Parole de locuteur : performance et confiance en identification biométrique vocale

Juliette Kahn

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « Sciences et Agrosciences »

Laboratoire Informatique d'Avignon (EA 931)

Parole de locuteur : performance et confiance en identification biométrique vocale

par

Juliette KAHN

Soutenue publiquement le 19 décembre 2011 devant un jury composé de :

M^{me}	Régine ANDRE-OBRECHT	Professeur, IRIT, Toulouse	Rapporteur
M.	Anders ERIKSSON	Professeur, Université de Göteborg, Göteborg (Suède)	Rapporteur
M^{me}	Martine ADDA-DECKER	Professeur, LPP, Paris	Présidente du jury
M.	Edouard GEOFFROIS	Ingénieur, DGA, Paris	Examinateur
M^{me}	Solange ROSSATO	Maître de Conférences, LIG, Grenoble	Co-encadrante
M.	Jean-François BONASTRE	Professeur, LIA, Avignon	Directeur de thèse

---

Remerciements

Le travail que vous vous apprêtez à lire est le fruit de nombreuses rencontres et échanges. L'aventure n'aurait pas été aussi grisante, positive et constructive sans toutes les possibilités qui m'ont été offertes par de nombreuses personnes que je voudrais remercier ici.

Mes premiers remerciements s'adressent aux membres de mon jury qui m'ont fait l'honneur de me consacrer un peu de leur temps.
Je souhaite remercier Martine Adda-Decker qui a bien voulu présider ce jury. Ses travaux sur l'utilisation de grand corpus pour la caractérisation de la parole sont un véritable exemple pour moi.
Je suis très reconnaissante à Régine André-Obrecht et à Anders Erikson qui ont accepté d'être rapporteurs de ce travail. Merci pour votre lecture constructive et vos suggestions.
Je tiens à remercier Edouard Geoffroy pour sa relecture rigoureuse du manuscrit.
Merci à Jean-François Bonastre et Solange Rossato pour leurs encouragements de tous les instants et leur rigueur scientifique. Merci de m'avoir fait assez confiance pour me laisser cet espace de liberté dont j'avais besoin. Ces trois années ont été riches de travaux et d'émotions et c'est ce qui les a rendu aussi précieuses. Merci encore.

Un travail de thèse n'est pas possible sans un financement solide. Je tiens à remercier la Direction Générale de l'Armement pour avoir financer ce travail de thèse et m'avoir laissé toute liberté d'initiative.

J'ai eu la chance, pendant ces trois ans de me rendre dans trois laboratoires différents

pour y travailler. Cela m'a permis d'appréhender mon sujet par différents prismes et je tenais à remercier ici toutes les personnes qui m'ont accueillie et qui ont accepté de travailler avec moi.
Je tenais tout d'abord à remercier les directeurs successifs du LIA, Marc Elbèze et Georges Linarès pour leur politique de laboratoire qui permet aux doctorants d'entreprendre de nombreux projets. Merci à l'Equipe parole et plus particulièrement à ceux qui travaillent sur la problématique du locuteur : Corinne Fredouille, Driss Matrouf, Christophe Levy, Anthony Larcher, Pierre-Michel Bousquet, Gilles Pouchoulin pour leur patience et la place qu'ils m'ont laissée pour appréhender le locuteur par un autre chemin. Merci à Nicolas Audibert pour sa patience, sa gentillesse et sa rigueur. Tu m'as vraiment trop supportée ! ! !
Je tenais également à remercier les chercheurs du LIG et en particulier les membres de l'équipe GETALP. Merci à Laurent Besacier, Véronique Aubergé, Michel Vacher, François Portet, Benjamin Lecouteux, Pédro, Marta, Marion et Bassam pour leur accueil du second semestre...
Merci à Nicolas Scheffer qui m'a invitée à venir 3 mois au SRI en Californie. Son accueil et sa disponibilité m'ont permise de prendre conscience que je pouvais faire certaines choses loin des premiers laboratoires où j'ai travaillé. I want to thank the people of the STAR Laboratory in particular Kristin Precoda, Dimitra Vergyri, Lise Schriberg, Luciana Ferrer, Martin Graciarena and Yun Lei for their wholehearted welcome.
Durant mon séjour californien j'ai pu également me rendre à ICSI, Berkeley pour une discussion très enrichissante avec ses chercheurs. Thank you to Nikki Mirghafori and Lara Stoll for their valuable advices.
Tout au long de ma thèse, j'ai pu me rendre dans divers congrès où j'ai eu l'occasion de discuter avec des chercheurs sans qui ce manuscrit ne serait pas ce qu'il est. I want to thank Joe Campbell, Alvin Martin, Craig Greenberg and Georges Doddington for our discussions about my work despite my English.

Les trois ans pendant lesquels j'ai pu travailler sur ce sujet de thèse ont également été l'occasion de monter de nombreux projets pour essayer de rapprocher les doctorants et chercheurs des différentes disciplines qui s'intéresse à la parole.
Je tiens à remercier tous les organisateurs de MajecSTIC et des RJCP 2009 : l'AFCP et le comité de pilotage de MajecSTIC pour leur confiance, Claire Petiteau, Pierre Gotab, Raphaël Rubino, Florian Verdet, Christophe Servan, Rémi Lavalley, Marie-Jean Meurs,

Mickaël Rouvier, Florian Pinault, Gregory Senay pour tous ces moments partagés.
Un très grand merci aux personnes qui ont participé avec moi à la mise en place des JEFP, en particulier Laurianne Georgeton, Cécile Fougeron, Cédric Gendrot, Emmanuel Ferragne, Pierre Gotab et Audrey Acher.

Cette thèse se termine positivement également car j'ai pu continuer à travailler sur des questions d'évaluation de systèmes de Traitement Automatique de la Langue au sein du LNE : Merci à Ludovic Quintard, Olivier Galibert et Romuald Gorjup pour m'avoir laissé le temps de terminer ce travail.

Ce travail n'aurait pas pu aboutir sans les personnes que j'ai rencontré pendant mes études à Grenoble.
Merci aux membres du DIP de l'Université Stendhal, en particulier Georges Antoniadis, Thomas Lebarbé, Olivier Kraif, Claude Ponton, Virginie Zampa qui m'ont démontré, lors de mes études de Master, l'apport que pouvait avoir le TAL sur la Linguistique.
Merci également aux membres de l'équipe SLD du Gipsa-Lab, en particulier Nathalie Vallée, Elisabetta Carpitelli, Jean-Pierre Lay, Carole Chauvin pour m'avoir encouragéé depuis tant d'année.

Une thèse est un grand marathon impossible à réaliser sans le soutien de ses amis, merci donc à Isabelle R, Stéphane, Auriane, Gom, Mimi, Delphine, Alexis, Hélène, Mel, Isabelle E, Vannina, Raph, Rémi, Hug, Nico A, Thierry pour tous ces moments partagés, où on a bien rigolé, discuté et refait le monde...
Merci à ma famille qui fait ce que je suis aujourd'hui. Une bise à chacun d'entre vous.
Enfin merci à toi, Nico, pour ton soutien et ton amour.

Résumé

Ce travail de thèse explore l'usage biométrique de la parole dont les applications sont très nombreuses (sécurité, environnements intelligents, criminalistique, surveillance du territoire ou authentification de transactions électroniques). La parole est soumise à de nombreuses contraintes fonction des origines du locuteur (géographique, sociale et culturelle) mais également fonction de ses objectifs performatifs. Le locuteur peut être considéré comme un facteur de variation de la parole, parmi d'autres. Dans ce travail, nous présentons des éléments de réponses aux deux questions suivantes :

- Tous les extraits de parole d'un même locuteur sont-ils équivalents pour le reconnaître ?
- Comment se structurent les différentes sources de variation qui véhiculent directement ou indirectement la spécificité du locuteur ?

Nous construisons, dans un premier temps, un protocole pour évaluer la capacité humaine à discriminer un locuteur à partir d'un extrait de parole en utilisant les données de la campagne NIST-HASR 2010. La tâche ainsi posée est difficile pour nos auditeurs, qu'ils soient naïfs ou plus expérimentés. Dans ce cadre, nous montrons que ni la (quasi) unanimité des auditeurs ni l'auto-évaluation de leurs jugements ne sont des gages de confiance dans la véracité de la réponse soumise.

Nous quantifions, dans un second temps, l'influence du choix d'un extrait de parole sur la performance des systèmes automatiques. Nous avons utilisé deux bases de données, NIST et BREF ainsi que deux systèmes de RAL, ALIZE/SpkDet (LIA) et Idento (SRI). Les systèmes de RAL, aussi bien fondés sur une approche UBM-GMM que sur une approche i-vector montrent des écarts de performances importants mesurés à l'aide d'une mesure de variation relative autour de l'EER moyen, Vr (pour NIST, $Vr_{Idento} = 1.41$ et $Vr_{ALIZE/SpkDet} = 1.47$ et pour BREF, $Vr = 3.11$) selon le choix du fichier d'apprentissage utilisé pour chaque locuteur. Ces variations de performance, très importantes, montrent la sensibilité des systèmes automatiques au choix des extraits de parole, sensibilité qu'il

est important de mesurer et de réduire pour rendre les systèmes de RAL plus fiables.
Afin d'expliquer l'importance du choix des extraits de parole, nous cherchons les indices les plus pertinents pour distinguer les locuteurs de nos corpus en mesurant l'effet du facteur Locuteur sur la variance des indices (η^2). La F0 est fortement dépendante du facteur Locuteur, et ce indépendamment de la voyelle. Certains phonèmes sont plus discriminants pour le locuteur : les consonnes nasales, les fricatives, les voyelles nasales, voyelles orales mi-fermées à ouvertes.
Ce travail constitue un premier pas vers une étude plus précise de ce qu'est le locuteur aussi bien pour la perception humaine que pour les systèmes automatiques. Si nous avons montré qu'il existait bien une différence cepstrale qui conduisait à des modèles plus ou moins performants, il reste encore à comprendre comment lier le locuteur à la production de la parole. Enfin, suite à ces travaux, nous souhaitons explorer plus en détail l'influence de la langue sur la reconnaissance du locuteur. En effet, même si nos résultats indiquent qu'en anglais américain et en français, les mêmes catégories de phonèmes sont les plus porteuses d'information sur le locuteur, il reste à confirmer ce point et à évaluer ce qu'il en est pour d'autres langues.

Table des matières

Abréviations et notations

Nous listons ici les abréviations utilisées dans ce document. Nous utilisons les abréviations anglaises lorsqu'elles sont l'usage courant dans la communauté francophone.

ACP	Analyse en Composante Principale
DCT	Discrete Cosine Transform
DET	Detection Error Tradeoff
EER	Equal Error Rate
F0	Fréquence fondamentale
F1	Premier formant
F2	Deuxième formant
F3	Troisième formant
F4	Quatrième formant
FA	Fausse acceptation
FFT	Fast Fourier Transform
FR	Faux Rejets
GETALP	Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole
GMM	Gaussian Mixture Model
GSCC	Glottal Source Cepstral Coefficients
HASR	Human Assisted Speaker Recognition
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LFA	Latent Factor Analysis
LFCC	Linear Frequency Cepstral Coefficients
LIA	Laboratoire Informatique d'Avignon
LIG	Laboratoire Informatique de Grenoble

LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
NAQ	Normalized Amplitude Quotient
NIST	National Institute of Standard and Technology
NIST-SRE	NIST Speaker Recognition Evaluation
n.s	Non significatif
PLP	Prédiction Linéaire Perceptuelle ou Perceptual Linear Predictive
RAL	Reconnaissance Automatique du Locuteur
Shim.	Shimmer
SRI	Stanford Research Institute
SVM	Support Vector Machine
signif.	Significativité
UBM	Universal Background Model

Pour faciliter la lecture du manuscrit, la notation des nombres suit les règles suivantes.

- Les décimales sont séparées de la partie entière par un point.
- Le passage au millier, million et milliard est signalé par un blanc.

Introduction

L'idée selon laquelle il est possible de reconnaître quelqu'un à partir de sa voix est très répandue. Dès qu'un auditeur entend une personne parler, il peut y associer certaines caractéristiques comme son genre, son âge ou encore son origine géographique ou sociale, voire lui attribuer des traits de personnalité : la parole est toujours incarnée. Dans son cours de linguistique générale, Saussure (de Saussure, 1916) qualifie la parole de « partie individuelle du langage », en opposition avec la langue qu'il définit comme « un produit social de la faculté de langage et un ensemble de conventions nécessaires ». Cette dichotomie permet de séparer la langue et de ses instanciations particulières, incarnées, produites par un individu dans des circonstances spécifiques : la parole.
En 1962, Kersta affirme qu'il est possible d'identifier une personne à partir d'un enregistrement de sa parole, il utilise le terme « d'empreinte vocale » (Kersta, 1962). Cette empreinte est en réalité une représentation par spectrogramme du signal acoustique. Bolt *et al*, en réponse à Kersta, montrent la mutabilité des spectrogrammes et la non-pertinence de la métaphore proposée (Bolt et al., 1973). En effet, tandis que les empreintes digitales et les empreintes génétiques sont « une image du corps dont on peut effectuer un traitement statistique pour les enregistrer en base de données » (Galton, 1892), la parole, si elle est bien issue du corps, n'en est pas uniquement l'image et ne peut se résumer à une production biomécanique résultant du mouvement des articulateurs. En effet, des pans entiers de la linguistique s'intéressent aux variations diastratiques, diatopiques et diaphasiques qui s'observent dans la parole (Ducrot et Schaeffer, 1995). La sociolinguistique variationniste étudie, depuis Labov (Labov, 1972), l'influence de l'origine sociale des locuteurs sur leur façon de parler. La géolinguistique correspond à « la caractérisation des parlers en rapport avec leur localisation ». La pragmatique, en se focalisant sur les « usages du langage » (Austin, 1970), met en évidence que la parole est performative et contextualisée (Moeschler et Reboul, 1994). En fonction de ce que l'auditeur veut dire et faire faire à son interlocuteur, il utilise un mode d'expres-

sion différent dans lequel l'état émotionnel du locuteur joue également un rôle (Scherer, 1984). La parole est donc soumise à de nombreuses contraintes fonction des origines du locuteur (géographique, sociale et culturelle) mais également fonction de ses objectifs performatifs. Ainsi la parole ne répond pas aux critères de donnée biométrique tels qu'entendus pour les empreintes digitales ou génétiques. Pourtant, le locuteur peut être considéré comme un facteur de variation de la parole, parmi d'autres. La variation idiosyncratique ne peut s'abstraire de l'origine sociale et géographique du locuteur et est influencée par le contexte : chaque enregistrement de la voix d'un locuteur s'effectue dans un contexte d'énonciation donné.
Deux questions émergent de ce constat et constituent l'objet de ce travail de thèse.

- Tous les extraits de parole d'un même locuteur sont-ils équivalents pour le reconnaître ? Il s'agit de quantifier la possibilité de reconnaissance d'un locuteur à partir d'un extrait de parole.
- Comment se structurent les différentes sources de variation qui véhiculent directement ou indirectement la spécificité du locuteur ? Les indices idiosyncratiques proposés étant souvent étudiés pour décrire d'autres phénomènes de variation, la question de l'influence relative du locuteur sur ces différents indices se pose.

Issus de la reconnaissance des formes, les outils de vérification du locuteur sont programmés pour établir si deux enregistrements de parole ont été prononcés par le même locuteur. Les performances de ces systèmes, évaluées par des campagnes comme celles du National Institut of Standard and Technology (NIST) (NIST, 2011), ont beaucoup progressé cette dernière décennie. Ces outils peuvent sembler sûrs dans le sens où leurs taux d'erreurs sont très bas. Cependant, lors de ces campagnes d'évaluation, se pose rarement la question de l'influence des extraits de parole utilisés. La prise en compte de la variabilité due au contexte d'énonciation est souvent négligée car en règle générale, à un locuteur correspond un extrait de parole. L'importance de l'extrait de parole lui-même sur les performances des systèmes automatiques est étudiée dans ce travail et renforce la nécessité d'attribution d'une mesure de confiance dans la décision fournie. Cette nécessité n'est pas réservée aux systèmes automatiques puisqu'elle a été également soulignée dans le cas des expertises judiciaires (French et Harrison, 2007), (Bonastre et al., 2003) ou (Eriksson, 2006). La définition d'une mesure de confiance impose de relier les erreurs commises pour les systèmes automatiques et/ou les auditeurs humains aux caractéristiques des extraits de parole utilisés. Il s'agit d'identifier et de classifier les indices extraits de la parole qui sont influencés par le locuteur afin de vérifier

si nous pouvons avoir plus confiance dans la décision prise à partir de ces segments.

Cadre de la thèse

Située à la croisée de l'informatique et de la phonétique, cette thèse, financée par le système de bourses DGA/CNRS, s'est déroulée au sein de deux laboratoires français, le Laboratoire Informatique d'Avignon (thématique Langage) et le Laboratoire Informatique de Grenoble (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole (GETALP)). Nos analyses ont été menées principalement à l'aide du système ALIZE/SpkDet, majoritairement développé au sein du LIA. Les questions que ce travail a suscitées nous ont conduit à nous rendre pour un séjour de trois mois au SRI International, Stanford, Californie, au sein du Speech Technology And Research (STAR) Laboratory afin de vérifier nos résultats sur un autre système de Reconnaissance Automatique du Locuteur (RAL), Idento.
Pour évaluer les systèmes automatiques, nous avons également bénéficié des données fournies lors des campagnes NIST- Speaker Recognition Évaluation (SRE) auxquelles ont participé les laboratoires LIA et SRI. Nous avons profité de la création d'une nouvelle tâche lors de NIST-SRE 2010 qui implique une évaluation des capacités humaines à discriminer des locuteurs.

Organisation du document

La partie I constitue un état de l'art qui énumère les éléments envisagés pour retrouver le locuteur dans le signal de parole.
Le chapitre 1 présente la vérification du locuteur par les systèmes automatiques. Le mode d'évaluation ayant des conséquences importantes sur les solutions envisagées en RAL, nous avons choisi de commencer par la présentation des évaluations NIST, avant de décrire les techniques de modélisation généralement utilisées. Il s'agit dans ce chapitre de comprendre les approches qui sous-tendent les techniques de RAL et non de détailler les techniques en elles-mêmes. Les paramètres acoustiques font, eux, l'objet d'une description plus précise.
Le chapitre 2 concerne la capacité humaine à reconnaître les personnes à partir d'un extrait de parole. Une distinction entre les performances obtenues pour les personnes connues et les nouvelles personnes est effectuée. Les processus cognitifs qui sous-tendent

cette capacité sont présentés et confrontés à un cas de phonoagnosie. Ce chapitre finit par la liste des indices qui ont été utilisés pour reconnaître le locuteur. La problématique précisant les objectifs de ce travail de thèse clôt cette première partie et introduit les contributions présentées en parties II et III.

La partie II du document interroge les modes d'évaluation afin de mettre en évidence la variabilité des performances en vérification du locuteur.
Les capacités humaines sont évaluées en s'appuyant sur notre participation à la campagne *Human Assisted Speaker Recognition* (HASR) et présentées chapitre 3. La recherche d'une mesure de confiance à attribuer aux réponses des auditeurs constitue l'objectif de ce chapitre.
Le chapitre 4 présente les expérimentations menées sur différentes bases de données. Ces expériences permettent, d'une part, de faire émerger des profils de locuteur. Elles permettent, d'autre part, d'estimer la variabilité des performances des deux systèmes de RAL, ALIZE/SpkDet et Idento, en fonction des extraits de parole utilisés.

La partie III met en évidence l'influence du locuteur sur des indices considérés comme idiosyncratiques. Nous cherchons à comprendre pourquoi des écarts de performance très importants sont observés en fonction des extraits de parole.
Le chapitre 5 présente la méthodologie utilisée pour établir l'influence relative du locuteur sur les indices et pour vérifier si ces indices permettent d'attribuer une confiance dans la décision prise.
Le chapitre 6 présente les résultats pour les indices issus de la phonation tandis que le chapitre 7 se focalise sur les indices relevant de l'articulation de la parole.
Le chapitre 8 est consacré aux coefficients cepstraux, utilisés par tous les systèmes de RAL. L'influence du locuteur sur ces coefficients est estimée pour chaque catégorie phonémique en français et en anglais et une première étude sur la sélection des vecteurs acoustiques est menée.

Nous concluons ce travail par un résumé des contributions de cette thèse ainsi que par une présentation des perspectives qui en découlent.

Première partie

Retrouver le locuteur par le signal de parole

Chapitre 1

Vérification du locuteur, systèmes de RAL et évaluation de la performance

Résumé : *Nous présentons dans ce chapitre la vérification du locuteur. Nous nous intéressons d'abord à la méthode commune d'évaluation adoptée pour rendre compte des performances des systèmes automatiques au cours de cette tâche. Après avoir décrit les principes de modélisation en prenant comme exemple les systèmes que nous avons utilisés au cours de notre thèse, nous revenons sur les paramètres idiosynchratiques qui ont été envisagés en RAL. Nous terminons ce chapitre en insistant sur le rôle central de l'évaluation dans le développement de la recherche en vérification automatique du locuteur.*

Sommaire

1.1 Vérification du locuteur

Reconnaître une personne à l'aide d'un enregistrement de sa parole est une tâche courante qui intervient dans de nombreuses situations. Par exemple, un auditeur peut chercher à savoir qui lui parle au téléphone, il peut aussi rechercher dans la foule une personne en particulier. Reconnaître le locuteur permet d'améliorer la sécurité dans une large variété d'applications, allant du contrôle d'accès à la sécurisation de transactions électroniques. Savoir quel locuteur parle offre également un grand intérêt dans le cadre d'applications de dialogue homme-machine, en termes de performance et de convivialité par exemple. L'application judiciaire, où l'enregistrement de parole est parfois la seule preuve accessible, est également un cas où il est important de savoir reconnaître une personne grâce à un enregistrement de parole.
Ces exemples d'application montrent bien la nécessité de comprendre comment extraire et organiser l'information sur le locuteur présente dans le signal de parole. Plusieurs tâches de reconnaissance peuvent être envisagées.

1.1.1 Tâches de reconnaissance

Identification du locuteur

La première tâche envisagée en reconnaissance du locuteur est **l'identification du locuteur**. Dans ce cas, il s'agit de retrouver l'identité d'un locuteur parmi un panel de personnes.
Dans le cadre d'une évaluation perceptive, il est possible que l'auditeur connaisse déjà le locuteur à identifier (personne connue) ou au contraire que l'auditeur ait dans un premier temps à apprendre la voix du locuteur (personne nouvelle).
L'identification peut être ouverte, dans le sens où il est possible que le locuteur ne soit pas présent dans le panel, ou bien fermée, lorsque le locuteur est obligatoirement dans le panel.
L'identification en milieu ouvert se retrouve dans des applications comme la recherche

d'un locuteur dans une foule. La technique du panel, qui consiste à demander à l'auditeur d'identifier le locuteur dans un ensemble d'enregistrements, a longtemps été utilisée pour les expertises judiciaires (French et Harrison, 2007).

Segmentation et structuration en locuteur

Une seconde tâche consiste à déterminer dans un signal de parole le nombre de locuteurs présents et le moment où ils interviennent. Il s'agit de **la segmentation et de la structuration en locuteur**. Cette tâche est essentielle pour un système domotique qui doit, par exemple, évaluer quand le propriétaire de la maison parle ou bien en recherche d'information lorsqu'il s'agit de retrouver un locuteur.

Vérification du locuteur

La tâche sur laquelle nous allons nous concentrer dans ce manuscrit est la vérification du locuteur (terme pour l'informatique) ou authentification du locuteur (terme pour les tests perceptifs). Cette tâche est souvent celle demandée lors des expertises judiciaires (Boë et al., 1999) où les experts doivent déterminer si la pièce à conviction a bien été prononcée par la personne suspectée. Elle intervient également dans les systèmes de sécurité de type contrôle d'accès.
La vérification du locuteur consiste à déterminer si deux enregistrements de parole ont été prononcés par le même locuteur. Cette tâche appartient donc aux tests de discrimination. Le premier enregistrement mis à disposition est appelé signal d'apprentissage ou pièce de comparaison tandis que le second est connu sous le terme de signal de test ou pièce de question.
Nous parlons de « comparaison cible » lorsque les deux enregistrements ont été produits par le même locuteur et de « comparaison imposteur » lorsque les deux enregistrements ont été prononcés par des locuteurs différents.

1.1.2 Réponse et erreurs en vérification du locuteur

La réponse à la tâche de vérification du locuteur est de type **binaire** : soit les deux enregistrements ont été prononcés par le même locuteur, soit deux locuteurs différents ont produit les enregistrements. Deux cas d'erreurs sont possibles.
Lorsque la décision est, en **comparaison cible**, que les deux enregistrements n'ont pas été produits par le même locuteur, nous parlons de **Faux Rejet**.
En **comparaison imposteur**, lorsque la réponse est que les deux enregistrements ont été produits par le même locuteur, nous parlons de **Fausse Acceptation**.
La décision prise doit toutefois être pondérée à l'aide d'**un score de confiance en la réponse fournie**. Pour les systèmes automatiques, ce score consiste en une mesure de vraisemblance d'appartenance et sert à la prise de décision (Campbell et al., 2005). Dans ce cadre, la réponse n'est donc pas uniquement binaire.

1.2 Un cadre d'évaluation : NIST-SRE

Depuis plus de 10 ans [1], le *National Institute of Standard and Technology* (NIST) propose d'évaluer les systèmes de traitements automatiques de la parole, notamment les systèmes de vérification du locuteur [2], en organisant une série de campagnes d'évaluation intitulée *NIST Speaker Recognition Evaluation* (NIST-SRE). Ces campagnes jouent un rôle important, voire fondamental, dans l'orientation des travaux de recherche de ces dernières années. Les avancées sur des questions comme l'influence du mode d'enregistrement ((Auckenthaler et al., 2000), (Campbell et al., 2007), (Kenny et al., 2005), (Dehak et al., 2009)) ou la durée des signaux ((Fauve et al., 2008),(Reynolds et al., 2000)) ont été possibles parce que NIST-SRE a proposé de nombreux corpus où ces problèmes étaient posés. Cette campagne est aujourd'hui une référence, c'est pour cette raison que nous allons commencer par présenter les principes métrologiques qui sous-tendent cette évaluation.

1. La première évaluation a eu lieu en 1996 mais la description de l'évolution des évaluations NIST s'appuie sur les plans d'évaluation accessibles sur http://www.itl.nist.gov/iad/mig/tests/spk/
2. D'autres tâches, comme le suivi de locuteur, sont également organisées par NIST, dans notre travail nous nous focalisons uniquement sur la tâche de vérification.

1.2.1 Les règles du protocole d'évaluation

La procédure d'évaluation proposée par NIST-SRE consiste à **mettre à disposition plusieurs centaines de milliers de comparaisons cible et imposteur**.
Pour évaluer les systèmes dans les mêmes conditions, certaines règles communes ont été édictées par le NIST (NIST, 2011). Ces règles ont évolué au cours du temps.

Indépendance des comparaisons

Le premier principe de NIST-SRE consiste en l'indépendance des comparaisons les unes par rapport aux autres. Ainsi, **il est interdit d'utiliser les autres comparaisons fournies par NIST-SRE pour décider si les deux enregistrements en question proviennent du même locuteur**.
Il est entendu que les signaux fournis lors des campagnes précédentes peuvent être utilisés pour développer les systèmes automatiques, en complément des données limitées fournies par NIST lors de l'évaluation en cours.

Informations fournies

Toutes les informations accessibles dans les en-têtes des fichiers d'enregistrement peuvent être utilisées pour la prise de décision.
Les **modes d'enregistrements** des fichiers ont évolués au cours du temps et cette information est actuellement bien spécifiée. Depuis 2008, deux façons d'enregistrer les locuteurs sont possibles dans le corpus (NIST, 2008). Il s'agit de données dites *interview* et de données dites *téléphoniques*. Si pour les données *interview*, les locuteurs sont enregistrés à l'aide de différents microphones dans une salle peu bruitée, pour les données *téléphoniques*, les locuteurs sont mis en relation de manière aléatoire et doivent converser entre eux à travers le téléphone. Le type de microphone ou de téléphone utilisé n'est pas connu, mais les participants savent s'il s'agit de données interview ou des données téléphoniques (NIST, 2008).
Le **sexe** du locuteur ainsi que la **langue**[3] parlée sont fournis. Si jusqu'en 2001 la **tran-**

3. De 1997 à 1999 ainsi qu'en 2010, l'ensemble des enregistrements étaient en langue anglaise, cette information n'avait donc pas lieu d'être. Il était tout de même indiqué si les locuteurs étaient natifs de

scription orthographique des signaux était interdite (Campbell et Reynolds, 1999), elle est depuis cette date autorisée. Il est à noter qu'au début de l'introduction de ces transcriptions, la transcription manuelle était autorisée. Aujourd'hui, l'écoute des signaux est interdite, les transcriptions peuvent être fournies par le NIST-SRE ou obtenues à l'aide de systèmes automatiques de reconnaissance de la parole (Doddington et al., 2000).
La nouvelle tâche *Human Assisted Speaker Recognition* (HASR) proposée en 2010 (Greenberg et al., 2011b) est une exception à l'**interdiction d'écoute des enregistrements** puisqu'il s'agit justement de voir quelles améliorations peut fournir l'écoute de certaines comparaisons par l'humain.

Soumission

Pour chacune des comparaisons, les participants doivent soumettre une **décision** qui spécifie si l'hypothèse selon laquelle les deux enregistrements ont été produits par le même locuteur est vraie ou fausse. Ils doivent également soumettre **un score** qui suit la règle « plus le score est grand, plus la confiance dans le fait que les deux enregistrements proviennent du même locuteur est grande. »(NIST, 2010).

1.2.2 La métrique d'évaluation

A partir de la décision

Plutôt qu'un taux de réponses correctes sur la totalité des comparaisons soumises, **la performance du système est évaluée pour les comparaisons cible d'une part et imposteur d'autre part**. Le pourcentage de fausses acceptations (FA) et le pourcentage de faux rejets (FR) sont mesurés tels que définis par les équations 1.1 et 1.2.

$$FR = \frac{Nombre\ de\ comparaisons\ résultant\ en\ un\ Faux\ Rejet}{Nombre\ de\ comparaisons\ cible} \tag{1.1}$$

$$FA = \frac{Nombre\ de\ comparaisons\ résultant\ en\ une\ Fausse\ Acceptation}{Nombre\ de\ comparaisons\ imposteur} \tag{1.2}$$

l'anglais ou non.

Chaque participant peut ainsi être positionné dans un plan FA-FR. Cette représentation est utilisée pour comparer les réponses des participants dans le cadre de HASR. En revanche, l'évaluation des performances des systèmes automatiques ne se fait pas uniquement à l'aide des décisions binaires mais également à partir du score fourni.

A partir du score

Chaque participant doit indiquer un score qui respecte la règle selon laquelle plus l'hypothèse que les deux enregistrements ont été produits par le même locuteur est sure, plus ce score est élevé. Dans ce cadre, la distribution de scores obtenus pour les comparaisons cible doit logiquement montrer une moyenne plus élevée que la distribution des scores des comparaisons imposteur.
Dans un cas idéal, les deux distributions sont complètement séparées. Dans la réalité, elles se chevauchent et **ce chevauchement donne lieu à des erreurs qui peuvent être quantifiées en termes de FA et de FR**. Une représentation schématique de ces distributions est donnée en figure 1.1.

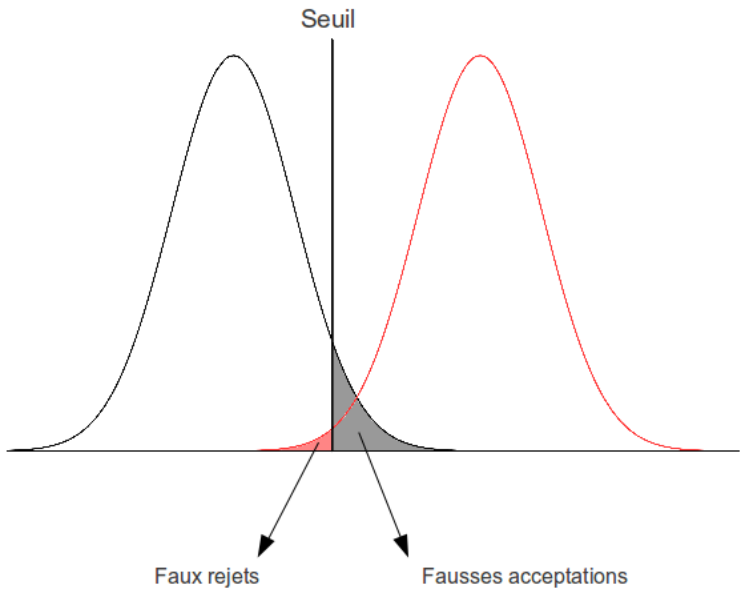


FIGURE *1.1* – *Représentation schématique des distributions de scores : importance du seuil*

Toutes les comparaisons dont le score est supérieur à un seuil donné sont considérées comme provenant du même locuteur, tandis que les comparaisons dont le score est in-

férieur à ce même seuil sont considérées comme provenant de deux locuteurs différents (décision).

Le seuil choisi a donc une grande influence sur les FA et de FR. Chaque seuil définit un point dans un plan FA-FR. La courbe représentant ces points est dénommée courbe DET (*Detection Errore Tradeoff*) (Martin et al., 1997)) et est illustrée par la figure 1.2 revenant ainsi à une représentation dans un espace usuel.

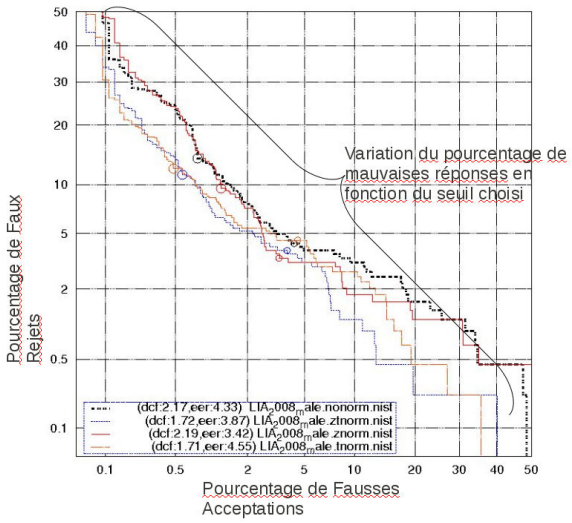


FIGURE 1.2 – *La courbe DET : une représentation de l'évolution des taux d'erreur en fonction du seuil choisi*

Des points particuliers sur cette courbe permettent également de comparer plus simplement les systèmes entre eux. Le premier d'entre eux est le **DCF** qui sert de référence principale dans le cadre de NIST-SRE (NIST, 2010). Il s'agit d'un point qui permet de prendre en compte les priorités de NIST en terme de coût des erreurs. Il est défini par l'équation 1.3.

$$DCF = Coût(FR) * P(cible) * FR + Coût(FA) * P(imposteur) * FA \tag{1.3}$$

où $Coût(FR)$ et $Coût(FA)$ sont respectivement les coûts d'un Faux Rejet et d'une Fausse Acceptation et $P(cible)$ et $P(imposteur)$ sont les probabilités des comparaisons cibles et des comparaisons imposteur.
Dans les campagnes NIST-SRE, le coût des erreurs et les probabilités a priori d'apparition des différentes comparaisons ont évolué au cours des années comme reporté dans le tableau 1.1. L'évolution des coûts et des probabilités *a priori* suggère que les organisateurs de NIST veulent que les systèmes évitent de réaliser des FA au détriment des FR.

Période	$Coût(FR)$	$Coût(FA)$	$P(cible)$	$P(imposteur)$
1997-2010	10	1	0.01	1-$P(cible)$=0.99
2010 [4]	1	1	0.001	1-$P(cible)$=0.999

TABLE 1.1 – *Valeurs des coûts des erreurs et des probabilités d'apparitions des types de comparaisons.*

Le **taux d'égal erreur ou *Equal Error Rate* (EER)** qui correspond au point où $FR = FA$ est un autre point de référence illustré par la figure 1.3.

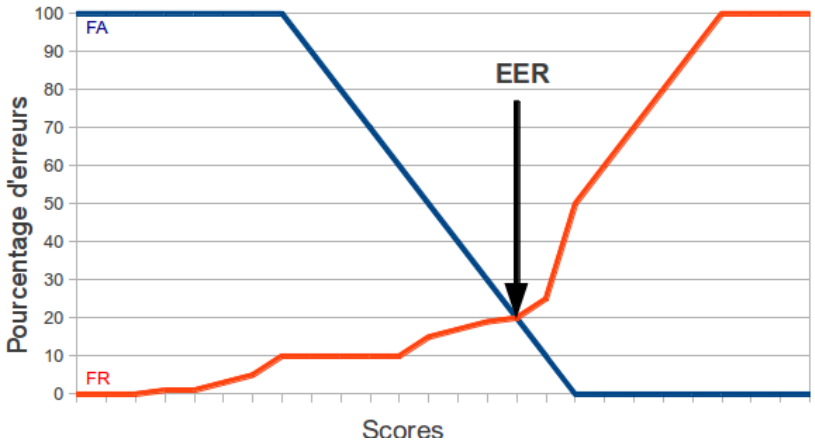


FIGURE 1.3 – *EER, score pour lequel* $p(FA) = p(FR)$

Cette mesure est pratique pour comparer les systèmes mais, comme elle est calculée *a posteriori*, elle ne permet pas d'évaluer les techniques de détermination du seuil optimal pour tenir compte de la fonction de coût définie par l'application visée. Dans ce travail, nous évaluons les systèmes de vérification du locuteur étudiés en donnant le même

4. Pour certains jeux de comparaisons

poids au FA et au FR, aussi les performances des systèmes sont, elles, mesurées à l'aide de leur EER.

1.2.3 Résultats de l'évaluation de la performance des systèmes automatiques

Nous observons depuis la création des évaluations NIST-SRE, une **amélioration indiscutable des performances des systèmes automatiques de vérification du locuteur**, notamment depuis l'introduction des techniques d'analyses factorielles. En 2010, les meilleurs systèmes étaient en dessous de 2% d'EER pour la condition avec 2 minutes 30 de signal en apprentissage et en test (Greenberg et al., 2011a). **Plusieurs éléments de variation de performance ont été étudiés à l'occasion des campagnes NIST-SRE**. L'influence de ces facteurs sur les performances a été étudiée en comparant les résultats des différentes cohortes qui possèdent certaines caractéristiques que nous allons à présent décrire.

La durée des enregistrements

Dès le début des campagnes NIST-SRE, **la durée des enregistrements a été envisagée comme un facteur pouvant faire varier les performances des systèmes**. Si au départ la durée des enregistrements variait entre 30 secondes et 2 minutes, aujourd'hui les cohortes peuvent contenir des enregistrements de 10 secondes à plus de 20 minutes. Ceci a évidemment une influence sur la quantité d'information extraite par les systèmes pour effectuer la comparaison. Comme illustré par la table 1.4 [5], avec **l'accroissement de la durée des enregistrements une amélioration globale des résultats est observée**, les taux d'EER variant de plus de 18% d'EER pour des durées de 10 secondes d'enregistrement à moins de 3% d'EER pour 20 minutes d'enregistrement.

Ces tendances se retrouvent en 2010 où l'EER varie de 0.5% à 17% pour respectivement les fichiers de 20 minutes et ceux de 10 secondes (Greenberg et al., 2011a) [6]. Ces résultats sont confirmés, lors d'une étude utilisant les outils ALIZE/SpkDet (Fauve et al.,

5. L'ensemble des courbes est issu de http://www.itl.nist.gov/iad/mig/tests/spk/2008/official_results/index.html

6. Dans cet article, seules les courbes DET sont fournies, les EER sont déduits de la lecture de ces courbes.

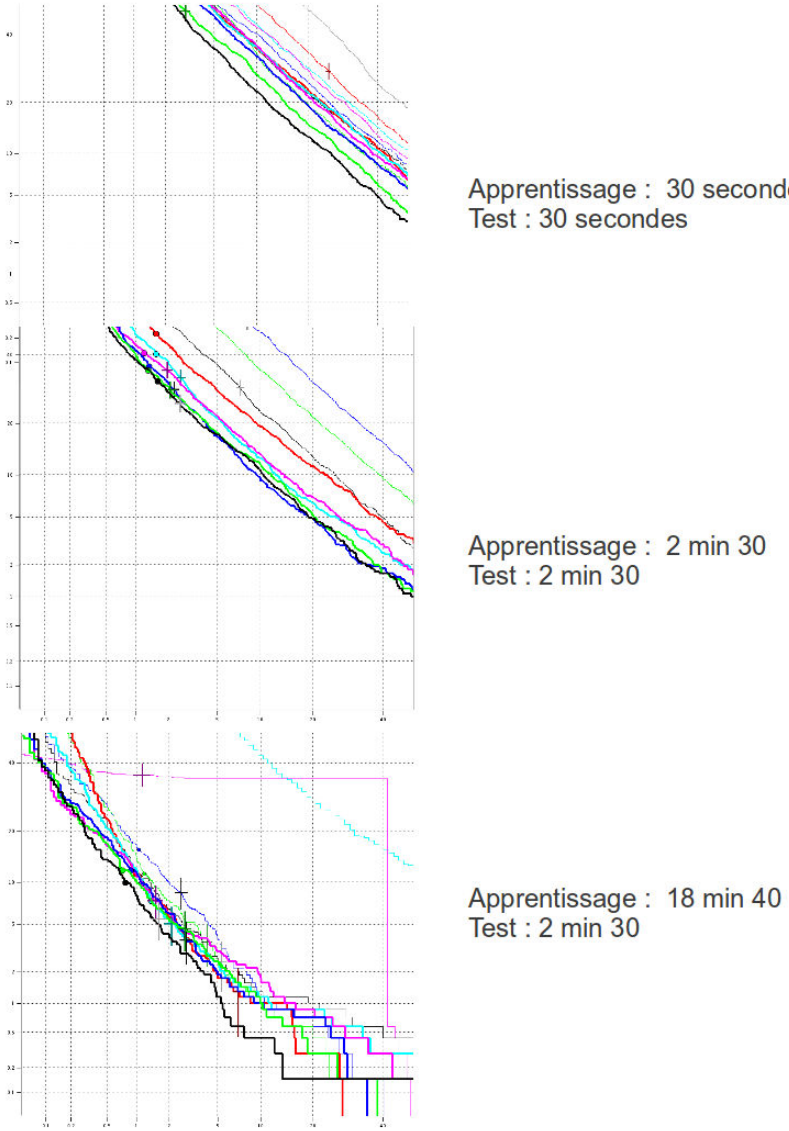


FIGURE 1.4 – *Courbes DET résumant les performances obtenues par les différents systèmes ayant participé à NIST-SRE 2008 selon la durée des enregistrements en parole conversationnelle téléphonique (10 secondes en apprentissage et 10 secondes en test ; 2 minutes 30 secondes en apprentissage et 2 minutes 30 secondes en test ; 20 minutes en apprentissage et 2 minutes 30 secondes en test).*

2007) en analysant pour les outils développés dans ALIZE/SpkDet les différences de performance. Dans ce cas, les auteurs observent pour le système SVM une variation de 9.43% à 30.1% d'EER pour des enregistrements de 2 minutes 30 secondes d'une part et de 10 secondes d'autre part.

Les techniques d'enregistrement

La question **des différentes techniques d'enregistrement de la parole du locuteur est un problème qui a été soulevé très tôt par NIST-SRE**. En 1997, par exemple, tous les locuteurs étaient déjà enregistrés avec deux types de téléphones filaires. En 2002, les téléphones cellulaires sont rajoutés aux filaires. En 2008, les enregistrements *interview* sont rajoutés. Les performances des systèmes sont sensibles aux différences de canal comme l'illustre les courbes DET de la Figure 1.2 [7].

Les données *interview* conduisent à de meilleurs résultats que les données *téléphoniques* ($EER_{interview} = 3\%$ vs $EER_{telephonique} = 5\%$). Lorsque ce sont les **mêmes microphones** qui ont enregistrés les éléments utilisés en apprentissage et en test, **les performances globales des systèmes sont également meilleures** ($EER_{interview\text{même}micro} = 1.3\%$ vs $EER_{interviewmicro\text{différents}} = 3\%$). Il est toutefois à noter que ce ne sont pas exactement les mêmes comparaisons qui sont effectuées en données interview et en données téléphoniques par exemple ce ne sont pas les mêmes auditeurs qui sont enregistrés et le nombre de comparaisons utilisées pour calculer la performance est différent. Les résultats ne sont pas parfaitement comparables mais ils rendent bien compte de l'influence du mode d'enregistrement sur les performances du système.

La langue des locuteurs

Entre 1997 et 1999, l'ensemble des enregistrements étaient en langue anglaise. En 2010, la difficulté résidait dans le fait que certains locuteurs étaient natifs et d'autres non. Pour les autres campagnes, plusieurs langues étaient représentées : c'est d'abord l'espagnol qui a été ajouté, puis le mandarin et le russe pour arriver en 2008 à plus de 20 langues différentes représentées. La difficulté susceptible d'être ajoutée par les changements de langue est étudiée en terme de performance globale. Il s'agit d'estimer si les performances se dégradent globalement lorsque plusieurs langues sont représentées dans la cohorte. Dans le cas de 2008 [8], **les performances lorsqu'il n'y a que des locuteurs anglophones sont meilleures que pour les conditions où les lo-**

7. L'ensemble des courbes de résultats est issu de http://www.itl.nist.gov/iad/mig/tests/spk/2008/official_results/index.html
8. L'ensemble des courbes de résultats est issu de http://www.itl.nist.gov/iad/mig/tests/spk/2008/official_results/index.html

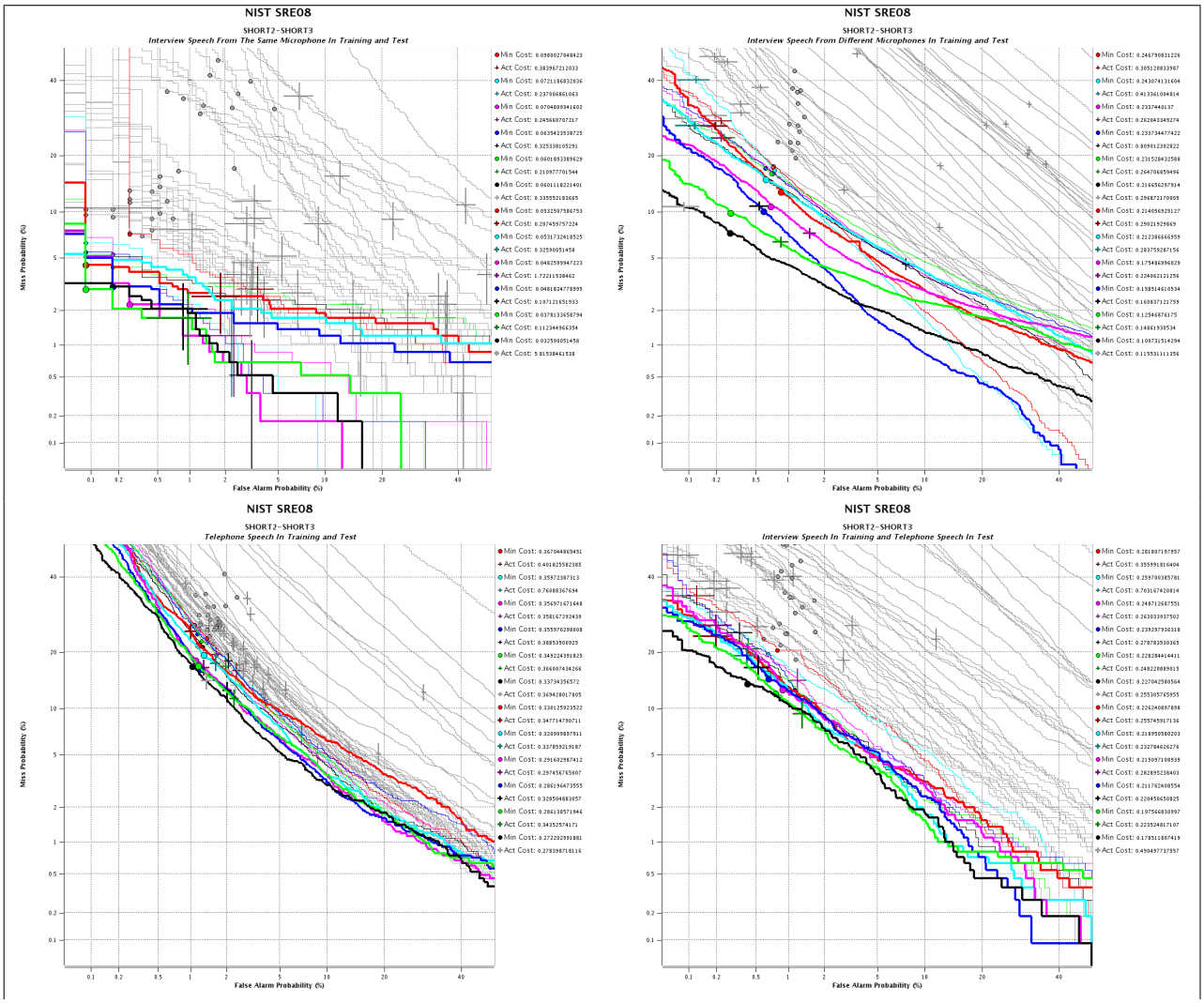


TABLE 1.2 – *Courbes DET résumant les performances obtenues par les différents systèmes ayant participé à NIST-SRE 2008 selon le type d'enregistrement utilisé. Tous les signaux d'apprentissage et de test durent 2 minutes et 30 secondes (Données d'interview avec le même microphone en apprentissage et en test ;Données d'interview avec des microphones différents ;Données téléphoniques ; Données d'interview en apprentissage et téléphonique en test.)*

cuteurs parlent plusieurs langues (pour le meilleur système, $EER_{Anglais} = 2.1\%$ vs $EER_{Touteslanguesconfondues} = 5\%$). Il est tout de même à noter que l'ajout de locuteurs non natifs de l'anglais n'a que peu d'influence sur les performances globales (Pour le meilleur système, $EER_{Anglaisnatifs} = 2.0\%$ vs $EER_{Anglaisavecnonnatifs} = 2.1\%$).

Autres pistes

En 2010, pour la première fois, un corpus permettant de tester l'influence de **l'effort vocal** des locuteurs sur les performances des systèmes a été proposé. Trois séries sont alors comparées. Dans la première, qui sert de témoin, le signal d'apprentissage correspond à une session où les locuteurs parlent « normalement ». Dans la seconde, appelée *low effort*, le signal d'apprentissage correspond à une session où il est demandé aux locuteurs de parler doucement. Enfin, pour la troisième série appelée « *high effort* », les fichiers d'apprentissage correspondent à une session où les locuteurs avaient un casque sur les oreilles par lequel du bruit était diffusé. Ils devaient se faire entendre d'un autre interlocuteur qui lui aussi avait un casque sur les oreilles. La conséquence de ce protocole est que les locuteurs ont tendance à parler plus fort. Ce sont les mêmes locuteurs qui ont participé à ces trois séries. Les fichiers de test sont considérés comme comprenant de la parole « normale ». Dans ce cadre, les performances des systèmes varient légèrement en fonction de la série. Si pour la série *low effort*, l'EER est autour de 1%, il est de 2.1% pour la série *high effort*. La série témoin a un EER de 1.9% (Greenberg et al., 2011a)[9]. L'effort vocal a donc une légère influence sur les performances des systèmes. Il est surprenant que les meilleurs résultats ne soient pas pour la série où les efforts vocaux en apprentissage et en test sont équivalents. Une analyse des signaux pourrait peut-être expliquer ces différences.
Un autre élément a commencé à être analysé en 2010, il s'agit de **la constance de la voix à travers le temps**. Ainsi, des personnes qui avaient été enregistrées lors des années précédentes ont été de nouveau enregistrées afin d'évaluer la capacité des systèmes à trouver une constance dans la voix au cours du temps. Les résultats de ces séries n'ont toutefois pas encore été publiés à notre connaissance.

1.3 Principes de modélisation d'un locuteur par un système de RAL

Ces dernières années, une nette amélioration des performances de systèmes de RAL a été mise en évidence lors des campagnes NIST-SRE et l'étude de l'influence de facteurs tels que la durée des enregistrements ou les modes d'enregistrement ont amené

9. Dans ce papier, les résultats sont uniquement présentés en terme de courbes DET, les EER sont déduits de ces courbes

les participants à proposer des modélisations des locuteurs qui prennent en compte ces difficultés. Les principes de modélisation de ces systèmes relèvent soit d'**approches génératives** soit d'**approches discriminantes** et nécessitent un **apprentissage supervisé**. La description des deux systèmes que nous avons utilisés lors de ce travail, ALIZE/SpkDET (Bonastre et al., 2008) et Idento (Scheffer et al., 2011), permet de présenter les principes des modélisations les plus utilisées dans le domaine, sans prétendre à l'exhaustivité.

1.3.1 Décomposition en trois phases

Trois étapes distinctes sont identifiées dans un systèmes automatique de vérification du locuteur.
La première consiste à extraire du signal des vecteurs de paramètres, x, à intervalles réguliers (trames), c'est la **phase de paramétrisation**. Cette étape est effectuée aussi bien pour le signal d'apprentissage que pour le signal de test. Nous la décrirons en détail dans la partie 1.4. La seconde consiste à construire un modèle à partir des paramètres extraits du signal d'apprentissage afin d'**obtenir le modèle du locuteur**. La troisième, **la phase de décision**, consiste à décider du statut du signal de test : l'hypothèse qu'il ait été produit par le même locuteur que celui qui a été enregistré dans le signal d'apprentissage est-elle vraie ? Dans cette partie nous présentons ces deux dernières étapes.
Le processus de décision en vérification du locuteur repose sur un test d'hypothèses. Connaissant un signal de parole, S, et un locuteur L, deux hypothèses peuvent être envisagées.

- Le parole a été prononcée par le locuteur (H_0).
- La parole n'a pas été prononcée par le locuteur (H_1).

Il s'agit d'**estimer la probabilité de chacune des deux hypothèses**. Le rapport entre les deux hypothèses H_0 et H_1 s'écrit comme défini par l'équation 1.4 :

$$LR(S, L) = \frac{p(S|H_0)}{p(S|H_1)} \tag{1.4}$$

Ce rapport correspond au score à partir duquel la décision est prise : il est comparé à un seuil, θ, pour prendre la décision. Lorsque le ratio est inférieur à θ, l'hypothèse selon laquelle les deux enregistrements ont été prononcés par deux locuteurs différents est retenue. Si le ratio est supérieur à θ, alors l'hypothèse qu'il s'agit du même locuteur est considérée comme vraie.

Nous allons nous attacher à décrire ici les deux systèmes avec lesquels nous avons travaillé. Le premier, ALIZE/SpkDet (Bonastre et al., 2008) repose sur une approche générative tandis que Idento (Scheffer et al., 2011) s'appuie sur une méthode discriminante.

1.3.2 ALIZE/SpkDet : exemple d'une approche UBM-GMM

ALIZE/SpkDet est une plate-forme biométrique libre de droit principalement développée au Laboratoire Informatique d'Avignon. **Ce système repose sur une approche générative qui utilise des modèles multi-gaussien (GMM) dont la construction s'appuie sur un *Universal Background Model* ou modèle du monde (UBM)**.

Modèle multi-gaussien et Modèle du monde

Le modèle du locuteur est résumé par un **mélange de gaussiennes** (GMM) (Reynolds, 1995) de dimension M [10]. Le locuteur est ainsi caractérisé par une loi de densité de probabilité dans l'espace des paramètres acoustiques. Pour un vecteur de paramètres x de dimension D, la loi de densité est définie par l'équation 1.5.

$$p(x, H) = \sum_{i=1}^{M} w_i p_i(x) \tag{1.5}$$

où w_i est le poids attribué à la gaussienne i tel que $\sum_{i=1}^{M} w_i = 1$ avec $p_i(x)$, la probabilité pour x d'appartenir à la gaussienne i définie par l'équation 1.6.

$$p_i(x) = \frac{1}{2\pi^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{1.6}$$

Ce modèle repose sur l'hypothèse que les vecteurs de paramètres, x, provenant d'un enregistrement prononcé par un locuteur suivent une loi de probabilité propre à ce locuteur.
Nous l'avons vu précédemment, la durée des enregistrements est assez courte, en majorité de 2.5 minutes soit environ 15 000 trames. **L'estimation du GMM avec uniquement ces trames n'est pas très précise**.
Pour remédier à cela, le modèle du locuteur est construit par adaptation d'un modèle

10. Pour ALIZE/SpkDet, la dimension du GMM varie de 256 à 2048 gaussiennes (Larcher et al., 2010)

du monde (UBM). Le modèle du monde est un mélange de gaussiennes construit à partir de plusieurs milliers de fichiers de parole prononcés par des centaines de locuteurs dans différentes conditions. L'idée ici est d'**obtenir une représentation précise de ce qu'est la parole afin de structurer l'espace des paramètres autour des lieux où se concentrent les échantillons de parole**. Ce modèle est unique pour l'apprentissage de tous les locuteurs. Il est tout de même souvent construit un modèle du monde par genre.
L'apprentissage du modèle du monde repose sur l'algorithme d'Espérance Maximisation (EM) (Laird, 1993). Cet algorithme cherche à optimiser les paramètres du modèle pour maximiser la vraisemblance des données avec ce même modèle.
Le modèle du locuteur consiste à **modifier, en fonction des paramètres extraits du fichier d'apprentissage, les moyennes de l'UBM**. La méthode la plus utilisée en reconnaissance du locuteur est celle du *Maximum a Posteriori* (MAP) (Gauvain et Lee, 1994). La figure 1.5 illustre le processus d'apprentissage à partir d'un UBM.

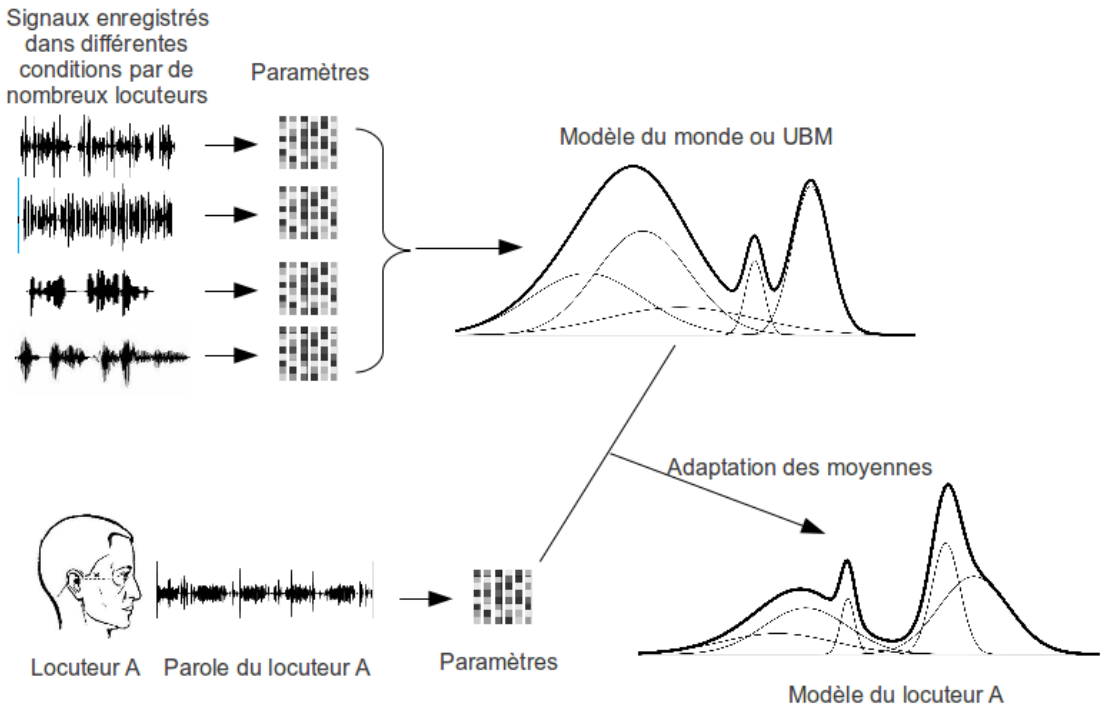


FIGURE 1.5 – *Apprentissage du modèle de locuteur : une adaptation du modèle du monde selon les paramètres extraits du signal d'apprentissage.*

Dans le cadre d'un modèle UBM-GMM, $p(S|H_0)$ correspond à la vraisemblance moyenne du signal test d'appartenir au modèle du locuteur.
$p(S|H_1)$ peut être approximée grâce à une cohorte d'imposteurs (Rosenberg et al., 1992) mais c'est généralement la vraisemblance d'appartenance au modèle du monde (Carey et Parris, 1992) qui est utilisée pour estimer la contre-hypothèse.
Si (Higgins et al., 1991) préconise d'utiliser pour la construction de l'UBM des signaux de parole considérés comme proches de celui du signal d'apprentissage, (Reynolds, 1995) considère qu'il faut utiliser tous les types d'enregistrement de manière à avoir accès à des signaux proches et d'autres plus éloignés.

Factor Analysis

La variation des conditions d'enregistrement des signaux de parole est un des problèmes majeurs que les participants à NIST-SRE ont cherché à résoudre. La technique du Factor Analysis est une des solutions qui a permis d'améliorer très significativement (d'un facteur deux en moyenne) les performances des systèmes [11]. Cette technique est implémentée dans ALIZE/SpkDet (Matrouf et al., 2008a).
Le Factor Analysis (Kenny et al., 2005) repose sur l'hypothèse que **l'enregistrement de parole dépend de différents facteurs dont deux principaux, l'un rendant compte du locuteur, l'autre correspondant à des informations sur le canal**. Ici le canal est entendu comme information pouvant varier dans le signal de parole et qui n'est pas due au locuteur comme les conditions d'enregistrement ou la langue. Le Factor Analysis admet comme seconde hypothèse que **chaque facteur peut être représenté dans un sous-espace spécifique de dimension réduite**. Il s'agit donc dans cette technique de ne **conserver pour la modélisation que la partie contenant l'information sur le locuteur**.

