment et d'une attestation de mission.

2.1.5 Comparer les voix de façon empirique

La première affaire relative à l'identification d'une personne par la voix et dont il reste des traces remonterait au XVII^{e} siècle (Eriksson, 2012). Au vu du niveau technique de l'époque, l'identification des personnes à leur voix ne pouvait reposer que sur la perception auditive.

Cas réel · Meurtre de Charles I^{er} d'Angleterre

Le 30 janvier 1649, Charles I^{er} d'Angleterre est victime d'un régicide par décapitation. Le meurtrier a utilisé des artifices tels qu'un masque et une perruque pour camoufler son

visage. Un témoin affirme qu'il s'agit d'un individu portant le nom de William Hulet. Il l'a reconnu « à sa voix ». Suite à ce témoignage, William Hulet a été jugé coupable de haute trahison et a été condamné à mort.

Depuis le XVII^{e} siècle, la reconnaissance des personnes à leur voix a connu des avancées scientifiques et techniques notables entraînant des changements dans les approches et pratiques. Durant le XX^{e} siècle, plusieurs méthodes de comparaison de voix ont été utilisées. Cependant, la voix d'un individu connaît une très grande variabilité, ce qui ajoute de la difficulté à la comparaison de voix. De plus, le support et les conditions d'enregistrement ajoutent davantage de variabilité, et ce, quelle que soit la méthode employée. Ces facteurs de variabilité sont présentés dans la figure 2.4.

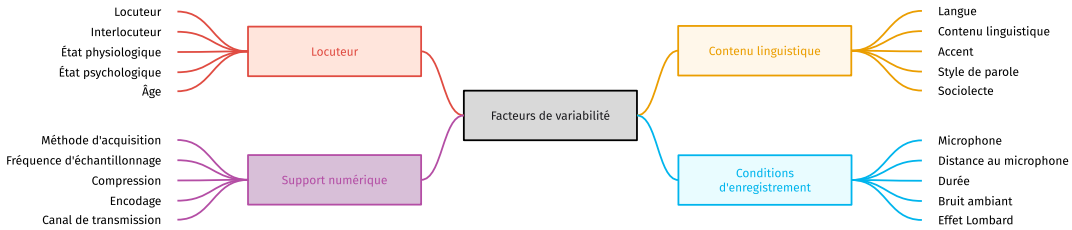


Figure 2.4 – *Exemples de facteurs de variabilité de la voix*

2.1.5.1 De la fiabilité du tapissage vocal

La question de la reconnaissance des voix, et plus particulièrement sa fiabilité, est une des problématiques rencontrée en criminalistique. C'est notamment le cas du tapissage vocal. Si un témoin auditif a entendu la voix du criminel, mais n'est pas en mesure de l'identifier formellement, un tapissage vocal peut être demandé. Cela consiste à écouter la voix de plusieurs individus et de la confronter à celle à laquelle le témoin a été exposé pour déterminer s'il s'agit de la même voix. Plusieurs études estiment qu'entre 9 et 24 % des témoins auditifs sont capables d'identifier la voix avec succès (Kerstholt *et al.*, 2006 ; Öhman *et al.*, 2010 ; Öhman *et al.*, 2013a ; Öhman *et al.*, 2013b ; Olsson *et al.*, 1998 ; Yarmey *et al.*, 1994).

La fiabilité du témoin dépend de plusieurs facteurs. Premièrement, comme énoncé précédemment, les faits et la séance d'identification ne doivent pas être trop éloignés dans le temps ainsi que le suggère l'expérience de McGehee, mais aussi celles de Yarmey *et al.* et Kerstholt *et al.* (Kerstholt *et al.*, 2004 ; McGehee, 1937 ; Yarmey *et al.*, 1994). La langue et l'accent ont également un effet sur le taux de bonne identification ainsi que le montrent les études de Thompson, de Stevenage *et al.* et de Philippon *et al.* (Kerstholt *et al.*, 2006 ; Philippon *et al.*, 2007 ; Stevenage *et al.*, 2012 ; Thompson, 1987). Le temps d'exposition à la voix du criminel est également un facteur qui influe sur le taux de bonne identification (Kerstholt *et al.*, 2004). Kerstholt *et al.* montre qu'un taux d'identification correct à 24 % lorsque les auditeurs ont l'opportunité d'indiquer qu'aucune des voix ne semble correspondre à la voix entendue. Ce taux grimpe à 34 % dès lors qu'elles ou ils sont contraints de faire un choix (Kerstholt *et al.*, 2006). En conditions expérimentales, le taux de mauvaise identification peut atteindre 98 % ou 99 % (van Wallendael *et al.*, 1994 ; Yarmey *et al.*, 1994), bien que d'autres études montrent un taux moindre de 50 % à 85 % (Kerstholt *et al.*, 2004 ; Öhman *et al.*, 2011 ; Philippon *et al.*, 2007 ; Smith et Baguley, 2014 ; Stevenage *et al.*, 2012 ; Yarmey, 2001).

Cas réel · David Shepard

Une mauvaise identification peut amener à l'emprisonnement d'une personne sur la base d'éléments faux ou peu fiables. Ce fut le cas de David Shephard, accusé de viol, cambriolage, port d'armes et menaces terroristes.

Le 24 novembre 1983, une femme est enlevée sur un parking par deux hommes qui quittent les lieux avec la voiture de la victime. Ils la violent, puis la font descendre de la voiture avant de s'enfuir avec. La victime procède à une séance d'identification de voix et identifie David Shephard. Ce dernier est condamné à 30 ans de prison.

En 1992, Shepard demande une analyse ADN sur les échantillons prélevés sur la victime et ses vêtements. Alors que les échantillons prélevés sur la victime ne permettent pas de le disculper, les échantillons prélevés sur les vêtements indiquent que Shepard ne fait pas partie des deux hommes qui ont commis l'assaut. Shepard sort de prison le 25 avril 1985 après y avoir passé plus de 11 ans (Innocence Project, 2016b).

Innocence Project

L'organisme *Innocence Project*[3] a pour mission de reprendre les affaires des personnes condamnées et demander des (contre-)expertises en vue d'invalider les éléments à charge mettant ainsi fin à leur condamnation. À ce jour, aux États-Unis, l'organisme recense 8 personnes libérées alors qu'elles furent condamnées sur la base d'identifications vocales :

James Waller	24 ans de prison	(Innocence Project, 2016e)
Michael Anthony Williams	24 ans de prison	(Innocence Project, 2016g)
Sedrick Courtney	16 ans de prison	(Innocence Project, 2016h)
Eduardo Velasquez	13 ans de prison	(Innocence Project, 2016d)
Dean Cage	12 ans de prison	(Innocence Project, 2016c)
David Shepard	11 ans de prison	(Innocence Project, 2016b)
Kerry Kotler	10 ans de prison	(Innocence Project, 2016f)
Brian Piszczek	3 ans de prison	(Innocence Project, 2016a)

2.1.5.2 L'« empreinte vocale » : l'identité des locuteurs en une image

En 1962, Kersta introduit le terme « empreinte vocale[4] », qui par analogie à l'empreinte digitale, permet une identification formelle des individus à leur voix (Kersta, 1962). Ce que Kersta appelle « empreinte vocale » correspond à la représentation du signal de parole sous forme de spectrogramme. Un spectrogramme est une représentation du signal qui permet de visualiser les fréquences présentes dans le signal, leur intensité ainsi que leur distribution dans le temps. Pour le signal de parole, le spectrogramme permet de rendre compte de l'articulation. La figure 2.5 montre le spectrogramme pour l'énoncé « j'ai dix-neuf ans ». L'axe horizontal représente la temporalité, l'axe vertical représente les fréquences du signal. La coloration plus ou moins prononcée des fréquences représente leur intensité.

3. https://innocenceproject.org
4. Traduction de l'anglais *voiceprint*

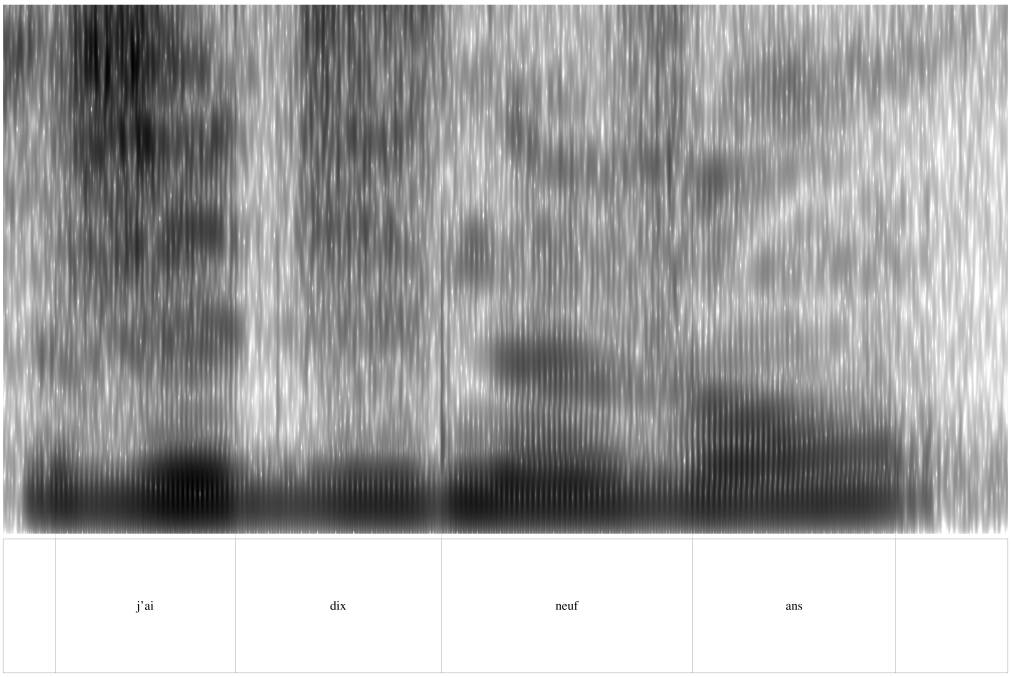


Figure 2.5 – *Spectrogramme pour l'énoncé « j'ai dix-neuf ans »*

En 1973, plusieurs scientifiques ont émis une critique des travaux de Kersta en remettant en cause la fiabilité de l'« empreinte vocale » (Bolt *et al.*, 1973). Les auteurs indiquent que les ressemblances entre les spectrogrammes résultent majoritairement du contenu linguistique et de l'anatomie de l'appareil phonatoire. Tosi a publié en 1979 une étude montrant que les êtres humains sont en mesure d'identifier les locuteurs en visualisant les spectrogrammes de leurs productions orales (Tosi, 1979).

À partir de la fin des années 1990, certains membres de la communauté scientifique s'accordent à écarter l'« empreinte vocale » des tribunaux pour son manque de fiabilité et de fondements scientifique. C'est le cas d'associations telles que l'Association francophone de la communication parlée (AFCP) (Association Francophone de la Communication Parlée (AFCP), 2002), dans la lignée de la motion de 1990 suivie de la pétition lancée par le Groupe francophone de la communication parlée (GFCP) en 1999 (GFCP, 1999). Plusieurs scientifiques insistent également sur la nécessité d'utiliser des méthodes validées scientifiquement pour

la comparaison de voix (Boë, 2000 ; Boë *et al.*, 1999 ; Bonastre, 2020 ; Bonastre *et al.*, 2003). En 2009, Campbell *et al.* indiquent que le terme « empreinte vocale » ne doit pas être utilisé dans le cadre judiciaire (Campbell *et al.*, 2009).

De nos jours, malgré que l'avis de la communauté scientifique soit que le concept d'empreinte ne s'applique pas à la voix, le terme « empreinte vocale » est encore très largement utilisé même s'il ne fait plus référence aux spectrogrammes, il fait aujourd'hui référence aux modèles conçus pour la vérification du locuteur. Il est possible de le retrouver dans des encyclopédies grand public (Gaurenne *et al.*, 2022). Le terme est également présent dans des solutions de traitement de la parole construites par des entreprises telles que Google [5] ou encore Amazon [6].

2.1.5.3 Analyse phonético-linguistique de la voix

En criminalistique, la comparaison de voix a également été effectuée par une analyse linguistique combinée à une analyse phonétique. L'enjeu pour les phonéticiens est de déceler d'éventuelles particularités spécifiques aux locuteurs telles que des erreurs grammaticales, des troubles du langage ou encore un accent ou un dialecte non standards. Au niveau phonétique, les formants et de la fréquence fondamentale sont étudiés. Le résultat d'une comparaison de voix en utilisant l'approche phonético-linguistique se présente sous la forme d'un *LR* dont la valeur est entièrement subjective et dépendante de l'analyste.

2.1.6 Batvox : un système de comparaison de voix pour la criminalistique

Batvox est un logiciel édité par l'éditeur espagnol Agnitio pour la comparaison de voix dans le cadre judiciaire. Le logiciel intègre de multiples fonctionnalités pour la comparaison de voix en criminalistique :

- gestion des affaires ;
- description de la qualité des enregistrements audio afin de déterminer s'ils sont exploitables ou non ;

5. https://support.google.com/assistant/answer/7394306
6. https://aws.amazon.com/fr/connect/voice-id

- banque d'échantillons de voix enregistrés dans des conditions variées en vue de former un modèle aussi proche que possible des données à analyser ;
- génération d'un rapport récapitulatif.

La première version de Batvox est basée sur les GMM. Elle intègre une normalisation des signaux audio permettant une meilleure performance. La deuxième version du logiciel permet une amélioration de la performance grâce à des méthodes de compensation du canal. La troisième version du logiciel intègre de nombreuses fonctionnalités :

- approche UBM-GMM ;
- JFA pour limiter les effets du locuteur et du canal ;
- identification du genre du locuteur ;
- diarization.

En 2012, la quatrième version de Batvox est publiée. Elle comprend un changement majeur : l'utilisation de l'architecture *i-vector* pour la RAL. Le modèle JFA est également amélioré. La vitesse d'exécution de Batvox en est grandement réduite. La performance de Batvox 3 (UBM-GMM) et de Batvox 4 (*i-vector*) est évaluée dans des conditions semblables à celles rencontrées en criminalistique (van der Vloed, 2016 ; Zhang et Tang, 2018). Les résultats montrent que Batvox 4 est plus performant que Batvox 3. En France, c'est à partir de la version 4 que Batvox est utilisé dans le cadre de la criminalistique.

En plus de Batvox, le logiciel Vocalise développé par la société *Oxford Wave Research* est également utilisé à des fins criminalistiques.

Cas réel · L'affaire Cahuzac

En 2012, Jérôme Cahuzac est ministre délégué au Budget auprès du ministre de l'Économie et des Finances. Le 4 décembre 2012, le journal indépendant Mediapart révèle 2021 que Cahuzac possède un compte non déclaré en Suisse. Le journal mentionne un écrit d'un agent du fisc adressé à sa hiérarchie dans lequel il est mention d'un « compte bancaire à numéro en Suisse » ouvert par Cahuzac. Ce dernier dément les accusations.

Le lendemain, Mediapart met en ligne un enregistrement audio présenté comme une conversation entre Jérôme Cahuzac et son gestionnaire de fortune. Alors interrogé sur le sujet, le ministre continue de nier les faits au sein de l'Assemblée nationale.

Le 7 décembre, le parquet de Paris ouvre une enquête préliminaire suite à la plainte du ministre contre Mediapart pour diffamation.

Un mois plus tard, le 8 janvier 2013, à la demande d'Edwy Plenel, fondateur du journal Mediapart, une enquête préliminaire pour « blanchiment de fraude fiscale » est ouverte par le parquet de Paris.

Le 24 janvier 2013, une expertise de la police scientifique, menée sur l'enregistrement sonore, indique que ce dernier n'est pas un montage. Une seconde expertise menée en février 2013 conclut que les voix sont celles de Jérôme Cahuzac et son gestionnaire de fortune, Hervé Dreyfus. Cette expertise repose sur une comparaison de voix effectuée avec le logiciel Batvox ainsi qu'une comparaison phonétique.

> *« Sur 3 minutes 40, il y a quelques secondes où cela pourrait être moi, mais il se trouve que ce n'est pas moi. [...] Des proches ont écouté la bande et ne m'ont pas reconnu. »*
> – Jérôme Cahuzac, à propos de l'enregistrement audio

2.2 Des méthodes scientifiques au service de la justice

2.2.1 De l'expertise à l'expertise scientifique : les normes Frye et Daubert

Dans le cadre légal, une expertise scientifique découle directement de savoirs et techniques scientifiques. En conséquence, la méthode scientifique est appliquée : une hypothèse est dressée puis testée. Néanmoins, les méthodes employées dans les laboratoires doivent respecter certaines règles et standards pour être admissibles devant un tribunal. Ces règles et standards sont définis par les normes Frye et Daubert.

Cas réel · Affaire Trayvon Martin

La perception de l'identité du locuteur peut se retrouver au cœur d'affaires criminelles comme le montre l'affaire Trayvon Martin. Le 26 février 2012, dans un quartier résidentiel, Trayvon Martin, un adolescent afro-américain âgé de 17 ans, meurt d'un coup de feu tiré par George Zimmerman, un policier latino-américain. L'affaire Trayvon Martin prit de l'ampleur notamment sur le caractère supposé raciste du crime. Peu avant le coup de feu, une altercation a lieu entre les deux hommes. Des riverains, alertés par la situation, appellent les secours. Les enregistrements effectués par les services de secours mettent en évidence des hurlements d'appels à l'aide provenant de l'altercation. Cependant, la personne à l'origine de ces hurlements n'est pas identifiée. La famille de Martin soutient qu'il s'agit de la voix de l'adolescent, et la famille de Zimmerman maintient qu'il s'agit du policier.

Plusieurs expertises, mandatées par les deux parties et des médias, ont été conduites sur les enregistrements audio afin de déterminer l'auteur des appels à l'aide. Les conclusions des expertises, présentées lors du procès de Zimmerman, sont mitigées. Tom Owen, l'un des experts qui ont comparé un échantillon de voix parlée de Zimmerman avec les cris issus des enregistrements des secours, indique que Zimmerman n'est pas à l'origine des hurlements. Alan Reich, un autre expert, affirme que les cris proviennent de Martin. Peter French et George Doddington, deux scientifiques spécialisés dans la comparaison de voix, soulignent l'« absurdité » que représente la comparaison entre une voix hurlée et une voix parlée.

La question de la fiabilité des expertises vocales et plus particulièrement de leur acceptation au sein de la communauté scientifique se pose. La considération de ces expertises est laissée à l'appréciation de la juge :

> *There is no evidence to establish that their scientific techniques have been tested and found reliable.*[7]
> – Debra S. Nelson, juge lors du procès de George Zimmerman à propos des expertises de comparaison de voix

De plus, la juge indique qu'une comparaison entre une voix parlée et une voix hurlée ne sont pas envisageables.

> *The Court accepts the opinions of [defense witnesses] that reliable comparison of normal speech to the screams in the 911 call is not possible.*[8]
> – Debra S. Nelson, juge lors du procès de George Zimmerman à propos des expertises de comparaison de voix

Les expertises de comparaison de voix sont alors écartées des éléments de preuve du procès. Inculpé pour homicide non prémédité, Zimmerman plaide non coupable, ce que le jury reconnaîtra après seize heures de délibération.

2.2.1.1 La norme Frye

Cas réel · *Frye v. United States*

En 1923, James Alphonzo Frye, accusé de meurtre, fait appel à la Cour fédérale de Columbia (États-Unis)[9]. Au cours du procès, un expert intervient au nom de la défense en indiquant que l'accusé n'avait pas menti. Pour cela, il a recours à un détecteur de mensonges basé sur la mesure de la pression systolique. Selon l'expert, le fait de dire la vérité ne suppose aucun effort alors que le mensonge demande un certain effort qui se reflète dans la pression artérielle. Ainsi, le mensonge induit une élévation caractéristique de la pression systolique.

Cependant, la Cour fédérale de Columbia conteste la recevabilité de l'expertise, car la méthode employée n'était pas généralement admise par la communauté scientifique.

> *Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the*

7. « Rien ne permet d'établir que les techniques scientifiques utilisées ont été testées et demeurent fiables. »

8. « La Cour accepte [le point de vue de la défense] quant au manque de fiabilité d'une comparaison entre une voix normale et les hurlements de l'appel au 911. »

> *evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.*
> *– Frye v. United States*

La norme Frye définit les conditions de recevabilité d'une expertise scientifique et indique que les méthodes employées doivent être « généralement acceptées » par un ensemble significatif de chercheurs de la discipline associée. Son objectif est de déterminer si une méthode est suffisamment fiable pour être admise devant un tribunal. Cela présuppose deux choses :

1. définir l'ensemble significatif de chercheurs ;
2. évaluer les publications scientifiques dans le but de déterminer le consensus sur les méthodes employées.

Il s'agit du premier standard d'admissibilité des expertises scientifiques.

2.2.1.2 *Federal Rule of Evidence 702*

En 1975, la Cour de justice états-unienne revient sur la définition de la norme Frye dans l'affaire *United States v. Downing* [10] et la remplace avec la *Federal Rule of Evidence 702*. En effet, les tribunaux états-uniens ont tendance à interpréter la norme Frye de façon restrictive, ce qui a pour conséquence de limiter l'admissibilité des expertises scientifiques. Par conséquent, la Cour de justice états-unienne décide de la remplacer pour définir de nouvelles conditions de recevabilité des expertises scientifiques :

- l'expert scientifique, technique ou le sachant doit rendre compréhensible son rapport ou témoignage aux yeux de la cour ;
- le témoignage doit être étayé d'un nombre suffisant de faits ou de données ;
- le témoignage est le fruit de méthodes et principes fiables ;

9. https://law.justia.com/cases/district-of-columbia/court-of-appeals/1923/no-3968.html
10. https://law.justia.com/cases/federal/appellate-courts/F2/753/1224/

- l'expert a appliqué avec rigueur les méthodes et principes.

2.2.1.3 La norme Daubert

La définition de la *Federal Rule of Evidence 702* est reprise par la Cour Suprême états-unienne dans l'affaire *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[11] en 1993.

Cas réel · *Daubert v. Merrell Dow Pharmaceuticals, Inc.*

Jason Daubert et Eric Schuller sont deux enfants nés avec des malformations congénitales. Leurs parents imputent ces malformations à la prise du médicament Bendectin par leurs mères pendant la grossesse. Ils intentent un procès contre Merrell Dow Pharmaceuticals Inc., une filiale de Dow Chemical Company, devant la Cour fédérale de Californie. Merrell Dow Pharmaceuticals Inc. demande le transfert du procès devant la Cour fédérale des États-Unis.

Au cours du procès, Merrell Dow Pharmaceuticals Inc. demande un jugement sommaire en sa faveur arguant par une expertise que le médicament Bendectin n'est pas à l'origine des malformations congénitales. Les parents de Jason Daubert et Eric Schuller présentent un expert qui indique que le médicament Bendectin est à l'origine des malformations congénitales. Cependant, l'expertise est basée sur des études sur des animaux *in vivo* et *in vitro*, des études pharmacologiques et une relecture d'études publiées. Ces méthodes n'ont pas encore été acceptées par la communauté scientifique.

La Cour fédérale de Californie approuve la demande de jugement sommaire de Merrell Dow Pharmaceuticals Inc. et Jason Daubert et Eric Schuller font appel devant la Cour d'appel des États-Unis pour le neuvième circuit. La Cour d'appel des États-Unis pour le neuvième circuit confirme la décision de la Cour fédérale de Californie en indiquant que l'expertise présentée par les parents de Jason Daubert et Eric Schuller n'est pas recevable, car les méthodes employées n'ont pas été acceptées par la communauté scientifique. De plus, la Cour d'appel des États-Unis pour le neuvième circuit doute de la fiabilité de l'expertise parce qu'elle semble avoir été réalisée en préparation du procès.

11. https://www.law.cornell.edu/supct/html/92-102.ZS.html

> Sans cette expertise, la Cour d'appel des États-Unis pour le neuvième circuit doute que les parents de Jason Daubert et Eric Schuller puissent prouver lors d'un procès que le médicament Bendectin est à l'origine des malformations congénitales.

Suite à cette décision, la Cour Suprême états-unienne définit la norme Daubert pour déterminer les conditions d'admissibilité des expertises scientifiques. La norme Daubert est définie par quatre critères :

1. le juge est le gardien de la fiabilité de l'expertise ;
2. l'expertise délivrée doit être fiable et pertinente dans la résolution du litige ;
3. l'expert doit appliquer des méthodes scientifiques ;
4. les méthodes utilisées sont dites « scientifiques » si :
   - elles sont acceptées par la communauté scientifique ;
   - elles font l'objet d'une publication dans une revue scientifique à comité de lecture ;
   - elles ont été soumises à des tests et des évaluations ;
   - leur marge d'erreur est connue ;
   - elles ont été menées indépendamment du procès ou dépendante de l'intention de fournir le témoignage.

Selon Dixon et Gill, la norme Daubert a eu pour effet de refuser l'admissibilité de 70 % des expertises scientifiques présentées devant les tribunaux états-uniens entre juin 1996 et juin 1997, soit deux ans après son adoption (contre 53 % avant l'adoption de la norme) (Dixon et Gill, 2001).

2.2.2 Analyses ADN et approche bayésienne

Au milieu des années 1990, le domaine de la criminalistique est bouleversé avec l'émergence des comparaisons ADN. Ces dernières ont fait leur première apparition dans les années 1980 et ont été très largement acceptées dans les années 1990 au sein du monde criminalistique (Committee on DNA Technology in Forensic Science and National Research Council

and Life Sciences Commission and Division on Earth and Life Studies and Commission on Life Sciences and Division on Earth and Life Studies Staff and National Research Council Staff and others, 1992; National Research Council and others, 1996).

Les analyses ADN ont permis d'observer une évolution des pratiques en criminalistique. En effet, alors que ces dernières reposaient essentiellement sur des méthodes déterministes ou catégorielles, les analyses ADN utilisent majoritairement des méthodes probabilistes (Buckleton *et al.*, 2018). Lors d'une analyse, la trace est analysée et comparée de manière à pouvoir indiquer la probabilité de validité d'une de ces deux hypothèses :

1. H_0 : la trace et la pièce de comparaison proviennent de la même source ;
2. H_1 : la trace et la pièce de comparaison proviennent de sources différentes.

Le résultat du calcul de probabilité entre les deux hypothèses correspond au *LR*.

2.2.2.1 Interprétation du LR

Dans un rapport technique, qu'il s'agisse d'un rapport d'expertise ou d'un rapport de réquisition, le *LR* apparaît sous forme numérique. Cependant, pour apporter davantage de compréhension à ce résultat pour les personnes non expertes (dont les enquêtrices, enquêteurs, magistrates et magistrats), le *LR* est souvent exprimé sur une échelle verbale dont un exemple est montré dans le tableau 2.2 :

LR	**Échelle verbale**
1	Non concluant
1 à 10	Faiblement en faveur de l'hypothèse H_0
10 à 100	Modérément en faveur de l'hypothèse H_0
100 à 1000	Fortement en faveur de l'hypothèse H_0
1000 à 1000000	Très fortement en faveur de l'hypothèse H_0
> 1000000	Extrêmement fortement en faveur de l'hypothèse H_0

Tableau 2.2 – *Exemple d'échelle verbale du LR*

Le calcul des probabilités de chaque hypothèse varie en fonction de la probabilité *a priori* et de la probabilité *a posteriori*.

Probabilité *a priori* La probabilité *a priori* permet de favoriser une hypothèse par rapport à l'autre. Lorsqu'il existe une équiprobabilité entre les deux hypothèses, la probabilité *a priori* est de 0.5.

Probabilité *a posteriori* La probabilité *a posteriori* est la probabilité de l'hypothèse après analyse des données, en ayant eu connaissance de nouvelles informations.

> **Note**
>
> Les conclusions du rapport technique sont purement factuelles et répondent à la mission de la réquisition à personne qualifiée ou de l'ordonnance de commission d'expertise. Elles ne permettent pas d'indiquer la culpabilité ou non d'un individu ; elles indiquent seulement la probabilité que l'individu soit la source de la trace. L'interprétation des résultats et le poids qui lui est donné reviennent au requérant.

2.2.3 Le rapport de vraisemblance appliqué à la comparaison de voix

L'arrivée des analyses ADN a permis de faire évoluer les pratiques en criminalistique. En effet, l'approche probabiliste a permis de mieux comprendre les résultats des analyses et de les interpréter de manière plus objective et répond aux critères de la norme Daubert. Ainsi, les analyses ADN sont aujourd'hui largement acceptées par la communauté scientifique et judiciaire. Ceci dit, la différence d'interprétation est toujours possible et les analyses peuvent présenter des erreurs.

En 1998, à Avignon, lors du *Workshop on Speaker Recognition and its Commercial and Forensic Applications*, puis dans leur publication en 2000, Champod et Meuwly présente une approche bayésienne pour la comparaison de voix (Bernardo et Smith, 2009 ; Champod et Meuwly, 2000). En 2000, le *LR* est pour la première fois intégré dans un système de RAL (Meuwly, 2000 ; Meuwly et Drygajlo, 2001). Ces travaux sont secondés par ceux de Rose en 2002 et Rose *et al.* en 2003 en faveur de l'utilisation du *LR* dans les expertises de comparaison de voix (Rose, 2002 ; Rose *et al.*, 2003).

Pour qu'une méthode soit admissible dans un tribunal, Gonzalez-Rodriguez *et al.* proposent une approche transparente, testable et compréhensible par des personnes non expertes. Pour cela, ils s'appuient sur une approche bayésienne et recommandent l'utilisation du *LR* pour la comparaison de voix (Gonzalez-Rodriguez *et al.*, 2007). Cette recommandation est suivie par plusieurs scientifiques tels que (Aitken et Taroni, 2004 ; Evett *et al.*, 2000 ; Morriso, 2009). Elle apparaît également dans le *National Research Council Report* en 2009 (National Research Council and others, 2009).

En criminalistique, la comparaison de voix consiste à comparer un enregistrement de voix de source inconnue (trace) et un ou plusieurs enregistrements de voix de source connues (pièces de comparaison). Le *LR* est un outil statistique qui permet d'indiquer le degré de support d'une de ces deux hypothèses par rapport à l'autre :

1. l'hypothèse H_p : la trace et la pièce de comparaison proviennent de la même source ;
2. l'hypothèse H_d : la trace et la pièce de comparaison proviennent de différentes sources.

Soit deux enregistrements e_1 et e_2 ainsi que les hypothèses H_p et H_d, le *LR* est calculé ainsi qu'indiqué dans l'équation 2.1. Plus la valeur du *LR* s'éloigne de 1, plus la probabilité de l'hypothèse associée est renforcée.

$$LR = \frac{P(e_1, e_2 | H_p)}{P(e_1, e_2 | H_d)} \begin{cases} < 1 & H_d \text{ est plus probable que } H_p. \\ = 1 & H_d \text{ et } H_p \text{ sont équiprobables.} \\ > 1 & H_p \text{ est plus probable que } H_d. \end{cases} \tag{2.1}$$

2.2.3.1 Calibration des scores

La question de la calibration apparaît en vue de réduire les marges d'erreur des systèmes. Elle permet de « redonner un sens » aux scores en les repositionnant par rapport à 1 (ou 0 si échelle logarithmique). Plusieurs méthodes ont été proposées (Brummer et Van Leeuwen, 2006 ; Gonzalez-Rodriguez *et al.*, 2006 ; Gonzalez-Rodriguez et Ramos, 2007 ; Gonzalez-Rodriguez *et al.*, 2007 ; Kinoshita, Ishihara *et al.*, 2014 ; Nautsch *et al.*, 2016). La méthode de calibration linéaire, proposée par Brummer et Van Leeuwen a montré ses forces lors des campagnes NIST-SRE (Brummer et Van Leeuwen, 2006). La calibration linéaire vise

à appliquer une fonction affine sur les scores bruts LR_{brut} où le coefficient et l'ordonnée à l'origine sont calculés par régression logistique. Cette régression logistique repose sur l'entraînement de données qui doivent refléter suffisamment les conditions d'enregistrement des pièces de question et de comparaison. L'équation 2.2 décrit le calcul de calibration des scores.

$$LR_{cal} = LR_{brut} \times \text{coefficient} + \text{ordonnée} \tag{2.2}$$

En 2021, un groupe de scientifiques ont écrit un « consensus sur la validation de la comparaison de voix en criminalistique » (Morrison *et al.*, 2021). Les auteurs affirment que les systèmes de comparaison de voix doivent être en mesure d'effectuer une calibration des scores en utilisant un modèle statistique en sortie de système.

2.2.3.2 Coût du rapport de vraisemblance logarithmique (C_{llr})

Le C_{llr} est une fonction de coût qui permet d'évaluer le *LR*. Il se calcule en bits. À la différence de métriques telles que l'*EER*, celle-ci ne repose pas sur une décision binaire. Pour ce faire, une fonction logarithmique des *LR* permet d'obtenir *LLR*. Le C_{llr} est une fonction de perte qui estime la différence entre la valeur d'un *LLR* et le « sens » de cette valeur (Brummer et Preez, 2006). Pour chaque comparaison cible, plus le *LLR* est supérieur à 0 et plus le coût est faible. *A contrario*, plus le *LLR* est inférieur à 0 et plus le coût est élevé. Les comparaisons imposteurs suivent le schéma inverse : plus le *LLR* est inférieur à 0 et plus le coût est faible et plus le *LLR* est supérieur à 0 et plus le coût est élevé. Par conséquent, plus la valeur du C_{llr} est faible, plus le système est performant. Le calcul du C_{llr} est présenté dans l'équation 2.3.

$$C_{llr} = \frac{1}{2N_{tar}} \sum_{LR \in X_{tar}} log_2(1 + \frac{1}{LR}) + \frac{1}{2N_{non}} \sum_{LR \in X_{non}} log_2(1 + LR) \tag{2.3}$$

Morrison *et al.* indiquent qu'un système bien calibré, mais qui aurait une performance équivalente à celle du hasard aurait un C_{llr} de 1.18 bit (Morrison *et al.*, 2021).

Le C_{llr} peut être décomposé en deux parties :

- le C_{llr}^{cal} qui correspond la perte lors de la calibration ;
- le C_{llr}^{min} qui correspond à la perte minimum pour atteindre un système parfaitement calibré.

Le C_{llr} d'un système parfaitement calibré correspond au C_{llr}^{min}. Ce dernier peut également être utilisé comme métrique pour évaluer la performance d'un système de comparaison de voix. La figure 2.6 présente la valeur de C_{llr} pour trois systèmes/conditions différents de RAL. Les graphiques de gauche et du milieu présentent deux systèmes différents. Les graphiques du milieu et de droite présentent le même système sous deux conditions différentes. Chacun des graphiques fait apparaître le C_{llr}^{cal} et le C_{llr}^{min} (*discrimination loss*).

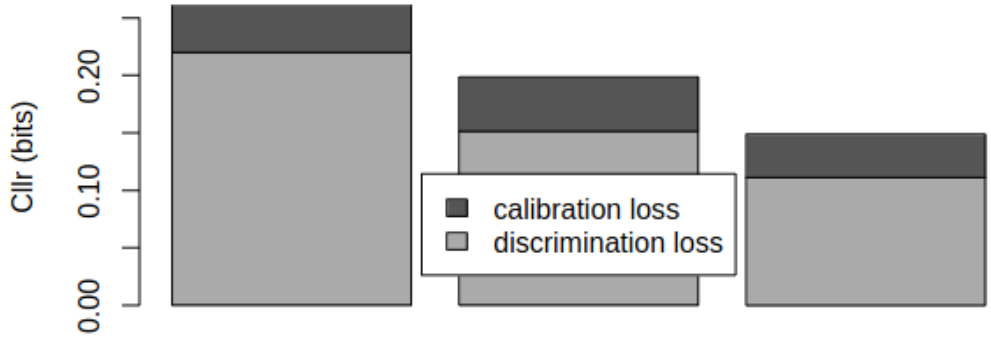


Figure 2.6 – C_{llr} *pour trois systèmes/conditions différentes de RAL. Les graphiques de gauche et du milieu présentent deux systèmes différents. Les graphiques du milieu et de droite présentent le même système sous deux conditions différentes. D'après (van Leeuwen et Brümmer, 2007).*

2.2.3.3 Validation des systèmes de comparaison de voix

Pour s'assurer de la fiabilité du résultat d'une comparaison de voix, le système est soumis à validation. Celle-ci repose sur des données extérieures aux données utilisées pour l'entraînement du modèle, la calibration des scores ou encore les enregistrements de la comparaison de voix. Selon Morrison *et al.*, cette validation doit suivre un protocole strict (Morrison *et al.*, 2021) :

- les données de validation des systèmes doivent comprendre les paires d'enregistrements ainsi que le score *LR* délivré ;

- certaines comparaisons doivent être cibles et d'autres doivent être imposteurs;
- le système à valider ne doit pas savoir à l'avance si les comparaisons sont cibles ou imposteurs;
- le résultat de la validation est un ensemble de *LR* pour les comparaisons cibles et un ensemble de *LR* pour les comparaisons imposteurs;
- La performance est considérée comme bonne si les valeurs des scores des comparaisons cibles sont élevées et les valeurs des scores des comparaisons imposteurs sont faibles.