Normalisations

Dans les campagnes NIST-SRE, il est demandé aux participants de ne fournir qu'un seul seuil. Par ailleurs, les campagnes ont très tôt imaginé des cohortes où fichiers d'apprentissage et de test sont enregistrés dans des conditions très différentes. Ces contraintes ont eu, notamment, comme conséquence d'amener les participants à proposer

11. $EER_{sansFA} = 17\%$ vs $EER_{avecFA} = 7.1\%$ (Kenny et Dumouchel, 2004)

une normalisation des scores. **Les techniques les plus courantes consistent à centrer et réduire les scores obtenus en fonction de cohortes d'imposteurs**. Trois types de normalisations sont observées en 2010 dans les systèmes participants à la campagne NIST-SRE.

z-norm permet de préciser le rapport du signal d'apprentissage à des signaux dont nous savons qu'ils sont des imposteurs (Reynolds, 1997). Cette transformation se résume par l'équation 1.7.

$$Score = \frac{log(p(Signal_{Apprentissage}|H_0)) - \mu_I}{\sigma_I} \tag{1.7}$$

où μ_I et σ_I sont respectivement la moyenne et la variance des scores obtenus par le signal d'apprentissage face à des fichiers test que nous savons tous imposteur.

t-norm permet de préciser le rapport du signal de test à des signaux dont nous savons qu'ils sont des imposteurs (Auckenthaler et al., 2000). Cette transformation se résume par l'équation 1.8.

$$Score = \frac{log(p(Signal_{test}|H_0)) - \mu_I}{\sigma_I} \tag{1.8}$$

zt-norm est l'application conjointe et successive des deux normalisations précédentes. Ces normalisations permettent une légère amélioration des résultats comme l'illustre la figure 1.6 pour ALIZE/SpkDet.

En 2008, pour ALIZE/SpkDet, sur les fichiers d'apprentissage et de test de 2.5 minutes, les EER sans normalisation, avec znorm, avec tnorm et ztnorm sont respectivement de 4.33%, 3.42%, 4.55% et 3.87%.

1.3.3 Idento : exemple de système fondé sur les i-vectors

Idento (Scheffer et al., 2011) est le **système i-vectors** développé par le SRI. Il repose sur un **classifieur Machine à Vecteurs Support (SVM)**.

I-vector

Dans la continuité du Factor Analysis (Kenny et al., 2005), (Dehak et al., 2009) met au point la technique des i-vectors en **utilisant le Factor Analysis comme une « technique d'extraction des paramètres »**. Si la technique du Factor Analysis telle que décrite

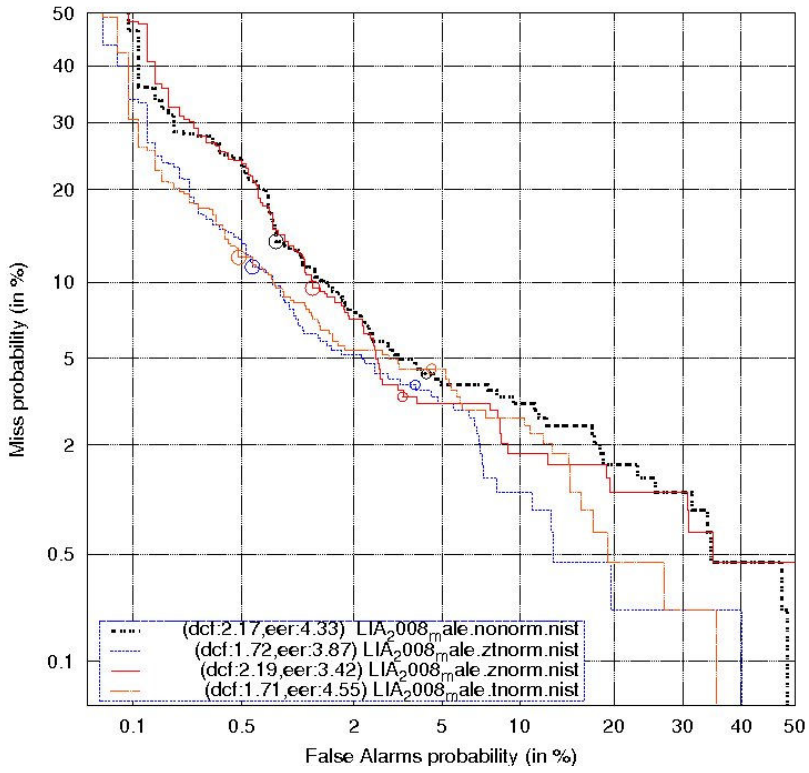


FIGURE 1.6 – *Performances pour ALIZE/SpkDet sur les données NIST-SRE 08 en fonction de la normalisation utilisée :* $EER_{nonorm} = 4.33$, $EER_{znorm} = 3.42$, $EER_{tnorm} = 4.55$, $EER_{ztnorm} = 3.87$

dans (Kenny et al., 2005) considère deux espaces séparés entre le canal et le locuteur, l'hypothèse pour la construction des i-vectors est qu'**il existe un espace de *variabilité totale* qui contient à la fois la variation due au canal et au locuteur**. Les **i-vectors correspondent à un vecteur qui contient les coordonnées du locuteur dans *l'espace de variabilité totale***. Une compensation des différences de mode d'enregistrement est effectuée dans ce cadre à l'aide d'une Analyse Linéaire Discriminante (LDA) dans l'espace de variabilité totale.

L'avantage de ces techniques est de réduire grandement la dimension des vecteurs de paramètres et donc d'améliorer la durée d'exécution des calculs. Elles améliorent également les performances avec une diminution d'un facteur deux [12]. Dans ce cadre, pour estimer les probabilité de H_0 et de H_1, un SVM est utilisé.

12. Pour les données hommes anglophones de NIST-08, $EER_{FactorAnalysis} = 2.64\%$ vs $EER_{i-vectors} = 1.12\%$ (Dehak et al., 2009)

Machine à vecteur support

Les SVM sont des **classifieurs développés pour permettre la séparation de données complexes dans des espaces de grandes dimensions** (Wan et Campbell, 2000). Plutôt que de construire une fonction qui associe directement l'espace d'entrée à la classe souhaitée, la classification consiste à trouver une fonction dont le signe donne l'appartenance à la classe (Scheffer, 2006). Il s'agit dans ce cas de **trouver un hyperplan qui permet au mieux de séparer les données de chacune des deux classes**. Cet hyperplan est estimé de façon à maximiser la marge entre les données des deux classes comme l'illustre la figure 1.7.

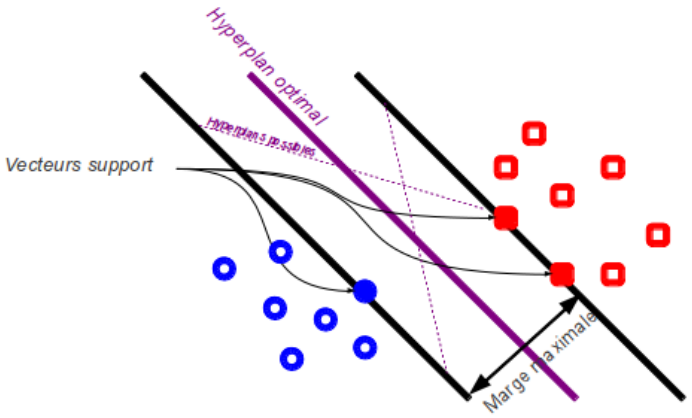


FIGURE 1.7 – *Représentation schématique d'un SVM d'après (Scheffer, 2006)*

L'attrait pour les classifieurs SVM tient en leur capacité à traiter des problèmes non-linéaires. Le principe sous jacent est qu'un problème non séparable linéairement peut être transformé en un problème séparable linéairement dans un espace de dimension suffisamment grande.
Dans le cas du système i-vectors, **les i-vectors sont projetés dans le SVM**. La distance est estimée à l'aide d'un *cosine kernel*.

Normalisation

Les mêmes types de normalisation que celles décrites en 1.3.2 peuvent être appliquée sur les scores produits par Idento.

1.4 Focus sur la phase de paramétrisation

Nous avons vu précédemment les principales techniques de modélisation du locuteur. Nous souhaitons ici décrire les paramètres acoustiques qui ont été utilisés afin de mieux appréhender où se situe l'information propre au locuteur. **L'étape de paramétrisation consiste, à partir d'un signal de parole numérisé, à extraire des vecteurs de paramètres à intervalle de temps régulier**, trames (Bimbot et al., 2004). Plusieurs paramètres ont été envisagés pour reconnaître le locuteur les plus courants étant les paramètres cepstraux. A titre d'exemple, le tableau 1.3 résume l'ensemble des paramètres utilisés par les différents systèmes ayant participé à la dernière campagne d'évaluation des systèmes NIST-SRE 2010 (NIST, 2010) [13].

	MFCC	PLP	LPCC	LFCC	Paramètres Prosodiques	Autre
Nombre de systèmes	43	14	11	6	3	5

TABLE *1.3* – *Paramètres utilisés par les différents systèmes ayant participé à la campagne NIST-SRE 2010.*

Nous allons ici nous attacher à décrire les principes qui fondent ces différents paramètres. La Figure 1.8 illustre le calcul de ces différents paramètres (Furui, 1981).

1.4.1 Analyses cepstrales

Les paramètres issus d'analyses cepstrales sont utilisés dans tous les systèmes de vérification du locuteur. Nous pouvons en distinguer différents types (Furui, 1981) : **les MFCC, les LFCC, les LPCC et les PLP**. L'atout majeur de ces paramètres est qu'ils sont **décorrélés entre eux** ((Haton et al., 2006), (Larcher, 2009)). La décorrélation des

13. Ces statistiques sont tirées des descriptions de systèmes fournies par les participants de NIST-SRE 2010 lors du workshop NIST-SRE de Brno les 24 et 25 juin 2010

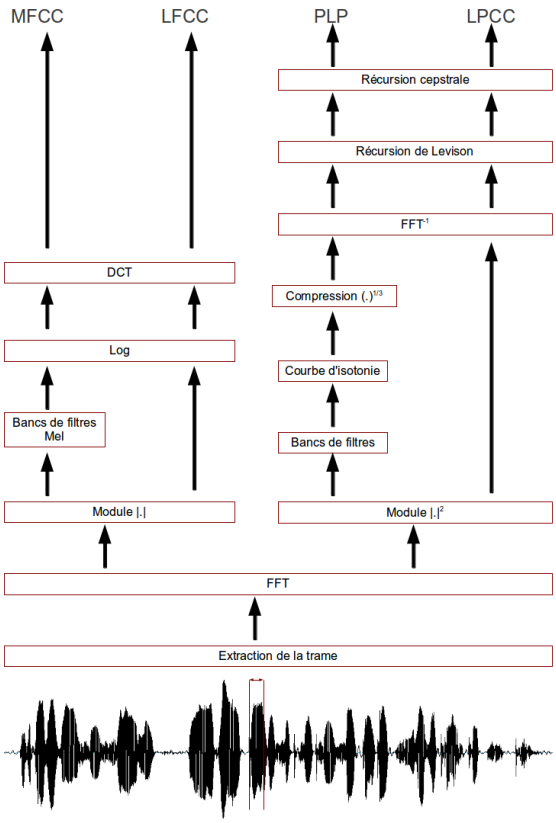


FIGURE *1.8* – *Processus de calcul des différents paramètres cepstraux utilisés*

coefficients permet de limiter le nombre de coefficients nécessaires pour définir l'espace à modéliser puisque toutes les valeurs sont indépendantes les unes des autres. **Tous ces coefficients reposent sur une analyse fréquentielle trame par trame du signal**. L'**information temporelle est abandonnée** dans les systèmes de RAL contrairement aux systèmes de reconnaissance de la parole qui utilisent égelent ces coefficients (Haton et al., 2006).

Sélection de trames

Il est important de souligner que **toutes les trames du signal ne sont pas utilisées**. Une technique de détection d'activité vocale est employée par tous les systèmes ayant participé à NIST-SRE 2010. Cette technique consiste à classifier chacune des trames en trame de **parole ou de non-parole en utilisant la quantité d'énergie**.
La majorité des systèmes utilisent pour leur modèle l'ensemble des trames détectée

comme parole, mais certains systèmes utilisent des sélections plus précises. Par exemple, le système du SRI International (Scheffer et al., 2011) combine plusieurs systèmes dont un n'utilise que **les trames issues des phonèmes nasals**. Ils utilisent alors un système de transcription de la parole pour déterminer les trames qui seront utilisées. Dans le même ordre d'idée, un autre de leurs systèmes ne sélectionne que **les trames en fin de groupe de souffle**. Tous ces choix partent des hypothèses que ces zones sont plus porteuses d'information sur le locuteur [14].

LFCC et MFCC

Les MFCC (Davis et Mermelstein, 1980) et LFCC sont le résultat d'une analyse fréquentielle du signal de parole réalisée à l'aide de calcul de spectres de Fourier à court terme sur une fenêtre temporelle. Pour les MFCC, une fois dans l'espace des fréquences, le module du spectre est filtré par un banc de filtre dont les fréquences centrales sont fixées par l'échelle Mel (Stevens et al., 1937). Cette transformation, illustrée par la Figure 1.9, permet de mieux **rendre compte de la perception des fréquences par l'oreille humaine**. Dans le cas des LFCC, ce filtre n'est pas appliqué.

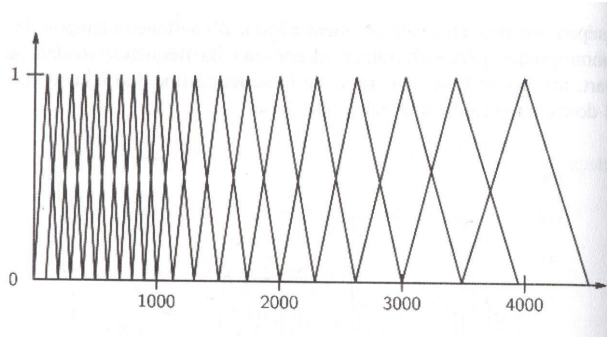


FIGURE *1.9* – *Banc de filtres à l'échelle Mel d'après (Haton et al., 2006)*

Le logarithme de ces valeurs est ensuite calculé. La dernière étape consiste à appliquer une transformée inverse en cosinus discrète (Calliope, 1989). La durée des fenêtres varie entre 20 millisecondes et 250 millisecondes selon les systèmes. La grande majorité des systèmes utilise des fenêtres de Hamming d'une durée de 20 millisecondes avec un pas

14. Les nasales sont produites en faisant passer de l'air dans la cavité nasale dont la forme est propre au locuteur ; en fin de groupe de souffle, on observe souvent un relâchement des articulateurs qui peut donner une information sur la taille du conduit vocal (Shriberg et Stolcke, 2008)

de 10 millisecondes. Ce choix permet de respecter l'hypothèse de stationnarité du signal mais a pour conséquence d'utiliser des trames dont la durée est bien inférieure à un segment de parole [15]. Le tableau 1.4 résume les différentes durées de fenêtres pour les paramètres cepstraux. Il est à noter que le pas de trame correspond le plus souvent à la moitié de la durée des trames.

Longueur de la trame (millisecondes)	Nombre de systèmes
20	19
25	14
30	9
32	2
250	3
Non renseigné	10

TABLE *1.4* – *Longueur de trames pour le calcul des coefficients cepstraux*

La dimension des vecteurs varie en fonction des systèmes. Si le nombre de coefficients cepstraux fluctue entre 12 et 20, la majorité des systèmes ayant participé à NIST-SRE 2010 utilise 19 coefficients cepstraux ainsi que c0 qui correspond à l'énergie présente dans la trame de signal analysé. Il est à noter qu'en reconnaissance de la parole, où nous sommes principalement intéressé à rendre compte de l'enveloppe spectrale pour reconnaître le phone prononcé, les systèmes utilisent en moyenne 12 coefficients cepstraux (Haton et al., 2006). En tenant compte d'un nombre de coefficients plus grand, nous supposons récupérer également l'information sur la source du signal de parole.

Coefficients issus d'une Prédiction linéaire

Les LPCC, quant à eux, s'appuient sur la technique de prédiction linéaire (Markel et Gray, 1976). Cette technique se fonde sur **la corrélation entre les échantillons successifs de parole, corrélation qui peut être attribuée aux résonances du conduit vocal d'après la théorie source-filtre** (Fant, 1970). Ces coefficients rendent compte de l'enveloppe spectrale. Une fois les coefficients de prédiction linéaires calculés, ceux-ci sont

15. A titre d'exemple, (Marchal, 2007) cite une durée entre 50 et 150 ms pour les occlusives (Ladefoged, 2005) illustre son propos avec des spectrogramme où les voyelles ont une durée autour de 150 millisecondes

transformés dans l'espace cepstral. La majorité des systèmes de NIST-SRE 2010 qui utilisent des LPCC a opté pour des vecteurs de 18 coefficients.
Les coefficients PLP (Furui, 1981), (Hermansky, 1990) permettent de **tenir compte de la perception non linéaire des fréquences par l'oreille humaine**. Ils reposent eux aussi sur une analyse fréquentielle du signal de parole, dont les valeurs seront transformées à l'aide d'une échelle perceptive Bark (Zwicker et Feldtkeller, 1981). Une transformée de Fourier inverse est ensuite appliquée sur ces coefficients afin d'obtenir les PLP. Les systèmes utilisent entre 12 et 15 coefficients PLP. Cette analyse peut être renforcée par une analyse spectrale relative (RASTA) qui simule l'insensibilité de l'oreille aux variations temporelles lentes (Hermansky et al., 1991).

Informations dynamiques

La majorité des systèmes (72% des systèmes pour NIST-SRE 2010) souhaitent tenir compte d'une information dynamique dans leurs vecteurs de paramètres. Pour cela, ils utilisent **les variations immédiates des paramètres acoustiques** en calculant les dérivées temporelles première (Δ) et seconde ($\Delta\Delta$) (Furui, 1981). La dérivée première peut être liée à **la vitesse de variation du spectre** tandis que la dérivée seconde rend compte de **l'accélération**. Nous retrouvons ces calculs de Δ et de $\Delta\Delta$ avec tous les types de paramètres (cepstraux ou PLP).

1.4.2 Autres paramètres utilisés

Éléments prosodiques

Peu de systèmes proposent d'utiliser des paramètres prosodiques [16]. Ceux qui les utilisent espèrent ainsi **capter des informations sur la source de production qui ramène notamment à la morphologie des plis vocaux** [17] (Ferrer et al., 2010). **Ces paramètres sont, dans tous les cas, combinés avec un système utilisant des paramètres acoustiques, généralement des coefficients cepstraux**.
Un des systèmes les plus connus utilisant ces paramètres est le système NERF développé par le SRI International (Scheffer et al., 2011). Dans ce système, les éléments prosodiques

16. Lors de NIST-SRE 2010, seulement 3 systèmes disent utiliser ce type de paramètres
17. Nous parlons ici de plis vocaux et non de cordes vocales car les plis vocaux contiennent diffréents éléments anatomiques (muqueuse, muscles et cartilage) qui rendent la métaphore de la corde caduque.

extraits sont la **courbe de la fréquence fondamentale** (F0, fréquence de vibrations des plis vocaux) et **la courbe d'intensité** ainsi que **la durée des segments** d'où sont extraits les éléments précédents. Les paramètres utilisés pour rendre compte des deux premiers éléments sont les coefficients de Legendre introduit par (Dehak et al., 2007) pour la RAL. Il s'agit de décrire ces courbes à l'aide de polynômes d'ordre 5. Différentes zones ont été testées, les éléments ayant été calculés sur **chaque syllabe**, sur les **vallées des éléments voisés** et sur des zones uniformes telles que définies dans (Kockmann et al., 2010).
Un des autres systèmes de NIST-SRE 2010 utilisant des indices prosodiques calcule les courbes de F0 et d'intensité sur des fenêtres fixes de 300 millisecondes, les trames non voisées étant rejetées. L'information de durée dans ce système consiste à compter le nombre de trames voisées sur un intervalle de 30 trames soit 300 millisecondes. Cette durée de trame correspond à peu près à la durée moyenne d'une syllabe.
Il est à noter que, précédemment, d'autres paramètres avaient été envisagés pour rendre compte des éléments prosodiques comme **la durée des mots**, **la durée des états** permettant de décrire les phones ou **la durée des pauses** (Shriberg et Stolcke, 2008). La combinaison de ces paramètres avec les systèmes fondés sur les coefficients cepstraux permet d'améliorer les performances. Par exemple, sur les données de NIST 2008, le système du SRI sans les paramètres prosodiques en données téléphoniques a un EER de 2.7% qui chute à 2.4% après la fusion avec le système utilisant des paramètres prosodiques (Ferrer et al., 2010).

Niveau glottique

Certains participants à NIST-SRE 2010 ont utilisé des paramètres rendant compte de la **source glottale** en utilisant des *Glottal Source Cepstral Coefficients* (GSCC)(Mazaira et al., 2010). Ces coefficients reposent sur une transformée inverse du signal de manière à ne décrire que la source du signal. Les auteurs notent une amélioration des performances en utilisant ces paramètres plutôt que les MFCC. Il est toutefois à noter que la performance initiale de leur système MFCC sont assez éloignée de l'état de l'art ($EER = 27.15$ en parole téléphonique, 2.5 minutes).

Niveau segmental, réalisation phonétique

La **déviance de la réalisation de certains phones** par rapport à un prototype est considérée comme un bon indice sur le locuteur par certains concepteurs de systèmes. Ainsi, la fréquence des phones absents de la langue parlée par le locuteur, a été envisagée comme paramètre par (Campbell et al., 2004). Dans ce cas, les performances du système sont améliorées (de 21.8% d'EER à 13.4% pour des fichiers de 2.5 minutes). (Klusacek et al., 2003) observent, quant à eux, la réalisation de certains phones selon leur contexte. Par cette technique, l'EER passe de 2.8% à 1.7% pour des fichiers de 20 minutes.

Niveau lexical

D'autres systèmes rendent compte du **vocabulaire employé par les locuteurs**, considérant qu'une information idiosyncratique est contenue dans le choix des idiomes utilisés par le locuteur. Si (Scheffer et al., 2011) utilisent les fréquences de mots comme paramètres, (Mirghafori et al., 2005) ne comptabilise que certains mots comme *actually, anyway, like, see, well, now* et des expressions telles que *you know, you see, I think, I mean* ou certains bruits de bouches tels que *um, uh*. Ces éléments permettent d'améliorer légèrement les performances. L'ajout de la fréquence des mots dans le système du SRI a permis de diminuer la DCF [18] de 0.471 à 0.298 lors de l'évaluation 2010 en condition téléphonique (Scheffer et al., 2011).

Synthèse : le rôle fondamental de l'évaluation

Le **paradigme de l'évaluation est un outil incontournable pour faire avancer la recherche**. Les avancées observées et les questions scientifiques abordées cette dernière décennie sont, dans leur majorité, dues au mode d'évaluation des systèmes. Ainsi, **en vérification du locuteur, d'importants efforts ont été consacrés à la question de la robustesse des systèmes aux modes d'enregistrement ou à la durée des signaux qui sont des questions centrales de l'évaluation NIST-SRE**. Les techniques du Factor Analysis et des i-vectors ont permis une progression drastique sur ces questions. Un autre exemple concerne les transcriptions dont l'utilisation a permis d'explorer des paramètres

18. Le taux d'EER n'est pas disponible dans ce papier, les performances du système n'étant exprimées qu'en terme de DCF.

dépendants du message délivré par le locuteur.
De plus en plus de soumissions à NIST-SRE combinent des approches différentes afin d'obtenir de meilleures performances. Si certaines tentent d'introduire des données liées à la connaissance de la production de la parole ou faire des hypothèses sur la localisation l'information idiosyncratique, **une grande majorité préfère combiner de nombreux systèmes en utilisant différents paramètres cepstraux et en jouant sur les dimensions des modèles**.
Il est à noter que la majorité des paramètres acoustiques utilisés en RAL sont également utilisés en reconnaissance automatique de la parole où le but est de transcrire ce qui est dit dans le signal de parole. Ceci met en évidence que ces **paramètres rendent plus compte du signal de parole en général que de spécificités du locuteur en tant que tel**. Par ailleurs, **l'analyse des erreurs commises par les systèmes peut permettre de comprendre où se situe l'information pertinente pour la vérification du locuteur.**. Comme l'évaluation des systèmes de vérification du locuteur se fait de manière globale, l'analyse des erreurs n'est pas beaucoup développée. **Il est actuellement difficile de comprendre l'influence des éléments de variations de la parole sur les performances des systèmes car la constitution des corpus ne permet pas toujours de rendre compte de l'influence exacte d'un facteur, toute chose n'étant pas égale par ailleurs**.
Ces campagnes d'évaluation ont déjà permis de poser des **profils de locuteurs différents**. (Doddington et al., 1998) met en évidence qu'il existe des comportements de locuteurs et propose de décrire ces comportements à l'aide d'une ménagerie métaphorique. Les locuteurs au comportement normal [19] sont appelés *moutons*. Les locuteurs dont les enregistrements utilisés en apprentissage donnent lieu à de nombreuses FA sont appelés *agneaux* tandis que ceux qui donnent lieu à de nombreux FR sont appelés *chèvres*. Enfin, les locuteurs dont les enregistrements utilisés en tests donnent lieu à de nombreux FA sont appelés des *loups*. La « chasse aux loups » a notamment été étudiée par (Stoll et Doddington, 2010). Dans ces travaux, **la distinction entre fichiers d'enregistrement et locuteur n'est pas toujours évidente**, pourtant un locuteur peut produire plusieurs enregistrements différents.
Cette distinction entre enregistrement et locuteur est complètement niée dans l'approche des i-vectors qui considère qu'un enregistrement est équivalent à un locuteur. L'enregistrement reste le support sur lequel la parole du locuteur a été enregistrée et **le défi de la vérification du locuteur est de conserver les indices sur le locuteur tout en s'extrayant des éléments de variation** comme la langue ou l'effort vocal. Il est in-

19. tel qu'estimé à partir d'exemples d'enregistrements

téressant de noter que l'introduction, dans les campagnes NIST SRE, de nouveau type d'enregistrement se focalisant sur l'effort vocal ou l'influence de l'évolution de la voix dans le temps ouvrent actuellement de nouvelles perspectives en ce sens.
La question des indices idiosyncratiques et de leur relative indépendance avec les autres facteurs de variation de la parole est centrale et nécessite de poser la distinction entre enregistrement et locuteur lors des évaluations des systèmes de RAL.

Chapitre 2

Reconnaître son interlocuteur : une capacité humaine à évaluer

Résumé : *Ce chapitre est consacré à la capacité humaine à reconnaître des voix. Cette capacité est évaluée par différents protocoles depuis de nombreuses années. Les expériences menées montrent que cette capacité est variable selon les conditions : il est indispensable de séparer les tâches où les personnes sont déjà connues des auditeurs de celles où il s'agit de nouvelles personnes car les performances sont très différentes selon les conditions. D'autres éléments comme la longueur des extraits, la langue des locuteurs ou l'état émotionnel du locuteur jouent un rôle important dans les performances des auditeurs. Suite à la description de ces capacités, nous présentons les modèles cognitifs qui ont été envisagés pour expliquer la perception des voix. Enfin, nous établissons la liste des paramètres idiosynchratiques qui ont été mis en évidence par tests perceptifs ou par analyses acoustiques.*

Sommaire

2.1 Capacité humaine à reconnaître un locuteur

L'évaluation des capacités humaines à reconnaître une personne à partir de sa parole est une question qui intéresse les phonéticiens depuis longtemps. Cet intérêt n'a tout de même pas donné lieu, contrairement à l'évaluation des systèmes automatiques, à un véritable protocole de référence même si des propositions dans ce sens ont été faites (Broeders et Amelsvoort, 1999).

2.1.1 Protocoles d'évaluation et métriques

Les protocoles d'évaluation des capacités humaines à reconnaître son interlocuteur sont très variés et pas toujours comparables (Van Lancker et Kreiman, 1987). Avant de présenter les performances obtenues par l'humain, nous nous intéresserons aux métriques employées en fonction de la tâche choisie.

Similarité entre les voix et test d'authentification

Une première approche pour évaluer la capacité humaine à reconnaître une personne à partir de sa voix consiste à présenter aux auditeurs une paire d'enregistrements vocaux et de les interroger sur la similarité entre les locuteurs (Erikson, 2007). La réponse des auditeurs peut être de type binaire : il s'agit du même locuteur ou bien il s'agit de deux locuteurs différents. Ceci correspond à la tâche de vérification du locuteur telle que définie en informatique. Elle a notamment été utilisée par (Schlichting et Sullivan, 1996). Si elle reste minoritaire pour les tests perceptifs, elle est très courante pour les analyses acoustiques (Rogers, 1998). La métrique généralement utilisée dans ce cadre est le taux de réussite global, TG. Ce taux consiste à comptabiliser toutes les réponses correctes (qu'elles soient issues de comparaisons cible ou de comparaisons imposteur) par auditeur puis de faire la moyenne de ces taux, comme illustré par les équations 2.1 et 2.2.

$$TG = \frac{\sum_{k=1}^{n} TG_k}{n} \tag{2.1}$$

où TG_k est le taux de réussite pour chaque auditeur (indicé par k) calculé selon l'équation 2.2 et où n est le nombre d'auditeurs.

$$TG_k = \frac{\text{Nombre de réponses correctes pour l'auditeur } k}{\text{Nombre de comparaisons écoutées par l'auditeur } k} \tag{2.2}$$

La réponse des auditeurs peut également être fournie à l'aide d'échelles de proximité où la plus petite valeur indique qu'il s'agit du même locuteur tandis que la plus élevée signifie que les locuteurs sont très différents. Ceci est notamment utilisé par (Remez et al., 2004) ou (Kahn et Rossato, 2009). Dans ce cadre, ce sont des matrices de distances qui sont analysées. Ce type de tâche, même s'il se développe de plus en plus, reste toutefois marginal par rapport à la tâche d'identification.
Une autre métrique utilisée en discrimination est la mesure d-prime (Macmillan et Creelman, 1991). Cette mesure tient compte de la tendance des auditeurs à répondre systématiquement de la même façon (toujours oui ou toujours non). Elle obtenue par l'équation 2.3.

$$d - prime = z(p(\text{"oui"}|H_0)) - z(p(\text{"oui"}|H_1)) \tag{2.3}$$

où $z()$ est une normalisation des données pour les centrer et les réduire, $p(\text{"oui"}|H_0)$ est la proportion de fois où l'auditeur a répondu qu'il s'agissait du même locuteur alors qu'il s'agissait du même locuteur (Vraie Acceptation) et $p(\text{"oui"}|H_1)$ est la proportion de fois où l'auditeur a répondu qu'il s'agit du même locuteur alors que les enregistrements proviennent de deux locuteurs différents (Fausses Acceptations).
Dans ce cadre, si $d - prime = 0$, cela signifie que l'auditeur a répondu au hasard. Un $d - prime$ de 1 signifie que 69% des comparaisons imposteur et comparaisons cible ont été correctement reconnues. Le maximum qui puisse être atteint est 6.93. La figure 2.1 illustre le lien entre d-prime et une courbe ROC.

Identification du locuteur

L'identification du locuteur consiste à retrouver un locuteur dans un panel d'enregistrement. Cette tâche, majoritaire, est notamment utilisée par (Schlichting et Sullivan, 1998) ou (Lavner et al., 2000). Des expériences où il faut directement nommer le locuteur ont également été réalisées (Van Lancker et al., 1985). Si la voix est nouvelle pour les auditeurs, une phase d'apprentissage est prévue. Il se peut que soit demandé aux

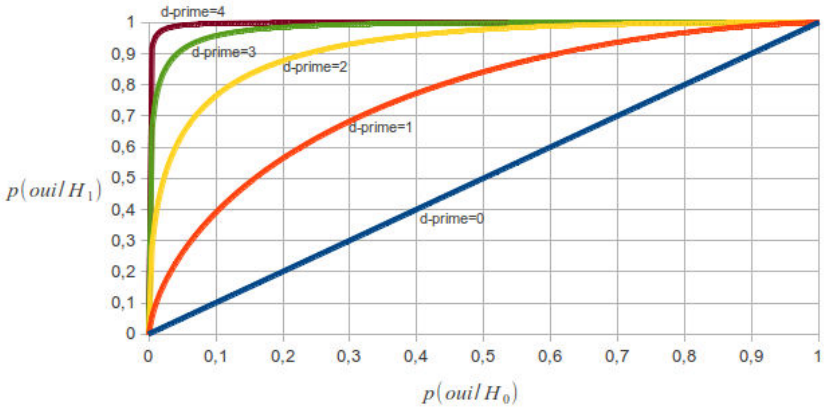


FIGURE 2.1 – *Lien entre les valeurs de d-prime et la proportion de Fausses Acceptations et de Vraies Acceptations.*

auditeurs de donner leur confiance dans leur réponse à l'aide d'une échelle de confiance. Le chiffre le plus haut désigne généralement qu'ils sont très confiants dans leur réponse.

Plusieurs métriques ont été envisagées pour ce type de tâche. La première est le taux d'identification ou son homologue le taux d'erreurs. Le taux d'identification, $T_{identification}$, tient compte du nombre de fois où l'auditeur a identifié correctement le locuteur comme le retranscrit l'équation 2.4.

$$T_{identification} = \frac{Nombre\ d'identifications\ correctes}{Nombre\ total\ d'identifications\ à\ réaliser} \tag{2.4}$$

Le taux d'erreurs, $T_{erreurs}$, correspond au pourcentage de fois où les auditeurs se sont trompés dans leur identification : ils ont déclaré que l'enregistrement de parole provenait d'un autre locuteur que celui qui a réellement effectué l'enregistrement. Ce taux est l'autre face du taux d'identification (cf. équation 2.5).

$$T_{erreurs} = 1 - T_{identification} \tag{2.5}$$

La performance des auditeurs peut également être exprimée en comptabilisant le nombre de fois où les auditeurs ont identifié un locuteur alors que ce n'était pas lui (fausse acceptation, $FA_{identification}$, équation 2.6) et le nombre de fois où le locuteur a été pris pour un autre (faux rejet, $FR_{identification}$, équation 2.8). Ces taux sont obtenus à partir de

la matrice de confusion des réponses, MC.

$$FA_{identification} = \sum_{i=1}^{n} FA_{loc_i} \tag{2.6}$$

avec FA_{loc_i} défini par l'équation 2.7

$$FA_{loc_i} = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} t_{i,j}}{n} \tag{2.7}$$

$$FR_{identification} = \sum_{j=1}^{n} FR_{loc_j} \tag{2.8}$$

avec FR_{loc_j} défini par l'équation 2.9

$$FR_{loc_j} = \frac{\sum_{\substack{i=1 \\ i \neq j}}^{n} t_{i,j}}{n} \tag{2.9}$$

où i et j sont respectivement l'indice de la colonne et de la ligne de la matrice de confusion MC et n est le nombre de locuteurs. La Figure 2.2 illustre ces éléments.

		Réponses des auditeurs			
		Loc. 1	Loc. 2	Loc. 3	Loc. 4
Références	Loc. 1	Case prise en compte pour le calcul de Taux identification du Loc. 1	Cases prises en compte pour le calcul de faux rejets pour le Loc 1.		
	Loc. 2	Cases prises en compte pour le calcul de fausses acceptations pour Loc. 1			
	Loc. 3				
	Loc. 4				

FIGURE 2.2 – *Éléments pris en compte lors de l'expression de la performance à l'aide d'un* $FA_{identification}$ *et d'un* $FR_{identification}$

2.1.2 Des performances inégales

Identifier des personnes connues

L'humain sait dès son plus jeune âge reconnaître des personnes à partir d'un enregistrement de parole (Mehler et Dupoux, 1990). A quatre semaines, l'enfant est plus sensible à la voix de sa mère qu'à la voix d'une autre femme lorsqu'il entend des enregistrements réalisés alors que ces femmes s'adressent à leurs enfants (Miles et Meluish, 1974). Cette sensibilité est mesurée à l'aide de la technique de la succion non-nutritive (Siqueland et DeLucia, 1969) qui consiste à mesurer à l'aide d'une tétine à capteur, le nombre de succions réalisées par le nourrisson. Une augmentation du nombre de succions est considérée comme un indicateur du fait que l'enfant est intéressé par ce qui se passe autour de lui. Dans le cas de cette expérience, le taux de succions augmente de 18% lorsque l'enfant entend la voix de sa mère après avoir entendu la voix d'une autre femme, alors que le taux de succion diminue de 35% lorsque l'enfant entend une voix inconnue après avoir entendu la voix de sa mère. **La voix de sa mère est donc connue de l'enfant dès ces premières semaines de vie et il ne réagit pas de la même façon à cette voix qu'à une autre.**

L'humain a d'ailleurs une facilité à reconnaître, à partir de parole, des personnes qui lui sont connues. Une personne connue peut être aussi bien une célébrité (homme politique, star de variété) (Schlichting et Sullivan, 1996) ou (Van Lancker et al., 1985), qu'un membre de la famille (Ladefoged, 1980) ou un collègue de travail (Hollien et Doherty, 1982). Il s'agit de personnes que les auditeurs ont l'habitude d'entendre. **Dans ce cas, l'humain reconnaît très bien les personnes**, les taux d'identification s'élèvent selon les circonstances (nombre d'auditeurs étudiés, longueur des enregistrements, panel ou nomination...) autour de 98% pour (Hollien et Doherty, 1982), 83% pour (Ladefoged, 1980) ou 69.9% pour (Van Lancker et al., 1985).

La durée des enregistrements joue un rôle important dans les différences de performances. Plus les enregistrements ont une durée importante, plus les taux d'identification sont élevés. Ainsi, pour (Ladefoged, 1980), les taux d'identification sont de 31%, 66% et 83% lorsque le locuteur prononce respectivement « *hello* », une phrase simple ou un message d'une durée de 30 secondes.

L'état émotionnel influence également le choix des auditeurs. (Saslove et Yarmey, 1980) notent que les auditeurs reconnaissent mieux les locuteurs lorsque, dans les deux enregistrements à comparer, l'état émotionnel est le même plutôt que lorsqu'un des en-

registrements est réalisé avec le locuteur en colère. Lorsque les 15 auditeurs réalisent cette tâche d'identification, tous répondent parfaitement quand le locuteur a une voix neutre tandis que seulement quatre auditeurs obtiennent 100% de bonnes réponses lorsque le locuteur s'exprime avec colère. Cette différence est significative et montre que l'état émotionnel du locuteur peut changer les capacités d'identification de l'auditeur. Ce phénomène est d'ailleurs également souligné par (Scherer et al., 1998) dans sa revue générale sur l'influence de l'état émotionnel du locuteur sur la parole.
Au delà de la familiarité, c'est **la question du contexte** où la voix est entendue qui est posée. Par exemple, (Bricker et al., 1976) ont enregistré leurs collègues de travail au laboratoire et ont demandé à d'autres collègues de les identifier. Que se serait-il passé si les locuteurs avaient été enregistrés sur leurs lieux de vacances et avaient été identifiés par leurs collègues ? Des résultats similaires auraient-ils été trouvés ? A notre connaissance, aucune expérience d'identification de personnes à partir de la voix n'a posé la question dans ces termes. Toutefois, (Campbell et Erickson, 2004) en suivant une femme pendant trois ans dans différents contextes de communication montre que les indices prosodiques varient en fonction de la personne à qui elle s'adresse. De même, (de Looze et al., 2011) montre que les courbes prosodiques des interlocuteurs ont tendance à se rejoindre lorsqu'il y a connivence dans le dialogue. Dans ce cadre, nous pouvons nous demander si la familiarité de la voix n'est pas dépendante du contexte. Une première réponse peut se trouver dans la réaction du nourrisson qui ne réagit plus à la voix de sa mère lorsque celle-ci lit un texte destiné à un adulte (Mehler et al., 1976).

Une difficulté accrue pour les nouvelles voix

La connaissance préalable du locuteur par l'auditeur est un facteur qui influe grandement les performances des auditeurs. Ainsi, (Hollien et Doherty, 1982) par un test d'identification sur un groupe de dix locuteurs, montre que « alors que le taux d'erreur est de 60% pour des auditeurs qui ne connaissent pas les locuteurs, il tombe à 2% pour des auditeurs familiers ». Des résultats équivalents sont trouvés par (Bricker et al., 1976) ou (Ladefoged, 1980) dans des contextes quelque peu différents. Les performances des auditeurs, lorsque ceux-ci n'ont entendu la voix de cette nouvelle personne que quelques instants avant de réaliser l'identification, varient grandement : 26.6% pour (Van Lancker et al., 1985), 46% pour (Ladefoged, 1980) et jusqu'à 82% (Blatchford, 2006) de taux d'identifications correctes. Cette capacité à apprendre à re-

connaître de nouvelles personnes à partir de leur parole est d'ailleurs détectée par (Spence, 1996) qui, toujours en mesurant des taux de succions, montre que les nourrissons d'un mois ne réagissent pas de la même manière face à l'enregistrement d'une infirmière qui s'occupe d'eux depuis trois jours et face à l'enregistrement d'une nouvelle infirmière. Le nombre de succions est significativement différent entre les deux types d'infirmières ($F(2,21) = 5.9, p < 0.01$ avec Nb succions$_{\text{infirmières connues}} = 12$ vs Nb succions$_{\text{infirmières non-connues}} = 10$ [1]).

La durée des enregistrements joue aussi dans l'identification de nouvelles voix un rôle important. Dans le cas de nouvelles personnes, (Ladefoged, 1980) montre que les taux d'identification diminuent si la durée des enregistrements diminue. Dans ce cas, ils sont de 17%, 36% et 46% respectivement pour « *hello* », une phrase simple et 30 secondes de parole. Ce résultat est confirmé par (Blatchford, 2006) : les auditeurs arrivent mieux à reconnaître la locutrice quand celle-ci prononce un énoncé du type *Face down on the ground and hands behind your back now !* ($T_{identification} = 82\%$) que lorsqu'elle prononce un énoncé du type *Get him* ($T_{identification} = 46\%$). Il est toutefois à noter que (Legge et al., 1984) montre que, pour les personnes inconnues, lorsque la longueur des enregistrements dépasse la minute, les taux de bonnes identifications diminuent ($T_{identification} = 70\%$ avec 1 minute de parole vs $T_{identification} = 58\%$ avec 2 minutes de parole). Trouver la bonne durée d'enregistrement, celle qui permet à l'auditeur d'être dans les conditions optimales pour reconnaître la personne, n'est donc pas chose si aisée. Il ne suffit pas que l'auditeur soit exposé à l'enregistrement le plus long de la parole du locuteur pour maximiser sa performance.

Par ailleurs, **le temps intervient aussi dans la mémorisation des indices**. Quand l'enregistrement de la personne n'est pas ré-entendu, la performance des auditeurs diminue avec le temps comme le montre l'expérience de (McGehee, 1937) où les auditeurs doivent identifier une personne après avoir entendu un extrait de parole il y a un jour ($T_{identification} = 83\%$), deux jours ($T_{identification} = 83\%$), deux semaines ($T_{identification} = 68\%$), trois semaines ($T_{identification} = 51\%$) et cinq mois ($T_{identification} = 13\%$). (McGehee, 1937) ne présente pas aléatoirement ses stimuli ; ce que lui reproche (Saslove et Yarmey, 1980) qui ne trouve pas de différences significatives ($F < 1.0$) entre les performances des deux cohortes testées, l'une réalisant la tâche d'identification juste après avoir entendue la parole de la personne à identifier, l'autre effectuant l'identification 24 heures plus tard.

1. Significativité mesurée à l'aide d'un test de Fisher, la notation se lie de la manière suivante : F(Degrès de liberté du facteur, Degrès de liberté du résidu)=F-mesure (Wonnacott et Wonnacott, 1991)

Au delà de la présentation aléatoire ou non-aléatoire, certaines études autorisent les auditeurs à écouter autant de fois qu'ils le souhaitent les enregistrement tandis que d'autres les obligent à répondre après une seule écoute. Nous n'avons pas trouvé d'expérience qui mesure l'impact du protocole sur les performances des auditeurs, mais nous tenions à souligner que les deux procédures existent dans la littérature.

Variations de performance selon les enregistrements

La grande majorité des résultats présentés sur les capacités humaines de reconnaissance des locuteurs a été obtenue avec des signaux dont les enregistrements sont effectuées dans des conditions comparables. Les enregistrements à comparer ont été enregistrés dans les mêmes conditions, le plus souvent en chambre sourde avec un microphone. Ces conditions expérimentales sont très différentes de celles que nous avons pu observées pour les systèmes de RAL où la variabilité due aux conditions d'enregistrement est un phénomène étudié depuis le début des campagnes d'évaluation. (Alexander et al., 2004) compare les performances des humains et des systèmes au cours d'un test d'authentification avec des enregistrements qui sont systématiquement enregistrés avec du matériel différent. Dans ce cas, la performance des humains est de 35% de EER contre 20% lorsque les enregistrements sont faits avec le même matériel.
La langue est un filtre qui joue également un rôle important dans le processus de reconnaissance des locuteurs. Par exemple, des auditeurs anglophones reconnaissent mieux des locuteurs anglophones que des locuteurs germanophones (Goggin et al., 1991) ($T_{identification} = 40\%$ pour les anglophones vs $T_{identification} = 12\%$ pour les germanophones). De même, des auditeurs de langue seconde suédoise reconnaissent moins bien un locuteur suédois qu'un natif du suédois (Schlichting et Sullivan, 1996) et (Sullivan et Schlichting, 2000) lorsque la personne à reconnaître est le premier ministre suédois. $T_{identification}$ pour les natifs est de 93% contre 56% pour les non-natifs ayant une connaissance du suédois. Les apprenant du suédois le reconnaissent néanmoins mieux que des auditeurs n'ayant aucune connaissance du suédois (Sullivan et Schlichting, 2000) ($T_{identification}$ pour les auditeurs sans connaissance du suédois est de 16%). Il faut tout de même noter que, dans ce cadre expérimental, pour les auditeurs qui ne parlent pas du tout le suédois, le premier ministre suédois ne fait sûrement pas partie des personnes connues, alors que c'est le cas pour les suédois. L'accent avec lequel parle le locuteur peut également suffire pour tromper les auditeurs. (Sjostrom et al., 2006) après

avoir enregistré un locuteur qui parle deux dialectes de Suède, montre que les performances des auditeurs sont meilleures lorsque l'identification du locuteur est faite avec deux enregistrements où le locuteur parle de la même manière que lorsqu'il n'y a pas adéquation de dialecte entre les deux extraits de parole.

Variation de performance selon les auditeurs

Tous les humains ne réussissent pas dans les mêmes proportions à reconnaître une personne à partir d'un extrait de parole. La spécificité des phonéticiens (à la fois à l'écoute et en lecture spectrographique) a beaucoup été étudiée et les résultats restent controversés.
D'une part, (Schiller et Köster, 1998) montre qu'un panel de phonéticiens réussit légèrement mieux qu'un panel d'auditeurs naïfs à identifier des locuteurs. Les auditeurs naïfs ont obtenu un taux d'identification de 92% avec 2% de $FA_{identification}$ et 7% de $FR_{identification}$ tandis que les phonéticiens ont obtenus 98% de Taux d'identification, 1% de $FA_{identification}$ et 1,6% de $FR_{identification}$.
Les résultats de (Reich et al., 1976) tendent à montrer que les performances sont similaires lorsque les enregistrements sont réalisés sans trucage ($T_{Identification} = 92\%$). Les auteurs soulignent que les phonéticiens discriminent mieux les locuteurs que les naïfs lorsque les locuteurs ont cherché à déguiser leur voix en utilisant par exemple des mouchoirs sur leur bouche ($T_{identification} = 59\%$ pour les naïfs contre $T_{identification} = 81\%$ pour les phonéticiens).
(Stevens et al., 1957) observent quant à eux des différences de performance importantes entre leurs auditeurs qui sont tous naïfs. Dans le cas d'identification auditive de voix familières en milieu fermé, les auteurs soulignent une grande disparité entre les auditeurs : en effet, le meilleur réalise moins de 5% de taux d'erreur tandis que le moins bon réalise environ 16% de taux d'erreurs.

.L'humain sait donc reconnaître des locuteurs à partir d'extrait de parole mais cette capacité dépend de certaines circonstances. Ces éléments ont conduits les chercheurs à proposer des modèles sur les processus cognitifs impliqués dans la reconnaissance.

2.2 Les processus cognitifs impliqués

Nous avons vu dans la section précédente que l'humain sait reconnaître des locuteurs à partir d'extraits de parole mais que cette capacité dépend de nombreux facteurs. Ces éléments ont conduit les chercheurs à proposer des modèles sur les processus cognitifs impliqués dans la reconnaissance des locuteurs.

2.2.1 Prototypes

Le parallèle entre la reconnaissance des locuteurs par leur parole et par leur visage est notamment proposé par (González et al., 2011). Selon cette théorie, dans les deux cas, l'auditeur encode un prototype (de parole ou de visage) qui sert de référence pour la reconnaissance. En perception multimodale, le sujet se sert à la fois des indices visuels et auditifs pour reconnaître la personne. La figure 2.3 illustre ce modèle.

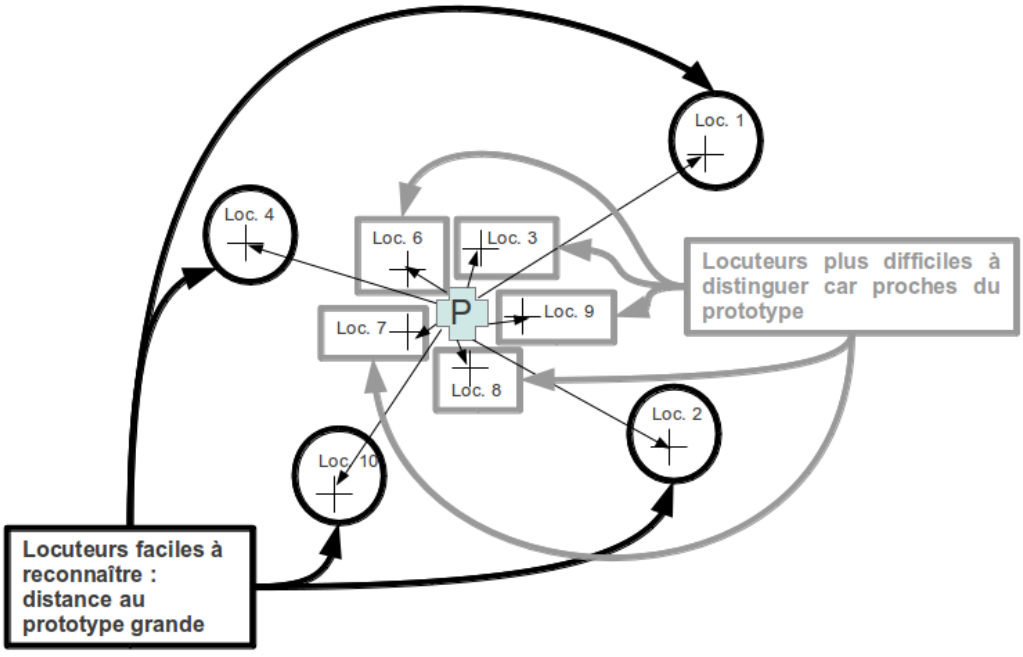


FIGURE 2.3 – *Théorie des prototypes*

Reconnaissance des visages

Pour les visages, cette théorie s'appuie sur différentes expériences. Tout d'abord, (Goren et al., 1975) montre que les nourrissons, dès leurs premières heures, sont plus sensibles à des patterns qui respectent l'organisation d'un visage que des patterns qui, bien que composés des mêmes éléments, ne respectent pas l'organisation du positionnement des yeux, du nez et de la bouche. Ils suivent plus le pattern où l'ordre des éléments du visage sont respectés que les autres ($p < 0.001$, le nourrisson tourne sa tête à 166.75 pour suivre le pattern qui respecte l'organisation du visage contre 141.25 degrés pour celui où les éléments sont mélangés). Les nourrissons semblent donc présenter une prédisposition à regarder des visages.
Le prototype se construit peu à peu à partir de l'ensemble de visages observés. Ainsi, (Valentine et al., 1995) montre que les sujets ont tendance à mieux reconnaître les visages qui sont de la même couleur de peau que la leur. Alors qu'ils ont peu de contact avec la personne blanche à identifier visuellement, les sujets blancs réussissent à 95% l'identification alors que les sujets noirs réussissent à 72%. Ce phénomène est bien expliqué par le fait que les sujets sont plus souvent exposés à des visages de la même couleur que la leur. (Sangrigoli et al., 2005) confirme cette hypothèse en mettant en évidence que des enfants asiatiques adoptés très tôt en France obtiennent les mêmes résultats dans l'identification de visages de type caucasien que des enfants d'origine caucasienne ($T_{identification} = 96\%$ pour les adoptés vs $T_{identification} = 94\%$ pour les Français de type caucasien).

Reconnaissance des voix

Pour les extraits de parole, selon la théorie du prototype, **l'auditeur construit à travers toutes les paroles qu'il a entendues, un prototype générique de ce qu'est la parole des locuteurs**. Lorsqu'il rencontre une nouvelle personne, il se réfère à ce prototype pour apprendre la nouvelle voix, en retenant en quoi l'enregistrement de la parole diffère du prototype (Lavner et al., 2001). Plus la personne lui est familière, moins il fait appel au prototype comme référence et il encode uniquement les différences au prototype. Ceci expliquerait pourquoi les humains sont plus rapides à reconnaître une personne connue qu'une personne inconnue (Papcun et al., 1989).
Le prototype est construit à partir de la parole déjà rencontrée, ainsi l'auditeur con-

struit son prototype majoritairement à l'aide de parole de personnes qui parlent la même langue que lui. Ceci pourrait expliquer les différences de performance observée en fonction des langues.
Par ailleurs, (Lavner et al., 2001) montre que tous les locuteurs inconnus ne sont pas tous aussi faciles à reconnaître pour un panel d'auditeur. Il suggère que certains indices sur le locuteur s'écartent plus que d'autres du prototype, ce qui expliquerait que certains locuteurs soient plus faciles à reconnaître que d'autres (Sullivan et Schlichting, 2000).
Ce modèle, majoritaire chez les psychologues, se rapproche en informatique de l'approche UBM-GMM (Reynolds et al., 2000) qui se fonde sur un modèle générique de ce qu'est la parole et qui adapte ensuite à l'enregistrement du locuteur les moyennes de ce modèle.

2.2.2 Jeux de paramètres acoustiques

(Bricker et al., 1976) suggère plutôt que **l'auditeur extrait des indices du signal acoustique et compare ces indices avec ceux des paroles de locuteurs qu'il connaît déjà pour évaluer la similarité entre l'extrait à évaluer et ceux qu'il a stocké**. Dans ce cas, l'auditeur ne fait pas appel à un prototype mais compare directement le nouvel exemplaire avec les exemplaires mémorisés.
La fréquence d'apparition de la voix du locuteur joue alors un rôle primordial. Plus l'auditeur a été soumis à un nombre important d'exemples de parole d'un locuteur, plus sa connaissance de la voix est précise. Ici encore, une explication de la différence entre voix connue et voix inconnue est donnée.
Un argument avancé par les teneurs de cette théorie est que les auditeurs ne répondent pas à la même vitesse quand il s'agit de discriminer deux locuteurs et de discriminer des phonèmes (Cutler et al., 2011). Le traitement du segment est plus rapide que celui du locuteur. Le fait que le processus soit plus long pour la voix suggère que des traitements préalables doivent être réalisés. L'hypothèse, dans ce cadre, est que les traitements préalables en question concerne l'extraction de paramètres pertinents pour mémoriser les locuteurs. Ces paramètres seraient différents de ceux utilisés dans la parole.
La figure 2.4 illustre cette théorie.

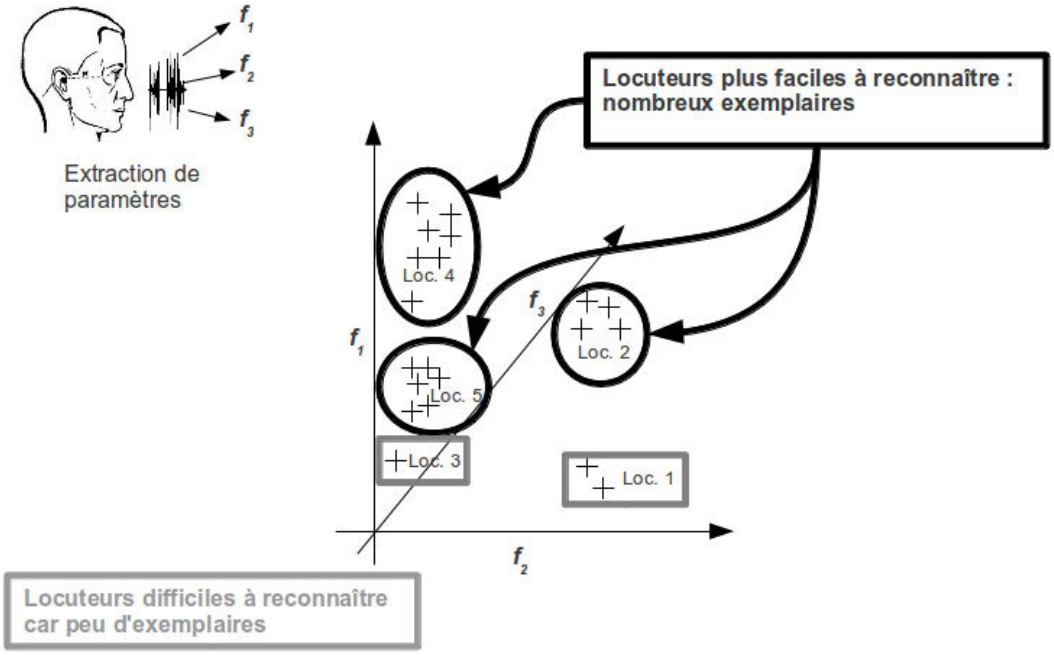


FIGURE 2.4 – *Reconnaissance des locuteurs par les jeux de paramètres acoustiques*

Ce modèle rejoint l'approche par modèles d'ancrage (Mami, 2003) qui cherche à situer les locuteurs les uns par rapport aux autres et non à voir comment le locuteur diffère de la moyenne.

2.2.3 Ce que nous apprend la phonoagnosie

La phonoagnosie est un symptôme qui se traduit par le fait que le patient a des difficultés pour reconnaître les personnes de sa propre famille à partir d'un extrait de parole (Lancker et Canter, 1982). Cette maladie est l'occasion de mieux comprendre comment l'humain utilisent les indices pour reconnaître ses interlocuteurs.
Comme cette maladie est assez rare, (Garrido et al., 2009) étudie les capacités d'une seule patiente anglophone atteinte de phonoagnosie.

Une déficience pour reconnaître les personnes à partir d'un extrait de parole

Dans un premier temps, (Garrido et al., 2009) montrent que leur patiente a plus de difficultés pour réaliser certaines tâches où intervient un enregistrement vocal qu'un panel de huit auditrices témoins (anglophones et ayant le même âge que la patiente).
Ainsi, elle n'arrive pas déterminer si la voix qu'elle entend lui est connue ou non alors qu'elle dit connaître les personnes à identifier. Par ailleurs, elle n'arrive pas à identifier la personne lorsqu'elle dit connaître la voix (3% d'identification correcte pour la patiente alors que les taux d'identification des sujets témoin varient de 40% à 65%).
Elle arrive également moins bien que le panel à apprendre de nouvelles voix anglophones. Dans ce cas, les auditeurs ont à apprendre les six voix de locuteurs inconnus tous originaires de la même région anglophone en associant à chaque voix un prénom. Les auditeurs écoutent ensuite un autre enregistrement, ils doivent dire s'il s'agit de la même personne ou non que le prénom indiqué. Cette expérience est répétée quatre fois de suite. Si pour le premier bloc de répétitions, la patiente réussit aussi bien que le panel de contrôle, au fur et à mesure de l'expérience la patiente arrive de moins en moins à discriminer les voix. Par la suite, elle n'arrive pas à identifier les voix qu'elle vient d'apprendre et elle n'est pas capable de distinguer les voix qu'elle vient d'apprendre de nouvelles voix anglophones (12% d'identification pour la patiente contre entre 40% et 55% pour le panel témoin).

Aucune différence avec d'autres auditeurs pour...

Si la patiente n'arrive pas à reconnaître des personnes connues grâce à un enregistrement de parole, elle sait reconnaître les gens à partir de leur visage aussi bien que le panel.
La patiente sait aussi bien que les autres auditrices discriminer des locuteurs qui parlent dans une langue étrangère ou qui parlent dans le bruit ($TG = 61\%$ pour la patiente vs $TG = 62\%$ pour le panel témoin). Elle sait également aussi bien qu'elles reconnaître le genre des locuteurs en milieu bruité.
La patiente ne présente pas de problème particulier en terme de compréhension des énoncés : elle réussit aussi bien que le panel témoin à associer un énoncé oral à la situation qu'il décrit ou à répéter des mots après les avoir entendus. Elle n'a d'ailleurs aucun mal à mémoriser des listes de mots lors d'un *Rey Auditory Verbal Learning Test*

(Rey, 1964).[2] Le niveau segmental ne lui pose pas plus de problème puisqu'elle sait aussi bien que le panel identifier des voyelles.
La patiente sait identifier, aussi bien que le panel, des émotions. Les émotions testées sont les six émotions suivantes : la joie, la tristesse, la peur, la surprise, la colère et dégoût. Ces tests sont réalisés à l'aide de données verbales dans un premier temps puis à l'aide de bruits de bouche dans un second temps. Les émotions sont identifiées à 82% et 72% par le panel témoin contre 83% et 70% par la patiente pour respectivement le premier et le second cas.
Les sons qui ne sont pas des sons de parole ne lui posent pas de problème particulier : elle sait aussi bien que le panel associer un bruit à une image ou nommer des objets ou des animaux à partir de sons. Elle sait également discriminer des instruments de musique.
Les aptitudes de cette patiente atteinte de phonoagnosie montrent tout d'abord que le processus cognitif de reconnaissance de locuteur à partir de la parole est différent de celui employé pour reconnaître le message délivré par la parole. Elle n'a pas de difficulté pour comprendre ce qui est dit ou pour identifier des phonèmes. Il ne s'agit donc pas uniquement de percevoir comment le locuteur dit les choses. Par ailleurs, la reconnaissance des locuteurs n'est pas traitée comme celle d'instruments de musique. Il ne s'agit donc pas uniquement d'une question de mélodie. Enfin, les processus pour reconnaître les voix et les visages ne sont pas exactement les mêmes puisque la patiente n'est pas atteinte de protoagnosie.
Ces études indiquent bien que la reconnaissance de la personne par un extrait de parole est un processus cognitif à part entière qui ne se fondent pas sur les mêmes processus que la reconnaissance de la parole ou la reconnaissance des émotions. Quels sont les indices utilisés pour reconnaître les locuteurs ?