Le jeu de données de validation est construit de sorte que les conditions d'enregistrement soient représentatives des conditions d'enregistrement des enregistrements de la comparaison de voix à résoudre. Les données de validation doivent également être représentatives des locuteurs impliqués dans la comparaison de voix. Elles doivent aussi contenir un nombre suffisant de locuteurs afin que les résultats de la validation soient représentatifs des conditions de la comparaison de voix.

2.3 Vers un standard pour la comparaison de voix en criminalistique

La comparaison de voix en criminalistique est un domaine d'expertise dont les pratiques ne sont pas encore standardisées. Les méthodes et techniques peuvent varier d'un expert à l'autre et les résultats obtenus peuvent être différents. Alors que les systèmes de comparaison de voix possèdent des indicateurs afin d'évaluer leur performance sur une comparaison donnée, ces derniers ne sont pas toujours utilisés par les experts. De plus, ces indicateurs ne permettent pas de déterminer si le résultat d'une comparaison est fiable ou non. L'ensemble de ces éléments montrent l'importance de mettre en place des standards pour encadrer les pratiques des experts sur la comparaison de voix en criminalistique.

2.3.1 Accréditation des laboratoires spécialisés dans la comparaison de voix

Le processus d'accréditation répond à une volonté européenne d'harmoniser les processus, méthodes et compétences entre les laboratoires des différents pays. Dans le cas des laboratoires dédiés à la criminalistique, l'accréditation concerne la norme ISO/CEI 17025 [12]. Certains laboratoires de police scientifique sont déjà accrédités selon la norme ISO/CEI 17025 : c'est notamment le cas des plateaux techniques en identité judiciaire (recherche de traces papillaires, signalisation, *etc.*). La norme ISO/CEI 17025 comporte notamment des exigences sur le respect des protocoles qualité et l'habilitation des personnels aux méthodes et techniques utilisées.

L'objectif de l'accréditation est de valider et de promouvoir les matériels et modes opératoires utilisés par les laboratoires, apportant ainsi davantage de crédibilité quant aux travaux et rapports techniques des personnels habilités.

À l'heure actuelle, en France, aucun laboratoire de criminalistique ne possède d'accréditation conformément à la norme ISO/CEI 17025 pour la comparaison de voix. L'un des objectifs du projet Voxcrim est de définir un ensemble de conditions dans lesquelles le résultat d'une comparaison de voix est considéré comme étant fiable en vue de mettre en place un protocole pour répondre aux exigences de cette accréditation que seul le Comité français d'accréditation (Cofrac) peut délivrer en France.

Le travail de ce manuscrit s'inscrit dans la volonté d'accréditation des laboratoires de comparaison de voix. Une première étape vers l'accréditation des laboratoires de comparaison de voix est d'établir un cadre scientifique dans lequel la fiabilité d'une comparaison de voix est connue. Cela implique de définir la notion de « locuteur », d'en comprendre la reconnaissance aussi bien par les êtres humains que par les systèmes automatiques et enfin de mettre en place un cadre scientifique dans lequel la fiabilité d'une comparaison de voix est connue.

12. https://www.iso.org/fr/standard/66912.html

2.3.2 Délimiter l'ensemble de règles pour une comparaison de voix fiable et pertinente

Cas réel · Affaire Élodie Kulik

Dans la nuit du 10 au 12 janvier 2002, Élodie Kulik rentre chez elle après avoir passé la soirée avec des amis. En chemin vers son domicile, elle est projetée sur le bas-côté de la route par un véhicule et appelle les secours lorsqu'elle est tirée hors de son véhicule, entraînée dans une autre voiture qui la conduit sur un chemin isolé sur la commune de Tertry dans la Somme. Elle y sera violée et assassinée. Son corps est retrouvé le 11 janvier 2002, partiellement brûlé.

Alors qu'aucun témoin n'a assisté à la scène, une trace ADN complète, quatre autres incomplètes et une trace papillaire sont retrouvées sur les lieux du crime... et l'enregistrement de l'appel passé par Élodie Kulik aux secours. Les traces ADN sont comparées au FNAEG, mais ne donnent aucun résultat. En janvier 2012, soit dix ans après les faits, une nouvelle comparaison ADN est effectuée par parentèle et permet d'identifier, au final, Grégory Wiart, décédé d'un accident de voiture en 2003.

Un an plus tard, sept personnes proches de Grégory Wiart sont placées en garde à vue. Deux d'entre elles seront relâchées, les quatre autres seront mises en examen. Parmi les personnes restantes, Willy Bardon est mis en examen pour enlèvement, séquestration, viol en réunion et meurtre avant d'être incarcéré.

L'enregistrement de l'appel aux secours dure 26 secondes et ne contient que 4 secondes de signal potentiellement exploitable pour une comparaison de voix après nettoyage audio. Pas moins de 14 expertises ont été effectuées sur les enregistrements, aussi bien pour transcrire le contenu linguistique que pour identifier les locuteurs. Concernant la comparaison de voix, les expertises sont unanimes : la qualité de l'enregistrement est insuffisante pour une comparaison de voix. De plus, même au niveau des transcriptions, les experts ne sont pas d'accord sur le contenu linguistique de l'enregistrement.

En 2013, un rapport d'expertise d'un laboratoire privé part du postulat que l'être humain possède une faculté à reconnaître les voix « familières ». Six auditeurs, proches de Willy Bardon, ont été soumis à des tests pour évaluer si la voix de ce dernier figure bel et bien dans l'enregistrement. Pour cela, ils ont écouté l'enregistrement et indiqué s'ils reconnaissaient ou non les voix présentes. En cas de réponse positive, ils devaient indiquer qui en était l'auteur avec un score de fiabilité allant de 1 à 10. Dans un second temps, un fichier audio contenant 22 échantillons de voix leur était présenté. Quatre de ces échantillons ont été prononcés par Willy Bardon. Pendant l'écoute sans coupure du fichier audio, ils devaient faire signe lorsqu'ils entendaient la voix de Willy Bardon. Lorsque la première écoute n'avait pas permis de localiser tous les échantillons de voix de Willy Bardon, une seconde écoute sans coupure était proposée. En cas d'erreur sur cette seconde écoute, une troisième était mise en place. Cette fois, les échantillons de voix étaient présentés un à un. Enfin, ils ont écouté un fichier audio contenant un échantillon de voix de Willy Bardon, un cri de la victime puis une des voix masculines de l'enregistrement (supposément appartenant à Willy Bardon). Willy Bardon lui-même a été soumis aux mêmes tâches. Il indique qu'il ne reconnaît pas sa voix, mais que la voix entendue ressemble à la sienne. L'expertise conclut que les six auditeurs proches de Willy Bardon ont reconnu sa voix dans l'enregistrement avec une fiabilité de 87 %.

Fin novembre 2019, lors du procès de Willy Bardon, des membres de l'AFCP sont intervenus en qualité de témoin. Ils ont indiqué que les méthodes employées par l'expert ne relèvent pas du domaine scientifique, et qu'elle véhicule des contre-vérités. En effet, la faculté à reconnaître les voix « familières » n'est pas prouvée scientifiquement. De plus, reconnaître la voix d'une personne plusieurs années après l'avoir entendue est une tâche difficile, voire impossible. Enfin, la méthode employée induit de nombreux biais cognitifs, notamment les biais de confirmation et d'autorité, qui a conduit les auditeurs à reconnaître la voix de Willy Bardon dans l'enregistrement.

Malgré l'avis des experts de la police scientifique, de l'IRCGN et de l'AFCP, après écoute des enregistrements, la présence de la voix de Willy Bardon a été laissée à la libre appréciation des jurés.

À l'issue de son procès, Willy Bardon est condamné à 30 ans de réclusion criminelle.

2.3.2.1 La *box-rule* : un ensemble de règles et de facteurs dans lequel le résultat d'une comparaison de voix est considéré comme fiable

Un premier pas pour la standardisation des pratiques en comparaison de voix est de définir un ensemble de règles et de facteurs dans lequel le résultat d'une comparaison de voix est considéré comme fiable. La construction de la *box-rule* passe par l'expérimentation et la validation de chaque facteur, d'abord individuellement puis en combinaison avec les autres facteurs. Le but est de déterminer les limites de chaque facteur afin de ne pas fournir de rapport d'expertise si la comparaison de voix outrepasse les limites des différents facteurs. Par exemple, si la comparaison de voix est effectuée entre deux locuteurs de genre différent, le résultat de la comparaison ne sera pas fiable.

Quels facteurs pour définir la *box-rule* ? Pour définir l'ensemble des facteurs de la *box-rule*, il est nécessaire de déterminer les facteurs qui influent sur la comparaison de voix. Ces facteurs portent sur les enregistrements (durée, fréquence d'échantillonnage, *etc.*), les locuteurs (genre, âge, langue, *etc.*) et les conditions d'enregistrement (environnement, matériel, *etc.*). La multiplicité des facteurs et leur combinaison rendent la définition de la *box-rule* complexe.

Quelles sont les limites de chaque facteur ? Une fois les facteurs définis, il est nécessaire de déterminer les limites de chaque facteur, d'abord individuellement puis en combinaison avec les autres facteurs. Par exemple, la combinaison de deux facteurs peut avoir une influence sur la comparaison de voix alors que ces deux facteurs n'ont pas d'influence individuellement.

Les facteurs ont-ils le même poids ? Une fois les limites de chaque facteur déterminées, il est nécessaire de déterminer le poids de chaque facteur. Même si un facteur peut être déterminant pour une comparaison de voix, un autre peut être négligeable.

L'objectif de ce manuscrit est de travailler sur ces trois questions afin de poser les premières règles et facteurs de la *box-rule*. Nous proposons de travailler sur les facteurs suivants :

- Locuteurs : genre, âge, nationalité.

- Enregistrements : durée, écart entre les enregistrements.

Ces facteurs sont évalués individuellement puis deux par deux.

Synthèse du chapitre

La comparaison de voix est une tâche qui consiste à comparer un enregistrement de voix de source inconnue (trace) et un ou plusieurs enregistrements de voix de source connues (pièces de comparaison). Au cours de l'histoire, elle a été utilisée dans la résolution d'affaires judiciaires, dans un premier temps par des experts (ou non) humains puis par des systèmes automatiques. Cependant, elle est contrainte par les pièces dont la variabilité peut en influer sur le résultat.

Les facteurs de variabilité sont nombreux et peuvent être liés aux locuteurs, au contenu linguistique, au fichier numérique ou aux conditions d'enregistrement. Cela pose la question de la fiabilité des résultats de comparaison de voix et certaines affaires réelles ont montré l'importance de cette question. Bien que certaines normes et standards soient déjà mis en place, il est nécessaire de poursuivre les travaux de recherche afin de répondre aux problématiques de la comparaison de voix en contexte criminalistique.

Deuxième partie

Contributions

3 | Quelles données pour la comparaison de voix en criminalistique ?

Résumé : *Dotée de plus de 7000 locuteurs, la base de données Voxceleb est très largement utilisée au niveau international pour l'entraînement des systèmes de RAL. Cependant, dans le domaine criminalistique, Voxceleb reste très éloignée des conditions d'enregistrement rencontrées. En effet, les conditions d'enregistrement sont très différentes des enregistrements utilisés lors d'affaires judiciaires et la variabilité intralocuteur est très peu représentée. Ces deux problèmes peuvent entraîner des biais dans l'évaluation des performances des systèmes de RAL et, par conséquent, dans l'interprétation des résultats. Pour pallier ces problèmes, nous avons conçu deux bases de données : FABIOLE 2 et PTSVOX. Ces bases de données mettent l'accent sur la variabilité intralocuteur et les conditions d'enregistrement rencontrées en criminalistique.*

Sommaire

3.1 Utilisation de données pour la comparaison de voix

En informatique, et plus particulièrement en apprentissage automatique, les données sont une part importante des processus. En RAL et en comparaison de voix, des jeux de données distincts sont nécessaires pour l'apprentissage des modèles, la calibration des scores et la validation des systèmes de comparaison. L'évaluation des systèmes de RAL repose aussi sur des jeux de données dédiés. Ces différents jeux de données doivent par conséquent posséder des caractéristiques distinctes.

3.1.1 Représenter la variabilité inter- et intralocuteur

Plusieurs bases de données sont largement utilisées dans la RAL. C'est le cas de TIMIT (Garofolo *et al.*, 1993), Switchboard (Godfrey *et al.*, 1992), Aurora (Pearce, 1998), CHAINS (Cummins *et al.*, 2006), CSLU (Cole *et al.*, 1998), YOHO (Higgins, 1990). Aujourd'hui, l'état de l'art utilise presque exclusivement la base de données VoxCeleb pour l'apprentissage des modèles.

Les bases de données de parole doivent posséder un grand nombre de locuteurs que ce soit pour la modélisation de la variabilité interlocuteur lors de l'apprentissage des modèles que pour son test. Ces bases de données peuvent posséder jusqu'à plusieurs milliers de locuteurs, ce qui est idéal pour représenter la variabilité interlocuteur. En revanche, la variabilité intralocuteur est souvent oubliée et cela se manifeste par un nombre d'enregistrements et de contextes d'enregistrement par locuteur très restreint.

3.1.2 Couvrir les différents contextes d'enregistrement

Ainsi que nous l'avons évoqué précédemment, les données utilisées pour la calibration des scores et la validation des systèmes doivent se rapprocher au maximum des fichiers audio utilisés pour la comparaison de voix. Or, en comparaison de voix, toutes les conditions ne sont pas contrôlées. En revanche, avec les informations déjà connues sur le locuteur (genre, langue, *etc.*) et sur le contexte d'enregistrement (matériel, *etc.*), il est possible de détermi-

ner les conditions d'enregistrement les plus probables. À ce jour, il n'existe pas de base de données de parole qui couvre l'ensemble des conditions d'enregistrement possibles.

3.2 Voxceleb : base de données de référence

En plus des corpus fournis par le NIST, les dernières recherches sur la reconnaissance du locuteur ont recours à la base de données Voxceleb (Nagrani *et al.*, 2019 ; Nagrani *et al.*, 2017). Composées de vidéos issues du site YouTube, les bases de données Voxceleb regroupent plus de 7000 locuteurs pour 1 million d'énoncés, totalisant ainsi plus de 2000 heures de parole. L'essentiel de ces bases est en langue anglaise et seuls le genre et la nationalité des locuteurs sont renseignés. Les enregistrements sont de diverses origines, et il est possible d'observer des environnements bruités. Les bases de données Voxceleb sont aujourd'hui très largement utilisées en reconnaissance automatique du locuteur, notamment dans l'apprentissage des modèles.

3.2.1 Description

Voxceleb correspond à deux bases de données réalisées par l'université d'Oxford, *Voxceleb1* et *Voxceleb2*, issues d'extraits de vidéos YouTube (Nagrani *et al.*, 2019). Les auteurs indiquent que Voxceleb est conçu de manière à couvrir un maximum de diversité que ce soit au niveau langagier (accent) ou au niveau social (ethnie, profession, âge). De plus, les conditions d'enregistrement sont variées dans des environnements qui peuvent être bruités. La base de données Voxceleb peut être aussi bien utilisée pour la détection de visage que pour la RAL. Le tableau 3.1 résume le contenu des bases de données *Voxceleb1* et *Voxceleb2*.

Note · Sélection des locuteurs

La sélection des locuteurs a été réalisée à partir de la base de données VGG Face Dataset (Parkhi *et al.*, 2015), qui elle-même est issue des personnalités les plus recherchées sur *Freebase Knowledge Graph* et *The Internet Movie Database*. Les vidéos sont ensuite fil-

trées pour ne garder que les trames qui contiennent le locuteur en train de parler. Enfin, une étape de reconnaissance des visages est réalisée pour ne garder que les passages où le visage du locuteur est correctement reconnu.

Jeu de données	***Voxceleb1***	***Voxceleb2***
Locuteurs	690	3761
Locutrices	561	2351
Vidéos	22 496	150 480
Énoncés	153 516	1 128 246
Durée moyenne des énoncés	8.2s	7.8s

Tableau 3.1 – *Description de* Voxceleb1 *et* Voxceleb2. *D'après (Nagrani et al., 2019)*

Nagrani *et al.* donnent un complément d'information sur Voxceleb2 (Nagrani *et al.*, 2019). La distribution de la durée des énoncés montre que plus de 500 000 d'entre eux durent moins de 6 secondes et que 29 % des locuteurs sont de nationalité états-unienne. Ces informations sont résumées dans la figure 3.1.

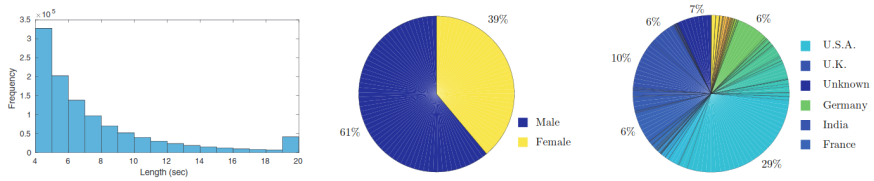


Figure 3.1 – *Complément d'information sur* Voxceleb2. *À gauche se trouve la distribution des durées des énoncés. Au milieu, la proportion de locuteurs. À droite, la répartition de la nationalité des locuteurs (Nagrani et al., 2019)*

3.2.1.1 Manque de métadonnées sur les locuteurs

Les bases de données Voxceleb sont fournies avec deux fichiers tabulaires (un par base de données) qui contiennent les noms, le genre et la nationalité des locuteurs. Ces informations sont issues des pages Wikipédia des locuteurs. Elles ne reflètent cependant pas la langue maternelle ou la langue d'expression des signaux, ni le niveau de bruit ou encore

le contexte dans lesquels les enregistrements ont été réalisés. Aussi, les fichiers tabulaires ne contiennent pas de métadonnées temporelles. Par conséquent, il n'est pas possible de savoir si les enregistrements ont été réalisés à des périodes différentes de la vie des locuteurs. Pour résumer, les informations disponibles sur les locuteurs et les enregistrements ne permettent pas de savoir si les corpus d'entraînement et de test sont équilibrés en termes de genre, de nationalité, de langue maternelle, de langue d'expression, de niveau de bruit ou encore de contexte d'enregistrement, ni même de pouvoir corréler ces informations avec les performances des modèles entraînés sur ces corpus.

3.2.2 Protocoles et jeux de données

Nagrani *et al.* définissent également des protocoles d'évaluation pour chacune des bases. Les données sont séparées comme indiqué dans le tableau 3.2. Au sein d'une même base de données, il n'y a pas de recoupement de locuteur entre le corpus d'entraînement et celui de test. En revanche, il existe un recoupement en locuteurs entre le corpus d'entraînement de *Voxceleb1* et le corpus de test de *Voxceleb2* ainsi qu'entre les deux corpus de test.

Voxceleb1	**Entraînement**	**Test**
Locuteurs	1211	40
Vidéos	21 819	677
Énoncés	128 642	4874
Voxceleb2	**Entraînement**	**Test**
Locuteurs	5994	118
Vidéos	145 569	4911
Énoncés	1 092 009	36 237

Tableau 3.2 – *Répartition des données de* Voxceleb1 *et Voxceleb2 d'après (Nagrani* et al.*, 2019)*

Voxceleb admet également trois jeux de données d'évaluation : *Voxceleb1 test*, *Voxceleb1-H* et *Voxceleb1-E*.

Voxceleb1 test Le jeu de données *Voxceleb1 test* est composé de 40 locuteurs dont le nom commence par un E pour assurer une répartition égale entre les locuteurs masculins et fé-

minins. Ce jeu de données étant limité en nombre de locuteurs, une bonne performance sur ce jeu de données ne garantit pas une bonne performance sur un autre jeu de données.

Voxceleb1-E Le jeu de données *Voxceleb1-E* est composé de paires d'enregistrements formées aléatoirement à partir du jeu de données de test entier. Il comprend 581 480 paires impliquant les 1251 locuteurs de *Voxceleb1*.

Voxceleb1-H Le jeu de données *Voxceleb1-H* est composé de paires d'enregistrements des locuteurs ayant le même genre et la même nationalité. Ce jeu de données comprend 1190 locuteurs et 552 536 de paires d'enregistrements.

3.2.3 Utilisation dans le domaine de la reconnaissance automatique du locuteur

Depuis plusieurs années, Voxceleb est devenue incontournable dans le domaine de la RAL (Chen *et al.*, 2022 ; Desplanques *et al.*, 2020 ; Lin et Mak, 2020 ; Snyder *et al.*, 2019 ; Snyder *et al.*, 2018 ; Tak *et al.*, 2021 ; Villalba *et al.*, 2020b ; Yang *et al.*, 2021 ; Zhai *et al.*, 2021). La base de données est utilisée soit pour l'apprentissage des modèles, soit pour l'évaluation de la performance de ces derniers, voire les deux. Le tableau tableau 3.3 résume les travaux qui utilisent Voxceleb pour l'apprentissage des modèles et ceux qui l'utilisent pour l'évaluation de la performance de ces derniers.

Publication	**Entraînement**	**Test**	*EER*
(Snyder *et al.*, 2018)	SWBD + *Voxceleb1*	SITW Core + SRE16	4.16
(Snyder *et al.*, 2019)	Voxceleb	SITW	1.7
(Tak *et al.*, 2021)	Voxceleb	Voxceleb	1.12
(Zhai *et al.*, 2021)	TIMIT + Voxceleb	TIMIT + Voxceleb	15.7
(Desplanques *et al.*, 2020)	Voxceleb	*Voxceleb1*	0.87
(Lin et Mak, 2020)	Voxceleb	*Voxceleb1*	1.81
(Chen *et al.*, 2022)	Gigaspeech + Vox Populi	*Voxceleb1*	0.38

Tableau 3.3 – *Études utilisant Voxceleb pour l'apprentissage des modèles et/ou pour l'évaluation de la performance de ces derniers*

3.2.4 Sources de biais dans les bases de données

Alors que Nagrani *et al.* mettent en avant la diversité des locuteurs dans les bases de données Voxceleb, Hutiri et Ding ont mis au point un plan d'évaluation pour estimer les biais contenus dans ces bases de données, notamment en termes de représentativité (Hutiri et Ding, 2021 ; Hutiri et Ding, 2022 ; Nagrani *et al.*, 2019). Pour cela, Hutiri et Ding cherchent à répondre à deux questions :

1. Équité [1] : la performance d'un modèle défini varie-t-elle en fonction des catégories de locuteurs ?
2. Comparaison : comment comparer l'équité parmi les différents modèles de vérification du locuteur ?

Hutiri et Ding testent deux modèles du même type (ResNetSE34V2 (Heo *et al.*, 2020) et ResNetSE34L (Chung *et al.*, 2020)) sur les données Voxceleb pour évaluer et comparer leur équité (Hutiri et Ding, 2021). Le modèle ResNetSE34V2 a été conçu pour améliorer la performance et le modèle ResNetSE34L pour réduire le temps d'inférence. Les auteurs indiquent des disparités dans les résultats du test du modèle ResNetSE34V2 sur le jeu de données *Voxceleb1-H*. La performance est calculée en utilisant le C_{Det} *ratio* ainsi que défini dans la campagne d'évaluation NIST-SRE 2019 (National Institute of Standards and Technology, 2019). Ce ratio est défini comme le rapport entre le nombre de locuteurs dont la performance est supérieure à un seuil C_{Det} et le nombre total de locuteurs. Le seuil C_{Det} est défini comme la valeur de C_{Det} pour laquelle le taux de faux rejets est égal au taux de fausses acceptations.

Dans un premier temps, ils mettent l'accent sur la différence de performance entre les locuteurs et les locutrices. En effet, la performance des locutrices est généralement inférieure à la performance globale, indiquant que le modèle semble favoriser les voix masculines. Seules les locutrices états-uniennes et irlandaises ont une performance supérieure à la performance globale. Cependant, le jeu de données de test ne comprend que 5 locutrices de nationalité irlandaise, ce qui ne permet pas de tirer de conclusions sur la performance des locutrices irlandaises. Les locutrices états-uniennes sont au nombre de 368. Les locutrices d'origine indienne obtiennent la moins bonne performance.

1. *Fairness* en anglais

Du côté des locuteurs, la majorité des nationalités ont une performance supérieure à la performance globale. Le modèle est moins adapté aux 13 locuteurs de nationalité norvégienne.

La comparaison de l'équité entre les modèles ResNetSE34V2 et ResNetSE34L ne montre pas de différence significative. En effet, les deux modèles font état d'une très bonne performance pour les locuteurs des États-Unis, mais cette dernière est moindre pour les locuteurs originaires de l'Inde et de Norvège. Les locuteurs mexicains ont obtenu une performance supérieure à la performance globale pour le modèle ResNetSE34V2, mais inférieure pour le modèle ResNetSE34L.

Hutiri et Ding concluent que les modèles de vérification du locuteur peuvent être biaisés en défaveur des locutrices et des locuteurs de certaines nationalités. L'équité entre les modèles varie de manière inconsistante, favorisant certains sous-groupes au détriment d'autres sous-groupes. Cependant, dans ces expériences, cette variation n'apparaît pas corrélée au genre, à la nationalité ou à la taille du sous-groupe.

Hutiri et Ding mettent en avant les différents biais qui peuvent intervenir dans la RAL notamment avec l'utilisation des bases de données Voxceleb pour l'entraînement des modèles (Hutiri et Ding, 2022). Ils reprennent les travaux de Suresh et Guttag qui ont identifié sept sources de biais (Suresh et Guttag, 2021).

Biais historique Les biais historiques reproduisent les biais existants dans la société. La construction de la base de données Voxceleb 1 repose sur la sélection de locuteurs provenant de la base VGG Face Dataset (Parkhi *et al.*, 2015), qui est elle-même issue des personnalités les plus recherchées sur *Freebase Knowledge Graph* et *The Internet Movie Database*. Après plusieurs étapes de sélection des vidéos pour ne garder que les trames qui contiennent le locuteur en train de parler, intervient une étape de reconnaissance des visages pour inclure uniquement les passages où le visage est correctement reconnu.

Biais de représentation Les biais de représentation peuvent arriver lorsque certains groupes sont sous-représentés. Cela entraîne une mauvaise performance du modèle pour ces groupes. Pour illustrer ce biais, les auteurs évoquent les résultats du SVEVA Challenge 2021 (Hutiri et

Ding, 2021) qui montrent que les modèles de vérification du locuteur ont une performance inférieure pour les locuteurs originaires d'Inde et de Norvège.

Biais de mesure Les biais de mesure sont liés à la manière dont les données sont étiquetées et annotées pour la tâche définie. Dans le cas de Voxceleb 1, la nationalité des locuteurs provient de leur page Wikipédia. Pour les concepteurs de Voxceleb, la nationalité est un indicateur de l'accent des locuteurs. Or, la nationalité n'est pas toujours un indicateur fiable de l'accent. En effet, deux locuteurs d'un même pays peuvent avoir des accents différents et deux locuteurs d'origines différentes peuvent avoir le même accent. Les auteurs soulignent également que le genre est vu comme une donnée binaire dont l'annotation ne semble pas suivre de méthode particulière.

Biais de masse[2] Les biais de masse peuvent arriver quand les données sont traitées en comme un ensemble alors qu'elles nécessitent d'être traitées individuellement ou en sous-groupes. Les auteurs reprennent l'exemple de la performance des modèles de vérification du locuteur en regroupant les locuteurs par genre et par nationalité (Hutiri et Ding, 2021). Ils montrent que la performance varie en fonction du genre et de la nationalité des locuteurs.

Biais d'apprentissage Les biais d'apprentissage sont liés à la manière dont le modèle est entraîné et aux conséquences de cet entraînement sur les performances du modèle. Les auteurs ont utilisé les deux modèles décrits dans (Hutiri et Ding, 2021) et montrent que l'architecture des modèles accentue les différences de performance et ont tendance à avoir un biais en défaveur des locutrices et des sous-groupes sous-représentés.

Biais d'évaluation Les biais d'évaluation interviennent quand la population de référence pour l'évaluation diffère de la population cible et que les métriques utilisées ne montrent qu'un aspect de la performance du modèle. Les auteurs ont testé le modèle ResNetSE34V2 sur les trois jeux de données d'évaluation issus de Voxceleb 1. Ils indiquent que le jeu de données *Voxceleb 1 test* est bien trop réduit pour évaluer la performance du modèle. En revanche, le jeu de donnée *Voxceleb 1-H* admet une performance bien moindre par rapport au jeu de données *Voxceleb 1-E*. Pour les auteurs, cela montre que la performance du modèle est liée à la manière dont les données sont sélectionnées pour l'évaluation. Ils s'intéressent

2. *Aggregation bias* en anglais.

également aux métriques utilisées, et plus particulièrement au *EER* ainsi qu'à la valeur minimale de la fonction de coût de détection sur l'ensemble des tests ($C_{Det}(\theta_{@\text{ overall min}})$). Les auteurs présentent l'*EER* comme étant une « vision ultra-simplifiée de la performance du modèle », car faux positifs et faux négatifs sont considérés comme équivalents alors que les systèmes de vérification du locuteur tendent soit vers un *FAR*, soit vers un *FRR* très faible. En ce qui concerne le $C_{Det}(\theta_{@\text{ overall min}})$, les auteurs indiquent qu'il ne montre que la performance du modèle pour un seuil de décision donné.

Biais de déploiement Le biais de déploiement se produit quand le contexte applicatif diffère du contexte d'entraînement. La vérification du locuteur est utilisée dans de nombreux contextes différents (assistants personnels, centre d'appels, *etc.*) et les auteurs indiquent que les modèles de vérification du locuteur sont entraînés sur des données provenant de contextes différents. Selon l'application, l'évaluation se concentrera sur le *FAR* ou le *FRR*. Par exemple, dans le cas d'un assistant personnel, un *FAR* bas sera privilégié pour éviter que l'assistant ne réponde à une personne qui n'a pas la permission d'utiliser l'assistant. À l'inverse, dans le cas d'un centre d'appels, un *FRR* bas sera privilégié pour que l'appelant soit redirigé vers un opérateur humain qui validera par la suite son identité par d'autres moyens.

3.3 FABIOLE 2 : dans la continuité du projet FABIOLE

3.3.1 De FABIOLE à FABIOLE 2

3.3.1.1 La variabilité intralocuteur, oubliée dans les bases de données de voix

L'évaluation des systèmes de comparaison de voix passe essentiellement par des bases de données dédiées. Celles du NIST-SRE peuvent être utilisées pour évaluer les systèmes de comparaison de voix, mais ces dernières font peu apparaître la variabilité intralocuteur. En effet, les bases de données du NIST-SRE sont composées de locuteurs ayant peu d'enregistrements. Par exemple, pour le plan d'évaluation de 2021, les locuteurs avaient 10 enregistrements dont une partie est bruitée (Jones *et al.*, 2022 ; Sadjadi *et al.*, 2021). Ce problème est également présent dans les bases de données vocales telles que VoxCeleb, TIMIT (Garofolo

et al., 1993) ou encore Switchboard (Godfrey *et al.*, 1992)). Les bases de données spécialement conçues pour la comparaison de voix connaissent le même problème (Morrison *et al.*, 2015 ; Ramos *et al.*, 2008 ; van der Vloed *et al.*, 2014).

La base de données FABIOLE 2 s'inscrit dans la continuité du projet FABIOLE [3] au cours duquel la base de données FABIOLE a été mise en place. L'objectif de la base de données est de déterminer la fiabilité des systèmes de comparaison de voix en étudiant la variabilité intralocuteur (Ajili *et al.*, 2016b). Pour cela, le nombre d'enregistrements par locuteur dans des conditions d'enregistrement différentes est primordial. Aussi, afin de limiter les risques de biais liés à d'autres facteurs que la variabilité intralocuteur, les conditions d'enregistrement sont contrôlées.

Pour répondre à ces contraintes, les enregistrements de FABIOLE sont exclusivement issus d'émissions de radio ou de télévision françaises. Les enregistrements contiennent un minimum de 30 secondes de parole nette de tout bruit.

3.3.1.2 Locuteurs et jeux de données

En raison de contraintes de temps et de budget, la base de données FABIOLE est composée de 130 locuteurs exclusivement masculins. Ces locuteurs sont intervenus au sein de 10 émissions de radio ou de télévision différentes. FABIOLE se présente en 2 jeux de données distincts :

1. set *T* : 30 locuteurs ayant chacun 100 enregistrements comprenant plus de 30 secondes de parole utile ;
2. set *I* : 100 locuteurs ayant chacun 1 enregistrement

Chaque enregistrement provient d'une émission différente.

Ces jeux de données ont permis de mettre en place deux protocoles expérimentaux.

Protocole du set *T* Le premier protocole est centralisé sur le set *T* et comprend plus de 150 000 comparaisons cibles et plus de 4,5 millions de comparaisons imposteurs. Chacun des 30 locuteurs possède 4950 comparaisons cibles et 290 000 comparaisons imposteurs.

3. Projet financé par l'ANR portant le numéro ANR-12-BS03-001

Protocole des sets *T* et *I* Le second protocole utilise les données des deux jeux de données. Ce protocole permet d'utiliser les données du jeu de données *I* pour former les comparaisons imposteurs (sans utiliser les données du jeu de données *T*). Le nombre de comparaisons cibles est le même que pour le premier protocole, mais le nombre de comparaisons imposteurs est de 300 000, soit 10 000 par locuteur du set *T*.

3.3.2 La base de données FABIOLE 2

3.3.2.1 Près de 400 locuteurs

Dans la continuité de FABIOLE, la base de données FABIOLE 2 en reprend les principes et contraintes. La base de données FABIOLE 2 a été conçue dans le cadre du projet ANR Voxcrim [4]. Elle suit les mêmes contraintes d'enregistrement que FABIOLE mais est plus richement dotée en locuteurs et en émissions.

Tout comme FABIOLE, les enregistrements de FABIOLE 2 sont fournis avec des fichiers au format Transcriber qui comprennent la séparation en tours de parole des locuteurs. Cette dernière a été faite manuellement pour l'intégralité des enregistrements.

La base de données FABIOLE 2 comprend 112 locutrices et 287 locuteurs. Les informations sur les locuteurs sont présentées dans le tableau B.1 en annexe. Les dates de naissance, nationalités et professions ont été récupérées depuis Internet, principalement depuis Wikipédia, parfois depuis des sites officiels ou LinkedIn. Lorsque seule l'année de naissance est connue, la date de naissance du locuteur est fixée au 1er janvier de l'année en question. Seul le locuteur LOC226 est réellement né un 1er janvier. Certains locuteurs possèdent une double nationalité. Celle-ci est alors renseignée en tant que telle. La durée correspond à la durée totale de parole nette en secondes. Le nombre de contextes correspond au nombre d'enregistrements différents dans lesquels le locuteur est intervenu.

Définition · Contexte

4. VoxCrim, ANR-17-CE39-0016 - https://voxcrim.univ-avignon.fr

Nous utilisons la notion de « contexte » pour désigner une unique situation d'enregistrement. Par exemple, si un locuteur est enregistré dans plusieurs fois dans la même émission, mais à des jours différents, il est considéré comme ayant participé à autant de contextes que de jours.

Les tableaux 3.4, 3.5 and 3.6 présentent respectivement la nationalité, la profession et le nombre de contextes des locuteurs.

	Française	**Double nationalité**	**Autres**
Locutrices	98	3	11
Locuteurs	268	2	17
Total	**366**	**5**	**28**

Tableau 3.4 – *Nationalité des locuteurs par genre*

	Journaliste	**Politicien**	**Autres**
Locutrices	78	23	11
Locuteurs	160	81	46
Total	**238**	**104**	**57**

Tableau 3.5 – *Profession des locuteurs par genre*

	Nombre de contextes		
	3-19	**20-49**	**50+**
Locutrices	48	29	35
Locuteurs	78	115	94
Total	**126**	**144**	**129**

Tableau 3.6 – *Nombre de contextes des locuteurs par genre*

3.3.2.2 Plus d'un millier d'émissions

Les enregistrements proviennent plus d'un millier d'émissions télévisées et radiophoniques. Ces dernières comprennent aussi bien des débats que des entretiens, des journaux d'actualité, *etc*. Les émissions proviennent très majoritairement de média français. Cependant, certaines émissions proviennent de pays issus de la francophonie (Belgique, Canada, Suisse, Maroc et Algérie). Seulement 8 émissions proviennent d'Israël. La répartition des émissions par média et par pays est présentée dans les tableaux 3.7 and 3.8.

Média	**Émissions**
Télévision	567
Radio	379
Internet	259
Télévision et Radio	4
Total	**1209**

Tableau 3.7 – *Nombre d'émissions par type de média*

Pays	**Émissions**
France	1027
Maroc	10
Israël	8
Suisse	6
Algérie	1
Belgique	1
Canada	1
Total	**1216**

Tableau 3.8 – *Nombre d'émissions par pays*

3.4 PTSVOX : des données au plus près du terrain

En France, le nombre de demandes de comparaisons de voix croît d'année en année. Cela est dû à la démocratisation des téléphones mobiles d'une part, mais également de l'évolution des usages. Depuis quelques années, les applications de messagerie instantanée permettent d'envoyer des messages vocaux. Cette nouvelle fonctionnalité permet aux utilisateurs de gagner du temps dans la transmission des messages. Elle est même devenue une pratique courante. De même, de plus en plus d'utilisateurs prennent le réflexe de filmer pour témoigner d'un fait ou d'une parole (altercations, menaces). Ces vidéos sont parfois partagées sur les réseaux sociaux.