2.3 A la recherche d'indices idiosynchratiques

Dès 1927, (Sapir, 1927) distingue cinq éléments pour décrire la parole d'un locuteur. Il s'agit de « qualité de voix, la prosodie [(rythme et intensité)], la prononciation [(les éléments phonétique et phonologique)], le vocabulaire et le style ». Dans cette partie, nous souhaitons lister les indices acoustiques, qui ont été proposés depuis cette descrip-

2. Ce test consiste à faire apprendre une première liste de mots en répétant 5 fois chaque élément de la liste puis de demander au patient de redonner les mots. Le patient doit, par la suite, apprendre une seconde liste de mots. Enfin, il doit répéter la première liste qu'il a apprise.

tion, suite à des analyses du signal de parole et/ou des tests perceptifs.Il s'agit pour nous d'établir une liste des indices pertinents afin de pouvoir par la suite les utiliser pour expliquer à quoi est sensible le système de RAL.

2.3.1 Fréquence fondamentale

La fréquence fondamentale (F0) correspond à la fréquence à laquelle vibrent les plis vocaux lorsque la glotte est fermée (phonation). Elle ne peut être mesurée que sur les sons dits voisés car lorsque la glotte est ouverte (sons non-voisés), les plis vocaux ne vibrent pas.
Cette fréquence dépend à la fois de critères morphologiques (non contrôlés) comme la masse ou l'épaisseur des plis vocaux et de critères contrôlés par le locuteur comme le débit d'air issus des poumons (Marchal, 2007). **La F0 est un paramètre idiosyncratique énoncé à de nombreuses reprises dans la littérature (Nolan, 2001) ou (Kunzel, 1994)**. D'ailleurs, (Spence et Freeman, 1996) montre que des nouveaux nés ne reconnaissent pas leur mère lorsque celle-ci chuchote. Avec de la voix chuchotée, nous n'avons accès qu'aux éléments d'articulation et au rythme mais pas au voisement. La F0 joue donc un rôle important pour le nouveau né pour reconnaître sa mère par la parole. **L'intonation de la voix joue un rôle primordial dans la reconnaissance des personnes**. Par exemple, (Mehler et al., 1976) montre qu'en fonction de l'intonation l'enfant ne réagit pas de la même manière à la voix de sa mère. De même, (Van Dommelen, 1987) montre « l'influence du rythme et du F0 en matière de reconnaissance du locuteur » en faisant écouter des signaux laryngés à des auditeurs. En modifiant la durée des portions voisées et non voisées, le contour de F0 et de la hauteur de F0, il montre que les auditeurs confondent les locuteurs et que ces paramètres sont important dans la reconnaissance des locuteurs. Toutefois, (Van Lancker et al., 1985) montre que la tâche d'identification des locuteurs connus est moins bien réussie si les signaux sont écoutés non pas dans le sens classique d'écoute mais à l'envers, ce qui fait que les locuteurs ne peuvent pas comprendre ce qui est dit ($T_{identification} = 69.9$ en écoute classique avec 2 secondes de signaux vs $T_{identification} = 57.5\%$ en écoute à l'envers avec 4 secondes de signaux). L'information sur le locuteur ne réside pas uniquement dans la valeur moyenne de F0, sa variation est porteuse d'information.
Toutefois, comme le souligne (Cappé, 1995), la F0 moyenne se situe, par exemple pour 35% des locuteurs masculins allemands, entre 110Hz et 120Hz. Par conséquent, elle

ne « fournit [...] des résultats exploitables [que] lorsqu'un des locuteurs [...] s'éloigne fortement des caractéristiques moyennes ». De plus, il semble important de souligner que la F0 varie également en fonction d'autres facteurs comme l'état émotionnel du locuteur (Hollien, 1990).
Nous tenions à souligner qu'à notre connaissance aucune étude n'a été faite sur l'influence des tons, qui sont des éléments phonologiques qui s'appuient sur les courbes de F0, sur la reconnaissance des locuteurs. Nous ne savons pas si ces résultats sont reproductibles lorsque la courbe de F0 sert d'élément phonologique comme c'est le cas dans les langues à tons.

2.3.2 Jitters et shimmers

La qualité de voix est mise en avant comme influençant la reconnaissance du locuteur (Erikson, 2007) ou (Zetterholm, 2006). Elle consiste à décrire le mode de phonation en terme de voix plus ou moins tendue, laryngalisées, soufflées, murmurées ou chuchotées (Laver, 1980). Ces description sont avant tout auditives ou articulatoires.
Parmi les indices acoustiques proposés pour mesurer la qualité de voix se trouvent le jitter et le shimmer. Le jitter correspond aux modulations de durée des cycles glottaux tandis que le shimmer correspond aux modulations d'amplitude des cycles. Par exemple, (Bohm et Shattuck-Hufnagel, 2007) montre que les jitters mesurés sur des /*a*/ sont différents en fin de phrases en fonction de la locutrice.
D'autres paramètres acoustiques ont également été proposés pour décrire la qualité de voix. Ainsi, (Campbell et Mokhtari, 2003) mesure le degrés de *breathy voice* via le *Normalized Amplitude Quotient* (NAQ) et montre des différences importantes en fonction des interlocuteurs chez une locutrice suivie pendant trois ans. Ce paramètre, complexe à extraire n'a pas été utilisé, à notre connaissance comme un indice différenciant les locuteurs.

2.3.3 Mesures formantiques

Les transitions formantiques ont été repérées comme idiosyncratiques par (McDougall, 2006). Son expérience est une analyse acoustique où elle montre que les transitions formantiques sont spécifiques au locuteur. Au delà de la valeur moyenne des formants, le projet de (Nolan et al., 2006) propose d'utiliser la dynamique des formants

qui rendrait compte de l'articulation. Chaque locuteur a sa propre « stratégie articulatoire »pour réaliser la suite de phonèmes. Ainsi, (de Jong et al., 2007) montrent que les trajectoires formantiques pour F1, F2, F3 sont significativement différentes en fonction du locuteur.
Par ailleurs, **les valeurs de F3 et F4 sont présentées par (Lavner et al., 2001) comme dépendantes du locuteur**. (Nolan, 2001) suggère notamment d'étudier la dispersion de ces valeurs formantiques. (Gendrot et Adda-Decker, 2007) mettent bien en avant que F3 joue un rôle important en français pour distinguer /i/ de /y/. L'utilisation des formants comme critère idiosynchratique est donc sûrement dépendant des langues.

2.3.4 L'information sur le locuteur inégalement répartie dans le signal de parole

Tous les phonèmes ne sont pas aussi discriminants pour reconnaître les locuteurs. Ainsi, (Rose, 2011) met en évidence à l'aide d'une analyse acoustique, l'utilité des **fricatives** dans la reconnaissance des locuteurs déjà envisagée par (Nolan, 2001). Des différences importantes de valeurs de centre de gravité [3] en fonction des locuteurs sont d'ailleurs également soulignées par (Schmid, 2011) pour les affriquées des parlers romanches.
(Amino et al., 2006) met en évidence que des auditeurs japonais reconnaissent plus facilement des locuteurs qu'ils connaissent après l'écoute de phrases contenant des **nasales** que des phrases contenant des orales. En revanche, comme le lien entre articulatoire et acoustique des nasales est un débat qui reste en cours, aucune mesure précise n'est suggérée comme indice sur le locuteur. Certains phonèmes semblent donc être plus porteurs que d'autres d'indices sur le locuteur.

2.3.5 Les autres niveaux du langage

D'autres niveaux ont également été identifiés. (Butcher, 2002) souligne l'intérêt de décrire le style employé par le locuteur d'après la classification de (Mitchell et Delbridge, 1965). Ce style peut être en partie décrit à l'aide du **lexique**. Les **accents** semblent pertinents pour (Schwartz et al., 2009) comme moyen d'identifier les locuteurs. Ils

3. Le centre de gravité est la valeur fréquentielle pour laquelle, dans le spectre, autant d'énergie se situe au-dessous et au-dessus de cette fréquence.

proposent une méthodes de reconnaissance des locuteurs où ils comparent notamment la similarités des systèmes phonologiques et de la fréquence fondamentale.

Synthèse du chapitre

Les auditeurs sont capables de reconnaître des personnes à partir d'enregistrements de parole. **Cette capacité dépend de la connaissance qu'a l'auditeur du locuteur, de la langue dans laquelle parle le locuteur et de la durée des enregistrements**.
La **diversité des protocoles d'évaluation** utilisés pour estimer les capacités humaines à reconnaître un interlocuteur a pour conséquence que certaines conclusions peuvent sembler contradictoires. Cette diversité rend difficiles les comparaisons des résultats obtenus. **Une évaluation commune avec une explication du protocole comme cela est fait en informatique apporterait sans doute une amélioration de nos connaissances**.
Les modèles cognitifs proposés pour expliquer comment l'auditeur mémorise les locuteurs, notamment le modèle de prototype, rejoignent des approches proposées en RAL.
L'étude de la phonoagnosie nous conduit à conclure que **le processus de reconnaissance des personnes à partir d'un enregistrement de parole n'est pas le même que le processus de reconnaissance du message transmis**.
Les paramètres qui permettent de reconnaître un locuteur à partir d'un signal de parole se situent à différents niveaux du langage : si la phonation, la prosodie ou l'articulation sont étudiées plus particulièrement par la phonétique, la phonologie ou la lexicologie ont aussi un rôle à jouer dans l'identification du locuteur. **Un indice à lui seul ne permet pas de reconnaître une personne**. C'est la combinaisons des indices qui semble la solution la plus pertinente. Dans ce chapitre, nous avons également montré que **l'information idiosynchratique n'est pas uniquement contenue dans les sons voisés** où la voix, c'est-à-dire le résultat de la phonation, peut être entendue. En ce sens, **notre travail ne consiste pas uniquement à faire une reconnaissance du locuteur à l'aide de sa voix mais à l'aide de l'ensemble de la parole enregistrée**.

Parole, voix et locuteur

Les mesures de performance adoptées en vérification du locuteur sont moyennées sur un grand nombre de comparaisons cible et imposteur. Ces mesures peuvent permettre de rendre compte de la variabilité des performances en fonction de données contrôlées qui se différencient par un facteur dont l'impact peut être étudié. La notion de locuteur est parfois réduite dans ce type d'évaluation, par une confusion courante entre locuteur et enregistrement. En effet, dans NIST-SRE, toutes les comparaisons sont considérées comme équivalentes, la mention du locuteur n'étant pas utilisée pour évaluer la performance. Or, la parole est non seulement incarnée mais également soumise à d'autres contraintes qui interfèrent sur les indices idiosyncratiques. Nous voulons reconnaître le locuteur à partir de sa voix mais la parole s'en mèle.
Nous faisons le choix, dans ce manuscrit, de limiter l'utilisation du terme de voix. En effet, dans leurs ouvrages, (Garde, 1954) et (Cornut, 2009), tous les deux intitulés *la voix*, les auteurs décrivent principalement le phénomène de phonation, c'est-à-dire la création d'une onde périodique complexe (bourdonnement) par la vibration des plis vocaux. Nous emploierons donc, dans ce travail, la terminologie suivante.

- **Parole** : onde dynamique produite par le conduit vocal et qui répond à des contraintes d'ordre biomécanique, linguistique, sociologique, psychologique et performatif. La parole est motivée, produite pour être entendue.
- **Enregistrement ou Extrait de parole** : support par lequel nous pouvons avoir accès à la parole d'un locuteur. Un locuteur peut, heureusement, produire plusieurs enregistrements ou extraits de parole.
- **Locuteur** : personne qui parle, il s'agit de l'objet d'étude de cette thèse.

Dans cette thèse, nous cherchons, tout d'abord, à **proposer des méthodologies permettant de rendre compte de la variabilité des performances en vérification du locuteur.** Nous établissons, dans un premier temps, un protocole pour évaluer la capacité humaine à discriminer un locuteur à partir d'un extrait de parole. Cette évaluation

est accompagnée d'une estimation de la confiance en la réponse des auditeurs. Nous souhaitons, dans un second temps, quantifier l'influence du choix d'un extrait de parole sur la performance des systèmes automatiques. Ce plan d'évaluation doit donc quantifier l'impact d'un changement d'extrait de parole pour représenter le locuteur, en tenant compte du fait que tous les locuteurs ne sont peut-être pas aussi aisément reconnus.

Il s'agit, par la suite, **d'établir quels sont les indices les plus pertinents pour distinguer les locuteurs** de nos corpus. Pour répondre à cette question, nous nous appuyons sur les indices qui ont déjà été identifiés par les linguistes, les informaticiens et les psychologues comme porteurs d'informations idiosyncratiques. Si les techniques du RAL utilisent dans leur grande majorité des paramètres rendant compte de l'ensemble du signal de parole, les indices mis en évidence par les tests perceptifs ou les études acoustiques caractérisent aussi bien la phonation que l'articulation de la parole. Cette étude sera l'occasion d'établir si ces indices nous permettent de **différencier les extraits de parole** pertinents pour les systèmes de RAL de ceux qui obtiennent de moins bons résultats.

Deuxième partie

Quantifier la possibilité de reconnaître un locuteur

Chapitre 3

Évaluation perceptive dans le cadre du *Human Assisted Speaker Recognition* de NIST

Résumé : *En 2010, pour la première fois, le NIST proposait une évaluation où les humains pouvaient intervenir dans le processus de décision. Nous avons profité de ce protocole commun pour évaluer la capacité humaine à discriminer des locuteurs et estimer la confiance que nous pouvons avoir dans la réponse des auditeurs. Nous montrons que la tâche ainsi posée est difficile pour nos auditeurs, qu'ils soient naïfs ou plus expérimentés. Ce résultat est également valable pour les autres participants à HASR, un seul ayant réussi à faire mieux que le hasard sur cette tâche. Nous montrons par ailleurs que la quasi unanimité des auditeurs dans une réponse ou l'auto-évaluation ne sont pas des gages de confiance dans la réponse soumise. Enfin, nous mettons en évidence que toutes les paires de signaux n'ont pas toutes la même difficulté.*

Sommaire

3.1 Protocole HASR défini par NIST

En 2010, pour la première fois, **le NIST a proposé une tâche, nommée *Human Assisted Speaker Recognition* (HASR), dont le but est d'« évaluer la complémentarité qu'il peut exister entre expertise humaine et traitement automatique »** (Greenberg et al., 2011b). L'introduction d'une évaluation humaine dans cette tâche de vérification est un moyen d'établir une référence pour mesurer la capacité des humains à discriminer des personnes à partir d'extraits de parole. En participant à cette campagne, nous avons accès à un nombre important de données qui correspondent à celles utilisées lors des campagnes d'évaluation des systèmes automatiques. Le contexte de l'évaluation nous autorise une comparaison avec d'autres sites qui participent également à cette évaluation.

Dans le cadre de la tâche HASR, les participants peuvent **combiner des solutions impliquant des humains et des systèmes de vérification du locuteur**. Ils sont donc autorisés à écouter les signaux et/ou à leur appliquer différents traitements pour améliorer l'intelligibilité du signal, contrairement aux autres tâches pour lesquelles l'écoute des signaux est interdite.

Pour chaque comparaison, les participants doivent fournir une réponse composée d'**une décision** (Oui/Non [1]) et d'**un score de confiance dans cette décision**. Il est important de souligner que lors de cette évaluation, les participants n'ont pas accès directement à l'ensemble de la cohorte de comparaisons. Ils doivent **fournir une réponse à la première comparaison pour avoir accès à la seconde** et ainsi de suite jusqu'à la dernière paire.

L'ensemble des paires est issu de l'évaluation NIST-SRE 2010 Mixer 6 (**?**), elles sont donc en **langue anglaise** et chaque enregistrement a une durée d'environ **2 minutes 30 sec-**

1. Oui : les deux enregistrements ont été prononcés par le même locuteur, Non : il s'agit de deux personnes différentes

ondes. Il est à noter que les fichiers qui composent une paire ont toujours été enregistrés dans des **conditions différentes l'un par rapport à l'autre**. L'un est toujours issu des données interview tandis que l'autre provient de données téléphoniques. Les microphones et les téléphones utilisés varient également d'une paire à l'autre. Des paires avec différents efforts vocaux (cf. 1.2.3) ont également été incluses dans cette tâche.
Les paires d'enregistrements ont été sélectionnées par NIST en fonction de leur **difficulté évaluée par le système de vérification automatique du locuteur de ICSI-Berkeley** (Morgan, 2010). Les comparaisons cible sélectionnées sont celles qui obtiennent les scores les plus bas tandis que les comparaisons imposteur choisies correspondent à celles qui donnent lieu aux scores les plus élevés.
Nous avons décidé d'évaluer la capacité humaine à distinguer les paires d'enregistrements proposées dans le cadre de l'évaluation HASR2 soit sur les **150 comparaisons** de parole afin d'avoir accès à un nombre important de stimuli. En participant à cette évaluation, nous souhaitons **connaître la confiance qui peut être attribuée à la réponse (oui/non) fournie**. Pour répondre à cette question, **deux pistes sont explorées**. La première consiste à **quantifier la confiance que chaque auditeur attribue à sa réponse** sur une échelle de 0 (pas confiant dans sa réponse) à 5 (très confiant dans sa réponse). La seconde approche consiste à faire appel à un panel d'auditeurs (avec un nombre impair). Dans ce cas, **le taux de confiance correspond à l'accord inter-juges**, c'est-à-dire au nombre d'auditeurs qui ont donné la même réponse.
Par ailleurs, nous ne nous attendons pas à ce que notre panel d'auditeurs soit spécifiquement performant pour discriminer les personnes car **la tâche proposée est une tâche difficile**. D'une part, les auditeurs auxquels nous avons fait appel **ne sont pas natifs de l'Anglais**. D'autre part, **ils ne connaissent pas les locuteurs qu'ils vont devoir discriminer** et la durée d'enregistrement accessible (2 minutes 30 secondes) ne permet pas de faire un apprentissage préalable des locuteurs par les auditeurs [2]. Enfin, **les conditions d'enregistrement des deux extraits à comparer sont systématiquement différentes**.

2. A titre d'exemple, (Legge et al., 1984) utilise, pour « apprendre une voix non familière », des enregistrements dont la durée varie entre 10 minutes et plus de 2h20. (Papcun et al., 1989) utilisent, quant à eux, des enregistrements de plus de 5 minutes pour la phase d'apprentissage des nouveaux locuteurs.

3.2 Évaluation de la performance

La tâche se compose de deux cohortes nommées HASR1 et HASR2 décrites par le tableau 3.1. HASR1 se compose de 15 paires de comparaison (6 comparaisons cible contre 9 comparaisons imposteur) tandis que HASR2 comporte 150 paires à comparer (51 comparaisons cible et 99 comparaisons imposteur). Il est à noter que les comparaisons de HASR1 sont incluses dans HASR2. Ainsi, si les participants choisissent de participer à HASR2, ils participent automatiquement à HASR1. Un déséquilibre entre femmes et hommes est observé. 114 et 9 paires sont des femmes et 36 et 6 paires sont des hommes respectivement pour HASR2 et HASR1.

Protocole	Type de Comparaison	Hommes	Femmes	Total
HASR1	Cible	5	1	6
	Imposteur	1	8	9
HASR2	Cible	20	31	51
	Imposteur	16	83	99

TABLE 3.1 – *Répartition des comparaisons cible et imposteur pour HASR*

Contrairement au protocole proposé par HASR, outre **FA et FR**, l'évaluation des performances est également observée de **manière globale à l'aide d'un taux de réussite**. Il s'agit d'exprimer le pourcentage de réponses correctes indépendamment du type de comparaison cible et imposteur. **Les taux de réussite sont comparés aux taux de réponses correctes qui pourraient être fournies si les réponses étaient fournies au hasard**, et ce à l'aide d'un test binomial (Wonnacott et Wonnacott, 1991). Les différents degrés de significativité retenus pour caractériser la différence entre les taux obtenus par le panel avec ceux obtenus au hasard sont les suivants.

- Si $p \geqslant 0.05$, la différence n'est pas considérée comme significative.
- Si $p < 0.05$, nous considérons la différence avec le hasard comme légèrement significative.
- Si $p < 0.01$, nous considérons la différence avec le hasard comme significative.
- Si $p < 0.001$, nous considérons la différence avec le hasard comme très significative.

Il est à noter que dans un test binomial, nous jugeons la significativité de la différence et non la nature de la différence (« meilleur » ou « moins performant » que le hasard). C'est la valeur du taux qui indique si la performance est meilleure ou non que le hasard. Enfin, nous souhaitons **vérifier l'hypothèse selon laquelle plus une décision est unanime, plus elle a des chances d'être correcte**. Pour cela, nous comptabilisons pour chaque stimuli l'accord entre les auditeurs et regardons si un accord entraîne plus de réponses correctes. **L'influence du score de confiance donné par chaque auditeur est également étudiée**.

3.3 Première étude perceptive lors de l'évaluation HASR

3.3.1 Méthodologie adoptée

Panel d'auditeurs

L'obligation, dans le cadre de la participation à HASR, de fournir une réponse à la comparaison précédente pour avoir accès à une autre comparaison, implique que tous les auditeurs de notre panel doivent répondre en même temps. En effet, l'avis de l'ensemble des participants nous est nécessaire pour soumettre la décision finale. Ceci nécessite une grande coordination entre les différents auditeurs et c'est pour cette raison que notre premier **panel est assez restreint**.
Trois auditeurs (un homme âgé de 31 ans et deux femmes âgées de 25 et 36 ans) **natifs du français** ont participé à cette première expérience. **Tous ces auditeurs ont l'habitude d'analyser du signal de parole et n'ont pas de problème d'audition connu**. Ils pratiquent l'anglais dans le cadre de leurs activités professionnelles mais n'ont jamais séjourné plus d'un an dans un pays anglophone.
Avec un panel aussi petit, il ne nous est pas possible d'étudier l'influence du stimulus. Cette question sera abordée dans une étude post-campagne, une fois le panel d'auditeurs élargi.

Manipulation de stimuli

Chaque paire de comparaison fournie par NIST se compose de deux enregistrements d'une durée d'environ **2 minutes 30 secondes** chacun. Cette durée ne permet pas d'-

effectuer un apprentissage préalable des locuteurs par les auditeurs [3] et dans le même temps elle est beaucoup **plus longue que celle habituellement utilisée en test perceptif** [4]. Nous avons donc décidé de ne pas présenter aux auditeurs les deux enregistrements tels quels, ce qui les auraient conduit à écouter pendant environ 2 minutes 30 secondes un premier locuteur puis d'entendre le second locuteur pour enfin prendre une décision.

Notre choix a été le suivant : les auditeurs entendent des extraits plus courts de chaque enregistrement en **alternant les enregistrements toutes les six secondes** afin de confronter les auditeurs à chacun des deux enregistrements.

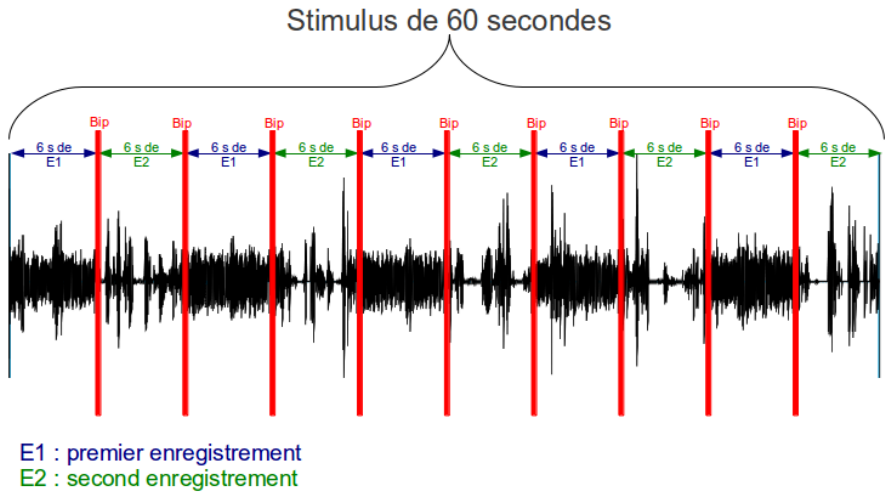


FIGURE 3.1 – *Un stimulus est la concaténation d'extraits de 6 secondes de chacun des deux enregistrements séparés par un bip.*

Pour chaque enregistrement de 2 minutes 30 secondes, nous sélectionnons des portions de 6 secondes qui contiennent **le plus d'énergie** à l'aide de l'outil EnergyDetector d'ALIZE/SpkDet (Matrouf et al., 2008b) afin de conserver des zones de parole. Nous concaténons alternativement un extrait de 6 secondes du premier enregistrement et un extrait de 6 secondes du second enregistrement en séparant chaque extrait d'un bip jusqu'à obtenir **un stimuli d'une durée d'environ 1 minute**. Ainsi, si les auditeurs écoutent les enregistrements jusqu'au bout, ils ont accès à 30 secondes de parole issues

3. cf note de base de page numéro 2 de ce chapitre

4. (Lavner et al., 2000) effectue l'identification acoustique des auditeurs sur les voyelles isolées. (Schlichting et Sullivan, 1998) présente des phrases courtes d'une durée de 6 secondes environ. (Saslove et Yarmey, 1980) utilisent des stimuli de 11 secondes. (Legge et al., 1984) utilisent des stimuli entre 6 et 60 secondes. Les stimuli présentés par (Papcun et al., 1989) ont une durée moyenne de 1.58 minutes

du premier enregistrement et 30 secondes du second enregistrement. Le **bip permet aux auditeurs de savoir qu'ils ont changé d'enregistrement**. La figure 3.2 présente un stimulus ainsi construit.
Les auditeurs peuvent prendre leur décision dès qu'ils ont entendu un extrait de chaque enregistrement soit au bout de 12 secondes, il peuvent attendre la fin du stimulus et sont autorisés à réécouter le stimuli autant de fois qu'ils le souhaitent. Nous autorisons également les auditeurs à filtrer le signal audio afin de diminuer l'effet de certaines fréquences, certaines conditions d'enregistrement étant extrêmement bruitées comme illustré par la figure 3.2. L'écoute est faite à l'aide du logiciel Praat (Boersma et Weenink, 2009) et les filtrages ont également été réalisés avec les filtres de ce logiciel.

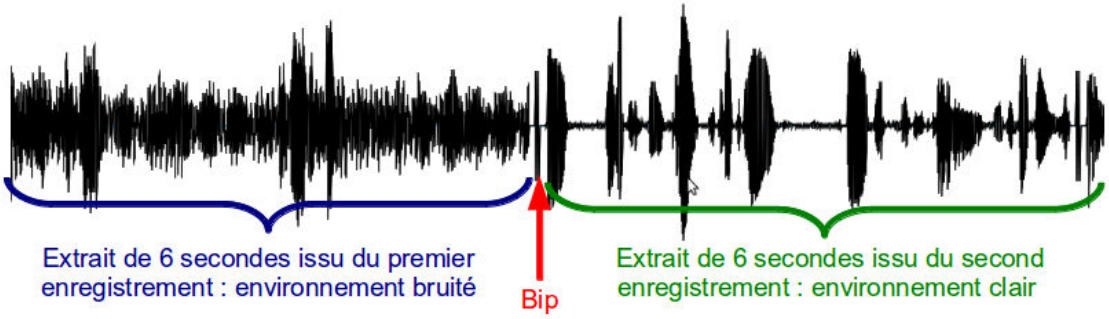


FIGURE 3.2 – *Différence de qualité d'enregistrement entre 2 extraits de parole pour une même locutrice.*

Prise de décision et score de confiance

Comme imposé par le protocole HASR (3.1), tous les auditeurs ont écouté les stimuli dans le même ordre. Chaque locuteur, indépendamment des autres, indique si les deux enregistrements qui lui sont présentés ont été produits par le même locuteur. Sa réponse est de type binaire : oui ou non. Il indique, par la suite, la confiance qu'il a dans sa réponse à l'aide d'une échelle de 0 à 5, 0 signifiant qu'il n'a pas du tout confiance en sa réponse et 5 qu'il est très confiant. Ce score de confiance n'a pas été soumis à NIST mais il permet de rendre compte de l'impression des auditeurs quant à leurs réponses. **Il s'agit par la suite d'établir la pertinence de ce score de confiance comme prédiction de réponses correctes**.
La réponse soumise aux organisateurs de HASR correspond au vote majoritaire des trois auditeurs. Ainsi, si au moins deux auditeurs répondent qu'il ne s'agit pas du

même locuteur, la réponse soumise est « non ». Au contraire, si au moins deux auditeurs répondent qu'il s'agit du même locuteur alors la réponse soumise est « oui ».
Afin d'effectuer, par la suite, des fusions de scores avec un système automatique, nous avons décidé d'utiliser les distributions de scores obtenus par le système ALIZE/SpkDet (Matrouf et al., 2008b) lors de NIST-SRE 2008 pour définir les scores de confiance soumis à HASR. Ce score n'est qu'une correspondance entre le nombre de réponses positives et les distributions de scores du système. Le calcul du score de confiance demandé par les organisateurs est défini à partir du nombre de locuteurs ayant pris la même décision comme illustré par la figure 3.3.
Quatre scores de confiance sont attribués en fonction de l'accord entre les auditeurs. Les écarts entre les scores ont été choisis afin de permettre, lors de la fusion avec un système automatique, d'avoir des écarts importants si les auditeurs sont unanimes.

- Si les trois auditeurs ont tous décidé que les deux enregistrements ont été produits par le même locuteur, le score soumis à HASR correspond à la moyenne des scores issus de comparaisons imposteur de NIST-SRE 2008 moins deux fois l'écart type de cette distribution.
- Si deux auditeurs affirment qu'il s'agit de deux locuteurs différents, le score soumis correspond à la moyenne des scores issus de comparaisons imposteur de NIST-SRE 2008.
- Si seulement deux auditeurs sur trois décident qu'il s'agit du même locuteur, le score de confiance soumis correspond à la moyenne des scores issus des comparaisons cible de NIST-SRE 2008.
- Si les trois auditeurs sont d'accord pour dire qu'il s'agit du même locuteur, alors le score soumis correspond à la moyenne des scores issus de comparaisons cible de NIST-SRE 2008 plus deux fois l'écart type de cette distribution.

Système automatique pour la comparaison

Les résultats obtenus par les trois auditeurs sont comparés avec ceux obtenus par **le système basé sur la technique du SVM (Chang et Lin, 2011) issu des outils libres ALIZE/SpkDet** (Matrouf et al., 2008b) en appliquant les techniques du Factor Analysis (Matrouf et al., 2007). Le seuil choisi pour évaluer le système est celui utilisé pour NIST-SRE 2010 (Larcher et al., 2010).

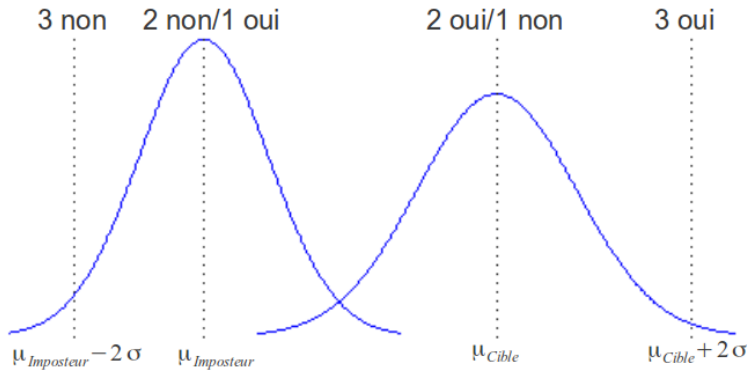


FIGURE 3.3 – *Choix des valeurs des scores de confiance soumis à HASR en fonction des distributions de scores de ALIZE/SpkDet sur les cohortes de NIST-SRE 2008.*

3.3.2 Performance et confiance dans le panel d'auditeurs

Performances

Le taux de réponses correctes des trois auditeurs est de 59%. Ce taux est significativement légèrement différent du hasard (test binomial : $p < 0.05$). La moyenne des scores de confiance indiqués par les auditeurs sur l'ensemble des stimuli est de 2.2 sur une échelle de 0 (pas du tout confiant dans la réponse) à 5 (très confiant dans la réponse). Ce score de confiance indique que **les auditeurs ont une confiance assez moyenne en leur réponse**. Il retranscrit bien la performance globale des auditeurs qui est légèrement meilleure que le hasard mais qui n'est pas très élevée.
Les auditeurs ont répondu correctement à 65% des comparaisons cible (soit un taux de FR de 35%) et à 56% des comparaisons imposteur (soit un taux de FA de 44%). Si le taux de FR est significativement différent du hasard ($p < 0.05$), le taux de FA est équivalent à celui obtenu par le hasard ($p = 0.3149$). Le score de confiance moyen est de 2.0 pour les comparaisons cible et de 2.3 pour les comparaisons imposteur. Il est surprenant que le score de confiance des auditeurs soit plus élevé lorsque leurs réponses ne sont pas différentes du hasard (cas des comparaisons imposteur) que lorsque leur taux de réponses correctes atteint 65% (cas des comparaisons cible). Est-il sûr de prendre en compte le ressenti des auditeurs sur leur propre décision ?

Confiance dans le score de confiance

Si les auditeurs ont confiance en leur jugement (moyenne du score de confiance supérieure à 2.5), **leur taux de réussite est-il plus élevé que lorsqu'ils sont peu confiants** (moyenne du score de confiance inférieure à 2.5) dans leur réponse ?
Pour cette analyse, nous avons séparé les comparaisons en fonction de leur score de confiance : d'un côté nous retrouvons toutes les comparaisons où la moyenne des scores de confiance est inférieure à 2.5 (cohorte *faible confiance*) et de l'autre toutes les comparaisons où le moyenne des scores de confiance est supérieure à 2.5 (cohorte *confiance plus haute*).
Dans ce cas, une claire asymétrie est observée dans la répartition des comparaisons cible et imposteur comme résumé par le tableau 3.2.

score de confiance	Nombre de comparaisons imposteur	Nombre de comparaisons cible	Nombre total de comparaison
[0 ;2.5]	55	37	92
]2.5 ;5]	44%	14	58
	99	51	

TABLE 3.2 – *Répartition des comparaisons pour les cohortes faible confiance et confiance plus haute.*

Si 61% des comparaisons se retrouvent dans la cohorte *faible confiance*, les proportions sont bien plus importantes pour les comparaisons cible que pour les comparaisons imposteur (72.5% vs 55.5%).
Les performances globales se montrent légèrement meilleures pour la cohorte *confiance plus haute* que pour la cohorte *faible confiance*. Le taux de réponses correctes est de 62% pour la cohorte *confiance plus haute* contre 57% pour la cohorte *faible confiance*. Le taux de FR de la cohorte *faible confiance* est de 32% tandis que celui de la cohorte *confiance plus haute* est de 43%. En comparaisons cible, le score de confiance des auditeurs en leur réponse n'est pas un gage de qualité de la réponse car ils se trompent plus souvent quand ils ont confiance dans leur réponse.

Le taux de FA de la cohorte *faible confiance* est de 51% tandis que celui de la cohorte *confiance plus haute* est de 36%. En comparaison imposteur, le ressenti des auditeurs semble

	Faible confiance	*Confiance plus haute*
Cible	FR=32%	FR=43%
Imposteur	FA=51%	FA=36%
Taux de réussite global	59%	59%

TABLE 3.3 – *Performance pour la cohorte où les auditeurs sont confiants dans leur réponse et celle où ils ne sont pas confiants dans leur réponse.*

être plus indicatif. Le tableau 3.3 retranscrit ces résultats.

Au vu de ces résultats, il semble qu'il existe un lien entre la confiance des auditeurs et le type de comparaisons auxquelles ils sont confrontés. Ils sont plus confiants lorsqu'ils affirment qu'il s'agit de deux locuteurs différents. Étant donné que le type de comparaison n'est pas connu à l'avance, le score de confiance des auditeurs en leur réponse n'est pas un bon prédiction de réponses correctes.

Confiance dans l'unanimité

Un autre indicateur de confiance repose sur **le nombre de réponses semblables attribuées par les auditeurs**. Plus les auditeurs sont d'accord, plus le score de confiance attribué à la réponse est grand.
Les trois auditeurs ont pris la même décision dans 51% des cas (coefficient Kappa=0.34 [5]). **Ce taux est supérieur à celui obtenu par le hasard** [6]. Les auditeurs sont donc unanimes plus souvent que si les trois réponses étaient tirées au hasard. **Pouvons-nous avoir plus confiance dans les réponses unanimes que dans les autres réponses ?**
Pour vérifier cette hypothèse nous pouvons comparer les taux de réussite (et les taux de FA et FR) de l'ensemble des comparaisons où les trois auditeurs sont unanimes à celui de l'ensemble des comparaisons où un des auditeurs n'a pas répondu comme les deux autres. La cohorte *unanimes* se compose de 23 comparaisons cible et de 53 compara-

5. Dans notre cas (3 annotateurs, 150 comparaisons, 213 réponses positives), si les auditeurs donnent systématiquement la même réponse le coefficient Kappa aurait été de 1. Dans le pire des cas (où ils n'auraient jamais répondu la même chose), le coefficient Kappa aurait été de -0.33. Si les auditeurs avaient répondu au hasard, le coefficient Kappa=0)

6. Sachant qu'il y a trois auditeurs, la probabilité que les trois auditeurs soient d'accord est la somme de la probabilité qu'ils répondent tous « oui » et de la probabilité qu'ils répondent tous « non ». Soit une probabilité de 0.35. D'après (Wonnacott et Wonnacott, 1991), ce taux peut varier de 0.07. Ainsi, le hasard est compris entre 28% et 42% d'accord.

isons imposteur tandis que la cohorte *désaccord* se compose de 28 comparaisons cible et 46 comparaisons imposteur. Les résultats sont résumés par le tableau 3.4.

	Accord	Désaccord
Cible (FR)	35%	36%
Imposteur (FA)	30%	61%
Taux de réussite global	49%	68%

TABLE 3.4 – *Performances en fonction de l'unanimité des auditeurs.*

Le taux de réussite global de la cohorte *unanimes* est de 49% ($p = 0.9076$) tandis que celui de la cohorte *désaccord* est de 68% ($p < 0.01$). Le taux de FR de la cohorte *unanimes* est de 35% ($p = 0.1849$) et celui de la cohorte *désaccord* est de 36% ($p = 0.2100$). Le taux de FA pour la cohorte *unanimes* est de 30% ($p = 0.1839$) tandis que celui de la cohorte *désaccord* est de 61% ($p < 0.01$).
Contrairement à notre hypothèse, les auditeurs ont une meilleure performance globale lorsqu'ils ne sont pas d'accord. Nous ne pouvons a priori pas utiliser l'unanimité des auditeurs pour prédire la performance globale des auditeurs. En comparaisons cible, il n'y a pas de différence de performance si les auditeurs sont unanimes ou non. En comparaison imposteur, l'unanimité des auditeurs ne rend pas les performances meilleures que le hasard ; en revanche, lorsqu'ils sont en désaccord, la performance est pire que le hasard.
L'unanimité n'est donc pas un bon indicateur pour prédire la confiance à accorder à la réponse donnée. Ici encore, nous pouvons nous interroger sur l'influence des différences de modes d'enregistrement sur l'unanimité.
Comme l'illustre la figure 3.4, nous observons en comparaison imposteur, un nombre de réponses unanimes correctes plus important que de réponses non-unanimes correctes (36 vs 17).

A l'inverse, en comparaison cible (figure 3.5), le nombre de réponses correctes non-unanimes sont plus importantes que le nombre de réponses correctes unanimes (18 vs 15).
Ainsi, **lorsqu'ils ont raison, les auditeurs sont unanimes en comparaison imposteur mais pas en comparaison cible**.

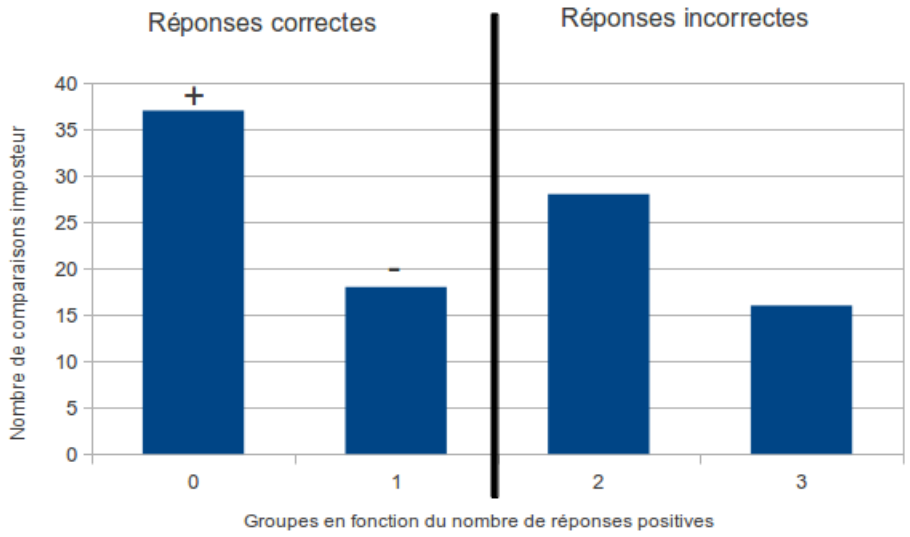


FIGURE 3.4 – *Nombre de réponses en fonction du nombre d'auditeurs en comparaison imposteur*

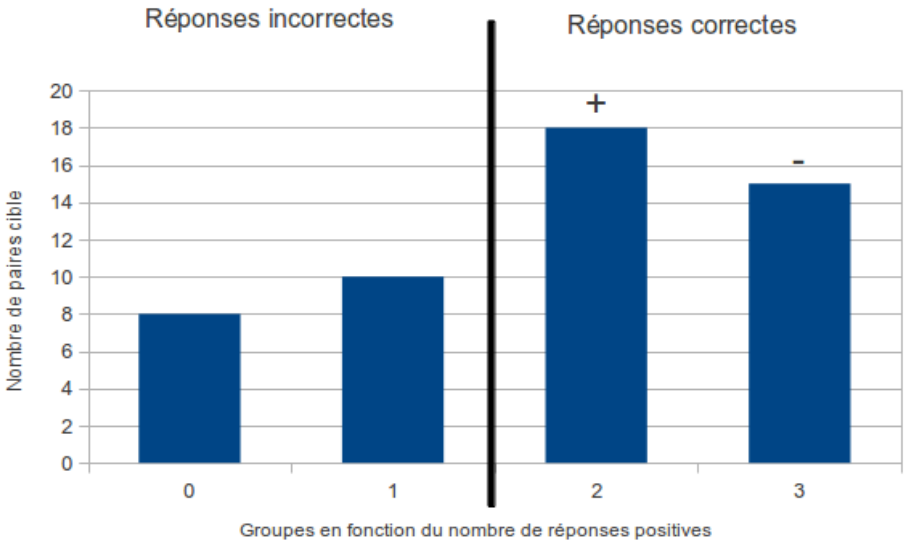


FIGURE 3.5 – *Nombre de réponses en fonction du nombre d'auditeurs en comparaison cible.*

Le fait que les modes d'enregistrement des personnes soient systématiquement différents pourrait expliquer ce phénomène. En effet, en comparaison imposteur, la distance entres les indices idiosynchratiques des deux locuteurs peut être renforcée par la différence de canal. Au contraire, en comparaison cible, cette dernière peut créer l'impression de personnes distinctes alors qu'il n'en est rien. Cette explication peut justifier que les auditeurs soient plus confiants en comparaison imposteur qu'en com-

paraison cible et que la majorité soit plus importante en comparaison imposteur qu'en comparaison cible.

Les performances de ce panel d'auditeurs très restreint ne sont pas très élevées. La tâche proposée est bien une tâche difficile. L'unanimité et les taux de confiance qui auraient pu être utilisés comme critère de confiance dans les réponses données semblent biaisés par le type de comparaisons à effectuer. Ce biais peut être expliqué par les différences de mode d'enregistrement systématiquement présentes dans les comparaisons. Il s'agit maintenant de savoir si les auditeurs ont tous des performances équivalentes ou si certains sont meilleurs que d'autres pour cette tâche de discrimination.

3.3.3 Performance par auditeur

Les trois auditeurs ont des performances quelque peu différentes comme illustré par le tableau 3.5.

	Nb réponses correctes	signif.	%FA	signif.	%FR	signif.
Auditeur 1	80	ns	48%	ns	43%	ns
Auditeur 2	100	***	28%	***	43%	ns
Auditeur 3	86	ns	46%	ns	35%	*
Vote Majoritaire	88	ns	44%	ns	35%	*

TABLE 3.5 – *Proportion d'erreurs des 3 auditeurs pour HASR2*

Seul un des trois auditeurs a un taux de réussite meilleur que le hasard (taux de réussite $=$ 67% et $p < 0.001$), les deux autres auditeurs ayant une performance équivalente à celle obtenue en répondant au hasard ($p = 0.08607$ et $p = 0.4625$). Néanmoins, si le taux de FA obtenu par le seul auditeur « performant » est meilleur que le hasard ($FA = 28\%$ et $p < 0.001$), son taux FR n'en diffère pas ($p = 0.4011$). Cet auditeur semble répondre souvent « non » et sa relative bonne performance globale est due, en partie, au déséquilibre entre comparaisons cible et imposteur.

Les performances des trois auditeurs ne sont finalement pas si éloignées les unes des autres : ils ont du mal à discriminer les locuteurs dans ces conditions. Les auditeurs

semblent avoir développé des stratégies de réponses différentes : un des auditeurs a tendance à répondre qu'il s'agit de locuteurs différents tandis que les deux autres se rapprochent d'un tirage aléatoire. La stratégie du premier auditeur est quelque peu masquée par le déséquilibre entre les comparaisons cible et imposteur. Notre panel d'auditeurs ne semble pas très bon pour discriminer les locuteurs mêmes si ses performances globales sont légèrement meilleures que le hasard. Il est intéressant de comparer ces résultats avec ceux des autres participants à HASR2.

3.3.4 Comparaisons avec les autres propositions à HASR

Système SVM d'ALIZE/SpkDet

La performance obtenue par test perceptif peut tout d'abord être comparée à celle du système automatique. Celui-ci obtient un taux de FA de 20% et un taux de FR de 35%. Cette performance est significativement meilleure que le hasard (Taux d'erreur $=$ 25% et $p < 0.001$) et ne dépend pas d'une stratégie de réponse ($p_{\text{Acceptations correctes}} < 0.001$, $p_{\text{Rejets corrects}} < 0.001$).
La cohorte d'évaluation semble être une cohorte particulièrement difficile aussi bien pour les humains que pour le système automatique. A titre d'exemple, la performance du système automatique pour l'ensemble des comparaisons de NIST-SRE 2010 était de 15.9% en FA et de 4.5% en FR pour le seuil choisi. Les performances obtenues sur la cohorte HASR sont donc bien en deçà de celles obtenues habituellement.

Autres participants

Sept autres sites ont participé à HASR2. L'ensemble de leurs résultats bruts est présenté en annexe B.1. Nous n'avons cependant accès qu'à leur taux de FA et de FR et non à leurs scores de confiance. Nous ne pourrons pas vérifier les phénomènes de confiance que nous avons mis en évidence pour notre panel d'auditeurs.
Un seul site a obtenu une performance réellement différente du hasard ($p_{\text{Réponses correctes}} < 0.001$) et ne dépendant pas d'une stratégie de réponse ($FR = 23.5\%$ et $p_{\text{Acceptations correctes}} < 0.001$, $FA = 16.2\%$ et $p_{\text{Rejets corrects}} < 0.001$).
Trois des participants ont un taux de FA qui est significativement pire que le hasard ($58\% < FA < 76\%$ et $p < 0.001$) et un taux de FR qui est significativement meilleur que le hasard ($4\% < FR < 26\%$ et $0.05 < p < 0.001$). Ces participants ont donc tendance à

répondre qu'il s'agit du même locuteur.
Au contraire, un des sites a un taux de FA significativement meilleur que le hasard ($FA = 3\%$ et $p < 0.001$) et un taux de FR significativement pire que le hasard ($FR = 71\%$ et $p < 0.001$). Ce site a donc tendance à répondre qu'il s'agit de personnes différentes lorsqu'il est confronté à une paire de signaux.

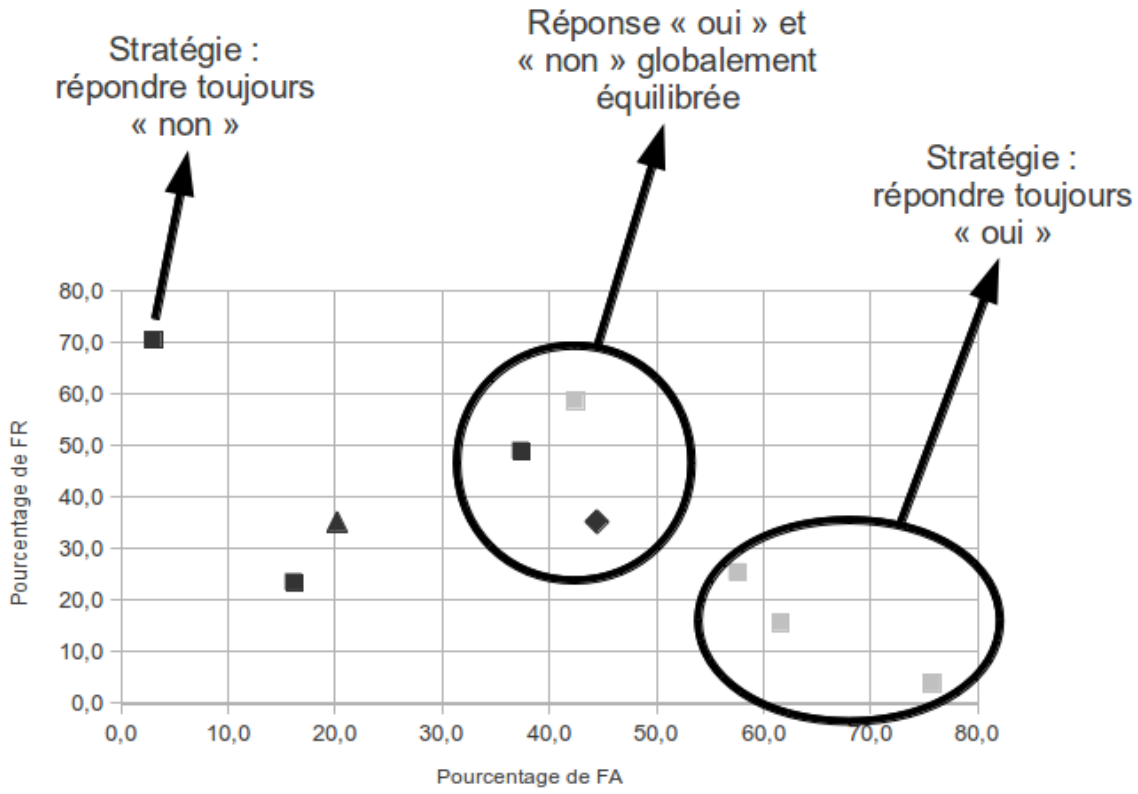


FIGURE 3.6 – *Résultats obtenus pour HASR2, triangle : SVM ($FA = 20\%$, $FR = 35\%$), losange : Auditeurs ($FA = 44\%$, $FR = 35\%$), carrés : autres participants. Éléments foncés : taux de réussite différent du hasard, Éléments clairs : taux de réussite équivalent à celui obtenu par le hasard.*

Il est à noter que si chaque site du groupe qui a tendance à répondre « oui » a une performance globale qui n'est pas différente du hasard ($0.3692 < p < 0.8066$), le site qui a tendance à répondre « non » a, au contraire, une performance globale très significativement différente du hasard ($p < 0.001$). Ceci est du au fait que le nombre de comparaisons cible et imposteur n'est pas équilibré. Ainsi, la stratégie de répondre souvent « non » s'avère logiquement plus payante que celle de répondre souvent « oui ».

La cohorte HASR2 est une cohorte difficile : seul un participant est réellement meilleur que le hasard. Les autres participants ont des performances qui sont proches du hasard ou ont développé des stratégies de réponses qui les conduisent à un déséquilibre entre leur taux de FA et de FR. Ces stratégies sont quelque peu masquées par le déséquilibre du nombre de comparaisons cible et imposteur. Ceci nous invite à équilibrer la cohorte de comparaisons.

3.3.5 Limites du protocole HASR et de notre soumission

Les corpus issus des protocoles HASR1 et HASR2 sont des corpus comportant de **nombreuses sources de variation** (parole conversationnelle, mode d'enregistrement, parole native/parole non native...). Ceci fait partie de la difficulté et de l'intérêt de la tâche proposée. Certaines limites auxquelles nous pouvons remédier facilement sont cependant observées dans le protocole que nous venons de présenter.
Un **déséquilibre est observé entre les comparaisons cible et les comparaisons imposteur** (HASR1 : 6 vs 9 ; HASR2 : 51 vs 99). Comme nous l'avons montré précédemment, ce déséquilibre masque parfois quelque peu les stratégies de réponse des participants. Par ailleurs, il est à noter que dans la majorité des tests perceptifs où la réponse est binaire, les expérimentateurs préfèrent équilibrer le type de comparaisons.
De plus, le protocole tel qu'il est proposé **empêche de présenter aléatoirement les stimuli aux auditeurs**. Ceci biaise sérieusement les tests perceptifs puisque les auditeurs entendent dans le même ordre les stimuli. Il n'est alors pas possible d'annihiler l'effet de l'ordre de présentation des stimuli.
Par ailleurs, **notre première étude n'a été réalisée qu'avec trois auditeurs** qui ont l'habitude d'annoter des signaux de parole. Nous pouvons également nous interroger sur l'influence de l'habitude d'annotation sur la performance à reconnaître des locuteurs.

Une fois la campagne HASR terminée, il est possible de remédier au déséquilibre de la répartition des comparaisons et à l'ordre de présentation des stimuli. Sans les contraintes de temps, il nous est également possible d'élargir le panel d'auditeurs afin d'avoir une analyse plus fine des résultats avec un test perceptif mieux contrôlé. Cette démarche nous permet notamment d'étudier les performances par stimuli.

3.4 Extension du protocole HASR

Une fois la campagne HASR terminée, il nous a semblé important de continuer notre étude en élargissant notre panel d'auditeurs de manière à répondre aux questions suivantes.

- Existe-t-il des différences de performances entre auditeurs naïfs et auditeurs plus expérimentés ?
- Les paires de stimuli sont-elles toutes équivalentes pour effectuer une discrimination ?

3.4.1 Quelques changements méthodologiques

Panel d'auditeurs

Le panel de locuteurs est fortement élargi. Deux panels différents d'auditeurs ont été étudiés afin d'étudier l'influence des signaux de parole sur les performances globales dans cette tâche de discrimination de locuteurs à partir de leur parole. Le **premier** se compose de 29 auditeurs considérés comme **non-expérimentés**, dans le sens où ils n'ont pas l'habitude d'écouter et de manipuler des signaux de parole. Les 20 femmes et les 9 hommes qui composent ce groupe ont étudié en moyenne 9.3 ans l'anglais (leur moyenne d'age est de 29.3 ans). Aucun d'entre eux n'a vécu plus d'un an dans un pays anglophone.
Le **second groupe** se compose de 18 auditeurs considérés comme **expérimentés** dans le sens où, en tant que phonéticiens, ils ont l'habitude d'écouter et d'annoter des signaux de parole. Aucun d'eux ne se présente comme un spécialiste de la reconnaissance du locuteur. Aucun de ces phonéticiens n'a vécu plus d'un an dans un pays anglophone.

Stimuli

Les stimuli sont, comme pour la soumission à HASR, le fruit d'**une concaténation des deux signaux comme expliqué en 3.3.1**. L'ordre de présentation des stimuli est différente pour chaque auditeur afin de minimiser l'effet d'ordre. Les auditeurs peuvent prendre leur décision à partir du moment où ils ont entendu au moins un extrait des deux enregistrements au moins une fois, soit après 12 secondes d'écoute, mais ils ne

sont pas autorisés à réécouter le stimulus une fois qu'ils l'ont écouté dans son intégralité, soit après 1 minute. Nous avons fait ce choix expérimental car lors du test HASR, les participants prenaient en moyenne leur décision après une seule écoute. De plus, cela permet de réduire la durée du test perceptif. Ce dernier se déroule dans un environnement calme et chaque auditeur est doté d'un casque audio.
Le test perceptif auquel a participé le groupe **des auditeurs non-expérimentés**, que nous nommerons HASR2-ext, consiste en **une cohorte de 102 comparaisons**. Il se compose des 51 comparaisons cible de HASR2 et de 51 comparaisons imposteur choisies selon deux critères. Nous avons gardé les 9 comparaisons imposteur de HASR1 et avons sélectionné les 42 comparaisons imposteur les plus difficiles pour le système ALIZE/SpkDet que nous avons utilisé pour estimer la complémentarité entre les réponses des auditeurs et du système automatique. **Le test perceptif est donc équilibré entre comparaisons cible et comparaisons imposteur** comme indiqué dans le tableau 3.6. En revanche, cette cohorte n'est pas équilibrée entre locuteurs masculins (27 comparaisons) et féminins (75 comparaisons). Ceci ne nous permet pas d'étudier précisément l'influence du genre sur les performances, ce qui est regrettable. La durée du test perceptif est de l'ordre d'une heure et trente minutes. Ainsi, chaque auditeur a passé le test en deux sessions.
Afin de réduire cette durée pour **les auditeurs expérimentés, plus difficiles à trouver, nous avons limité le panel de stimuli** pour ce groupe d'auditeurs. Le test auquel ont participé le groupe d'auditeurs expérimentés, que nous nommerons HASR1-ext, se compose de l'ensemble des stimuli de HASR1 ainsi que des trois comparaisons cible les plus difficiles pour ALIZE/SpkDet de HASR2 afin d'équilibrer le nombre de comparaisons cible et imposteur à 9.

Protocole	Type de Comparaison	Hommes	Femmes	Total
HASR1-ext	Cible	5	4	9
	Imposteur	1	8	9
HASR2-ext	Cible	20	31	51
	Imposteur	7	44	51

TABLE 3.6 – *Répartition des comparaisons cible et imposteur pour HASR-ext.*

Prise de décision et score de confiance

La décision finale est, comme précédemment, prise par vote majoritaire. Ici encore, nous souhaitons définir un score pour effectuer des comparaisons avec le système

automatique SVM d'ALIZE/SpkDet décrit en 3.3.1. Ce score dépend du taux d'accord entre les auditeurs. Plus ceux-ci sont en accord pour dire qu'il s'agit du même locuteur dans les deux enregistrements plus le score est élevé comme l'illustre la figure 3.7. Les valeurs de score s'appuient une nouvelle fois sur les distributions de scores imposteur et cible du système comme l'illustre la figure 3.7. Lorsque jusqu'à un tiers des auditeurs affirment qu'il s'agit du même locuteur, les valeurs de la distribution de scores obtenus en comparaison imposteur sont retenues pour le mapping, un tiers des auditeurs répondant « oui » correspondant à la moyenne des scores des comparaisons imposteur. De même, si entre 2/3 et 100% des auditeurs sont d'accord pour dire qu'il s'agit du même locuteur, ce sont les valeurs des scores obtenus en comparaison cible qui sont utilisées, la moyenne de la distribution des scores de ALIZE/SpkDet étant la valeur choisie lorsque deux tiers des auditeurs affirment qu'il s'agit du même locuteur. Enfin lorsque entre 1/3 et 2/3 des auditeurs déclarent qu'il s'agit du même locuteur, nous utilisons une interpolation des deux distributions.

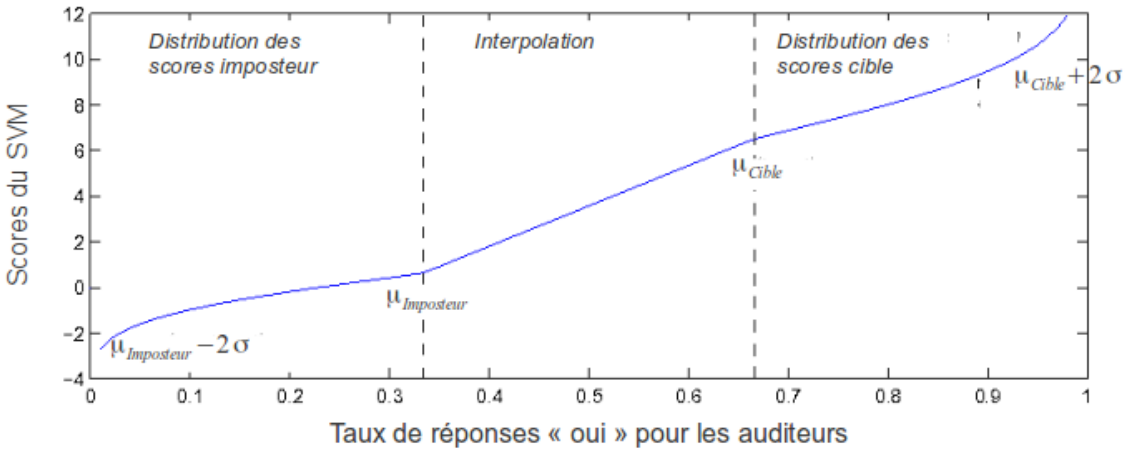


FIGURE 3.7 – *Correspondance entre la proportion de réponses des auditeurs estimant qu'il s'agit du même locuteur et les valeurs de scores attribuées selon les distributions d' ALIZE/SpkDet.*

3.4.2 Performance globale

Les auditeurs inexpérimentés ont un taux de réussite de 58.8%, ce qui n'est pas significativement différent du hasard ($p = 0.092$).
Le taux de FA est de 43.1% et celui de FR est de 39.2%. La performance des auditeurs n'est pas différente de celle du hasard aussi bien pour les comparaisons cible

($p = 0.161$) que pour les comparaisons imposteur ($p = 0.401$).
Les auditeurs expérimentés ont, quant à eux, un taux de réussite de 39%. Le taux de FA est de 44% et celui de FR est de 78%.
Dans ce cas aussi, les taux ne sont pas différents du hasard ($p_{\text{Réponses correctes}} = 0.4807$, $p_{\text{FR}} = 0.1797$ et $p_{\text{FA}} = 1$). Il est toutefois à noter que le nombre de comparaisons pour le groupe des expérimentés est très réduit. Il est difficile de définir des tendances uniquement avec 18 comparaisons.

Les auditeurs, aussi bien expérimentés qu'inexpérimentés, ne semblent pas faire mieux que le hasard dans cette tâche de discrimination. Ce résultat confirme que la cohorte est difficile.