Lors d'une comparaison de voix, les conditions d'enregistrement de la trace, soit l'enregistrement prélevé, ne sont jamais pleinement contrôlées. De ce fait, les fichiers à comparer peuvent avoir été enregistrés dans des contextes différents. De plus, les locuteurs peuvent ne pas se montrer coopératifs lors du prélèvement de voix et le contenu linguistique de la trace, prélevée sur la scène du crime ou du délit, peut s'avérer trop pauvre pour effectuer une comparaison de voix. À cela s'ajoute que la voix des locuteurs peut être modifiée par une affection au niveau du système respiratoire, un état psychologique intense ou encore en vue de la camoufler. Sans compter que les enregistrements peuvent ne pas contenir uniquement la voix de l'individu d'intérêt, mais peuvent également contenir un dialogue entre plusieurs personnes (voix qui se chevauchent, variations de volume sonore entre les voix, *etc.*). De plus, les enregistrements peuvent également présenter de la saturation, des bruits ambiants ou des bruits parasites. Enfin, la bande passante du téléphone filtre une bande de fréquences réduite entre 400 et 3400 Hz, ce qui a pour effet de détériorer le signal de parole.

Bien que la performance des systèmes de reconnaissance du locuteur n'ait cessé de s'améliorer au fil des campagnes d'évaluation lancées par le NIST, ces campagnes présentent très peu d'enregistrement pour un même locuteur, ce qui ne permet pas d'estimer l'impact de la variabilité intralocuteur sur les performances des systèmes. Le contexte socioculturel dans lequel évolue un individu a un impact sur ses productions langagières et ces informations ne sont pas disponibles lors des campagnes d'évaluation.

Afin de mieux pouvoir répondre aux besoins spécifiques de la criminalistique, le SNPS a mis en place la base de données PTSVOX avec l'objectif de mesurer l'influence des différents facteurs de variabilité sur la performance des systèmes de comparaison de voix.

3.4.1 Des prélèvements de voix à la base de données

La base de données PTSVOX est le résultat de campagnes de prélèvement de voix effectuées à l'école nationale de police de Nîmes et au centre de formation de police de Chassieu. Les prélèvements de voix suivent le protocole strict du SNPS. Un total de 369 personnes (144 femmes et 225 hommes) ont été enregistrées. La base est composée d'enregistrements microphoniques et téléphoniques réalisés sous forme d'entretien. L'objectif est d'atteindre la parole la plus spontanée possible. Par conséquent, le contenu linguistique et la durée des enregistrements peuvent varier d'un individu à l'autre et d'une session à l'autre. La majorité des locuteurs n'a été enregistrée qu'une seule fois alors qu'un sous-ensemble de la base, composé de 12 femmes et 12 hommes, a été enregistré à plusieurs reprises. Ces 12 locutrices et 12 locuteurs ont également été enregistrés en lecture.

Chaque session d'enregistrement contient au moins deux enregistrements de parole spontanée, l'un effectué au microphone et le second au téléphone. La base compte un total de 952 fichiers audio, ce qui correspond à plus de 80 heures de données. Les enregistrements ont été effectués avec un enregistreur de type H4n, paramétré sur une fréquence d'échantillonnage à 44100 Hz, en stéréo (car il possède deux microphones), avec une résolution de 16 bits. Trois téléphones ont été utilisés pour effectuer les enregistrements téléphoniques : un Huawei Ascend Y550 et deux Wiko Cink Slim. L'application Call Recorder, développée par Appliqato, sous la version 4.1.1 d'Android, est utilisée pour enregistrer directement les fichiers audio sur l'appareil. Les enregistrements sont paramétrés sur une fréquence d'échantillonnage de 44100 Hz, en mono, avec une résolution de 16 bits. Cependant, 27 fichiers ont été enregistrés avec une fréquence d'échantillonnage de 8000 Hz.

Pour unifier les caractéristiques techniques des enregistrements, ceux-ci sont également proposés après rééchantillonnage à 8000 Hz, 16000 Hz ou 44100 Hz, en mono. Les expériences présentées ci-après utilisent les données monocanal et rééchantillonnées à 16000 Hz.

3.4.1.1 Profil des locuteurs

Tous les locuteurs ont signé un formulaire de participation et un formulaire de consentement. Du fait que les prélèvements de voix ont été effectués dans des centres de formation, la tranche d'âge des locuteurs est assez restreinte. Ainsi 253 des 369 locuteurs sont âgés de 18 à 24 ans. Seulement 5 locuteurs avaient plus de 30 ans au moment du prélèvement. Le tableau 3.9 récapitule la répartition des locuteurs par tranche d'âge.

Tranche d'âge	Locuteurs
18-20	94
20-22	80
22-24	79
24-26	53
26-28	37
28-30	21
30+	5

Tableau 3.9 – *Nombre de locuteurs en fonction de la tranche d'âge (au moment de l'enregistrement)*

Ainsi qu'indiqué dans le tableau 3.10, le français est la langue maternelle de 346 des 369 locuteurs. D'autres langues telles que le shimaore (langue parlée à Mayotte), les créoles guyanais, guadeloupéen et réunionnais sont également mentionnées. Seize locuteurs ont aussi déclaré avoir le turc, le portugais, le berbère, le malgache, l'arabe, le bushi tongo (langue du Suriname), le kurde, le guinéen ou l'italien pour langue maternelle.

Avant chaque session d'enregistrement, les locuteurs devaient indiquer si leur voix était susceptible d'être affectée par une quelconque condition ou état de santé.

3.4.1.2 Transcriptions orthographiques et alignement phonémique

Les enregistrements ont été transcrits manuellement en utilisant le logiciel Praat (Boersma et Weenink, 2001). La transcription orthographique a été faite en préparation d'un alignement

Langue	Locuteurs
Français	346
Shimaoré	5
Créoles	8
Turc	3
Portugais	3
Berbère	2
Malgache	2
Arabe	2
Bushi tongo	1
Kurde	1
Guinéen	1
Italien	1

Tableau 3.10 – *Langues maternelles des locuteurs de PTSVOX*

phonétique automatique grâce à un outil développé par le LIA. L'alignement phonétique a ensuite été corrigé manuellement par des réservistes citoyennes de la Police nationale.

3.4.2 Protocoles utilisés

La base PTSVOX est composée de 2 sous-ensembles.

Set Intra Ce sous-ensemble est composé des 24 locuteurs enregistrés à plusieurs reprises, en parole spontanée et en lecture.

Set Inter Ce jeu de données est composé de 345 locuteurs enregistrés lors d'une seule session, en parole spontanée.

La composition des jeux de données est résumée dans le Tableau 3.11.

Pour les besoins de ce travail de recherche, les enregistrements de PTSVOX ont été découpés en « tours de parole ». Pour cela, nous avons considéré qu'un silence de 1.25 seconde ou plus marquait les frontières d'un tour de parole. Le tableau 3.12 représente le nombre de tours de parole par jeu de données.

	Femmes	Hommes	Sessions	Durée (en secondes)	
				Microphone	Téléphone
Intra	12	12	2 à 4	11755	10439
Inter	132	213	1	38341	34071

Tableau 3.11 – *Jeux de données de la base PTSVOX*

	Tours de parole	
	Microphone	**Téléphone**
Intra	2159	1094
Inter	11120	5095

Tableau 3.12 – *Jeux de données de la base PTSVOX*

La description de PTSVOX a fait l'objet d'un article présenté aux Journées d'études sur la parole à Nancy (Chanclu *et al.*, 2020).

Synthèse du chapitre

Les données utilisées sont cruciales pour l'entraînement et le test des systèmes de RAL. L'état de l'art montre que la base de données Voxceleb est aujourd'hui devenue incontournable. Dotée de plus de 7000 locuteurs, elle est largement utilisée dans la RAL dans l'apprentissage des modèles que dans leur évaluation. Toutefois, nous mettons en évidence les limites de Voxceleb pour le contexte criminalistique. D'une part, les conditions d'enregistrement sont très différentes des enregistrements utilisés lors d'affaires judiciaires. D'autre part, la variabilité intralocuteur est très peu représentée. De plus, la base de données Voxceleb présente des biais dans la représentation des locuteurs et favorise la reconnaissance de certaines populations au détriment d'autres. Ces différents problèmes peuvent entraîner des biais dans l'évaluation des performances des systèmes de RAL et, par conséquent, dans l'interprétation des résultats.

Pour répondre au manque de données permettant de modéliser la variabilité intralocuteur, nous avons conçu la base de données FABIOLE 2, dans le cadre du projet Voxcrim. Elle compte 399 locuteurs enregistrés dont plus d'une centaine de locuteurs ont été enregistrés dans au moins 30 contextes différents. Les enregistrements sont issus de plus d'un millier d'émissions de radio et de télévision. Nous mettons cependant en évidence que les profils des locuteurs sont assez homogènes puisqu'il s'agit surtout de journalistes et de politiciens.

Toujours dans le cadre du projet Voxcrim, nous avons conçu la base de données PTSVOX pour travailler sur des enregistrements plus proches des conditions réelles de la comparaison de voix. En effet, les enregistrements sont issus de campagnes de prélèvements de voix tels que généralement effectués au SNPS, ce qui n'est pas le cas des bases de données Voxceleb et FABIOLE 2. PTSVOX comprend 369 locuteurs enregistrés dans des conditions variées en termes de matériel d'enregistrement et de style de parole. Pour étudier la variabilité intralocuteur, une partie des locuteurs a été enregistrée à plusieurs reprises aussi bien en lecture qu'en parole spontanée, bien que le nombre d'enregistrements pour ces locuteurs soit limité. Néanmoins, nous relevons le même biais que pour FABIOLE 2, à savoir une variabilité des profils de locuteurs limitée puisque les locuteurs sont essentiellement des élèves policiers.

4 | Premiers éléments de la *box-rule*

Résumé : *La* box-rule *définit les conditions dans lesquelles la fiabilité d'une comparaison de voix est connue. À ce jour, cette* box-rule *n'existe pas. Pour initier sa définition, nous avons identifié des facteurs utiles qui sont la durée des enregistrements, la différence de durée entre les enregistrements, le genre, l'âge, et l'écart temporel entre les enregistrements à comparer. Pour ce faire, la base de données FABIOLE 2 est utilisée pour tester un système de comparaison de voix basé sur l'architecture ECAPA-TDNN et entraîné sur les données VoxCeleb. Nos résultats montrent que la durée des enregistrements, la différence de durée entre les enregistrements, le genre, l'âge, et l'écart temporel entre les enregistrements à comparer ont une influence sur la performance du système. Nous observons cependant que ces facteurs n'ont pas tous la même influence sur la performance du système et que tous les locuteurs ne sont pas influencés de la même manière par ces facteurs. Ces facteurs ne sont pas suffisants pour définir une* box-rule *complète, mais ils sont un premier pas vers une définition plus précise de la* box-rule.

Sommaire

4.1 Fiabilité en reconnaissance du locuteur et comparaison de voix

La question de la fiabilité des systèmes de RAL et de comparaison de voix est un travail qui a déjà été amorcé par Kahn et Ajili (Ajili, 2017 ; Kahn, 2011). Les travaux de Kahn ont permis de mettre en évidence l'importance de la variabilité intralocuteur pour la modélisation des locuteurs. En effet, les enregistrements ne modélisent pas le locuteur de manière équivalente. Certains phonèmes tels que les voyelles nasales, les voyelles semi-ouvertes et semi-fermées, les consonnes nasales et les consonnes fricatives portent davantage d'informations sur le locuteur. Aussi, une parole contrôlée contient moins d'informations discriminantes sur le locuteur qu'une parole plus spontanée. Kahn insiste sur la nécessité de distinguer locuteur et enregistrement de parole et propose la mise en place d'une mesure de confiance pour expliquer les performances des systèmes de RAL.

Ces travaux ont été poursuivis en 2017 par Ajili qui montrent que les consonnes fricatives contiennent des informations sur le locuteur dans les fréquences supérieures à 4000 Hz. Il met également en évidence le pouvoir discriminant des voyelles, des consonnes nasales et des consonnes liquides dans les comparaisons imposteur ainsi que l'importance des voyelles orales pour modéliser la variabilité intralocuteur. *A contrario*, les voyelles et consonnes nasales contiennent peu de variabilité intralocuteur. L'auteur propose d'effectuer des prédictions sur des profils de locuteurs.

Du côté de la comparaison de voix, plusieurs facteurs ont été identifiés comme pouvant en influencer la performance. C'est notamment le cas du contenu phonétique, du rythme, de la variabilité intralocuteur, mais aussi le cas de la durée et de la distance entre le locuteur et le microphone (Ajili *et al.*, 2016a ; Ajili *et al.*, 2017 ; Ajili *et al.*, 2018 ; Kahn *et al.*, 2010 ; Nandwana *et al.*, 2019).

Avec l'objectif de construire la *box-rule*, il est nécessaire de définir les conditions dans lesquelles la fiabilité des résultats d'une comparaison de voix est assurée. Pour cela, nous travaillons sur un ensemble de facteurs réduits afin de définir les conditions dans lesquelles une comparaison de voix est possible. En étudiant les variations de ces facteurs, il est possible d'observer leur influence sur la performance d'un système de comparaison de voix.

Cela permettra de définir un cadre dans lequel la fiabilité d'une comparaison de voix est considérée comme connue.

4.2 Protocole expérimental

4.2.1 Système ECAPA-TDNN

Le système de comparaison de voix utilisé repose sur ECAPA-TDNN (Desplanques *et al.*, 2020). L'architecture utilisée obtient une meilleure performance sur les protocoles de Voxceleb comparativement aux TDNN et TDNN étendus (Desplanques *et al.*, 2020). Desplanques *et al.* indiquent un *EER* égal à 0.87 sur le jeu de test *Voxceleb1 test* en utilisant 1024 canaux dans les couches convolutives.

Les réseaux de neurones ECAPA-TDNN sont essentiellement composés de couches convolutives ainsi que présenté dans la figure 4.1. Le modèle est entraîné sur les données d'entraînement de Voxceleb en utilisant la recette proposée par SpeechBrain (Ravanelli *et al.*, 2021). Les paramètres d'entrée du système comprennent 80 coefficients MFCC avec une fréquence d'échantillonnage de 16 kHz. Les coefficients MFCC sont calculés sur des trames de 25 ms avec un décalage de 10 ms. Une étape de normalisation des coefficients MFCC au niveau de la moyenne. Les couches convolutives possèdent 1024 canaux C et la taille du module d'attention est de 128. La couche entièrement connectée FC possède 192 neurones. Un vecteur de représentation des enregistrements est obtenu en sortie après la couche entièrement connectée suivie d'une normalisation en lot (FC + BN sur la figure 4.1).

4.2.2 Métriques de la performance du système

Les scores fournis par le système sont obtenus en utilisant la distance cosinus qui est interprétée comme étant un *LR*. Elle est calculée selon l'équation 4.1 où x et y représentent tous deux des vecteurs de représentation du signal.

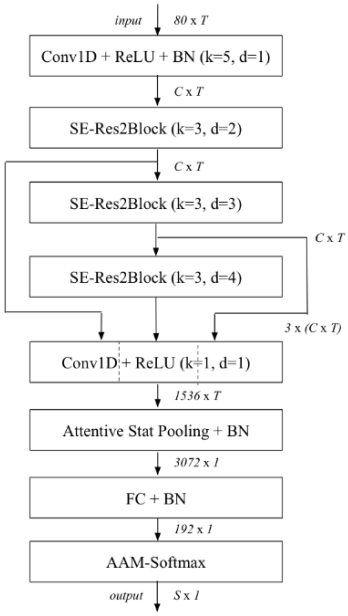


Figure 4.1 – *Architecture du réseau de neurones ECAPA-TDNN. D'après (Desplanques et al., 2020)*

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{4.1}$$

Après obtention des scores, ces derniers sont calibrés en suivant la méthode de Brummer et Van Leeuwen (Brummer et Van Leeuwen, 2006), puis le C_{llr}^{min} est calculé. Cette métrique est utilisée pour évaluer la performance du système pour minimiser l'impact des conditions de test sur la performance. Plus le C_{llr}^{min} est faible, meilleure est la performance du système.

4.2.3 Données utilisées

Dans les expériences ci-après requérant la base de données FABIOLE 2, des conditions supplémentaires sont ajoutées afin de correspondre au mieux au contexte de la comparaison de voix appliquée à la criminalistique.

Tours de parole et contextes Lors d'une comparaison de voix, qu'elle soit cible ou imposteur, les enregistrements proviennent de contextes différents. Cette condition est reproduite, quel que soit le type de comparaison. Ainsi, les deux enregistrements d'une comparaison proviennent d'enregistrements captés lors d'une émission différente et/ou un jour différent.

Équilibre entre les comparaisons cibles et les comparaisons imposteurs Pour mieux étudier la performance du système de comparaison de voix, les comparaisons cibles et imposteurs sont équilibrées, et ce, pour chacune des locutrices et chacun des locuteurs. Ainsi, chaque locuteur dispose de 400 comparaisons cibles et 400 comparaisons imposteurs. Cela permet d'interpréter les scores délivrés sans avoir à prendre en compte le biais que peut induire un fort déséquilibre entre les deux types de comparaison.

Sélection des locuteurs Les locuteurs d'intérêt sont intervenus dans au moins 30 contextes différents. Les locuteurs restants sont utilisés pour les comparaisons imposteurs. Pour chaque genre, 50 locuteurs forment les comparaisons cibles. Les locuteurs diffèrent selon les facteurs.

Même nombre de locuteurs par genre Pour chaque genre, le nombre de locuteurs est équilibré. En effet, chaque genre dispose de 50 locuteurs. Les comparaisons sont effectuées entre locuteurs du même genre.

Construction des tours de parole Les tours de parole sont définis par les transcriptions fournies avec FABIOLE 2. Afin d'obtenir un nombre de tours de parole suffisant pour les durées les plus longues, les tours de parole pour une même émission sont concaténés jusqu'à obtenir la durée désirée. *A contrario*, si la durée du tour de parole initial est supérieure à la durée désirée, le tour de parole est tronqué.

4.2.4 Facteurs étudiés

Chaque comparaison est effectuée entre une paire d'enregistrements. Afin de pouvoir étudier l'influence de différents facteurs sur la performance, les enregistrements sont sélectionnés selon les facteurs suivants.

Durée des signaux Les durées étudiées sont 3, 5, 7, 10, 20 et 30 secondes (± 10 %). Nous nous attendons à ce que la performance du système augmente avec la durée des signaux. En effet, les signaux de parole plus longs et contiennent potentiellement davantage d'informations.

Écart de durée entre les signaux à comparer En comparant des enregistrements qui ne possèdent pas nécessairement la même durée, il est possible d'étudier l'influence de l'écart de durée entre les signaux à comparer sur la performance. L'hypothèse que nous avançons est que plus l'écart de durée entre les signaux à comparer est important, plus la performance est faible.

Âge du locuteur En étudiant les scores des comparaisons des signaux de 30 secondes, nous calculons si la performance varie en fonction de l'âge du locuteur.

Genre du locuteur Les comparaisons étant effectuées entre locuteurs du même genre, il est possible d'étudier l'influence du genre du locuteur sur la performance.

Écart entre les enregistrements Puisque les comparaisons cibles sont effectuées entre des enregistrements provenant de contextes différents, il est possible d'étudier l'influence de l'écart entre les enregistrements sur la performance.

4.3 Influence de la variation de la durée de parole sur la performance

4.3.1 Trouver la durée de parole minimale pour une comparaison de voix

L'objectif de cette première expérience est de mesurer l'influence de la durée des enregistrements sur la performance de la comparaison de voix. La durée des enregistrements est un facteur important pour la comparaison de voix. Il s'agit d'un facteur facilement contrôlable déjà utilisé pour déterminer la faisabilité d'une comparaison de voix. Par exemple, le logiciel Batvox ne permet pas de comparer des enregistrements de moins de 7 secondes.

Dans cette expérience, nous étudions l'influence de la durée des enregistrements sur la performance de la comparaison de voix. Pour cela, nous nous focaliserons sur des enregistrements de 3, 5, 7, 10, 20 et 30 secondes avec l'hypothèse que la performance augmente avec la durée des enregistrements.

Différence entre durée et contenu des signaux

Pour chacune des durées étudiées, les enregistrements ont été sélectionnés aléatoirement. Bien que la durée de certains signaux soit plus longue que d'autres, cela ne signifie pas que le contenu desdits signaux est plus riche. En effet, il est possible que ces signaux plus longs contiennent autant, voire moins d'information que des signaux plus courts.

Ajili *et al.* ont montré que l'importance du contenu linguistique et phonétique sur la fiabilité de la comparaison de voix (Ajili *et al.*, 2016a ; Ajili *et al.*, 2018). Ajili *et al.* ont introduit le concept de d'homogénéité entre le contenu acoustique des signaux à comparer (Ajili *et al.*, 2015). Cette mesure d'homogénéité a permis de démontrer qu'un manque d'homogénéité entre les signaux à comparer peut dégrader la performance de la comparaison de voix (Ajili *et al.*, 2017).

La base de données FABIOLE 2 ne contient aucune information sur le contenu des signaux, il n'est donc pas possible de contrôler ce facteur.

4.3.1.1 Une stabilité de la performance à partir de 7 secondes

La figure 4.2 présente le C_{llr}^{min} obtenu pour l'ensemble des locuteurs en fonction de la durée des tours de parole. Nous pouvons constater une amélioration de la performance avec la durée des enregistrements, allant d'un C_{llr}^{min} à 0.146 pour les tours de parole de 3 secondes à 0.127 pour les tours de parole de 30 secondes. Cependant, cette amélioration n'est pas linéaire et se stabilise à partir de 7 secondes de parole où le C_{llr}^{min} varie entre 0.130 et 0.127.

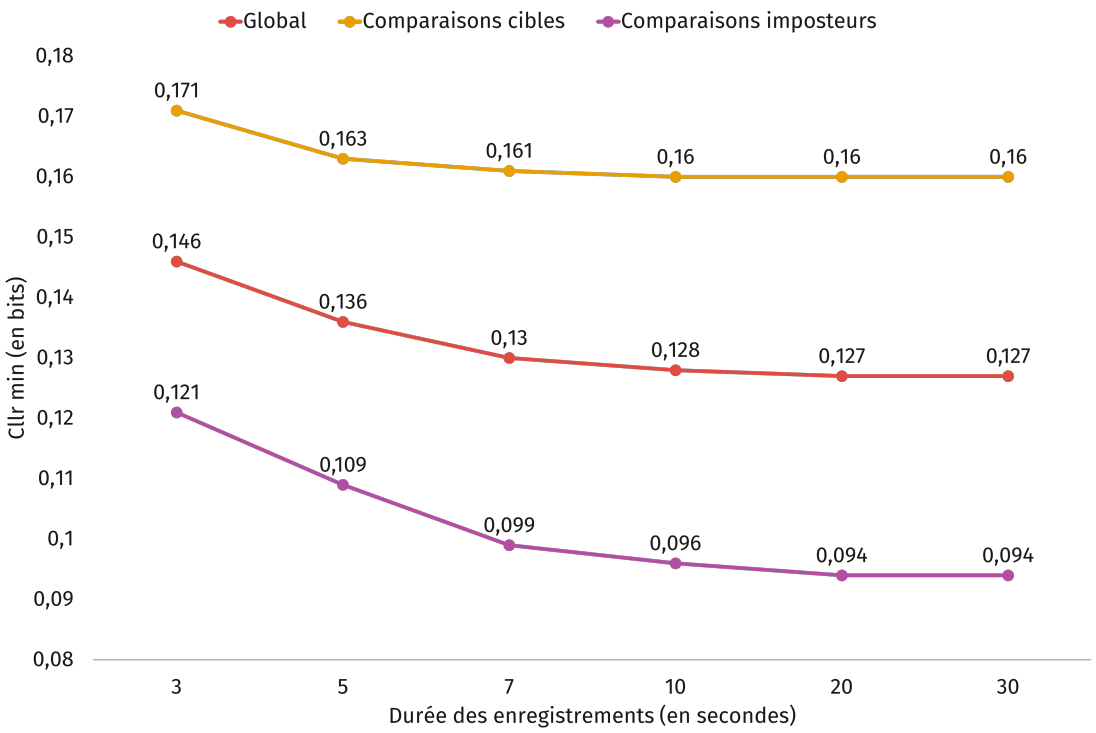


Figure 4.2 – C_{llr}^{min} *en fonction de la durée des enregistrements pour les locutrices et locuteurs*

Le calcul du C_{llr}^{min} pour les comparaisons cibles et imposteur montre un déséquilibre entre les deux types de comparaisons et indique une meilleure performance sur les comparaisons imposteurs. La différence de performance augmente avec la durée des enregistrements.

4.3.1.2 Des disparités suivant le genre

La figure 4.3 présente le C_{llr}^{min} obtenu pour les locutrices et locuteurs en fonction de la durée des tours de parole. Pour chaque genre, la tendance globale du C_{llr}^{min} est décroissante selon la durée des enregistrements. Aussi, pour chacun des deux genres, nous observons que le C_{llr}^{min} est bien plus élevé pour les comparaisons cibles que pour les comparaisons imposteurs. De manière très générale, le système a davantage de difficulté à comparer les

locutrices que les locuteurs. En effet, les locutrices présentent un C_{llr}^{min} plus élevé que les locuteurs pour toutes les durées de parole excepté pour les enregistrements de 30 secondes.

Une première différence de performance s'observe entre 3 et 5 secondes pour chacun des deux genres. Cette différence est plus marquée chez les femmes où le C_{llr}^{min} passe de 0,154 à 0,092, alors que chez les locuteurs, le C_{llr}^{min} chute de 0.135 à 0.103. En revanche, pour les locuteurs, une différence nette est observée entre 7 et 10 secondes où le C_{llr}^{min} passe de 0,131 à 0,084. Du côté des locutrices, on observe une stabilité du C_{llr}^{min} entre 0,115 et 0,083 pour les tours de parole de 5 à 20 secondes avant d'atteindre 0,099 pour les enregistrements de 30 secondes.

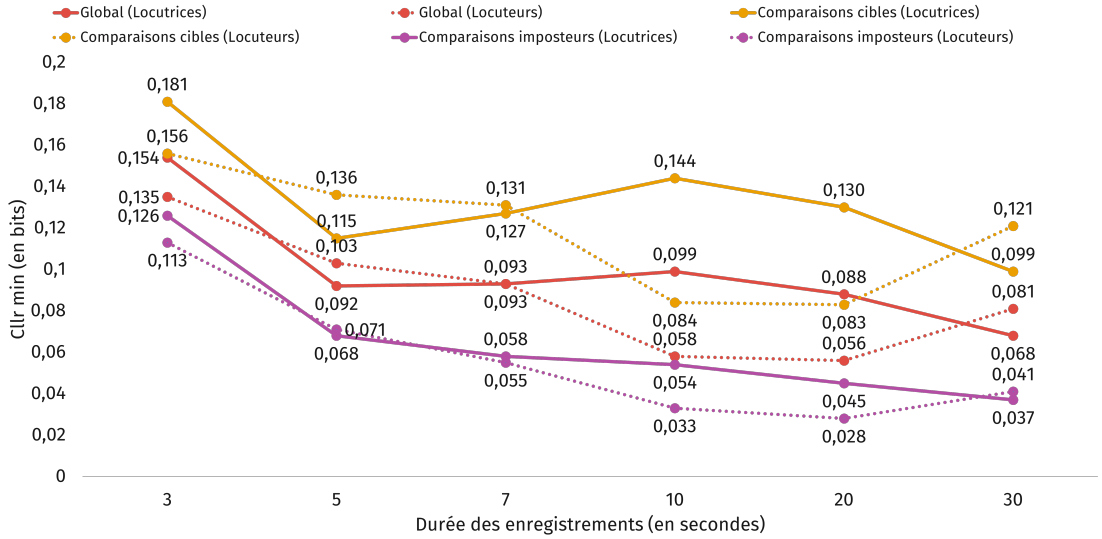


Figure 4.3 – C_{llr}^{min} *en fonction de la durée des enregistrements, en fonction du genre*

4.3.2 Comparer des enregistrements de durées différentes

4.3.2.1 Composition du corpus

En comparaison de voix, la durée de parole utile pour chaque enregistrement ne se situe pas forcément dans le même ordre de grandeur. Ainsi, certains enregistrements de quelques secondes peuvent être comparés à des enregistrements de plusieurs minutes. L'objectif de

ce chapitre est de déterminer l'influence de cet écart de durée sur la performance de la comparaison de voix. Pour cela, nous utilisons les tours de parole de FABIOLE 2 sans aucun regard sur la durée de parole de chacun des enregistrements. Pour calculer l'écart entre les enregistrements, nommé ci-après R_d, nous procédons au calcul présenté dans l'équation 4.2 où d_1 et d_2 sont les durées des deux enregistrements à comparer.

$$R_d = \frac{\max(d_1, d_2)}{\min(d_1, d_2)} \tag{4.2}$$

Les enregistrements sont ensuite répartis en 6 groupes en fonction de leur écart de durée ainsi qu'indiqué dans le tableau 4.1.

R_d	**Comparaisons**
1-1.27	13 334
1.27-1.65	13 333
1.65-2.22	13 333
2.22-3.23	13 333
3.23-5.64	13 333
5.64-261.75	13 334
Total	80 000

Tableau 4.1 – *Répartition des comparaisons en fonction de l'écart de durée entre les enregistrements à comparer.*

4.3.2.2 Privilégier de courts écarts de durée

La figure 4.4 présente le C_{llr}^{min} obtenu pour l'ensemble des locuteurs en fonction de R_d. De manière globale, le C_{llr}^{min} est égal à 0.098. Cependant, des variations de la performance sont observées en fonction de R_d. En effet, entre 1 et 3.23, le C_{llr}^{min} est plutôt stable et varie entre 0.07 et 0.078. À partir de 3.23, le C_{llr}^{min} augmente à 0.105 pour atteindre 0.17 lorsque R_d se situe entre 5.64 et 261.75.

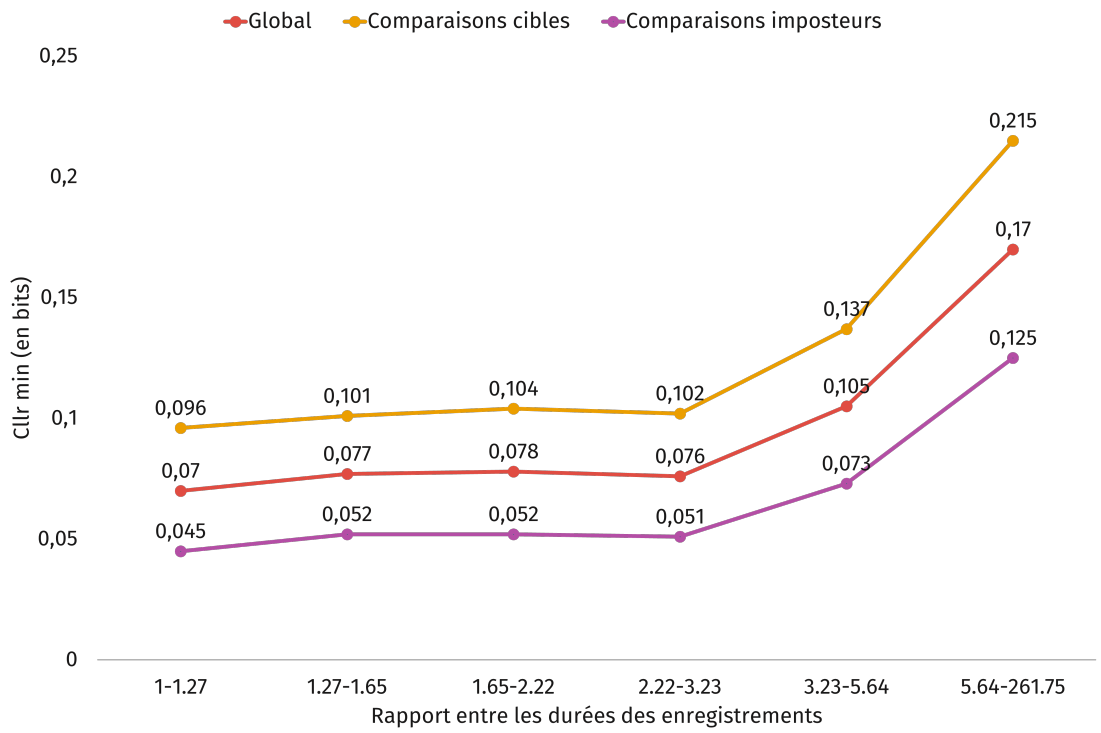


Figure 4.4 – C_{llr}^{min} *en fonction de l'écart de durée des enregistrements pour les locutrices et locuteurs*

4.4 Âge, genre et performance par locuteur

4.4.1 Comparaisons entre les enregistrements de 30 secondes

Pour évaluer la performance des facteurs non liés à la durée des enregistrements, nous utilisons les résultats des comparaisons des enregistrements de 30 secondes. Pour chaque genre, 50 locuteurs sont utilisés pour les comparaisons cibles. Ces locuteurs sont apparus dans au moins 30 contextes différents. Les autres locuteurs sont utilisés pour les comparaisons imposteurs. Pour chaque locuteur, il y a 400 comparaisons cibles et 400 comparaisons imposteurs.

4.4.2 Une meilleure performance pour les locutrices

La performance du système utilisé est présentée dans la figure 4.5 sous forme de C_{llr}^{min}. Les résultats sont présentés en fonction du genre des locuteurs. Les locutrices obtiennent un C_{llr}^{min} de 0,068 alors que les locuteurs obtiennent un C_{llr}^{min} de 0,081.

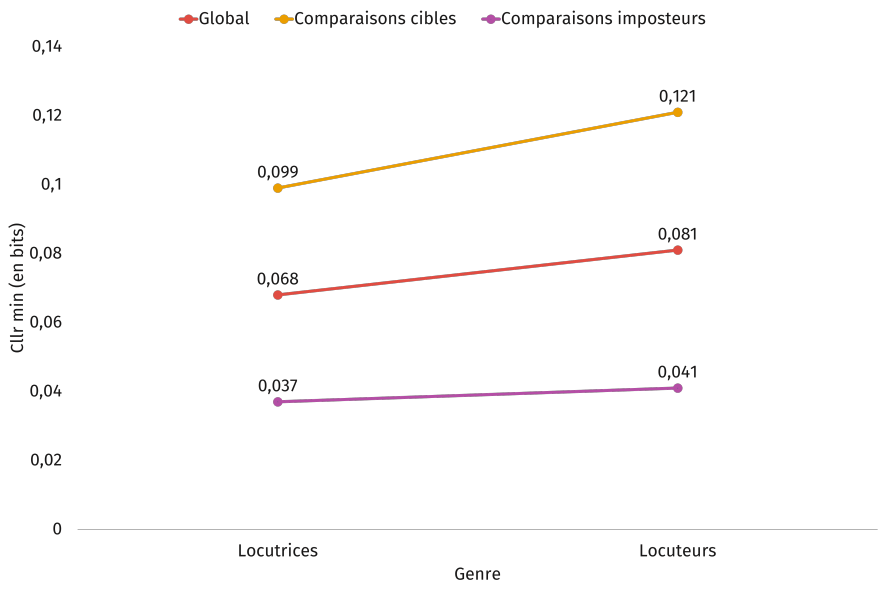


Figure 4.5 – C_{llr}^{min} *en fonction du genre*

4.4.3 Influence de l'âge sur la performance

4.4.3.1 Découpage en tranches d'âge

En vue d'évaluer l'influence de l'âge sur la performance, nous avons divisé les locuteurs en 6 groupes d'âge présentés dans le tableau 4.2. L'âge ne concerne que le locuteur cible aussi bien sur les comparaisons cibles que sur les comparaisons imposteurs. Un même locuteur peut être dans deux tranches d'âge différentes. Enfin, les résultats présentés excluent les locuteurs dont l'âge n'est pas connu. C'est la raison pour laquelle seules 66 400 comparaisons sont utilisées pour cette analyse sur les 80 000 comparaisons disponibles.

Tranche d'âge	Comparaisons
27-40	11 898
40-45	11 130
45-50	11 058
50-54	10 760
54-60	11 410
60-68	10 144
Total	66 400

Tableau 4.2 – *Répartition des comparaisons en fonction de l'écart de durée entre les enregistrements à comparer.*

4.4.3.2 Les quadragénaires sont les moins bien reconnus

La performance par tranche d'âge est présentée dans la figure 4.6. La performance est particulièrement bonne pour les locuteurs âgés de 50 à 54 ans avec un C_{llr}^{min} à 0.043. Les comparaisons qui incluent des locuteurs âgés de 40 à 50 ans obtiennent les moins bonnes performances avec un C_{llr}^{min} à 0.119 pour la tranche d'âge 40-45 et 0.092 pour celle de 40-45 ans. Le C_{llr}^{min} des locuteurs en deçà de 40 ans est de 0.064.

4.4.4 Des disparités entre les locuteurs

La performance n'est pas égale entre les locuteurs d'un même genre. Les scores de 26 locutrices et 25 locuteurs, soit la moitié du corpus, ne présentent aucun recoupement entre les comparaisons cibles et imposteurs, ce qui indique que pour ces locuteurs, aucune erreur n'a été relevée étant donné le seuil de décision choisi. La performance des autres locuteurs est présentée dans la figure 4.7 pour les locutrices et dans la figure 4.8 pour les locuteurs.