3.4.3 Performance par auditeur

Auditeurs inexpérimentés

Pour le groupe des auditeurs inexpérimentés, dont les résultats individuels sont présentés en Annexe B.2, le taux de réponses correctes (FA et FR confondus) varie de 44% à 66% en fonction de l'auditeur. 86% des auditeurs ont des performances globales qui ne se distinguent pas de celles obtenues en répondant au hasard. Ainsi, seuls quatre auditeurs font mieux que le hasard ($p < 0.05$) pour cette tâche comme l'illustre la figure 3.8.

Il est à noter que le taux de réponses correctes est légèrement corrélé aux nombres d'années d'étude de l'anglais par les auditeurs ($r = 0.437$,$p = 0.016$). Ce résultat confirme l'influence de la connaissance de la langue sur la performance des auditeurs (Goggin et al., 1991).
Les taux de FR et de FA varient respectivement de 22% à 73% et de 20% à 80%. Aucun auditeur ne voit à la fois son taux de FA et son taux de FR différer du hasard. Parmi les quatre auditeurs dont le taux est légèrement différent du hasard, deux ont un taux de FA équivalent au hasard et un taux de FR légèrement différent du hasard tandis que les deux autres ont un taux de FA différent du hasard et un taux de FR équivalent au hasard.

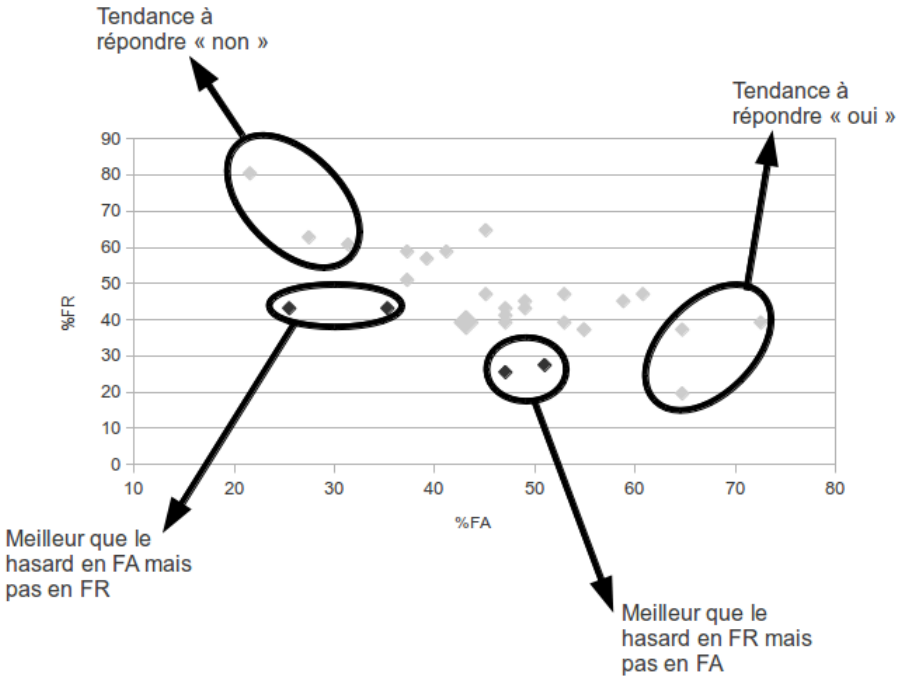


FIGURE 3.8 – *Taux de FA et de FR pour chaque auditeur inexpérimentés, seuls les éléments foncés ont un taux de réussite différent du hasard*

Les auditeurs semblent développer des stratégies de réponses positives ou négatives comme l'illustre la figure 3.9. Quatre auditeurs ont un taux de réponses « oui » inférieur à 40% et différent du hasard ce qui signifie qu'ils ont plutôt tendance à répondre « non ». Un des auditeurs développe clairement une stratégie de réponse négative ($FA = 22\%$ avec $p < 0.001$ et $FR = 80\%$ avec $p < 0.001$). Au contraire, cinq auditeurs ont un taux de réponse de « oui » supérieur à 60% et significativement différent du hasard. Les autres auditeurs semblent équilibrer leurs réponses positives et négatives. Ces différentes stratégies sont confirmées par un test de Cochran ($Q(29) = 171.219$, $p < .0001$) qui montre clairement un effet du locuteur sur les performances.

Auditeurs expérimentés

Les taux de réussite individuels du groupe dit expérimenté varient de 28% à 56% comme l'illustre la figure 3.10. Aucun de ces auditeurs ne fait mieux que le hasard dans ses réponses ($0.09625 < p < 1$).

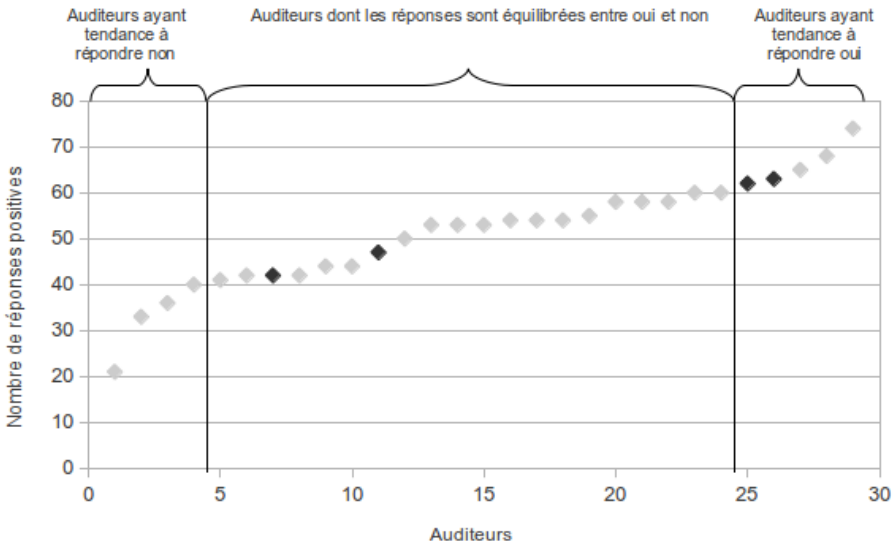


FIGURE 3.9 – *Nombre de réponses positives pour les auditeurs inexpérimentés : les losanges foncés correspondent aux auditeurs qui ont un taux de réussite global différent du hasard*

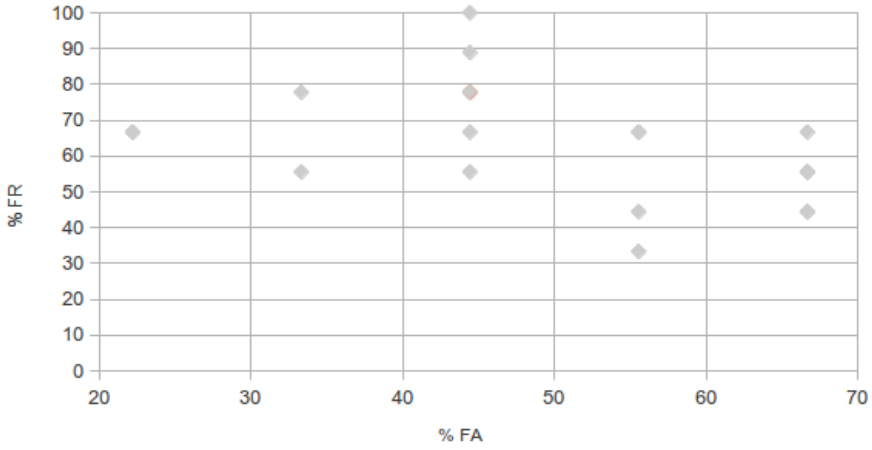


FIGURE 3.10 – *Taux de FA et de FR pour chaque auditeur expérimenté : aucun ne fait mieux que le hasard (tous les losanges sont clairs)*

Ce groupe n'ayant été évalué que sur 18 paires, il est difficile de mettre en évidence des stratégies de réponses par auditeur. Cinq auditeurs ont tendance à développer une stratégie de réponse plutôt négative face aux stimuli (un a une différence significative du hasard, les quatre autres sont proches du seuil de significativité).

Si certains auditeurs obtiennent des performances significativement meilleures que le hasard, ce n'est pas le cas lors de l'analyse des deux types comparaison. Il apparaît que les auditeurs expérimentés ne montrent pas de meilleures performances globales dans cette tâche que les auditeurs inexpérimentés. Ces résultats rendent bien compte du fait que la tâche demandée aux auditeurs est une tâche très difficile. Cette difficulté n'est pas a priori la même sur toutes les comparaisons proposées.

3.4.4 Performance par stimuli

La comparaison des performances en fonction des stimuli est faite **uniquement avec les résultats des auditeurs inexpérimentés** car nous avons alors 51 comparaisons cible et 51 comparaisons imposteur. Les résultats chiffrés par stimuli sont accessibles en Annexe B.3.
La performance des auditeurs selon le stimulus varie drastiquement comme l'illustrent la figure 3.11 pour les comparaisons cible et la figure 3.12 pour les comparaisons imposteur. Le taux de FR varie de 10% à 87% et celui de FA de 3% à 90%.

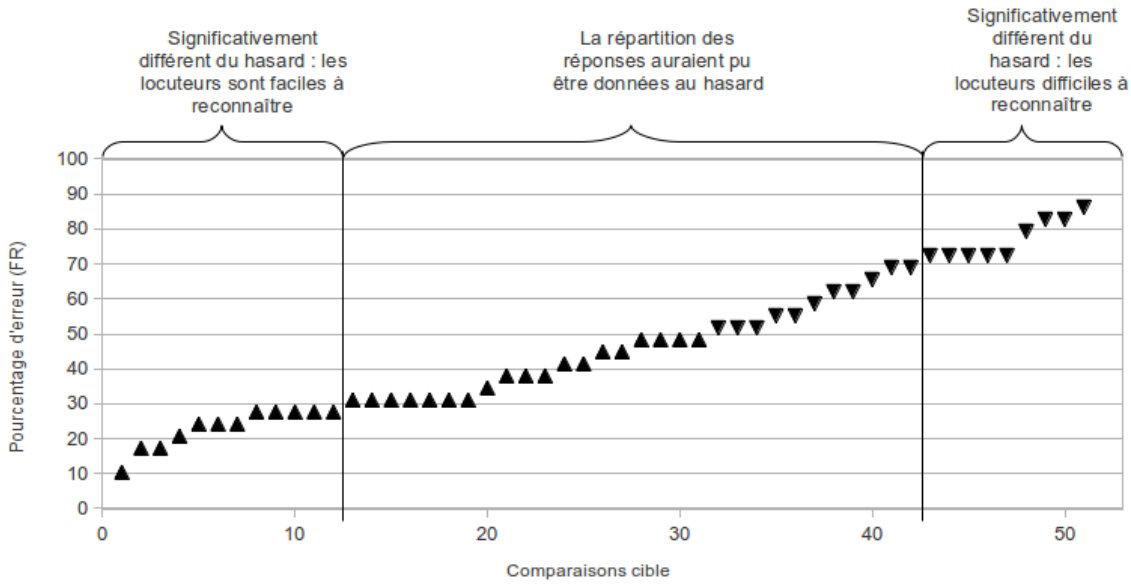


FIGURE *3.11* – *Taux de réussite en fonction des stimuli en comparaison cible*

Certains stimuli permettent donc une comparaison bien plus facile que d'autres. Cependant, aussi bien en comparaison cible qu'en comparaison imposteur, une grande ma-

jorité des stimuli ont des distributions de réponses qui ne diffèrent pas du hasard (respectivement 59% et 61%).

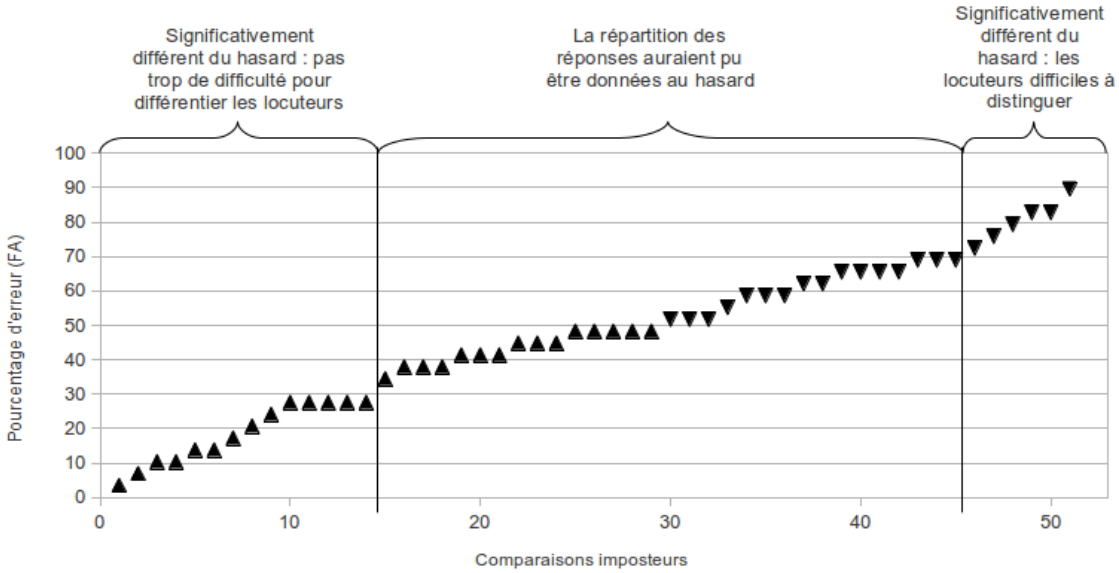


FIGURE 3.12 – *Taux de réussite en fonction des stimuli en comparaison imposteur*

Les auditeurs ont répondu correctement dans une proportion supérieure à celle du hasard pour 24% des comparaisons cible et 12% des comparaisons imposteur. Il est toutefois à noter que, pour certains stimuli, les auditeurs donnent des réponses significativement différentes du hasard mais fausses. Ainsi pour 24% des comparaisons cible, les auditeurs ont donné majoritairement et ce de façon significativement différente du hasard une réponse « non » . C'est le cas également pour 12% des comparaisons imposteur. Le fait que pour une comparaison donnée, les réponses des auditeurs soient différentes du hasard n'est pas un gage de confiance dans la réponse comme l'illustre le tableau 3.7.

TABLE 3.7 – *Nombre de comparaisons pour lesquelles le nombre de réponses est significativement différent du hasard.*

	Comparaisons cible		Comparaisons imposteur	
	Réponses correctes	Réponses incorrectes	Réponses correctes	Réponses incorrectes
>hasard	12	9	14	6

Tous les stimuli n'ont donc pas la même difficulté pour les auditeurs : 25% d'entre eux sont correctement traités par les auditeurs (taux de réponses correctes significativement différent du hasard), au contraire, 15% des stimuli posent problème (taux de réponses incorrectes significativement différent du hasard). Enfin, la tâche étant très difficile, 60% des stimuli ont des résultats qui ne sont pas différents de ceux obtenus par le hasard.
Il nous semble intéressant d'analyser le contenu des stimuli que nous avons repérés comme faisant une certaine unanimité entre les auditeurs pour comprendre sur quels indices se sont appuyés les auditeurs pour prendre leur décision qu'elle soit correcte ou non. A l'écoute, il apparaît que certains stimuli sont tellement bruités qu'une analyse de la parole est rendue extrêmement difficile voire impossible. Nous ne pouvons pas sur ces signaux explorer plus avant les possibles indices mesurables dans les extraits de parole qui permettent aux auditeurs de déterminer s'il s'agit ou non de la même personne.

3.4.5 Complémentarité entre les réponses automatiques et celles obtenues par tests perceptifs

Dans le même esprit que celui de l'évaluation HASR proposé par NIST, nous avons voulu étudier **la possible complémentarité entre les systèmes de RAL et les auditeurs**. Le taux de réussite sur le panel HASR-2 du système SVM est de 63% avec un taux de FA de 39% et un taux de FR de 35%. Ce taux global est significativement différent du hasard ($p < 0.05$). Si le taux de FR est significativement différent du hasard ($p < 0.05$), le taux de FA ne l'est pas ($p = 0.1608$). Ces résultats sont en deçà de ceux habituellement obtenus par ce système, la cohorte étant particulièrement difficile.
Les erreurs effectuées par le système automatique et par le panel d'auditeurs ne se situent pas au niveau des mêmes comparaisons comme l'illustre la figure 3.13.

En comparaisons cible, seules 41% des paires sont discriminées correctement à la fois par le système automatique et par le panel d'auditeurs. 20% des paires ne sont correctement discriminées que par les humains et 25% des paires ne le sont que par le système. Enfin 16% des paires ne sont discriminées correctement ni par le système automatique ni par la cohorte d'humains.
Des résultats similaires sont observés en comparaison imposteur, 35% des paires sont

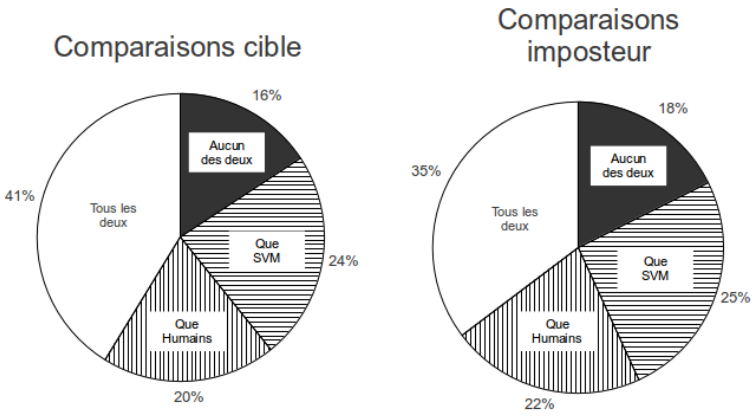


FIGURE 3.13 – *Évaluation de la complémentarité des réponses entre la cohorte des auditeurs inexpérimentés et ALIZE/SpkDet*

correctement caractérisées à la fois par le système et la cohorte d'auditeurs. 22% des paires ne sont correctement discriminées que par les humains et 25% des paires ne le sont que par le système. Enfin 18% des paires ne sont discriminées correctement ni par le système automatique ni par le panel d'humains.

Une certaine complémentarité est observée entre la cohorte humaine et le système automatique. Cette complémentarité peut être exploitée en donnant un poids à chacune décision prise à l'aide d'une régression logistique (Wonnacott et Wonnacott, 1991) par exemple, mais étant donné que les réponses sont pour une grande majorité des stimuli, peu différentes du hasard, les désaccords entre système automatique et panel d'auditeurs peuvent être attribués à cette part d'aléa des réponses.

Synthèse du chapitre

Nous avons cherché à évaluer **la capacité d'auditeurs natifs du français à discriminer les locuteurs** issus du corpus HASR2 et à **comparer les réponses des auditeurs avec celles d'un système automatique** classiquement utilisé en vérification du locuteur. Cette tâche se présente comme **une tâche difficile** pour les raisons suivantes.

- Les auditeurs ne sont pas natifs de la langue parlée par les locuteurs à discriminer.
- Les auditeurs ne connaissent pas les locuteurs qu'ils vont devoir discriminer et la

durée d'enregistrement (2 minutes 30 secondes) accessible ne permet pas de faire un apprentissage préalable des locuteurs par les auditeurs (Saslove et Yarmey, 1980).

– Les conditions d'enregistrement des deux extraits à comparer sont systématiquement différentes.

Sur l'ensemble des comparaisons proposées, **les résultats par vote majoritaire ne sont pas meilleurs que les performances qu'il est possible d'obtenir en tirant au hasard les réponses fournies**. D'ailleurs, seuls quatre auditeurs obtiennent une performance meilleure que celle obtenue par le hasard. La tâche proposée est donc très difficile pour **les auditeurs, qui ne présentent pas d'aptitude particulière qu'ils soient expérimentés ou non**. Certains auditeurs montrent des **stratégies de réponses** en répondant plutôt « oui » ou plutôt « non ». Ces stratégies globales se retrouvent pour les autres participants à HASR.

Les performances des auditeurs varient drastiquement en fonction des stimuli. Toutes les paires ne sont donc pas équivalentes pour les auditeurs. Il est important de comprendre ce qui différencie les paires qui font l'unanimité pour les auditeurs des comparaisons pour lesquelles ils répondent au hasard : des indices propres au locuteur sont peut-être très présents dans les premiers et moins présents dans les seconds.

Ce genre de procédures est tout de même très long, les tests perceptifs prenant beaucoup de temps pour finalement n'avoir accès qu'à peu de données par rapport au nombre de comparaisons auxquelles nous pouvons avoir accès en n'utilisant que des systèmes automatiques. Les signaux de parole de HASR sont très bruités et empêchent une analyse des possibles indices propres au locuteur. Ne serait-il pas possible de repérer les enregistrements pour lesquels les systèmes automatiques obtiennent de bonnes performances et ceux pour lesquels les performances sont moins bonnes afin d'établir par la suite quelles différences de contenus ont ces enregistrements ?

Chapitre 4

Sensibilité des systèmes

Résumé : *Les métriques actuellement utilisées pour évaluer la performance d'un système de RAL sont fondées sur une mesure moyenne des erreurs sur un très grand nombre de comparaisons. Quelques études se sont attachées à étudier l'influence du locuteur sur les performances de systèmes (Doddington et al., 1998) démontrant que certains locuteurs entraînent plus d'erreurs que d'autres. Dans ce chapitre nous nous interrogeons sur l'influence sur les performances de systèmes de l'enregistrement de parole utilisé en apprentissage pour modéliser le locuteur. Notre méthode consiste, dans un premier temps, à tester sur les mêmes données de nombreux enregistrements produits par le même locuteur. Nous sélectionnons, dans un second temps, pour chaque locuteur le fichier qui minimise le taux FA+FR au seuil de l'EER et le fichier qui le maximise. Enfin, nous comparons les performances obtenus par ces fichiers sur une cohorte identique. Cette méthode est appliquée sur deux bases de données (NIST-SRE 2008 (Martin et Greenberg, 2009) et BREF 120 (Lamel et al., 1991)), et pour deux systèmes (ALIZE/SpkDet (Bonastre et al., 2008) et Idento (Scheffer et al., 2011)). Les taux de variation relative observés sont de 1.4 pour NIST et de 2.6 pour les locuteurs de BREF.*

Sommaire

4.1 Hypothèses d'étude

Dans ce chapitre nous souhaitons étayer les hypothèses suivantes.

– S'il existe bien des profils de locuteur qui influencent les performances des systèmes de vérification du locuteur (Doddington et al., 1998), les indices idiosynchratiques issus de la parole sont des paramètres qui sont également utilisés pour décrire d'autres fonctions de la parole. **Les indices dépendent donc non seulement du locuteur mais également du contexte d'énonciation** qui est notamment dépendant de l'interaction entre les locuteurs, ou du contenu sémantique délivré. Les techniques d'enregistrement ou la durée des signaux, très étudiés pour les systèmes automatiques, sont des éléments de variation importants. Dans ces conditions, **tous les enregistrements d'un même locuteur ne sont pas équivalents pour le modéliser et nous devrions même à durée d'enregistrement équivalente observer des écarts de performances**.

– Cette variation de performance étant due à la **nature même des données**, nous devrions observer des variations de performances **quelque soit le système** utilisé.
– **Plus une base de données est variable** (parole conversationnelle vs parole lue), **plus la variation de performance** en fonction des extraits de parole sélectionnés **doit être importante**.
– Plus la durée des enregistrements est longue, moins la variation de performance est importante.

Pour apporter des éléments de réponse à ces hypothèses, nous devons, avant toute chose, **disposer de nombreux enregistrements d'un même locuteur et comparer les performances obtenues pour chacun des modèles**. Nous commencerons d'abord par

présenter les bases de données et les systèmes avec lesquels nous avons travaillé puis nous nous interrogerons sur les possibles méthodes pour mesurer la sensibilité des systèmes à cette variabilité. Enfin, nous présenterons les résultats que nous avons obtenus.

4.2 Bases de données

Nous avons travaillé avec deux bases de données, NIST 08 et BREF 120 qui sont deux bases fournies lors d'évaluation de technologies de la parole dont nous disposons au LIA.

4.2.1 NIST 08 : téléphone, conversationnel, multilingue

La première base de données est issue du corpus NIST 2008 (Martin et Greenberg, 2009). Nous nous sommes centré sur la partie **téléphonique**.
Dans ce cadre, des locuteurs, vivants aux États-Unis et qui ne se connaissent pas, sont mis en relation deux à deux et aléatoirement. **Les locuteurs sont enregistrés une ou plusieurs fois mais dans le second cas, ils changent d'interlocuteur**. Aucune consigne ne leur est donnée sur les thèmes à aborder ni sur la langue à utiliser dans la communication.
Une conversation dure en moyenne 311 secondes (environ 5 minutes). Chaque locuteur est enregistré sur une piste différente. Ainsi il est **possible d'extraire l'enregistrement d'un seul locuteur**.
Les téléphones à partir desquels les locuteurs communiquent sont de **différents types** (cellulaire, filaire...) et les signaux sont échantillonnés à 8000 Hz.
221 locuteurs-homme ont été enregistrés lors de ces conversations téléphoniques. Les enregistrements ainsi recueillis ont permis de générer les 11 636 comparaisons imposteur et les 874 comparaisons cible qui constituent l'évaluation NIST 08 pour cette tâche téléphonique.
Lors de ces conversations, **dix-huit langues** sont parlées par ces locuteurs comme résumé par le tableau 4.1.

Les langues présentes appartiennent à des familles de langues différentes (6 indo-européennes, 1 austronésienne, 5 sino-tibétaines, 3 altaïques, 1 chamito-sémitique, 1 tai-Kadaï et 1

TABLE 4.1 – *Langues présentes dans la base données dans M-08*

Langues	Nombre de fichiers	Nombre de locuteurs
Arabe Égyptien	3	1
Bengali	8	5
Chinois Min Nan	6	1
Chinois Min Nan et Mandarin	1	1
Mandarin	12	10
Mandarin et Chinois cantonnais	1	1
Anglais (Natif, non-natif)	554	164
Farsi	3	3
Hindi	48	24
Hindi.Anglais Indien	7	7
Italien	6	4
Japonais	19	7
Coréen	12	6
Russe	23	17
Tagalog	4	2
Thaï	41	15
Ouzbek	3	2
Vietnamien	29	12
Wu (Chinois de Shanghai)	6	2
Chinois cantonnais	30	13

austroasiatique). La localisation de ces langues, illustrée par la figure 4.1, montre qu'une grande proportion d'entre elles sont d'Asie du sud-est.

La langue majoritairement utilisée reste l'anglais (seulement 4% des locuteurs n'ont aucun enregistrement en anglais). 55% des locuteurs parlent deux langues. Cependant, 35% des locuteurs n'ont des enregistrements que dans une seule langue, l'anglais, tandis que 10% des locuteurs parlent 3 langues (cf tableau 4.2).

TABLE 4.2 – *Nombre de locuteurs en fonction du nombre de langues parlées dans M-08*

Nombre de langues parlées par locuteur	1	2	3
Nombre de locuteurs	60	95	16

Ce corpus contient donc de nombreuses langues avec des locuteurs plurilingues. Il est vraisemblable que les performances du système de RAL soient sensibles à cette diver-

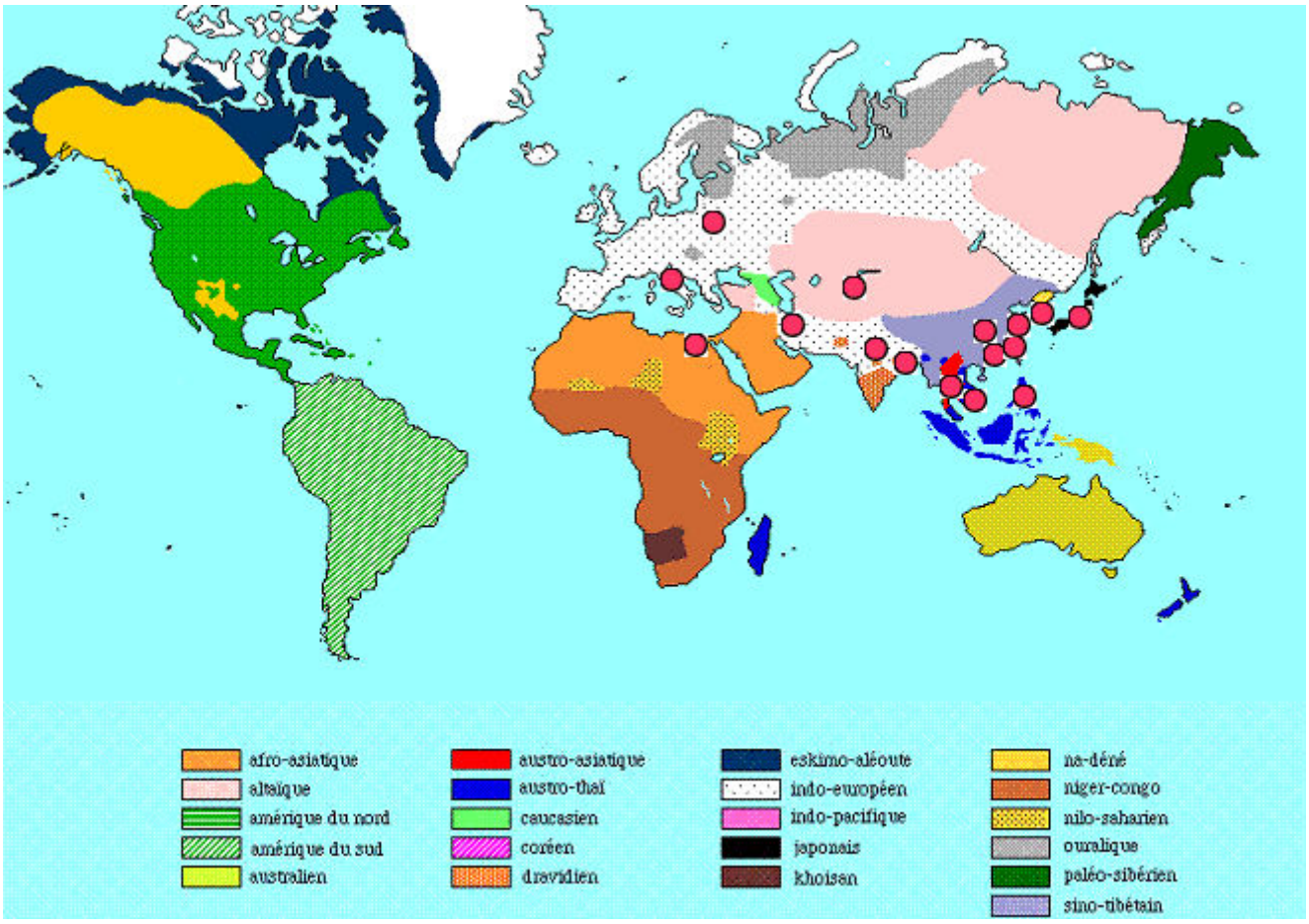


FIGURE 4.1 – *Localisation des langues présentes dans M-08, fond de carte de Vallée et Arnal 2000*

sité linguistique.

L'objectif de notre étude étant d'**analyser l'influence du choix du signal d'apprentissage sur les performances d'un système de RAL**, nous n'avons conservé que les **171 locuteurs ayant au moins 3 enregistrements différents**.
Afin de profiter de l'ensemble des fichiers présents dans la base de données pour mener notre étude, nous avons utilisé une **procédure de *Leave-one-out***, c'est-à-dire que chaque fichier a été utilisé pour construire un modèle de locuteur puis a été testé avec l'ensemble des fichiers présents dans la base de données exceptés celui qui a servi à l'apprentissage du modèle. L'ensemble des comparaisons créé par cette procédure est appelé dans notre document M-08. Cette cohorte se compose de 3 624 comparaisons cibles et de 661 416 comparaisons imposteur. Dans ces comparaisons cible et imposteur, 816 en-

registrements différents sont exploités. Le nombre de modèles par locuteur varie entre 3 et 20. Les caractéristiques de NIST 08 et de M-08 sont résumées dans le tableau 4.3.

Base	Locuteurs	Modèles	Comparaisons de fichiers	
			Imposteurs	Cibles
NIST 2008	221	648	11 636	874
M-08	171	816	661 416	3 624

TABLE 4.3 – *Locuteurs et Modèles pour NIST 08 et M-08.*

En utilisant les mêmes enregistrements de 2.5 minutes élaborés dans le cadre de la campagne NIST-SRE 08, nous avons construit une cohorte de comparaisons cible et imposteur, M-08, qui permet d'étudier les variations de performances dues à des enregistrements de diverses conversations (types de téléphone, langues, distribution phonétique, nombre de trames sélectionnées par le système de RAL, hésitations, rire et bruit de bouches dues au mode conversationnel...). Il est à noter que si la durée moyenne des enregistrements est de 2 minutes 30 secondes, cela ne signifie pas que le système utilise les 15 000 trames du signal pour construire le modèle. En effet, comme nous l'indiquions en 1.4.1 une sélection de trames est toujours effectuée par le système. Ainsi en moyenne, ALIZE/SpkDet sélectionne 73 secondes de signal utilisé. Cette sélection peut varier de façon importante en fonction des enregistrements (sur M-08, entre 45 secondes et 1.8 minutes de paroles sélectionnées).
Il nous semble intéressant de travailler également avec une base de données où les différentes sources de variabilité sont mieux contrôlées notamment en vérifiant que le nombre de trames sélectionnées est stable.

4.2.2 BREF 120 : microphone, parole lue, français natif

La base de données BREF 120 (Lamel et al., 1991) a été constituée à l'origine pour construire un système de reconnaissance de la parole grand vocabulaire pour le français. Elle comporte 120 locuteurs qui ont lu des phrases issues du journal *Le Monde*. Tous les locuteurs ne lisaient pas exactement les mêmes phrases. Il est toutefois à noter que l'ensemble des phrases prononcées par un locuteur comportaient tous les phonèmes du français distribués comme dans la langue française. **Tous les locuteurs ont été enregistrés en une seule session avec le même matériel d'enregistrement**.
Pour notre étude, les enregistrements des 9 locuteurs francophones non-natifs ont été

écartés. **Nous n'avons conservé que les locuteurs natifs du français, soit 64 femmes et 47 hommes**. Nous appellerons cette base de données dans notre manuscrit BREF.
Pour chaque locuteur, nous avons concaténé des enregistrements de phrases qu'il avait prononcées de manière à obtenir **39 fichiers** pour lesquels environ 30 secondes de trames avaient été sélectionnées par le système de RAL. Contrairement à M-08, nous avons ici contrôlé le nombre de trames sélectionnées par le système. Les enregistrements ont une durée moyenne de 42 secondes.
Nous avons un nombre de fichiers par locuteur qui est beaucoup plus important que pour NIST-08, nous ne sommes pas obligé d'utiliser une procédure de *Leave-on-out*. Ainsi, pour constituer notre panel de comparaisons, **18 fichiers ont été sélectionnés aléatoirement comme fichiers d'apprentissage tandis que les 21 autres ont été utilisés comme signaux de test**. Ainsi, ont été réalisées, pour chaque locuteur homme, 17 766 comparaisons (378 ($1 \times 21 \times 18$) comparaisons cible et 17 388 ($46 \times 21 \times 18$) comparaisons imposteur) et, pour chaque locuteur femme 24 192 comparaisons (378 ($1 \times 21 \times 18$) comparaisons cible contre 23 814 ($63 \times 21 \times 18$) comparaisons imposteur) ; soit au total, 835 002 tests pour les hommes et 1 548 288 tests pour les femmes.
Dans un second temps, lorsque nous avons voulu observer **l'influence de la durée d'enregistrement sur les écarts de performance d'un système de RAL**, nous avons concaténé les fichiers d'apprentissage précédemment décrits de manière à obtenir des enregistrements où plus de 2 minutes 30 secondes de trames sont sélectionnées. Nous avons comparé ces nouveaux modèles avec les fichiers test utilisés précédemment. Dans ces conditions, pour chaque locuteur, seuls 3 fichiers sont disponibles en apprentissage. Ainsi, 4 032 et 2 961 comparaisons cible et 254 016 et 136 206 comparaisons imposteur ont été réalisées respectivement pour les femmes et les hommes dans cette condition appelée BREF-2min30svs30s. Il est à noter que pour un modèle donné, le nombre de comparaisons est le même dans les conditions BREF et BREF-2min30svs30s. Le tableau 4.4 résume le nombre de comparaisons pour chaque condition.

Base	Locuteurs	Modèles	Comparaisons de fichiers	
			Imposteurs	Cibles
BREF Femmes 30s	64	1 152	1 524 096	24 192
BREF Hommes 30s	47	846	817 236	17 776
BREF Femmes 2min30svs30s	64	192	1 524 096	24 192
BREF Hommes 2min30svs30s	47	141	817 236	17 776

TABLE 4.4 – *Nombre de comparaisons pour le corpus BREF selon les durées d'enregistrement et le genre des locuteurs.*

4.3 Systèmes utilisés

Nous avons testé 2 systèmes différents sur M-08 et un seul système sur BREF et BREF-2min30svs30s. Ces deux systèmes correspondent à deux systèmes état-de-l'art entre 2008 et 2011, ALIZE/SpkDet et Idento décrits précédemment. Nous spécifierons les bases de données ayant été utilisés à l'apprentissage des éléments constitutifs du modèle (modèle du monde, Matrice U ou Matrice de translation pour la PPCA...). En effet, les données utilisées pour l'apprentissage de ces structures sont primordiales dans l'élaboration des futurs modèles (Scheffer, 2006).

4.3.1 ALIZE/SpkDet

Comme décrit précédemment, ALIZE/SpkDet (Bonastre et al., 2008) est un système de RAL basé sur le **paradigme UBM/GMM** (Reynolds et al., 2000) développé notamment au LIA. Le modèle du monde est constitué de 1024 gaussiennes. Lors de l'adaptation des modèles, les matrices de co-variances sont conservées, seules les valeurs moyennes sont modifiées à l'aide d'un algorithme MAP.
ALIZE/SpkDet inclut les techniques de Factor Analysis (Kenny et al., 2005). La dimension de la matrice U utilisée pour le Factor Analysis est de 40.
Le vecteur de paramètres utilisé est de dimension 60. Il se compose de 19 *Linear Filter Cepstral Coefficients* (LFCC) ainsi que l'Énergie (C0) ainsi que des Delta et Delta-Delta. Ces paramètres sont calculés sur une fenêtre de 10 ms.
Lors de la configuration pour M-08, nous avons utilisé les données de langue anglaise issues de Fisher (Cieri et al., 2004) et NIST 2005 (NIST, 2005) pour constituer le modèle du monde. La matrice U du Factor Analysis a été obtenue à partir des données NIST 2006 (NIST, 2006). Cette configuration correspond à la soumission effectuée par le LIA lors de l'évaluation NIST 2008 (Matrouf et al., 2008b). Nous n'avons cependant pas effectué de normalisation des scores.
Pour BREF, le modèle du monde est calculé à partir des données de langue française enregistrées en studio et extraites des corpus ESTER-1 (Gravier et al., 2004) et ESTER-2 (Galliano et al., 2005). La technique du Factor Analysis n'a pas été utilisée dans ce cas là car tous les locuteurs ont été enregistrés dans les mêmes conditions, c'est-à-dire un

studio d'enregistrement. Aucune normalisation des scores n'a été prise en compte.

4.3.2 Idento

Idento (Scheffer et al., 2011) est un système de RAL développé au SRI basé sur la technique des i-vector (Dehak et al., 2009). Ce système a été utilisé lors d'un séjour de février à mai 2011 au sein de ce laboratoire californien.
Les corpus NIST 2004 (NIST, 2004), 2005 (NIST, 2005) et 2006 (NIST, 2006) ont servis à construire le modèle du monde, les matrices nécessaires à la construction des i-vectors. Le vecteur de paramètres est de dimension 60 et est composé de 20 Mel Filter Cepstral Coefficients (MFCC) ainsi que des 20 Delta et des 20 Delta-Delta.
Nous comparerons les résultats obtenus en No-Norm avec ceux obtenus en ZT-Norm afin de mesurer l'influence de la normalisation sur la variation de performance.

4.4 Performances par locuteur : à la recherche des agneaux et des chèvres

Les performances obtenues pour M-08 et BREF 30s sont étudiées en fonction des locuteurs afin de vérifier si certains locuteurs sont sujets à plus d'erreurs que d'autres. Les locuteurs que nous recherchons sont les *chèvres* (qui ont des taux de Faux Rejets importants) et les *agneaux* (qui ont des taux de Fausses Acceptations important) (Doddington et al., 1998).

4.4.1 Calcul de la performance

La performance par locuteur étudiée correspond à la performance obtenue par ce locuteur lorsque ses fichiers sont utilisés en apprentissage. Le taux de Faux Rejets et de Fausses Acceptations dépend du seuil choisi. Nous avons décidé de maintenir ce seuil fixe tout au long de nos expériences, en choisissant la valeur du seuil de l'EER sur l'ensemble des comparaisons possibles. C'est à partir de ce seuil que nous allons calculer la performance du système ALIZE/SpkDet pour chaque locuteur.
Nous comptabilisons pour chaque locuteur, indépendamment du fichier d'apprentissage utilisé, le nombre de Fausses Acceptations et de Faux Rejets. A partir de ces

nombres, nous calculons FA_i et FR_i par locuteur i. Nous quantifions également la proportion de FA et de FR attribuée à chaque locuteur. Les *agneaux* sont les locuteurs qui donnent lieu à FA_i élevé tandis que les *chèvres* sont les locuteurs pour lesquels FR_i est important.
Cette analyse est effectuée à partir des scores fournis par ALIZE/SpkDet pour M-08 d'une part et pour BREF d'autre part. Il est à noter que, pour BREF, les locuteurs sont testés sur le même nombre de comparaisons imposteurs d'une part (23 814 et 17 388 par locuteur respectivement pour les femmes et pour les hommes) et de comparaisons cible d'autre part (378 par locuteur pour les femmes et pour les hommes). Ainsi, la précision des taux obtenus est équivalente entre locuteur. En revanche, pour M-08, le nombre de comparaisons varient en fonction du locuteur. Certain taux seront donc plus précis que d'autres.

4.4.2 M-08

Lorsque ALIZE/SpkDet est configuré pour M-08, nous avons calculé pour chacun des 171 locuteurs les FA_i et FR_i avec une valeur de seuil fixé à -0.1549. Ces performances sont illustrées par la figure 4.2.

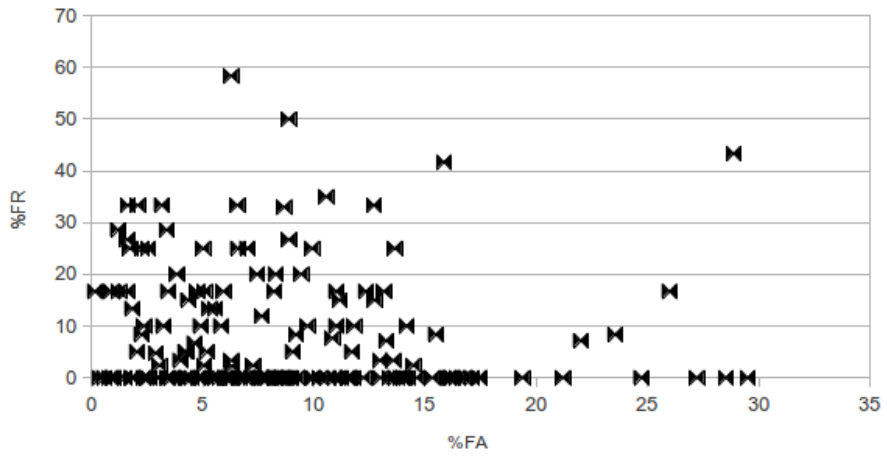


FIGURE 4.2 – *FA_i et FR_i des locuteurs de la base de données M-08.*

Peu de locuteurs ont à la fois des FA_i et des FR_i élevés. Par exemple, le locuteur correspondant au FR_i le plus élevé (58.3%) a un FA_i de 6.3%. De même, le locuteur qui a le FA_i le plus élevé (29.5%) a un FR_i de 0%. Seul un locuteur est clairement diffi-

cile à repérer à la fois en comparaison cible et en comparaison imposteur (FA=28.9% et FR=43.3%). Il est à noter que, pour les trois fichiers d'apprentissage disponibles de ce locuteur, le nombre moyen de trames sélectionnées est inférieur à la moyenne des trames habituellement sélectionnées (5 200 trames pour ce locuteur contre 7 200 en moyenne). Les modèles de ce locuteur sont donc construits avec beaucoup moins de trames que les autres modèles.

La moyenne des FA_i est de 8.8% ; ce taux varie de 0.15% à 29.5% ($\sigma = 6.03\%$) en fonction des locuteurs. La moyenne des FR_i est de 7.8%, ce taux varie de 0% à 58.3% ($\sigma = 11.66\%$). Les différences en fonction des locuteurs sont importantes, encore plus en comparaison cible qu'en comparaison imposteur. Il ne faut toutefois pas oublier que le nombre de comparaisons cible est nettement réduit par rapport à celui des comparaisons imposteur. Les FR_i sont beaucoup moins précis que les FA_i puisque les FR_i sont calculés sur 21 comparaisons cible en moyenne tandis que FA_i sont calculés sur 3 868 comparaisons imposteur en moyenne. Par ailleurs, il est à noter que tous les locuteurs ne sont pas évalués sur le même nombre de comparaisons, étant donné qu'ils n'ont pas tous le même nombre d'enregistrements disponibles.

Certains locuteurs semblent être à l'origine de nombreuses erreurs. Ainsi, 50% des FR sont attribués à seulement 9% des locuteurs, les *chèvres* de (Doddington et al., 1998). Ce phénomène est illustré par la figure 4.3 qui représente les FA cumulés et les FR cumulés en fonction des locuteurs. Pour les FR, une rupture claire apparaît entre les locuteurs. Les FA montrent une continuité depuis les *moutons* jusqu'au *agneaux*. Ce phénomène est sûrement dû au fait que le nombre de comparaisons n'est pas du tout le même entre comparaisons cible et comparaisons imposteur.

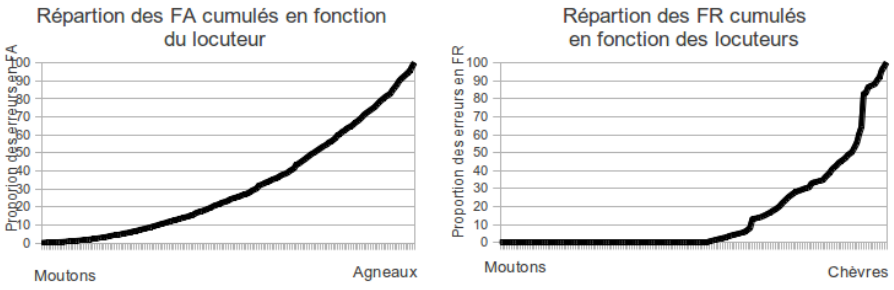


FIGURE 4.3 – *M-08 : FA et FR cumulé en fonction des locuteurs*

4.4.3 BREF

ALIZE/SpkDet, configuré pour le français est testé sur BREF. Les performances obtenues, FA_i et FR_i, sont étudiées séparément pour les locuteurs et les locutrices, comme l'illustrent les figures 4.4 et 4.5. Le seuil utilisé pour les hommes est 0.0162 et celui pour les femmes est 0.0174.

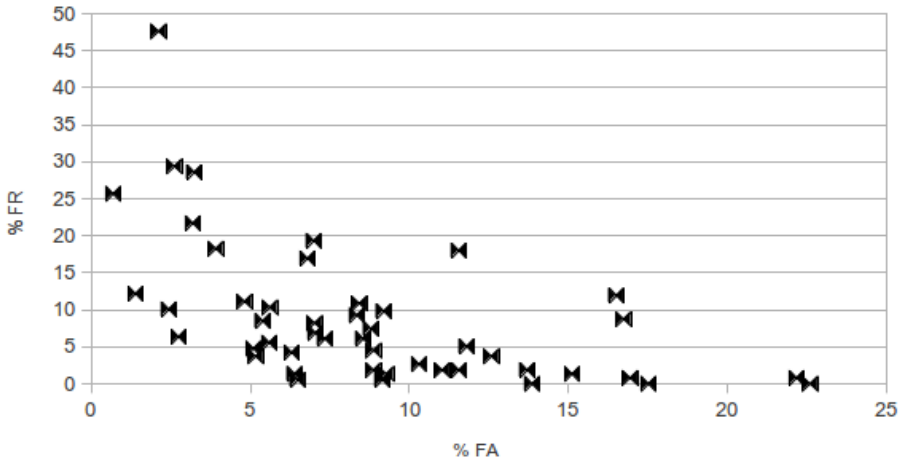


FIGURE 4.4 – *FA_i et FR_i des locuteurs de la base de données BREF*

Pour les hommes, la moyenne des FA_i est de 8.8%. Ce taux varie de 0.7% à 22.6% en fonction des locuteurs ($\sigma = 5.14$). La moyenne des FR_i est de 8.8%, avec une plage de 0% à 47.6% en fonction des locuteurs ($\sigma = 9.57$). La variation est plus importante pour les FR que pour les FA.
Les locuteurs présentent rarement des FA_i et des FR_i élevés. Par exemple, le locuteur qui a le FR_i le plus élevé (47.6%) a un FA_i bas (2.1%). Inversement, le locuteur qui a le FA_i le plus élevé (22.6%) a un taux de FR de 0%.

Pour les femmes, la moyenne des FA_i est de 10%. Il varie de 1.04% à 27.5% en fonction des locutrices ($\sigma = 5.74$). La moyenne des FR_i est de 10% également. Il varie de 0% à 35.2% ($\sigma = 9\%$).
Le FR_i maximal atteint (35.2%) par une locutrice correspond à un FA_i de 2% et le FA_i maximal atteint (27.5%) correspond à un FR_i de 0.3%.
Les FA_i et les FR_i sont calculés avec respectivement 23 814 comparaisons imposteur et 378 comparaisons cible. Un net déséquilibre existe encore entre les deux types de com-

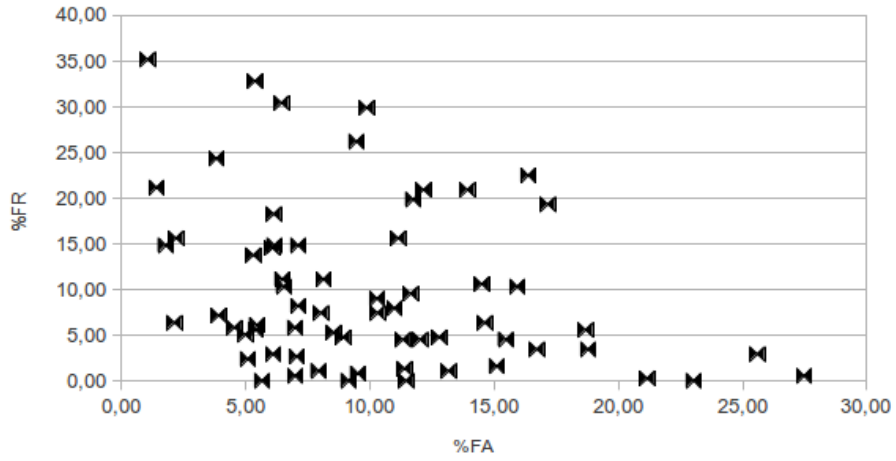


FIGURE 4.5 – *FA_i et FR_i des locutrices de la base de données BREF*

paraisons. Ceci pourrait expliquer les écarts observés entre FA et FR. Il est toutefois à noter que dans BREF, les locuteurs sont tous testés avec le même nombre de comparaisons, ainsi les écarts obtenus dans les mêmes conditions sont tout à fait comparables et ils montrent bien qu'il existe des différences en termes de performance importantes en fonction du locuteur.

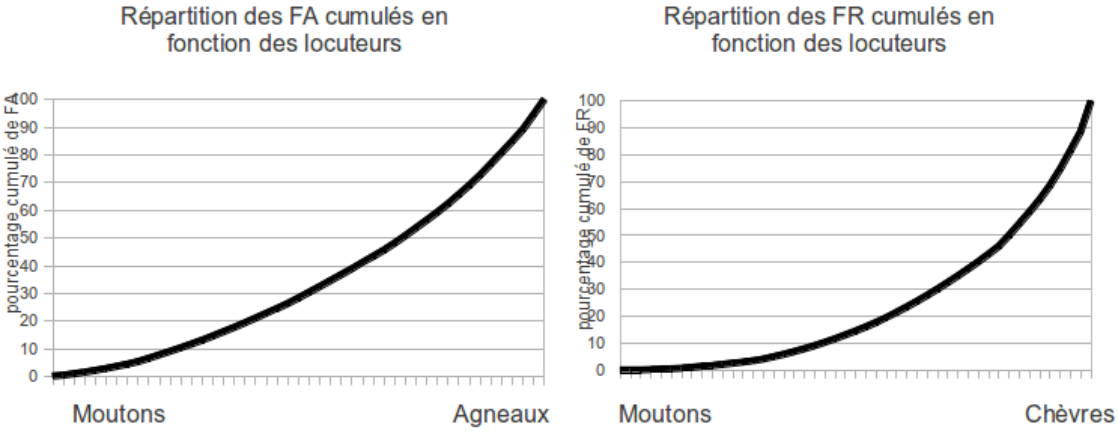


FIGURE 4.6 – *BREF : FA et FR cumulés en fonction des locuteurs*

Sur BREF, 50% des FA sont dus à 30% des locuteurs et 50% des FR sont dus à 19% des locuteurs. Pour les femmes, 50% des FR sont dus à 20% des locutrices et 50% des FA sont dus à 30% des locutrices. Ce phénomène est mis en évidence par les figures 4.6 et 4.7 qui illustrent la répartition des FA et de FR cumulés pour les hommes et les femmes.

Les relations ne sont pas linéaires et montrent là encore un continuum entre les profils de locuteurs que sont les *chèvres*, les *agneaux* et les *moutons*.

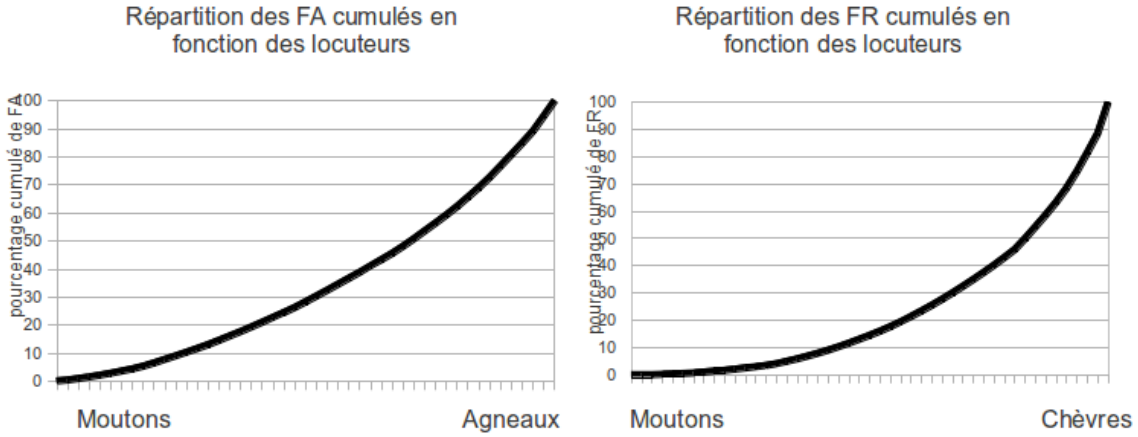


FIGURE 4.7 – *BREF : FA et FR cumulés en fonction des locutrices*

Pour les deux bases de données, nous avons bien trouvé parmi les locuteurs et locutrices étudiés des *chèvres* et des *agneaux*, mettant ainsi en évidence à la suite de (Doddington et al., 1998) qu'il existe des différences propres aux locuteurs. Ces résultats par locuteurs ont été calculés indépendamment du fichier d'apprentissage utilisé. Tous les enregistrements utilisés pour modéliser le locuteur sont-ils équivalents ? Quelle sensibilité a le système au fichier d'apprentissage utilisé ?

4.5 Mesurer la sensibilité des systèmes

Mesurer la sensibilité des systèmes automatiques aux fichiers d'apprentissage revient à **quantifier les différences en terme de performance qu'amène le changement de l'enregistrement utilisé en apprentissage**. Plusieurs angles d'étude peuvent être envisagés. Dans un premier temps, nous proposons d'utiliser comme élément de comparaison **les distributions de scores cible et imposteur obtenues pour chaque modèle**. Cette solution connaissant quelques limites, nous présenterons, dans une second temps, la méthode que nous avons adoptée et qui **s'appuie sur les FA et FR moyens**.

4.5.1 Analyse des distributions de scores

Une première approche pour rendre compte de la sensibilité des systèmes à la variabilité des fichiers d'apprentissage est de **comparer les distributions de scores obtenues en comparaison cible et imposteur pour un locuteur en fonction des enregistrements utilisés pour l'apprentissage du modèle**.
En effet, nous pouvons faire l'hypothèse que, pour un même locuteur, les distributions de scores observées pour une même série de comparaisons devraient être sensiblement les mêmes si les enregistrements d'un même locuteur sont équivalents pour le modéliser.
Les figures 4.8 et 4.9 sont des exemples de distributions de scores que nous avons obtenues pour la base de données M-08 en comparaison imposteur.

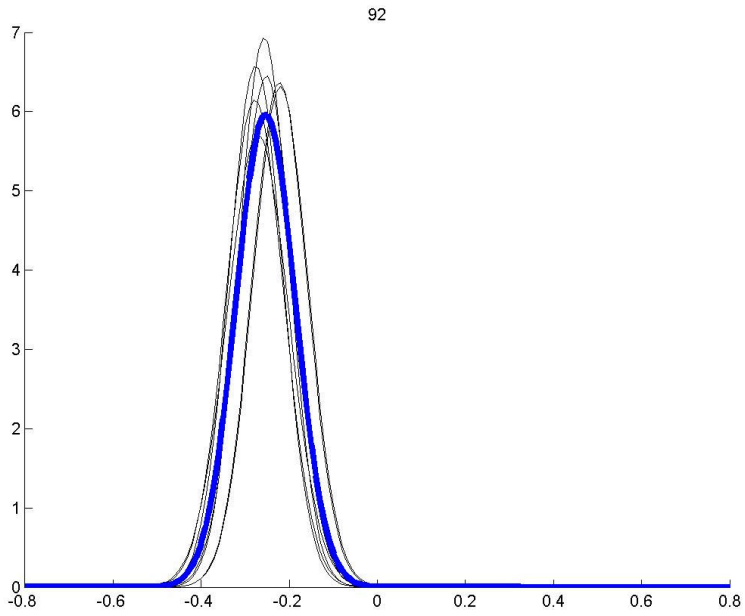


FIGURE 4.8 – *Distribution de scores imposteur pour les différents modèles d'un locuteur de la base de données M-08 : les distributions sont globalement les mêmes. La distribution au trait plus épais correspond à la moyenne des distributions*

Pour le premier locuteur (92), nous observons que les distributions sont très proches les unes des autres, tandis que pour le second (locuteur 209), nous voyons plutôt deux distributions différentes apparaître. Si deux distributions différentes apparaissent, cela signifie que les modèles de ce locuteur ne se comportent pas de la même manière face

aux tests.

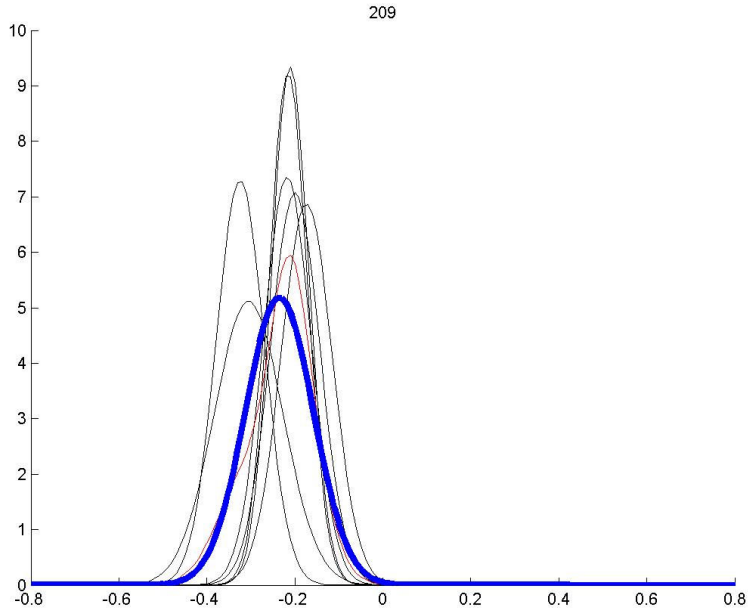


FIGURE 4.9 – *Distribution de scores pour les différents modèles d'un locuteur de la base de données M-08 : 2 groupes de distributions avec des moyennes différentes. La distribution au trait plus épais correspond à la moyenne des distributions*

Ces différences de distribution ont évidemment des conséquences sur les FA et les FR. Le seuil de décision à l'EER étant -0.1549, les modèles dont les distributions de scores sont les plus à gauche (moyennes les plus faibles) donneront lieu à moins de FA que les distributions de droite.

Par cette première représentation, nous avons mis en évidence que les fichiers d'apprentissage ne conduisent pas aux mêmes distributions de scores en comparaison imposteur. Il apparaît que tous les fichiers ne sont pas équivalents.
L'analyse des résultats en comparaison cible pose, en revanche, plus de problème. En effet, nous n'avons accès qu'à peu de comparaisons cible par locuteur (entre 3 et 20 pour M-08 et 21 pour BREF). Il n'est pas possible de comparer des distributions de scores dans ce cas.
Par ailleurs, une autre difficulté de cette méthode réside dans **l'obtention d'une mesure des variations observées qui soit comparable avec les métriques largement utilisées**

afin d'avoir des valeurs de référence. En effet, si nous voyons des différences entre les distributions, nous ne savons pas mesurer facilement leur implication. De plus, ce type de comparaisons sépare les comparaisons cible des comparaisons imposteur et il n'est pas facile de tenir compte du recouvrement entre les deux.
Une autre représentation pour mesurer cette sensibilité du système doit donc être envisagée.

4.5.2 Utiliser le meilleur et le pire modèle

Définir FA_{ij} et FR_{ij} sur la totalité des données pour la sélection du meilleur et du pire modèle

Comme nous l'avons fait pour les locuteurs, nous pouvons obtenir **le taux de Fausses Acceptations, FA_{ij} et de Faux Rejets, FR_{ij}, avec le seuil de l'EER pour chaque locuteur i et chaque modèle j**. Il est possible alors de **déterminer pour chaque locuteur quel est le meilleur et le pire modèle en fonction de ces taux**. Le meilleur modèle est celui qui minimise la somme de FA et de FR tandis que le pire maximise cette somme. La sélection du meilleur et du pire modèle est réalisée sur le plus grand nombre de comparaisons possibles (intégralité de M-08, de BREF ou de BREF 2min30svs30s).

Établir différentes séries de tests où seul change le fichier d'apprentissage

Une fois le meilleur et le pire modèle sélectionné pour chaque locuteur, il s'agit de **mesurer l'écart de performance entre les deux modèles du même locuteur**. Pour effectuer la comparaison, une cohorte de fichiers test est définie. Au lieu de rattacher chaque fichier test à un fichier d'apprentissage comme cela est fait habituellement pour définir une comparaison, nous rattachons chaque fichier test à un locuteur qui est considéré comme le locuteur d'apprentissage. Une comparaison est donc ici composée d'un locuteur « d'apprentissage » et d'un fichier de test. Nous pouvons définir de nombreuses comparaisons afin d'être certains que les différents locuteurs interviennent dans la cohorte et qu'ils seront tous testés en comparaison cible et imposteur. **Cette cohorte est le canevas qui répertorie la série de test**. En effet, pour chaque locuteur nous pouvons choisir un fichier d'apprentissage dont nous connaissons à priori la performance local $FA_{ij} + FR_{ij}$ (Meilleur, Pire ou aléatoire). Une série de tests correspond au

canevas tel que nous l'avons défini où le locuteur est modélisé à l'aide du fichier d'apprentissage de notre choix. Ainsi, **chaque série de tests se compose exactement des mêmes locuteurs et des mêmes fichiers de test, seul change le fichier d'apprentissage utilisé**. Une fois le fichier d'apprentissage sélectionné, un locuteur est représenté par le modèle élaboré à partir d'un seul fichier d'apprentissage dans toute la série.
Afin de mesurer l'influence du choix de ce fichier sur les performances du système, nous avons **réalisé plusieurs séries de tests**. Pour la première série, nous avons utilisé en apprentissage pour chaque locuteur son **meilleur modèle (série *Min*)**, puis nous avons réalisé la série de tests en utilisant en apprentissage **le pire modèle du locuteur (série *Max*)**. **Cette démarche nous permet de mesurer l'écart maximum de performance que nous pouvons observer pour les mêmes tests et les mêmes locuteurs lorsque seuls les fichiers d'apprentissage changent**. Pour établir la performance du système lorsque le fichier d'apprentissage n'est ni le meilleur ni le pire, **des tirages aléatoires de modèles** ont également été réalisés pour le corpus BREF. Ces modèles ont ensuite été testés dans les mêmes conditions que les séries *Min* et *Max*. Pour le corpus issus de NIST, nous avons conservé les fichiers qui étaient la référence lors de l'évaluation.

Comparer des performances globales

La performance globale est mesurée à l'aide d'une courbe DET et d'un taux d'EER pour chacune des séries. Nous pouvons ainsi comparer les performances obtenues pour chacune des séries et rendre compte de la variation de performance due au fichier d'apprentissage puisque c'est l'unique élément qui change entre nos séries de tests.

Pour M-08, la cohorte de tests choisie est celle proposée par NIST où les 171 locuteurs de M-08 sont testés en apprentissage. Comme certains fichiers que nous avons sélectionnés comme meilleurs ou comme pires étaient à l'origine utilisés comme fichiers test, nous avons du supprimer certaines comparaisons. Cette cohorte se compose de 511 comparaisons cible et 2 856 comparaisons imposteur.
Pour BREF et BREF 2min30svs30s, la cohorte de tests est composée de l'ensemble des tests déjà décrit en 4.2.2. Pour les femmes, la série se compose de 1 344 comparaisons cible et de 84 672 comparaisons imposteur. Pour les hommes, la série se compose de 987 comparaisons cible et 45 402 comparaisons imposteur. Toutes les conditions de com-

paraisons sont résumées par le tableau 4.5.

Bases de données		Comparaisons cible	Comparaisons imposteur	Nombre de locuteurs
NIST		511	2 856	171
BREF	Femmes	1 344	84 672	64
	Hommes	987	45 402	47

TABLE 4.5 – *Nombre de comparaisons utilisées pour mesurer la sensibilité des systèmes à la variabilité intra-locuteur sur les bases de données NIST et BREF*

La démarche que nous avons adoptée permet de mesurer les écarts maximaux que nous pouvons observer en fonction des fichiers d'apprentissage sélectionnés. Les comparaisons de performance sont faites à l'aide des mesures classiques.
La variation relative, Vr, entre les séries pour chaque système et chaque base de donnée testée est définie par l'équation 4.1. **Cette mesure nous permet de rendre compte de la variation due aux données d'apprentissage.**

$$Vr = \frac{EER_{Max} - EER_{Min}}{EER_{Moyen}} \tag{4.1}$$

4.6 Variabilité de la performance

Nous commencerons par décrire les résultats obtenus pour les séries de tests issues de NIST puis les résultats obtenus pour celles de BREF.

4.6.1 NIST

Approche UBM-GMM

La courbe DET de la figure 4.10 présente les résultats obtenus par ALIZE/SpkDet. Si la série *Min* obtient un EER à 4.1%, la série *Max* a un EER de 21.9%. La sélection correspondant à celle de NIST conduit à un EER de 12.1%. Dans ce cas, $Vr = 1.47$.

Nous observons un écart de performance de plus de 17 points absolus ce qui représente 34% d'erreur en plus. Le choix du modèle d'apprentissage a donc des conséquences très

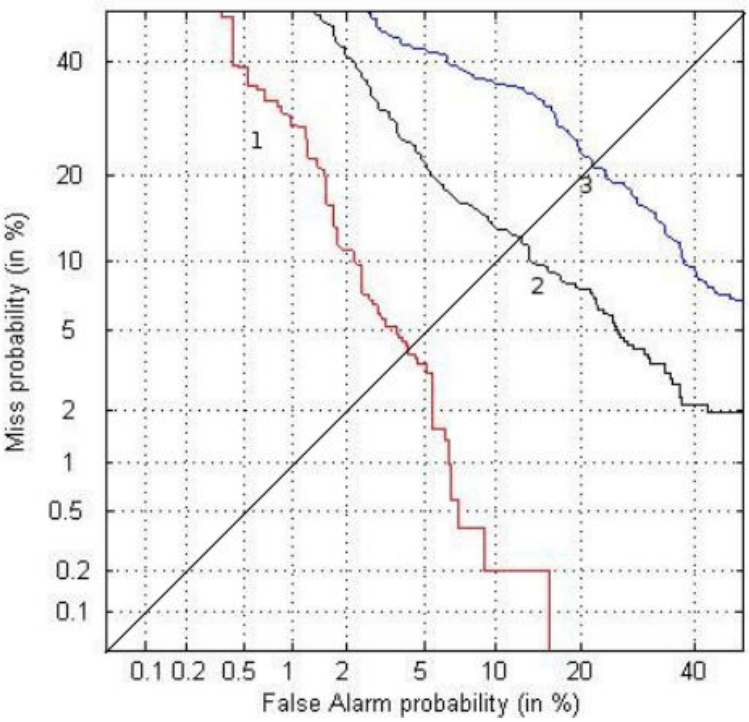


FIGURE *4.10* – *Courbes DET pour les séries Min, Max et aléatoire pour la base de données NIST 2008 (511 comparaisons cible et 2 856 comparaisons imposteur) testées sur le système ALIZE/SpkDet : fluctuation de performance de 4.1% à plus de 21.9%*

importantes sur les performances du système. Ceci est indépendant du type de locuteur puisque ce sont exactement les mêmes locuteurs qui sont comparés dans les deux séries. La série correspondant à NIST montre que si les fichiers d'apprentissage sont tirés aléatoirement, la performance du système se situe entre les deux série *Min* et *Max* et rend compte d'une performance moyenne en lissant les écarts importants dus au choix du fichiers d'apprentissage.

Le système ALIZE/SpkDet est sensible au choix du fichier d'apprentissage pour modéliser le locuteur. Tous les fichiers d'apprentissage ne sont donc pas équivalents pour ce système. Afin de conforter ces résultats et de vérifier que ces écarts sont dus, non pas au système, mais au contenu des enregistrements de parole choisi pour représenter le locuteur, nous avons, lors de notre séjour au SRI, réalisé les mêmes séries de test avec Idento.

Approche i-vector

Les courbes des figures 4.11 et 4.12 présentent les résultats obtenus en utilisant Idento. La figure 4.11 illustre les résultats sans normalisation des scores tandis que la figure 4.12 présente les résultats avec une normalisation ZT.

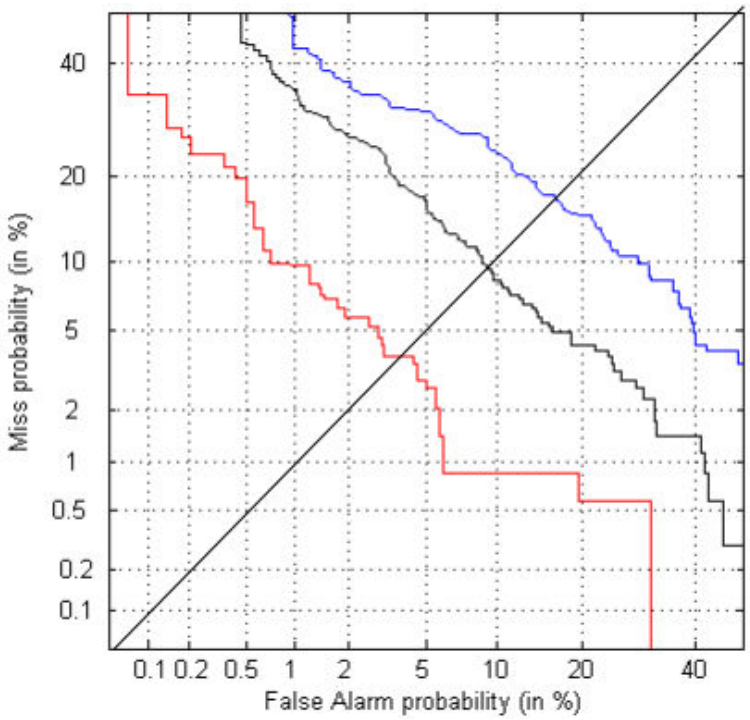


FIGURE 4.11 – *Courbes DET pour les séries Min, Max et aléatoire pour les séries de tests issues de NIST 2008 (511 comparaisons cible et 2 856 comparaisons imposteur) testées sur le système Idento sans normalisation : fluctuation de performance de 3.8% à plus de 16.8%*

Sans normalisation, l'EER varie de 3.8% pour la série *Min* à 16.8% pour la série *Max*. La série où les modèles correspondent à ceux choisi par NIST a un EER de 9.2%. Dans ce cas, $Vr = 1.41$. Des écarts de performance s'observent donc également pour un système basé sur les i-vectors, dans des proportions semblables à celles obtenues pour ALIZE/SpkDet.

La normalisation ne corrige que partiellement ces écarts de performance. En ZT-norm, l'EER varie de 3.1% avec la série *Min* à 13.8% pour *Max*, le tirage correspondant à NIST conduisant à un EER de 7.3%. *Vr* est alors égal à 1.46. Les écarts de performances exis-

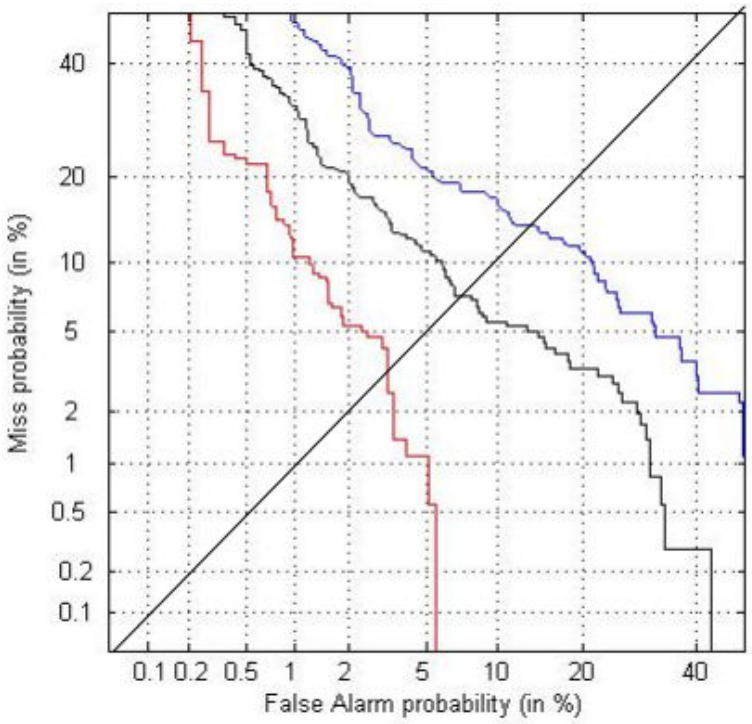


FIGURE 4.12 – *Courbes DET pour les séries Min, Max et aléatoire pour la base de données NIST 2008 (511 comparaisons cible et 2 856 comparaisons imposteur) testées sur le système Idento avec une normalisation ZT-norm : fluctuation de performance de 3.1% à plus de 13.8%*

tent toujours malgré la normalisation.

Sur les séries de tests issues de NIST, nous observons des variations relatives de performance entre 1.41 et 1.47. Il existe des écarts de performances aussi bien pour l'approche UBM-GMM que pour l'approche i-vector et la normalisation des scores ne permet pas de corriger ces écarts de variation. Ceux-ci sont donc à attribuer aux enregistrements utilisés pour modéliser le locuteur. Certains fichiers contiennent plus d'indices idiosynchratiques que d'autres. En étudiant les fichiers de chaque série *Min* et *Max* sélectionnés pour ALIZE/SpkDet et Idento, il est apparu que seul 30% des fichiers qui sont considérés comme les pires pour le système ALIZE/SpkDet le sont aussi pour Idento. De même, 30% des fichiers qui sont considérés comme les meilleurs pour le système ALIZE/SpkDet le sont pour Idento. Il semble donc qu'**il existe une certaine variabilité entre le systèmes pour déterminer le meilleur et le pire enregistrement**. Ceci est sans doute dû à la mesure de $FRij$ qui est calculée sur peu de comparaisons cible.

La cohorte de NIST connaît de nombreuses sources de variation comme la langue, les conditions d'enregistrement ou encore le nombre de trames sélectionnées pour construire le modèle. En effet, si les enregistrements ont une durée moyenne de 2 minutes 30 secondes, cela ne signifie pas que le système sélectionne la même quantité de données pour construire le modèle du locuteur. D'ailleurs, il existe une différence significative entre le nombre de trames sélectionnées entre la série *Min* et *Max* d'ALIZE/SpkDet ($F = 11.11$, $p < 0.001$). Ce sont peut-être ces critères qui expliquent les écarts de performance. Voyons quelles sont les variations de performance lorsque la base de données est plus contrôlée.

4.6.2 BREF

Seul le système ALIZE/SpkDet a été évalué à l'aide de la base de données BREF. Ici encore, les fichiers de tests sont exactement les mêmes pour chacun des locuteurs, seul change entre les séries le fichier d'apprentissage utilisé pour représenter le locuteur. Nous avons utilisé, dans un premier temps, des fichiers d'enregistrements qui comportaient entre 30 et 33 secondes de trames sélectionnées puis nous avons travaillé avec des fichiers de 2.5 minutes de trames sélectionnées afin de **mesurer l'influence de la quantité des données d'apprentissage sur la variation de performance**.

30 secondes de trames sélectionnées

Les courbes DET des figures 4.13 et 4.14 sont respectivement calculées sur 1 344 comparaisons cible et 84 672 comparaisons imposteur pour les femmes et sur 987 comparaisons cible et 45 402 comparaisons imposteur pour les hommes.
Pour les femmes, l'EER varie de 1.1% pour la série *Min* à 28.5% pour la série *Max*. Les 10 séries pour lesquelles nous avons choisi aléatoirement le fichier d'apprentissage de chacune des locutrices ont un EER qui varie de 8.8% à 11.5% avec une moyenne de 10.3% et un écart type de 1.1. En utilisant un EER moyen de 10.3%, *Vr* est égal à 2.66.

Pour les hommes, l'EER fluctue de 1.0% à 33.0% entre la série *Min* et la série *Max*. L'EER est en moyenne de 9.0% avec un écart type de 1.4, il fluctue entre 6.3% et 11.6% pour les dix séries où le choix du fichier d'apprentissage se fait aléatoirement. Avec un EER

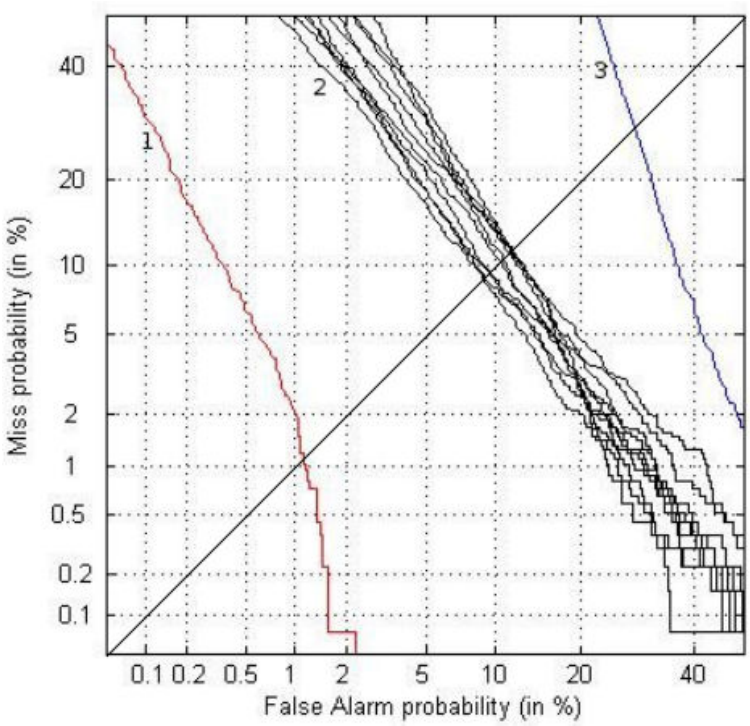


FIGURE 4.13 – *Courbes DET pour les séries Min, Max et 10 séries aléatoires pour les séries de tests issues de BREF, avec 64 femmes (1 344 comparaisons cible et 84 672 comparaisons imposteur) et des enregistrements de 30 secondes de trames sélectionnées en apprentissage et en test, testées sur le système ALIZE/SpkDet : fluctuation de performance de 1.1% à 28.5%*

moyen de 9.0%, *Vr* est ici égal à 3.55.

Dans un contexte plus contrôlé (même microphone, langue unique, parole lue...), des écarts de performances très importants sont également observés. Il apparaît clairement que les enregistrements utilisés ne comportent pas tous la même quantité d'information nécessaire pour modéliser le locuteur par un système de RAL. Nous avons ensuite étudié l'influence de la durée des enregistrements sur ces écarts de performance.

2.5 minutes de trames sélectionnées

Les courbes DET des figures 4.15 et 4.16 sont calculées comme précédemment sur 1 344 comparaisons cible et 84 672 comparaisons imposteur pour les femmes et sur 987

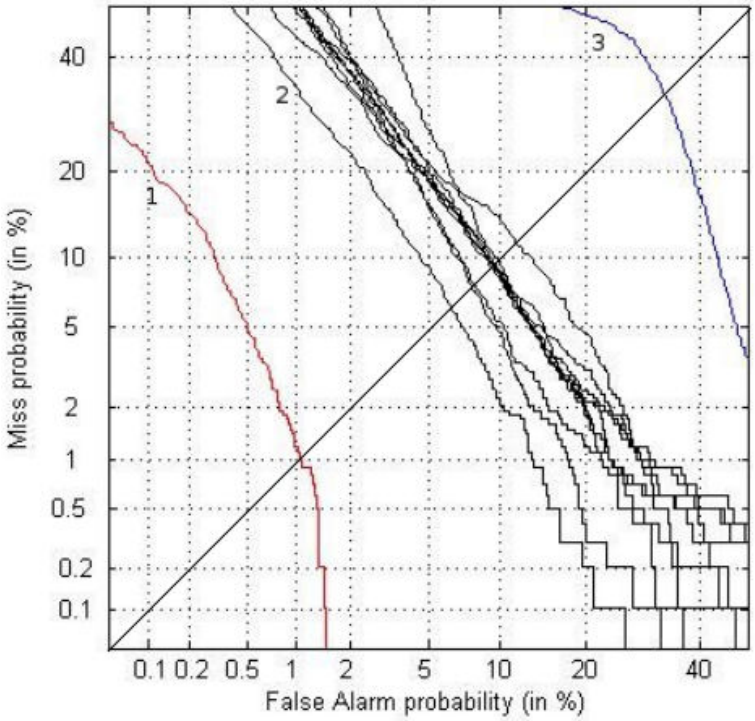


FIGURE 4.14 – *Courbes DET pour les séries Min, Max et 10 séries aléatoires pour les séries de tests issues de BREF, avec 47 hommes (987 comparaisons cible et 45 402 comparaisons imposteur) et des enregistrements de 30 secondes de trames sélectionnées en apprentissage et en test, testées sur le système ALIZE/SpkDet : fluctuation de performance de 1.0% à 33.0%*

comparaisons cible et 45 402 comparaisons imposteur pour les hommes.
Pour les femmes, l'EER varie de 0.9% à 6.0%.

Le même phénomène est observé pour les hommes : l'EER fluctue de 1.0% à 5.8%.

Si l'augmentation du nombre de trames sélectionnées a une influence importante sur les pires modèles (l'EER passant de 33% à 6%), il est toutefois possible avec seulement 30 secondes de parole d'obtenir des résultats équivalents à ceux obtenus avec des enregistrements 5 fois plus longs. La quantité de trames sélectionnées, aspect quantitatif, joue donc un rôle primordial dans la constitution du modèle mais la question de la pertinence de l'information, aspect qualitatif, est également soulevée par cette expérience.
Les résultats présentés dans cette section sont résumés par le tableau 4.6.

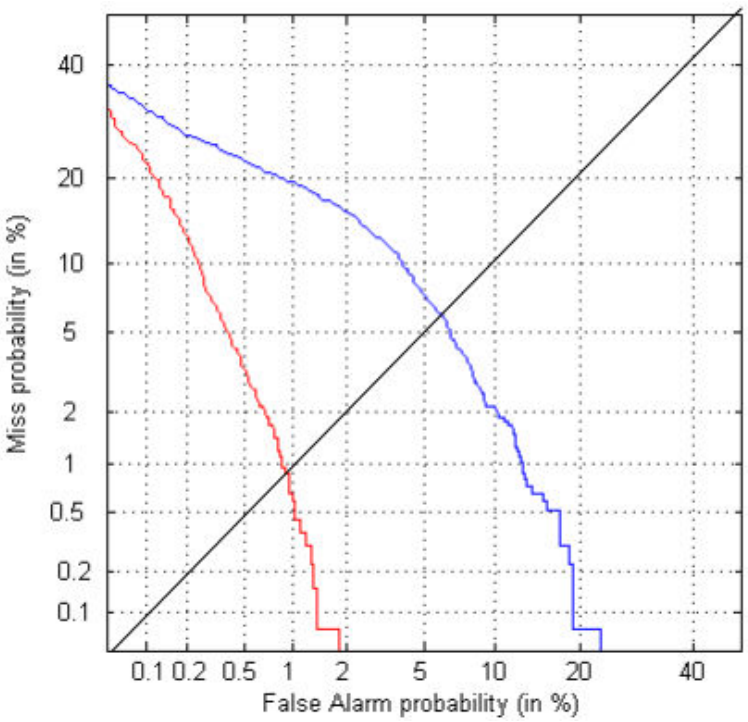


FIGURE 4.15 – *Courbes DET pour les séries Min, Max pour la base de données Bref, avec 64 femmes (1 344 comparaisons cible et 84 672 comparaisons imposteur) et des enregistrements de 2 minutes 30 de trames sélectionnées en apprentissage et de 30 secondes de trames sélectionnées en test, testées sur le système ALIZE/SpkDet : fluctuation de performance de 0.9% à 6.0%*

Conditions					*Min*	*Max*	Aléatoires	*Vr*
Bases de Données			Systèmes					
NIST			ALIZE/SpkDet		4.1%	21.9%	12.1%	1.47
			Idento	No norm	3.8%	16.8%	9.2%	1.41
				ZT norm	3.1%	13.8%	7.3%	1.46
BREF	Femmes	30 secondes	ALIZE/SpkDet		1.1%	28.5%	10.3% (1.1)	2.66
		2.5 minutes			0.9%	6.0%	non calculé	
	Hommes	30 secondes			1.0%	33.0%	9.0% (1.4)	3.55
		2.5 minutes			1.0%	5.8%	non calculé	

TABLE 4.6 – *Variation de performance selon les fichiers d'apprentissage choisis*

Nous avons déterminé pour deux bases de données, l'une régulièrement utilisée en vérification du locuteur et l'autre avec un contenu beaucoup plus contrôlé et pour deux systèmes, ALIZE/SpkDet et Idento, qu'une part importante de la performance des systèmes automatiques dépend du choix des fichiers d'apprentissage. En effet, alors que les locuteurs modélisés et les fichiers de test sont exactement les mêmes

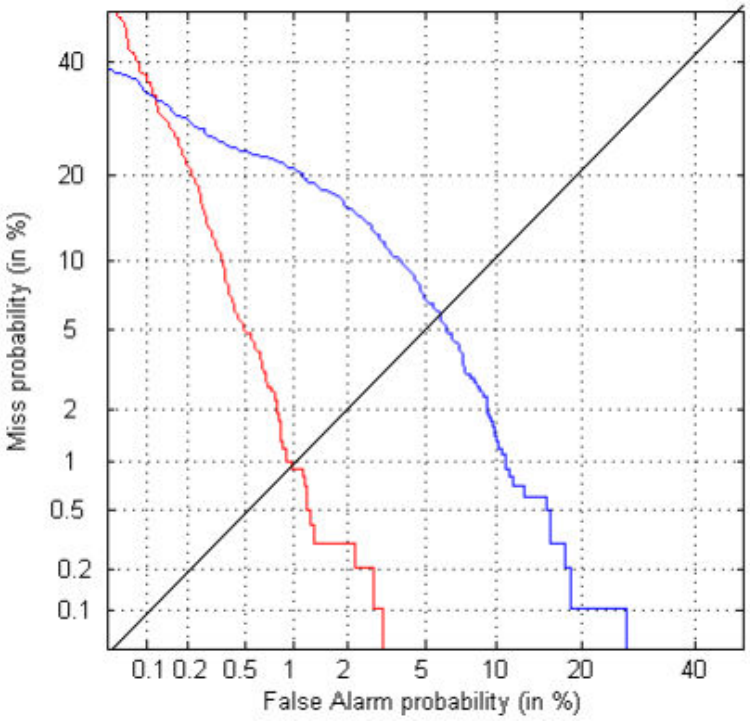


FIGURE 4.16 – *Courbes DET pour les séries Min, Max, avec 47 hommes (987 comparaisons cible et 45 402 comparaisons imposteur) et des enregistrements de 2 minutes 30 de trames sélectionnées en apprentissage et de 30 secondes de trames sélectionnées en test, testées sur le système ALIZE/SpkDet : fluctuation de performance de 1.0% à 5.8%*

nous pouvons observer des variations de performance très importantes. Cette variation semble dépendre de la base de données. A notre surprise, plus la base de données est contrôlée, plus la variation en fonction du fichier d'apprentissage est importante. Ceci peut également être attribué au fait que le nombre de trames sélectionnées est plus faible dans BREF que dans NIST. En effet, la quantité de trames sélectionnées pour construire le modèle a une influence importante sur les écarts de performance. Ceci montre que le choix du fichier d'apprentissage n'est pas anodin : il est important de pouvoir comprendre ce qui différencie les séries de fichiers ayant conduit aux meilleures et aux pires performances afin de pouvoir prédire si un fichier d'apprentissage contient les informations pertinentes permettant de bien modéliser le locuteur. Cet objectif est celui de la partie III de ce document.

4.7 Variation propre au système

Devant l'importance des écarts de performance dus au changement de fichier d'apprentissage, il nous a semblé indispensable de **vérifier la stabilité du système**. **Que se passe-t-il si les fichiers d'apprentissage sont équivalents ?** Pour cela, nous avons construits à partir des fichiers de 2 minutes et 30 secondes de trames sélectionnées du corpus BREF deux modèles différents en utilisant chacun des fichiers. **Le premier modèle comporte toutes les trames impaires du fichier tandis que le second modèle comporte toutes les trames paires**. Nous pouvons considérer que les informations utilisées pour construire les deux modèles sont équivalentes.
Comme précédemment, nous cherchons à déterminer quels sont les écarts maximum de performances que nous pouvons observer en fonction du modèle utilisé. Pour chaque locuteur, nous avons déterminé quel est **le meilleur modèle entre celui construit avec les trames paires et celui construit avec les trames impaire**s. Nous avons effectué la même série de comparaisons en prenant les meilleurs fichiers puis les pires fichiers. Les fichiers tests des comparaisons sont ceux utilisés en 4.6.2. Cette expérience a été menée avec les fichiers des séries *Min* et *Max* uniquement.
Le tableau 4.7 présente les EER dans chacune des conditions.

Genre	Catégorie d'origine des fichiers	Catégorie pour une trame sur deux	EER
Hommes	*Min* ($EER = 1.0\%$)	*Min*	2.1%
		Max	3.2%
	Max ($EER = 5.8\%$)	*Min*	2.7%
		Max	3.2%
Femmes	*Min* ($EER = 0.9\%$)	*Min*	1.2%
		Max	2.7%
	Max ($EER = 6.0\%$)	*Min*	1.2%
		Max	2.3%

TABLE 4.7 – *EER obtenus en prenant une trame sur deux des fichiers Min et Max de BREF 2min30svs30s*

Pour les hommes, lorsque les modèles sont construits à partir des fichiers de la série *Min* ($EER = 1.0$ lorsque l'intégralité des fichiers est utilisée), les meilleurs modèles obtiennent un EER de 2.1% tandis que les pires modèles obtiennent un EER de 3.2%. Lorsque les modèles sont construits à partir de la série *Max* ($EER = 5.8$ lorsque l'intégralité des fichiers est utilisée), les meilleurs modèles obtiennent un EER de 2.7% tandis

que les pires obtiennent un EER de 3.2%.
Pour les femmes, lorsque les modèles sont construits à partir des fichiers de la série *Min* ($EER = 0.9$ lorsque l'intégralité des fichiers est utilisée), les meilleurs modèles obtiennent un EER de 1.2% tandis que les pires modèles obtiennent un EER de 2.7%. Lorsque les modèles sont construits à partir de la série *Max* ($EER = 6.0$ lorsque l'intégralité des fichiers est utilisée), les meilleurs modèles obtiennent un EER de 1.2% tandis que les pires obtiennent un EER de 2.3%.

Que ce soit pour les hommes ou pour les femmes, nous observons un écart de performance de près d'un point d'EER alors que les modèles ont été construits à partir de jeux de données statistiquement équivalents. Lorsque l'on observe les distributions de scores obtenues (figure 4.17), nous voyons bien qu'il existe une différence entre les séries. Cet écart est dû au système mais est largement plus faible que les écart de performance observés précédemment.
Il est par ailleurs étonnant que les performances de la série *Max* soient si proches de celles de la série *Min*. Une analyse de la composition trame à trame des jeux de données utilisés pour l'apprentissage reste nécessaire pour mieux comprendre le comportement du système, qui pourrait être du à la présence de quelques données très spécifiques.

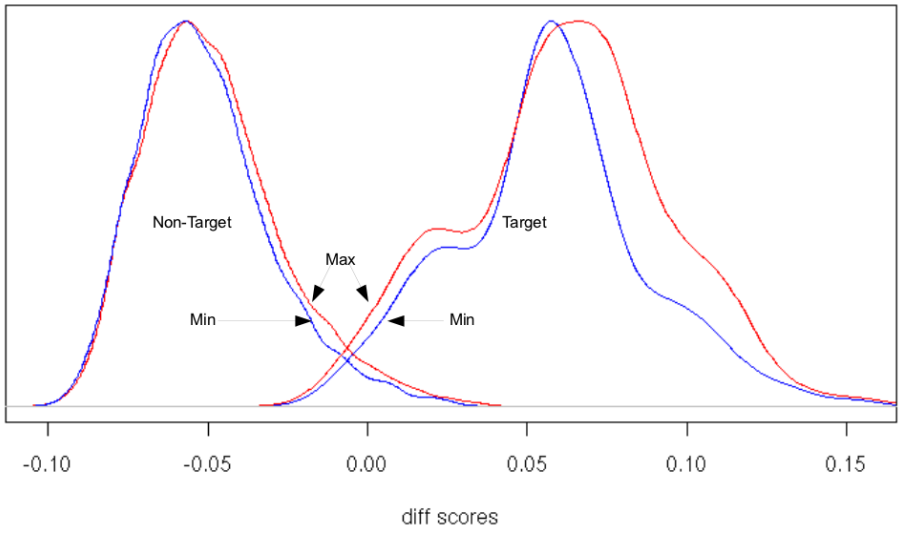


FIGURE 4.17 – *Distribution des scores pour les modèles issus de la série Min des hommes de BREF 2min30svs30s en sélectionnant une trame sur 2*

Synthèse du chapitre

Dans ce chapitre, nous avons montré les éléments suivants :

- Les systèmes de RAL, aussi bien fondés sur une approche UBM-GMM que sur une approche i-vector montrent des **écarts de performances importants** (Pour NIST, $Vr_{Idento} = 1.41$ et $Vr_{ALIZE/SpkDet} = 1.47$ et pour BREF, $Vr = 3.11$) selon le choix du fichier d'apprentissage utilisé pour chaque locuteur.
- La **normalisation des scores n'a que peu d'effet** sur les écarts de performance ($1.41 < Vr_{NoNorm} < 1.47$ et $Vr_{ZTNorm} = 1.46$).
- Ce phénomène est observé aussi bien sur une base de données où les enregistrements sont **très contrôlés** que sur de la **parole conversationnelle** : les variations relatives dépendent de la base de données (Pour ALIZE/SpkDet, $Vr_{NIST} = 1.47$ et $Vr_{BREF} = 3.11$).
- **Augmenter le nombre de trames utilisées pour l'apprentissage permet de diminuer les pires performances** (Pour les hommes de BREF, $EER_{30s} = 33\%$ et $EER_{2.5minutes} = 5.3\%$ pour les séries *Max*) mais n'a que peu d'effet sur les meilleures performances (Pour les femmes de BREF, $EER_{30s} = 1.1\%$ et $EER_{2.5minutes} = 0.9\%$ pour les séries *Min*).
- La sensibilité d'un système comme ALIZE/SpkDet est d'un point de EER lorsque les données d'apprentissage sont similaires.

Troisième partie

Localisation, dans le flux de parole, des indices idiosyncratiques en vue d'une prédiction de la performance

Chapitre 5

Le facteur locuteur comme source de variation

Résumé : *Dans ce chapitre, nous présentons la méthode que nous avons adoptée afin de partir à la recherche du locuteur à travers différentes mesures acoustiques identifiées comme idiosyncratiques. Cette étude nous permet d'une part de mieux comprendre où se situent les indices pertinents pour discriminer les locuteurs et expose des critères pour distinguer un fichier donnant lieu à un modèle pertinent de celui conduisant à un modèle peu performant.*

Sommaire

5.1 Questions

Rechercher le locuteur dans un extrait de parole revient à comprendre **comment la variation du signal peut être expliquée par le facteur locuteur en n'oubliant pas que d'autres contraintes influencent à la fois le locuteur et le signal de parole** comme sa culture, sa langue, sa situation sociale ou son état émotionnel. Il est d'autant plus

difficile de séparer les différentes contraintes que les **indices identifiés comme permettant de reconnaître le locuteur** (cf. chapitre 2.3) peuvent également **servir à caractériser d'autres facteurs linguistiques ou para-linguistiques**. La recherche du locuteur dans le signal de parole ne peut pas faire abstraction de ces éléments. Il est tout de même possible de **contrôler certains éléments de variation afin de comprendre le rôle et la place du locuteur**. L'identification des meilleurs indices pour repérer le locuteur dans cette situation donnée est alors envisageable.
Ce cadre étant posé, les trois questions auxquelles nous cherchons à répondre dans ce chapitre sont les suivantes.

- L'influence du locuteur sur l'indice et donc la pertinence de l'indice varie-t-elle en fonction des extraits de parole ?
- L'information sur le locuteur est-elle uniformément répartie dans le signal de parole ?
- Quelle part de variation peut alors être attribuée au locuteur ?

5.2 Corpus

Pour comprendre un phénomène, le choix du corpus est primordial. En effet, ce corpus est l'outil central de notre compréhension de la parole. Dans notre cas, il doit contenir **assez de locuteurs pour rendre compte de la variabilité inter-locuteur**. Mais **chaque locuteur doit avoir été enregistré assez longtemps pour observer la variabilité intra-locuteur**. Les conditions d'enregistrement doivent être bien connues de manière à être contrôlées.

5.2.1 Un corpus contrôlé

Les expériences que nous avons menées pour mesurer l'influence du fichier d'apprentissage sur les performances d'un système de RAL ont été menées sur deux corpus différents l'un de parole conversationnel et multilingue (NIST), l'autre de parole lue en français (BREF). Nous avons dans un premier temps mené nos analyses sur le **corpus BREF**, un corpus beaucoup plus contrôlé pour lequel :

- Les locuteurs sont originaires de la **région parisienne et francophones**.
- Tous les locuteurs ont été enregistrés dans les **mêmes conditions**, la question des conditions d'enregistrement est donc écartée.

– Le contexte de communication est de la **parole lue**, nous disposons d'une transcription orthographiques des phrases prononcées.

Il s'agit donc d'un cadre assez contrôlé pour lequel nous pouvons extraire les indices acoustiques assez facilement puisque nous disposons des transcriptions et de signaux de bonne qualité (enregistrement en chambre sourde par microphone).

Une référence

Le premier de nos objectifs est d'**établir l'influence qu'a le locuteur sur certains éléments du signal de parole déjà identifiés, dans la littérature, comme porteur d'information idiosyncratique**. Pour répondre à cet objectif, nous avons, pour chaque indice étudié, analysé l'influence du locuteur sur les fichiers ayant servi de corpus de test dans BREF. Nous disposons ainsi pour étudier l'influence du locuteur d'un ensemble d'enregistrements important d'environ **11.4 minutes de parole par locuteur**.

Min et *Max*

Le second objectif consiste à **déterminer si les indices sur le locuteur sont plus présents dans les fichiers qui conduisent à une modélisation performante du locuteur (série *Min*) que dans les fichiers qui conduisent à une modélisation moins performante du locuteur (série *Max*).**
Ainsi, nous mesurons l'influence du locuteur sur les indices dans les deux séries en espérant que l'influence du locuteur soit plus importante dans *Min* que dans *Max*. Enfin, nous souhaitons **vérifier si les valeurs des indices étudiés permettent de séparer nos deux séries**.

5.2.2 Premières études sur un corpus conversationnel

Nous avons également travaillé sur les **coefficients cepstraux** qui sont des paramètres porteurs d'information sur le locuteur et utilisés dans tous les systèmes de RAL. Nous sommes alors revenus en partie sur le **corpus NIST** afin de se confronter à d'autres facteurs de variation, présents dans la parole conversationnelle.

5.3 Indices étudiés

Nous avons décidé de nous concentrer sur un certain nombre d'indices acoustiques qui ont été décrits comme étant influencés par le locuteur dans le chapitre 2.3.
Nous commençons par des indices issus de la **phonation** à savoir la **F0 moyenne, le jitter et le shimmer**. Ces indices rendent compte notamment de la taille et de la tension des plis vocaux qui peuvent être dépendants du locuteur. Suivent des indices relatifs à l'**articulation** qui, dans une certaine mesure, est dépendante du locuteur (taille du conduit vocal, stratégie articulatoire). L'articulatoire nous donne aussi accès à la façon dont le locuteur prononce les phonèmes de sa langue. La **distribution des phonèmes** est étudiée pour les enregistrements *Min* et *Max*, afin de vérifier si cela ne pourrait pas expliquer que certains enregistrements soient plus pertinents que d'autres selon, tous les phonèmes n'étant pas tous porteurs de la même quantité d'information sur le locuteur. Nous étudions la réalisation des phonèmes par leurs **centres de gravités**, cette mesure nous indique où se situe l'énergie des fréquences présentes dans chaque phonème. Pour les voyelles orales, les **4 premiers formants** sont mesurés en **statique** (moyenne) et en **dynamique** (coarticulation, suivi de formants), afin de mesurer l'influence du locuteur sur chacun d'entre eux.
La variation des **coefficients cepstraux** due au locuteur est également étudiée sur les deux corpus BREF et NIST, permettant ainsi une étude de **plusieurs langues**.

5.4 Mesures

Les mesures des indices ont été réalisées, après un **alignement forcés de l'ensemble des enregistrements** à l'aide de Speeral (Linares et al., 2007), grâce au logiciel Praat (Boersma et Weenink, 2009) exceptés pour les coefficients cepstraux qui ont été extraits à l'aide de SPro (Gravier, 2011). Notre objectif est de comprendre **dans quelle mesure les indices rendent compte de caractéristiques idiosyncratiques**. En effet, le **meilleur indice est celui qui maximise les différences entre les locuteurs (différences inter-locuteur) et minimise les différences entre les enregistrements un même locuteur (différences intra-locuteur)**.
Nous pouvons définir une variation intra-locuteur, σ_{intra}, comme la moyenne des écarts-types de chaque locuteur. Cette variation est à comparer avec la variation inter-locuteur,σ_{inter}, définie par la variance des moyennes de chaque locuteur.

Ce problème peut être ramené au test F qui compare la variance expliquée par les différences entre les moyennes d'échantillon avec la variance inexpliquée au sein des échantillons.
L'influence du locuteur sur les indices consiste à **calculer la taille de l'effet d'un facteur lors d'une analyse de variance** (ANOVA ou MANOVA), notée η^2 (Levine et Hullett, 2002). Dans ce cadre, le facteur fixé est le locuteur tandis que la variable dépendante est l'indice étudié. La F0, le jitter, le shimmer ou les centres de gravités sont définis par une valeur unique. Dans ce cas, une ANOVA (Wonnacott et Wonnacott, 1991) est utilisée pour analyser la variance. En revanche, la distribution des phonèmes ou bien les valeurs des coefficients cepstraux sont multidimensionnels. Ici, ce sont des MANOVA qui sont utilisées pour l'analyse de la variance.

Nous souhaitons mesurer l'influence du facteur locuteur sur les mesures acoustiques que nous avons effectuées. Nous saurons ainsi si les paramètres que nous avons retenus sont de bons indices pour reconnaître le locuteur. Si le facteur locuteur explique une grande part de variance de nos mesures, cela signifie que les paramètres sont influencés par le locuteur et donc qu'ils sont de bons indices pour reconnaître le locuteur.

Chapitre 6

La phonation

Résumé : *Dans ce chapitre, l'influence du locuteur sur des indices de phonation, à savoir la fréquence fondamentale, le jitter et le shimmer, est étudiée. Si le facteur locuteur permet d'expliquer 18.9% pour les hommes et de 28.4% pour les femmes de la variation de F0, il ne permet pas de distinguer les deux séries Min et Max même si 12% des locuteurs et 7% des locutrices sont différentiables entre Min et Max. Le jitter et le shimmer ne sont pas très influencés par le locuteur (*$2.6\% < \eta^2 < 5.2\%$*).*

Sommaire

6.1 Précisions sur les éléments étudiés

La **fréquence fondamentale** (F0) fait partie des paramètres mis en évidence par (Van Dommelen, 1987) ou par (Kahn et Rossato, 2009) comme permettant aux humains d'identifier des voix. Les informations sur les courbes de F0 sont également utilisées par des systèmes de vérification automatique du locuteur (Scheffer et al., 2011). (Nolan, 2001) insiste également sur l'importance de la qualité de voix comme indice pour reconnaître le locuteur.
Afin de mesurer l'impact du locuteur sur la phonation, nous avons effectué trois mesures acoustiques identifiées comme pertinentes pour reconnaître le locuteur, à savoir **la F0 moyenne, le jitter et le shimmer**, résumées par la figure 6.1.

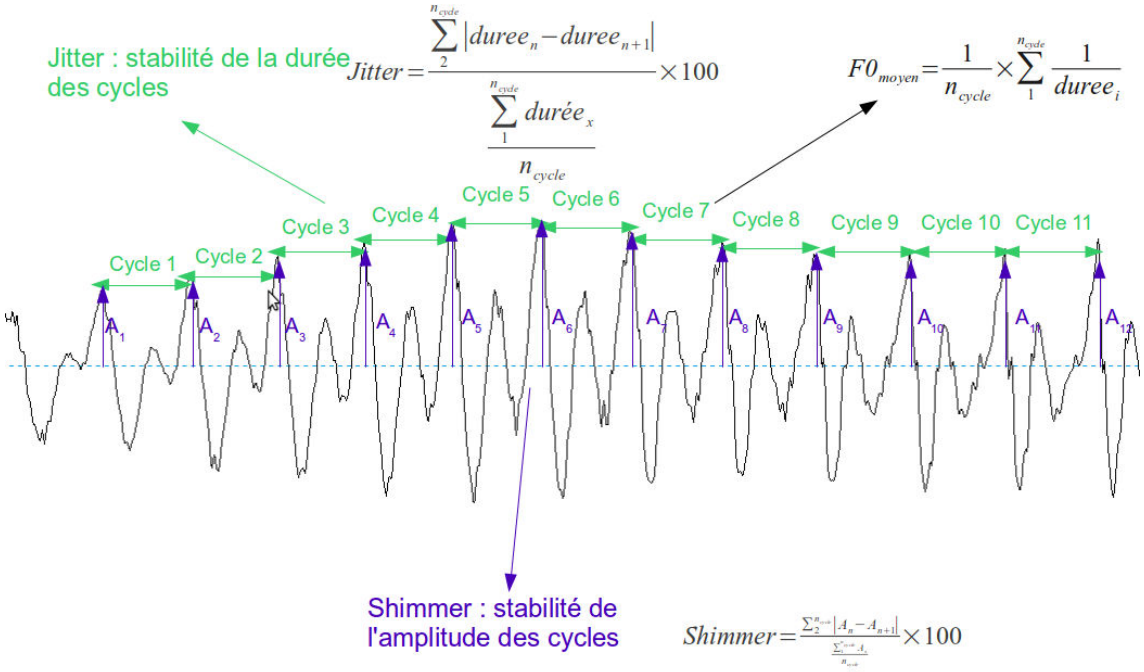


FIGURE 6.1 – *Calcul de la F0, du jitter et du shimmer*

Le **jitter et le shimmer** sont des indices couramment utilisés pour définir **la qualité de la phonation en orthophonie** (Marchal, 2007). Ainsi, même si ces deux indices n'ont été que très peu utilisés pour reconnaître les locuteurs, nous les avons utilisés pour rendre compte en partie de la qualité de voix des locuteurs.
Les paramètres ont été mesurés sur **les voyelles orales** de l'ensemble des fichiers util-

isés en test et sur les fichiers qui constituent les séries *Min* et *Max* sur la base de données BREF soit plus de 21 h de parole. Les mesures ont été effectuées à l'aide de praat (Boersma et Weenink, 2009).

6.2 Influence relative du locuteur sur les indices de la source indépendamment de la voyelle prononcée

6.2.1 Fréquence fondamentale

Variation inter-locuteur vs variation intra-locuteur

Sur l'ensemble des voyelles des fichiers de test, la médiane de la fréquence fondamentale (F0) s'élève à 140 Hz pour les hommes et de 238 Hz pour les femmes. Les distributions de F0 sont bien différentes entre les hommes et les femmes (T-test : $t(334\,093.9) = -709.8$).
La F0 connaît une variation importante en fonction du locuteur ($\sigma = 39\ Hz$ pour les hommes et $\sigma = 41\ Hz$) en fluctuant de 98 Hz à 181 Hz pour les hommes et de 183 Hz à 297 Hz pour les femmes. Les tableaux C.1.1 et C.1.2 mis en annexes indiquent ces valeurs pour, respectivement, les hommes et les femmes. La F0 semble donc a priori un bon indice pour distinguer les locuteurs.
Toutefois, comme l'illustrent les Figures 6.2 et 6.3, les écarts-types sont plus ou moins importants en fonction du locuteur même si les variances peuvent être considérées comme homogènes (Test de Levene : $t(46) = 95.538; p < 2.2 * 10^{-16}$ pour les hommes et $t(63) = 166.76; p < 2.2 * 10^{-16}$ pour les femmes).

Ainsi, pour les hommes, la variation moyenne intra-locuteur est de 33 Hertz tandis que la variation inter-locuteur est de 18 Hertz. Pour les femmes, la variation moyenne intra-locuteur est de 35 Hertz, la variation inter-locuteur est de 23 Hertz.

La variation intra-locuteur est donc plus importante que la différence inter-locuteur aussi bien pour les hommes que pour les femmes.
Cette variation intra-locuteur peut tout à fait être expliquée par le rôle que joue la F0

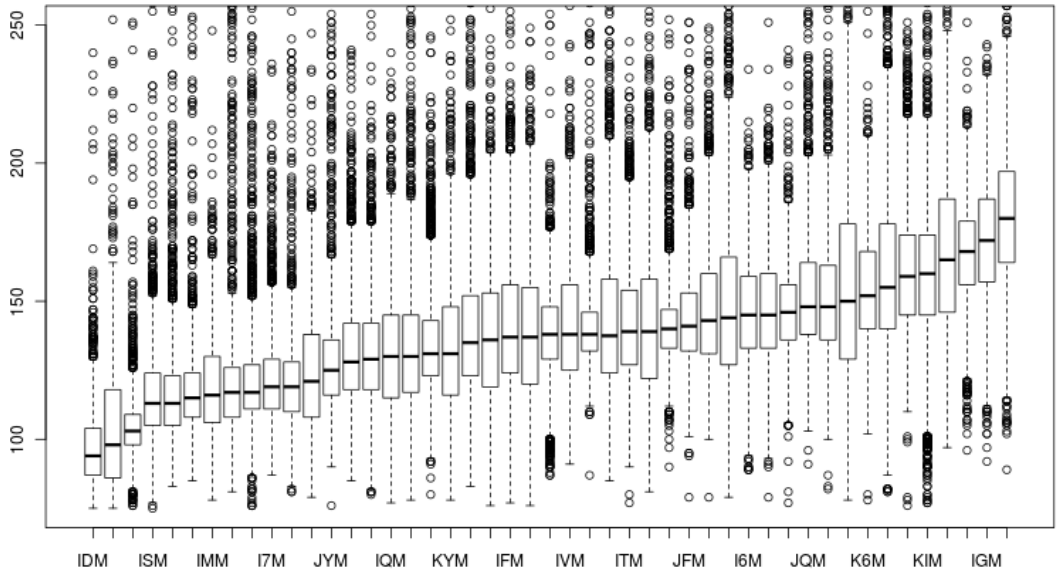


FIGURE 6.2 – *F0 mesurée sur les voyelles de BREF pour les 47 hommes composant le corpus*

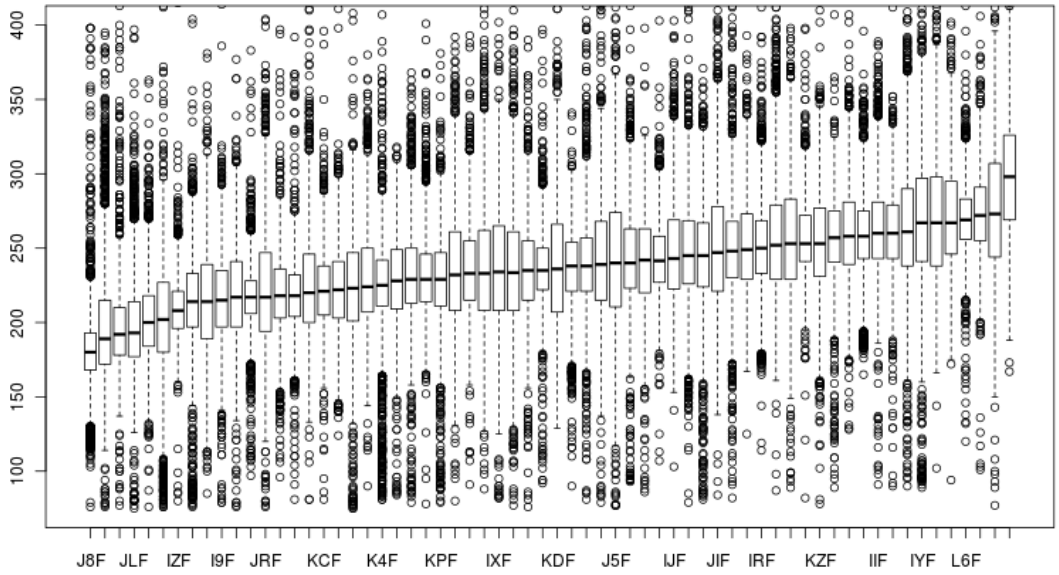


FIGURE 6.3 – *F0 mesurée sur les voyelles de BREF pour les 64 femmes composant le corpus*

au niveau syntaxique pour déterminer une forme interrogative et affirmative. Par exemple, « vraiment » est prononcé par une femme d'abord dans l'énoncé « il est vrai-

ment l'heure de partir » (Figure 6.4) puis dans l'énoncé « es-tu vraiment sur qu'il viendra ? »(Figure 6.5).

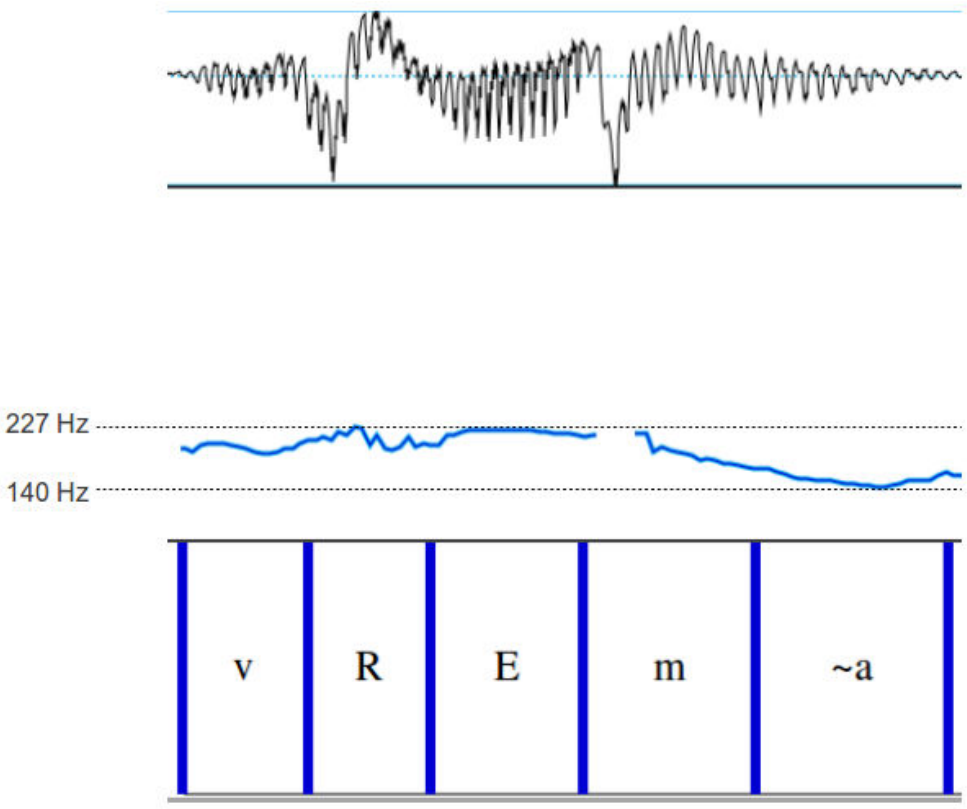


FIGURE 6.4 – *Variation de la F0 pour vraiment dans « Il est vraiment l'heure d'y aller »*

La valeur de F0 peut varier de 140 Hertz à 495 Hertz pour la même locutrice en fonction de la phrase.

Les valeurs de F0, même si elles varient plus chez un même locuteur qu'entre les locuteurs, semblent présenter tout de même une variation inter-locuteur importante qui peut être visualisée sur les figures 6.2 et 6.3.

Effet du locuteur sur les valeurs de F0

Une analyse de la variance par ANOVA est effectuée avec, comme variable dépendante, les valeurs de F0 moyennes pour chaque voyelle et, comme facteur fixé, les locuteurs afin de mesurer l'influence des locuteurs sur les valeurs de F0.

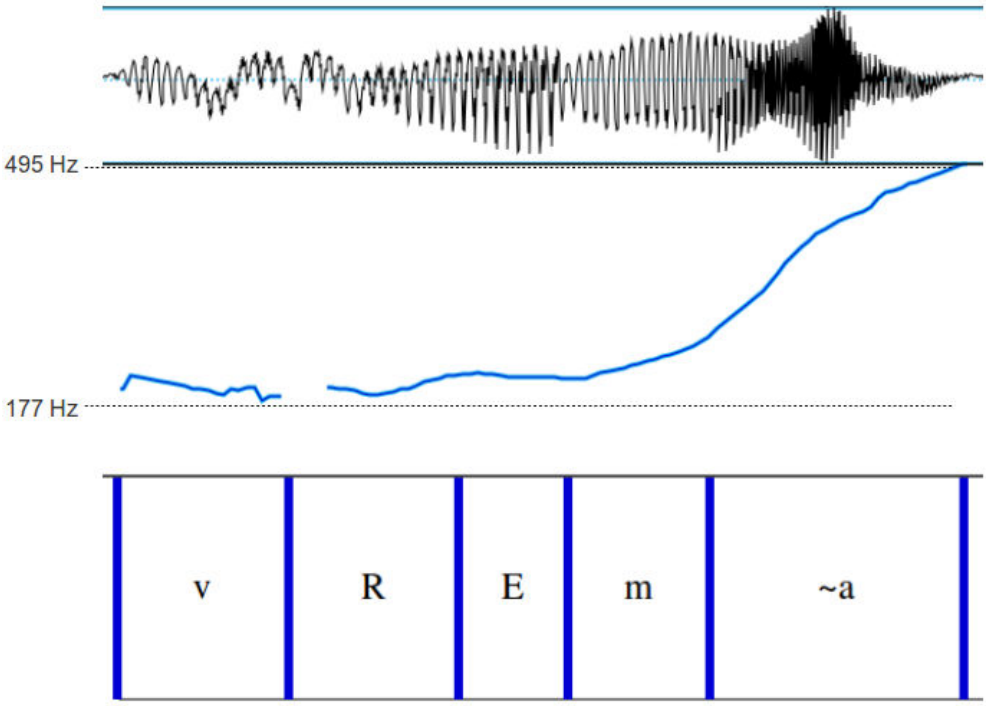


FIGURE 6.5 – *Variation de la F0 pour vraiment dans « Es-tu vraiment sûr qu'il viendra ? »*

Un effet du facteur locuteur sur la distribution des valeurs moyennes de F0 est observé aussi bien pour les hommes ($F(46, 10\ 428) = 763.08; p < 0.001$) que pour les femmes ($F(63, 199552) = 1258.3; p < 0.001$). La part de variance expliquée par le facteur locuteur, mesuré par η^2 est de 18.9% pour les hommes et de 28.4% pour les femmes.
La F0 semble donc être un facteur pertinent pour distinguer les locuteurs, plus particulièrement pour les femmes.

6.2.2 Jitter

Variation inter-locuteur vs variation intra-locuteur

La valeur médiane des jitters est de 1.4% pour les hommes et de 1.1% pour les femmes. Cette valeur médiane est légèrement au-dessus de la valeur de référence comme seuil de pathologie (1.04%). Toutefois dans (Yu et al., 2001) les voix pathologiques étudiées ont des valeurs de jitter bien au dessus de 2.07%. En fonction des locuteurs, la valeur médiane de jitter varie de 1.1% à 2.3% pour les hommes et de 0.8% à 1.7% pour les femmes. La variation inter-locuteur est de 0.24% pour les hommes et de 0.16% pour les femmes tandis que la variation intra-locuteur est respectivement de 1.7% et de 1.4%.

La variation intra-locuteur est donc plus grande que la variation inter-locuteur pour les valeurs de jitter. **Nous pouvons donc nous interroger sur la pertinence du jitter comme indice de différenciation des locuteurs.**

Effet du locuteur sur les valeurs de jitter

Afin de mesurer la part de variation des valeurs de jitter expliquée par le facteur locuteur, une ANOVA à un facteur avec comme variable dépendante les mesures de jitter pour l'ensemble des voyelles et comme facteur le locuteur a été effectuée. Si le facteur locuteur semble explicatif des distributions obtenues ($F(46, 141\,663) = 82.177; p < 0.001$ pour les hommes et $F(63, 227\,526) = 91.823; p < 0.001$ pour les femmes), il n'explique que 2.6% de variance totale pour les hommes et 2.5% de la variance totale pour les femmes. **Le jitter semble donc être un facteur qui explique seulement une faible proportion de variation due aux locuteurs de la base de données BREF.**

6.2.3 Shimmers

Variation inter-locuteur vs variation intra-locuteur

La valeur médiane de shimmer est de 9.2% pour les hommes et de 7.2% pour les femmes. Ces mesures sont assez éloignées du seuil de pathologie qui est de 3.8% (Teston, 2004). Il est toutefois à noter que ce seuil de pathologie est établi sur des voyelles tenues et non pas en parole continue, comme c'est le cas dans nos travaux.
Les valeurs de shimmer varient en fonction des locuteurs de 7.3% à 12.8% pour les hommes et de 5.5% à 10.4% pour les femmes. La variation inter-locuteur est de 1.27% pour les hommes et de 1.10% pour les femmes à comparer avec la variation intra-locuteur de 7.16% pour les hommes et de 5.42% pour les femmes.
La variation intra-locuteur est donc également plus grande que la variation inter-locuteur pour les valeurs de shimmer.

Effet du locuteur sur les valeurs de shimmer

Une ANOVA à un facteur avec comme variable dépendante les mesures de shimmer pour l'ensemble des voyelles et comme facteur le locuteur a été effectuée.