Chez les locutrices, 5 ont obtenu un C_{llr}^{min} supérieur à 0,15. Pour les locuteurs, ce nombre s'élève à 4. Cela dit, une plus grande variation du C_{llr}^{min} est observée entre ces 4 locuteurs.

Le locuteur LOC046 est celui qui obtient la performance la plus faible avec un C_{llr}^{min} de 0,457. Chez les femmes, il s'agit de LOC013 avec un C_{llr}^{min} de 0,183.

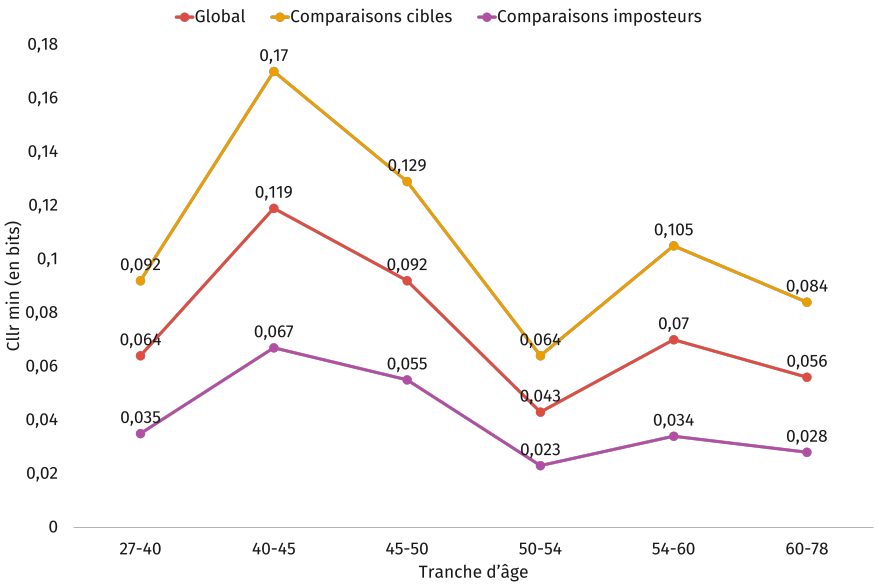


Figure 4.6 – C_{llr}^{min} *en fonction de l'âge, tous genres confondus*

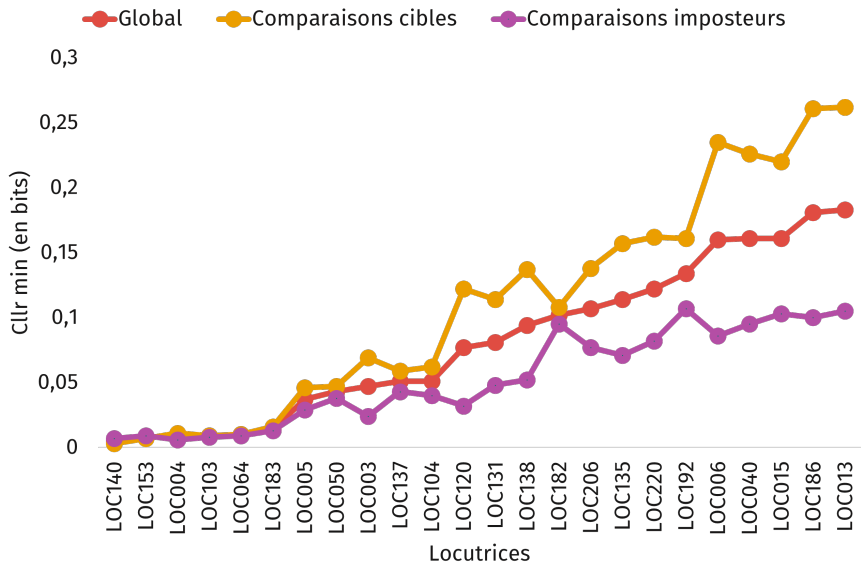


Figure 4.7 – *Performance par locutrice lorsque* $C_{llr}^{min} > 0$

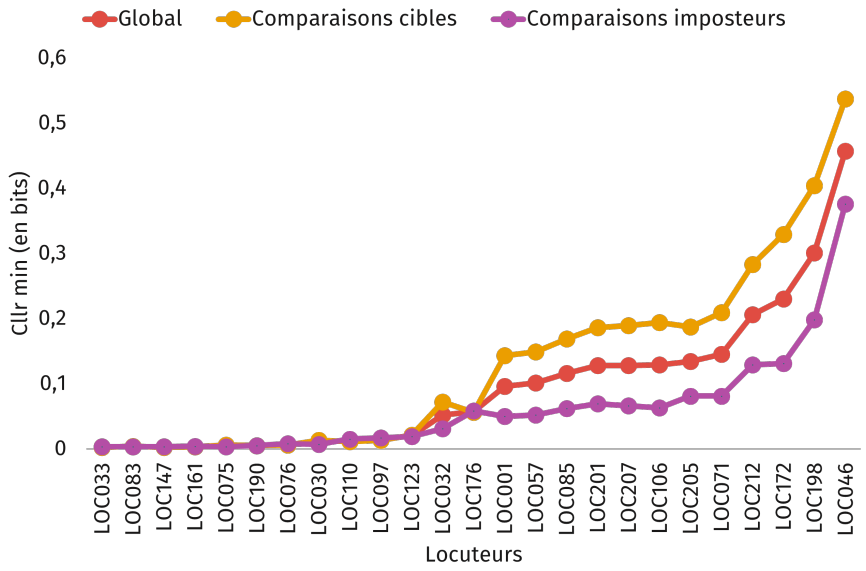


Figure 4.8 – *Performance par locuteur lorsque* $C_{llr}^{min} > 0$

4.5 Écart temporel entre les enregistrements à comparer

4.5.1 Certaines voix varient davantage dans le temps

La base de données FABIOLE 2 contient des enregistrements effectués entre le 30 novembre 2014 et le 16 novembre 2020, soit environ 6 années. Pour autant, cela ne signifie pas que, pour tous les locuteurs, les enregistrements couvrent l'entièreté de cette période. Nous étudions néanmoins l'influence de l'écart temporel entre les enregistrements sur les scores de comparaison de voix. Pour cela, nous travaillons sur les résultats des comparaisons cibles réalisées avec des enregistrements de 30 secondes. Cela représente 400 comparaisons cibles pour les 50 locuteurs de chaque genre, soit 40 000 comparaisons cibles au total.

Pour calculer l'écart temporel entre deux enregistrements, nous utilisons la date de l'enregistrement. L'écart temporel est calculé en jours, dont la distribution est présentée dans la figure 4.9. L'écart temporel maximum entre deux enregistrements est de 2055 jours, soit plus de 5 ans et demi.

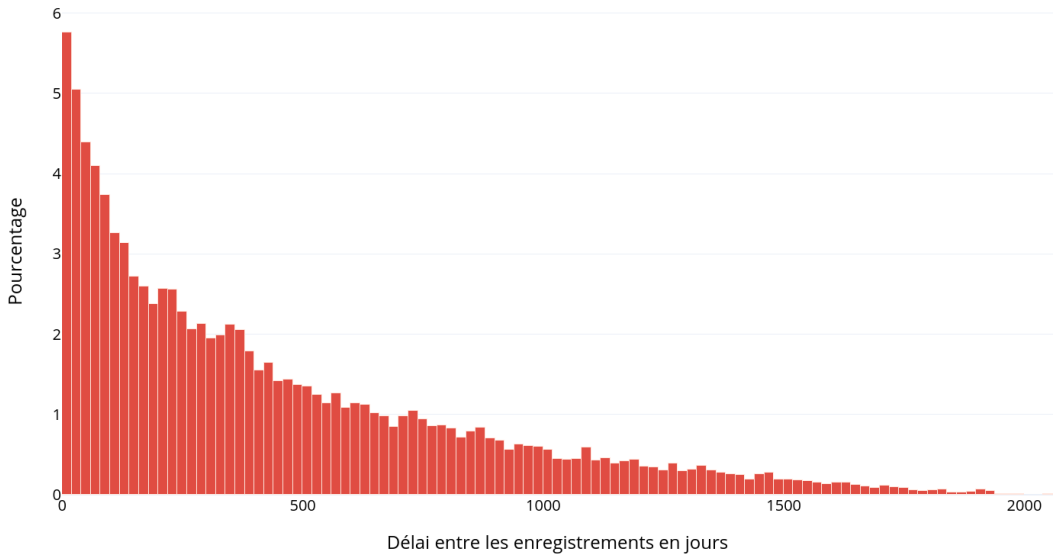


Figure 4.9 – *Distribution des écarts temporels entre les enregistrements*

Pour calculer l'influence de l'écart temporel entre les enregistrements, nous calculons le coefficient de corrélation de Pearson entre les scores de comparaison et l'écart temporel entre les enregistrements. Le coefficient de corrélation de Pearson est défini par l'équation 4.3 où x représente les scores obtenus et y l'écart temporel (en jours) entre les enregistrements. Il est compris entre -1 et 1 où -1 indique une corrélation négative parfaite, 0 indique une absence de corrélation et 1 indique une corrélation positive parfaite.

$$r_{xy} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.3}$$

Sur l'ensemble des comparaisons cibles, $r = -0,18$ ($p < 0.001$), soit une très faible corrélation négative. Le calcul du coefficient de corrélation de Pearson sur les fausses acceptations ne montre pas de corrélation significative ($r = -0,09$, $p = 0.05$).

Le coefficient de corrélation de Pearson est calculé pour chaque locuteur, d'abord sur l'ensemble des comparaisons cibles, puis sur les fausses acceptations s'il y a. Les locuteurs sont classés par ordre croissant de la valeur du coefficient de corrélation de Pearson sur l'ensemble des comparaisons cibles et sont présentés dans les tableaux 4.3 and 4.4.

Les résultats montrent un effet de l'écart temporel entre les enregistrements très significatif pour la moitié des locuteurs. Pour l'ensemble des comparaisons cibles, une corrélation négative significative est observée pour les locuteurs LOC109, LOC139, LOC039, LOC213, LOC148 et LOC054 avec un coefficient variant entre -0.70 pour le locuteur LOC109 et -0.50 pour le locuteur LOC054. La significativité décroît au fur et à mesure que les coefficients de corrélation de Pearson augmentent.

r	p	**Locuteurs**
$-0.70 \geq p > -0.19$	$p < 0.0001$	49
$-0.19 \geq p > -0.17$	$p < 0.001$	5
$-0.17 \geq p > -0.13$	$p < 0.01$	9
$-0.13 \geq p \geq -0.08$	$p < 0.1$	11
$-0.08 \geq p > 0.07$	$p \geq 0.1$	26

Tableau 4.3 – *Coefficients de corrélation r pour les comparaisons cibles*

Locuteur	*p*	*r*
LOC003	-1.0	$p \geq 0.1$
LOC138	**-0.76**	$p < 0.1$
LOC006	**-0.5**	$p < 0.1$
LOC057	-0.3	$p \geq 0.1$
LOC040	0.02	$p \geq 0.1$
LOC201	0.02	$p \geq 0.1$
LOC046	0.03	$p \geq 0.1$
LOC106	0.03	$p \geq 0.1$
LOC120	0.06	$p \geq 0.1$
LOC172	0.06	$p \geq 0.1$
LOC135	0.08	$p \geq 0.1$
LOC186	0.13	$p \geq 0.1$
LOC207	0.14	$p \geq 0.1$
LOC205	0.16	$p \geq 0.1$
LOC085	0.17	$p \geq 0.1$
LOC198	0.18	$p \geq 0.1$
LOC071	0.20	$p \geq 0.1$
LOC001	0.22	$p \geq 0.1$
LOC013	**0.47**	$p < 0.1$

Tableau 4.4 – *Coefficients de correlation r pour les FA*

4.6 Synthèse des premiers éléments de la *box-rule*

Jusqu'à présent, le concept de *box-rule* n'était pas défini. L'objectif de ce chapitre est donc d'évaluer l'influence des facteurs de variabilité dans la performance d'un système de comparaison de voix en vue de définir les premiers éléments de la *box-rule*. Le système utilisé est construit sur l'architecture ECAPA-TDNN via le toolkit SpeechBrain. Le modèle est entraîné sur les données d'entraînement de VoxCeleb. Pour tester ce modèle, nous avons utilisé les données de FABIOLE 2. Nous avons évalué la performance au moyen du C_{llr}^{min}. Les facteurs étudiés sont les suivants :

- la durée des enregistrements ;
- la différence de durée entre les enregistrements ;
- le genre ;
- l'âge ;

- le locuteur cible ;
- ’écart temporel entre les enregistrements à comparer pour les comparaisons cibles.

4.6.1 Durée des enregistrements

Pour évaluer l’influence de la durée des enregistrements, nous avons travaillé sur des durées de 3, 5, 7, 10, 20 et 30 secondes. Pour chacune de ces durées, 50 locutrices et 50 locuteurs ayant chacun un minimum de 30 contextes différents ont été sélectionnés aléatoirement. Puis, pour chacun de ces locuteurs, nous avons formé 400 comparaisons cibles et 400 comparaisons imposteurs. Les locuteurs ayant moins de 30 contextes différents ont été utilisés pour les comparaisons imposteurs.

Le travail sur la durée des enregistrements a montré un C_{llr}^{min} compris entre 0.127 et 0.146 pour l’ensemble des durées. À partir de 7 secondes d’enregistrement, la performance s’est stabilisée entre 0.130 et 0.127. Nous observons des différences entre les genres qui n’ont pu être expliquées par la seule condition de durée. D’une part, jusqu’à 10 secondes de parole, la performance a été meilleure sur les locuteurs que les locutrices. D’autre part, nous constatons qu’à partir de 10 secondes, la tendance s’inverse et les locutrices ont présenté une meilleure performance que les locuteurs. Enfin, la performance des locuteurs est repassée devant celle des locutrices pour les enregistrements de 30 secondes. Il est donc nécessaire de poursuivre les travaux sur l’influence de la durée des enregistrements en prenant en compte d’autres facteurs tels que le contenu phonétique.

4.6.2 Rapport entre les durées des enregistrements

Nous avons également travaillé sur les différences de durée entre les enregistrements. Nous avons pu montrer que les écarts de durée réduits permettent au système d’obtenir une meilleure performance. Ainsi, les meilleurs résultats sont obtenus pour les comparaisons dont les enregistrements ont une durée similaire, avec un rapport oscillant entre 1 et 1.27. En effet, lorsque l’écart se situe entre 1 et 1.27 seconde, le C_{llr}^{min} était de 0.07 et une hausse est observée à partir d’un 3.23 secondes d’écart où il a atteint 0.105. Ces résultats

restent à approfondir sur des durées plus courtes, notamment sur des durées inférieures à 10 secondes.

4.6.3 Genre et âge

Nous avons réalisé des études sur le genre et l'âge sur les comparaisons réalisées avec les enregistrements de 30 secondes utilisés pour l'étude sur la durée. Les résultats confortent ceux vus précédemment. Ainsi, pour le genre, le système a mieux reconnu les locutrices que les locuteurs, ce n'est pas le cas des durées de 10 et 20 secondes. Le manque d'information sur les locuteurs ainsi que sur le contexte d'enregistrement ne permet pas d'approfondir l'analyse pour expliquer les résultats. Par conséquent, nous encourageons à poursuivre les travaux pour expliquer cette différence de performance entre les durées.

Du côté des locuteurs, nous observons une disparité : tous les locuteurs cibles n'ont pas été reconnus de la même manière. En effet, bien que le C_{llr}^{min} pour la moitié des locuteurs soit égal à 0, 5 locutrices et 4 locuteurs ont obtenu une performance supérieure à 0,150. Nous mettons en évidence et que l'âge est également un facteur important puisqu'il permet d'observer une différence de performance entre les tranches d'âge. En effet, les personnes de 40 à 50 sont les moins bien reconnues avec un C_{llr}^{min} égal à 0.119. En revanche, les locuteurs âgés de 50 à 54 ans observent la meilleure performance avec un C_{llr}^{min} à 0.064. Toutefois, les tranches d'âge se basent uniquement sur le nombre de comparaisons et non sur des caractéristiques physiologiques. Il est donc nécessaire de poursuivre les travaux sur l'influence de l'âge sur la performance d'une comparaison de voix, en redéfinissant les tranches d'âge en fonction de caractéristiques physiologiques.

4.6.4 Écart temporel entre les enregistrements à comparer

Pour analyser l'influence de l'écart temporel entre les enregistrements à comparer, nous avons utilisé les comparaisons cibles réalisées avec des enregistrements de 30 secondes. Afin d'estimer cette influence, nous avons utilisé le coefficient de corrélation de Pearson. Ce dernier montre une corrélation négative sur l'ensemble des locuteurs, ce qui indique que

plus l'écart temporel entre les enregistrements à comparer sont longs, plus la performance décroît. Pour 49 locuteurs, nous observons une forte corrélation négative significative. Cependant, pour 37 locuteurs, la corrélation est très peu, voire pas du tout significative. D'autres travaux sont nécessaires pour comprendre cette différence entre les locuteurs, comprendre pourquoi certains locuteurs sont plus sensibles à l'influence de l'écart temporel que d'autres.

4.6.5 Lacunes dans la modélisation de la variabilité intralocuteur

Pour chacun des facteurs étudiés, nous avons remarqué que le système utilisé était plus performant sur les comparaisons imposteurs que les comparaisons cibles. Nous dressons l'hypothèse que le modèle est insuffisamment entraîné pour modéliser la variabilité intralocuteur, plus difficile à modéliser que la variabilité interlocuteur (Ajili *et al.*, 2017 ; Kahn *et al.*, 2010). Cela s'explique par le fait que les données utilisées pour l'entraînement du modèle contiennent assez peu d'enregistrements par locuteur.

4.6.6 Limites de la base de données FABIOLE 2

Nous avons défini les premiers éléments de la *box-rule*. Cependant, ces éléments doivent être approfondis par d'autres analyses. Les informations fournies avec la base de données FABIOLE 2 ne permettent pas de contrôler davantage de facteurs, notamment le contenu phonétique et le contexte d'enregistrement, ce qui limite l'analyse. Dans l'idéal, il aurait fallu que FABIOLE 2 couvre suffisamment tous les facteurs de variabilité entre les pièces à comparer afin que tous puissent être étudiés indépendamment et en interaction.

5 | Évaluer la capacité humaine à regrouper les locuteurs

Résumé : *En criminalistique, il peut être nécessaire de déterminer le nombre de locuteurs présents dans un ensemble d'enregistrements de parole. Pour évaluer la capacité humaine à identifier les locuteurs, nous avons conçu une expérience perceptive dans laquelle des enregistrements de parole sont présentés à des auditeurs avec la consigne de les regrouper en locuteurs. Nous avons choisi cette approche afin de minimiser les risques de biais induits par l'effet d'amorçage, rencontrés dans les comparaisons binaires. Pour évaluer la performance des auditeurs, nous avons introduit un nouveau protocole de test perceptif utilisant la notion de pureté de regroupement, comme utilisé pour la segmentation et le regroupement automatiques en locuteurs. Nos résultats montrent que les auditeurs ont obtenu une performance de 86.40 % en moyenne avec un effet significatif de la langue maternelle. Une comparaison avec une approche automatique montre que cette dernière est plus performante que les auditeurs avec une performance moyenne de 97.47 %.*

Sommaire

La perception humaine des locuteurs est variable en fonction de plusieurs facteurs tels que la personne considérée, l'écart temporel entre l'exposition à la voix et la tâche de reconnaissance ou encore la durée des stimuli. Cela s'explique notamment par l'effet d'amorçage où l'exposition d'un stimulus entraîne une réponse à un autre stimulus, et ce, de manière inconsciente (Bargh et Chartrand, 2014). L'effet d'amorçage se retrouve dans plusieurs études sur la parole ou sur la perception des locuteurs. À titre d'exemple, dans le cadre d'une tâche de reconnaissance de mots parlés, les résultats montrent que les mots sont reconnus plus rapidement lorsqu'ils ont été prononcés par la même personne (Craik et Kirsner, 1974). L'effet d'amorçage permet également d'augmenter la perception humaine des locuteurs ainsi que le montrent Schweinberger *et al.* (Schweinberger *et al.*, 1997). Dans leur expérience, les scientifiques allemands ont étudié l'effet d'amorçage dans une tâche d'identification des locuteurs. Pour cela, les auditeurs ont été soumis à plusieurs conditions d'amorçage : signaux de parole (personnalités cibles et imposteurs), portrait, nom et aucun amorçage. Les résultats de cette expérience montrent que, lorsque les auditeurs sont amorcés par la voix des personnalités cibles, le temps de réaction moyen est diminué, passant de 1.8 à 1.5 seconde ainsi que le taux d'erreur, allant de 27.7 % 18.5 %.

Pour évaluer la capacité humaine à identifier les personnes à leur voix, il est nécessaire de minimiser les risques de biais afin d'appréhender la tâche le plus objectivement possible. Afin de limiter les effets d'amorçage que peuvent induire les tests binaires, nous avons conçu une expérience perceptive dans laquelle des enregistrements de parole sont présentés aux auditeurs qui doivent les regrouper en locuteurs. Ainsi, tous les enregistrements sont présentés en même temps aux auditeurs qui peuvent les écouter et réécouter dans l'ordre souhaité et les comparer les uns par rapport aux autres.

Challenge VoicePrivacy L'expérience décrite dans ce chapitre a été mise en place dans le cadre du challenge VoicePrivacy, qui vise à grouper ensemble des signaux de voix anonymisée et non anonymisée d'une même personne (Tomashenko *et al.*, 2020). VoicePrivacy se concentre sur l'anonymisation de la voix avec l'objectif de laisser le moins d'informations personnelles sur le locuteur dans le signal de parole tout en préservant l'intelligibilité qui permet à la locutrice ou au locuteur d'effectuer les actions désirées (consulter son compte en banque par exemple). En parallèle, l'anonymisation de la voix doit également empêcher de retrouver le locuteur d'origine.

5.1 Catégoriser des enregistrements de parole en locuteurs

L'objectif de la tâche de regroupement en locuteurs proposée ici est de regrouper les enregistrements dans un à quatre groupes, où chaque groupe représente supposément un locuteur différent. Pour cela, une interface a spécialement été développée au sein du LIA pour répondre aux besoins de l'expérience. Elle permet de suivre les différentes actions des auditeurs :

- nombre d'écoutes pour chaque enregistrement ;
- durée totale d'écoute pour chaque enregistrement ;
- nombre de mouvements pour chaque enregistrement ;
- durée totale des mouvements pour chaque enregistrement ;
- nombre de changements de classe pour chaque enregistrement ;
- position finale de chaque enregistrement dans le canevas ;
- durée totale de l'expérience.

Après que les auditeurs se sont connectés avec les identifiants fournis, ils sont invités à lire les consignes de l'expérience, présentées sous forme de texte et sous forme de vidéo, aussi bien en français qu'en anglais. Après avoir pris connaissance des consignes, les auditeurs débutent la tâche de regroupement en locuteurs. Dès le début de l'expérience, les auditeurs ont accès à tous les enregistrements qui ont été disposés aléatoirement sur le canevas comme illustré dans la figure 5.1. L'interface leur laisse la possibilité de les écouter et de les changer de catégorie autant de fois que nécessaire.

5.1.1 Données

Les enregistrements sont issus de la base de données VCTK (Yamagishi *et al.*, 2019) composée d'enregistrements de phrases lues par 110 anglophones, dédiée à la synthèse de la parole, et plus particulièrement du jeu de données *VCTK-test (common)*. Le corpus comprend les énoncés 1 à 24 de 15 locutrices et 15 locuteurs, résultant un total de 700 énoncés [1]. Les 15 locutrices et les 15 locuteurs sont séparés de façon aléatoire en deux groupes :

1. Tous les énoncés n'ont pas été prononcés par tous les locuteurs.

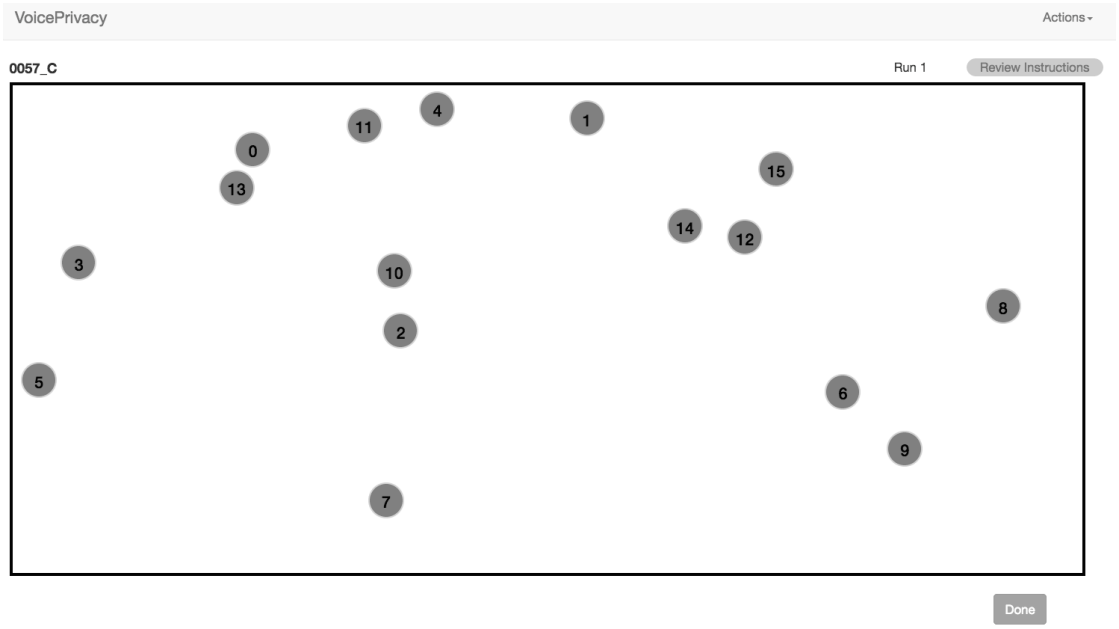


Figure 5.1 – *Capture d'écran de l'interface au début de l'expérience*

- 9 locuteurs de référence qui sont les locuteurs d'intérêt ;
- 6 distracteurs qui sont introduits pour apporter du « bruit » à l'expérience.

Le tableau 5.1 présente la répartition des locuteurs ainsi que leur nombre d'enregistrements.

Chacun des enregistrements a été tronqué après 3 secondes de parole. Pour améliorer le confort à l'écoute, un fondu en fermeture d'une durée de 250 millisecondes fait suite à ces 3 secondes. Les enregistrements tronqués sont dénommés ci-après stimuli. Chaque auditeur doit catégoriser 16 stimuli au contenu linguistique différent. Ces 16 stimuli sont prononcés par 4 personnes du même genre : 3 locuteurs de référence et 1 distracteur. Chaque locuteur de référence possède 2 à 6 stimuli et le distracteur n'en possède qu'un seul.

5.1.2 Auditeurs

Un total de 75 auditeurs (48 hommes et 27 femmes) ont participé à l'expérience. Une auditrice a été écartée du fait de très mauvais résultats. Ces personnes comptent 38 locuteurs

Locuteur	Genre	Énoncés
p225	Femme	23
p228	Femme	23
p229	Femme	24
p230	Femme	20
p231	Femme	23
p233	Femme	23
p236	Femme	21
p239	Femme	24
p244	Femme	23
p250	Femme	24
p257	Femme	24
p267	Femme	24
p268	Femme	22
p269	Femme	24
p226	Homme	24
p227	Homme	23
p232	Homme	24
p243	Homme	23
p254	Homme	24
p256	Homme	24
p258	Homme	24
p259	Homme	24
p270	Homme	24
p273	Homme	23
p274	Homme	23
p279	Homme	23
p286	Homme	24
p287	Homme	24

Tableau 5.1 – *Répartition des locuteurs. En couleur sont les distracteurs.*

natifs du français, 28 locuteurs natifs de l'anglais et 9 personnes natives d'une autre langue. Le tableau 5.2 décrit les informations sur les auditeurs.

Langue maternelle	Genre	Auditeurs
Français	Femme	11
Français	Homme	27
Anglais	Femme	9
Anglais	Homme	19
Italien	Femme	3
Allemand	Femme	1
Arabe	Homme	1
Kurde	Homme	1
Polonais	Femme	1
Portugais	Femme	1
Russe	Femme	1

Tableau 5.2 – *Répartition des locuteurs. En couleur sont les distracteurs.*

5.1.3 Déroulé d'une session de l'expérience

Une session de l'expérience se décompose en trois essais : un essai de contrôle et deux essais d'évaluation. L'essai de contrôle contient uniquement des stimuli de voix non anonymisée et permet d'obtenir une base de référence pour les essais d'évaluation. Les essais d'évaluation ont pour objectif d'estimer la performance de l'anonymisation de la voix des locuteurs. Chaque essai contient 16 stimuli au contenu linguistique différent. Ces stimuli sont répartis entre quatre locuteurs : 3 locuteurs de références et 1 distracteur. Au cours d'une session, un locuteur n'est rencontré qu'une seule fois, et tous les locuteurs sont du même genre. Pour chaque essai, chaque locuteur de référence possède 2 à 6 stimuli et le distracteur n'en possède qu'un seul.

Nous nous intéressons ici uniquement aux essais de contrôle pour étudier la perception de l'identité des locuteurs.

5.1.4 Métriques

Afin de respecter les consignes qui incitent les auditeurs à créer des groupes contenant 1 locuteur, nous introduisons une métrique que nous appelons la pureté du regroupement.

Cette métrique part du postulat que les auditeurs ont cherché à obtenir le meilleur résultat lors de leur regroupement, et que chaque groupe dispose d'une étiquette différente. Le calcul de la pureté du regroupement p est décrit dans l'équation 5.1. Soit :

- un test M ;
- un groupe d'enregistrements m dans M ;
- les différentes combinaisons d^k pour chaque locuteur unique assigné à un groupe d'enregistrements dans M ;
- nombre d'enregistrements N pour le test.

$$p(M) = \max_k \frac{1}{N} \sum_{m \in M} m \cap d_m^k \tag{5.1}$$

Les valeurs de la pureté du regroupement sont exprimées entre 0 et 100 % où 100 % représente un regroupement sans erreur.

La pureté du regroupement est mise en perspective avec la pureté présentée dans l'équation 5.2. Il y a deux différences fondamentales entre la pureté du regroupement et la pureté :

1. la pureté du regroupement n'admet qu'une seule étiquette par groupe ;
2. la pureté du regroupement n'admet pas la même étiquette pour plusieurs groupes différents.

$$purete(M) = \frac{1}{N} \sum_{m \in M} m \cap d \tag{5.2}$$

Exemple La figure 5.2 montre un exemple de regroupement et l'utilisation de la pureté du regroupement. Dans cette illustration, 16 enregistrements sont prononcés par 4 locuteurs différents ainsi définis : Le Dixième Docteur, le Onzième Docteur, un Cyberman et René Magritte. Les 16 enregistrements ont été répartis dans 4 groupes. La combinaison entre les groupes et les locuteurs qui maximise la pureté du regroupement est présentée dans la figure 5.2. Il est important de noter que René Magritte est l'étiquette d'un groupe dans lequel il n'y a aucun enregistrement de sa voix. Ici, la pureté du regroupement est égale à 0.75, alors que la pureté revient à 0.88.

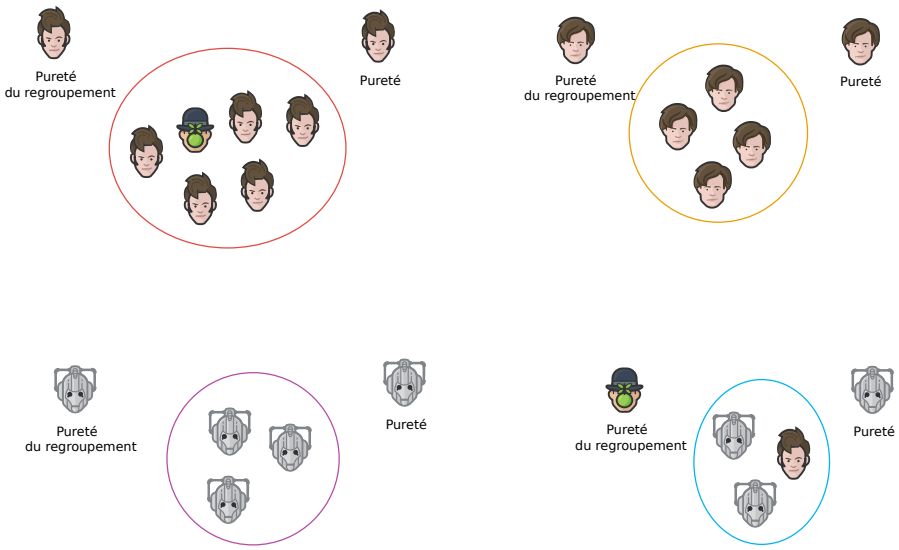


Figure 5.2 – *Étiquetage des groupes en fonction de la pureté du regroupement et de la pureté*

5.1.5 Comparaison avec une approche automatique

Afin de mettre en perspective la performance des auditeurs, nous comparons leurs résultats avec ceux d'une approche automatique, et par conséquent, objective. L'approche automatique a pour objectif de simuler autant que possible l'expérience perceptive de regroupement des locuteurs. Pour cela, les mêmes tests de l'expérience perceptive sont utilisés.

Le système utilisé repose sur l'approche *x-vector* (Snyder *et al.*, 2018) pour la représentation des enregistrements. Le modèle du système *x-vector* est appris sur le jeu de données d'apprentissage *LibriSpeech train-clean-360* (Panayotov *et al.*, 2015) dont la fréquence d'échantillonnage est de 16 000 Hz. La paramétrisation comprend 30 MFCC calculés sur des fenêtres de 25 ms avec un décalage de 10 ms sur une bande passante allant de 20 à 7600 Hz.

Afin de simuler le regroupement en locuteurs, nous utilisons le regroupement hiérarchique (*hierarchical clustering*) avec des liaisons de Ward et une distance euclidienne entre les groupes. Cette approche permet de simuler le nombre variable de groupes pour un test. À l'instar de l'expérience perceptive, la pureté du regroupement permet de déterminer le

nombre de groupes et l'étiquetage de ces derniers. Le regroupement hiérarchique est effectué à partir de *x-vectors* extraits directement des enregistrements audio.

Dendrogrammes Les dendrogrammes sont des représentations graphiques du regroupement hiérarchique. Ils représentent les différents groupes ainsi que la distance entre les groupes. Les dendrogrammes font apparaître les différents groupes possibles allant du plus large (un groupe contenant tous les enregistrements) au plus fin (un groupe par enregistrement). La distance est représentée par les lignes verticales : plus la ligne est longue, plus la distance entre deux groupes est élevée. La figure 5.3 montre le dendrogramme pour le test `0125_C`.

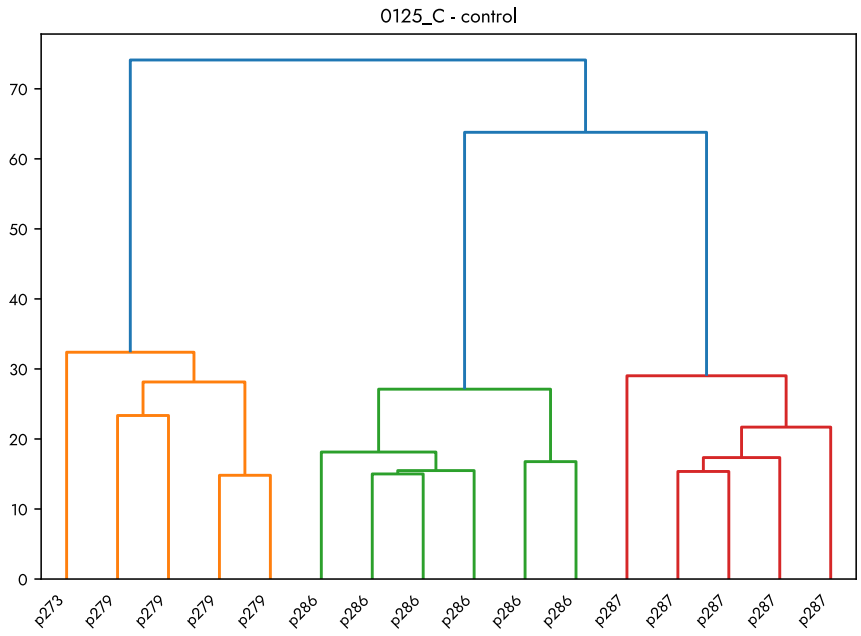


Figure 5.3 – *Exemple de dendrogramme mettant en évidence trois groupes*

5.2 Évaluer la capacité humaine à regrouper les enregistrements en locuteurs

Pour atteindre une performance optimale, chaque essai doit contenir quatre groupes contenant un seul et unique locuteur. Sur ces critères, la performance humaine est relativement bonne. En effet, ainsi que rapporté dans le tableau 5.3, les auditeurs ont formé en moyenne 3.79 groupes qui contiennent 1.33 locuteur. Cela représente 14 % de *FA*. Le nombre de changements de classe est également peu élevé (1.09 [2] en moyenne), ce qui indique que très peu de stimuli ont suscité une hésitation.