Le facteur locuteur explique des distributions obtenues 3.7% de variance totale pour les hommes ($F(46,141\ 663) = 119.80; p < 0.001$) et 5.2% de la variance totale pour les femmes ($F(63,227\ 526) = 198.68; p < 0.001$). **Le shimmer semble donc être un facteur peu pertinent**.

Parmi les trois mesures que nous avons expérimentées pour décrire la phonation, la fréquence fondamentale semble être l'indice le plus influencé par le locuteur, le facteur locuteur n'expliquant que peu de variance des valeurs de jitter et de shimmer.
Il est toutefois à noter que les mesures de jitter et de shimmer sont généralement réalisées sur la voyelle /a/ **tenue. (Brockmann et al., 2011) montre que la voyelle utilisée pour les mesures a une influence sur les valeurs de jitter et de shimmer obtenues. IL peut donc être important de prendre en compte le timbre de la voyelle pour effectuer nos analyses.**

6.3 Influence relative du locuteur sur les indices de la source en fonction de la voyelle prononcée

6.3.1 Importance relative du timbre de la voyelle et du locuteur

Nous pouvons mesurer à l'aide d'ANOVA à 2 facteurs **l'impact des deux facteurs locuteur et timbre de la voyelle sur les mesures de F0, jitter et shimmer**.

Fréquence Fondamentale

Le timbre de la voyelle explique 0.8% de la variance totale des valeurs de F0 pour les hommes et 1.6% de la variance pour les femmes. Comme nous l'avons vu précédemment, le facteur locuteur explique quant à lui entre 18.9% et 28.4% de la variance. Ainsi, au vu de ce ratio, **le timbre de la voyelle a une influence très relative sur la F0 par rapport à l'influence du locuteur**.

Jitter

2.7% et 3.6% de la variance des valeurs de jitter sont expliquées par le timbre de la voyelle pour respectivement les hommes et les femmes. L'influence du locuteur qui est de 2.6% et de 2.7% sur les **valeurs de jitter est équivalente à celle du timbre de la voyelle. Une analyse en fonction des voyelles semble donc nécessaire.**

Shimmer

Le timbre de la voyelle explique 2.8% et 2.9% de la variance des valeurs de shimmer pour les hommes et les femmes respectivement. **L'influence du locuteur sur les valeurs de shimmer est équivalente à ces résultats. Ici aussi une analyse en fonction du timbre de la voyelle peut s'avérer pertinent.**

6.3.2 Effet du locuteur sur les information de source en fonction du timbre de la voyelle

Pour mesurer l'impact du locuteur en fonction de la voyelle sur les mesures effectuées pour décrire la source, des ANOVA à un facteur ont été effectuées séparément pour les 10 voyelles orales du français à savoir /i/, /y/, /u/, /e/, /ø/, /o/, /ɛ/, /œ/, /ɔ/, /a/. L'ensemble des résultats obtenus (valeurs moyennes, σ_inter, σ_intra et η^2) par voyelle est résumé par le tableau C.4 en annexe, seuls les η^2 sont retranscrits dans le tableau 6.1 puisqu'il s'agit de la mesure la plus parlante.

Fréquence fondamentale

Un effet du locuteur sur la distribution des valeurs de F0 est observé pour chacune des voyelles ($p < 0.001$). Les η^2 varient de 15.9% à 29.4% pour les hommes et de 26.1% à 34.3% pour les femmes selon la voyelle. **Si des légères différences sont observées en fonction du timbre de la voyelle, il est difficile de trouver un critère qui permette d'expliquer les différences observées**. Les tendances générales (sans faire de différence entre timbre de la voyelle) se retrouvent ici : les valeurs de F0 semblent être un critère les locuteurs plus important pour les femmes que pour les hommes.

		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
H	F0	16.5%	21.4%	29.4%	25.9%	15.9%	21.0%	22.1%	24.6%	21.2%	16.3%
	Jitter	3.8%	3.3%	4.0%	3.7%	3.6%	2.6%	5.5%	4.6%	2.8%	3.0%
	Shim.	3.9%	3.8%	5.2%	6.0%	5.2%	6.3%	5.9%	4.3%	5.3%	5.4%
F	F0	31.7%	30.1%	28.7%	28.6%	26.1%	27.6%	28.9%	34.3%	29.4%	27.9%
	Jitter	3.5%	3.3%	4.1%	3.1%	3.3%	3.3%	6.9%	5.1%	4.3%	3.0%
	Shim.	8.8%	7.6%	5.4%	8.7%	6.5%	6.0%	10.7%	7.5%	5.5%	3.6%

TABLE *6.1* – η^2 *du facteur locuteur pour les indices de description de la source pour les voyelles du français*

Jitter

Un effet du locuteur sur la distribution des valeurs de jitter est observé pour chacune des voyelles ($p < 0.001$). Les η^2 du jitter varient de 2.6% à 5.5% pour les hommes et de 3.0% à 6.9% pour les femmes. Aussi bien pour les hommes que pour les femmes, c'est le phonème /œ/ qui est le plus influencé par le locuteur. Viennent ensuite les phonèmes les plus postérieurs.

Shimmer

Un effet du locuteur sur la distribution des valeurs de shimmer est observé pour chacune des voyelles ($p < 0.001$). Les η^2 du shimmer varient de 3.8% à 6.3% pour les hommes et de 3.6% à 10.7% pour les femmes. Il est difficile d'extraire une tendance en fonction de la catégorie phonémique de la voyelle car les tendance sont opposées entre hommes et femmes pour les shimmers.

Nous avons montré sur un grand nombre d'enregistrements que le locuteur influence de façon plus importante les valeurs de F0 que celles de jitter et de shimmer. Ceci peut s'expliquer par le fait que l'ensemble des locuteurs sont considérés comme n'ayant pas de pathologie et que les mesures de jitter et shimmer sont au départ un bon indice de dysphonie. Reste à savoir si la part de variation expliquée par le facteur locuteur varie d'un enregistrement à l'autre ce qui pourrait expliquer les différences de performance observées au chapitre II. La question suivante est donc de savoir, étant donné la grande variabilité intra-locuteur observée, la série *Min* permet de mieux distinguer les locuteurs que la série *Max*.

6.4 Pertinence des indices pour prédire la qualité d'un enregistrement

Les extraits de parole de la série *Min* permettent une meilleure modélisation du locuteur que ceux utilisés dans la série *Max* pour modéliser les mêmes locuteurs. **Une de nos hypothèses explicatives est que la part de variation due aux locuteurs dans la série *Min* est plus importante que pour *Max*.**
Afin de mesurer la part de variation expliquée par le facteur locuteur dans nos séries de fichiers, des ANOVA à un facteur ont été réalisées sur les mesures issues des fichiers de la liste *Min* d'une part et de *Max* d'autre part. L'ensemble des résultats est présenté dans le tableau 6.2

6.4.1 Effet du locuteur sur les valeurs de F0

La première variable dépendante observée est la F0 moyenne calculée indépendamment du timbre de la voyelle.
Un effet du facteur locuteur est observé dans les deux séries pour les hommes (*Min* : $F(46,7\ 423); p < 2.2e^{-16}$;*Max* : $F(46,7\ 249); p < 2.2e^{-16}$) et pour les femmes(*Min* : $F(63,7\ 423); p < 2.2e^{-16}$;*Max* : $F(63,7\ 249); p < 2.2e^{-16}$). La part de variation expliquée par le facteur locuteur est légèrement plus élevée pour *Min* que pour *Max* ($\eta^2_{Min} = 21.5\% vs \eta^2_{Max} = 20.4\%$) pour les hommes. En revanche, pour les femmes c'est la série *Max* qui obtient η^2 le plus élevé (($\eta^2_{Min} = 25.5\% vs \eta^2_{Max} = 27.6\%$)).
Contrairement à nos attentes, la part de variation expliquée par le locuteur est plus grande pour *Min* que pour *Max* chez les hommes mais ce n'est pas le cas pour les femmes.

6.4.2 Effet du locuteur sur les valeurs de jitters

La seconde variable dépendante étudiée est le jitter calculé sur chacune des voyelles séparément.
Un effet du facteur locuteur est observé dans les deux séries uniquement pour les voyelles /a/ et /e/ pour les hommes (/a/ : $Min : F(46,1\ 348); p < 0.001, Max : F(46,1\ 377); p < 0.001$; /e/ : $Min : F(46,1\ 026); p < 0.001, Max : F(46,958); p < 0.001$).

Pour les femmes, l'effet du locuteur n'est significatif que pour /a/, et /i/ (/a/ : *Min* : $F(63, 1\,848); p < 0.001$, *Max* : $F(63, 1\,813); p < 0.001$; /i/ : *Min* : $F(63, 1\,354); p < 0.001$, *Max* : $F(63, 1\,368); p < 0.01$).
L'effet du locuteur sur le jitter, quand il existe, est plus élevé pour les mesures issues des fichiers de la liste *Min* que pour celles issues des fichiers de la liste *Max*. Il reste toutefois bien peu élevé par rapport à l'effet de la F0 et localisé sur très peu de timbres vocaliques.

6.4.3 Effet du locuteur sur les valeurs de shimmers

La troisième variable dépendante est le shimmer calculé sur chacune des voyelles séparément.
Un effet du locuteur est observé pour toutes les voyelles exceptées /o/, /œ/ et /ɔ/ pour les hommes et /œ/, /u/ et /y/ pour les femmes. Les parts de variance expliquée par le locuteur sont plus importantes dans *Min* et dans *Max* que dans les valeurs de référence que nous avons établies en 6.3.2. La part de variance de shimmer expliquée par le locuteur est à peu près la même entre *Min* et *Max*. Il est difficile d'utiliser ce critère pour séparer nos deux séries.

Le facteur locuteur n'explique pas exactement la même proportion de variance entre *Min* et *Max* selon la variable dépendante utilisée. Pour la F0, si pour les hommes une part légèrement plus importante de variation est expliquée dans la série *Min*, ce n'est pas le cas pour les femmes. Pour les valeurs de jitter, le locuteur semble expliquer une part de variance plus importante pour *Min* que pour *Max*. Il est toutefois difficile de savoir si les différences observées de part de variance expliquée sont significatives ou non. Les paramètres pourraient avoir joué un rôle important pour certains locuteurs et pas pour d'autres. Il pourrait être intéressant de savoir le nombre de locuteurs pour lesquels les paramètres de source sont différents entre *Min* et *Max*.

			/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
F0	H	Effet	Calculé sur toutes les voyelles indistinctement ($p < 2.2e^{-16}$)									
		Min	21.5%									
		Max	20.4%									
	F	Effet	Calculé sur toutes les voyelles indistinctement ($p < 2.2e^{-16}$)									
		Min	25.5%									
		Max	27.6%									
Jitter	H	Effet	***			***						
		Min	9.4%	n.s	n.s	8.5%	n.s	n.s	n.s	n.s	n.s	n.s
		Max	8.2%			6.5%						
	F	Effet	***					***/**				
		Min	5.4%	n.s	n.s	n.s	n.s	9.4%	n.s	n.s	n.s	n.s
		Max	5.9%					7.0%				
Shim.	H	Effet	***/**	**		***	***	***			**	
		Min	7.5%	8.2%	n.s	12.5%	10.4%	11.1%	n.s	n.s	25.9%	n.s
		Max	5.5%	8.8%		10.3%	10.4%	9.3%			28.2%	
	F	Effet	***	***	*/**	***	***	***		**/***		
		Min	11.7%	13.8%	22.0%	12.3%	8.7%	11.1%	n.s	14.9%	n.s	n.s
		Max	11.2%	15.9%	21.3%	13.4%	13.8%	10.6%		18.9%		

TABLE 6.2 – η^2 *du facteur locuteur pour Min et pour Max sur les paramètres de la source.*

6.4.4 Différencier *Min* et *Max*

Méthodes

Pour évaluer **le nombre de locuteurs dont les valeurs relatives à la phonation changent entre *Min* et *Max***, nous avons comptabilisé le nombre de locuteurs pour lesquels la valeur de F0 diffère significativement entre *Min* et *Max*. Ainsi, des ANOVA à un facteur ont été réalisées par locuteur. Le facteur dépendant est l'appartenance à la série *Min* ou à la série *Max*, appelée plus loin Performance, et les paramètres dépendants sont la F0, le jitter et le shimmer. Pour la F0, l'effet de la voyelle est négligée, une seule ANOVA est donc réalisée. En revanche pour le jitter et le shimmer, une ANOVA doit être réalisée voyelle par voyelle.

Fréquence fondamentale

L'effet de la Performance est significatif ($p < 0.01$) pour 6 des 47 hommes, soit un peu plus de 12% et pour 5 des 64 femmes, soit 7.8%. Ainsi, la différence de Performance

entre *Min* et *Max* peut être expliquée pour quelques locuteurs par une différence de F0. Cette différence n'est toutefois pas générale puisque pour la grande majorité des locuteur, il n'existe pas de différence significative entre les distributions de F0 entre la série *Min* et *Max*. D'ailleurs, en analysant par t-test apparié l'effet de la Performance sur les valeurs médianes de F0 de chaque locuteur, aucune différence significative n'est observée aussi bien pour les hommes ($t(46) = -0.7955; p = 0.4304$) que pour les femmes ($t(63) = 0.7781; p = 0.4394$).
De même, si nous conservons l'ensemble des données et réalisons des ANOVA à mesures répétées avec comme facteur dépendant la Performance et comme variable les valeurs de F0 en tenant compte du facteur locuteur, un léger effet de la Performance est observé sur les distribution de F0 pour les hommes ($F(1, 14\,718) = 0.1563; p < 0.01$). Toutefois, nous observons également que $\eta^2 = 0.03\%$, le facteur Performance n'a donc quasi aucun effet sur les valeurs de F0. Par ailleurs, aucun effet de la Performance n'est observé sur les distributions de F0 pour les femmes($F(1, 18\,987) = 0.1563; p = 0.6926$).
Les valeurs de F0 à elles seules ne montrent pas de différence permettant d'expliquer les différences entre les signaux de parole qui conduisent à une modélisation performante et ceux de la série *Max*.

Jitter et shimmer

Les mêmes questions se posent pour les valeurs de jitter et de shimmer. Dans ce cas, les analyses doivent être réalisées voyelles par voyelles. Nous n'avons pas assez de données pour mesurer pour chaque locuteur et pour chaque voyelle l'effet des valeurs de jitter et de shimmer. Nous pouvons cependant tester s'il existe une différence entre les deux séries grâce à des ANOVA à mesures répétées. Le facteur dépendant est la Performance, la variable étudiée est le jitter en tenant compte du locuteur.
Aucun effet de la Performance n'est observé pour les valeurs de jitter des femmes quelque soit la voyelle ($0.1298 < p < 0.9295$). Pour les hommes, à l'exception de /a/, aucun effet de la Performance n'est observé ($0.1645 < p < 0.8964$). Pour /a/, si l'effet est légèrement significatif ($F(1, 2\,771) = 5.34; p < 0.05$), $\eta^2 = 0.19\%$, ce qui montre que l'effet du locuteur n'est vraiment pas important. La part de variation expliquée par le facteur Performance est donc extrêmement faible.
Aucun effet de la Performance n'est observé pour les valeurs de shimmer des hommes quelque soit la voyelle ($0.1348 < p < 0.9796$). Pour les femmes, à l'exception de /a/, aucun effet de la Performance n'est observé ($0.2575 < p < 0.9806$). Pour /a/, si l'effet

est légèrement significatif ($F(1, 3\,724) = 4.67; p < 0.05$), $\eta^2 = 0.001\%$. La part de variation expliquée par le facteur Performance est donc extrêmement faible.
Le jitter et le shimmer ne montrent pas de différence significative en fonction de la Performance.

Synthèse du chapitre

Si les valeurs de **F0 semblent être influencées par le locuteur** (η^2 autour de 20%), elles ne suffisent pas à expliquer les différences de performance observées entre les séries *Min* et *Max*.
La qualité de phonation, représentée par les mesures de jitter et de shimmer ne semblent pas être des indices très pertinents pour différencier les locuteurs de la base de données BREF dans laquelle les extraits enregistrés sont constitués de phrases extraites du *Monde* lue dans une chambre sourde.
Les paramètres acoustiques de la source mesurés ici ne permettent pas de différencier nos deux séries.

Chapitre 7

L'articulation

Résumé : *Dans ce chapitre, nous étudions comment le locuteur influence certains indices généralement attribués à des différences d'articulation. Nous observons tout d'abord si la répartition des phonèmes au sein des deux séries Min et Max pourrait expliquer les différences de performances. Très peu de différences sont observées. Nous étudions ensuite l'influence du locuteur sur les centres de gravité des phonèmes. Nous montrons que les centres de gravités des fricatives et des nasales sont plus sensibles au locuteur que ceux des plosives. Pour les voyelles orales, plus la voyelle est haute et fermée, moins l'effet du locuteur est élevé sur les valeurs des centres de gravité. En étudiant les formants de ces voyelles, nous montrons que les voyelles hautes sont moins dépendantes du locuteur que les autres voyelles. Nous montrons également que F3 et F4 sont plus sensibles au locuteur que F1 et F2. Ces valeurs ne nous permettent cependant pas de différencier les deux séries Min et Max. Les indices de co-articulation semblent moins pertinents que les mesures moyennes sur les formants. Toutes ces analyses sont évidemment effectuées phonème par phonème car l'effet de la catégorie phonétique est très important sur toutes ces mesures.*

Sommaire

7.1 Répartition des phonèmes dans les séries *Min* et *Max*

En 1995, (Magrin-Chagnolleau et al., 1995) ont mis en évidence des **fluctuations de performance en fonction du contenu phonétique** du fichier d'apprentissage pour des systèmes automatiques. En ne construisant leurs modèles qu'à partir de certains segments, ils ont montré que les nasales et certaines voyelles permettaient d'obtenir de meilleures performances que des modèles construits à partir de plosives ou de fricatives. Cette différence de performance en fonction des phonèmes utilisés a également été mis en évidence pour l'identification par écoute humaine par (Amino et al., 2006). Les **nasales** jouaient alors un rôle prépondérant dans la reconnaissance. **Le contenu phonétique jouerait-il encore un rôle important pour les systèmes actuels ? Certains segments seraient-ils plus porteurs d'information sur le locuteur ?** Une présence plus importante de ces segments spécifiques dans les fichiers composant la série *Min* que dans la série *Max* pourrait être un élément d'explication des différences de performance observées.

7.1.1 Méthode

Nous **comptabilisons le nombre de trames de chaque phonème** présent dans les fichiers *Max*, les fichiers *Min* et l'ensemble des fichiers choisis aléatoirement lors des expériences décrites en partie II de manière à vérifier l'équilibre de répartition phonologique de BREF. La **durée de chaque phonème, exprimée en nombre de trames,** est considérée comme paramètre pour estimer si certains phonèmes sont plus présents que d'autres

dans les séries.

7.1.2 Distribution globale

Pour savoir si globalement les distributions de phonèmes sont différentes, une MANOVA à mesures répétées est effectuée avec comme facteur fixé la Performance et comme variable dépendante la répartition en nombre de trames de chaque catégorie phonétique. Aucune différence significative n'a été constatée entre les deux séries aussi bien pour les hommes ($F(33,14) = 2.23, p = 0.056$) que pour les femmes ($F(33,31) = 1.40, p = 0.175$).
Il n'existe donc pas de différence globale entre les deux distributions, a priori ce n'est pas la présence de certains phonèmes qui explique les différence de performance.

7.1.3 Étude de chaque phonème

Phonèmes

En effectuant des ANOVA à mesure répétée à un facteur pour chacun des phonèmes, nous évaluons les différences de quantité de chaque phonème entre *Min* et *Max*.
Pour les hommes, il y a significativement plus de /k/ dans la série *Max* que dans le série *Min* ($F(1,46) = 7.58, p = 0.008$) et légèrement plus de /ʁ/ dans *Min* que dans *Max* ($F(1,46) = 5.83, p = 0.020$). Les autres phonèmes ne présentent pas de différences significatives. Pour les femmes, seul /t/ est significativement plus présent dans *Min* que dans *Max* ($F(1,63) = 4.84, p = 0.032$).
Les résultats obtenus confirment l'analyse faite sur la distribution globale. Il n'existe que très peu de différences entre les deux séries. La quantité de phonèmes, exprimée en nombre de trames, ne peut à elle seule expliquer les différences de performance observées entre *Min* et *Max*.

Mode articulatoire

Nous avons par la suite **regroupé les phonèmes selon leur mode articulatoire** (plosives non voisées, plosives voisées, fricatives non-voisées, fricatives voisées, latérales et approximantes, consonnes nasales, voyelles hautes, voyelles ouvertes et non parole)

afin de savoir si certaines catégories étaient plus présentes que d'autres.
L'évaluation des différences de la distribution de ces catégories entre les séries par MANOVA à mesures répétées, ne montre pas de différence significative aussi bien pour les hommes ($F(10,37) = 1.59, p = 0.149$) que pour les femmes ($F(10,54) = 1.92, p = 0.062$). Les comparaisons par classes de phonèmes indiquent que seules les consonnes nasales sont significativement plus présentes dans la série *Max* pour les femmes ($F(1,63) = 4.38, p = 0.040$). Les Figures 7.1 et 7.2 illustrent la distribution des trames en fonction des différentes catégories phonétiques respectivement pour les hommes et les femmes. Elles représentent également la quantité de trames observée pour chaque phonème.

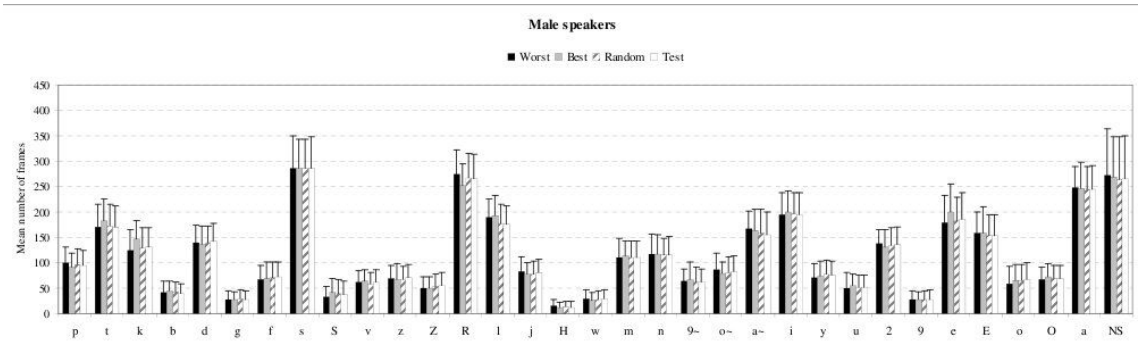


FIGURE 7.1 – *Distributions des catégories phonétiques pour les séries Min et Max des hommes*

Même en regroupant les phonèmes par catégorie, très peu de différence entre les deux séries est observé.

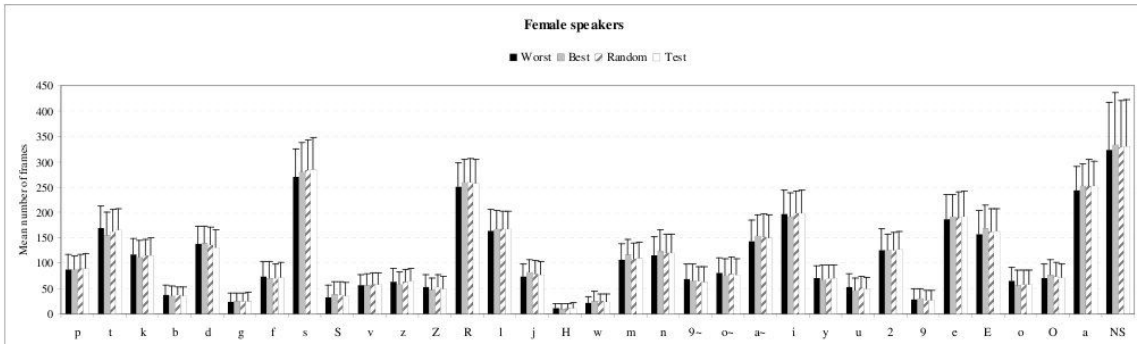


FIGURE 7.2 – *Distributions des catégories phonétiques pour les séries Min et Max des femmes*

Les séries ne présentent donc quasi aucune différence en terme de distribution de phonèmes. Le peu de différence observé ne pourrait, à lui seul, expliquer les

différences de performance mises en évidence dans le chapitre 4. Les qualités intrinsèques des phonèmes pourraient peut-être expliquer les différences de performances. Nous savons par exemple que le triangle vocalique est influencé par la taille du conduit vocal ou que les bruits de friction, mesurés par les centres de gravité, dépendent de la forme du palais du locuteur. Ainsi, au delà du nombre de phonèmes, ce sont les qualités des segments qui seraient spécifiques au locuteur.

7.2 Influence du locuteur sur les centres de gravité des phonèmes

7.2.1 Importance relative du locuteur sur les valeurs de centre de gravité

Il s'agit en premier lieu d'**établir l'importance relative du locuteur pour les valeurs de centre de gravité**. Nous savons déjà que le phonème sur lequel est mesuré le centre de gravité a une importance, il permet notamment de distinguer les fricatives entre elles (Gordon et al., 2002).

Effet de la catégorie phonémique et du locuteur

Pour **mesurer la part de variance des valeurs de centre de gravité expliquée par le facteur phonème et par le facteur locuteur**, une ANOVA à deux facteurs est réalisée. Les deux facteurs sont l'étiquette phonémique et le locuteur, la variable dépendante est la valeur du centre de gravité. Ces mesures sont effectuées sur l'ensemble des fichiers tests de BREF. Les occurrences de chaque phonème pour les hommes et pour les femmes sont consultables en annexe D.1.
Des différences significatives sur les valeurs de centres de gravité sont observées aussi bien pour le facteur phonème (Hommes : $F(33, 455\,101) = 16\,120.3; p < 0.001$, Femmes : $F(33, 662\,532) = 16\,366.18; p < 0.001$) que pour le facteur locuteur (Hommes : $F(46, 455\,101) = 678.32; p < 0.001$, Femmes : $F(63, 662\,532) = 587.31; p < 0.001$). Les Figures 7.3 et 7.4 illustrent la répartition des valeurs de centres de gravité à l'aide de boîte à moustaches, pour chaque phonèmes pour les hommes et pour les femmes respectivement.

Les valeurs médianes pour chaque phonème sont résumées en annexe D.2.

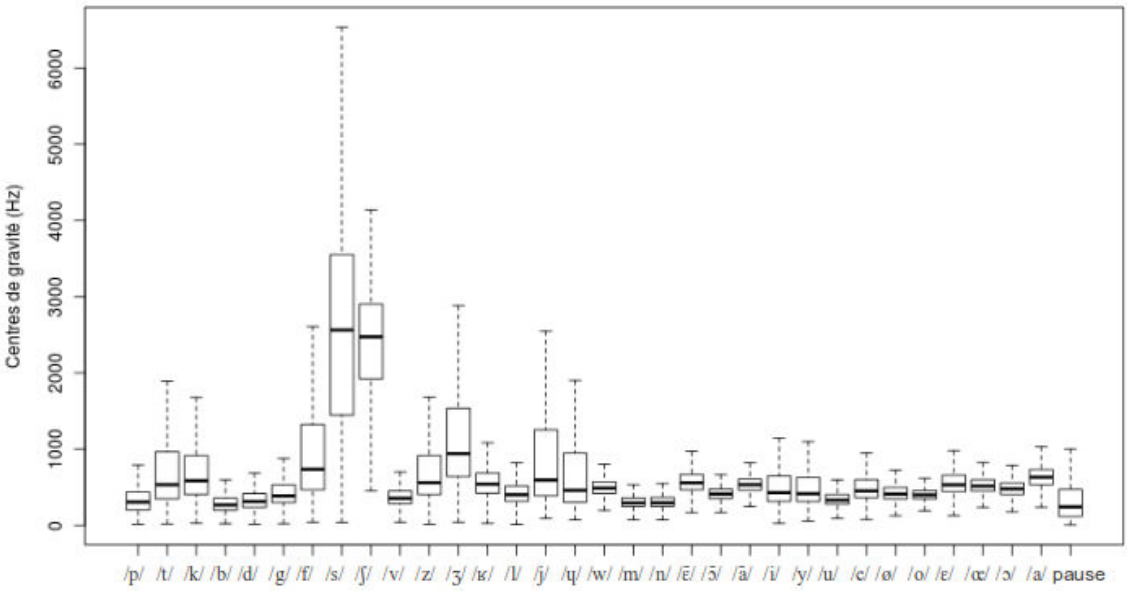


FIGURE 7.3 – *Répartition des valeurs de centre de gravité selon le phonème pour les hommes*

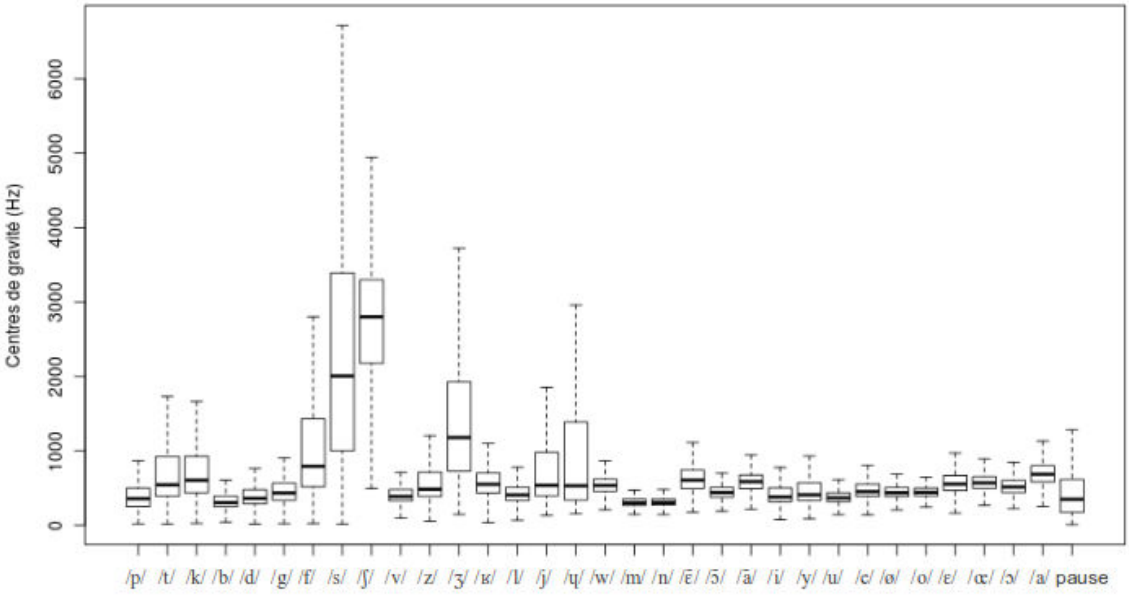


FIGURE 7.4 – *Répartition des valeurs de centre de gravité selon le phonème pour les femmes*

Les fricatives non-voisées montrent clairement les valeurs les plus élevées du centre de gravité. Pour les voyelles, plus la position de la langue est postérieure, plus la valeur du centre de gravité est basse. Pour les consonnes plosives, les labiales sont plus graves que les dentales, elles mêmes plus graves que les dorsales. Le même phénomène est observé pour les fricatives voisées. Pour les fricatives non-voisées, /f/ est le phonème

le plus grave, suivi de /ʃ/ et de /s/.
L'effet du phonème sur les valeurs de centre de gravité est beaucoup plus important que celui du locuteur aussi bien pour les hommes ($\eta^2_{Phoneme} = 52.2\%$ vs $\eta^2_{Locuteur} = 3.1\%$) que pour les femmes ($\eta^2_{Phoneme} = 43.6\%$ vs $\eta^2_{Locuteur} = 3.0\%$). **Une analyse phonème par phonème semble donc particulièrement intéressante pour mesurer l'influence du locuteur sur les centres de gravité.**

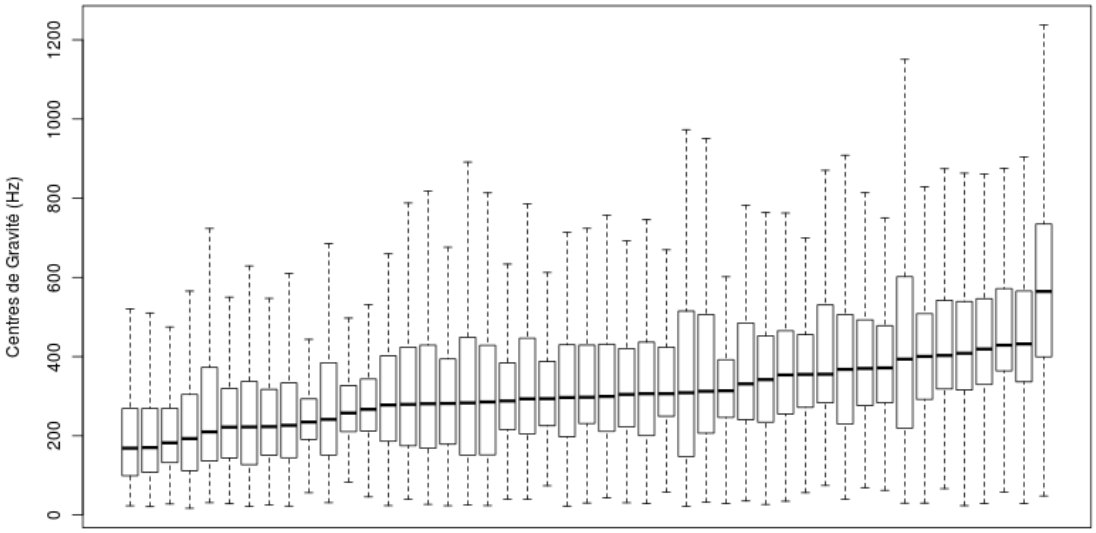
Analyse par phonème de l'effet du locuteur

Pour connaître **l'influence du locuteur sur les centres de gravité**, des ANOVA à un facteur sont réalisées pour chaque phonème. Le facteur est le locuteur tandis que la variable dépendante est la valeur du centre de gravité. Le tableau D.3 en annexe présente les η^2 obtenus pour chaque phonème.
Pour les hommes, la part de variance des valeurs de centre de gravité expliquée par

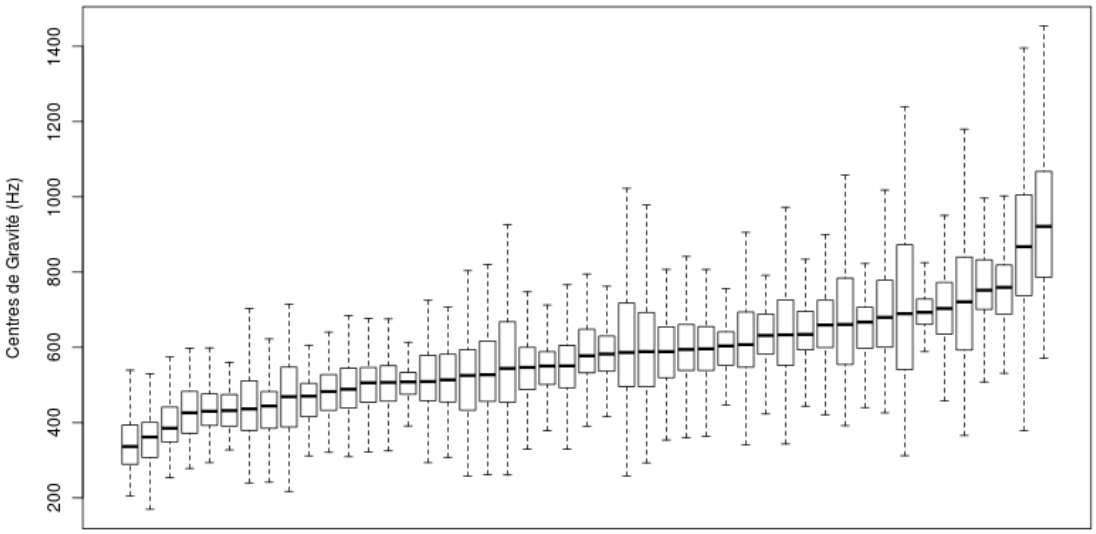
Catégories	η^2 pour les hommes	η^2 pour les femmes
Plosives non voisées	10.4%	8.1%
Plosives voisées	16.6%	11.6%
Fricatives non voisées	23.5%	18.1%
Fricatives voisées	23.9%	17.9%
Consonnes nasales	39.7%	26.2%
Approximantes et glides	15.3%	13.0%
Voyelles fermées	19.7%	16.8%
Voyelles mi-fermées	38.3%	34.0%
Voyelles mi-ouvertes	45.4%	48.7%
Voyelle ouverte	40.7%	51.9%
Voyelles nasales	48.0%	51.5%
Pauses	9.8%	8.6%

TABLE 7.1 – η^2 *du facteur locuteur sur les centres de gravité*

le locuteur varie de 7.9% pour /p/ à 54.6% pour /œ/. Pour les femmes, les valeurs de η^2 varient de 7.2% pour /w/ à 63% pour /ɛ̃/. Certains phonèmes sont plus dépendants du locuteur que d'autres. A titre d'exemple, le tableau 7.2, composé de deux figures, illustre la différence de répartition des centres de gravité pour les phonèmes /p/ et /ɛ̃/ pour les 47 hommes. Ces phonèmes ont une importante différence de η^2 ($\eta^2_p = 7.9\%$ vs $\eta^2_{\tilde{\varepsilon}} = 54.6\%$).



/p/



/ɛ̃/

TABLE 7.2 – *Répartition des valeurs de centres de gravité pour les phonèmes* /p/ *et* /ɛ̃/ *pour les 47 locuteurs hommes : des différences inter-locuteur beaucoup plus importantes et différences intra-locuteur réduites pour la voyelle nasale.*

En terme de catégories, les voyelles nasales (H : $\eta^2 = 48.0\%$ et F : $\eta^2 = 51.5\%$) sont les plus sensibles au locuteur ainsi que le montre les résultats moyens par catégories du tableau 7.1. Suivent les consonnes nasales et les voyelles orales puis les fricatives aussi bien non-voisées que voisées. Les plosives non-voisées et voisées ainsi que les approximantes semblent être les moins sensibles aux locuteurs avec des η^2 inférieurs à 15%. Fort heureusement, les pauses (H : $\eta^2 = 9.8\%$ et F : $\eta^2 = 8.6\%$) ne sont que peu influencées par le locuteur.
En ce qui concerne les voyelles orales, il semble que moins la voyelle est fermée et postérieure, plus elle est influencée par le locuteur. Ces tendances, illustrées par le tableau 7.1 et la Figure 7.5, sont observées aussi bien pour les hommes que pour les femmes. Plusieurs hypothèses peuvent être émises pour expliquer ces observations.

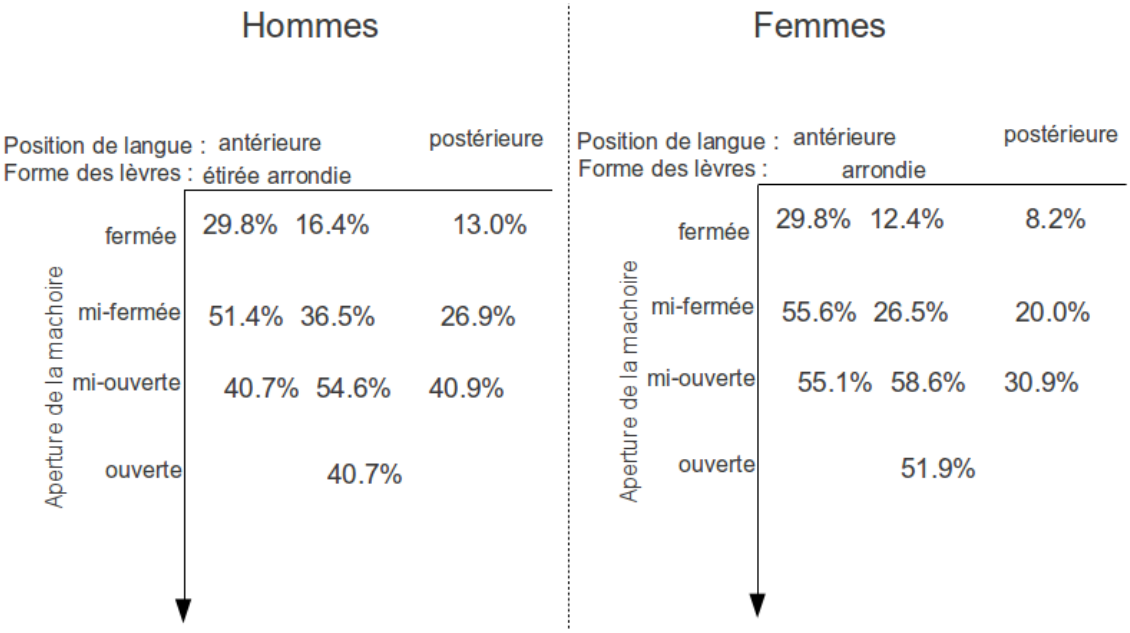


FIGURE 7.5 – η^2 *pour les voyelles orales en fonction de leurs lieux d'articulation : plus la voyelle est fermée et postérieure, moins elle est influencée par le locuteur.*

Lors de la réalisation de nasales, l'air passe par la cavité nasale dont la forme dépend énormément de l'individu. Les fricatives sont le fruit d'une constriction de la langue, le long du palais dur pour /s/, /z/ et /ʃ/, /ʒ/ et d'une constriction entre les lèvres et les dents pour /f/, /v/. Pour les quatre premières fricatives, nous pouvons faire l'hypothèse que la forme du palais dur est dépendante du locuteur (Toda, 2009). Pour les labio-dentale, la forme des dents peut, peut-être, jouer un rôle.

Par ailleurs, les consonnes dentales semblent également plus dépendantes du locuteur que les autres consonnes. La forme des dents peut à nouveau expliquer ces différences. Pour les voyelles, /u/ consiste à positionner le langue de manière à obtenir deux résonances d'Helmotz pour produire deux formants très bas. Pour réaliser ceci, les degrés de libertés possibles ne sont pas très nombreux pour l'appareil phonatoire humain, il n'est donc pas possible de discriminer les locuteurs à partir de ce type de voyelle.
Les degrés d'aperture de la voyelle peuvent faire partie d'une stratégie du locuteur. Il peut choisir de réaliser de grands gestes ou bien des petits gestes.

Les centres de gravités, en fonction du phonème, apparaissent comme des indices d'informations propres au locuteur. Similairement aux paramètres de description de la source, il s'agit à présent de savoir si cette mesure peut permettre de distinguer un enregistrement de la série *Min* d'un enregistrement de la série *Max*.

7.2.2 Les centres de gravité et les performances des modèles de locuteur

Influence du locuteur sur les centres de gravité extraits de *Min* et *Max*

Nous allons commencer par vérifier si **le facteur locuteur joue un rôle plus important dans la série *Min* que *Max***. Pour cela, nous mesurons les centres de gravité de chaque phonème présent dans les deux séries. Puis, des ANOVA à un facteur sont réalisées pour chaque phonème et pour chaque série. Le facteur fixé est donc le locuteur et la variable dépendante est la valeur du centre de gravité.
L'ensemble des résultats obtenus est résumé par le tableau D.4 en annexe. Nous retrouvons dans les deux séries, les mêmes tendances que celles que nous avons décrites en 7.2.1 à savoir que les nasales, les voyelles orales et les fricatives sont les catégories dont les mesures de centre de gravité sont les plus sensibles au locuteur. Toutefois, les glides ne présentent pas toujours des différences significatives, aussi bien dans la série *Min* que dans la série *Max*. Il s'agit d'une catégorie pour laquelle le locuteur n'expliquait pas beaucoup de variance d'après 7.2.1.
Il est également à noter que, pour un peu plus de 50% des phonèmes prononcés par les hommes, les valeurs de centre de gravité ont une part de variance expliquée par le facteur locuteur plus importante dans *Min* que dans *Max*. En moyenne, la part de variance expliquée par le locuteur dans la série *Min* est plus importante que dans la série *Max* ($\eta^2_{Min}Moyen = 35.6\%$ vs $\eta^2_{Max}Moyen = 32.8\%$). Pour les femmes, le même phénomène

est observé : en moyenne le facteur locuteur explique une plus grande part de variance dans la série *Min* que dans la série *Max* ($\eta^2_{Min} Moyen = 34.3\%$ vs $\eta^2_{Max} Moyen = 32.6\%$). Ce phénomène est illustré par les Figures 7.6 et 7.7 pour respectivement les hommes et les femmes.

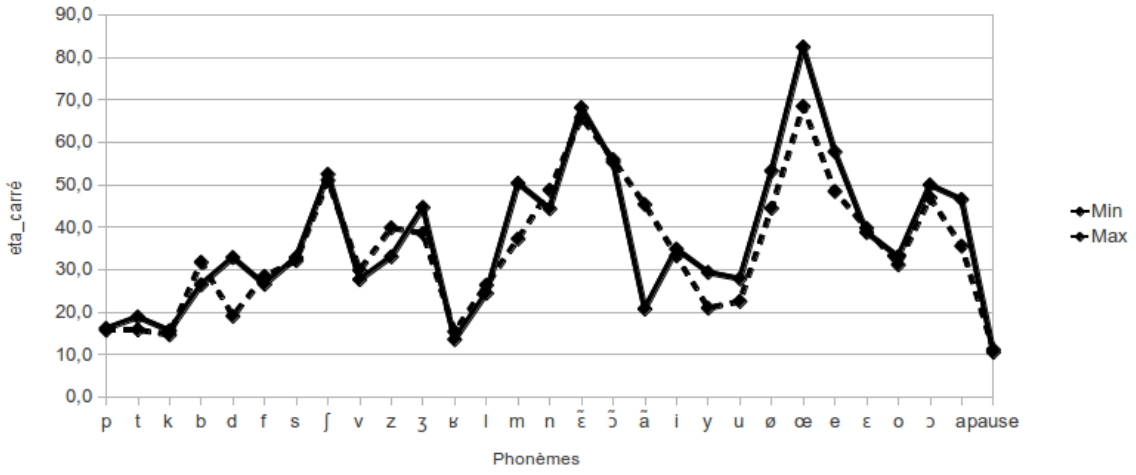


FIGURE 7.6 – *Valeurs des η^2 pour chaque phonème de Min et de Max pour les hommes*

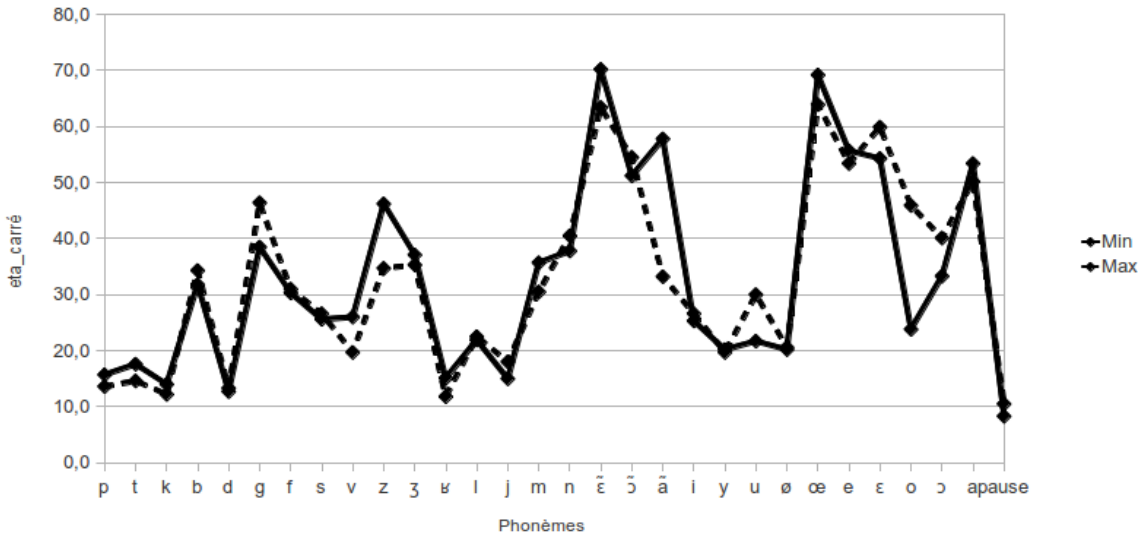


FIGURE 7.7 – *Valeurs des η^2 pour chaque phonème de Min et de Max pour les femmes*

Les centres de gravités semblent être légèrement plus influencés par le locuteur dans la série *Min* que dans la série *Max*. Reste à savoir si les valeurs des centres de gravité sont significativement différentes entre les deux séries.

Différencier *Min* et *Max* par les valeurs de centre de gravité

Des ANOVA à mesure répétée ont été réalisées avec comme variable dépendante, pour chaque phonème séparément, les valeurs de centre de gravité, comme facteur la Performance et comme facteur inter-sujet les locuteurs.
Comme l'illustre le tableau D.6 en annexe, seul le phonème /m/ ($p < 0.05$) pour les hommes et les phonèmes /ã/ et /ɛ̃/ pour les femmes connaissent des différences significatives entre les deux séries. Toutefois la variance expliquée par la Performance pour ces deux phonèmes reste très faible (pour les hommes : $\eta^2_m = 0.48\%$ et pour les femmes : $\eta^2_{\tilde{a}} = 0.33\%$ et $\eta^2_{\tilde{\varepsilon}} = 0.16\%$). Les centres de gravité ne permettent pas d'expliquer les différences de performance observées lors des tests d'évaluation.
Les centres de gravité permettent de résumer en une valeur comment se répartissent les fréquences qui composent le son concerné. Il semble que cette valeur, pour certains phonèmes (les nasales, les voyelles et les fricatives), soit dépendante de façon importante du locuteur. Cependant, les valeurs de centre de gravité n'expliquent pas les différences de performance observées entre les deux séries de fichiers.

7.3 Les voyelles orales par leurs valeurs de formants

7.3.1 Approche

Les formants sont un autre moyen de décrire la qualité des voyelles orales en rendant compte notamment de la taille du conduit vocal (Fant, 1970). Il semble donc judicieux d'**étudier l'influence du locuteur sur les quatre premiers formants**.
Les voyelles orales sont généralement décrites par leurs valeurs formantiques. Ces dernières dépendent en premier lieu de la position des articulateurs que sont la langue, la mâchoire ou les lèvres. Les modèles les plus courants de la production des voyelles représentent le conduit vocal en une suite de tubes. La longueur de ces tubes dépend de la position des articulateurs mais aussi de la longueur intrinsèque du conduit. Ce dernier facteur dépend de la morphologie du locuteur.

7.3.2 Impact du locuteur et de la catégorie vocalique sur les formants

En premier lieu, nous proposons d'**étudier l'impact du locuteur et du timbre de la voyelle sur les voyelles en considérant que chaque voyelle est représentée par la valeur moyenne de ses quatre premiers formants**. En effet, le timbre de la voyelle est le plus souvent décrit par les trois premiers formants alors que le quatrième formant est parfois rattaché au locuteur (Lavner et al., 2000).
Nous mesurons, pour chaque voyelle des enregistrements de tests issus de BREF, les valeurs moyennes des quatre premiers formants à l'aide de Praat (Boersma et Weenink, 2009). Les valeurs médianes des formants sont résumées dans les tableaux 7.4 et 7.5 respectivement pour les hommes et les femmes. Nous retrouvons globalement les valeurs de référence pour les voyelles du français (Gendrot et Adda-Decker, 2007).

Description de la voyelle par les quatre premiers formants

Afin de mesurer l'influence du locuteur et du timbre vocalique sur les mesures des quatre premiers formants, une MANOVA à deux facteurs est calculée sur l'ensemble de ces mesures. La variable dépendante est un vecteur composé des valeurs moyennes des quatre premiers formants, les deux facteurs sont le timbre de la voyelle et le locuteur.
Les deux facteurs ont un effet sur les valeurs des 4 premiers formants aussi bien pour les hommes (Timbre : $F(9,36) = 17\,460.8; p < 0.001$, Locuteur : $F(46,184) = 756.8; p < 0.001$) que pour les femmes (Timbre : $F(9,36) = 32\,456; p < 0.001$, Locuteur : $F(63,252) = 649; p < 0.001$). Comme nous nous y attendions, le timbre de la voyelle (pour les hommes : $\eta^2 = 26.8\%$ et pour les femmes : $\eta^2 = 32.5\%$) explique plus de variance que le facteur locuteur (pour les hommes $\eta^2 = 1.8\%$ et pour les femmes $\eta^2 = 1.2\%$).

Analyse de chaque formant

Des ANOVA à un facteur ont ensuite été réalisées pour connaître l'influence du locuteur et du timbre de la voyelle sur chacun des formants. Un effet des deux facteurs est observé pour chaque formant ($p < 0.001$). Les valeurs formantiques sont avant tout affectées par le timbre de la voyelle aussi bien pour F1, F2, F3 que F4 pour les hommes comme pour les femmes comme le retranscrit le tableau 7.3 qui contient les $\eta^2_{voyelle}$ et les $\eta^2_{locuteur}$. **Le locuteur joue un rôle moindre dans ce cas mais il semble que F3 et F4**

	F1		F2		F3		F4	
	$\eta^2_{voyelle}$	$\eta^2_{locuteur}$	$\eta^2_{voyelle}$	$\eta^2_{locuteur}$	$\eta^2_{voyelle}$	$\eta^2_{locuteur}$	$\eta^2_{voyelle}$	$\eta^2_{locuteur}$
Hommes	28.1%	7.0%	75.0%	1.7%	54.3%	7.8%	36.5%	15.6%
Femmes	58.7%	3.5%	76.3%	7.6%	51.5%	10.5%	45.5%	11.2%

TABLE 7.3 – *η^2 du timbre vocalique et du locuteur sur les quatre premiers formants*

soient plus dépendantes du locuteur que F1 et F2. Au vu de l'importance du timbre de la voyelle dans l'analyse de variance, une étude voyelle par voyelle semble incontournable pour comprendre l'influence du locuteur sur les formants.

7.3.3 Impact du locuteur sur les valeurs formantiques pour chaque voyelle

Variation intra-locuteur vs variation inter-locuteur

Dans un premier temps, pour chaque voyelle, les variations inter et intra-locuteurs ont été calculées pour les hommes 7.4 et pour les femmes 7.5. La variation intra-locuteur

		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
Occurrences		31 128	19 585	7 371	23 151	21 260	23 822	2 915	9 126	6 575	9 353
F1	Moyen	602	498	509	434	487	384	509	521	434	412
	Inter	39	28	75	33	97	31	37	54	38	35
	Intra	94	90	115	114	142	113	60	93	113	126
F2	Moyen	1 476	1 759	1 334	1 951	1 575	2 363	1 474	1 312	1 070	2 223
	Inter	49	62	225	77	97	84	58	87	40	64
	Intra	189	138	310	125	237	207	135	239	187	187
F3	Moyen	2 540	2 595	2 707	2 694	2 592	3 040	2 509	2 572	2 052	2 890
	Inter	113	98	144	82	117	78	102	118	89	99
	Intra	132	132	222	144	192	187	123	172	188	223
F4	Moyen	3 680	3 686	3 667	3 709	3 570	3 662	3 547	3 580	2 806	3 542
	Inter	163	157	124	150	121	92	144	122	80	91
	Intra	178	189	228	183	207	191	153	180	198	181
η^2 multivarié		30%	30%	25%	29%	23%	16%	30%	26%	13%	15%

TABLE 7.4 – *Valeurs moyennes, variation inter et variation intra des formants des voyelles orales pour les hommes*

est plus importante que la variation inter-locuteur. Cette variation peut s'expliquer par les phénomènes de co-articulation des segments.

		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
Occurrences		40 683	26 422	8 599	30 380	26 403	31 423	3 803	12 744	8 427	12 186
F1	Moyen	708	571	505	481	471	383	587	551	443	422
	Inter	45	29	30	29	27	29	35	32	33	29
	Intra	98	82	87	79	72	90	54	78	100	113
F2	Moyen	1 705	2 021	1 227	2 229	1 676	2 409	1 676	1 382	1 086	2 260
	Inter	85	94	49	101	67	80	84	59	41	63
	Intra	227	181	224	173	227	130	144	222	186	193
F3	Moyen	2 833	2 911	2 860	3 006	2 808	3 021	2 843	2 846	1 886	2 876
	Inter	147	131	141	120	126	70	148	153	68	58
	Intra	184	162	160	151	174	169	138	155	217	176
F4	Moyen	3 940	4 004	3 949	4 050	3 882	3 620	3 948	3 939	2 826	3 597
	Inter	212	220	144	192	135	73	172	151	72	70
	Intra	266	265	181	242	221	191	201	195	150	183
η^2 multivarié		28%	29%	27%	26%	21%	15%	37%	28%	11%	10%

TABLE 7.5 – *Valeurs moyennes, variation inter et variation intra des formants des voyelles orales pour les femmes*

Effet du locuteur sur les quatre premiers formants

Des MANOVA à un facteur ont été réalisées pour chaque voyelle afin de mesurer l'influence du locuteur sur chaque voyelle représentée par les quatre premier formants. Le facteur est le locuteur et la variable dépendante est un vecteur de dimension 4 composé de la moyenne de F1, F2, F3 et F4 pour la voyelle donnée. Un effet du locuteur est observé pour chaque timbre de voyelle ($p < 0.0001$). L'effet du locuteur varie cependant en fonction de la voyelle. Les voyelles fermées semblent être moins affectées par le locuteur que les voyelles plus ouvertes comme l'illustre le tableau 7.6. Pour /i/, /u/ et /y/, η^2 est compris entre 13% et 16% pour les hommes tandis que pour les autres voyelles η^2 se situe entre 23% et 30%. Le même phénomène est observé pour les femmes.

Effet du locuteur par formant

Après avoir étudié l'effet du locuteur sur le vecteur F1/F2/F3/F4, des ANOVA à un facteur par voyelle pour chaque formant ont été réalisées afin de mesurer l'effet du locuteur sur chaque formant et cela pour chaque voyelle. Les résultats sont également résumés dans le tableau 7.6

Un effet du locuteur est observé pour chaque formant de chaque voyelle (p<0.001). F3

η^2		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
H	F1	15.1%	9.7%	28.6%	7.8%	27.8%	7.4%	27.8%	21.9%	9.9%	6.9%
	F2	6.2%	17.1%	32.4%	26.7%	13.2%	13.7%	14.1%	9.9%	4.4%	10.7%
	F3	41.8%	37.0%	28.8%	24.0%	26.4%	14.7%	41.0%	30.0%	19.3%	16.1%
	F4	42.5%	38.7%	21.6%	37.9%	24.9%	19.4%	43.7%	30.2%	14.4%	20.3%
	η^2 multi-varié	30%	30%	25%	29%	23%	16%	30%	26%	13%	15%
F	F1	17.5%	11.0%	11.4%	12.2%	14.2%	9.9%	28.1%	14.4%	9.8%	6.5%
	F2	12.2%	20.4%	4.7%	24.3%	7.9%	26.8%	25.6%	6.4%	4.4%	9.4%
	F3	37.4%	39.0%	43.4%	37.7%	33.6%	14.6%	52.3%	46.6%	8.8%	9.5%
	F4	37.2%	41.3%	37.6%	36.1%	25.5%	12.6%	41.6%	37.0%	17.8%	12.1%
	η^2 multi-varié	28%	29%	27%	26%	21%	15%	37%	28%	11%	10%

TABLE 7.6 – η^2 *pour chaque voyelle décrite à l'aide des formants : les voyelles fermées sont moins influencées par le facteur locuteur que les autres voyelles.*

et F4 semblent être plus sensibles au locuteur que F1 et F2. Par exemple, /a/, prononcé par les hommes, voit 30% de sa variance expliquée par le locuteur lorsque une analyse des 4 formants est réalisée. F1 seul voit 15.1% de sa variance expliquée par le locuteur tandis que la variance de F4 est expliquée à 42.5% par le locuteur. Ce phénomène est observé pour chaque voyelle même si les voyelles fermées sont moins affectées que les autres voyelles.

Les voyelles fermées semblent être moins propices que les autres pour différencier les locuteurs.

7.3.4 Les formants des voyelles orales pour différencier *Min* de *Max*

La part de variance des valeurs formantiques expliquée par le facteur locuteur est-elle plus grande dans la série *Min* que dans la série *Max* ?

Effet du locuteur sur les quatre premiers formants des voyelles extraits de *Min* et *Max*

Comme nous savons que le timbre de la voyelle influence fortement les valeurs formantiques, nous avons effectué en premier lieu une MANOVA à un facteur sur l'ensemble des données de *Min*, d'une part, de *Max*, d'autre part, afin de mesurer l'effet des locuteur sur les valeurs formantiques des voyelles. Ici encore, la variable dépendante se compose d'un vecteur de dimension 4 composé des valeurs moyennes de F1, F2, F3 et F4 et le facteur fixe est le locuteur.
Un effet du locuteur ($p < 0.001$) est observé pour toutes les voyelles. Les voyelles fermées sont encore celles qui sont le moins influencées par le locuteur, aussi bien pour les hommes ($19.3 < \eta^2_{\text{voyelles fermées}} < 27.0$ vs $25.4 < \eta^2_{\text{voyelles plus ouvertes}} < 59.6$) que pour les femmes ($19.0 < \eta^2_{\text{voyelles fermées}} < 25.2$ vs $21.8 < \eta^2_{\text{voyelles plus ouvertes}} < 57.9$).
Trois voyelles -/ɛ/, /e/ et /i/- pour les hommes et quatre voyelles -/o/, /e/, /ø/ et /u/- pour les femmes sont plus influencées par le locuteur dans *Min* que dans *Max* (cf. les tableaux en annexe D.6.1 et D.6.2). Ces voyelles ne sont pas les mêmes pour les hommes et pour les femmes. Le pourcentage de variance expliquée par le locuteur pour les formants ne nous permet pas de distinguer la série *Min* de la série *Max* aussi bien pour les hommes que pour les femmes.