	Moyenne (μ)	**Écart-type (δ)**
Groupes	3.79	0.48
Loc. par groupe	1.33	0.55
Chang. de classe	1.09	0.37
Pureté du regr.	86.40 %	11.87
Pureté	89.70 %	10.35

Tableau 5.3 – *Performance générale des auditeurs*

5.2.1 Les locuteurs ne sont pas tous égaux dans leur reconnaissance

Alors que la performance générale est bonne, les résultats par locuteur montrent que certaines voix portent davantage à confusion. Les figures 5.4 and 5.5 montrent respectivement une très bonne reconnaissance générale des locuteurs pour cette tâche, non généralisable à la comparaison de voix en criminalistique. Cependant, la performance est plus élevée pour les locuteurs que les locutrices. La pureté du regroupement moyenne pour les locutrices est de 84.82 (δ = 11.43) et de 88.48 (δ = 12.13) pour les locuteurs.

Chez les locutrices, p236 obtient la meilleure précision avec 95 % de ses stimuli dans ses groupes. Les autres stimuli de cette locutrice se trouvent dans les groupes de la locutrice p240. En revanche, le contraire ne se vérifie pas. La locutrice p239 observe le *FR* le

2. L'attribution d'une classe compte pour un changement de classe.

plus élevé. En effet, celui-ci s'élève à 26 %. La présence des distractrices ont conduit les auditeurs à regrouper ses stimuli dans les groupes des distractrices. Nous pouvons observer une confusion réciproque entre les locutrices p231 et p269 où 6 stimuli de l'une ont été attribués à l'autre. Les groupes de la locutrice p240 accueillent également des stimuli de p236, p250 et p257. La locutrice p239 est celle qui obtient la moins bonne performance avec 76.47 % de ses stimuli dans les groupes qui lui sont dédiés. Elle a été confondue avec les distractrices p225, p244, p267 et p268.

La plupart des voix des locuteurs ont été regroupées entre elles. Les locuteurs p258 et p259 sont ceux qui possèdent la performance la plus basse avec respectivement 83 % et 77 % de stimuli bien classifiés. Ils ont été confondus mutuellement. La performance pour les autres locuteurs dépasse 91 % excepté pour p274 dont la performance est égale à 84 %.

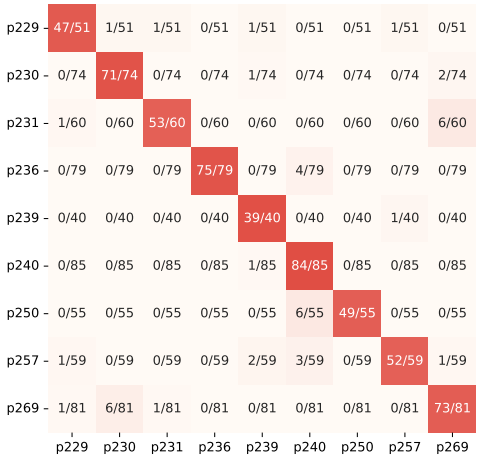


Figure 5.4 – *Matrice de confusion des locutrices après regroupement humain*

5.2.2 Performance individuelle et effets constatés

La performance est variable selon l'auditeur. Les auditeurs ont obtenu une performance légèrement plus élevée (μ = 86.73, δ = 11.55) que les auditrices (μ = 85.75, δ = 12.46). 17 auditeurs ont obtenu une pureté du regroupement égale à 100 %. À titre de comparaison, la pureté du

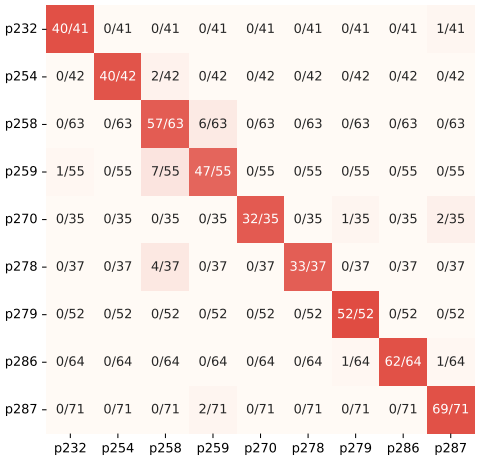


Figure 5.5 – *Matrice de confusion des locuteurs après regroupement humain*

regroupement de l'auditrice 0155 est égale à 56.25 %. Son regroupement est présenté dans la figure 5.6. L'auditrice a pu regrouper les deux tiers des stimuli du locuteur p254. En revanche, pour les autres locuteurs, la bonne catégorisation ne dépasse pas 50 %.

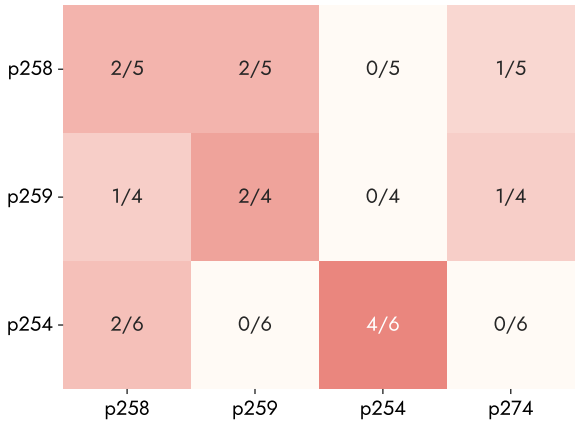


Figure 5.6 – *Matrice de confusion de l'auditrice 0155*

Afin de déterminer si les informations recueillies (genre et langue maternelle) sur les auditeurs présentent un effet sur la performance, nous utilisons le test U de Mann-Whitney (Mann et Whitney, 1947) et le test H de Kruskal-Wallis (Kruskal et Wallis, 1952). Ces deux tests évaluent l'hypothèse selon laquelle des ensembles de données appartiennent à la même distribution. Le test U de Mann-Whitney s'applique sur deux ensembles de données et le test H de Kruskal-Wallis s'applique sur au moins trois ensembles de données.

Genre des auditeurs Le test U de Mann-Whitney indique qu'il n'existe pas d'effet significatif du genre des auditeurs sur la performance ($\chi^2 = 152960$, $p = 0.24$). La figure 5.7 montre la répartition de la pureté du regroupement en fonction du genre des auditeurs.

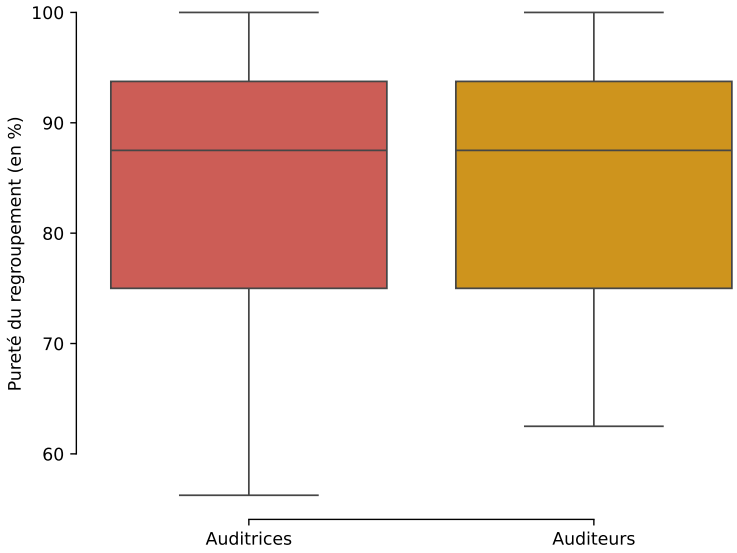


Figure 5.7 – *Répartition de la pureté du regroupement en fonction des auditeurs*

Langue maternelle Selon le test U, la langue maternelle des auditeurs a un effet sur la performance ($\chi^2 = 153216$ ($p < 0.0001$)). En effet, les auditeurs natifs du français ont particulièrement ont effectué la tâche avec une pureté du regroupement moyenne à 89.02 % ($\delta = 12.46$). Les auditeurs natifs de l'anglais se montrent un peu moins performants avec une pureté du regroupement moyenne de 82.97 % ($\delta = 11.25$). La figure 5.8 montre la répartition de la pureté du regroupement en fonction de la langue maternelle. Les locuteurs des autres langues ont une performance moyenne de 86.72 % ($\delta = 7.31$). Ces résultats sont surprenants

puisque les stimuli sont en langue anglaise. Cependant, nous pouvons soulever l'hypothèse que les locuteurs non natifs font abstraction du contenu linguistique pour se concentrer sur la voix.

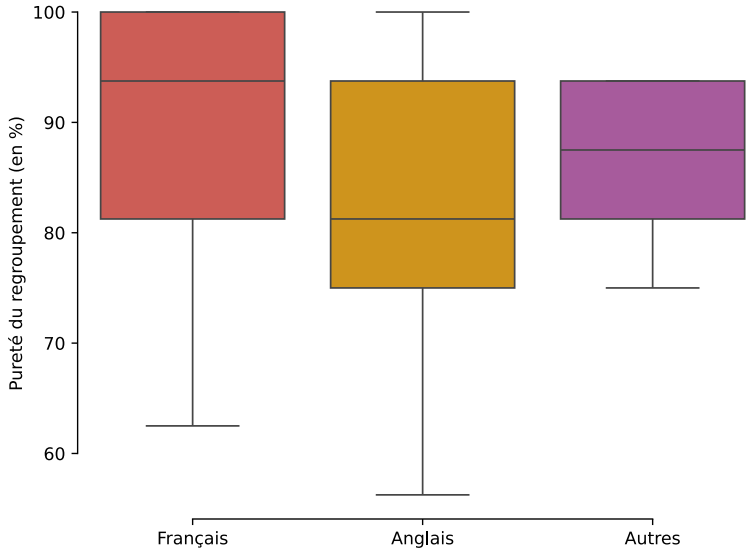


Figure 5.8 – *Répartition de la pureté du regroupement en fonction de la langue maternelle des auditeurs*

5.3 Le système surpasse la performance humaine

Comparativement à la performance humaine, déjà très bonne, le système utilisé se montre plus performant encore. Le tableau 5.4 montre qu'en termes de pureté du regroupement, la performance moyenne est égale à 97.47 % ($\delta = 0.32$). Le nombre de locuteurs par groupe est légèrement inférieur (1.11) et de manière générale, le nombre de groupes est légèrement inférieur au regroupement humain.

Le genre des locuteurs a un effet sur la performance du système ($\chi^2 = 112640$, $p < 0.001$). Les matrices de confusion présentées dans les figures 5.9 and 5.10 présentent les regroupements automatiques pour locuteurs. L'expérience entière montre que 4 stimuli sont mal classés pour les locuteurs et 5 le sont pour les locutrices.

	Humains		Système	
	μ	δ	μ	δ
Groupes	3.79	0.48	3.75	0.43
Loc. par groupe	1.33	0.55	1.11	0.32
Pureté du regr.	86.40 %	11.87	97.47 %	3.61
Pureté	89.70 %	10.35	97.64 %	3.09

Tableau 5.4 – *Comparaison de la performance entre humains et le système automatique*

Figure 5.9 – *Matrice de confusion des locutrices après regroupement automatique*

5.4 Discussion et conclusion

Dans le contexte criminalistique, il peut être nécessaire d'identifier le nombre de locuteurs parmi un ensemble d'enregistrements. Pour étudier cela, nous avons mis en place d'une tâche perceptive de regroupement d'enregistrements en locuteurs que nous avons introduite dans le cadre du challenge VoicePrivacy. Cette expérience perceptive vise à évaluer la capacité humaine à identifier et regrouper les locuteurs malgré l'anonymisation de certains enregistrements. 75 auditeurs, de langue maternelle française, anglaise ou autre,

Figure 5.10 – *Matrice de confusion des locuteurs après regroupement automatique*

sont soumis à 3 tests : 1 test de contrôle qui contient uniquement des stimuli originaux et 2 tests d'évaluation qui contiennent 50 % de stimuli anonymisés. Le test de contrôle permet d'obtenir une base de référence pour les tests d'évaluation. Ce chapitre ne contient que les résultats sur les tests de contrôle en vue d'étudier la perception humaine des locuteurs. Afin de répondre aux consignes de l'expérience, nous avons défini la pureté du regroupement pour mesurer la performance des auditeurs. Similaire à la pureté, elle ajoute cependant deux règles strictes : un groupe ne peut contenir qu'une seule étiquette et une étiquette ne peut être attribuée qu'à un seul groupe.

Les résultats de l'expérience montrent l'intérêt d'une tâche de regroupement en locuteurs, car elle permet aux auditeurs d'identifier des locuteurs tout en s'affranchissant d'éventuels biais cognitifs rencontrés sur les comparaisons binaires. L'expérience permet également de mieux comprendre la perception humaine du locuteur. En effet, les résultats montrent que la performance des auditeurs varie entre 0.5625 et 1, ce qui indique que la tâche n'est pas aisée pour tous. Plusieurs facteurs peuvent expliquer ces résultats. D'une part, certaines confusions sont observées, surtout entre les locuteurs p258 et p259. Parmi les auditeurs ayant eu une pureté du regroupement en deçà de 0.75, les locuteurs p258 et p259 sont systématiquement retrouvés. Du côté des auditrices, la locutrice p240 est re-

trouvée dans 4 des 5 regroupements présentant une pureté du regroupement la plus basse, confondue avec p250 ou les distractrices.

Nous avons également mis en évidence que la langue maternelle des auditeurs était un facteur déterminant. Elle a un effet significatif sur la performance des auditeurs en défaveur des auditeurs natifs de l'anglais, ce qui apparaît étonnant puisque les stimuli sont en anglais. Nous émettons l'hypothèse que les personnes non natives font abstraction du contenu linguistique pour expliquer cet effet. Cependant, cela reste à démontrer.

Enfin, nous avons comparé la performance humaine avec celle d'un système automatique basé sur l'approche *x-vector*. Les résultats montrent que le système automatique surpasse la performance humaine avec une pureté du regroupement moyenne de 97.47 % contre 86.40 % pour les humains. Moins d'une dizaine de stimuli ont été mal catégorisés. Nous expliquons cela par le peu de variabilité des stimuli dont les conditions d'enregistrement sont identiques, mais aussi par le fait que la tâche n'est pas tout à fait la même entre humains et système. En effet, le système automatique n'écoute pas les enregistrements à proprement parler, mais se base sur des représentations de ces derniers, ce qui peut expliquer la différence de performance. De plus, les stimuli contiennent des enregistrements de phrases lues, ce qui contraint grandement la voix et offre moins de variabilité que la parole spontanée. Enfin, nous proposons d'augmenter la variabilité des enregistrements pour rendre l'expérience plus intéressante aussi bien dans la perception humaine que dans le regroupement automatique.

> **Regrouper des enregistrements en locuteurs n'équivaut pas à comparer des voix en criminalistique**
>
> La tâche de regroupement d'enregistrements en locuteurs présentée dans ce chapitre a été réalisée dans des conditions contrôlées et idéales pour assurer une performance optimale. En effet, les stimuli sont issus d'un corpus de parole contrôlé et les enregistrements sont de bonne qualité, seule la durée des enregistrements a été réduite pour limiter celle de l'expérience globale. Dans le cas d'une comparaison de voix en criminalistique, les conditions d'enregistrement sont rarement contrôlées pour l'ensemble

des fichiers à comparer. De ce fait, les résultats présentés dans ce chapitre ne sont pas généralisables à une comparaison de voix en criminalistique.

Ce travail a fait l'objet d'un article présenté à la conférence internationale *Interspeech 2021* à Brno en République tchèque (O'Brien *et al.*, 2021).

6 | Vers la caractérisation de la voix : le cas du type de phonation

Résumé : *La caractérisation des voix est un enjeu important en criminalistique, car elle permet d'ajouter de la finesse dans les processus de comparaison de voix. Pour aborder cette problématique, nous nous intéressons au type de phonation qui peut apporter des indices sur l'identité des locuteurs du fait d'un usage différentiel entre les locuteurs. L'objectif est de pouvoir détecter automatiquement le type de phonation global des locuteurs en vue d'indiquer s'ils ont plutôt une voix modale, craquée ou soufflée. Pour ce faire, nous proposons une architecture neuronale composée de PASE pour l'extraction des paramètres et d'un MLP pour la classification. Pour cela, nous mettons en place deux classifieurs binaires que nous utilisons en cascade : le premier classifieur distingue les trames modales des trames non modales, et le second classifieur distingue les trames craquées des trames soufflées. Dans un premier temps, nous évaluons la performance de notre système sur les voyelles prépausales de PTSVOX. Notre nouveau système parvient à reconnaître les trames modales à 86 %, les trames non modales à 76 %, les trames craquées à 85 % et les trames soufflées à 88 %, ce qui est supérieur au système de référence dans les quatre cas. Forts de ces résultats, nous avons généralisé notre système sur l'ensemble des phonèmes voisés de PTSVOX. Nos résultats ont permis de mettre en évidence des usages différents du type de phonation entre les locuteurs. En effet, une première différence est observée entre les hommes et les femmes, les hommes ayant une tendance à utiliser la voix non modale et les femmes la voix modale avec*

une hétérogénéité entre les locuteurs. Cela confirme l'intérêt de la caractérisation de la voix dans le cadre de la comparaison de voix.

Sommaire

Podesva définit la qualité de voix comme étant des « propriétés extra-grammaticales et suprasegmentales de la parole, inhérentes à la configuration de l'appareil phonatoire » (Podesva, 2007). Elle regroupe plusieurs caractéristiques telles que la nasalité, la clarté, le relâchement ou encore le type de phonation. La qualité de voix est caractérisée par la physiologie du locuteur d'une part, et par son contexte sociolinguistique d'autre part.

Le type de phonation est un descripteur de qualité de voix qui se traduit par un continuum articulatoire basé sur l'aperture des plis vocaux, allant d'une glotte fermée pour la voix craquée à une glotte ouverte pour la voix soufflée (Gordon et Ladefoged, 2001). Le type de phonation normal, dit modal, se trouve entre la voix craquée et la voix soufflée. Les types de phonation non modaux entraînent des perturbations de la F_0.

Des variations de qualité de voix sont observées dans plusieurs langues combinant type de phonation et nasalité (DiCanio, 2009 ; Garellek et Keating, 2011 ; Gordon, 2001 ; Gordon et Ladefoged, 2001). Par exemple, le système vocalique du mazatèque de Jalapa[1] repose sur les variations de la qualité de voix de 5 voyelles différentes : [i], [æ], [a], [o] et [u] (Silverman *et al.*, 1995). En français, l'utilisation de la voix soufflée peut permettre de marquer une fin prosodique (Smith, 1999).

Les études sur la qualité de voix de manière générale, et sur le type de phonation plus particulièrement reposent sur la perception des phonéticiens en raison des recherches assez peu développées sur la mesure des paramètres pertinents. Par conséquent, il apparaît nécessaire de mettre en place un système qui puisse détecter automatiquement les variations du type de phonation dans la voix des locuteurs.

6.1 Architecture proposée : PASE-MLP

6.1.1 Extraction des paramètres avec PASE

Le système utilisé pour la détection du type de phonation est composé de deux étapes : la paramétrisation et la classification. Dans un premier temps, le système doit se montrer

1. Langue amérindienne parlée dans les états de Veracrux et Oaxaca, au Mexique

performant sur un jeu de données annotées de taille restreinte, puisque l'annotation manuelle est une tâche longue et fastidieuse. La généralisation sur un plus grand ensemble de données non annotées est également un critère important pour permettre au système d'acquérir de nouvelles connaissances. Par conséquent, l'extracteur de paramètres doit être capable de traiter plusieurs paramètres à la fois et de s'adapter à de nouveaux paramètres afin de déterminer les plus pertinents pour la tâche. Pour cette raison, l'utilisation de PASE paraît adaptée.

PASE est un auto-encodeur qui permet d'obtenir une représentation unique d'un signal audio à partir de plusieurs paramètres (Pascual *et al.*, 2019). Appliqué au traitement du signal, un auto-encodeur est un réseau de neurones qui permet d'encoder un signal dans un espace latent de dimension réduite, puis de le décoder pour retrouver le signal d'origine. L'entraînement de PASE est à visée multitâches : l'encodeur est optimisé pour atteindre différents objectifs, matérialisés par un ensemble de *workers*, tels que la récupération de la forme d'onde, des coefficients cepstraux, des phonèmes ou encore de certains paramètres prosodiques. Ainsi, la représentation vectorielle est une information compressée capable de satisfaire chaque *worker*. Les représentations délivrées par PASE obtiennent de meilleures performances dans l'identification des locuteurs, la détection d'émotions ainsi que la reconnaissance de la parole par rapport à l'utilisation des MFCC seuls et des banques-filtres seules (Pascual *et al.*, 2019). Il est possible d'orienter l'apprentissage du modèle vers une tâche précise en *fine-tunant* l'encodeur (Pascual *et al.*, 2019). L'architecture neuronale de PASE est présentée dans la figure 6.1.

Une version évoluée, intitulée PASE+, permet d'ajouter davantage de paramètres dans la représentation des signaux dont les dérivées et dérivées secondes (Δ et ΔΔ) (Ravanelli *et al.*, 2020). Les contextes de la fenêtre des paramètres de PASE+ sont plus larges que PASE (7 trames à la place de 5), et il est également possible d'élargir la taille des fenêtres à 200 millisecondes (à la place de 25) pour le spectre de puissance, les MFCC, les banques de filtres et les gammatones. Le tableau 6.1 résume les paramètres que PASE et PASE+ peuvent extraire.

Le résultat de l'extraction des paramètres est une représentation vectorielle du signal, de 256 dimensions, qui regroupe l'ensemble des informations fournies par les différents paramètres.

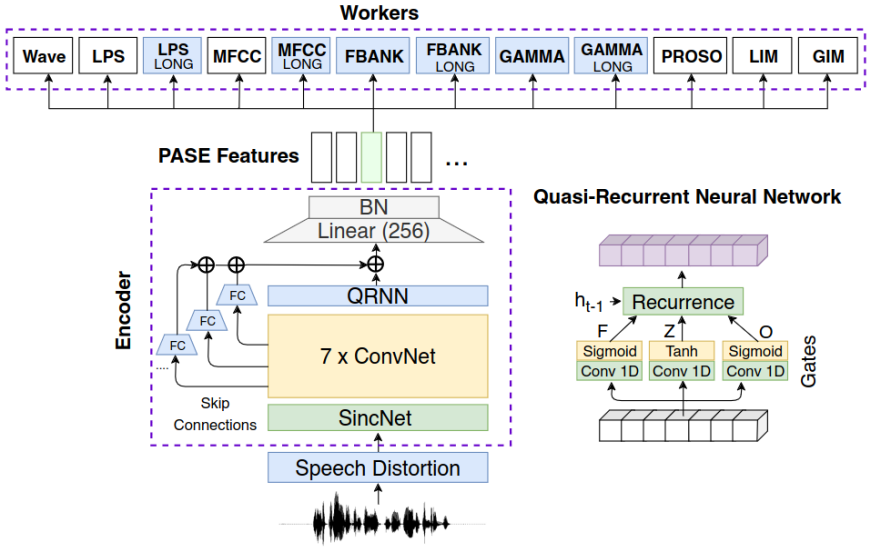


Figure 6.1 – *Architecture de PASE et PASE+ (Ravanelli et al., 2020)*

Paramètre	**PASE**	**PASE+**
Forme d'onde	✓	✓
Spectre de puissance	✓	✓
MFCC	✓	✓
Banque de filtres		✓
Gammatone		✓
Prosodie	✓	✓
Information maximale locale	✓	✓
Information maximale globale	✓	✓

Tableau 6.1 – *Liste des paramètres que peuvent extraire PASE et PASE+*

Pour une tâche de détection du type de phonation, l'autoencodeur PASE+ est utilisé avec les *workers* suivants : forme d'onde, MFCC, prosodie, information maximale locale et information maximale globale. L'avantage de PASE+ est qu'il permet d'ajouter ou de sélectionner d'autres *workers* sans avoir à réentraîner le modèle.

L'entraînement de PASE est effectué sur 6609 enregistrements de la base de données PTSVOX (Chanclu *et al.*, 2020). Les représentations vectorielles délivrées par PASE+ ont une taille de 256 dimensions, sur des trames du 10 millisecondes.

6.1.2 Classification en cascade avec un MLP

La classification consiste à utiliser deux classifieurs binaires en cascade. Le premier classifieur assure la détection des trames modales et non modales (craquées ou soufflées). Le second classifieur s'intéresse à la catégorisation des trames craquées et soufflées. Les deux classifieurs sont identiques : il s'agit d'un MLP qui possède une couche cachée de 256 dimensions avec une fonction d'activation *Rectified linear unit* (ReLU) et un *dropout* de 0.15. La couche de sortie est une couche *Softmax*. L'apprentissage du modèle s'effectue sur 150 époques avec un taux d'apprentissage fixé à 0.0001 et une taille de lot de 50 trames. Pour chacun des classifieurs, une sélection aléatoire des données est faite sur la classe majoritaire de sorte qu'elle possède le même nombre de trames que la classe minoritaire, soit 33 554 trames modales et non modales pour le premier classifieur, et 16 113 trames craquées et soufflées pour le second classifieur ainsi que le montre le tableau 6.3.

Prise de décision Les classifieurs délivrent un score qui conditionne la prise de décision : si le score est supérieur à 0.5, le classifieur répond « oui » ; dans le cas contraire, il répond « non ». La performance au niveau des segments est également calculée en moyennant sur les scores des sur les trames du segment, en suivant la même règle de décision.

6.1.3 Système de référence : MFCC-SVM

L'architecture PASE-MLP est comparée avec un modèle de référence basé sur les MFCC pour la paramétrisation et une SVM pour la classification. La paramétrisation du système de référence comprend 30 coefficients MFCC sur une bande passante allant de 0 à 8000 Hz. La taille de la fenêtre est de 20 millisecondes avec un recouvrement de 10 millisecondes entre deux fenêtres consécutives. La classification est opérée par un SVM avec une fonction de base radiale du noyau gaussien. À l'instar de l'apprentissage du MLP, les données de la classe majoritaire sont sélectionnées aléatoirement pour correspondre à la quantité de données de la classe minoritaire et les règles de prise de décision sont identiques.

6.2 Détection en milieu fermé sur les voyelles prépausales

6.2.1 Annotation manuelle en type de phonation

Les voyelles prépausales sont particulièrement intéressantes en français, car elles occupent une position majeure dans la courbe prosodique et elles sont essentielles dans la perception de la communication parlée. Pour cette expérience, les enregistrements microphoniques du jeu de données *Intra* de la base de données PTSVOX, ré-échantillonnés à 16000 Hz, sont utilisés. Les voyelles précédant directement une pause minimale de 50 millisecondes sont annotées aussi bien à l'oreille qu'en visualisant les spectrogrammes. L'ensemble des voyelles prépausales ont été annotées par un phonéticien en quatre catégories : craqué, soufflé, modal, autre. Une voyelle peut posséder une ou plusieurs étiquettes suivant l'évolution du type de phonation. Pour des raisons de clarté, le terme segment est utilisé pour désigner une étiquette à l'intérieur de la voyelle. La catégorie *autre* regroupe les cas ambigus ou bruités. Pour ne pas apporter de biais à l'expérience, cette catégorie est exclue de l'expérience. Le tableau 6.2 montre la répartition des classes après annotation. Au total, 8909 voyelles ont été annotées en 10360 segments. Le corpus final comprend 8459 voyelles prépausales dont 704 contiennent 2 étiquettes, soit un total de 9867 segments.

6.2.2 Composition du corpus

Le set *Intra* de PTSVOX est composé de 12 locutrices et 12 locuteurs. Deux locutrices et deux locuteurs ont été sélectionnés aléatoirement pour former les données de test. Les locuteurs sélectionnés sont `LG001` et `LG009` pour les femmes et `LG012` et `LG024` pour les hommes. Parmi les autres locuteurs, 80 % des données sont dédiées à l'apprentissage des classifieurs, et les 20 % restants à la validation. Le tableau 6.3 indique le nombre de segments et leur durée en trames de 10 millisecondes.

Voyelle	Lecture	Spontané	Total
craqué	245	661	906
modal	1280	3771	5051
soufflé	189	862	1051
autre	80	370	450
craqué + modal	8	27	35
craqué + soufflé	5	44	49
craqué + autre	0	5	5
modal + craqué	223	553	776
modal + soufflé	87	405	492
modal + autre	7	16	23
soufflé + modal	0	4	4
soufflé + craqué	1	16	17
soufflé + autre	0	3	3
autre + modal	0	1	1
autre + craqué	1	33	34
autre + soufflé	1	11	12
Total	**2127**	**6782**	**8909**

Tableau 6.2 – *Annotation en type de phonation des voyelles prépausales*

	Entraînement + Validation		Test	
	Segments	Trames	Segments	Trames
modal	4 090	89 478	687	15 773
non modal	2 220	33 554	464	7 410
craqué	1 088	17 441	251	3 569
soufflé	1 132	16 113	213	3 841

Tableau 6.3 – *Quantité et durée des segments (par trames de 0.01 seconde) pour les données d'entraînement et de tests*

6.2.3 Une performance très satisfaisante

Le tableau 6.4 présente la performance de chacun des systèmes en termes de la classification des trames modales et non modales. Le système PASE-MLP a obtenu 86 % de bonne

	PASE-MLP		MFCC-SVM		Total
	Modal	Non Modal	Modal	Non Modal	
Modal	13 640 86 %	2 133 14 %	12 430 79 %	3 343 21 %	15 773
Non Modal	1 768 24 %	5 642 76 %	2 076 28 %	5 334 72 %	7 410

Tableau 6.4 – *Matrice de confusion entre les trames modales et non modales pour les deux systèmes*

classification sur les trames modales et 76 % de classification sur les trames non modales. La performance du système de référence est moindre avec 79 % de bonne catégorisation pour les trames modales et 72 % sur les trames non modales.

Le tableau 6.5 présente la matrice de confusion des deux systèmes concernant la classification des segments modaux et non modaux. La classification des segments modaux et non modaux est plus performante que la classification des trames modales et non modales et ce pour les deux systèmes. En effet, le système PASE-MLP obtient 93 % de bonne classification sur les segments modaux et 83 % sur les segments non modaux. Le système de référence obtient 87 % de bonne classification sur les segments modaux et 78 % sur les segments non modaux. La différence de performance entre les trames et les segments s'explique par le fait que la prise de décision est effectuée sur la moyenne des scores des trames du segment. Ainsi, un segment peut être classifié comme modal alors qu'il contient des trames non modales.

	PASE-MLP		MFCC-SVM		Total
	Modal	Non Modal	Modal	Non Modal	
Modal	638 93 %	49 7 %	595 87 %	92 13 %	687
Non Modal	80 17 %	384 83 %	102 22 %	362 78 %	464

Tableau 6.5 – *Matrice de confusion entre les segments modaux et non modaux pour les deux systèmes*

En ce qui concerne la classification des trames craquées et soufflées, le système PASE-MLP montre également une meilleure performance que MFCC-SVM. Alors que le système de référence reconnaît les trames soufflées avec 85 % de bonne reconnaissance, la performance est moindre pour les trames craquées (67 % de bonne reconnaissance). Le système PASE-MLP a un taux de reconnaissance de 85 % pour les trames craquées et 88 % pour les trames soufflées. Au niveau du segment, le système PASE-MLP montre un taux de classification moyen de 94 % et le système de référence MFCC-SVM de 77 %.

	PASE-MLP		**MFCC-SVM**		
	Craqué	**Soufflé**	**Craqué**	**Soufflé**	**Total**
Craqué	3 048 85 %	521 15 %	2 376 67 %	1 193 33 %	3 569
Soufflé	465 12 %	3 376 88 %	586 15 %	3 255 85 %	3 841

Tableau 6.6 – *Matrice de confusion entre les trames craquées et soufflées pour les deux systèmes*

	PASE-MLP		**MFCC-SVM**		
	Craqué	**Soufflé**	**Craqué**	**Soufflé**	**Total**
Craqué	235 94 %	16 6 %	193 77 %	58 23 %	251
Soufflé	11 5 %	202 95 %	19 9 %	194 91 %	213

Tableau 6.7 – *Matrice de confusion entre les segments craqués et soufflés pour les deux systèmes*

6.3 Généralisation sur l'ensemble des voyelles

Au vu des résultats encourageants de l'architecture PASE-MLP sur les voyelles prépausales, le système est généralisé sur l'ensemble des phonèmes voisés des locuteurs du set *Intra* avec l'objectif de dégager des groupes de voix. Ainsi, nous souhaitons pouvoir regrou-

Locuteur	Voy. orales	Voy. nasales	Plosives	Fricatives	Liquides	Nasales	Glides	Total
LG001	55 623	11 483	6 685	6 860	12 015	6 076	3 203	135 110
LG002	86 339	17 542	8 805	6 450	13 354	9 858	4 105	187 096
LG003	53 545	12 830	6 188	4 567	10 098	5 952	2 656	124 070
LG004	10 904	2 140	1 787	1 482	3 710	1 245	844	30 491
LG005	12 016	2 243	1 729	1 484	3 352	1 231	912	30 566
LG006	5 438	978	942	698	1 562	613	397	14 892
LG007	11 296	2 082	1 593	1 189	3 197	1 150	880	28 344
LG008	58 930	12 905	6 982	5 879	12 225	6 686	3 098	141 336
LG009	58 589	10 277	6 470	4 804	10 156	6 284	2 885	127 969
LG010	59 698	12 819	6 975	5 517	12 168	8 829	3 143	145 439
LG011	65 381	12 837	7 051	5 487	12 501	9 009	3 974	150 229
LG012	64 328	16 436	7 240	6 426	10 080	8 277	3 249	153 523
LG013	51 235	12 096	7 616	5 609	12 298	6 945	3 246	134 212
LG014	58 884	13 058	7 762	6 054	12 971	7 267	3 475	145 284
LG015	64 475	13 384	7 449	6 288	14 514	7 388	3 533	154 129
LG016	70 043	18 075	8 528	6 381	13 210	10 409	4 085	168 536
LG017	39 530	7 345	6 699	5 188	10 772	6 120	2 670	106 508
LG018	39 608	7 925	4 914	4 315	7 655	5 080	2 104	92 692
LG019	45 009	8 719	5 969	5 146	9 877	5 451	2 565	111 853
LG020	31 083	5 978	4 178	3 051	6 759	3 977	2 128	77 471
LG021	29 660	6 740	4 964	3 086	7 618	4 799	2 350	80 823
LG022	28 226	5 338	3 213	2 401	5 569	3 308	1 726	65 430
LG023	50 676	10 698	6 013	5 516	10 036	5 661	2 571	121 386
LG024	51 124	8 796	6 563	5 816	9 901	6 088	2 873	122 935

Tableau 6.8 – *Nombre de trames pour les phonèmes voisés de chaque locuteur du set* Intra *de PTSVOX*

per les locuteurs en fonction de leur type de phonation global. Le tableau 6.8 présente le nombre de trames pour les phonèmes voisés par locuteur, soit un total de 2 650 324 trames.

Les prédictions fournies par le système permettent de calculer pour chaque locuteur le pourcentage de trames modales, non modales, craquées et soufflées. Ces informations permettent d'estimer une première étape dans la caractérisation de la voix. La figure 6.2 montre la répartition des locuteurs à partir des prédictions de l'architecture PASE-MLP.

Sur l'ensemble des voyelles du set *Intra* de PTSVOX, nous pouvons remarquer une tendance différente entre hommes et femmes : là où les locutrices adoptent plutôt une voix modale, les locuteurs adoptent plutôt une voix non modale, aussi bien craquée que soufflée. Une exception est la locutrice LG006 qui possède une voix craquée. Ainsi, 8 locuteurs sur 12 possèdent une voix non modale : 3 ont une voix soufflée (dont 1 à la voix très soufflée), 4 ont une voix craquée (1 à la voix très craquée) et 1 utilise autant la voix soufflée que la voix craquée. Du côté des femmes, seule la locutrice LG006 possède une voix craquée qui peut être expliquée par son tabagisme. Les autres locutrices possèdent une voix modale

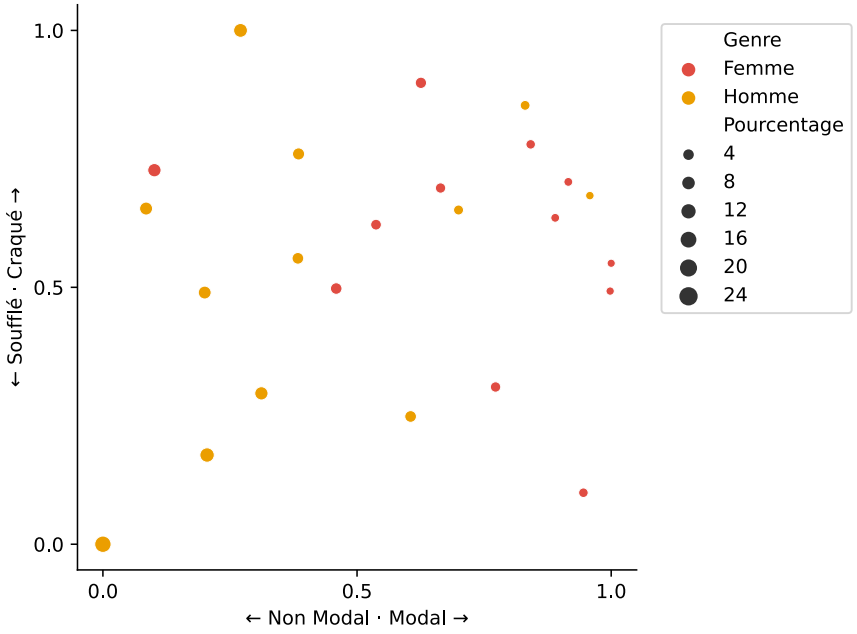


Figure 6.2 – *Répartition des locuteurs du set* Intra *en fonction du type de phonation*
Le pourcentage correspond au ratio entre le nombre de trames non modales de la classe majoritaire et le nombre total de trames non modales

avec majoritairement une voix craquée pour les trames non modales à l'exception de deux locutrices qui possèdent une voix soufflée dans les phonèmes non modaux.