Effet du locuteur pour chaque formant des voyelles extraits de *Min* et *Max*

Des ANOVA à un facteur pour chaque voyelle pour la série *Min*, d'une part, et pour la série *Max*, d'autre part, sont calculées de manière à évaluer l'effet du locuteur sur les valeurs formantiques des voyelles des deux séries. Le facteur fixé est le locuteur et la variable est la valeur moyenne du formant étudié.
L'analyse de chaque formant séparément donne les résultats suivants. Quatre voyelles (/a/, /e/, /ɔ/ et /y/) pour les hommes et huit voyelles (/ɛ/, /o/, /e/, /ø/, /i/, /œ/, /ɔ/ et /y/) pour les femmes ont une part de variance des valeurs de F1 plus importante pour la série *Min* que pour la série *Max*.
Trois voyelles (/i/, /u/ et /y/) pour les hommes et cinq voyelles (/o/, /ø/, /ɔ/, /u/ et /y/) pour les femmes ont une part de variance des valeurs de F2 plus importante pour la série *Min* que pour la série *Max*.
Deux voyelles (/i/, /y/) pour les hommes et trois voyelles (/ø/, /œ/ et /u/) pour les

femmes ont une part de variance des valeurs de F3 plus importante pour la série *Min* que pour la série *Max*.
Six voyelles (/a/, /ɛ/, /e/, /ø/, /i/ et /œ/) pour les hommes et sept voyelles (/a/, /o/, /e/, /ø/, /i/, /u/ et /y/) pour les femmes ont une part de variance des valeurs de F4 plus importante pour la série *Min* que pour la série *Max*.
La comparaison des résultats obtenus pour *Min* et pour *Max* est illustrée par la figure

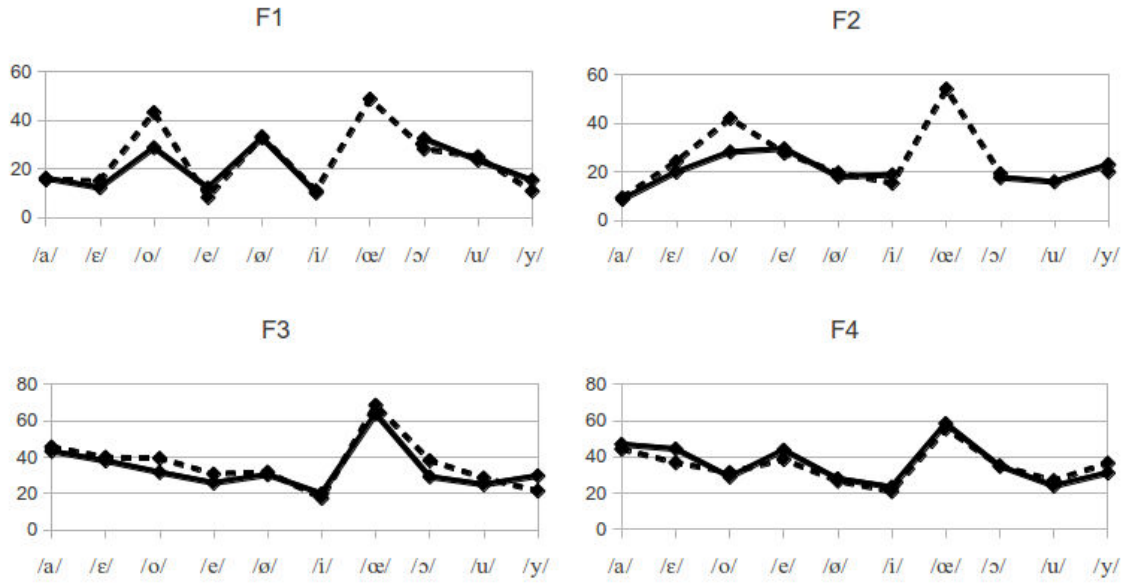


FIGURE 7.8 – *Valeurs de η^2 pour les séries Min (trait plein) et Max (trait pointillé) en fonction de chaque voyelle pour les hommes*

7.8 pour les hommes et la figure 7.9 pour les femmes. **Si pour F4, nous observons, pour plus de la moitié des voyelles, une part de variance plus importante expliquée par le facteur locuteur, les différences entre *Min* et *Max* ne sont pas très importantes**. Nous allons vérifier cela grâce à une ANOVA à mesures répétées.

Différence de valeurs formantiques

Cette analyse par ANOVA à mesures répétées est réalisée pour chaque phonème et chaque formant séparément. Le facteur fixé est la Performance, la variable inter-sujet est le locuteur et la variable dépendante est la valeur formantique. En ce qui concerne F1, nous ne trouvons aucune différence significative pour les hommes. Pour les femmes, seul le F1 du /ɛ/ est significativement différent entre *Min* et *Max* ($p < 0.05$). L'étude de F2 montre, pour les hommes, comme seule différence le F2 du /e/ ($p < 0.05$). Aucune

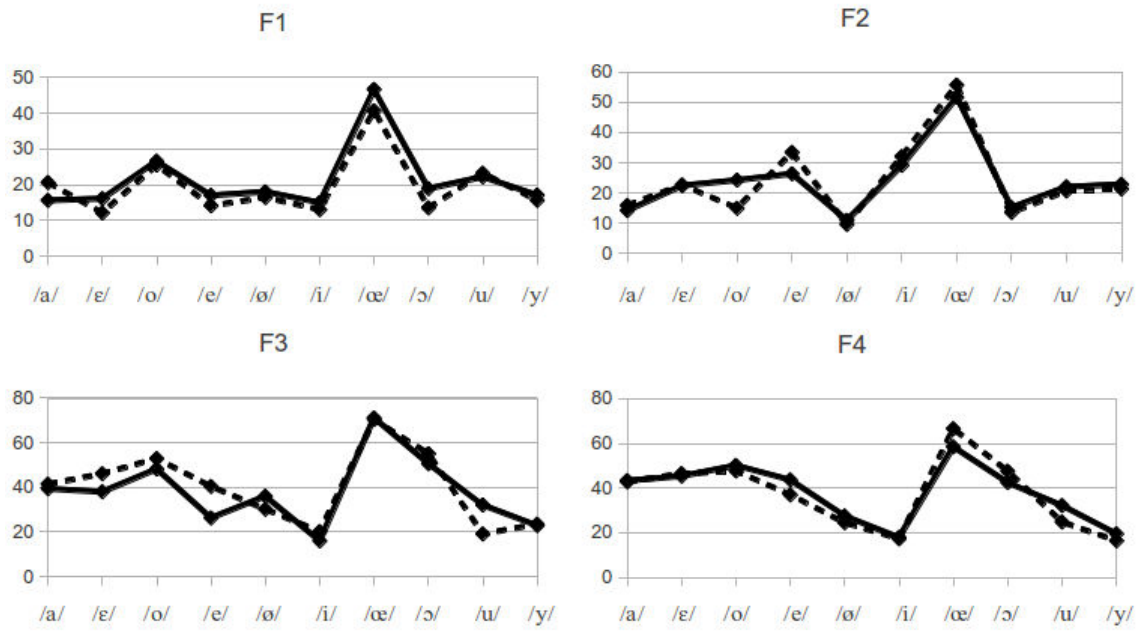


FIGURE 7.9 – *Valeurs de η^2 pour les séries Min (trait plein) et Max (trait pointillé) en fonction de chaque voyelle pour les femmes*

différence n'est trouvée pour les femmes. Aucun F3 n'est significativement différent que ce soit pour les hommes ou pour les femmes. Enfin, pour F4, des différences significatives s'observent pour /œ/ ($p < 0.05$), /u/ ($p < 0.05$) et /y/ ($p < 0.05$) pour les femmes. Pour les hommes, seul /e/ a un F4 significativement différent entre *Min* et *Max* ($p < 0.05$). Le tableau 7.7 résume les différences au niveau formantique pour les voyelles orales.

	F1	F2	F3	F4
Hommes	-	/e/*	-	/e/*
Femmes	/ɛ/*	-	-	/œ/* /u/* /y/*

TABLE 7.7 – *Résumé des phonèmes pour lesquels les mesures formantiques sont significativement différentes entre Min et Max*

7.3.5 L'aire des triangles pour différencier *Min* de *Max*

Méthode de calcul

Une autre mesure permettant de rendre compte des valeurs formantiques est le **calcul de l'aire du triangle vocalique sur F1/F2** tel que réalisé dans (Bradlow et al., 1996). Pour calculer cette aire, A, il suffit de connaître les coordonnées des voyelles extrêmes du triangle à savoir /i/, /a/ et /u/. A partir de ces coordonnées, nous pouvons calculer les 3 longueurs qui composent le triangle telles que définit par l'équation 7.1.

$$l_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{7.1}$$

où l_{ab} est la longueur qui sépare le point a et le point b, x_a et x_b sont respectivement l'abscisse du point a et b et y_a et y_b sont respectivement les ordonnées des points a et b. Il suffit par la suite de calculer l'aire obtenues par le triangle formé par ces 3 points à l'aide de la formule 7.2.

$$A = \sqrt{s(s - l_{iu})(s - l_{ua})(s - l_{ai})} \tag{7.2}$$

où s est la demie somme définie par l'équation 7.3 et A l'aire du triangle.

$$s = \frac{1}{2} \times (l_{iu} + l_{ua} + l_{ai}) \tag{7.3}$$

L'aire du triangle acoustique est ainsi calculée pour chaque locuteur dans la série *Min* d'une part et la série *Max* d'autre part. Nous avons par la suite comparé les valeurs d'aire obtenues pour savoir s'il existe une différence entre les séries. Pour effectuer cette comparaison, un t-test apparié est réalisé pour chaque genre.

Hypothèses

Nous pouvons faire **l'hypothèse que dans la série *Min* les sons sont plus hyperarticulés et donc que l'aire du triangle est plus grande que dans *Max***.

Comparaison des aires pour les séries *Min* et *Max*

Les triangles issus des séries *Min* et *Max* sont représentés par les figures 7.10 et 7.11 respectivement pour les hommes et les femmes. Pour les hommes, la moyenne

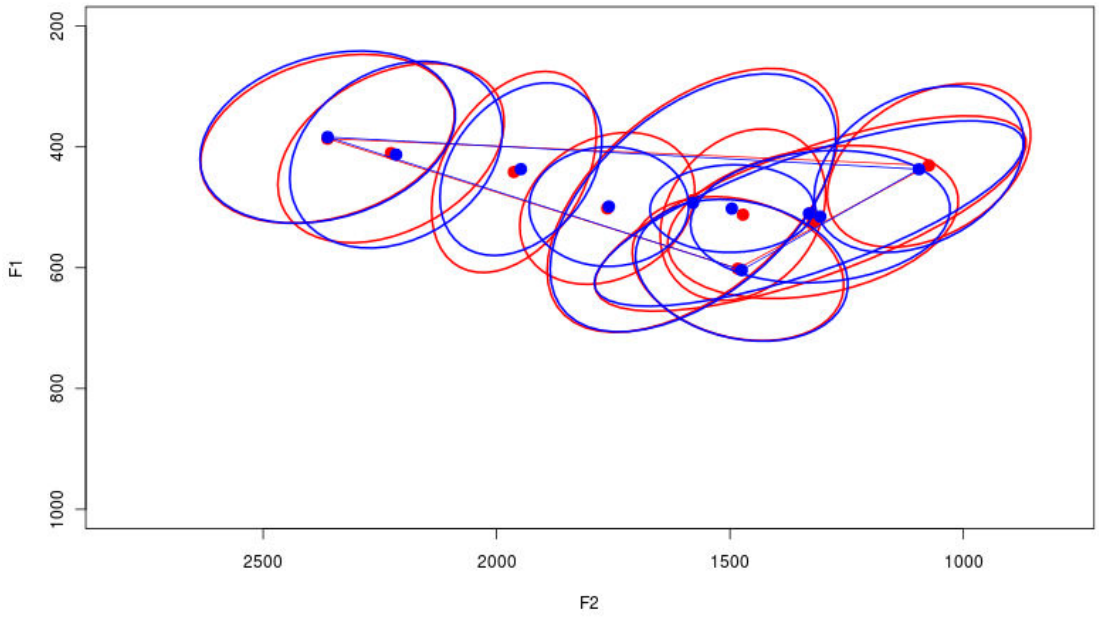


FIGURE 7.10 – *Triangles F1/F2 pour la série Min (Rouge) et Max (Bleu) pour les hommes*

des aires obtenue pour *Min* est de 65 271Hz^2 tandis que l'aire moyenne obtenu pour *Max* est de 61 061Hz^2 mais il n'existe pas de différence significative entre ces valeurs ($t(46) = 1.3001; p = 0.2005$). Pour les femmes, l'aire moyenne de *Min* est égale à 102 017Hz^2 tandis que l'aire moyenne de *Max* est égale 102177Hz^2, sans différence significative ($t(63) = -0.0493; p = 0.9609$). Les aires des triangles vocaliques de *Min* et *Max* ne nous permettent pas de différencier les deux séries.

Le locuteur semble plus influencer les valeurs formantiques de F3 et F4. D'autre part, les voyelles fermées semblent être moins sensibles au locuteur que les autres voyelles. Les degrés d'aperture de la mâchoire pourraient être dépendants du locuteur qui selon sa personnalité ouvrirait plus ou moins la mâchoire. Toutefois, les différences de valeurs formantiques, brutes ou ramenées à des aires de triangles, ne permettent pas de différencier les deux séries et donc d'expliquer les différences de performance observées. Les différences se situent peut-être au niveau de la coarticulation comme le souligne (Nolan et Grigoras, 2005).

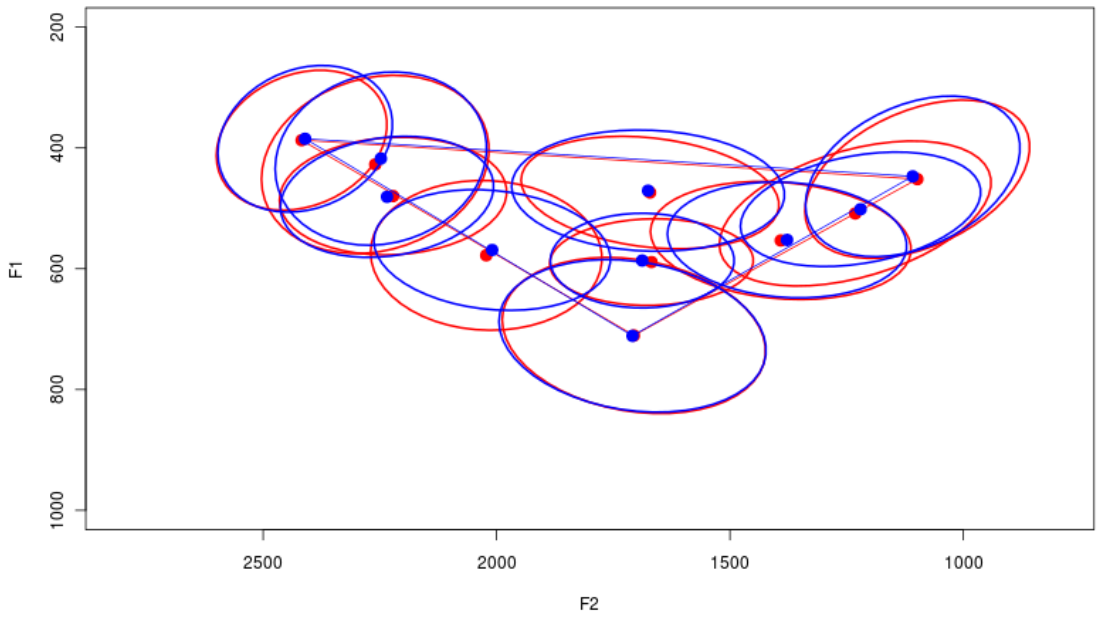


FIGURE 7.11 – *Triangles F1/F2 pour la série Min (Rouge) et Max (Bleu) pour les femmes*

7.4 Importance de la co-articulation

(McDougall, 2006) a mis en évidence que les **transitions formantiques sont un indice qui permet de discriminer les locuteurs entre eux**. Ces transitions formantiques sont par ailleurs très dépendantes de la co-articulation. Ainsi les transitions ne sont pas les mêmes selon le lieu d'articulation de la consonne qui précède la voyelle étudiée. Dans un premier temps, nous avons donc étudié les **distributions des trigrammes** présents dans les deux séries afin de voir si certains étaient plus présents que d'autres. Puis nous nous sommes interrogé sur le pouvoir discriminant des transitions formantiques pour séparer les locuteur de la base de données BREF.

7.4.1 Distributions de trigrammes

Nous n'avons étudié que les trigrammes pour lesquels le second élément est identifié comme une voyelle puisque, par la suite, nous nous intéresserons aux transitions formantiques. Les éléments de gauche et de droite de ces trigrammes peuvent pren-

dre différentes valeurs selon le mode et le lieu d'articulation du phonème. Nous avons retenu 6 catégories : les consonnes labiales, les consonnes coronales, les consonnes dorsales, les voyelles nasales, les voyelles orales et les pauses, ces catégories n'influençant théoriquement pas de la même manière la voyelles en terme de co-articulation (Vaissière, 2006). Pour chaque voyelle, nous avons calculé la fréquence de chacun des trigrammes pour chaque locuteur de chacune des séries. Une MANOVA avec comme facteur fixé la Performance et comme variable dépendante un vecteur contenant le nombre de fois où chaque trigramme a été rencontré est effectuée afin de savoir si les distribution de trigrammes sont différentes entre *Min* et *Max*.
Aucun effet de la Performance n'est observé sur les distributions de trigrammes aussi bien pour les hommes ($F(36, 1\,314) = 0.63889; p = 0.9527$) que pour les femmes ($F(36, 1\,790) = 0.71861; p = 0.8924$). La figure 7.12 illustre la distribution des trigrammes pour les femmes. **Les distributions de trigrammes entre *Min* et *Max* ne sont pas différentes**

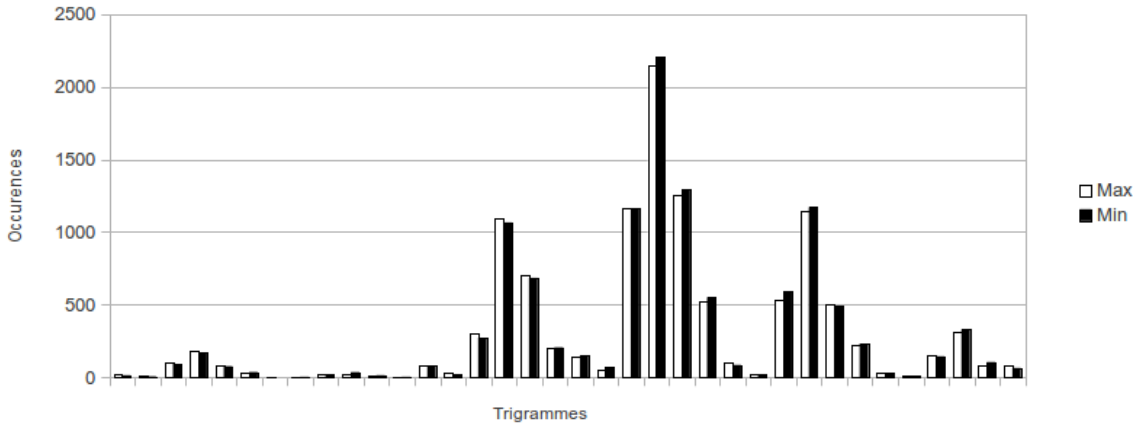


FIGURE 7.12 – *Distributions des trigrammes pour la série Min (noir) et la série Max (blanc), le cas des locutrices*

aussi bien pour les hommes que pour les femmes. Les effets de la co-articulation doivent donc être équivalents entre les deux séries.

7.4.2 Étude du locus

Il s'agit d'étudier si **les valeurs de transitions formantiques pour chaque voyelle en fonction du lieu d'articulation permettent de mieux discriminer les locuteurs** présents dans la base de données BREF, le locus étant une mesure de l'influence de la consonne précédente sur les valeurs de F2 de la voyelle.

Méthode

La transition formantique se mesure généralement sur le second formant. Il s'agit de mesurer la valeur de F2 à 10% de la voyelle et à 50% de la voyelle puis d'observer la droite de régression linéaire qui lie ces deux valeurs. Cette droite de régression se calcule pour chacun des lieux de la consonne qui précède la voyelle étudiée (Sussman et al., 1992). **Nous avons donc étudié, séparément, les voyelles précédées de consonnes labiales, de consonnes coronales et de consonnes dorsales.**
Trois valeurs peuvent être étudiées dans cette régression. La première est le R^2 qui indique comment la droite de régression suit les données. La seconde est le coefficient de la droite de régression, noté a, et le troisième est l'origine de la droite, noté b. a indique si les consonnes sont bien coarticulées dans les voyelles. Plus la pente est proche de 1 plus les consonnes sont coarticulées.
L'analyse est effectuée pour les hommes et pour les femmes séparément. La comparaison entre *Min* et *Max* est effectuée par t-test pour chaque R^2, a et b de chaque lieu d'articulation.

Résultats

L'ensemble des résultats obtenus est présenté dans le tableau 7.8.
Pour les hommes, R^2 varie, en fonction des lieux d'articulation, de 0.50 à 0.59 pour *Min*

		Hommes			Femmes		
		Moyenne *Min*	Moyenne *Max*	p	Moyenne *Min*	Moyenne *Max*	p
R^2	labiale	0.50	0.53	0.51	0.62	0.67	0.08
	coronale	0.54	0.52	0.41	0.73	0.77	0.07
	dorsale	0.59	0.59	0.96	0.58	0.61	0.2933
a	labiale	0.76	0.79	0.42	0.58	0.61	0.2884
	coronale	0.71	0.70	0.70	0.87	0.90	0.09
	dorsale	0.70	0.73	0.53	0.86	0.90	0.123
b	labiale	350	325	0.56	115	50	0.23
	coronale	449	465	0.73	254.7	197.1	0.0885
	dorsale	432	410	0.68	245	149	0.10

TABLE 7.8 – *Valeurs moyenne de R^2, a et b afin de définir le locus, selon les séries Min et Max*

et de 0.52 à 0.59 pour *Max*. Aucune différence de distribution n'est observée pour les R^2 ($0.41 < p < 0.96$). Pour les femmes, R^2 varie en fonction des lieux d'articulation de 0.58 à 0.73 pour *Min* et de 0.61 à 0.67 pour *Max*. Aucune différence de distribution n'est observée pour les R^2 ($0.07 < p < 0.2933$). R^2 ne permet pas de différencier les deux séries. Il est tout de même à noter que les valeurs des femmes sont plus élevées que celles des hommes.
Pour les hommes, *a* varie, en fonction des lieux d'articulation, de 0.70 à 0.76 pour *Min* et de 0.70 à 0.79 pour *Max*. Aucune différence de distribution n'est observée pour les *a* ($0.42 < p < 0.70$). Pour les femmes, *a* varie en fonction des lieux d'articulation de 0.86 à 0.95 pour *Min* et de 0.90 à 0.99 pour *Max*. Aucune différence de distribution n'est observée pour les *a* ($0.09 < p < 0.2884$). *a* ne permet pas de différencier les deux séries. Les valeurs de *a* indiquent que les consonnes sont plus coarticulées par les femmes que par les hommes.
Pour les hommes, *b* varie, en fonction des lieux d'articulation de 350 à 449 pour *Min* et de 325 à 465 pour *Max*. Aucune différence de distribution n'est observée pour les *b* ($0.56 < p < 0.73$). Pour les femmes, *b* varie en fonction des lieux d'articulation de 115 à 254.7 pour *Min* et de 50 à 197.1 pour *Max*. Aucune différence de distribution n'est observée pour les *b* ($0.0885 < p < 0.23$). Le paramètre *b* ne permet pas non plus de différencier les deux séries.

Les valeurs de locus ne permettent pas de distinguer les deux séries *Min* et *Max*. Une autre manière de mesurer la coarticulation est le suivi de formant. Ce procédé nous permettra-t-il de distinguer nos deux séries d'enregistrements ?

7.4.3 Effet du locuteur sur les courbes formantiques

Nous souhaitons vérifier si, pour notre corpus, le **suivi de formant tel que préconisé par (McDougall, 2006)** montre des différences en fonction des locuteurs de BREF puis s'il permet de séparer la série *Min* de la série *Max*.

Mesures et analyses

Pour chaque voyelle, nous avons extrait tous les 10% la valeur des quatre premiers formants. Chaque formant est donc représenté par neuf valeurs de 10 à 90% de la durée

de la voyelle. Nous souhaitons d'abord vérifier si ces valeurs nous permettent de discriminer les locuteurs. En d'autres termes, nous souhaitons mesurer la part de variance de ces valeurs qui peut être expliquée par le locuteur. Pour cela une MANOVA a été réalisée pour chaque voyelle et chaque formant. Le facteur fixé est dans ce cas le locuteur et la variable dépendante est un vecteur se composant des 9 valeurs du formant.

Effet du locuteur sur les courbes de formants

Un effet du locuteur est observé pour toutes les voyelles et pour tous les formants ($p < 0.001$). Pour les hommes, l'effet du locuteur sur F1 fluctue entre 2.1% (/e/) et 7.9% (/œ/). Pour F2, il varie de 1.4% (/a/) à 6.3% (/œ/). Pour F3, les valeurs de η^2 s'échelonnent de 2.6% (/i/) à 9.5% (/œ/). Enfin, pour F4, elles se situent entre 3.6% (/i/) et 10.3% (/œ/).
Pour les femmes, l'effet du locuteur sur F1 fluctue entre 2.1% (/ɛ/) et 6.4% (/œ/). Pour F2, il varie de 1.4% (/ø/) à 5.3% (/œ/). Pour F3, les valeurs de η^2 s'échelonnent de 2.5% (/y/) à 9.8% (/œ/). Enfin, pour F4, elles se situent entre 2.4% (/i/) et 8.5% (/œ/).
Le suivi de formant ne semble pas très influencé par le locuteur même si F3 et F4 montre plus d'influence du locuteur que F1 et F2. Il est tout de même à noter que nos mesures sont effectuées automatiquement aussi bien pour positionner les frontières de phonèmes que pour l'extraction des valeurs de formant. Ceci peut expliquer des erreurs de mesures pour une analyse de l'évolution temporelle des formants.

7.4.4 Séparer *Min* et *Max* par les courbes formantiques

Nous souhaitons vérifier si les valeurs des courbes formantiques nous permettent de distinguer *Min* de *Max*.

Méthode

Dans un premier temps, nous vérifions si la part de variance expliquée par le locuteur est plus grande dans *Min* que dans *Max*. Pour cela nous réalisons une MANOVA pour chaque voyelle et chaque formant pour la série *Min* d'une part et *Max* d'autre part. Puis dans un second temps, nous vérifions si les valeurs trouvées nous permettent de

distinguer *Min* de *Max* à l'aide d'une MANOVA où le facteur fixé est la Performance et la variable dépendante est un vecteur se composant des neuf valeurs du formant.

Effet du locuteur sur les courbes formantiques de *Min* et *Max*

L'ensemble des résultats est résumé par les tableaux D.8.1 et D.8.2 en annexe pour respectivement les hommes et les femmes.
Aussi bien pour les hommes que pour les femmes, 45% des mesures d'effet effectuées sur les voyelles sont plus élevées dans la série *Min* que dans la série *Max*. La part de variation des suivis de formant expliquée par le locuteur est équivalente entre les deux séries.
Il est à noter que les parts de variation sont plus importantes que dans notre référence calculée sur l'ensemble des tests. Ce résultat peut être expliqué par la différence en terme de quantité de données analysées.

Différentier *Min* et *Max* par les courbes formantiques

« Est-il possible de différencier *Min* de *Max* à l'aide des suivi de formant ? » est la dernière question à laquelle nous souhaitons répondre. Des MANOVA ont été effectuée pour chaque voyelle et chaque formant où le facteur indépendant est la performance et la variable est le vecteur des 9 valeurs de formants mesurées tous les dix pour-cents.
L'ensemble des résultats est résumé en annexe par le tableau D.9 pour respectivement les hommes et les femmes.
Pour les hommes, aucun formant d'aucune voyelle n'est significativement différent entre la série *Min* et *Max*. Le même résultat est observé pour les femmes excepté pour les troisième ($p < 0.001$) et quatrième ($p < 0.05$) formants du /i/ et le quatrième formant du /œ/. Les effets de la Performance restent tout de même peu importants puisque η^2 est égal à 0.14% pour F3 de /i/, à 0.09% pour F4 de /i/ et à 0.72% pour le F4 de /œ/

Il est donc difficile de distinguer nos deux séries à l'aide des courbes formantiques ainsi que nous l'attendions étant donnés les résultats sur l'ensemble des données de test.

Synthèse du chapitre

La manipulation d'un corpus de grande taille nous a amené à **distinguer la significativité de l'effet d'un facteur définie par la mesure p de la taille de l'effet d'un facteur définie par la mesure η^2**. En effet, au vu de la quantité de données que nous manipulions lors de nos expériences sur l'ensemble des fichiers de test, la significativité de l'effet est toujours avérée. En revanche, les analyse sur *Min* et *Max*, qui réduisent drastiquement le nombre de données accessibles limitent la significativité des effets notamment lorsqu'il s'agit de différencier les deux séries. **C'est la taille de l'effet du facteur qui nous permet de mesurer si une variable en est vraiment dépendante**.
En suivant cette méthodologie, nous avons montré que les indices idiosyncratiques se situent aussi bien au niveau de la **phonation** (fréquence fondamentale moyenne) qu'au niveau de l'**articulation** (centre de gravité et formants).
Nous avons également montré que **tous les phonèmes ne sont pas tous aussi porteurs d'information sur le locuteur**. Par exemple, si nous utilisons les **centres de gravités** comme descripteurs des phonèmes, ce sont les segments **nasals suivis des voyelles orales et des fricatives** qui sont les plus influencés par le locuteur. Au niveau des **voyelles orales, plus une voyelle est fermée et postérieure moins elle porte de l'information sur le locuteur**. Les formants peuvent être également de bons indices sur le locuteur particulièrement les **troisième et quatrième formants**. Mais cette pertinence dépend du timbre de la voyelle car **les voyelles fermées semblent être moins influencées par le locuteur que les autres voyelles**. L'information sur le locuteur n'est donc pas également répartie dans le signal de parole.
Même si nous avons su identifier et caractériser les indices que nous pourrions utiliser dans ce cadre de parole lue, ces indices ne nous permettent pas de distinguer les deux séries de fichiers *Min* et *Max*.

Chapitre 8

Les paramètres utilisés en RAL

Résumé : *Dans ce chapitre nous revenons sur les paramètres cepstraux utilisés par les systèmes automatiques de vérification du locuteur afin de mesurer quelle est l'influence du locuteur sur ces paramètres. Une analyse de la variance des vecteurs de paramètres montre que, pour la quasi totalité des phonèmes, les vecteurs sont différents entre les séries Min et Max. Dans un second temps, nous établissons la part de variance des cepstres qui peut être attribuée au facteur locuteur pour les fichiers de BREF et de NIST 2010. Les catégories fricatives, approximantes et glides ainsi que les nasales semblent être les catégories les plus influencées par les locuteurs. Des expériences à l'aide d'Idento montrent que les performances sont améliorées lorsque seules ces catégories sont utilisées ou lorsque les plosives sont éliminées.*

Sommaire

Jusqu'à présent, les analyses que nous avons menées pour comprendre ce qui amène l'importante différence de performance observée à partir des enregistrements de la série *Min* et de ceux de la série *Max* ont été infructueuses. Nous souhaitons vérifier si les vecteurs de coefficients cepstraux qui sont utilisés pour construire les modèles de locuteur sont différents entre les deux séries. Dans un second temps, nous proposons d'analyser la part de variation des coefficients cepstraux expliquée par le facteur locuteur.

8.1 Peut-on différencier Min et Max à l'aide des LFCC/MFCC ?

8.1.1 Méthode

Les coefficients utilisés par ALIZE/SpkDet sont des LFCC normalisés auxquels sont associés les coefficients Delta et double Delta. Dans notre analyse, nous considérons **ces trois types de paramètres séparément**. Les LFCC retranscrivent de l'information cepstrale tandis que les deltas et les delta-delta reflètent la dynamique de ces coefficients.
Pour chaque phonème, les coefficients extraits de toutes les trames de *Min* et de *Max* sont comparés à l'aide de MANOVA en séparant les hommes et les femmes. Les 20 coefficients correspondant aux LFCC, les 20 Delta et les 20 double Delta sont les variables dépendantes tandis que les groupes *Min* et *Max* sont les facteurs indépendants.
Nous étudions par la suite les vecteurs par coefficients afin d'identifier quels sont les coefficients qui diffèrent entre les deux séries. Nous comparons, pour chaque phonème, chaque coefficient à l'aide d'ANOVA. Le facteur fixé est la performance et la variable dépendante est la valeur de chaque coefficient.

8.1.2 Différence cepstrale en fonction des phonèmes

Mis à part /v/ prononcé par les femmes, les LFCC diffèrent significativement entre *Min* et *Max* pour tous les phonèmes. Les valeurs des Delta diffèrent pour 33% des phonèmes et semblent correspondre entre hommes et femmes excepté pour les consonnes nasales et la majorité des plosives. Les valeurs de delta-delta ne sont pas significativement différentes entre *Max* et *Min* excepté pour /j/ pour les deux genres, /l/ pour les femmes et /a/ et /i/ pour les hommes. Le tableau 8.1 résume les comparaisons

obtenues pour l'ensemble des MANOVA.

Les coefficients cepstraux sont donc différents entre les deux séries étudiées, ceci

	F			M		
Phon.	LFCC	Delta	D-D	LFCC	Delta	D-D
/p/	*****	****	n.s.	**	**	n.s.
/t/	******	**	n.s.	******	****	n.s.
/k/	****	n.s.	n.s.	****	***	n.s.
/b/	******	****	n.s.	******	*	n.s.
/d/	***	**	n.s.	***	*****	n.s.
/g/	******	*	n.s.	******	***	n.s.
/f/	******	**	n.s.	***	n.s.	n.s.
/s/	****	***	n.s.	*****	**	n.s.
/ʃ/	******	n.s.	n.s.	******	n.s.	n.s.
/v/	n.s.	**	n.s.	***	**	n.s.
/z/	***	n.s.	n.s.	******	*	n.s.
/ʒ/	******	n.s.	n.s.	******	n.s.	n.s.
/ʁ/	***	**	n.s.	******	***	n.s.
/l/	***	n.s.	*	******	***	n.s.
/j/	******	*	*	******	n.s.	*
/ɥ/	******	*	n.s.	******	n.s.	n.s.
/w/	******	**	n.s.	****	n.s.	n.s.
/m/	******	****	n.s.	******	*	n.s.
/n/	******	******	n.s.	******	*	n.s.
/ɛ̃/	******	*	n.s.	******	n.s.	n.s.
/ɔ̃/	*	**	n.s.	******	**	n.s.
/ɑ̃/	******	n.s.	n.s.	******	***	*
/i/	*****	**	n.s.	******	*****	**
/y/	****	n.s.	n.s.	******	n.s.	n.s.
/u/	******	n.s.	n.s.	******	*	n.s.
/ø/	******	*	n.s.	******	n.s.	n.s.
/œ/	******	****	n.s.	*	n.s.	n.s.
/e/	******	***	n.s.	******	**	n.s.
/ɛ/	******	n.s.	n.s.	******	****	n.s.
/o/	******	**	n.s.	******	**	n.s.
/ɔ/	******	n.s.	n.s.	***	******	n.s.
/a/	******	*****	n.s.	**	**	n.s.
NS	***	n.s.	n.s.	***	n.s.	n.s.

TABLE ***8.1*** **–** *Significativité des comparaisons des LFCC, Delta et Delta-Delta pour Max vs. Min pour chaque genre ****** : p<.000001 ; ***** : p<.00001 ; **** : p<.0001 ; *** : p<.001 ; ** : p<.01 ; * : p<.05 ; n.s. : non significatif.*

explique en partie pourquoi les performances des modèles diffèrent entre les deux séries.

8.1.3 Différence par coefficient

Selon le phonème étudié, ce ne sont pas les mêmes coefficients qui diffèrent, comme le retranscrit la Figure 8.1. Toutefois, les 20 coefficients sont bien utiles car chacun, en fonction du phonème étudié, permet de différencier *Min* de *Max*.

	LFCC1	LFCC2	LFCC3	LFCC4	LFCC5	LFCC6	LFCC7	LFCC8	LFCC9	LFCC10	LFCC11	LFCC12	LFCC13	LFCC14	LFCC15	LFCC16	LFCC17	LFCC18	LFCC19	LFCC20
/a/	0,2937	0,1801	0,5757	0,3314	0,0630	0,3896	0,0388	0,1799	0,7264	0,3103	0,0269	0,9296	0,8140	0,1298	0,0036	0,1305	0,0569	0,8010	0,7640	0,1342
/e/	0,2475	0,1163	0,8964	0,6240	0,0232	0,0000	0,2141	0,9676	0,6927	0,0118	0,1274	0,1255	0,0988	0,7759	0,0924	0,4950	0,3997	0,0019	0,0296	0,5768
/o/	0,1997	0,1608	0,0007	0,4495	0,0130	0,0031	0,0389	0,6683	0,0481	0,2383	0,0000	0,0290	0,4385	0,0170	0,0491	0,1348	0,8290	0,0621	0,0020	0,1539
/ɛ/	0,1486	0,0185	0,0012	0,0670	0,8934	0,5313	0,0102	0,3496	0,0117	0,5801	0,6200	0,1950	0,0035	0,3849	0,3795	0,6145	0,1830	0,0003	0,0028	0,0971
/ø/	0,1657	0,8694	0,0085	0,0811	0,7700	0,2728	0,0810	0,0020	0,0016	0,0073	0,3404	0,0080	0,0355	0,9858	0,8760	0,5732	0,6680	0,0501	0,0836	0,0002
/i/	0,0006	0,0000	0,0000	0,0867	0,7667	0,1813	0,0002	0,0000	0,0000	0,4877	0,0123	0,2815	0,6328	0,0468	0,0259	0,5841	0,1331	0,4917	0,0597	0,6454
/œ/	0,0450	0,6070	0,0002	0,4011	0,1345	0,1368	0,5058	0,3483	0,1651	0,1181	0,4331	0,2226	0,9350	0,9803	0,2208	0,6323	0,9669	0,1727	0,3497	0,0806
/ɔ/	0,8625	0,3940	0,2920	0,9206	0,8510	0,1918	0,0012	0,3261	0,6752	0,0858	0,4184	0,3017	0,3268	0,0000	0,8167	0,3268	0,8313	0,3163	0,2349	0,9089
/u/	0,4398	0,0040	0,0698	0,0264	0,5987	0,2128	0,3508	0,0339	0,2249	0,0000	0,5828	0,3331	0,0107	0,1462	0,5367	0,0020	0,0135	0,6378	0,2424	0,6151
/y/	0,6891	0,3313	0,8629	0,0009	0,8389	0,0000	0,2381	0,0058	0,0566	0,7960	0,0122	0,9294	0,6411	0,1120	0,0138	0,9402	0,0025	0,4295	0,2195	0,3141
/ɑ̃/	0,0151	0,4900	0,0945	0,9923	0,1679	0,0574	0,4249	0,2695	0,1782	0,5884	0,0831	0,1476	0,5831	0,0338	0,0000	0,1358	0,0003	0,9480	0,0043	0,3062
/ɔ̃/	0,2512	0,1346	0,7433	0,9629	0,1196	0,9817	0,0000	0,6598	0,0066	0,0000	0,0275	0,1052	0,2261	0,3695	0,1989	0,0091	0,0527	0,7732	0,3127	0,2321
/œ̃/	0,0766	0,1554	0,3575	0,0033	0,3147	0,5920	0,6763	0,1766	0,1287	0,1072	0,0021	0,0024	0,0012	0,0119	0,4777	0,0520	0,0028	0,0460	0,0028	0,1282
/ɛ̃/	0,1241	0,5656	0,0008	0,4199	0,5476	0,3643	0,9452	0,1293	0,0369	0,0409	0,3713	0,9031	0,4154	0,3136	0,0000	0,6317	0,7549	0,0912	0,0471	0,0000
/p/	0,9448	0,0336	0,4835	0,1195	0,7864	0,0604	0,3324	0,7879	0,8137	0,1945	0,0604	0,6632	0,4446	0,4298	0,5361	0,1877	0,0017	0,4181	0,3073	0,5032
/b/	0,4183	0,0595	0,3641	0,0372	0,2635	0,1190	0,1734	0,1367	0,0041	0,0000	0,0001	0,9433	0,0044	0,7370	0,6795	0,0000	0,4882	0,7521	0,1894	0,0629
/m/	0,0205	0,0666	0,1173	0,9277	0,0697	0,7124	0,6316	0,1183	0,0314	0,0002	0,8244	0,1495	0,7867	0,4218	0,6197	0,0310	0,4805	0,0025	0,7196	0,5775
/t/	0,0442	0,2762	0,0000	0,4421	0,1792	0,0067	0,1721	0,8910	0,0188	0,0121	0,2490	0,9657	0,5953	0,1087	0,0044	0,2590	0,4537	0,1156	0,9670	0,1505
/d/	0,3595	0,2756	0,8818	0,0369	0,1363	0,3950	0,0555	0,0027	0,0209	0,6666	0,0030	0,3290	0,0381	0,1464	0,3412	0,6512	0,0298	0,2595	0,2971	0,8129
/n/	0,0041	0,0147	0,1510	0,8459	0,0003	0,2375	0,0196	0,0000	0,3932	0,0668	0,0362	0,9852	0,6697	0,8394	0,2804	0,0005	0,0000	0,0252	0,0003	0,0033
/k/	0,0060	0,1729	0,2732	0,0027	0,6412	0,2527	0,3059	0,1126	0,6483	0,9971	0,5759	0,0033	0,4321	0,2352	0,6354	0,0029	0,0210	0,1085	0,3482	0,4515
/g/	0,2758	0,9617	0,0373	0,4957	0,5688	0,2257	0,0137	0,3221	0,3205	0,0287	0,5568	0,3901	0,4842	0,0000	0,1186	0,2563	0,0258	0,0906	0,2912	0,4757
/f/	0,2587	0,9281	0,6262	0,0205	0,3949	0,9484	0,5686	0,7049	0,8993	0,4357	0,1122	0,0202	0,4898	0,2170	0,5050	0,0562	0,1377	0,0322	0,5482	0,8286
/v/	0,7220	0,1869	0,7692	0,0000	0,0379	0,1338	0,1806	0,2172	0,2717	0,1968	0,9830	0,9274	0,0442	0,8544	0,6854	0,5120	0,0311	0,8037	0,1448	0,1606
/s/	0,8424	0,0046	0,2768	0,9391	0,0582	0,0994	0,9804	0,7342	0,6226	0,1157	0,3527	0,0106	0,4847	0,2346	0,6621	0,9809	0,7891	0,0817	0,7891	0,0185
/z/	0,4650	0,8611	0,0008	0,0366	0,0006	0,0000	0,0003	0,9858	0,3299	0,5497	0,8718	0,0697	0,0099	0,0154	0,5255	0,2163	0,0119	0,9332	0,7624	0,1047
/ʃ/	0,0137	0,1817	0,2008	0,8052	0,9090	0,0000	0,6064	0,5794	0,0155	0,0097	0,9980	0,5981	0,0000	0,0008	0,0011	0,0064	0,1454	0,8280	0,8110	0,4799
/ʒ/	0,5072	0,8251	0,0001	0,0001	0,2932	0,5610	0,4494	0,1732	0,0126	0,9293	0,6602	0,1329	0,0682	0,0297	0,2154	0,5483	0,3640	0,2310	0,2394	0,0001
/l/	0,0305	0,0001	0,2335	0,1745	0,0711	0,0000	0,0000	0,1094	0,4443	0,2607	0,2641	0,0989	0,0558	0,0059	0,5157	0,0915	0,0221	0,0000	0,0409	0,7294
/ʁ/	0,0107	0,0000	0,0015	0,0205	0,0126	0,0038	0,0312	0,5722	0,6093	0,2134	0,0004	0,0001	0,0258	0,2412	0,7437	0,5212	0,6216	0,4745	0,1711	0,4934
/w/	0,7560	0,0880	0,1220	0,4409	0,7060	0,0008	0,0091	0,0677	0,0000	0,8042	0,2431	0,4513	0,4980	0,0014	0,5450	0,1457	0,3262	0,0191	0,4123	0,5828
/j/	0,3695	0,9965	0,2185	0,1190	0,5034	0,4438	0,0134	0,0315	0,1775	0,7706	0,1108	0,0010	0,0360	0,0086	0,0176	0,0000	0,0001	0,0051	0,0002	0,8120
/ɥ/	0,0122	0,3467	0,0193	0,0135	0,5979	0,0651	0,0071	0,0646	0,6889	0,1601	0,2606	0,0715	0,0081	0,0287	0,1257	0,2009	0,0003	0,0046	0,7236	0,1957
pause	0,6776	0,2035	0,0181	0,0218	0,0091	0,5950	0,0192	0,9023	0,2714	0,2791	0,0162	0,0749	0,2907	0,9962	0,0516	0,5909	0,1837	0,5916	0,0034	0,7376

FIGURE 8.1 – *p value évaluant la significativité des différences pour chaque coefficient LFCC en fonction du phonème pour les séries Min-femmes et Max-femmes*

Les coefficients cepstraux employés dans *Min* et dans *Max* semblent différés les uns des autres. Reste à savoir quelle part de variation de ces valeurs cepstrales peut être expliquée par le facteur locuteur.

8.2 La part de variation des valeurs de MFCC expliquée par le locuteur varie-t-elle en fonction des segments ?

8.2.1 Élargissement de la question

Les coefficients cepstraux, qu'ils soient des LFCC ou des MFCC, sont les paramètres les plus couramment utilisés en vérification du locuteur. Aussi bien ALIZE/SpkDet que

Idento utilisent ces coefficients.
Il nous a semblé intéressant d'**étendre l'analyse de ces coefficients à l'anglais**, qui est la langue majoritaire des campagnes NIST. Cette étude préliminaire nous permettra de **vérifier si ce sont les mêmes classes de phonèmes qui sont porteuses d'information sur le locuteur pour deux langues différentes**.

8.2.2 Méthode

Corpus

Comme pour les autres indices, nous souhaitons rendre compte de la part de variation des coefficients cepstraux qui peut être attribuée au locuteur. Nous avons, dans ce cas, travaillé avec à la fois les données de BREF et des données issues des campagnes de NIST.
Les données de NIST sont issues de la campagne d'évaluation 2010 car nous souhaitons par la suite vérifier nos analyses par un test effectué sur les données de NIST 2008. Tous les fichiers ont une durée moyenne de 2 minutes 30 secondes. Les coefficients cepstraux sont extraits pour l'ensemble des fichiers et sont étiquetés, à l'aide du système de transcription du SRI (Shriberg et Stolcke, 2011). Pour l'anglais, la base de données se compose de 14 139 fichiers prononcés par 194 hommes.

Analyse

L'analyse de la part de variance des coefficients cepstraux expliquée par le facteur locuteur est effectuée pour chacune des deux langues séparément. Pour chaque langue, des MANOVA à mesures répétées sont effectuées pour chaque phonème. La variable dépendante est un vecteur de 20 coefficients cepstraux et le facteur indépendant est le locuteur. Les η^2 sont calculés pour chaque phonème afin d'évaluer l'effet du facteur locuteur sur les vecteurs de coefficients cepstraux. La moyenne de η^2 est calculée par classes phonémiques en regroupant les phonèmes selon leur mode articulatoire. Ce sont ces résultats qui sont retranscrits dans ce manuscrit.

8.2.3 Effet du locuteur sur les coefficients cepstraux

Un effet du locuteur est observé aussi bien pour la base de données NIST que pour BREF ($p < 0.001$). L'ensemble des résultats par classes phonémiques est présenté dans le tableau 8.2. Le lecteur trouvera les résultats par phonème à l'annexe E.

Les diphtongues et les glides sont la classe phonémique qui semble la plus influencée par le locuteur ($\eta^2_{NIST} = 39.8\%$ et $\eta^2_{BREF} = 12.5\%$). Suivent les fricatives ($\eta^2_{NIST} = 22.6\%$ et $\eta^2_{BREF} = 8.0\%$) et les nasales ($\eta^2_{NIST} = 18.2\%$ et $\eta^2_{BREF} = 13.4\%$). Arrivent enfin les voyelles orales ($\eta^2_{NIST} = 10.8\%$ et $\eta^2_{BREF} = 6.7\%$) et les plosives ($\eta^2_{NIST} = 9.9\%$ et $\eta^2_{BREF} = 4.7\%$).

Les catégories phonémiques les plus influencées par le locuteur sont les mêmes pour les deux langues excepté une inversion entre les nasales et les fricatives. Il est toutefois à noter que la catégorie nasale de BREF comporte des voyelles et des consonnes alors que pour NIST, seules des consonnes sont présentes, l'anglais n'ayant pas de voyelles nasales.

La parole dans NIST étant conversationnelle, des rires ont été identifiés par le système de transcription. Les trames caractérisées comme du rire ont un η^2 de 21.7%. Cet η^2 est très élevé et montre que le rire peut être un indicateur pertinent pour discriminer les locuteurs.

Les pauses ont un η^2 pour NIST de 15.2% et de 5.7% pour BREF. Ces valeurs sont assez élevées, elles peuvent peut-être être expliquées par la méthode de caractérisation des pauses.

Parmi les catégories les plus influencées par le locuteur, nous retrouvons les fricatives et les nasales. Les glides et les diphtongues font leur apparition. Ces éléments sont très influencés par la co-articulation et révèlent peut-être des stratégies articulatoires propres aux locuteurs révélées par les coefficients cepstraux. Les fricatives mises en évidence pour les centres de gravité, le sont également pour les coefficients cepstraux. En revanche, nous ne retrouvons pas les différences entre voyelles hautes et les autres voyelles.

L'analyse des coefficients cepstraux montre que les glides et diphtongues, les fricatives et les nasales sont les catégories phonémiques les plus influencées par le locuteur, tandis que la variation des plosives n'est que très peu expliquée par le facteur locuteur. Ces observations se retrouvent-elles dans les résultats d'un système de RAL en fonction des trames utilisées pour la modélisation ?

Catégorie	Conditions	
	Anglais conversationnel (NIST 2010)	Français lu (BREF)
Diphtongues et glides	39.8%	12.5%
Nasales	18.2%	13.4%
Fricatives non voisées	16.7%	9.3%
Fricatives voisées	29.7%	7.6%
Voyelles orales hautes	10.5%	8.1%
Autres voyelles orales	10.5%	7.3%
Plosives non voisées	9.9%	4.7%
Plosives voisées	9.9%	4.7%
Rires	21.7%	non mesuré
Pauses	15.2%	5.7%

TABLE 8.2 – *Proportion de variance des valeurs de MFCC par type de phonème expliquée par le facteur locuteur pour tous les phonèmes* $p < 0.001$

8.3 η^2 permet-il de prédire les phonèmes pertinents ?

8.3.1 Objectifs

L'analyse des effets du locuteur sur les coefficients cepstraux met en évidence que certaines catégories phonémiques sont plus influencées par le locuteur que d'autres. Nous souhaitons **vérifier si cette information peut être utile pour les systèmes automatiques**.

8.3.2 Méthode

L'ensemble des comparaisons effectuées pour vérifier la pertinence de nos mesures de η^2 correspond à la cohorte de développement du SRI issue de NIST 2008. Cette cohorte se compose de 395 comparaisons cible et 394 211 comparaisons imposteur. Les fichiers sont tous issus de conversations **téléphoniques en anglais**. Pour vérifier la pertinence des mesures de η^2, plusieurs expériences à l'aide d'Idento ont été conduites en utilisant toujours cette cohorte extraite de NIST-08 comme élément de comparaison.
L'expérience témoin correspond à la sélection de toutes les trames de parole où les rires et les pauses ont été exclues.
La première expérience consiste à ne **sélectionner, pour construire le modèle de locuteur que les vecteurs acoustiques des catégories phonémiques dont η^2 est le plus**

élevé, à savoir, les glides ou diphtongues, les nasales et les fricatives. Cette série est nommée **FGDN**. Nous avons, par ailleurs, **construit des modèles à partir des mêmes fichiers en sélectionnant aléatoirement le même nombre de trames que celles sélectionnées dans FGDN pour chaque fichier**. Ceci nous permet de vérifier si la sélection opérée connaît de meilleurs résultats qu'une sélection aléatoire des trames.
La seconde expérience consiste à **sélectionner tous les vecteurs acoustiques de parole exceptés les vecteurs acoustiques des catégories qui obtiennent le η^2 le plus élevé**, c'est-à-dire toutes les trames de parole que nous n'avions pas utilisées dans FGDN. Cette série est appelée **-FDGN**. Ici encore, une **sélection aléatoire est également menée de manière à comparer les résultats obtenus avec -FGDN** à ceux obtenus avec une sélection aléatoire.
La troisième expérience consiste à **sélectionner tous les vecteurs acoustiques de parole exceptés les vecteurs acoustiques des catégories qui obtiennent le η^2 le plus bas**, à savoir les plosives. Cette série est appelée **TSP**. Là encore, **cette série est comparée avec une série construite avec les mêmes fichiers dont sont extraits aléatoirement** exactement le même nombre de trames par fichier que TSP. Nous comparons **les taux d'EER** obtenus pour chaque expérience de manière à vérifier si le η^2 est un bon estimateur de la quantité d'information sur le locuteur.

8.3.3 Résultats

Lorsque toutes les vecteurs acoustiques de parole sont sélectionnés, l'EER est de 3.3% avec en moyenne 16 323 vecteurs acoustiques sont sélectionnés pour construire un modèle. Cette valeur est notre référence.

FGDN vs -FGDN

Lorsque les nasales, les fricatives, les glides et les nasales sont sélectionnées, les modèles sont construits à partir de 4 030 vecteurs acoustiques en moyenne. Dans ce cas, l'EER est de 4.5%. Dans le même temps, en sélectionnant aléatoirement 4 030 vecteurs, le système obtient un EER de 5.5%.
Les résultats obtenus en sélectionnant les vecteurs considérés comme les plus influencés par le locuteur sont meilleurs que lorsque les vecteurs acoustiques sont sélectionnés aléatoirement. Ceci nous encourage à considérer que notre sélection est per-

tinente.
Par ailleurs, lorsque tous les vecteurs acoustiques de parole sont sélectionnés exceptés les vecteurs acoustiques catégorisés comme nasales, glides, diphtongue ou fricatives, un EER de 4.5% est observé alors que 12 293 vecteurs en moyenne sont sélectionnés. En sélectionnant aléatoirement le même nombre de vecteurs, un EER de 5.0% est obtenu. **Avec un nombre de vecteurs 4 fois plus important, nous observons une performance moyenne équivalente en sélectionnant les vecteurs acoustiques dont le η^2 est le plus élevé. Cette expérience confirme que notre sélection est pertinente.**

TSP

En sélectionnant tous les vecteurs acoustiques exceptés ceux caractérisés comme plosive, l'EER est de 3.5% alors que 13 239 vecteurs acoustiques sont sélectionnés. En utilisant le même nombre de vecteurs acoustiques, mais en les sélectionnant aléatoirement, l'EER passe à 5.0%.
Alors que le résultat obtenu sans les plosives est comparable à celui obtenu en sélectionnant tous les vecteurs acoustiques de parole, le résultat obtenu en sélectionnant aléatoirement les vecteurs acoustiques est plus élevé. Les plosives ne semblent pas apporter d'information supplémentaire sur le locuteur comme l'indiquait le η^2. L'ensemble des résultats obtenus est résumé par le tableau 8.3.

Conditions	Nombre de trames	EER
Toutes les trames de parole (pas les rires)	16 323	3.3%
Fricatives, Glides, diphtongues et nasales (FGDN)	4 030	4.5%
Même nombre de trames que FGDN sélectionnées aléatoirement	4 030	5.5%
Tout sauf les segments de FGDN (-FGDN)	12 293	4.5%
Même nombre de trames que -FGDN sélectionnées aléatoirement	12 293	5.0%
Toutes les trames sauf les plosives (TSP)	13 239	3.5%
Même nombre de trames que TSP sélectionnées aléatoirement	13 239	5.0%

TABLE *8.3 – EER obtenus selon les trames utilisées pour le système Idento en condition téléphonique*

Cette série d'expérience montre que η^2 est une mesure qui peut être retenue pour caractériser les vecteurs acoustiques afin de connaître leur pertinence pour discrim-

iner le locuteur. Il est toutefois à noter que les meilleurs résultats sont obtenus en sélectionnant toutes les trames. La quantité d'information est donc un facteur essentiel. Cependant, certaines trames n'apportent que peu d'information.

Synthèse du chapitre

Les **coefficients cepstraux sont différents entre les séries *Min* et *Max***. Les coefficients pourraient donc peut-être être utilisés pour caractériser la pertinence d'un enregistrement pour reconnaître le locuteur.
Lorsqu'elles sont représentées par des coefficients cepstraux, les **catégories phonémiques ne sont pas toutes influencées par le locuteur de la même manière que ce soit pour l'anglais ou le français**. Les **diphtongues, les nasales et les fricatives** sont les catégories pour lesquelles l'effet du locuteur est le plus important. D'autre part, les **plosives ne sont que peu porteuses d'indices sur le locuteur.** La validité de prédiction des η^2 est confirmée par une série d'expériences où les performances du système ne sont que peu affectées alors que certaines catégories phonémiques sont absentes.

Quatrième partie

Conclusions et Perspectives

Contributions

L'évaluation est un outil incontournable pour faire avancer la recherche en ouvrant de nouvelles perspectives. En vérification du locuteur, d'important efforts ont été consacrés à la question de la robustesse des systèmes aux modes d'enregistrement ou à la durée des signaux qui sont des questions centrales de NIST-SRE. En 2010, les meilleurs systèmes étaient en dessous de 2% d'EER pour la condition avec 2 minutes 30 de signal en apprentissage et en test (Greenberg et al., 2011a). Le plan d'évaluation a aussi un rôle fondamental dans l'analyse des performances en vérification du locuteur. Dans le cadre de NIST-SRE, la métrique des performances consiste à comptabiliser le nombre de fois où le système s'est trompé indépendamment de l'auteur de l'extrait de parole. L'influence du locuteur sur la difficulté de reconnaissance n'est donc pas prise en compte. D'autre part, l'influence des extraits de parole utilisés pour chaque locuteur n'est pas non plus étudiée.

Les protocoles d'évaluation que nous avons mis en œuvre **prennent en compte la variabilité intra-locuteur des extraits de parole et évaluent la performance pour chaque locuteur**. Nos protocoles sont élaborés afin de répondre à la question scientifique posée en introduction : **Tous les extraits de parole d'un même locuteur sont-ils équivalents pour le reconnaître ?** D'autres questions découlent de cette première interrogation :

- Quelles sont les sources de variabilité qui influencent la capacité à reconnaître une personne à partir d'un enregistrement de parole ?
- Quelle confiance peut être attribuée à la réponse fournie ?

Tout au long de ce travail, nous avons mis en œuvre des protocoles qui nous permettent d'apporter des éléments de réponses à ces questions.

Perception humaine : des difficultés pour établir une mesure de confiance

La tâche proposée lors de HASR est très difficile puisqu'un seul des participants de la campagne HASR fait réellement mieux que le hasard. Certains participants développent des stratégies de réponse, à savoir répondre très souvent « oui » ou « non ».Lors de notre participation à HASR, **nous avons développé deux approches pour établir la confiance qu'il est possible d'attribuer à la réponse fournie**. La première consiste à demander aux auditeurs d'accompagner leur réponse d'un score de confiance (**auto-évaluation**). La seconde revient à **mesurer l'accord entre les auditeurs** et à considérer que plus les auditeurs sont unanimes, plus la confiance dans la réponse est élevée.
Dans le premier cas, nous avons comparé les taux de réussite globaux, les FA et les FR pour, d'une part, la cohorte où les auditeurs avaient en moyenne un score de confiance supérieur à 2.5 (cohorte *confiance plus haute*) et d'autre part la cohorte où les auditeurs avaient indiqué en moyenne des scores de confiance inférieur à 2.5 (cohorte *faible confiance*). Le taux de réponses correctes est de 62% pour la cohorte *confiance plus haute* contre 57% pour la cohorte *faible confiance*. Pourtant, le FR de la cohorte *faible confiance* est de 32% tandis que celui de la cohorte *confiance plus haute* est de 43%. En comparaisons cible, le score de confiance des auditeurs en leur réponse n'est pas un gage de qualité de la réponse car ils se trompent plus souvent quand ils ont confiance dans leur réponse. Le FA de la cohorte *faible confiance* est de 51% tandis que celui de la cohorte *confiance plus haute* est de 36%. En comparaison imposteur, le ressenti des auditeurs semble être plus indicatif.
Dans le second cas, nous avons comparé les taux de réussite globaux, les FA et les FR de la cohorte pour laquelle tous les auditeurs avaient répondu de la même manière (cohorte *unanime*) et ceux de la cohorte où ils n'étaient pas tous d'accord (cohorte *désaccord*). Le taux de réussite global de la cohorte *unanime* est de 49% tandis que celui de la cohorte *désaccord* est de 68%.
Dans le cadre difficile qu'est HASR, la confiance dans les réponses des auditeurs ne peut se faire ni par mesure majoritaire ni par un score de confiance fourni par l'auditeur. Nous proposons d'expliquer ces résultats par la **différence de conditions d'enregistrement**. En effet, en comparaison imposteur, la distance entre les indices idiosynchratiques des deux locuteurs peut être renforcée par la différence de canal. Au contraire, en comparaison cible, cette dernière peut créer l'impression de personnes distinctes alors qu'il n'en est rien.

Par ailleurs, les résultats obtenus par les auditeurs expérimentés et non expérimentés ne différant pas du hasard dans ce cadre expérimental, **nous ne pouvons pas faire plus confiance à des auditeurs expérimentés qu'à des naïfs**. Des **stratégies individuelles** sont observées chez nos auditeurs quelque soit leur catégorie. Seul quatre auditeurs (non-expérimentés) ont des réponses significativement différentes du hasard.

Systèmes automatiques : des écarts de performance fonction de l'extrait de parole du locuteur

Tous les enregistrements d'un même locuteur ne sont pas équivalents pour le modéliser : suivant le choix de l'enregistrement de référence, la performance des systèmes varie de 1% d'EER à 30% pour les mêmes locuteurs et les mêmes fichiers test. Cette variation, mesurée à l'aide d'un taux de variation relatif, Vr, est similaire pour un système fondé sur l'approche UBM-GMM-FA ou pour un système fondé sur l'approche i-vector (pour NIST, $Vr_{Idento} = 1.41$ et $Vr_{ALIZE/SpkDet} = 1.47$). La normalisation des scores ne semble pas permettre de limiter la variabilité des performances. **Un des apports de cette thèse est de montrer l'absolue nécessité de distinguer locuteur et extrait de parole**, distinction qui n'est pas faite au cours des évaluations NIST.