6.3.1 Limites de la solution proposée

Cette expérience met en avant les limites du processus aussi bien au niveau de l'annotation que du système.

Annotation L'annotation a été faite par un seul expert. Bien que spécialiste, le manque d'annotateurs multiples ne permet pas de calculer l'accord interannotateur (kappa de Cohen). Ainsi, certains segments ont pu être mal étiquetés, notamment du fait de la qualité de voix globale du locuteur (tendance à surannoter une voix craquée ou soufflée par exemple).

Apprentissage sur un corpus restreint L'apprentissage du modèle a été effectué sur les voyelles prépausales qui, certes, contiennent beaucoup d'informations du fait de leur position prosodique, mais qui ne sont pas représentatives de l'ensemble des phonèmes voisés.

Ces limites s'observent par exemple avec le locuteur LG017 dont la quasi-totalité des voyelles sont considérées comme étant soufflées. Les annotations manuelles de ce locuteur ne montrent pas une surreprésentation des voyelles soufflées.

6.4 Discussion et conclusion

Dans un contexte criminalistique, une caractérisation fine des voix n'est pas effectuée à ce jour. Cependant, la caractérisation des voix peut apporter des informations sur le locuteur et permettre des comparaisons entre des voix qui possèdent des caractéristiques similaires. Dans cette optique, le travail présenté dans ce chapitre vise à étudier différents types de phonation : voix modale et voix non modale (craquée et soufflée). Certains locuteurs font un usage préférentiel d'un des types de phonation non modale, ce qui pourrait porter des indices quant à leur identité.

Dans un premier temps, nous étudions les voyelles prépausales, très intéressantes en français, car elles occupent une position majeure dans la courbe prosodique et elles sont essentielles dans la perception de la communication parlée. Pour cela, nous avons annoté les voyelles d'intérêt dans les enregistrements au microphone de parole spontanée du jeu de données *Intra* de la base de données PTSVOX en quatre étiquettes : craqué, soufflé, modal et autre. Cette annotation a été effectuée par un phonéticien. La catégorie *autre* a été écartée afin d'éviter les biais.

Nous proposons une nouvelle architecture neuronale pour parvenir à cette tâche. Cette nouvelle architecture est composée de l'extracteur de caractéristiques PASE et d'un MLP pour la classification, profitant de la souplesse de PASE pour ajouter ou retirer des *workers* sans avoir à ré-entraîner le modèle. Cette nouvelle architecture est comparée avec un système de référence basé sur les MFCC et un SVM. Nous avons entraîné et validé les deux systèmes sur une partie des voyelles prépausales de PTSVOX de 20 locuteurs et testé les voyelles pré-

pausales des 4 locuteurs restants. Pour les deux systèmes, nous effectuons la classification aussi bien au niveau de la trame que du segment (en moyennant les scores des trames du segment). Dans un premier temps, les trames modales et non modales sont catégorisées, puis les trames non modales sont classifiées en trames craquées et soufflées.

Nos résultats montrent que l'architecture PASE-MLP est plus performante que le système de référence MFCC-SVM. En effet, pour PASE-MLP, la performance atteint 86 % pour les trames modales et 88 % pour les trames soufflées. Au niveau du segment, la performance atteint 93 % pour les segments modaux et jusqu'à 95 % pour les segments soufflés. Ces résultats montrent le potentiel de l'architecture en cascade, particulièrement robuste pour la détection des segments craquées et soufflées. La robustesse de l'architecture en cascade est à confirmer sur des ensembles de données plus grands et plus variés, appliquée à d'autres paramètres.

Ces résultats prometteurs nous ont permis de généraliser notre protocole à l'ensemble des phonèmes voisés – consonnes comme voyelles – des locuteurs du set *Intra* de PTSVOX. Cette généralisation vise à caractériser les locuteurs en fonction de leur type de phonation global. Pour cela, nous avons utilisé le système PASE-MLP, appris sur les voyelles prépausales, pour prédire le pourcentage de trames modales, non modales, craquées et soufflées pour chaque locuteur. Nos résultats montrent qu'il est possible de caractériser les locuteurs en fonction de leur type de phonation global. En effet, nous avons pu observer une hétérogénéité entre les locuteurs dans l'usage du type de phonation, avec des tendances différentes entre femmes et hommes : là où les locutrices adoptent plutôt une voix modale, les locuteurs adoptent plutôt une voix non modale. Les locuteurs étudiés montrent un usage différent du type de phonation : l'usage de la voix non modale est présent chez 10 locuteurs (dont 8 hommes) et se répartit de manière disparate entre une voix très soufflée et une voix très craquée. Cependant, ces résultats sont à nuancer du fait de l'annotation monoannotateur et du corpus d'apprentissage restreint. Ce travail permet d'ouvrir des perspectives pour la caractérisation de la voix et une éventuelle application de ces nouvelles connaissances en criminalistique. Toutefois, il est à confirmer par une annotation plus stricte, effectuée par plusieurs annotateurs afin de calculer l'accord interannotateur. Au niveau du modèle, il est également nécessaire de le renforcer avec des données plus variées et d'origines diverses.

Cette expérience est à compléter avec d'autres études sur la caractérisation de la voix, en l'étendant à d'autres paramètres tels que la nasalité, amorcé par (Kim et Gendrot, 2022 ; Kim *et al.*, 2023) tout en étudiant la variabilité intralocuteur de chacun des paramètres.

Cette étude a fait l'objet d'un article présenté à la conférence internationale *Interspeech 2021* à Brno en République tchèque (Chanclu *et al.*, 2021).

Troisième partie

Conclusions et perspectives

Apports de la thèse

Dans ce manuscrit de thèse, nous nous sommes intéressés aux différents aspects de la reconnaissance des locuteurs et au cas particulier de la comparaison de voix en criminalistique. L'un des objectifs du projet Voxcrim, dans lequel ce travail de thèse s'inscrit, est de mettre en place un protocole d'accréditation pour les laboratoires de comparaison de voix. Pour cela, il est nécessaire de définir un cadre scientifique pour garantir la fiabilité des résultats d'une comparaison de voix dans le cadre criminalistique. À cet effet, les bases de données PTSVOX et FABIOLE 2 ont été mises en place. Nous avons également introduit le concept de *box-rule* pour désigner les conditions dans lesquelles la fiabilité des résultats d'une comparaison de voix est connue.

En outre, nous avons mis en place une expérience perceptive de regroupement de voix en locuteurs dans le cadre du challenge VoicePrivacy en vue d'étudier la capacité humaine à percevoir les locuteurs, pour limiter les biais cognitifs que peuvent entraîner les comparaisons binaires. Pour cela, nous avons introduit une métrique, la pureté du regroupement, qui permet d'évaluer la performance d'un système de regroupement de voix en locuteurs en maximisant la précision des auditeurs.

Enfin, nous avons amorcé un travail de caractérisation de la voix par la détection du type de phonation avec l'objectif de caractériser les voix automatiquement. À cet effet, nous avons proposé un nouveau système ainsi qu'une architecture en cascade pour effectuer deux classifications binaires.

De l'importance des données pour la comparaison de voix

Aujourd'hui, la base de données Voxceleb est très largement utilisée en RAL en raison de son très grand nombre de locuteurs de différentes nationalités. Alors que Voxceleb est présentée par ses auteurs comme étant très variée, la base fait preuve de peu de diversité. En effet, les locuteurs sont majoritairement de sexe masculin et de nationalité américaine, ce qui tend à favoriser leur reconnaissance au détriment des femmes et des autres nationalités. Bien que Voxceleb soit une base de données très intéressante pour la RAL, elle montre de très grandes lacunes pour la modélisation de la variabilité intralocuteur, ce qui entraîne des erreurs.

La base de données FABIOLE 2, conçue dans le cadre du projet Voxcrim, est spécialement dédiée à l'étude de la variabilité intralocuteur. C'est la raison pour laquelle elle compte un plus petit nombre de locuteurs que Voxceleb – 399 contre 7363 – mais un nombre d'enregistrements par locuteur bien plus important s'étendant sur plusieurs années. Cependant, nous montrons que la base de données FABIOLE 2 ne permet pas d'étudier en profondeur la variabilité intralocuteur, car elle ne contient pas suffisamment d'informations sur les contextes d'enregistrement ni le contenu phonétique. Aussi, les locuteurs sont assez peu variés : la plupart d'entre eux sont journalistes ou issus de la classe politique.

Afin de nous rapprocher davantage des conditions réelles de la comparaison de voix, nous avons mis en place la base de données PTSVOX. Également conçue dans le cadre du projet Voxcrim, elle suit les protocoles de prélèvements de voix du SNPS. Elle est composée de 369 locuteurs enregistrés au téléphone et au microphone, 24 d'entre eux ont été enregistrés à plusieurs reprises, aussi bien en parole spontanée qu'en lecture. Ce jeu de données est particulièrement intéressant pour amorcer une étude sur la variabilité intralocuteur dans des conditions qui ne sont pas toujours optimales.

Premiers éléments de la *box-rule*

Nous avons défini les contours de la *box-rule* grâce aux résultats obtenus en étudiant l'influence des différents facteurs sur la performance du système ECAPA-TDNN. En effet, nous montrons que la durée, la différence de durée entre les enregistrements, le genre des locuteurs ainsi que l'âge du locuteur influent sur la performance.

La performance augmente avec la durée des enregistrements jusqu'à atteindre un plateau à partir de 7 secondes d'enregistrement. Néanmoins, nous observons des différences entre locutrices et locuteurs, notamment sur les enregistrements de 10 et 20 secondes où les écarts se creusent et les locuteurs observent une meilleure performance que les locutrices. À 30 secondes, l'écart se resserre et la performance des locutrices est légèrement supérieure à celle des locuteurs.

Nous confirmons que la différence de durée entre les enregistrements a également un effet sur la performance. Plus la différence de durée entre les enregistrements est importante, plus la performance diminue. Cependant, cette diminution n'est pas linéaire. Tant que le rapport de durée entre les enregistrements est inférieur à 3.23, le C_{llr}^{min} est compris entre 0.70 et 0.076. Au-delà, la C_{llr}^{min} augmente ensuite pour atteindre 0.170 pour les rapports de durée supérieurs à 5.64 secondes.

Nous soulignons également que le genre des locuteurs a un effet sur la performance. Comme nous l'avons vu précédemment, il existe une différence de performance entre les locutrices et les locuteurs sur les durées de 10 et 20 secondes. Ces résultats ne peuvent pas être expliqués à partir des informations sur les locuteurs. Aucune tendance n'a pu être dessinée sur les locutrices et les locuteurs en fonction de leur âge, de leur nationalité ou encore de leur profession.

La performance est aussi influencée par l'âge des locuteurs. Ceux dont l'âge est compris entre 50 et 54 ans sont mieux reconnus que les autres. En revanche, ce n'est pas le cas des quadragénaires, qui observe la performance la plus faible. Ni leur nationalité ou leur profession n'ont permis d'expliquer ces résultats aussi bien sur les quadragénaires que sur les locuteurs âgés de 50 à 54 ans.

Enfin, l'écart temporel entre les enregistrements à comparer influence également la performance. Ce paramètre a été étudié sur les comparaisons cibles uniquement et montre que plus l'écart temporel entre les enregistrements est important, plus la performance diminue. Ce phénomène est très significatif sur la moitié des 100 locuteurs et la corrélation est particulièrement marquée pour 6 locuteurs.

Nos résultats montrent que la performance d'un système de comparaison de voix est influencée par de nombreux facteurs. Bien que pris individuellement, ces facteurs influencent la performance, cette dernière est le résultat d'une combinaison entre plusieurs facteurs. Il est donc nécessaire de les étudier simultanément et avec d'autres paramètres tels que le contenu phonétique ou encore le style de parole.

La détection du type de phonation montre une possibilité de caractériser les locuteurs automatiquement avec un système aussi robuste que PASE-MLP. Dans le contexte de la *box-rule*, cette expérience permet d'envisager de détecter le type de phonation des fichiers à comparer.

Regrouper les voix en locuteurs

Dans le cadre du challenge VoicePrivacy, nous avons proposé une tâche de regroupement de voix en locuteurs en lieu et place d'une comparaison binaire entre deux voix. Nous avons introduit une métrique, la pureté du regroupement, qui permet d'évaluer la performance d'un système de regroupement de voix en locuteurs en maximisant la précision des auditeurs. Les 74 participants ont obtenu une pureté moyenne de regroupement de 86.40 %, dont la performance individuelle oscille entre 56.25 % et 100 %. Certains locuteurs ont été plus difficiles à regrouper que d'autres et certains ont apporté davantage de confusion que d'autres. Parmi les tests avec la performance la plus faible, ces locuteurs sont souvent retrouvés.

Ce que montre également cette expérience, c'est l'effet de la langue maternelle sur le regroupement des locuteurs. En effet, les personnes de langue maternelle anglaise ont obtenu une pureté moyenne de 82.97 %, tandis que les personnes de langue maternelle française ont

obtenu une pureté moyenne de 89.02 % et celles de langue maternelle autre ont obtenu une pureté moyenne de 86.72 %. Cela pourrait indiquer que le contenu linguistique n'est pas évalué de la même manière entre natifs et non natifs. Cette hypothèse est à vérifier avec des analyses plus poussées.

Cette expérience a été reproduite par un système automatique basé sur l'architecture *x-vector*. Nos résultats montrent que le système automatique obtient une pureté moyenne du regroupement de 97.47 %. En effet, seuls 9 stimuli ont été mal catégorisés. Ces résultats de l'approche automatique peuvent s'expliquer par le fait que les enregistrements utilisés présentent assez peu de variabilité. Il s'agit effectivement d'enregistrements de lecture de textes, qui est un contexte très particulier de parole où les locuteurs adoptent un style de parole très contrôlé, notamment pour des raisons d'intelligibilité.

Caractériser automatiquement les voix

En détectant automatiquement le type de phonation, nous avons montré qu'il est possible d'entamer une démarche de caractérisation des voix. En proposant un nouveau système ainsi qu'une architecture en cascade pour effectuer deux classifications binaires, nous avons rendu possible la détection du type de phonation. Le système proposé, PASE-MLP, comparé au système de référence MFCC-SVM, montre de bien meilleurs résultats aussi bien au niveau de la trame que du segment. L'architecture en cascade, quant à elle, permet de simplifier la classification en évaluant dans un premier temps la présence ou l'absence de la voix modale, puis en évaluant l'aspect soufflé ou craqué de la voix.

Les résultats que nous avons obtenus sur les voyelles prépausales montrent la supériorité du système proposé, mais aussi l'intérêt de la classification en cascade, puisque la précision avoisine les 95 % sur les trames et segments craqués et soufflés. La généralisation du système sur les phonèmes voisés de PTSVOX montre une première étape dans la caractérisation des voix puisque plusieurs tendances ont été observées. En effet, les femmes adoptent une voix modale et les hommes une voix non modale, aussi bien craquée que soufflée. Ces derniers résultats doivent être confirmés par des analyses plus poussées, car le modèle est

entraîné sur un nombre restreint de données qui ont été annotées par un seul annotateur, et peuvent comporter des biais.

Perspectives

Les travaux présentés dans ce manuscrit ont permis d'ouvrir des pistes pour aller plus loin dans le domaine de la comparaison de voix en contexte criminalistique. Cependant, plusieurs questions restent en suspens et des travaux restent à mener pour y répondre. Dans cette partie, nous présentons les perspectives qui découlent des travaux présentés dans ce manuscrit.

Perfectionner la *box-rule*

La plus grosse limite pour la mise en place de la *box-rule* est le manque d'information sur les données, notamment en ce qui concerne le contenu des enregistrements. En effet, les données utilisées ne permettent pas d'étudier un grand nombre de facteurs. De plus, l'essentiel des erreurs de comparaison de voix est rencontré lors des comparaisons cibles et il n'était pas possible d'en analyser les causes, alors que le nombre de données est suffisant.

La *box-rule* est amenée à évoluer au fur et à mesure des avancées scientifiques. Alors qu'aujourd'hui, elle ne comporte qu'un nombre restreint de paramètres, il est nécessaire de l'enrichir avec de nouveaux paramètres. Dans l'idéal, elle devrait être capable de prendre en compte tous les facteurs de variabilité entre les pièces pour déterminer les valeurs pour lesquelles la performance est optimale. Pour évaluer l'influence de la variation de chacun des facteurs, il est nécessaire de pouvoir les annoter. Cette annotation peut être réalisée par des experts ou par des systèmes automatiques, en suivant des protocoles stricts. La caractérisation automatique des voix pourra permettre d'introduire une mesure de distance entre

les voix de deux enregistrements, en prenant en compte un certain nombre de paramètres (telles que le type de phonation, la nasalité, *etc.*), et ainsi indiquer si leur comparaison est pertinente ou non.

Toujours dans l'objectif de compléter la *box-rule*, Ben Amor et Bonastre proposent un cadre pour expliquer le *LR* (Ben Amor et Bonastre, 2022). Pour cela, ils proposent d'annoter le signal de parole en fonction de la présence ou non d'attributs et de calculer un *LR* par attribut.

La base de données idéale pour la comparaison de voix

Les bases de données PTSVOX et FABIOLE 2 ont permis d'ouvrir des pistes pour aller plus loin dans la comparaison de voix dans un contexte criminalistique. Cependant, ces bases de données ne sont pas suffisamment riches pour pouvoir couvrir l'ensemble des facteurs de variabilité.

Dans le cas de FABIOLE 2, le manque d'annotation empêche de pouvoir étudier l'influence de nombreux facteurs, mais aussi de trouver la cause de certaines erreurs. Aussi, la date de naissance, la nationalité et la profession sont le fruit de recherches menées après réception des données, et peuvent par conséquent être erronées. Le manque d'annotation aussi bien phonétique que sur le type de parole ou d'émission empêche de rendre compte de la variabilité réelle pour un locuteur. Ces lacunes ont limité le nombre de facteurs étudiés dans le cadre de cette thèse ainsi que la profondeur de leur analyse.

Pour la base de données PTSVOX, bien qu'il y ait davantage de métadonnées que FABIOLE 2 et que le contenu linguistique soit également annoté, le nombre de locuteurs est très restreint, surtout pour étudier la variabilité intralocuteur. En effet, seuls 24 locuteurs ont été enregistrés à plusieurs reprises, entre 2 et 4, ce qui représente très peu de données pour modéliser ou évaluer la variabilité intralocuteur.

Dans un contexte de comparaison de voix en criminalistique, la base de données idéale devrait être en mesure de prendre en compte tous les facteurs de variabilité entre les pièces afin de les modéliser suffisamment pour limiter les erreurs. De nombreuses informations

sur les locuteurs, les contextes d'enregistrement et le contenu linguistique devraient être disponibles en vue de mieux comprendre l'origine des erreurs de comparaison de voix d'une part, et de pouvoir les résoudre d'autre part. Cela demande néanmoins une très grande quantité de données dont la collecte peut être très coûteuse. Les systèmes d'annotation automatique peuvent également permettre d'annoter à moindre coût des bases de données telles que Voxceleb ou FABIOLE 2 et ainsi pouvoir travailler sur d'autres facteurs pour étudier leur influence sur la performance d'un système de comparaison de voix.

Tableaux et figures

Liste des tableaux

Liste des figures

Glossaire

Acronymes

ADN	Acide désoxyribonucléique
AFCP	Association francophone de la communication parlée
ANR	Agence nationale de la recherche
APJ	Agent de police judiciaire
APJA	Agent de police judiciaire adjoint
ARCEP	Autorité de régulation des communications électroniques, des postes et de la distribution de la presse
ARPTS	Antenne régionale de police technique et scientifique
CNIL	Commission nationale de l'informatique et des libertés
Cofrac	Comité français d'accréditation
CPP	Code de procédure pénale
DCF	Detection Cost Function
DET	*Detection Error Trade-off*
ECAPA	*Emphasized Channel Attention, Propagation and Aggregation*
FAED	Fichier automatisé des empreintes digitales
FNAEG	Fichier national automatisé des empreintes génétiques
GFCP	Groupe francophone de la communication parlée
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markhov Model*
IRCGN	Institut de recherche criminelle de la Gendarmerie nationale
JFA	Joint Factor Analysis
LIA	Laboratoire Informatique d'Avignon
LNE	Laboratoire nationale de métrologie et d'essais

LPL	Laboratoire Parole et Langage
LPP	Laboratoire Phonétique et Phonologie
MFCC	*Mel-frequency cepstral coefficient*
MLP	Perceptron multi-couches
NIST	*National Institute of Standards and Technology*
OPJ	Officier de police judiciaire
PASE	*Problem-Agnostic Speech Encoder*
PNIJ	Plate-forme nationale des interceptions judiciaires
RAL	Reconnaissance automatique du locuteur
ReLU	*Rectified linear unit*
SNPS	Service national de Police scientifique
SRE	*Speaker Recognition Evaluation*
SRPTS	Service régional de police technique et scientifique
SVM	machine à support de vecteur
TDNN	*Time Delay Neural Network*
TLFi	Trésor de la langue française informatisé
UBM	*Universal Background Model*

Symboles

C_{llr}	Coût du rapport de vraisemblance logarithmique
C_{llr}^{min}	Coût du rapport de vraisemblance logarithmique minimum
EER	Taux d'erreur égale
F_0	Fréquence fondamentale
FA	Fausses acceptations
FAR	Taux de fausses acceptations
FR	Faux rejets
FRR	Taux de faux rejets
LLR	Logarithme du rapport de vraissemblance
LR	Rapport de vraissemblance
μ	Moyenne
δ	Écart-type

Références bibliographiques

Bibliographie personnelle

Chanclu, A., Ben Amor, I., Gendrot, C., Ferragne, E., & Bonastre, J.-F. (2021). Automatic Classification of Phonation Types in Spontaneous Speech : Towards a New Workflow for the Characterization of Speakers & Voice Quality. *Proc. Interspeech 2021*, 1015-1018. DOI : 10.21437/Interspeech.2021-1765

Chanclu, A., Georgeton, L., Fredouille, C., & Bonastre, J.-F. (2020). PTSVOX : une base de données pour la comparaison de voix dans le cadre judiciaire. In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla & S. Schneider (Éd.), *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). Volume 1 : Journées d'Études sur la Parole* (p. 73-81). ATALA. https://hal.archives-ouvertes.fr/hal-02798519

Gendrot, C., Ferragne, E., & Chanclu, A. (2022). Analyse phonétique de la variation interlocuteurs au moyen de réseaux de neurones convolutifs : voyelles seules et séquences courtes de parole. *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, 891-899. DOI : 10.21437/JEP.2022-94

O'Brien, B., Tomashenko, N., Chanclu, A., & Bonastre, J.-F. (2021). Anonymous speaker clusters : Making distinctions between anonymised speech recordings with clustering interface. *Proc. Interspeech 2021*, 3580-3584. DOI : 10.21437/Interspeech.2021-1588

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., & Maouche, M. (2022a). The VoicePrivacy 2020 Challenge : Results and findings. *Computer Speech & Language*, *74*, 101362. DOI : 10.1016/j.csl.2022.101362

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., & Maouche, M. (2022b). *Supplementary material to the paper The VoicePrivacy 2020 Challenge : Results and findings* [Supplementary material to the paper "The VoicePrivacy 2020 Challenge : Results and findings" (https://hal.archives-ouvertes.fr/hal-03332224) submitted to CSL.]. https://hal.archives-ouvertes.fr/hal-03335126

Bibliographie

Aitken, C. G., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists* (T. 26). Wiley Chichester.

Ajili, M. (2017). *Reliability of voice comparison for forensic applications* (thèse de doct.) [2017AVIG0223]. http://www.theses.fr/2017AVIG0223/document

Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., & Kahn, J. (2016a). Phonetic content impact on Forensic Voice Comparison. *2016 IEEE Spoken Language Technology Workshop (SLT)*, 210-217. DOI : 10.1109/SLT.2016.7846267

Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016b). FABIOLE, a Speech Database for Forensic Speaker Comparison. In N. C. (Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Éd.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (p. 726-733). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/summaries/35.html

Ajili, M., Bonastre, J.-F., Kheder, W. B., Rossato, S., & Kahn, J. (2017). Homogeneity Measure Impact on Target and Non-Target Trials in Forensic Voice Comparison. *INTERSPEECH*, 2844-2848.

Ajili, M., Bonastre, J.-F., Rossato, S., Kahn, J., & Lapidot, I. (2015). Homogeneity Measure for Forensic Voice Comparison : A Step Forward Reliability. In A. Pardo & J. Kittler (Éd.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (p. 135-142). Springer International Publishing.

Ajili, M., Rossato, S., Zhang, D., & Bonastre, J.-F. (2018). Impact of rhythm on forensic voice comparison reliability. *Odyssey 2018 The Speaker and Language Recognition Workshop*. https://hal.archives-ouvertes.fr/hal-01962531

Aphatie, J.-M., & Feltin-Palas, M. (2020). *J'ai un accent, et alors ?* Michel Lafon.

Arnold, A. (2015). *La voix genrée, entre idéologies et pratiques – Une étude sociophonétique* (thèse de doct.). Sorbonne Paris Cité. http://www.theses.fr/2015USPCA148

Assal, G., Zander, E., Kremin, H., & Buttet, J. (1976). Voice discrimination in patients with cerebral cortical lesions. *Schweizer Archiv fur Neurologie, Neurochirurgie und Psychiatrie = Archives suisses de neurologie, neurochirurgie et de psychiatrie*, *119*(2), 307-315. http://europepmc.org/abstract/MED/1006205

Association Francophone de la Communication Parlée (AFCP). (2002). *Motion adoptée à l'unanimité par le bureau du GCP (groupe de la communication parlée) de la SFA, reconduite intégralement par le GFCP de la SFA en 1997 et par l'AFCP en 2002*. http://www.afcp-parole.org/doc/MOTION_1990.pdf

Bach, A. C., Lederer, F. L., & Dinolt, R. (1941). Senile changes in the laryngeal musculature. *Archives of Otolaryngology*, *34*(1), 47-56. DOI : 10.1001/archotol.1941.00660040057006

Baken, R. J. (2005). The Aged Voice : A New Hypothesis. *Journal of Voice*, *19*(3), 317-325. DOI : 10.1016/j.jvoice.2004.07.005

Baker, P. (2008). *Sexed texts : language, gender and sexuality.* Equinox.

Bargh, J. A., & Chartrand, T. L. (2014). The mind in the middle : A practical guide to priming and automaticity research., 311-344.

Beck, J. M. (2010). Organic variation of the Vocal Apparatus. *The Handbook of Phonetic Sciences : Second Edition*, 153-201. DOI : 10.1002/9781444317251.ch5

Ben Amor, I., & Bonastre, J.-F. (2022). BA-LR : Binary-Attribute-based Likelihood Ratio estimation for forensic voice comparison. *2022 International Workshop on Biometrics and Forensics (IWBF)*, 1-6. DOI : 10.1109/IWBF55382.2022.9794542

Berhuet, S., Bléhaut, M., Brice-Mansencal, L., Croutte, P., Millot, C., & Müller, J. (2021). Baromètre du numérique 2021. *Autorité de régulation des communications électroniques, des postes et de la distribution de la presse (ARCEP)*. https://www.economie.gouv.fr/cge/barometre-numerique-2021

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (T. 405). John Wiley & Sons.

Bilmes, J. A., *et al.*. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International computer science institute*, *4*(510), 126.

Blanchet, P. (2016). *Discriminations : combattre la glottophobie*. Textuel.

Blanchet, P. (2021). Glottophobie. *Langage et société*, *Hors série*(HS1), 155-159. DOI : 10.3917/ls.hs01.0156

Boë, L.-J. (2000). Forensic voice identification in France. *Speech Communication*, *31*(2-3), 205-224. DOI : 10.1016/S0167-6393(99)00079-5

Boë, L.-J., Bimbot, F., Bonastre, J.-F., & Dupont, P. (1999). De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Langues*, *2*(4), 270-288.

Boë, L.-J., Contini, M., & Rakotofiringa, H. (1975). Étude statistique de la fréquence laryngienne. *Phonetica*, *32*(1), 1-23. DOI : 10.1159/000259683

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot. International*, *5*(9), 341-345.

Bolt, R. H., Cooper, F. S., David Jr, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1973). Speaker identification by speech spectrograms : some further observations. *The Journal of the Acoustical Society of America*, *54*(2), 531-534. DOI : 10.1121/1.1911935

Bonastre, J.-F. (2020). 1990-2020 : retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire. In G. Adda, M. Amblard & K. Fort (Éd.), *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). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)* (p. 38-47). ATALA. https://hal.archives-ouvertes.fr/hal-02750225

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J. P., Reynolds, D. A., & Magrin-Chagnolleau, I. (2003). Person authentication by voice : A need for caution. *Eighth European Conference on Speech Communication and Technology*.

Boughton, Z., & Armstrong, N. (1998). Identification and evaluation responses to a French accent : some results and issues of methodology. *Revue Parole*, 27-60.

Bricker, P. D., & Pruzansky, S. (1976). Speaker Recognition (N. J. Lass, Éd.), 295-326. DOI : 10.1016/B978-0-12-437150-7.50015-4

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal Attractiveness Increases by Averaging. *Current Biology*, *20*(2), 116-120. DOI : 10.1016/j.cub.2009.11.034

Brummer, N., & Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*, 230-275. DOI : 10.1016/j.csl.2005.08.001

Brummer, N., & Van Leeuwen, D. A. (2006). On calibration of language recognition scores. *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, 1-8. DOI : 10.1109/ODYSSEY.2006.248106

Buchtel, H. A., & Stewart, J. D. (1989). Auditory agnosia : Apperceptive or associative disorder ? *Brain and Language*, *37*(1), 12-25. DOI : 10.1016/0093-934X(89)90098-9

Buckleton, J. S., Bright, J.-A., & Taylor, D. (2018). *Forensic DNA evidence interpretation*. CRC press.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, *26*(2), 95-103. DOI : 10.1109/MSP.2008.931100

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE signal processing letters*, *13*(5), 308-311.

Candea, M. (2017). La notion d'« accent de banlieue » à l'épreuve du terrain. *GlottopoL*, (29), 13-26.

Cantril, H., & Allport, G. W. (1935). The psychology of radio.

Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PLOS ONE*, *7*(2), 1-12. DOI : 10.1371/journal.pone.0031353

Cerrato, L., Falcone, M., & Paoloni, A. (2000). Subjective age estimation of telephonic voices. *Speech Communication*, *31*(2), 107-112. DOI : 10.1016/S0167-6393(99)00071-0

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2), 193-203. DOI : 10.1016/S0167-6393(99)00078-3

Chen, G.-t. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, *2*(2), 159-171. http://www.jstor.org/stable/23752908

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518. DOI : 10.1109/JSTSP.2022.3188113

Chevrie-Muller, C., Salomon, D., & Ferrey, G. (1971). Contribution à l'établissement de quelques constantes physiologiques de la voix parlée de la femme adolescente, adulte et âgée. *Journal Français d'Oto-Rhino-Laryngologie*, *16*, 433-455.

Childers, D. G., & Wu, K. (1991). Gender recognition from speech. Part II : Fine analysis. *The Journal of the Acoustical Society of America*, *90*(4), 1841-1856. DOI : 10.1121/1.401664

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., & Han, I. (2020). In Defence of Metric Learning for Speaker Recognition. *Proc. Interspeech 2020*, 2977-2981. DOI : 10.21437/Interspeech.2020-1064

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2 : Deep Speaker Recognition. *INTERSPEECH*.

Cohen, J., Crystal, T., House, A., & Neuburg, E. (1980). Weighty voices and shaky evidence : A critique. *The Journal of the Acoustical Society of America*, *68*(6), 1884-1886. DOI : 10.1121/1.385178

Cole, R. A., Noel, M., & Noel, V. (1998). The CSLU speaker recognition corpus. *Fifth International Conference on Spoken Language Processing*.

Coleman, R. O. (1983). Acoustic correlates of speaker sex identification : Implications for the transsexual voice. *The Journal of Sex Research*, *19*(3), 293-295. DOI : 10.1080/00224498309551189

Committee on DNA Technology in Forensic Science and National Research Council and Life Sciences Commission and Division on Earth and Life Studies and Commission on Life Sciences and Division on Earth and Life Studies Staff and National Research Council Staff and others. (1992). *DNA technology in forensic science*. National Academy Press.

Cornut, G. (2019). *La voix*. Que sais-je.

Craik, F. I., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *The Quarterly Journal of Experimental Psychology*, *26*(2), 274-284. DOI : 10.1080/14640747408400413

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). The CHAINS speech corpus : Characterizing individual speakers. *Proc of SPECOM*, 1-6.

Dallaston, K., & Docherty, G. (2020). The quantitative prevalence of creaky voice (vocal fry) in varieties of English : A systematic review of the literature. *PLOS ONE*, *15*(3), 1-18. DOI : 10.1371/journal.pone.0229960

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection* [or the Preservation of Favored Races in the Struggle for Life]. Murray.

Davies, S. (2012). *The artful species : Aesthetics, art, and evolution*. OUP Oxford.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding : Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174-1176. DOI : 10.1126/science.7375928

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798. DOI : 10.1109/TASL.2010.2064307

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, *39*(1), 1-22. DOI : 10.1111/j.2517-6161.1977.tb01600.x

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN : Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, 3830-3834. DOI : 10.21437/Interspeech.2020-2650

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association*, *39*(2), 162-188.

Dixon, L., & Gill, B. (2001). *Changes in the Standards for Admitting Expert Evidence in Federal Civil Cases Since the Daubert Decision*. RAND Corporation.

Doddington, G. R. (1971). A Method or Speaker Verification. *The Journal of the Acoustical Society of America*, *49*(1A_Supplement), 139-139. DOI : 10.1121/1.1975906

Dolson, M. (1994). The pitch of speech as a function of linguistic community. *Music Perception*, *11*(3), 321-331. DOI : 10.2307/40285626

Ehrlich, P., & Feldman, M. (2003). Genes and Cultures : What Creates Our Behavioral Phenome ? *Current Anthropology*, *44*(1), 87-107. DOI : 10.1086/344470

Ellis, S. (1994). The Yorkshire ripper enquiry : part I. *International Journal of Speech Language and the Law*, *1*(2), 197-206.

Enzinger, E., & Morrison, G. S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, *277*, 30-40. DOI : 10.1016/j.forsciint.2017.05.007

Eriksson, A. (2012). Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work. In A. Neustein & H. A. Patil (Éd.), *Forensic Speaker Recognition : Law Enforcement and Counter-Terrorism* (p. 41-69). Springer New York. DOI : 10.1007/978-1-4614-0263-3_3

Evett, I., Jackson, G., Lambert, J., & McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & justice : journal of the Forensic Science Society*, *40*(4), 233-239. DOI : 10.1016/s1355-0306(00)71993-9

Ferguson, I. f. D. A., John D. (1980). *Symposium on the application of hidden Markov models to text and speech* (English language ed.). Institute for Defense Analyses, Communications Research Division Princeton, NJ.

French, P., Harrison, P., & Lewis, J. W. (2006). R v John Samuel humble : The Yorkshire ripper hoaxer trial. *The International Journal of Speech, Language and the Law*, *13*(2), 255-273.

Fryar, C. D., Kruszan-Moran, D., Gu, Q., & Ogden, C. L. (2018). Mean body weight, weight, waist circumference, and body mass index among adults : United States, 1999–2000 through 2015–2016.

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, *41*(2), 185-205.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1, 27403.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., Schweinberger, S. R., Warren, J. D., & Duchaine, B. (2009). Developmental phonagnosia : a selective deficit of vocal identity recognition. *Neuropsychologia*, *47*(1), 123-131. DOI : 10.1016/j.neuropsychologia.2008.08.003

Gaurenne, R., Ludes, B., & Pfitzinger, H. (2022). Police scientifique [Consulté le 20 avril 2022]. In *Encyclopædia Universalis*. https://www.universalis.fr/encyclopedie/police-scientifique/

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291-298. DOI : 10.1109/89.279278

GFCP. (1999). *Pétition pour l'arrêt des expertises vocales tant qu'elles n'auront pas été validées scientifiquement*. http://www.afcp-parole.org/doc/petition_GFCP.pdf

Gibbs, R. W., & Van Orden, G. C. (2010). Adaptive cognition without massive modularity. *Language and Cognition*, *2*(2), 149-176. DOI : 10.1515/langcog.2010.006

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD : Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, *1*, 517-520.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition [Odyssey 2004 : The speaker and Language Recognition Workshop]. *Computer Speech & Language*, *20*(2), 331-355. DOI : 10.1016/j.csl.2005.08.005

Gonzalez-Rodriguez, J., & Ramos, D. (2007). Forensic Automatic Speaker Classification in the "Coming Paradigm Shift". In C. Müller (Éd.), *Speaker Classification I : Fundamentals, Features, and Methods* (p. 205-217). Springer Berlin Heidelberg. DOI : 10.1007/978-3-540-74200-5_11

Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., & Ortega-Garcia, J. (2007). *Emulating DNA : Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition*. DOI : 10.1109/TASL.2007.902747

Gordon, M. (2001). Linguistic aspects of voice quality with special reference to Athabaskan. *Proceedings of the 2001 Athabaskan languages conference*, 163-178.

Gordon, M., & Ladefoged, P. (2001). Phonation types : a cross-linguistic overview. *Journal of phonetics*, *29*(4), 383-406.

Graddol, D. (1986). Discourse specific pitch behaviour. In *Intonation in discourse* (p. 221-238). London/Sidney : Croom Helm.

Gray, H. (1918). Anatomy of the human body. *Annals of surgery*, *68*(5), 564-566.

Greenberg, C. S., Mason, L. P., Sadjadi, S. O., & Reynolds, D. A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech & Language*, *60*, 101032.

Grossi, G., Kelly, S., Nash, A., & Parameswaran, G. (2014). Challenging dangerous ideas : a multidisciplinary critique of evolutionary psychology. *Dialectical Anthropology*, *38*(3), 281-285. DOI : 10.1007/s10624-014-9358-x

Hailstone, J. C., Ridgway, G. R., Bartlett, J. W., Goll, J. C., Buckley, A. H., Crutch, S. J., & Warren, J. D. (2011). Voice processing in dementia : a neuropsychological and neuroanatomical analysis. *Brain*, *134*(9), 2535-2547. DOI : 10.1093/brain/awr205

Hamilton, R. (2008). The Darwinian Cage : Evolutionary Psychology as Moral Science. *Theory, Culture & Society*, *25*(2), 105-125. DOI : 10.1177/0263276407086793

Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). End-to-end text-dependent speaker verification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5115-5119. DOI : 10.1109/ICASSP.2016.7472652

Heo, H. S., Lee, B.-J., Huh, J., & Chung, J. S. (2020). Clova Baseline System for the VoxCeleb Speaker Recognition Challenge 2020.

Herzog, H. (1933). Stimme und Persönlichkeit. *Zeitschrift für Psychologie*, *190*, 300-369.

Higgins, A. (1990). YOHO speaker verification. *Speech Research Symposium, Baltimore, MD*.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*(2), 139-148. DOI : 10.1016/S0095-4470(19)30953-2

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech and Hearing Research*, *15*(1), 155-159. DOI : 10.1044/jshr.1501.155

Hornibrook, J., Ormond, T., & Maclagan, M. (2018). Creaky voice or extreme vocal fry in young women. *New Zealand Medical Journal*, *131*(1486), 36-40.

Hughes, S. M., Dispenza, F., & Gallup, G. G. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. *Evolution and Human Behavior*, *25*(5), 295-304. DOI : 10.1016/j.evolhumbehav.2004.06.001

Hughes, S. M., & Harrison, M. A. (2017). Your Cheatin' Voice Will Tell on You : Detection of Past Infidelity from Voice. *Evolutionary Psychology*, *15*(2). DOI : 10.1177/1474704917711513

Huntley, R., Hollien, H., & Shipp, T. (1987). Influences of listener characteristics on perceived age estimations. *Journal of Voice*, *1*(1), 49-52.

Huteau, M. (2021). *Psychologie différentielle*. Dunod.

Hutiri, W. T., & Ding, A. Y. (2021). SVEva Fair : A Framework for Evaluating Fairness in Speaker Verification. *arXiv preprint arXiv:2107.12049*.

Hutiri, W. T., & Ding, A. Y. (2022). Bias in Automated Speaker Recognition. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 230-247. DOI : 10.1145/3531146.3533089

Ikeno, A., & Hansen, J. H. (2007). The effect of listener accent background on accent perception and comprehension. *EURASIP Journal on Audio, Speech, and Music Processing*, *2007*, 1-8.

Innocence Project. (2016a). Brian Piszczek [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/brian-piszczek/

Innocence Project. (2016b). David Shephard [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/david-shephard/

Innocence Project. (2016c). Dean Cage [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/dean-cage/

Innocence Project. (2016d). Eduardo Velasquez [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/eduardo-velasquez/

Innocence Project. (2016e). James Waller [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/james-waller/

Innocence Project. (2016f). Kerry Kotler [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/kerry-kotler/

Innocence Project. (2016g). Michael Anthony Williams [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/michael-anthony-williams/

Innocence Project. (2016h). Sedrick Courtney [Consulté le 9 novembre 2021]. https://innocenceproject.org/cases/sedrick-courtney/

Jain, A., Bolle, R., & Pankanti, S. (1999). *Biometrics : personal identification in networked society* (T. 479). Springer Science & Business Media.

Jain, A. K., Flynn, P., & Ross, A. A. (2007). *Handbook of biometrics*. Springer Science & Business Media.

Johns-Lewis, C. M. (1986). Prosodic differentiation of discourse modes. In *Intonation in discourse* (p. 199-220). London - Sidney : Croom Helm.

Jones, K., Walker, K., Caruso, C., Wright, J., & Strassel, S. (2022). WeCanTalk : A New Multi-language, Multi-modal Resource for Speaker Recognition. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3451-3456. https://aclanthology.org/2022.lrec-1.369

Kahane, J. C. (1981). Anatomic and physiologic changes in the aging peripheral speech mechanism. *Aging, Communication Process and Disorders*.

Kahane, J. C. (1990). Age-related changes in the peripheral speech mechanism : Structural and physiological changes. *ASHA Reports*, *19*, 75-87.

Kahn, J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale* (thèse de doct.) [2011AVIG0187]. http://www.theses.fr/2011AVIG0187/document

Kahn, J., Audibert, N., Rossato, S., & Bonastre, J.-F. (2010). Intra-speaker variability effects on Speaker Verification performance. *Odyssey*, 21.

Kenny, P., & Dumouchel, P. (2004). Disentangling speaker and channel effects in speaker verification. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I-37. DOI : 10.1109/ICASSP.2004.1325916

Kersta, L. G. (1962). Voiceprint identification. *The Journal of the Acoustical Society of America*, *34*(5), 725-725.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses : effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327-336. DOI : 10.1002/acp.974

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses : effects of accent, retention and telephone. *Applied Cognitive Psychology*, *20*(2), 187-197. DOI : 10.1002/acp.1175

Kim, L., & Gendrot, C. (2022). Classification automatique de voyelles nasales pour une caractérisation de la qualité de voix des locuteurs par des réseaux de neurones convolutifs. *Journées d'Etude de la Parole 2022 (JEP 2022)*. https://shs.hal.science/halshs-03980367

Kim, L., Gendrot, C., Elmerich, A., Amelot, A., & Maeda, S. (2023). Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques. In C. Servan & A. Vilnat (Éd.), *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues* (p. 101-108). ATALA. https://hal.science/hal-04130227

Kinoshita, Y., Ishihara, S., *et al.*. (2014). Background population : how does it affect LR based forensic voice comparison ? *The International Journal of Speech, Language and the Law*, *21*(2), 191-224.

Kisilevsky, B. S., Hains, S. M., Lee, K., Xie, X., Huang, H., Ye, H. H., Zhang, K., & Wang, Z. (2003). Effects of Experience on Fetal Voice Recognition. *Psychological Science*, *14*(3), 220-224. DOI : 10.1111/1467-9280.02435

Kitzing, P. (1979). *Glottografisk frekvensindikering : En undersökningsmetod för mätning av röstläge och röstomfang samt framställning av röstfrekvensdistributionen*. Oronkliniken i Malmö.

Klatt, D. H. (1986). Detailed spectral analysis of a female voice. *The Journal of the Acoustical Society of America*, *80*(S1), S97-S97. DOI : 10.1121/1.2024070

Klatt, D. H. (1987). Acoustic correlates of breathiness : First harmonic amplitude, turbulence noise, and tracheal coupling. *The Journal of the Acoustical Society of America*, *82*(S1), S91-S91. DOI : 10.1121/1.2025051

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, *87*(2), 820-857. DOI : 10.1121/1.398894

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner : voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B : Biological Sciences*, *279*(1738), 2698-2704. DOI : 10.1098/rspb.2012.0311

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*(6), 618-625. DOI : 10.1016/S0022-1031(02)00510-3

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies : An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583-621. DOI : 10.1080/01621459.1952.10483441

Künzel, H. J. (1989). How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight ? *Phonetica*, *46*(1-3), 117-125. DOI : 10.1159/000261832

Kwan, Q. Y. (1977). *Inference of the identity of source* (thèse de doct.). University of California, Berkeley.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Ladefoged, P. (1980). The ability of listeners to identify voices. *UCLA Work. Pap. Phonet.*, *49*, 43-89.

Lancaster, R. N. (2003). *The Trouble with Nature : Sex in Science and Popular Culture*. University of California Press. DOI : 10.1525/9780520936799

Lass, N. J., Barry, P. J., Reed, R. A., Walsh, J. M., & Amuso, T. A. (1979a). The effect of temporal speech alterations on Speaker height and weight identification. *Language and Speech*, *22*(2), 163-171. DOI : 10.1177/002383097902200206

Lass, N. J., Beverly, A. S., Nicosia, D. K., & Simpson, L. A. (1978). An investigation of speaker height and weight identification by means of direct estimations. *Journal of Phonetics*, *6*(1), 69-76. DOI : 10.1016/S0095-4470(19)31085-X

Lass, N. J., Brong, G. W., Ciccolella, S. A., Walters, S. C., & Maxwell, E. L. (1980a). An investigation of speaker height and weight discriminations by means of paired comparison judgments. *Journal of Phonetics*, *8*(2), 205-212. DOI : 10.1016/S0095-4470(19)31465-2

Lass, N. J., & Colt, E. G. (1980). A comparative study of the effect of visual and auditory cues on speaker height and weight identification. *Journal of Phonetics*, *8*(3), 277-285. DOI : 10.1016/S0095-4470(19)31464-0

Lass, N. J., & Davis, M. (1976). An investigation of speaker height and weight identification. *The Journal of the Acoustical Society of America*, *60*(3), 700-703. DOI : 10.1121/1.381142

Lass, N. J., Dicola, G. A., Beverly, A. S., Barbera, C., Henry, K. G., & Badali, M. K. (1979b). The effect of phonetic complexity on speaker height and weight identification. *Language and Speech*, *22*(4), 297-309. DOI : 10.1177/002383097902200401

Lass, N. J., Kelley, D. T., Cunningham, C. M., & Sheridan, K. J. (1980b). A comparative study of speaker height and weight identification from voiced and whispered speech. *Journal of Phonetics*, *8*(2), 195-204.

Lass, N. J., Phillips, J. K., & Bruchey, C. A. (1980c). The effect of filtered speech on speaker height and weight identification. *Journal of Phonetics*, *8*(1), 91-100. DOI : 10.1016/S0095-4470(19)31453-6

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. *Social markers in speech*, *1*, 32.

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology*, 4(1), 63-74. DOI : 10.1023/A:1009656816383

Lewis, J. W. (1989). The Yorkshire Ripper Hoax Tape. http://www.yek.me.uk/ykrprhoaxtp.html

Li, K. P., Dammann, J. E., & Chapman, W. D. (1966). Experimental Studies in Speaker Verification, Using an Adaptive System. *The Journal of the Acoustical Society of America*, *40*(5), 966-978. DOI : 10.1121/1.1910221

Li, S. Z., & Jain, A. (2015). *Encyclopedia of biometrics*. Springer Publishing Company, Incorporated.

Lickliter, R., & Honeycutt, H. (2003). Developmental dynamics : toward a biologically plausible evolutionary psychology. *Psychological bulletin*, *129*(6), 819. DOI : 10.1037/0033-2909.129.6.819

Lin, W.-W., & Mak, M.-W. (2020). Wav2Spk : A Simple DNN Architecture for Learning Speaker Embeddings from Waveforms. *INTERSPEECH*, 3211-3215.

Linville, S. E. (1981). *Acoustic characteristics of adult women's voices with advancing age : A production and perception study* (thèse de doct.). Northwestern University.

Linville, S. E. (1995). Vocal aging. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *3*(3), 183-187. https://journals.lww.com/co-otolaryngology/Fulltext/1995/06000/Vocal_aging.6.aspx

Linville, S. E., & Fisher, H. B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *The Journal of the Acoustical Society of America*, *78*(1), 40-48. DOI : 10.1121/1.392452

Linville, S. E., Korabic, E. W., & Rosera, M. (1990). Intraproduction variability in jitter measures from elderly speakers. *Journal of Voice*, *4*(1), 45-51. DOI : 10.1016/S0892-1997(05)80081-5

Linville, S. E., Skarin, B. D., & Fornatto, E. (1989). The interrelationship of measures related to vocal function, speech rate, and laryngeal appearance in elderly women. *Journal of Speech, Language, and Hearing Research*, *32*(2), 323-330.

Locard, E. (1931). *Traité de criminalistique*. J. Desvignes.

Mahler, D. A. (1983). Pulmonary aspects of aging. In S. R. Gambert (Éd.), *Contemporary geriatric medicine* (p. 45-85). Springer US. DOI : 10.1007/978-1-4684-4514-5_2

Maiers, W. (2001). Psychological theorising in transdisciplinary perspective. In *Theoretical issues in psychology* (p. 275-287). Springer.

Malinowski, A. (1967). Shape, dimensions and process of calcification of the cartilaginous framework of the larynx in relation to age and sex in the Polish population. *Folia morphologica*, *26*(2), 121-132. http://europepmc.org/abstract/MED/5299694

Maltoni, D., Maio, D., Jain, A. K., & Prabhakar, S. (2009). *Handbook of fingerprint recognition*. Springer Science & Business Media.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50-60. DOI : 10.1214/aoms/1177730491

Martinet, A. (1960). Eléments de linguistique générale.

Matějka, P., Glembek, O., Novotný, O., Plchot, O., Grézl, F., Burget, L., & Cernocký, J. H. (2016). Analysis of DNN approaches to speaker identification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5100-5104. DOI : 10.1109/ICASSP.2016.7472649

Matsui, T., Kanno, T., & Furui, S. (1996). Speaker recognition using HMM composition in noisy environments. *Computer Speech & Language*, *10*(2), 107-116. DOI : 10.1006/csla.1996.0007

McGehee, F. (1937). The Reliability of the Identification of the Human Voice. *The Journal of General Psychology*, *17*(2), 249-271. DOI : 10.1080/00221309.1937.9917999

McGlone, R. E., & Hollien, H. (1963). Vocal Pitch Characteristics of Aged Women. *Journal of Speech and Hearing Research*, *6*(2), 164-170. DOI : 10.1044/jshr.0602.164

Mclaren, M., Vogt, R., Baker, B., & Sridharan, S. (2007). A Comparison of Session Variability Compensation Techniques for SVM-Based Speaker Recognition. In R. van Son & H. V. hamme (Éd.), *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)* (p. 790-793). Casual Productions Pty Ltd. https://eprints.qut.edu.au/7594/

Mello, L., & Nymouce, J. (2021). #SeMEME 2 : Memes, reality shows e gastronomia com Luiza Mello e Jeannot Nymouce (PT/FR) [Consulté le 21 février 2022]. *YouTube*. https://www.youtube.com/watch?v=_zegEPbP2Rw#t=15m00s

Meuwly, D. (2000). *Reconnaissance de locuteurs en sciences forensiques : l'apport d'une approche automatique* (thèse de doct.). Université de Lausanne, Faculté de droit et des sciences criminelles.

Meuwly, D. (2006). Forensic individualisation from biometric data. *Science & Justice*, *46*(4), 205-213.

Meuwly, D., & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). *2001 : A Speaker Odyssey-The Speaker Recognition Workshop*.

Mills, M., & Melhuish, E. (1974). Recognition of mother's voice in early infancy. *Nature*, *252*(5479), 123-124. DOI : 10.1038/252123a0

Morris, R., & Brown Jr, W. (1988). Age-related differences in F0 and pitch sigma among females. *IASCP Bull*, *2*, 36-40.

Morriso, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*(4), 298-308. DOI : 10.1016/j.scijus.2009.09.002

Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., Ypma, R. J., *et al.*. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 299-309. DOI : 10.1016/j.scijus.2021.02.002

Morrison, G. S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B. K., Souza, S. D., Cummins, N., & Chow, D. (2015). Forensic database of voice recordings of 500+ Australian English speakers. http://databases.forensic-voice-comparison.net/

Mulac, A., & Giles, H. (1996). 'Your're Only As Old As You Sound' : Perceived Vocal Age and Social Meanings. *Health Communication*, *8*(3), 199-215.

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *The Journal of the Acoustical Society of America*, *98*(6), 3080-3095. DOI : 10.1121/1.413832

Mysak, E. D., & Hanley, T. (1958). Aging processes in speech : Pitch and duration characteristics. *Journal of Gerontology*, *13*, 309-313. DOI : 10.1093/geronj/13.3.309

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2019). Voxceleb : Large-scale speaker verification in the wild. *Computer Science and Language*.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb : A Large-Scale Speaker Identification Dataset, 2616-2620. DOI : 10.21437/Interspeech.2017-950

Nandwana, M. K., Ferrer, L., McLaren, M., Castan, D., & Lawson, A. (2019). Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems. *INTERSPEECH*, 4325-4329.

National Institute of Standards and Technology. (2019). NIST 2019 Speaker Recognition Evaluation Plan. https://www.nist.gov/itl/iad/mig/nist-2019-speaker-recognition-evaluation

National Research Council and others. (1996). The evaluation of forensic DNA evidence.

National Research Council and others. (2009). *Strengthening forensic science in the United States : a path forward*. National Academies Press.

Nautsch, A., Saeidi, R., Rathgeb, C., & Busch, C. (2016). Robustness of Quality-based Score Calibration of Speaker Recognition Systems with respect to low-SNR and short-duration conditions. *Odyssey*, 358-365.

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and cognition*, 44(3), 342-366. DOI : 10.1006/brcg.1999.1196

O'Connor, J. J., Re, D. E., & Feinberg, D. R. (2011). Voice pitch influences perceptions of sexual infidelity. *Evolutionary Psychology*, *9*(1), 147470491100900109. DOI : 10.1177/147470491100900109

Öhman, L., Eriksson, A., & Granhag, P. A. (2010). Mobile phone quality vs direct quality : How the presentation format affects the identification accuracy. *European Journal of Psychology Applied to Legal Context*, *2*(2).

Öhman, L., Eriksson, A., & Granhag, P. A. (2011). Overhearing the planning of a crime : Do adults outperform children as earwitnesses ? *Journal of Police and Criminal Psychology*, *26*(2), 118-127. DOI : 10.1007/s11896-010-9076-5

Öhman, L., Eriksson, A., & Granhag, P. A. (2013a). Angry Voices from the Past and Present : Effects on Adults' and Children's Earwitness Memory. *Journal of Investigative Psychology and Offender Profiling*, *10*(1), 57-70. DOI : 10.1002/jip.1381

Öhman, L., Eriksson, A., & Granhag, P. A. (2013b). Enhancing Adults' and Children's Earwitness Memory : Examining Three Types of Interviews. *Psychiatry, Psychology and Law*, *20*(2), 216-229. DOI : 10.1080/13218719.2012.658205

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology : Applied*, 4(2), 101-118. DOI : 10.1037/1076-898X.4.2.101

Orton, H., Dieth, E., of Leeds. Department of English Language, U., & Literature, M. E. (1962–1971). *Survey of English Dialects : B, the Basic Material*. University of Leeds.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech : an asr corpus based on public domain audio books. *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206-5210.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 1-12.

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., & Bengio, Y. (2019). Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks, 161-165. DOI : 10.21437/Interspeech.2019-2605

Paulson, W. (2018). *Literary Culture in a World Transformed : A Future for the Humanities*. Cornell University Press. DOI : 10.7591/9781501729348

Pear, T. H. (1931). Voice and personality.

Pearce, D. (1998). *Aurora Project : Experimental framework for the performance evaluation of distributed speech recognition front-ends* (rapp. tech.). ETSI working paper.

Pegoraro Krook, M. I. (1988). Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis. *Folia Phoniatrica et Logopaedica*, *40*(2), 82-90. DOI : 10.1159/000265888

Pélissier, M., & Ferragne, E. (2022). The N400 reveals implicit accent-induced prejudice. *Speech Communication*, *137*, 114-126. DOI : 10.1016/j.specom.2021.10.004

Pépiot, E. (2013). *Voix de femmes, voix d'hommes : différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones* (thèse de doct.). Université Paris VIII Vincennes-Saint Denis.

Pépiot, E. (2015). Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, (HS-16). DOI : 10.4000/corela.3783

Pernet, C. R., & Belin, P. (2012). The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, *3*, 23. DOI : 10.3389/fpsyg.2012.00023

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance : the effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, *21*(4), 539-550. DOI : 10.1002/acp.1296

Plénat, M. (1985). Morphologie du largonji et des loucherbems. *Langages*, (78), 73-95. DOI : 10.2307/41682029

Plotkin, H. (2008). *Evolutionary thought in psychology : A brief history*. John Wiley & Sons.

Podesva, R. J. (2007). Phonation type as a stylistic variable : The use of falsetto in constructing a persona 1. *Journal of sociolinguistics*, *11*(4), 478-504.

Podesva, R. J. (2011). Gender and the social meaning of non-modal phonation types. *Annual Meeting of the Berkeley Linguistics Society*, *37*(1), 427-448.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., & Khudanpur, S. (2018). Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. *Interspeech*, 3743-3747.

Pruzansky, S. (1963). Pattern-Matching Procedure for Automatic Talker Recognition. *The Journal of the Acoustical Society of America*, *35*(3), 354-358. DOI : 10.1121/1.1918467

Ptacek, P. H., & Sander, E. K. (1966). Age recognition from voice. *Journal of speech and hearing Research*, *9*(2), 273-277. DOI : 10.1044/jshr.0902.273

Ptacek, P. H., Sander, E. K., Maloney, W. H., & Jackson, C. R. (1966). Phonatory and related changes with advanced age. *Journal of speech and hearing research*, *9*(3), 353-360.

Ramig, L. A., & Ringel, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech, Language, and Hearing Research*, *26*(1), 22-30. DOI : 10.1044/jshr.2601.22

Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., & Lucena-Molina, J. J. (2008). Addressing database mismatch in forensic speaker recognition with Ahumada III : a public real-casework database in Spanish. *Ninth Annual Conference of the International Speech Communication Association*.

Rappaport, W. (1958). Über Messungen der Tonhöhenverteilung in der deutschen Sprache. *Acta Acustica United with Acustica*, *8*(4), 220-225.

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., … Bengio, Y. (2021). SpeechBrain : A General-Purpose Speech Toolkit [arXiv:2106.04624].

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., & Bengio, Y. (2020). Multi-task self-supervised learning for robust speech recognition. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6989-6993.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1), 19-41. DOI : 10.1006/dspr.1999.0361

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, *3*(1), 72-83. DOI : 10.1109/89.365379

Robertson, B., Vignaux, G. A., & Berger, C. E. (2016). *Interpreting evidence : evaluating forensic science in the courtroom*. John Wiley & Sons.

Rodeno, M., Sánchez-Fernández, J., & Rivera-Pomar, J. (1993). Histochemical and morphometrical ageing changes in human vocal cord muscles. *Acta Oto-Laryngologica*, *113*(3), 445-449. DOI : 10.3109/00016489309135842

Roncallo, P. (1948). Researches about ossification and conformation of the thyroid cartilage in men. *Acta Oto-Laryngologica*, *36*(2), 110-134. DOI : 10.3109/00016484809124245

Rose, H., & Rose, S. (2010). *Alas poor Darwin : Arguments against evolutionary psychology*. Random House.

Rose, P. (1991). How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency ? *Speech Communication*, *10*(3), 229-247. DOI : 10.1016/0167-6393(91)90014-K

Rose, P. (2002). *Forensic speaker identification*. CRC Press.

Rose, P., *et al.*. (2003). The technical comparison of forensic voice samples. In *Expert Evidence*. The Law Book Company.

Rose, R., Hofstetter, E., & Reynolds, D. (1994). Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 245-257. DOI : 10.1109/89.279273

Rossi, A., Ganassini, A., Tantucci, C., & Grassi, V. (1996). Aging and the respiratory system. *Aging Clinical and Experimental Research*, *8*(3), 143-161. DOI : 10.1007/BF03339671

Roswandowitz, C., Kappes, C., Obrig, H., & von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. *Brain*, *141*(1), 234-247.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, *24*(19), 2348-2353. DOI : 10.1016/j.cub.2014.08.048

Rouvier, M., Bousquet, P.-M., & Favre, B. (2015). *Speaker diarization through speaker embeddings*. DOI : 10.1109/EUSIPCO.2015.7362751

Ruti, M. (2015). *The age of scientific sexism : How evolutionary psychology promotes gender profiling and fans the battle of the sexes*. Bloomsbury Publishing USA.

Ryan, W. J., & Burk, K. W. (1974). Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, *7*(2), 181-192.

Sadjadi, O., Greenberg, C., Singer, E., Mason, L., & Reynolds, D. (2021). NIST 2021 Speaker Recognition Evaluation Plan. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932697

Sataloff, R. T., & Linville, S. E. (2005). The effects of age on the voice.

Saussure, F. d. (1916). Cours de linguistique générale. *C. Bally and A. Sechehaye*.

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory Long term Memory : Repetition Priming of Voice Recognition. *The Quarterly Journal of Experimental Psychology : Section A*, *50*(3), 498-517. DOI : 10.1080/713755724

Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., Von Rueden, C., Krauss, A., & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B : Biological Sciences*, *277*(1699), 3509-3518.

Shipp, T., & Hollien, H. (1969). Perception of the aging male voice. *Journal of Speech and Hearing Research*, *12*(4), 703-710. DOI : 10.1044/jshr.1204.703

Sigüenza, B. G. (2008). BATVOX : sistema automático de reconocimiento de locutor. *Estudios de fonética experimental*, 303-316.

Silverman, D., Blankenship, B., Kirk, P., & Ladefoged, P. (1995). Phonetic structures in Jalapa Mazatec. *Anthropological linguistics*, 70-88.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, *3*(2), 621-640. DOI : 10.1111/j.1749-818X.2009.00125.x

Smith, C. (1999). Marking the boundary : utterance-final prosody in French questions and statements. *Proceedings of the 14th International Congress of Phonetic Sciences*, *5*, 1181-1184.

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America*, *117*(1), 305-318. DOI : 10.1121/1.1828637

Smith, H. M. J., & Baguley, T. (2014). Unfamiliar voice identification : Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students*, *5*(1), 59-68. DOI : 10.5334/jeps.bs

Smyth, R., Jacobs, G., & Rogers, H. (2003). Male voices and perceived sexual orientation : An experimental and theoretical approach. *Language in Society*, *32*(3), 329-350. DOI : 10.1017/S0047404503323024

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. *Interspeech*, *2017*, 999-1003.

Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019). Speaker recognition for multi-speaker conversations using x-vectors. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5796-5800.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors : Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329-5333. DOI : 10.1109/ICASSP.2018.8461375

Spencer, L. E. (1988). Speech Characteristics of Male-to-Female Transsexuals : A Perceptual and Acoustic Study. *Folia Phoniatrica et Logopaedica*, *40*(1), 31-42. DOI : 10.1159/000265881

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, *24*(6), 647-653.

Sullivan, K. P., & Schlichting, F. (2000). Speaker discrimination in a foreign language : first language environment, second language learners. *International Journal of Speech Language and the Law*, *7*(1), 95-112. DOI : 10.1558/sll.2000.7.1.95

Sulpizio, S., Fasoli, F., Antonio, R., Eyssel, F., Paladino, M. P., & Diehl, C. (2020). Auditory gaydar : Perception of sexual orientation based on female voice. *Language and speech*, *63*(1), 184-206. DOI : 10.1177/0023830919828201

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*. DOI : 10.1145/3465416.3483305

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., & Larcher, A. (2021). End-to-End anti-spoofing with RawNet2. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6369-6373. DOI : 10.1109/ICASSP39728.2021.9414234

Takefuta, Y., Jancosek, E. G., & Brunt, M. (2017). A statistical analysis of melody curves in the intonation of American English. In A. Rigault & R. Charbonneau (Éd.), *Proceedings of the seventh International Congress of Phonetic Sciences / Actes du Septième Congrès international des sciences phonétiques* (p. 1035-1039). De Gruyter Mouton. DOI : 10.1515/9783110814750-142

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121-131. DOI : 10.1002/acp.2350010205

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., *et al.*. (2020). Introducing the VoicePrivacy initiative, 1693-1697. DOI : 10.21437/Interspeech.2020-1333

Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of communication disorders*, *42*(5), 324-333. DOI : 10.1016/j.jcomdis.2009.03.001

Tosi, O. (1979). *Voice identification : theory and legal applications*. University Park Press Baltimore.

Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*, *11*.

Troubetzkoï, N. S. (1949). Principes de phonologie. *Paris : Éditions Klincksieck*.

van Bezooijen, R. (1987). Transcription of long-term speech characteristics. *Zeitschrift fur Dialektologie und Linguistik*, *54*, 111-140.

van der Vloed, D. (2016). Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, *85*, 127-130. DOI : 10.1016/j.specom.2016.10.001

van der Vloed, D., Bouten, J., & van Leeuwen, D. A. (2014). NFI-FRITS : A forensic speaker recognition database and some first experiments.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834. DOI : 10.1016/0028-3932(87)90120-5

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985a). Familiar voice recognition : patterns and parameters Part II : Recognition of rate-altered voices. *Journal of Phonetics*, *13*(1), 39-52. DOI : 10.1016/S0095-4470(19)30724-7

Van Lancker, D., Kreiman, K., & Emmorey, K. (1985b). Familiar voice recognition : patterns and parameters Part I : Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19-38. DOI : 10.1016/S0095-4470(19)30723-5

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and cognition*, *1*(2), 185-195. DOI : 10.1016/0278-2626(82)90016-1

van Wallendael, L. R., Surace, A., Parsons, D. H., & Brown, M. (1994). 'Earwitness' voice recognition : Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology*, *8*(7), 661-677. DOI : 10.1002/acp.2350080705

van Leeuwen, D. A., & Brümmer, N. (2007). An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In C. Müller (Éd.), *Speaker Classification I :*

Fundamentals, Features, and Methods (p. 330-353). Springer Berlin Heidelberg. DOI : 10.1007/978-3-540-74200-5_19

Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4052-4056. DOI : 10.1109/ICASSP.2014.6854363

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., Dehak, R., *et al.*. (2020a). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech & Language*, *60*, 101026.

Villalba, J., Zhang, Y., & Dehak, N. (2020b). x-Vectors Meet Adversarial Attacks : Benchmarking Adversarial Robustness in Speaker Verification. *INTERSPEECH*, 4233-4237.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, *37*(3), 328-339.

Whiteside, S. P. (1998). Identification of a speaker's sex : a study of vowels. *Perceptual and Motor Skills*, *86*(2), 579-584. DOI : 10.2466/pms.1998.86.2.579

Wilcox, K. A., & Horii, Y. (1980). Age and changes in vocal jitter. *Journal of Gerontology*, *35*(2), 194-198.

Woehrling, C., & Boula de Mareüil, P. (2006). Identification of regional accents in French : perception and categorization. *Ninth International Conference on Spoken Language Processing*. DOI : 10.21437/Interspeech.2006-303

Wolfe, V. I., Ratusnik, D. L., Smith, F. H., & Northrop, G. (1990). Intonation and Fundamental Frequency in Male-to-Female Transsexuals. *Journal of Speech and Hearing Disorders*, *55*(1), 43-50. DOI : 10.1044/jshd.5501.43

Wu, K., & Childers, D. G. (1991). Gender recognition from speech. Part I : Coarse analysis. *The Journal of the Acoustical society of America*, *90*(4), 1828-1840. DOI : 10.1121/1.401663

Yamagishi, J., Veaux, C., MacDonald, K., *et al.*. (2019). CSTR VCTK Corpus : English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). DOI : 10.7488/DS/2645

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W.,

Watanabe, S., Mohamed, A., & Lee, H.-y. (2021). SUPERB : Speech Processing Universal PERformance Benchmark. DOI : 10.21437/Interspeech.2021-1775

Yarmey, A. D. (2001). Earwitness descriptions and speaker identification. *International Journal of Speech Language and the Law*, *8*(1), 113-122. DOI : 10.1558/sll.2001.8.1.113

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, *8*(5), 453-464. DOI : 10.1002/acp.2350080504

Yuasa, I. P. (2010). Creaky voice : A new feminine voice quality for young urban-oriented upwardly mobile American women ? *American Speech*, *85*(3), 315-337. DOI : 10.1215/00031283-2010-018

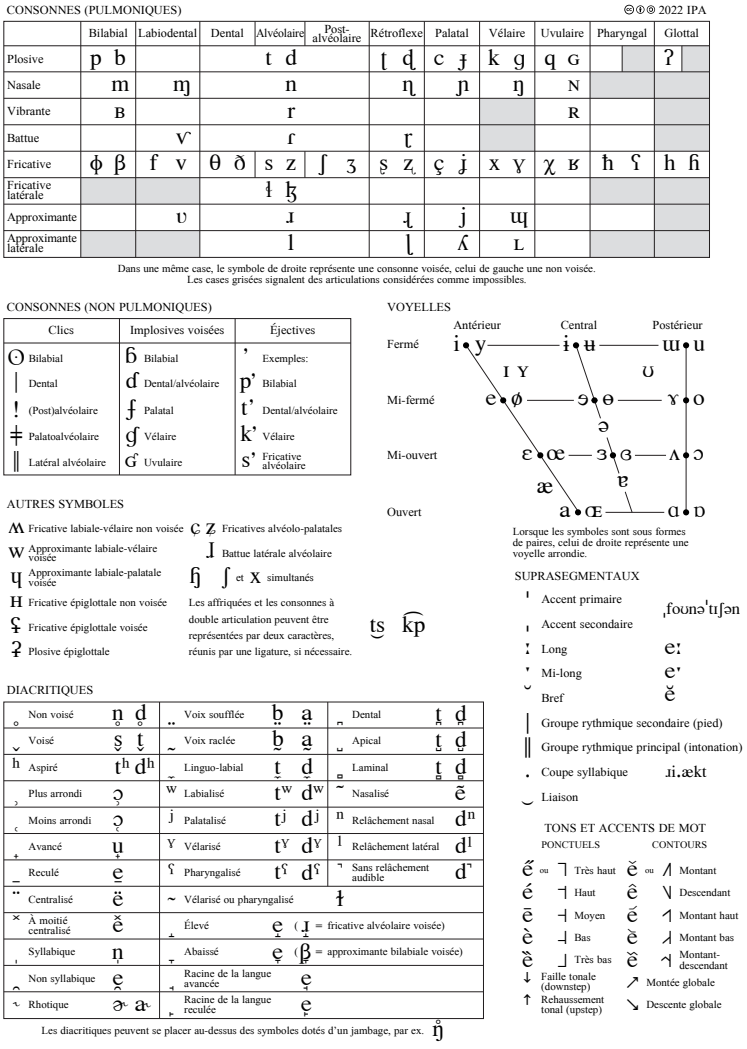
Zhai, T., Li, Y., Zhang, Z., Wu, B., Jiang, Y., & Xia, S.-T. (2021). Backdoor Attack Against Speaker Verification. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2560-2564. DOI : 10.1109/ICASSP39728.2021.9413468

Zhang, C., & Tang, C. (2018). Evaluation of Batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication*, *100*, 13-17. DOI : 10.1016/j.specom.2018.04.008

Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good : The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, *13*(2), 67-82. DOI : 10.1007/BF00990791

Annexes

A | Alphabet phonétique international



CONSONNES (PULMONIQUES) ©①◎ 2022 IPA

	Bilabial	Labiodental	Dental	Alvéolaire	Post-alvéolaire	Rétroflexe	Palatal	Vélaire	Uvulaire	Pharyngal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasale	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Vibrante	ʙ			r					ʀ		
Battue		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Fricative latérale				ɬ ɮ							
Approximante		ʋ		ɹ		ɻ	j	ɰ			
Approximante latérale				l		ɭ	ʎ	ʟ			

Dans une même case, le symbole de droite représente une consonne voisée, celui de gauche une non voisée.
Les cases grisées signalent des articulations considérées comme impossibles.

CONSONNES (NON PULMONIQUES)

Clics	Implosives voisées	Éjectives
ʘ Bilabial	ɓ Bilabial	’ Exemples:
ǀ Dental	ɗ Dental/alvéolaire	pʼ Bilabial
ǃ (Post)alvéolaire	ʄ Palatal	tʼ Dental/alvéolaire
ǂ Palatoalvéolaire	ɠ Vélaire	kʼ Vélaire
ǁ Latéral alvéolaire	ʛ Uvulaire	sʼ Fricative alvéolaire

VOYELLES

Lorsque les symboles sont sous formes de paires, celui de droite représente une voyelle arrondie.

AUTRES SYMBOLES

ʍ Fricative labiale-vélaire non voisée
ɕ ʑ Fricatives alvéolo-palatales
w Approximante labiale-vélaire voisée
ɺ Battue latérale alvéolaire
ɥ Approximante labiale-palatale voisée
ʡ Fricative épiglottale non voisée
ʢ Fricative épiglottale voisée
ʔ Plosive épiglottale
ʃʒ ʃ et ʒ simultanés
Les affriquées et les consonnes à double articulation peuvent être représentées par deux caractères, réunis par une ligature, si nécessaire.
t͡s k͡p

SUPRASEGMENTAUX

ˈ Accent primaire
ˌfoʊnəˈtɪʃən
ˌ Accent secondaire
ː Long eː
ˑ Mi-long eˑ
˘ Bref ĕ
| Groupe rythmique secondaire (pied)
‖ Groupe rythmique principal (intonation)
. Coupe syllabique ɹi.ækt
‿ Liaison

DIACRITIQUES

◌̥ Non voisé	n̥ d̥	◌̤ Voix soufflée	b̤ a̤	◌̪ Dental	t̪ d̪
◌̬ Voisé	s̬ t̬	◌̰ Voix raclée	b̰ a̰	◌̺ Apical	t̺ d̺
◌ʰ Aspiré	tʰ dʰ	◌̼ Linguo-labial	t̼ d̼	◌̻ Laminal	t̻ d̻
◌̹ Plus arrondi	ɔ̹	◌ʷ Labialisé	tʷ dʷ	◌̃ Nasalisé	ẽ
◌̜ Moins arrondi	ɔ̜	◌ʲ Palatalisé	tʲ dʲ	◌ⁿ Relâchement nasal	dⁿ
◌̟ Avancé	u̟	◌ˠ Vélarisé	tˠ dˠ	◌ˡ Relâchement latéral	dˡ
◌̠ Reculé	e̠	◌ˤ Pharyngalisé	tˤ dˤ	◌ˀ Sans relâchement audible	dˀ
◌̈ Centralisé	ë	◌̴ Vélarisé ou pharyngalisé	ɫ		
◌̽ À moitié centralisé	e̽	◌̝ Élevé	e̝ (ɪ̝ = fricative alvéolaire voisée)		
◌̩ Syllabique	n̩	◌̞ Abaissé	e̞ (β̞ = approximante bilabiale voisée)		
◌̯ Non syllabique	e̯	◌̘ Racine de la langue avancée	e̘		
◌˞ Rhotique	ə˞ a˞	◌̙ Racine de la langue reculée	e̙		

Les diacritiques peuvent se placer au-dessus des symboles dotés d’un jambage, par ex. n̊ŋ̊

TONS ET ACCENTS DE MOT

PONCTUELS		CONTOURS	
ő ou ˥	Très haut	ě ou ˩˥	Montant
é ˦	Haut	ê ˥˩	Descendant
ē ˧	Moyen	ě ˨˦	Montant haut
è ˨	Bas	ě ˩˧	Montant bas
ȅ ˩	Très bas	ě ˦˩	Montant-descendant
↓	Faille tonale (downstep)	↗	Montée globale
↑	Rehaussement tonal (upstep)	↘	Descente globale

B | Base de données FABIOLE 2

Locuteurs

Les 399 locuteurs de FABIOLE 2 sont décrits dans le tableau B.1.

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC001	01/01/1962	224	Homme	Française	Journaliste	15290.368
LOC002	15/06/1949	183	Homme	Française	Journaliste	18520.714
LOC003	04/10/1964	140	Femme	Française	Journaliste	9942.137
LOC004	27/10/1979	147	Femme	Française et Libanaise	Journaliste	8950.869
LOC005	01/12/1960	146	Femme	Française	Journaliste	20307.213
LOC006	14/05/1972	139	Femme	Française	Journaliste	13352.737
LOC007	08/09/1958	119	Homme	Française	Journaliste	14607.141
LOC008	05/05/1971	109	Homme	Française	Journaliste	9562.464
LOC009	01/01/1962	61	Homme	Française	Journaliste	5207.144
LOC010	13/04/1964	83	Homme	Française	Journaliste	6679.665
LOC011	18/04/1974	63	Homme	Française	Politicien	20266.768
LOC012	19/12/1964	72	Homme	Française	Journaliste	8461.085
LOC013	02/06/1980	71	Femme	Française	Journaliste	9664.859
LOC014	11/10/1982	64	Homme	Française	Politicien	25912.756
LOC015	05/09/1960	67	Femme	Canadienne	Acteur	28328.864
LOC016		67	Homme	Française	Journaliste	3098.349
LOC017	05/05/1964	56	Homme	Française	Politicien	16261.209
LOC018	17/08/1968	65	Homme	Française	Politicien	8916.749
LOC019	25/01/1967	66	Homme	Française	Journaliste	6264.976
LOC020	02/07/1973	66	Homme	Française	Journaliste	3556.706
LOC021	16/06/1968	61	Homme	Française	Journaliste	21253.694

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC022	07/09/1962	62	Homme	Française	Journaliste	4645.357
LOC023	17/10/1973	64	Homme	Française	Journaliste	3838.905
LOC024	30/01/1965	65	Homme	Française	Journaliste	3476.599
LOC025	03/05/1969	63	Homme	Française	Journaliste	6824.277
LOC026	26/06/1967	58	Homme	Française	Politicien	5287.993
LOC027	21/12/1977	62	Homme	Française	Politicien	4734.227
LOC028	26/09/1963	61	Homme	Française	Journaliste	4231.264
LOC029	24/09/1962	62	Homme	Française	Journaliste	5646.266
LOC030	11/01/1971	60	Homme	Française	Politicien	4911.36
LOC031		62	Femme	Française	Journaliste	4437.945
LOC032		62	Homme	Française	Journaliste	3482.734
LOC033	17/08/1971	57	Homme	Française	Politicien	16027.529
LOC034	08/02/1968	58	Homme	Française	Journaliste	9490.022
LOC035	20/05/1971	60	Homme	Française	Journaliste	5613.035
LOC036	26/08/1951	57	Homme	Française	Journaliste	2579.02
LOC037		49	Homme	Marocaine	Journaliste	15442.004
LOC038	28/02/1974	57	Homme	Française	Journaliste	4998.121
LOC039	03/04/1966	54	Homme	Française	Journaliste	4116.404
LOC040	24/03/1975	59	Femme	Française	Journaliste	4122.007
LOC041	13/08/1962	59	Homme	Française	Politicien	3656.674
LOC042	18/08/1963	57	Homme	Française	Journaliste	2545.08

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC043	04/03/1954	55	Homme	Française	Politicien	6269.464
LOC044	29/05/1960	58	Homme	Française	Journaliste	5855.3
LOC045		58	Femme	Marocaine	Journaliste	3591.844
LOC046	28/12/1972	55	Homme	Française	Journaliste	2994.519
LOC047		53	Homme	Française	Journaliste	2759.392
LOC048	05/08/1968	56	Femme	Française	Politicien	25642.284
LOC049	21/12/1977	55	Homme	Française	Politicien	23388.973
LOC050	19/02/1980	55	Femme	Française	Journaliste	8029.929
LOC051	28/09/1965	51	Homme	Française	Politicien	4212.452
LOC052		55	Homme	Française	Chercheur	3890.349
LOC053	11/06/1978	53	Homme	Française	Politicien	3580.246
LOC054	05/09/1965	56	Homme	Française	Journaliste	3575
LOC055	01/01/1988	56	Homme	Française	Chercheur	2752.964
LOC056	01/07/1957	54	Homme	Française	Journaliste	22824.321
LOC057		55	Homme	Marocaine	Journaliste	22410.993
LOC058	30/12/1958	54	Homme	Française	Politicien	15767.045
LOC059	28/11/1970	53	Homme	Française	Politicien	7954.419
LOC060		53	Femme	Française	Journaliste	6201.942
LOC061	13/05/1967	53	Homme	Française	Économiste	5276.184
LOC062	30/09/1978	55	Femme	Française	Politicien	5446.583
LOC063	12/04/1975	53	Homme	Française	Politicien	3822.13

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC064	01/01/1962	52	Femme	Française	Journaliste	3525.721
LOC065	13/10/1969	55	Homme	Française	Politicien	3461.844
LOC066	08/05/1963	55	Homme	Française	Journaliste	2925.948
LOC067		54	Homme	Française	Journaliste	2423.291
LOC068	03/02/1948	51	Homme	Française	Politicien	10483.057
LOC069	26/02/1963	54	Homme	Française	Journaliste	9141.901
LOC070	15/04/1969	52	Homme	Française	Politicien	5261.9
LOC071	08/05/1969	54	Homme	Française	Journaliste	5090.671
LOC072	04/04/1945	52	Homme	Française	Politicien	3994.542
LOC073	12/10/1976	53	Homme	Française	Journaliste	4031.919
LOC074	01/11/1943	51	Homme	Française	Économiste	3655.415
LOC075		51	Homme	Française	Chercheur	3362.714
LOC076	03/01/1966	54	Homme	Française	Politicien	3485.225
LOC077	01/01/1980	53	Femme	Française	Journaliste	3184.698
LOC078	03/08/1948	52	Homme	Française	Politicien	33915.812
LOC079	01/11/1962	53	Femme	Française	Politicien	31072.875
LOC080	23/05/1990	52	Homme	Française	Politicien	25769.827
LOC081	02/05/1965	51	Homme	Française	Politicien	23424.907
LOC082	03/02/1960	53	Homme	Française	Politicien	22496.142
LOC083	29/01/1956	53	Homme	Française	Politicien	17157.168
LOC084	27/07/1967	47	Homme	Française	Politicien	8127.619

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC085	01/01/1975	50	Homme	Haïtienne et Française	Journaliste	7458.712
LOC086	25/05/1951	50	Homme	Française	Politicien	6446.925
LOC087	31/08/1958	51	Homme	Française	Journaliste	5733.797
LOC088		51	Femme	Française	Journaliste	4742.178
LOC089	15/05/1965	53	Homme	Française	Journaliste	4189.03
LOC090	24/10/1981	52	Homme	Française	Politicien	3836.845
LOC091	30/10/1962	53	Homme	Française	Politicien	3678.731
LOC092	11/06/1968	52	Homme	Française	Politicien	2875.027
LOC093	25/05/1963	53	Homme	Française	Journaliste	2358.408
LOC094		51	Homme	Marocaine	Journaliste	40337.046
LOC095	03/01/1953	51	Homme	Française	Politicien	31532.877
LOC096	13/05/1958	50	Homme	Française	Politicien	27333.784
LOC097	09/04/1952	52	Homme	Française	Politicien	27461.558
LOC098	24/10/1967	52	Homme	Française	Politicien	23542.832
LOC099	03/01/1974	51	Homme	Française	Politicien	21833.571
LOC100		51	Femme	Marocaine	Journaliste	17373.959
LOC101	13/11/1986	50	Femme	Française	Politicien	14363.581
LOC102	01/01/1945	49	Homme	Française	Journaliste	13398.468
LOC103	26/05/1973	52	Femme	Française	Politicien	13006.36
LOC104	31/01/1975	50	Femme	Française	Journaliste	8998.556
LOC105	29/12/1977	52	Homme	Française	Politicien	6487.845

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC106	19/08/1951	52	Homme	Française	Politicien	4621.739
LOC107	18/01/1949	52	Homme	Française	Journaliste	4380.911
LOC108	11/05/1951	52	Femme	Française	Politicien	4201.58
LOC109	05/07/1970	52	Homme	Française	Économiste	3831.497
LOC110	17/12/1962	51	Homme	Française	Journaliste	3331.362
LOC111	28/10/1960	50	Homme	Française	Journaliste	3054.52
LOC112	19/06/1985	50	Femme	Française	Politicien	21709.273
LOC113	11/10/1985	49	Homme	Française	Politicien	7813.78
LOC114	12/06/1956	51	Homme	Française	Psychiatre	6268.423
LOC115	22/09/1953	51	Femme	Française	Politicien	4567.513
LOC116	07/03/1961	48	Homme	Française	Politicien	4301.916
LOC117		48	Homme	Française	Journaliste	3979.634
LOC118	29/03/1950	51	Homme	Française	Journaliste	3706.282
LOC119	30/05/1963	50	Femme	Française	Journaliste	3571.804
LOC120		51	Femme	Française	Journaliste	3597.999
LOC121	23/02/1968	51	Femme	Française	Journaliste	3597.292
LOC122	18/11/1982	51	Femme	Française	Politicien	3369.547
LOC123		51	Homme	Française	Journaliste	2974.852
LOC124	26/07/1937	51	Homme	Française	Psychiatre	2917.13
LOC125	16/10/1967	51	Homme	Française	Journaliste	2476.002
LOC126	08/06/1971	48	Homme	Française	Politicien	25798.319

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC127	01/07/1955	49	Homme	Française	Politicien	23973.796
LOC128	01/01/1965	49	Homme	Marocaine	Journaliste	15083.339
LOC129	18/10/1975	47	Homme	Française	Politicien	11165.523
LOC130	20/11/1960	49	Homme	Française	Politicien	11509.895
LOC131		50	Femme	Française	Journaliste	8823.065
LOC132	14/05/1973	48	Femme	Française	Politicien	8249.22
LOC133	04/12/1964	50	Homme	Française	Politicien	4265.218
LOC134	10/03/1958	50	Femme	Française	Journaliste	4099.74
LOC135	27/11/1965	50	Femme	Française	Politicien	3537.122
LOC136	19/06/1959	50	Femme	Française	Politicien	3371.279
LOC137	25/04/1973	50	Femme	Française	Politicien	3154.853
LOC138	24/02/1971	49	Femme	Française	Politicien	3130.023
LOC139		50	Femme	Française	Journaliste	2980.969
LOC140	01/04/1975	50	Femme	Française	Politicien	2892.081
LOC141	01/01/1966	49	Homme	Française	Journaliste	2719.542
LOC142		50	Homme	Française	Journaliste	2554.384
LOC143		49	Homme	Française	Journaliste	2422.591
LOC144		49	Homme	Française	Journaliste	2263.406
LOC145	05/11/1968	50	Homme	Française	Journaliste	2235.109
LOC146		49	Homme	Française	Journaliste	1898.657
LOC147	11/08/1941	46	Homme	Française	Politologue	4450.231

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC148	05/03/1960	48	Femme	Française	Journaliste	3918.988
LOC149	18/08/1968	46	Homme	Française	Politicien	2987.219
LOC150	12/12/1940	47	Homme	Française et Libanaise	Journaliste	2571.328
LOC151	16/09/1957	49	Homme	Française	Politicien	2426.807
LOC152		49	Femme	Française	Journaliste	2190.052
LOC153	14/07/1967	48	Femme	Française	Politicien	24765.468
LOC154	02/12/1964	42	Homme	Française	Journaliste	3546.584
LOC155	14/09/1951	48	Homme	Française	Politicien	3336.837
LOC156		48	Homme	Française	Journaliste	3330.694
LOC157	27/05/1976	48	Homme	Française	Journaliste	3298.835
LOC158	15/11/1974	48	Femme	Française	Politicien	2781.582
LOC159	05/04/1958	47	Homme	Française	Journaliste	2612.945
LOC160	16/06/1953	47	Homme	Française	Économiste	3088.244
LOC161	27/10/1968	47	Homme	Française	Syndicaliste	2978.33
LOC162	02/02/1967	46	Homme	Française	Journaliste	2539.501
LOC163	03/06/1962	45	Homme	Française	Journaliste	2498.837
LOC164	13/04/1973	47	Homme	Française	Politicien	2567.56
LOC165	22/12/1957	43	Femme	Française	Politicien	14339.919
LOC166	11/03/1957	46	Homme	Française	Politicien	9496.099
LOC167		46	Femme	Française	Journaliste	6464.884
LOC168	15/10/1951	45	Homme	Française	Économiste	4189.23

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC169	06/12/1973	42	Homme	Française	Politicien	3428.1
LOC170		43	Homme	Française	Journaliste	2981.417
LOC171	06/10/1958	45	Homme	Française	Politicien	2479.508
LOC172		39	Homme	Française	Journaliste	8069.815
LOC173	29/04/1957	44	Homme	Française	Politicien	7230.665
LOC174	22/07/1964	43	Homme	Française	Journaliste	3625.172
LOC175	20/06/1947	44	Homme	Française	Politicien	2224.323
LOC176	24/07/1947	41	Homme	Française	Économiste	2075.489
LOC177		43	Homme	Marocaine	Journaliste	14225.354
LOC178	08/10/1973	44	Homme	Française	Journaliste	11554.45
LOC179	01/01/1957	42	Homme	Française	Journaliste	6389.254
LOC180	12/07/1980	44	Femme	Française et Gabonaise	Politicien	5053.093
LOC181	01/04/1974	44	Homme	Française	Météorologiste	2964.735
LOC182	29/07/1970	42	Femme	Française	Journaliste	2350.994
LOC183	17/04/1974	43	Femme	Française	Politicien	6784.647
LOC184	08/09/1979	41	Homme	Française	Politicien	2145.989
LOC185	31/03/1968	42	Homme	Française	Auteur	4435.136
LOC186	25/04/1969	41	Femme	Française	Journaliste	1497.617
LOC187	04/09/1969	41	Homme	Française	Politicien	5026.514
LOC188	30/06/1952	40	Homme	Française	Journaliste	4703.535
LOC189	01/01/1983	39	Homme	Française	Journaliste	7109.096

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC190	02/06/1963	39	Homme	Française	Politicien	5238.981
LOC191		40	Homme	Française	Journaliste	4274.494
LOC192	01/01/1981	39	Femme	Française	Syndicaliste	6660.093
LOC193	01/02/1969	38	Homme	Française	Journaliste	5572.711
LOC194	13/05/1958	37	Homme	Française	Syndicaliste	4307.81
LOC195	30/04/1955	34	Homme	Française	Politicien	2894.248
LOC196	05/03/1955	39	Homme	Française	Politicien	2221.092
LOC197	24/01/1977	37	Homme	Française	Politicien	5057.273
LOC198	01/01/1960	38	Homme	Française	Journaliste	1831.195
LOC199	01/04/1961	36	Homme	Française	Syndicaliste	11554.779
LOC200	31/05/1940	37	Homme	Française	Journaliste	1933.641
LOC201	23/05/1966	36	Homme	Française	Journaliste	10385.514
LOC202		34	Homme	Française	Journaliste	4892.421
LOC203	08/10/1988	35	Homme	Française	Politicien	4464.646
LOC204	31/07/1947	30	Homme	Française	Politicien	3390.697
LOC205	01/11/1951	35	Homme	Française	Acteur	2311.326
LOC206	26/06/1973	35	Femme	Française	Journaliste	2212.204
LOC207	01/01/1952	33	Homme	Française	Journaliste	7620.429
LOC208	31/07/1963	33	Homme	Française	Entrepreneur	3266.23
LOC209	14/01/1948	32	Homme	Française	Journaliste	2775.111
LOC210	19/03/1961	32	Homme	Française	Journaliste	1673.652

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC211	05/08/1981	30	Homme	Française	Journaliste	3802.826
LOC212	01/01/1984	32	Homme	Française	Journaliste	1985.192
LOC213	06/08/1979	32	Femme	Française	Journaliste	1779.913
LOC214	15/04/1949	30	Homme	Française	Essayiste	7938.844
LOC215	06/12/1963	30	Homme	Française	Politicien	3674.514
LOC216	23/07/1955	30	Homme	Française	Journaliste	2961.825
LOC217	01/01/1976	31	Homme	Française	Journaliste	2497.165
LOC218	01/01/1980	31	Homme	Française	Journaliste	2192.028
LOC219		30	Homme	Française	Politicien	1979.885
LOC220		28	Femme	Française	Journaliste	1379.27
LOC221		29	Homme	Française	Journaliste	3935.223
LOC222	12/08/1954	29	Homme	Française	Politicien	3921.681
LOC223		30	Homme	Française	Journaliste	3648.147
LOC224	22/11/1976	30	Femme	Française	Journaliste	2614.937
LOC225	13/06/1975	29	Femme	Française	Politicien	12102.808
LOC226	01/01/1959	27	Homme	Française	Philosophe	6446.706
LOC227		27	Homme	Française	Journaliste	5683.886
LOC228	02/09/1939	28	Homme	Française	Politicien	4681.853
LOC229		27	Femme	Française	Journaliste	3201.344
LOC230		29	Femme	Française	Journaliste	2518.511
LOC231		28	Femme	Française	Journaliste	1914.955

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC232		28	Homme	Française	Journaliste	1678.571
LOC233	17/04/1961	28	Homme	Française	Politicien	4955.068
LOC234		28	Femme	Française	Journaliste	4053.922
LOC235	30/08/1969	25	Homme	Française	Journaliste	1377.237
LOC236	13/09/1995	26	Homme	Française	Politicien	3820.607
LOC237	01/01/1982	27	Femme	Française	Journaliste	2117.427
LOC238	17/03/1970	27	Homme	Française	Journaliste	1959.533
LOC239	08/08/1961	25	Homme	Française	Journaliste	922.611
LOC240	04/02/1947	27	Homme	Française	Journaliste	913.412
LOC241	21/05/1968	26	Homme	Française	Essayiste	3872.563
LOC242	15/05/1969	26	Femme	Française	Économiste	3573.564
LOC243	02/05/1950	25	Homme	Française	Politologue	2870.256
LOC244	01/01/1975	26	Homme	Française	Journaliste	2780.132
LOC245		25	Homme	Française	Journaliste	2229.069
LOC246		25	Homme	Française	Journaliste	1751.705
LOC247	15/04/1982	23	Homme	Française	Entrepreneur	5851.074
LOC248	18/02/1942	23	Homme	Française	Homme d’affaires	3422.213
LOC249	01/01/1963	24	Femme	Française	Journaliste	2912.38
LOC250		23	Femme	Française	Journaliste	1586.283
LOC251		23	Homme	Française	Médecin	1398.015
LOC252	02/09/1951	23	Homme	Française	Journaliste	4659.01

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC253	29/04/1972	23	Femme	Française	Journaliste	2037.423
LOC254	17/04/1967	22	Homme	Française	Politicien	1635.172
LOC255		23	Homme	Française	Journaliste	1265.688
LOC256	02/06/1966	20	Homme	Française	Journaliste	3264.972
LOC257		21	Femme	Française	Journaliste	3139.068
LOC258	06/12/1960	22	Homme	Française	Journaliste	3038.574
LOC259	16/08/1978	21	Homme	Française	Politicien	2900.412
LOC260	21/05/1983	22	Homme	Française	Journaliste	2237.852
LOC261	01/01/1987	21	Homme	Française	Journaliste	1997.034
LOC262	10/11/1961	21	Homme	Française	Journaliste	1747.24
LOC263		20	Homme	Française	Journaliste	1260.314
LOC264		22	Homme	Française	Journaliste	1102.573
LOC265	23/10/1966	21	Homme	Française	Politicien	3645.93
LOC266	01/01/1962	19	Homme	Française	Journaliste	2740.867
LOC267		19	Homme	Française	Journaliste	3017.112
LOC268	28/07/1964	20	Homme	Française	Journaliste	2424.503
LOC269	26/08/1960	21	Homme	Française	Journaliste	1906.684
LOC270	19/01/1951	21	Homme	Française	Médecin	1707.702
LOC271	30/06/1949	19	Homme	Française	Philosophe	1464.453
LOC272	14/11/1953	21	Homme	Française	Politicien	1310.074
LOC273	26/04/1956	20	Homme	Française	Journaliste	4070.978

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC274	02/07/1972	20	Homme	Française	Journaliste	2451.781
LOC275		20	Femme	Française	Journaliste	1134.581
LOC276		20	Homme	Française	Journaliste	932.164
LOC277		19	Homme	Marocaine	Journaliste	12705.179
LOC278	14/09/1952	19	Homme	Française	Journaliste	2133.663
LOC279	26/06/1966	17	Homme	Française	Acteur	1593.031
LOC280	15/04/1975	18	Femme	Française	Journaliste	1662.613
LOC281	19/08/1963	18	Homme	Française	Médecin	2199.485
LOC282	25/04/1969	18	Homme	Française	Journaliste	1695.897
LOC283	11/11/1982	15	Homme	Française	Journaliste	916.396
LOC284	19/04/1970	18	Homme	Française	Journaliste	806.049
LOC285	28/07/1960	17	Homme	Française	Journaliste	4192.542
LOC286	23/04/1980	16	Homme	Française	Journaliste	1955.908
LOC287	17/05/1977	16	Homme	Française	Journaliste	1641.694
LOC288	03/08/1959	15	Homme	Française	Journaliste	1589.657
LOC289		17	Homme	Française	Journaliste	1559.572
LOC290		15	Homme	Française	Entrepreneur	3247.86
LOC291		16	Homme	Française	Auteur	2013.686
LOC292		16	Femme	Française	Journaliste	1588.401
LOC293	04/09/1963	16	Homme	Française	Journaliste	1238.167
LOC294	07/07/1948	16	Homme	Française	Journaliste	1062.862

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC295	27/08/1936	15	Homme	Française	Journaliste	1864.939
LOC296	24/02/1960	14	Femme	Française	Journaliste	1283.108
LOC297	15/08/1945	14	Homme	Française	Politicien	775.78
LOC298		14	Femme	Française	Journaliste	2496.443
LOC299	13/01/1959	14	Homme	Française	Journaliste	1978.277
LOC300		12	Homme	Française	Journaliste	1457.72
LOC301	23/01/1979	14	Femme	Française	Journaliste	1369.868
LOC302	01/01/1987	14	Homme	Française	Journaliste	1202.663
LOC303	28/01/1955	13	Homme	Française	Politicien	1841.665
LOC304	07/10/1964	13	Homme	Française	Journaliste	1746.182
LOC305	01/01/1978	12	Femme	Française	Journaliste	1397.659
LOC306	05/11/1948	13	Homme	Française	Philosophe	620.139
LOC307		12	Homme	Marocaine	Politologue	9035.98
LOC308	05/04/1980	12	Homme	Française	Politicien	1390.94
LOC309		12	Femme	Française	Journaliste	1265.835
LOC310		12	Femme	Française	Journaliste	1115.273
LOC311		11	Homme	Française	Journaliste	475.576
LOC312	01/01/1970	10	Femme	Française	Journaliste	1577.565
LOC313	01/01/1973	8	Homme	Française	Journaliste	1022.673
LOC314		10	Homme	Française	Journaliste	988.76
LOC315	01/01/1969	10	Homme	Française	Journaliste	890.949

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC316	31/12/1977	10	Femme	Belge	Journaliste	777.15
LOC317	20/06/1928	9	Homme	Française	Journaliste	520.621
LOC318	13/10/1971	10	Homme	Française	Journaliste	611.663
LOC319		10	Femme	Française	Journaliste	528.762
LOC320	01/06/1959	9	Femme	Française	Météorologiste	438.739
LOC321		9	Homme	Marocaine	Politologue	6256.224
LOC322		9	Homme	Marocaine	Économiste	6113.498
LOC323	15/07/1968	9	Homme	Sénégalaise	Joueur de football	2351.008
LOC324	03/02/1970	14	Homme	Française	Journaliste	2708.359
LOC325	01/01/1973	9	Homme	Française	Journaliste	1477.382
LOC326		9	Homme	Française	Journaliste	632.568
LOC327		9	Homme	Française	Journaliste	612.175
LOC328		9	Femme	Française	Journaliste	529.921
LOC329	27/10/1950	8	Homme	Française	Journaliste	1357.986
LOC330		8	Femme	Française	Journaliste	1336.88
LOC331	25/08/1985	8	Femme	Française	Journaliste	721.114
LOC332		8	Femme	Française	Journaliste	615.577
LOC333		8	Homme	Française	Journaliste	596.306
LOC334		8	Homme	Française	Journaliste	520.057
LOC335	27/03/1985	8	Femme	Française	Journaliste	492.561
LOC336		8	Femme	Française	Journaliste	459.953

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC337	01/01/1970	8	Femme	Française	Journaliste	406.308
LOC338	18/11/1991	8	Femme	Française	Journaliste	323.188
LOC339		7	Homme	Marocaine	Chercheur	5197.682
LOC340		7	Homme	Marocaine	Journaliste	4435.37
LOC341		7	Homme	Marocaine	Journaliste	3421.495
LOC342		7	Femme	Française	Journaliste	2039.173
LOC343		7	Femme	Française	Journaliste	1387.138
LOC344		7	Femme	Française	Journaliste	1033.165
LOC345	16/09/1956	7	Homme	Française	Journaliste	445.625
LOC346		6	Femme	Française	Journaliste	248.075
LOC347		6	Femme	Marocaine	Psychologue	3637.908
LOC348		6	Homme	Marocaine	Entrepreneur	3189.182
LOC349	01/01/1973	5	Homme	Française	Journaliste	1409.777
LOC350		6	Homme	Française	Journaliste	843.703
LOC351		5	Femme	Française	Journaliste	635.69
LOC352	12/04/1976	6	Homme	Française	Politicien	746.665
LOC353	01/01/1983	6	Femme	Française	Journaliste	669.836
LOC354	16/03/1989	5	Homme	Française	Politicien	441.081
LOC355		6	Homme	Française	Journaliste	447.828
LOC356		6	Homme	Française	Journaliste	388.602
LOC357	21/03/1965	6	Homme	Française	Politicien	309.492

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC358		5	Femme	Marocaine	Journaliste	2929.451
LOC360		4	Homme	Française	Journaliste	630.766
LOC361		5	Femme	Française	Journaliste	621.165
LOC362		5	Femme	Française	Journaliste	580.471
LOC363		5	Homme	Française	Journaliste	479.215
LOC364	14/06/1986	5	Femme	Française	Journaliste	437.569
LOC365	01/01/1995	4	Homme	Française	Journaliste	356.717
LOC366	01/01/1980	5	Homme	Française	Journaliste	359.496
LOC367	01/01/1986	5	Femme	Française	Journaliste	343.762
LOC368	01/01/1985	5	Homme	Française	Journaliste	292.736
LOC369	18/02/1978	5	Femme	Française et Marocaine	Politicien	282.598
LOC370		4	Homme	Marocaine	Professeur	3093.255
LOC371		4	Homme	Française	Chercheur	2540.149
LOC372		4	Homme	Marocaine	Psychologue	2444.137
LOC373		4	Femme	Marocaine	Médecin	2408.264
LOC374		4	Femme	Marocaine	Psychologue	2202.158
LOC375		4	Femme	Marocaine	Psychologue	1685.365
LOC376	16/05/1970	4	Femme	Marocaine	Psychologue	1577.044
LOC377		4	Femme	Marocaine	Coach	1494.708
LOC378	25/03/1969	4	Homme	Espagnole et Marocaine	Joueur de football	1303.523
LOC379		4	Femme	Française	Journaliste	1086.135

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée
LOC380	27/01/1947	4	Homme	Française	Journaliste	1073.217
LOC381	14/08/1966	4	Homme	Française	Politicien	1011.937
LOC382		4	Femme	Française	Journaliste	983.755
LOC383		3	Femme	Française	Journaliste	981.244
LOC384		4	Femme	Française	Journaliste	952.877
LOC385	14/08/1971	3	Homme	Française	Chercheur	665.445
LOC386	28/06/1963	4	Homme	Française	Politicien	715.743
LOC387		4	Homme	Française	Journaliste	683.066
LOC388		4	Femme	Française	Journaliste	679.118
LOC389		4	Homme	Française	Journaliste	644.664
LOC390	26/04/1976	4	Femme	Française	Politicien	619.953
LOC391		4	Femme	Française	Journaliste	614.125
LOC392		4	Homme	Française	Journaliste	561.395
LOC393	01/01/1974	4	Homme	Française	Journaliste	530.164
LOC394		4	Homme	Française	Journaliste	521.849
LOC395		4	Femme	Française	Journaliste	497.743
LOC396	11/09/1973	4	Femme	Française	Météorologiste	472.077
LOC397	12/07/1966	4	Homme	Française	Journaliste	454.105
LOC398		4	Femme	Française	Journaliste	434.922
LOC399	30/09/1968	3	Homme	Française	Joueur de football	3160.82
LOC400	26/05/1984	4	Homme	Française	Politicien	415.584

Locuteur	Date de naissance	Contextes	Genre	Nationalité	Profession	Durée

Tableau B.1 – *Informations sur les locuteurs de FABIOLE 2*

C | Base de données PTSVOX

Lieux de résidence

Les lieux de résidence des locuteurs de PTSVOX sont répartis ainsi que présenté dans le tableau C.1.

État de santé au moment de l'enregistrement

Avant chaque session, les locuteurs devaient indiquer si leur voix était susceptible d'être affectée par une quelconque condition ou état de santé. L'interrogatoire a montré que :

- 130 locuteurs ont déclaré fumer ;
- 89 locuteurs ont dit être malades le jour de l'enregistrement (nez bouché, toux, mal de gorge) ;
- 20 locuteurs ont indiqué avoir eu recours à de l'orthophonie ;
- 7 locuteurs ont subi une opération dans la zone ORL.

Textes lus

Région/Country	Locuteurs
France métropolitaine	
Provence-Alpes-Côte d'Azur	126
Auvergne-Rhône-Alpes	119
Occitanie	66
Île-de-France	27
Grand Est	20
Nouvelle-Aquitaine	12
Hauts-de-France	11
Bourgogne-Franche-Comté	9
Corse	9
Pays de la Loire	6
Normandie	5
Bretagne	4
Centre-Val de Loire	1
Départements et territoires d'outre-mer	
La Réunion	14
Guyane	10
Mayotte	8
Guadeloupe	7
Martinique	4
Tahiti	1
France (région inconnue)	28
Europe	
Allemagne	4
Royaume-Uni	2
Autriche	1
Belgique	1
Espagne	1
Irlande	1
Portugal	1
Ukraine	7
Roumanie	1
Russie	1
Afrique	
Côte d'Ivoire	1
Tunisie	1
Amériques	
Brésil	1
États-Unis	1
République dominicaine	1
Asie	
Chine	1
Turquie	1
Asie (pays inconnu)	1

Tableau C.1 – *Lieux de résidence des locutrices et locuteurs*

Texte 1

Au nord du pays, on trouve une espèce de chats dont la queue est très courte. Ils sont noirs avec deux taches blanches sur le dos. Leur poil est beau et doux. Juste à côté vit une colonie d'oiseaux dont les nids sont accrochés au bord de la falaise. Ils doivent faire attention à ne pas faire tomber leurs œufs dans la mer. Ma sœur n'a qu'à traverser la rue pour rencontrer ces deux espèces vivant en harmonie au cœur d'un parc naturel. Régulièrement, sur le coup de midi, après avoir pris un bon thé, nous sortons de chez elle pour aller observer ces animaux.

Texte 2

Ma sœur est venue chez moi hier pour prendre le thé. Elle me parlait de ses vacances en mer du Nord, lorsque, dans notre dos, tomba un petit oiseau. Ses deux ailes étaient blessées, et il avait reçu un coup violent sur la queue ; son cœur battait très vite, mais il était en vie. Son plumage était beau et doux. Je m'approchais du bord de la fenêtre pour regarder dans la rue. Un chat s'éloignait d'un nid perché sur un arbre. Il avait dû faire fuir l'oiseau après l'avoir attaqué.

Texte 3

La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort, quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau. Ils sont tombés d'accord que celui qui arriverait le premier à faire ôter son manteau au voyageur serait regardé comme le plus fort. Alors, la bise s'est mise à souffler de toute sa force, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui et à la fin, la bise a renoncé à le lui faire ôter. Alors le soleil a commencé à briller et au bout d'un moment, le voyageur, réchauffé a ôté son manteau. Ainsi, la bise a dû reconnaître que le soleil était le plus fort des deux.

D | Système ECAPA-TDNN - Recette SpeechBrain

Hyperparamètres utilisés

```
# Feature parameters
n_mels: 80

# Pretrain folder (HuggingFace)
pretrained_path: speechbrain/spkrec-ecapa-voxceleb

# Output parameters
out_n_neurons: 7205

# Model params
compute_features: !new:speechbrain.lobes.features.Fbank
    n_mels: !ref <n_mels>

mean_var_norm: !new:speechbrain.processing.features.InputNormalization
    norm_type: sentence
    std_norm: False

embedding_model: !new:speechbrain.lobes.models.ECAPA_TDNN.ECAPA_TDNN
    input_size: !ref <n_mels>
    channels: [1024, 1024, 1024, 1024, 3072]
    kernel_sizes: [5, 3, 3, 3, 1]
```

```
    dilations: [1, 2, 3, 4, 1]
    attention_channels: 128
    lin_neurons: 192

classifier: !new:speechbrain.lobes.models.ECAPA_TDNN.Classifier
    input_size: 192
    out_neurons: !ref <out_n_neurons>

mean_var_norm_emb: !new:speechbrain.processing.features.InputNormalization
    norm_type: global
    std_norm: False

modules:
    compute_features: !ref <compute_features>
    mean_var_norm: !ref <mean_var_norm>
    embedding_model: !ref <embedding_model>
    mean_var_norm_emb: !ref <mean_var_norm_emb>
    classifier: !ref <classifier>

label_encoder: !new:speechbrain.dataio.encoder.CategoricalEncoder

pretrainer: !new:speechbrain.utils.parameter_transfer.Pretrainer
    loadables:
        embedding_model: !ref <embedding_model>
        mean_var_norm_emb: !ref <mean_var_norm_emb>
        classifier: !ref <classifier>
        label_encoder: !ref <label_encoder>
    paths:
        embedding_model: !ref <pretrained_path>/embedding_model.ckpt
        mean_var_norm_emb: !ref <pretrained_path>/mean_var_norm_emb.ckpt
        classifier: !ref <pretrained_path>/classifier.ckpt
        label_encoder: !ref <pretrained_path>/label_encoder.txt
```