Cette variabilité intra-locuteur dépend de la base de données. Ainsi, **plus les sources de variabilité sont nombreuses dans les enregistrements, moins la variation relative de performance est importante** (pour ALIZE/SpkDet, $Vr_{NIST} = 1.47$ et $Vr_{BREF} = 3.11$). Ce résultat met en avant **l'importance du contexte d'énonciation pour rechercher le locuteur**. Un locuteur ne fait pas varier sa parole de la même manière en fonction des circonstances. **Plus la parole est contrôlée, plus le locuteur est contraint dans son énonciation, plus le risque de ne pas trouver d'indices discriminants pour le locuteur est grand**.

La longueur des signaux permet de limiter les erreurs commises (pour les hommes de BREF, $EER_{30s} = 33\%$ et $EER_{2.5minutes} = 5.3\%$ pour les séries *Max*). L'augmentation de la durée des signaux d'apprentissage permet une diminution importante de la variabilité des performances.

Nous avons mis en évidence que **la variation de performance n'est pas uniquement un problème de différence de mode d'enregistrement, il existe dans la parole une variation intrinsèque qui fait que le locuteur est plus ou moins présent au sein des extraits de parole**. Ces résultats montrent qu'**il est essentiel de faire la différence entre enregistrement et locuteur et de ne pas se contenter d'une mesure moyenne de la performance**. Il est indispensable de prendre en compte cette variabilité intra-locuteur qui vient perturber les performances moyennes. **L'objectif est non seulement d'obtenir de meilleures performances moyennes mais également de minimiser l'impact de la variabilité intra-locuteur.**

La performance seule n'est pas suffisante, elle doit être accompagnée d'une mesure de confiance fondée sur l'acoustique qui nous permette d'expliquer (et de prédire) les performances obtenues à partir des enregistrements.

Des indices idiosyncratiques inégalement répartis dans la parole

Pour apporter des éléments de réponse nous avons, dans la partie III, **procédé à une analyse fine et détaillée des différents indices idiosyncratiques identifiés dans la littérature**. Cette analyse a été menée sur la base de données BREF qui regroupe plus de 21 heures de parole, 64 locutrices et 47 locuteurs. Les conditions d'enregistrement sont excellentes et permettent une analyse acoustique semi-automatique.

La manipulation d'un corpus de grande taille nous a amené à **distinguer la significativité de l'effet d'un facteur (p) de la taille de l'effet de ce facteur (η^2)**. En effet, au vu de la quantité de données que nous manipulons, la significativité de l'effet est toujours avérée. C'est la taille de l'effet qui nous permet d'établir si une variable est vraiment dépendante d'un facteur. La pertinence d'un indice idiosyncratique est donc mesurée à l'aide de la valeur de η^2 sur la base de données BREF.

Les indices étudiés pour discriminer le locuteur relèvent de la phonation et/ou de

l'articulation :

- Phonation : fréquence fondamentale, jitter, shimmer.
- Articulation : fréquence phonémique, centre de gravité, quatre premiers formant, aire du triangle vocalique, locus, suivis formantiques.

Concernant les indices relatifs à la phonation, le facteur locuteur explique 18.9% pour les hommes et de 28.4% pour les femmes de la variation de la fréquence fondamentale, et ce indépendamment de la voyelle.
Pour les indices relatifs à l'articulation, les η^2 les plus élevés sont observés pour **le centre de gravité de certains phonèmes et les troisième et quatrième formants des voyelles orales**. Les centres de gravités peuvent être mesurés pour tous les segments contrairement aux formants. Les segments les plus discriminants pour le locuteur sont les **consonnes nasales** ($\eta^2 = 39.7\%$), les **fricatives** ($\eta^2 = 23.0\%$), les **voyelles nasales** ($\eta^2 = 48.0\%$), les **voyelles orales mi-fermées à ouvertes** (η^2 de l'ordre de 40%). L'influence du locuteur sur les mesures formantiques est plus important pour les voyelles mi-fermées à ouvertes que pour les voyelles fermées. Les η^2 obtenus pour F3 sont de l'ordre de 16% pour les voyelles fermées alors qu'ils atteignent 39% pour les autres voyelles.
L'information sur le locuteur n'est pas également répartie dans le signal de parole, cette information est structurée. Certains phonèmes sont plus porteurs que d'autres d'information sur le locuteur.

Le rôle de ces phonèmes se retrouve lors de l'analyse des vecteurs acoustiques, sur BREF comme sur NIST. Si la hiérarchie est globalement respectée, les valeurs des η^2 diffèrent en fonction de la base de données. Ces différences de valeurs peuvent être attribuées à la différence de langue et/ou de contexte d'énonciation. D'ailleurs, **lorsque seuls les éléments qui sont les plus discriminants pour le locuteur sont utilisés pour construire un modèle de locuteur, les performances globales du système avec ces vecteurs acoustiques sélectionnés sont meilleures qu'avec des vecteurs acoustiques sélectionnées aléatoirement**. De plus, elles ne varient pas si les vecteurs acoustiques correspondant aux plosives, pour lesquelles l'effet du locuteur est le plus faible, sont ignorés.

En mettant en évidence que tous les enregistrements de parole ne sont pas équivalents pour modéliser le locuteur et en montrant que l'information sur le locuteur est structurée, nous avons proposé des premières pistes pour établir une mesure de confiance.[1cm] **Ce travail analytique, mené à l'aide de systèmes automatiques, est une donc première pierre pour comprendre la structuration de la variation de la parole. Ce thème est sans doute le défi majeur à relever pour tous les acteurs du monde de l'authentification vocale qui souhaiteraient mettre évidence la fiabilité des approches et non leur performance brute.**[1cm]

Perspectives

Le travail que nous avons mené connaît certaines limites et aurait pu explorer d'autres pistes.

Une première limite méthodologique à laquelle nous avons été confronté est que le nombre de comparaisons cible est toujours bien inférieur à celui des comparaisons imposteur. La base de données NIST contient de nombreux locuteurs mais peu d'extraits de parole par locuteur. A l'inverse, la base de données BREF nous permet d'obtenir 40 extraits de 30 secondes de parole par locuteur mais ne comporte que 47 hommes et 64 femmes. Il serait intéressant d'**étudier l'influence du nombre de locuteurs étudiés sur les performances, en n'omettant pas l'influence capitale du choix de l'extrait de parole.**

Les extraits de parole de BREF sont enregistrés dans un contexte très contrôlé de parole lue, limitant de fait la variabilité. L'objectif est de décorréler les informations propres au locuteur du contexte d'énonciation. Il est alors impossible d'**étudier l'impact des différents contextes d'énonciation dans le cadre de la reconnaissance du locuteur**. L'idéal serait d'enregistrer les locuteurs dans des contextes d'énonciation variés comme l'a entrepris Campbell *et al* (Campbell et Erickson, 2004) en suivant une locutrice durant trois ans. Les récents développements des évaluations NIST, comme l'introduction de l'effort vocal ou les réenregistrements de locuteurs déjà présents dans la base de données, permettront d'avoir plus d'extraits de parole pour un même locuteur dans des contextes plus variés que la lecture de BREF.

Une autre limite importante de ce travail est le nombre restreint de locuteurs étudiés

qui ne nous permet pas de rendre compte de la diversité des voix des locuteurs, impliquant une variabilité inter-locuteur relativement faible. Par ailleurs, en sélectionnant aléatoirement les personnes enregistrées, nous ne contrôlons pas **la proximité des voix des personnes étudiées. Il serait intéressant de se focaliser sur des voix considérées comme proches** : voix de frères, de pères et fils, d'amis d'enfance, collègues de travail.

Par ailleurs, dans ce travail, nous ne nous sommes intéressé qu'aux fichiers d'apprentissage en travaillant systématiquement avec les mêmes fichiers de tests. **L'influence de la paire de comparaisons n'a pas été abordée et reste une question ouverte**.

Au final, il s'agit de bien connaître la constitution des bases de données utilisées en évaluation de manière à pouvoir diagnostiquer là où le système commet des erreurs et comprendre ses difficultés.

Concernant les tests perceptifs, nous avons été confronté dans ce travail à la difficulté de trouver des auditeurs qui acceptent de passer plus d'une heure et demie à discriminer des voix. Par ailleurs, ces tests se limitent à des auditeurs francophones pour des stimuli en anglais, alors que nous savons que la langue a un rôle prépondérant dans l'identification des locuteurs. Un autre inconvénient est qu'il est difficile d'extraire de ces tests perceptifs des informations sur les paramètres acoustiques qui sous-tendent les décisions des auditeurs. Les critères de sélection des comparaisons de la campagne HASR sont en effet très défavorables à une analyse acoustique.

Dans cette thèse, l'utilisation d'outils de traitement automatique de la parole nous a permis de traiter plusieurs dizaines d'heures de parole. Nous avons réussi à mieux comprendre où se situe l'information pertinente sur le locuteur, il faut **continuer à exploiter les systèmes automatiques pour mieux définir les possibles paramètres acoustiques pertinents avant de les confirmer par des tests perceptifs si coûteux à mettre en place. Les systèmes automatiques sont de bons outils pour comprendre la parole et nous aider à caractériser les informations pertinentes dans les signaux de parole.**

Si dans notre analyse des indices, nous avons étudié un grand nombre de mesures souvent relatives à l'articulatoire, nous aurions pu affiner nos mesures sur la phonation en analysant par exemple les courbes de fréquence fondamentale ou des mesures de rythme et de qualité de voix. Un vaste champ reste à explorer étant donné que nous avons concentré notre analyse sur les segments. Les hypothèses invoquées pour expliquer pourquoi certains segments sont plus porteurs d'informations idiosyncratiques relèvent souvent de différences morphologiques : forme des fosses nasales, forme du palais dur et des dents, longueur des plis vocaux. Il est important de **vérifier ces hypothèses par des mesures articulatoires** qui permettraient de prédire les variations observées à partir de simulations par des conduits vocaux adaptés au locuteur.
Une source de variation idiosyncratique pourrait relever non pas de paramètres morphologiques mais de la **culture, de la sphère sociale et de la personnalité de cet individu. Ces éléments doivent être pris en compte et leur influence étudiée**, d'autant plus lorsque les futures applications de la reconnaissance du locuteur s'inscrivent dans la problématique des environnements intelligents où le système devra reconnaître le locuteur dans différentes circonstances communicationnelles.
Une autre question est celle de la langue parlée. **Les indices idiosyncratiques sont-ils les mêmes en fonction des langues ?** Qu'arrive-t-il si le système vocalique de la langue étudiée connaît moins de timbres ? Comment interagit une langue tonale avec les indices idiosyncratiques portés par la fréquence fondamentale ?

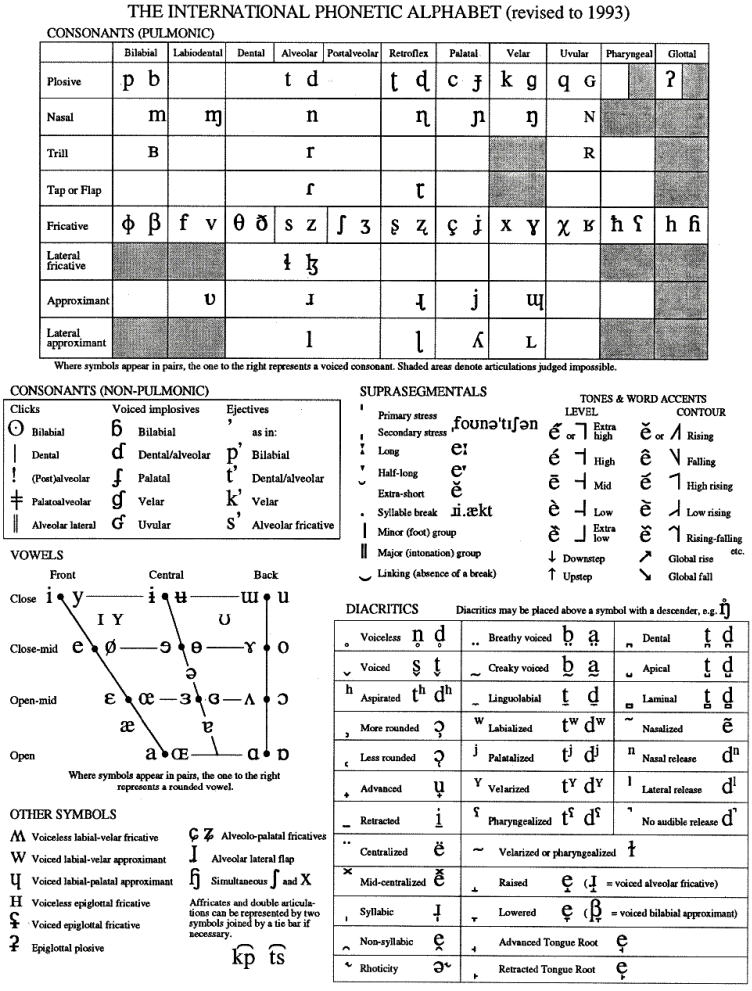
Cinquième partie

Annexes

Annexe A

Alphabet Phonétique International

Alphabet permettant de représenter l'ensemble des sons des langues du monde.



Annexe B

HASR

B.1 Performances des participants à HASR

Ce tableau indique les performances obtenues par tous les participants à HASR. Seul un participant a un taux de FA et de FR différent du hasard.

	Taux de réussite global	signif.	FR	signif.	FA	signif.
Site 1	81%	***	24%	***	16%	***
Site 2	74%	***	71%	**	3%	***
Site 3	59%	*	49%	ns	37%	*
Site 4	54%	ns	16%	***	62%	*
Site 5	53%	ns	25%	***	58%	ns
Site 6	52%	ns	59%	ns	42%	ns
Site 7	49%	ns	4%	***	76%	***

B.2 Groupe non-expérimentés : Résultats par auditeur

Ce tableau indique les performances obtenues pour chacun des 29 auditeurs inexpérimentés ayant participé au test perceptif. Seuls quatre auditeurs font mieux que le hasard.

	% réussite	p	signif.	FR	p	signif.	FA	p	signif.
A. 1	44%	0.276	ns	39%	0.160	ns	73%	0.002	**

	% réussite	p	signif.	Taux FR	p	signif.	Taux FA	p	signif.
A. 2	45%	0.372	ns	65%	0.048	*	45%	0.575	ns
A. 3	46%	0.488	ns	47%	0.779	ns	61%	0.160	ns
A. 4	48%	0.766	ns	45%	0.575	ns	59%	0.262	ns
A. 5	49%	0.921	ns	37%	0.091	ns	65%	0.04887	*
A. 6	49%	0.921	ns	80%	$1.47E^{-005}$	***	22%	$5.70E^{-005}$	***
A. 7	50%	1	ns	59%	0.262	ns	41%	0.262	ns
A. 8	50%	1	ns	47%	0.779	ns	53%	0.779	ns
A. 9	52%	0.766	ns	59%	0.262	ns	37%	0.091	ns
A. 10	52%	0.766	ns	57%	0.401	ns	39%	0.160	ns
A. 11	53%	0.620	ns	45%	0.575	ns	49%	1	ns
A. 12	53%	0.620	ns	45%	0.575	ns	49%	1	ns
A. 13	54%	0.488	ns	47%	0.779	ns	45%	0.575	ns
A. 14	54%	0.488	ns	61%	0.160	ns	31%	0.010	*
A. 15	54%	0.488	ns	37%	0.091	ns	55%	0.575	ns
A. 16	54%	0.488	ns	39%	0.160	ns	53%	0.779	ns
A. 17	54%	0.488	ns	43%	0.401	ns	49%	1	ns
A. 18	54%	0.488	ns	37%	0.091	ns	55%	0.575	ns
A. 19	55%	0.372	ns	63%	0.091	ns	27%	0.002	**
A. 20	55%	0.372	ns	43%	0.401	ns	47%	0.779	ns
A. 21	56%	0.276	ns	41%	0.262	ns	47%	0.779	ns
A. 22	56%	0.276	ns	51%	1	ns	37%	0.091	ns
A. 23	56%	0.276	ns	51%	1	ns	37%	0.091	ns
A. 24	57%	0.197	ns	39%	0.160	ns	47%	0.779	ns
A. 25	58%	0.137	ns	20%	$1.47e^{-005}$	***	65%	0.048	*
A. 26	61%	0.037	*	43%	0.401	ns	35%	0.048	*
A. 27	61%	0.037	*	27%	0.002	**	51%	1	ns
A. 28	64%	0.007	**	25%	0.001	***	47%	0.779	ns
A. 29	66%	0.002	**	43%	0.401	ns	25%	0.0006	***

B.3 Groupe non-expérimentés : Résultats par stimulus

B.3.1 Comparaisons cible

Ce tableau indique le taux d'erreur obtenu par les 29 auditeurs non-expérimentés pour chaque comparaison cible. Une variation importante est observée en fonction du stimulus.

Stimuli	VA	FR	%erreur	p	significativité
14	26	3	10%	$1.52e^{-005}$	***
94	24	5	17%	0.0005461	***
139	24	5	17%	0.0005461	***
21	23	6	21%	0.002316	**
66	22	7	24%	0.00813	**
97	22	7	24%	0.00813	**
102	22	7	24%	0.00813	**
44	21	8	28%	0.02412	*
65	21	8	28%	0.02412	*
67	21	8	28%	0.02412	*
133	21	8	28%	0.02412	*
136	21	8	28%	0.02412	*
31	20	9	31%	0.06143	ns
36	20	9	31%	0.06143	ns
54	20	9	31%	0.06143	ns
95	20	9	31%	0.06143	ns
107	20	9	31%	0.06143	ns
137	20	9	31%	0.06143	ns
147	20	9	31%	0.06143	ns
135	19	10	34%	0.136	ns
51	18	11	38%	0.2649	ns
56	18	11	38%	0.2649	ns
120	18	11	38%	0.2649	ns
79	17	12	41%	0.4583	ns
126	17	12	41%	0.4583	ns
99	16	13	45%	0.711	ns

Stimuli	VA	FR	%erreur	p	significativité
142	16	13	45%	0.711	ns
1	15	14	48%	1	ns
23	15	14	48%	1	ns
111	15	14	48%	1	ns
128	15	14	48%	1	ns
5	14	15	52%	1	ns
10	14	15	52%	1	ns
134	14	15	52%	1	ns
42	13	16	55%	0.711	ns
47	13	16	55%	0.711	ns
81	12	17	59%	0.4583	ns
61	11	18	62%	0.2649	ns
108	11	18	62%	0.2649	ns
92	10	19	66%	0.136	ns
77	9	20	69%	0.06143	ns
78	9	20	69%	0.06143	ns
15	8	21	72%	0.02412	*
32	8	21	72%	0.02412	*
86	8	21	72%	0.02412	*
96	8	21	72%	0.02412	*
106	8	21	72%	0.02412	*
141	6	23	79%	0.002316	**
25	5	24	83%	0.0005461	***
150	5	24	83%	0.0005461	***
7	4	25	86%	0.0001037	***

B.3.2 Comparaisons imposteur

Ce tableau indique le taux d'erreur obtenu par les 29 auditeurs non-expérimentés pour chaque comparaison imposteur. Une variation importante est observée en fonction du stimulus.

Stimuli	VR	FA	%erreur	p	significativité
93	28	1	3%	$1.12e^{-007}$	***
39	27	2	7%	$1.62e^{-006}$	***
60	26	3	10%	$1.52e^{-005}$	***
90	26	3	10%	$1.52e^{-005}$	***
52	25	4	14%	0.0001037	***
130	25	4	14%	0.0001037	***
115	24	5	17%	0.0005461	***
13	23	6	21%	0.002316	**
75	22	7	24%	0.00813	**
34	21	8	28%	0.02412	*
43	21	8	28%	0.02412	*
74	21	8	28%	0.02412	*
76	21	8	28%	0.02412	*
148	21	8	28%	0.02412	*
27	19	10	34%	0.136	ns
55	18	11	38%	0.2649	ns
104	18	11	38%	0.2649	ns
113	18	11	38%	0.2649	ns
28	17	12	41%	0.4583	ns
63	17	12	41%	0.4583	ns
143	17	12	41%	0.4583	ns
8	16	13	45%	0.711	ns
70	16	13	45%	0.711	ns
83	16	13	45%	0.711	ns
6	15	14	48%	1	ns
12	15	14	48%	1	ns
35	15	14	48%	1	ns
71	15	14	48%	1	ns
116	15	14	48%	1	ns
18	14	15	52%	1	ns
80	14	15	52%	1	ns
85	14	15	52%	1	ns
129	13	16	55%	0.711	ns

Stimuli	VR	FA	%erreur	p	significativité
3	12	17	59%	0.4583	ns
57	12	17	59%	0.4583	ns
127	12	17	59%	0.4583	ns
38	11	18	62%	0.2649	ns
68	11	18	62%	0.2649	ns
2	10	19	66%	0.136	ns
11	10	19	66%	0.136	ns
121	10	19	66%	0.136	ns
131	10	19	66%	0.136	ns
9	9	20	69%	0.06143	ns
89	9	20	69%	0.06143	ns
119	9	20	69%	0.06143	ns
145	8	21	72%	0.02412	*
124	7	22	76%	0.00813	**
62	6	23	79%	0.002316	**
41	5	24	83%	0.0005461	***
73	5	24	83%	0.0005461	***
4	3	26	90%	$1.52e^{-005}$	***

Annexe C

Données sur la source à partir de BREF

C.1 Valeurs moyennes de F0 par locuteur

C.1.1 Hommes

Les valeurs de F0 varient pour les hommes entre 117 Hz et 181 Hz.

Locuteur	F0 moyenne	Écart type
I0M	136	24.9
I6M	146	22.2
I7M	123	19.9
IAM	161	24.8
ICM	107	31.9
IDM	98	28.0
IFM	143	31.3
IGM	171	22.2
IHM	147	24.1
ILM	146	20.9
IMM	119	24.3
IOM	139	18.7
IQM	132	29.3

Locuteur	F0 moyenne	Écart type
ISM	117	27.2
ITM	141	21.9
IVM	143	30.8
J0M	139	31.4
J2M	131	22.5
J7M	133	26.8
JAM	118	78.1
JCM	161	57.2
JFM	145	29.0
JGM	138	31.9
JJM	147	22.2
JQM	153	29.7
JYM	138	64.0
K1M	146	43.8
K5M	167	29.9
K6M	154	21.8
K7M	181	31.7
KBM	167	18.9
KEM	151	25.4
KFM	117	26.0
KHM	161	34.0
KIM	164	46.2
KKM	139	52.3
KQM	123	42.6
KUM	126	34.2
KWM	143	30.6
KXM	122	30.4
KYM	133	26.6
L2M	131	61.8
L3M	153	49.1
L4M	142	41.9
L7M	143	31.4
L9M	119	34.9

Locuteur	F0 moyenne	Écart type
LBM	146	45.2

C.1.2 Femmes

Les valeurs de F0 varient pour les femmes entre 183 Hz et 297 Hz.

Locuteur	F0 moyenne	Écart type
I1F	199	40.5
I2F	219	33.3
I3F	237	26.9
I5F	222	29.1
I8F	232	33.2
I9F	216	31.1
IBF	249	35
IEF	220	28.5
IIF	263	31.9
IJF	247	37.9
IKF	260	32.6
INF	228	30.7
IPF	215	32.8
IRF	252	30.6
IUF	236	31.2
IWF	232	29.2
IXF	238	42.4
IYF	268	45.6
IZF	209	21.4
J1F	217	21.6
J3F	236	35.7
J4F	242	33.4
J5F	243	48.1
J6F	230	35.3
J8F	183	33.2

Locuteur	F0 moyenne	Écart type
J9F	202	30.1
JBF	203	41.4
JHF	235	41.9
JIF	251	44
JKF	216	41.1
JLF	199	34.9
JMF	270	44.9
JOF	277	44.4
JPF	248	35.5
JRF	223	43.3
JSF	257	40.3
JWF	258	26.9
JXF	265	41.6
JZF	225	37.1
K0F	243	25.5
K3F	297	40.4
K4F	225	35.7
K8F	258	32.1
K9F	240	33.1
KAF	237	44.1
KCF	222	27.3
KDF	239	43.6
KGF	224	38.2
KJF	259	29.6
KLF	243	41.9
KMF	257	42.6
KNF	252	33.8
KOF	237	26.4
KPF	229	32.4
KRF	198	35.9
KSF	237	46.7
KTF	244	33
KVF	245	36

Locuteur	F0 moyenne	Écart type
KZF	254	35.9
L0F	261	28.8
L5F	273	28.3
L6F	270	23
L8F	218	24.7
LAF	271	37

C.2 Jitter et shimmer par locuteur

C.2.1 Hommes

Locuteur	Jitter	Shimmer
I0M	1.5	8.6
I6M	1.2	7.3
I7M	1.8	8.8
IAM	1.8	8.4
ICM	2.1	10.2
IDM	2.3	10.1
IFM	1.5	8.1
IGM	1.2	7.6
IHM	1.6	9
ILM	1.5	8
IMM	1.6	10
IOM	1.4	7.8
IQM	1.6	9.1
ISM	1.8	7.8
ITM	1.6	8.3
IVM	1.6	9.9
J0M	1.6	7.3
J2M	1.6	9.7
J7M	1.3	8.1
JAM	1.8	11.4

Locuteur	Jitter	Shimmer
JCM	1.7	12.5
JFM	1.3	10.2
JGM	1.4	9.8
JJM	1.1	9.6
JQM	1.4	8.8
JYM	1.5	12.8
K1M	1.3	10.2
K5M	1.2	8.7
K6M	1.4	8.1
K7M	1.4	9.8
KBM	1.2	7.9
KEM	1.4	9.3
KFM	1.5	7.5
KHM	1.1	8.6
KIM	1.3	10.2
KKM	1.5	10.6
KQM	1.4	10.1
KUM	1.7	8.9
KWM	1.3	10.2
KXM	1.4	8.8
KYM	1.5	9.5
L2M	1.4	10.7
L3M	1.3	11.6
L4M	1.3	10.4
L7M	1.1	9.3
L9M	1.4	9.6
LBM	1.4	9.9

C.2.2 Femmes

Locuteur	Jitter	Shimmer
I1F	1.4	7.5
I2F	1.3	6.9

Locuteur	Jitter	Shimmer
I3F	1.1	6.1
I5F	1.1	6.4
I8F	1.1	6.2
I9F	1	6.7
IBF	1.1	5.5
IEF	1.1	5.6
IIF	1.1	6
IJF	0.9	5.5
IKF	0.9	6.4
INF	1	6.6
IPF	1.2	6.9
IRF	1.2	6.2
IUF	1.2	7.1
IWF	1.1	7.7
IXF	1.3	6.8
IYF	1.1	6.4
IZF	1.2	7.8
J1F	1.3	7.8
J3F	1	6.8
J4F	0.9	7.3
J5F	1.4	6.6
J6F	1	7.2
J8F	1.1	7.8
J9F	1.3	8
JBF	1.7	9.8
JHF	1.2	7.5
JIF	1.2	7.9
JKF	1.3	8.5
JLF	1.4	10.2
JMF	0.9	6.7
JOF	1.1	6.8
JPF	1	6.7
JRF	1	7.1

Locuteur	Jitter	Shimmer
JSF	1.2	6.6
JWF	0.8	5.9
JXF	1	7.4
JZF	1	8.2
K0F	1.1	8
K3F	1	7.6
K4F	1.2	8.9
K8F	1.3	7.7
K9F	1	10
KAF	1.2	8.6
KCF	1.2	8.3
KDF	1.1	6.6
KGF	1.3	6.7
KJF	1.1	6.8
KLF	1.2	7.7
KMF	1.1	10.4
KNF	1	6.2
KOF	1.3	6.6
KPF	1	6.9
KRF	1.1	10
KSF	1.7	9.6
KTF	1	7.6
KVF	1.2	7.5
KZF	1	7
L0F	0.8	6.6
L5F	0.9	8.4
L6F	1	6.1
L8F	1.1	7.7
LAF	1.1	6.9

C.3 Voyelles par locuteur pour l'extraction de F0

C.3.1 Hommes

Locuteurs	Nombre d'occurrences par segments									
	/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
I0M	701	368	161	553	419	564	67	210	165	217
I6M	615	364	137	445	421	446	59	188	103	164
I7M	587	347	165	449	423	487	73	177	143	177
IAM	701	392	197	515	404	522	66	196	171	206
ICM	669	315	165	557	441	521	53	184	133	225
IDM	769	507	172	610	420	654	74	259	167	253
IFM	674	366	173	534	482	518	56	172	134	195
IGM	653	379	145	451	401	472	62	200	132	173
IHM	724	454	167	553	504	533	71	226	148	199
ILM	644	350	119	538	489	501	45	167	115	217
IMM	668	354	171	582	474	568	51	145	130	210
IOM	712	516	158	438	447	527	61	224	181	197
IQM	564	314	130	390	386	400	60	148	116	155
ISM	720	411	186	548	453	507	82	241	182	214
ITM	649	381	176	463	458	466	71	187	159	205
IVM	698	553	174	384	431	535	57	214	161	202
J0M	546	266	141	490	454	416	56	194	113	169
J2M	523	221	114	469	385	442	50	164	116	172
J7M	656	421	147	503	457	493	68	211	118	203
JAM	590	323	128	421	395	470	55	132	121	185
JCM	692	363	139	558	550	458	71	158	136	215
JFM	717	561	119	505	486	563	59	217	114	243
JGM	713	363	106	551	441	466	51	213	139	210
JJM	746	487	144	496	460	556	65	254	169	223
JQM	683	498	145	478	500	546	52	221	151	217
JYM	724	427	195	635	466	562	58	173	144	206
K1M	641	393	150	452	420	499	59	164	128	195
K5M	605	352	152	483	403	422	56	144	115	171

Locuteurs	Nombre d'occurrences par segments									
	/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
K6M	693	489	133	459	457	478	58	214	123	212
K7M	636	362	185	535	429	487	62	155	154	201
KBM	561	301	77	493	407	418	35	187	100	165
KEM	614	346	138	533	486	429	49	134	129	177
KFM	680	614	143	458	483	531	63	220	143	215
KHM	565	449	151	322	477	457	56	189	114	192
KIM	778	579	177	518	489	624	74	246	175	235
KKM	697	488	170	454	374	497	77	177	145	176
KQM	651	467	192	404	413	486	79	183	150	175
KUM	661	474	160	567	482	551	56	221	164	219
KWM	622	356	181	459	604	467	55	163	137	151
KXM	676	414	175	546	447	507	55	169	153	204
KYM	591	389	154	499	433	512	79	191	127	187
L2M	645	480	181	414	446	468	73	190	131	202
L3M	658	396	169	493	411	498	74	229	132	198
L4M	693	480	151	531	557	574	65	231	135	237
L7M	698	585	135	321	472	581	57	246	154	202
L9M	706	474	205	450	451	582	72	218	154	191
LBM	692	381	215	620	453	561	66	171	154	195

C.3.2 Femmes

Locuteurs	Nombre d'occurrences par segments									
	/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
I1F	662	351	119	543	389	499	50	205	157	197
I2F	636	375	142	564	377	489	47	181	107	210
I3F	570	294	112	574	344	444	49	158	83	156
I5F	659	448	101	471	423	464	47	220	134	180
I8F	589	427	106	387	358	458	64	199	154	178
I9F	516	335	136	478	356	459	61	184	128	137
IBF	622	452	146	387	344	464	71	190	125	198
IEF	603	376	128	483	395	416	49	174	133	181

Locuteurs	Nombre d'occurrences par segments									
	/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
IIF	629	492	141	510	537	525	56	223	138	210
IJF	518	365	115	423	461	460	59	225	136	170
IKF	707	453	97	541	535	500	61	196	108	210
INF	585	374	99	427	401	491	55	194	88	175
IPF	666	475	117	513	489	523	74	228	126	204
IRF	611	332	126	466	487	487	43	183	146	190
IUF	576	372	147	466	385	465	50	164	121	189
IWF	697	394	119	558	416	533	81	213	123	219
IXF	611	449	139	442	464	486	68	191	132	185
IYF	581	442	113	484	412	482	49	183	142	189
IZF	668	367	139	555	388	462	65	182	136	184
J1F	719	445	126	547	438	523	74	228	165	222
J3F	526	279	113	429	307	382	38	196	128	162
J4F	655	450	87	432	587	492	55	174	128	201
J5F	565	281	88	467	366	405	45	155	97	151
J6F	597	333	152	487	396	399	54	130	151	156
J8F	628	359	156	500	484	422	64	153	113	168
J9F	691	435	152	521	376	516	54	199	138	208
JBF	549	318	95	465	395	415	40	195	109	136
JHF	649	416	127	521	429	461	53	194	131	190
JIF	701	462	146	476	434	530	67	203	149	197
JKF	589	357	133	411	397	459	67	169	123	188
JLF	607	346	179	523	368	464	60	156	132	182
JMF	536	390	144	287	378	415	49	143	132	139
JOF	722	474	161	468	421	560	76	232	149	239
JPF	671	367	175	503	351	508	67	187	147	211
JRF	622	431	144	408	418	479	40	178	122	181
JSF	673	389	148	487	430	511	57	205	152	193
JWF	533	409	123	391	429	498	62	182	95	175
JXF	598	402	123	495	343	476	54	219	132	179
JZF	659	538	85	370	347	484	59	207	112	199
K0F	615	477	136	439	385	453	55	208	150	184

Locuteurs	Nombre d'occurrences par segments									
	/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
K3F	705	589	129	356	399	529	48	217	139	202
K4F	687	413	154	483	455	496	74	159	124	208
K8F	743	453	107	571	454	513	71	244	108	216
K9F	669	521	106	414	456	515	67	218	102	212
KAF	603	362	85	580	520	494	38	212	109	167
KCF	661	350	121	569	360	443	55	191	132	178
KDF	660	349	89	579	417	463	62	206	131	194
KGF	583	376	124	406	407	468	45	187	128	173
KJF	614	486	133	464	310	502	73	252	137	184
KLF	719	417	186	573	407	548	61	214	154	228
KMF	702	460	164	440	374	557	78	243	149	222
KNF	658	453	144	429	389	539	75	225	144	193
KOF	590	333	165	556	376	487	57	204	129	180
KPF	607	438	142	444	388	515	61	216	124	176
KRF	595	409	166	349	411	465	67	189	140	191
KSF	539	342	113	384	367	471	53	151	104	148
KTF	709	494	185	503	444	541	76	244	151	209
KVF	615	458	137	434	368	479	66	210	150	211
KZF	635	399	135	438	453	497	49	183	131	216
L0F	716	474	140	424	422	552	77	242	145	193
L5F	629	522	143	440	380	499	69	227	140	193
L6F	596	367	169	409	331	475	59	176	125	177
L8F	722	415	194	520	445	542	64	212	134	210
LAF	679	459	152	459	431	570	56	257	149	205

C.4 Résultats pour analyse des indices sur la source en fonction de la voyelle

L'ensemble des ANOVA a un $p < 2.2e^{-16}$

C.4.1 Hommes

		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
Occurrences		31 101	19 570	7 368	23 127	21 241	23 822	2 913	9 117	6 578	9 352
F0(Hz)	Moyen	137	141	142	139	138	145	138	142	149	146
	Inter	16	18	18	17	17	19	16	18	20	19
	Intra	33	30	27	28	35	33	27	29	35	40
	η^2(%)	16.5	21.4	29.4	25.9	15.9	21.0	22.1	24.6	21.2	16.3
Jitter(%)	Moyen	1.8	1.8	1.9	1.9	2.1	2.2	1.5	1.9	2.4	2.3
	Inter	0.33	0.32	0.32	0.33	0.37	0.33	0.33	0.37	0.34	0.37
	Intra	1.48	1.56	1.44	1.59	1.79	1.93	1.17	1.58	1.90	2.05
	η^2(%)	3.8	3.3	4.0	3.7	3.6	2.6	5.5	4.6	2.8	3.0
Shimmer(%)	Moyen	9.9	10.0	10.2	10.5	11.5	12.6	8.8	9.9	12.4	13.6
	Inter	1.28	1.35	1.43	1.69	1.81	2.19	1.54	1.37	2.03	2.22
	Intra	5.95	6.50	5.87	6.60	7.57	8.14	5.77	6.11	8.04	8.89
	η^2(%)	3.9	3.8	5.2	6.0	5.2	6.3	5.9	4.3	5.3	5.4

C.4.2 Femmes

		/a/	/ɛ/	/o/	/e/	/ø/	/i/	/œ/	/ɔ/	/u/	/y/
Occurrences		40 447	26 270	8 518	30 123	26 204	31 119	3 790	12 685	8 351	12 109
F0(Hz)	Moyen	231	242	241	235	235	245	236	243	250	244
	Inter	22	23	22	22	21	22	24	24	24	23
	Intra	32	35	36	34	34	35	38	32	36	35
	η^2(%)	31.7	30.1	28.7	28.6	26.1	27.6	28.9	34.3	29.4	27.9
Jitter(%)	Moyen	1.2	1.4	1.5	1.6	1.7	1.8	1.0	1.5	2.1	2.2
	Inter	0.21	0.25	0.28	0.25	0.27	0.31	0.25	0.31	0.39	0.33
	Intra	1.01	1.24	1.23	1.29	1.44	1.61	0.84	1.25	1.79	1.83
	η^2(%)	3.5	3.3	4.1	3.1	3.3	3.3	6.9	5.1	4.3	3.0
Shimmer(%)	Moyen	7.9	7.8	8.1	8.1	9.2	9.4	6.5	7.8	10.1	10.8
	Inter	1.31	1.33	1.19	1.39	1.47	1.65	1.28	1.28	1.76	1.66
	Intra	4.00	4.50	4.81	4.48	5.42	6.37	3.58	4.42	7.00	8.38
	η^2(%)	8.8	7.6	5.4	8.7	6.5	6.0	10.7	7.5	5.5	3.6

C.5 Différences *Min, Max* par voyelles pour les mesures de jitter et de shimmer

C.5.1 Jitter

	Hommes		Femmes	
	effet	η^2	effet	η^2
/a/	$F(1,2\,771) = 1.77; p = 0.1837$		$F(1,3\,724) = 4.67; p < 0.05$	0.001%
/ɛ/	$F(1,1\,705) = 2.06; p = 0.1511$		$F(1,2\,441) = 0.87; p = 0.3503$	
/o/	$F(1,572) = 0.52; p = 0.4725$		$F(1,789) = 0.001; p = 0.973$	
/e/	$F(1,2\,030) = 0.04; p = 0.8436$		$F(1,2\,776) = 0.15; p = 0.6943$	
/ø/	$F(1,1\,706) = 7e^{-04}; p = 0.9796$		$F(1,2\,365) = 0.09; p = 0.7528$	
/i/	$F(1,1\,978) = 0.13; p = 0.7226$		$F(1,2\,785) = 1.28; p = 0.2575$	
/œ/	$F(1,231) = 0.002; p = 0.9614$		$F(1,333) = 0.46; p = 0.4993$	
/ɔ/	$F(1,777) = 0.17; p = 0.6775$		$F(1,1\,182) = 1.02; p = 0.312$	
/u/	$F(1,479) = 0.04; p = 0.8329$		$F(1,739) = 6e^{-04}; p = 0.9806$	
/y/	$F(1,692) = 2.24; p = 0.1348$		$F(1,1\,010) = 0.19; p = 0.6595$	

C.5.2 Shimmer

	Hommes		Femmes	
	effet	η^2	effet	η^2
/a/	$F(1,2\,771) = 5.34; p < 0.05$	0.19%	$F(1,3\,724) = 2.30; p = 0.1298$	
/ɛ/	$F(1,1\,705) = 1.87; p = 0.1713$		$F(1,2\,441) = 0.26; p = 0.6085$	
/o/	$F(1,572) = 0.96; p = 0.327$		$F(1,789) = 0.15; p = 0.6951$	
/e/	$F(1,2\,030) = 0.13; p = 0.7198$		$F(1,2\,776) = 0.29; p = 0.5901$	
/ø/	$F(1,1\,706) = 0.66; p = 0.4142$		$F(1,2\,365) = 1.76; p = 0.1851$	
/i/	$F(1,1\,978) = 0.63; p = 0.4277$		$F(1,2\,785) = 1.04; p = 0.3077$	
/œ/	$F(1,231) = 0.02; p = 0.8964$		$F(1,333) = 0.0166; p = 0.8976$	
/ɔ/	$F(1,777) = 1.04; p = 0.3087$		$F(1,1\,182) = 0.01; p = 0.9295$	
/u/	$F(1,479) = 1.94; p = 0.1645$		$F(1,739) = 0.15; p = 0.6943$	
/y/	$F(1,692) = 0.0631; p = 0.8018$		$F(1,1\,010) = 1.13; p = 0.2883$	

Annexe D

Données sur le filtre à partir de BREF

D.1 Occurrences des phonèmes étudiés

Phonèmes	Hommes	Femmes
/p/	14 465	20 738
/t/	22 761	33 362
/k/	16 225	23 297
/b/	4 899	7 131
/d/	20 456	28 811
/g/	3 257	4 743
/f/	6 294	9 050
/s/	25 557	26 555
/ʃ/	2 641	3 759
/v/	8 021	11 405
/z/	7 276	10 358
/ʒ/	4 925	6 895
/ʁ/	33 758	48 552
/l/	26 038	37 455
/j/	9 622	14 124
/ɥ/	1 538	2 276

Phonèmes	Hommes	Femmes
/w/	3 240	4 575
/m/	12 904	18 777
/n/	13 580	20 044
/ɛ̃/	6 532	9 352
/ɔ̃/	8 178	11 853
/ɑ̃/	14 397	20 539
/i/	23 761	34 426
/y/	9 348	13 343
/u/	6 570	9 285
/ø/	21 158	28 852
/œ/	2 907	4 177
/e/	23 069	33 094
/ɛ/	19 539	29 001
/o/	7 344	9 469
/ɔ/	9 095	14 069
/a/	30 969	44 450
pause	34 857	58 812

D.2 Valeurs médianes des centres de gravité selon le phonème (Hz)

Phonèmes	Hommes	Femmes
/p/	304	355
/t/	533	542
/k/	585	604
/b/	266	302
/d/	311	359
/g/	386	432
/f/	735	791
/s/	2 563	2 006

Phonèmes	Hommes	Femmes
/ʃ/	2 473	2 801
/v/	355	385
/z/	560	482
/ʒ/	941	1 178
/ʁ/	541	549
/l/	400	406
/j/	594	535
/ɥ/	458	528
/w/	488	533
/m/	293	298
/n/	295	302
/ɛ̃/	559	606
/ɔ̃/	411	439
/ɑ̃/	533	585
/i/	428	378
/y/	414	407
/u/	331	367
/ø/	410	436
/œ/	514	568
/e/	451	449
/ɛ/	532	552
/o/	398	441
/ɔ/	478	512
/a/	631	684
pause	240	347

D.3 Le locuteur comme facteur pour les valeurs de centre de gravité

Phonèmes	Hommes		Femmes	
	Effet	η^2	Effet	η^2
/p/	$F(46, 14\,418) = 27.091; p < 0.001$	7.9%	$F(63, 20\,674) = 29.116; p < 0.001$	8.1%
/t/	$F(46, 22\,714) = 77.061; p < 0.001$	13.5%	$F(63, 33\,298) = 47.484; p < 0.001$	8.2%
/k/	$F(46, 16\,178) = 38.668; p < 0.001$	9.9%	$F(63, 23\,233) = 31.563; p < 0.001$	7.9%
/b/	$F(46, 4\,852) = 16.686; p < 0.001$	13.7%	$F(63, 7\,067) = 14.999; p < 0.001$	11.8%
/d/	$F(46, 20\,409) = 70.843; p < 0.001$	13.8%	$F(63, 28\,747) = 49.002; p < 0.001$	9.7%
/g/	$F(46, 3\,210) = 20.02; p < 0.001$	22.3%	$F(63, 4\,679) = 11.507; p < 0.001$	13.4%
/f/	$F(46, 6\,247) = 32.5; p < 0.001$	19.3%	$F(63, 8\,986) = 24.162; p < 0.001$	14.5%
/s/	$F(46, 25\,510) = 241.31; p < 0.001$	30.3%	$F(63, 36\,491) = 152.88; p < 0.001$	20.9%
/ʃ/	$F(46, 2\,594) = 15.085; p < 0.001$	21.1%	$F(63, 3\,695) = 13.572; p < 0.001$	18.8%
/v/	$F(46, 7\,974) = 43.069; p < 0.001$	19.9%	$F(63, 11\,341) = 35.387; p < 0.001$	16.4%
/z/	$F(46, 7\,229) = 50.043; p < 0.001$	24.2%	$F(63, 10\,294) = 32.093; p < 0.001$	16.4%
/ʒ/	$F(46, 4\,878) = 40.697; p < 0.001$	27.7%	$F(63, 6\,831) = 28.868; p < 0.001$	21.0%
/ʁ/	$F(46, 33\,711) = 86.743; p < 0.001$	10.6%	$F(63, 48\,488) = 85.319; p < 0.001$	10.0%
/l/	$F(46, 25\,991) = 160.28; p < 0.001$	22.1%	$F(63, 37\,391) = 127.48; p < 0.001$	17.7%
/j/	$F(46, 9\,575) = 18.9; p < 0.001$	8.3%	$F(63, 14\,060) = 20.921; p < 0.001$	8.5%
/ɥ/	$F(46, 1\,491) = 4.0389; p < 0.001$	11.1%	$F(63, 2\,212) = 3.3307; p < 0.001$	8.7%
/w/	$F(46, 3\,193) = 6.886; p < 0.001$	9.0%	$F(63, 4\,511) = 5.5777; p < 0.001$	7.2%
/m/	$F(46, 12\,857) = 179.79; p < 0.001$	39.1%	$F(63, 18\,713) = 91.342; p < 0.001$	23.6%
/n/	$F(46, 13\,533) = 197.42; p < 0.001$	40.2%	$F(63, 19\,980) = 127.52; p < 0.001$	28.7%
/ɛ̃/	$F(46, 6\,485 = 163.81; p < 0.001$	53.7%	$F(63, 9\,288) = 251.51; p < 0.001$	63.0%
/ɔ̃/	$F(46, 8\,131) = 146.07; p < 0.001$	45.2%	$F(63, 11\,789) = 132.32; p < 0.001$	41.4%
/ɑ̃/	$F(46, 14\,350) = 221.45; p < 0.001$	41.5%	$F(63, 20\,475) = 324.82; p < 0.001$	50.0%
/i/	$F(46, 23\,714) = 219.21; p < 0.001$	29.8%	$F(63, 34\,362) = 231.07; p < 0.001$	29.8%
/y/	$F(46, 9\,301) = 39.661; p < 0.001$	16.4%	$F(63, 13\,279) = 29.704; p < 0.001$	12.4%
/u/	$F(46, 6\,523) = 21.205; p < 0.001$	13.0%	$F(63, 9\,221) = 13.069; p < 0.001$	8.2%
/ø/	$F(46, 21\,111) = 263.38; p < 0.001$	36.5%	$F(63, 28\,788) = 164.68; p < 0.001$	26.5%
/œ/	$F(46, 2\,860) = 74.913; p < 0.001$	54.6%	$F(63, 4\,113) = 92.372; p < 0.001$	58.6%
/e/	$F(46, 23\,022) = 528.43; p < 0.001$	51.4%	$F(63, 33\,030) = 657.04; p < 0.001$	55.6%
/ɛ/	$F(46, 19\,492) = 290.34; p < 0.001$	40.7%	$F(63, 28\,937) = 563.52; p < 0.001$	55.1%
/o/	$F(46, 7\,297) = 58.496; p < 0.001$	26.9%	$F(63, 9\,405) = 37.457; p < 0.001$	20.0%
/ɔ/	$F(46, 9\,048) = 136.04; p < 0.001$	40.9%	$F(63, 14\,005) = 99.178; p < 0.001$	30.9%
/a/	$F(46, 30\,922) = 461.06; p < 0.001$	40.7%	$F(63, 44\,386) = 759.81; p < 0.001$	51.9%

Phonèmes	Hommes		Femmes	
	Effet	η^2	Effet	η^2
pause	$F(46,34\,810) = 82.544; p < 0.001$	9.8%	$F(63,58\,748) = 87.242; p < 0.001$	8.6%

D.4 Influence du locuteur sur les centres de gravité des séries *Min* et *Max*

Phonèmes	Hommes				Femmes			
	Min		Max		Min		Max	
	Effet	η^2	Effet	η^2	Effet	η^2	Effet	η^2
/p/	***	16.1%	***	15.7%	***	15.7%	***	13.6%
/t/	***	18.8%	***	15.8%	***	17.6%	***	14.6%
/k/	***	15.6%	***	14.6%	***	14.0%	***	12.2%
/b/	**	26.4%	***	31.7%	**	31.8%	***	34.3%
/d/	***	32.8%	***	19.0%	***	12.7%	***	13.3%
/g/	n.s		n.s		*	38.5%	**	46.4%
/f/	***	26.5%	***	28.4%	***	30.3%	***	31.0%
/s/	***	32.8%	***	32.1%	***	25.6%	***	26.6%
/ʃ/	***	52.4%	**	51.0%	n.s		n.s	
/v/	***	27.6%	***	29.8%	***	26.0%	***	19.7%
/z/	***	33.0%	***	39.8%	***	46.2%	***	34.7%
/ʒ/	***	44.6%	***	38.5%	***	37.1%	***	35.3%
/ʁ/	***	13.5%	***	15.3%	***	15.1%	***	11.8%
/l/	***	24.4%	***	26.3%	***	21.9%	***	22.5%
/j/	n.s		**	16.0%	**	15.0%	***	18.0%
/ɥ/	n.s		n.s		n.s		n.s	
/w/	***	56.0%	n.s		n.s		n.s	
/m/	***	50.3%	***	37.2%	***	35.7%	***	30.5%
/n/	***	44.3%	***	48.7%	***	37.8%	***	40.5%
/ɛ̃/	***	68.1%	***	65.8%	***	70.2%	***	63.4%
/ɔ̃/	***	55.3%	***	55.9%	***	51.2%	***	54.5%

Phonèmes	Hommes				Femmes			
	Min		Max		Min		Max	
	Effet	η^2	Effet	η^2	Effet	η^2	Effet	η^2
/ɑ̃/	***	20.7%	***	45.3%	***	57.8%	***	33.2%
/i/	***	34.8%	***	33.2%	***	25.3%	***	26.6%
/y/	***	29.3%	***	20.9%	***	20.2%	***	19.7%
/u/	***	27.8%	**	22.5%	*	21.7%	***	30.0%
/ø/	***	53.2%	***	44.4%	***	20.2%	***	20.6%
/œ/	***	82.4%	***	68.4%	***	69.2%	***	63.9%
/e/	***	57.7%	***	48.4%	***	55.7%	***	53.4%
/ɛ/	***	38.7%	***	39.7%	***	54.3%	***	59.9%
/o/	***	33.0%	***	31.1%	**	23.8%	***	45.9%
/ɔ/	***	49.9%	***	46.9%	***	33.3%	***	40.1%
/a/	***	46.5%	***	35.5%	***	53.4%	***	50.2%
pause	***	10.5%	***	11.1%	***	8.3%	***	10.5%

D.5 Différentier *Min* et *Max* à l'aide des centres de gravité

Phonèmes	Centres de gravité			
	Hommes		Femmes	
	Effet	η^2	Effet	η^2
p	n.s		n.s	
t	n.s		n.s	
k	n.s		n.s	
b	n.s		n.s	
d	n.s		n.s	
g	n.s		n.s	
f	n.s		n.s	
s	n.s		n.s	
ʃ	n.s		n.s	
v	n.s		n.s	
z	n.s		n.s	
ʒ	n.s		n.s	
ʁ	n.s		n.s	
l	n.s		n.s	
j	n.s		n.s	
ɥ	n.s		n.s	
w	n.s		n.s	
m	$F(1, 1\,196) = 5.786; p < 0.05$	0.48%	n.s	
n	n.s		n.s	
ɛ̃	n.s		n.s	
ɔ̃	n.s		n.s	
ɑ̃	n.s		$F(1, 1\,736) = 5.7343; p < 0.05$	0.33%
i	n.s		n.s	
y	n.s		n.s	
u	n.s		n.s	
ø	n.s		n.s	
œ	n.s		n.s	
e	n.s		n.s	
ɛ	n.s		$F(1, 2\,461) = 3.9278; p < 0.05$	0.16%
o	n.s		n.s	
ɔ	n.s		n.s	
a	n.s		n.s	
pause	n.s		n.s	

TABLE *D.6* – *Comparaisons des centres de gravité selon le phonème entre Min et Max*

D.6 Influence du locuteur sur les voyelles : comparaison *Min* et *Max*

D.6.1 Hommes

		F1/F2/F3/F4		F1		F2		F3		F4	
		Min	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
/a/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	33.0	33.1	16.2	15.6	8.7	9.3	43.2	45.5	46.8	44.2
/ɛ/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	34.8	33.3	12.3	15.1	19.9	24.2	37.9	39.9	44.3	36.7
/o/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	31.8	37.8	28.8	43.3	28.2	42.0	31.6	39.3	28.9	31.4
/e/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	32.4	30.7	12.2	8.3	29.5	27.7	25.8	30.7	43.7	38.5
/ø/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	25.4	27.7	32.8	33.3	18.0	19.6	30.3	31.7	27.9	26.6
/i/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	20.6	19.3	10.2	11.1	18.8	15.3	19.9	17.4	23.3	20.9
/œ/	effet	***	***	ns	**	ns	***	***	***	***	***
	η^2(%)	52.9	59.6		48.8		54.1	63.8	68.5	58.3	55.4
/ɔ/	effet	***	***	***	***	***	***	***	***	***	***
	η^2(%)	31.1	32.7	32.5	28.2	17.6	19.3	29.2	38.1	34.8	35.0
/u/	effet	***	***	**	**	***	ns	***	***	***	***
	η^2(%)	23.9	25.2	23.6	25.1	15.9		24.9	28.5	23.9	26.9
/y/	effet	***	***	**	***	***	***	***	***	***	***
	η^2(%)	25.6	27.0	15.4	10.9	23.0	20.0	29.7	21.4	31.1	36.5

D.6.2 Femmes

		F1/F2/F3/F4		F1		F2		F3		F4	
		Min	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*
	effet	***	***	***	***	***	***	***	***	***	***
/a/	η^2(%)	30.9	32.2	15.6	20.6	14.1	15.8	39.4	41.4	43.3	42.7
	effet	***	***	***	***	***	***	***	***	***	***
/ɛ/	η^2(%)	31.3	32.9	16.1	12.1	22.4	22.6	38.0	46.1	45.3	46.4
	effet	***	***	***	***	**	***	***	***	***	***
/o/	η^2(%)	39.2	39.0	26.6	25.5	24.2	14.9	48.2	52.9	50.1	47.6
	effet	***	***	***	***	***	***	***	***	***	***
/e/	η^2(%)	31.8	29.7	17.0	14.1	26.3	33.4	26.3	40.4	43.6	36.9
	effet	***	***	***	***	***	***	***	***	***	***
/ø/	η^2(%)	25.3	21.8	18.0	16.4	10.9	9.5	36.0	30.2	27.4	24.4
	effet	***	***	***	***	***	***	***	***	***	***
/i/	η^2(%)	19.0	20.6	15.0	13.0	29.1	32.0	16.0	20.2	17.8	17.2
	effet	***	***	***	*	***	***	***	***	***	***
/œ/	η^2(%)	55.6	57.9	46.7	40.7	51.6	55.7	71.1	70.4	58.4	66.5
	effet	***	***	***	***	***	***	***	***	***	***
/ɔ/	η^2(%)	32.1	36.7	18.9	13.5	15.1	13.5	50.6	55.1	42.4	47.5
	effet	***	***	*	***	*	**	*	*	***	***
/u/	η^2(%)	25.2	21.8	22.2	23.2	22.0	20.6	32.1	19.1	32.1	24.7
	effet	***	***	***	*	***	***	***	***	***	**
/y/	η^2(%)	20.3	20.3	17.1	15.6	22.8	21.3	22.9	23.5	19.4	16.3

D.7 Influence du locuteur sur la dynamique du formant

D.7.1 Hommes

	F1		F2		F3		F4	
	effet	η^2(%)	effet	η^2(%)	effet	η^2(%)	effet	η^2(%)
/a/	***	3.5	***	1.4	***	7.3	***	6.7
/ɛ/	***	3.0	***	3.5	***	7.3	***	6.9
/o/	***	7.2	***	6.3	***	6.0	***	5.3
/e/	***	2.1	***	5.4	***	4.9	***	6.5
/ø/	***	5.0	***	2.4	***	4.7	***	4.9
/i/	***	3.4	***	2.4	***	2.6	***	3.6
/œ/	***	7.9	***	4.9	***	9.5	***	10.3
/ɔ/	***	4.1	***	2.4	***	5.7	***	6.0
/u/	***	4.0	***	2.7	***	4.2	***	4.0
/y/	***	3.8	***	2.5	***	3.3	***	4.2

D.7.2 Femmes

	F1		F2		F3		F4	
	effet	η^2(%)	effet	η^2(%)	effet	η^2(%)	effet	η^2(%)
/a/	***	3.3	***	1.9	***	5.8	***	5.9
/ɛ/	***	2.1	***	3.0	***	6.0	***	6.5
/o/	***	4.9	***	2.0	***	7.9	***	6.6
/e/	***	3.7	***	3.6	***	5.9	***	5.5
/ø/	***	3.0	***	1.4	***	5.2	***	4.2
/i/	***	3.3	***	4.0	***	2.4	***	2.4
/œ/	***	6.4	***	5.3	***	9.8	***	8.5
/ɔ/	***	3.1	***	1.7	***	7.7	***	6.3
/u/	***	3.5	***	2.1	***	2.6	***	3.8
/y/	***	4.3	***	2.4	***	2.5	***	3.1

D.8 Effet du locuteur sur les courbes formantiques dans *Min* et *Max*

D.8.1 Hommes

		Min		*Max*	
		effet	η^2 (%)	effet	η^2 (%)
/a/	F1	***	5.8	***	5.9
	F2	***	4.1	***	4.3
	F3	***	9.4	***	10.0
	F4	***	9.5	***	8.8
/ɛ/	F1	***	6.4	***	7.3
	F2	***	7.4	***	9.1
	F3	***	10.9	***	11.5
	F4	***	11.7	***	9.8
/o/	F1	***	16.8	***	20.4
	F2	***	14.2	***	19.8
	F3	***	15.3	***	19.8
	F4	***	15.7	***	21.6
/e/	F1	***	5.3	***	5.7
	F2	***	8.5	***	8.8
	F3	***	7.8	***	8.9
	F4	***	10.3	***	10.2
/ø/	F1	***	9.4	***	8.0
	F2	***	6.6	***	6.3
	F3	***	8.9	***	8.4
	F4	***	7.7	***	7.7
/i/	F1	***	5.9	***	5.7
	F2	***	6.0	***	5.9
	F3	***	6.0	***	6.2
	F4	***	6.9	***	6.2
/œ/	F1	***	44.8	***	36.7
	F2	***	39.1	***	35.9
	F3	***	43.0	***	37.5

	F4	***	37.4	***	31.6
/ɔ/	F1	***	13.3	***	15.6
	F2	***	12.5	***	12.8
	F3	***	14.4	***	15.8
	F4	***	12.1	***	15.5
/u/	F1	***	18.6	***	16.6
	F2	***	15.4	***	13.7
	F3	***	15.5	***	17.9
	F4	***	17.0	***	18.3
/y/	F1	***	11.1	***	10.5
	F2	***	12.3	***	13.5
	F3	***	13.1	***	13.0
	F4	***	13.3	***	15.7

D.8.2 Femmes

		Min		*Max*	
		effet	η^2 (%)	effet	η^2 (%)
/a/	F1	***	5.8	***	6.0
	F2	***	5.2	***	5.0
	F3	***	8.4	***	9.1
	F4	***	9.1	***	8.8
/ɛ/	F1	***	6.6	***	6.1
	F2	***	6.9	***	7.6
	F3	***	9.0	***	11.2
	F4	***	10.3	***	10.7
/o/	F1	***	16.1	***	18.1
	F2	***	15.6	***	14.6
	F3	***	21.6	***	21.8
	F4	***	20.9	***	19.9
/e/	F1	***	7.8	***	7.3
	F2	***	7.0	***	8.9

		Min		*Max*	
		effet	η^2 (%)	effet	η^2 (%)
	F3	***	8.9	***	9.5
	F4	***	9.9	***	9.3
/ø/	F1	***	7.5	***	7.8
	F2	***	6.3	***	6.7
	F3	***	10.4	***	9.6
	F4	***	8.0	***	8.7
/i/	F1	***	6.2	***	6.3
	F2	***	7.7	***	8.5
	F3	***	5.5	***	6.7
	F4	***	6.3	***	6.1
/œ/	F1	***	32.9	***	39.5
	F2	***	30.6	***	41.7
	F3	***	34.5	***	41.1
	F4	***	33.3	***	41.3
/ɔ/	F1	***	12.5	***	12.8
	F2	***	11.6	***	11.6
	F3	***	15.8	***	18.1
	F4	***	16.1	***	16.2
/u/	F1	***	18.1	***	16.6
	F2	***	17.2	***	16.0
	F3	***	18.5	***	14.0
	F4	***	18.7	***	16.9
/y/	F1	***	14.8	***	12.6
	F2	***	14.8	***	12.5
	F3	***	15.1	***	13.2
	F4	***	13.6	***	12.1

D.9 Différentier *Min* et *Max* par la courbe formantique

		Hommes		Femmes	
		effet	η^2 (%)	effet	η^2 (%)
/a/	F1	$F(9, 2\,990) = 1.8071; p = 0.06207$		$F(9, 3\,819) = 1.0330; p = 0.4106$	
	F2	$F(9, 2\,990) = 0.56792; p = 0.8244$		$F(9, 3\,819) = 0.71954; p = 0.6915$	
	F3	$F(9, 2\,990) = 1.2430; p = 0.2636$		$F(9, 3\,819) = 1.0782; p = 0.3753$	
	F4	$F(9, 2\,990) = 0.37885; p = 0.9457$		$F(9, 3\,819) = 1.2521; p = 0.2581$	
/ɛ/	F1	$F(9, 1\,888) = 1.3245; p = 0.2187$		$F(9, 2\,524) = 1.2025; p = 0.2887$	
	F2	$F(9, 1\,888) = 0.53095; p = 0.853$		$F(9, 2\,524) = 0.80745; p = 0.6094$	
	F3	$F(9, 1\,888) = 0.59745; p = 0.8$		$F(9, 2\,524) = 1.1773; p = 0.305$	
	F4	$F(9, 1\,888) = 0.95535; p = 0.4755$		$F(9, 2\,524) = 0.83794; p = 0.83794$	
/o/	F1	$F(9, 639) = 0.62617; p = 0.7752$		$F(9, 857) = 0.75598; p = 0.6575$	
	F2	$F(9, 639) = 0.42759; p = 0.9205$		$F(9, 857) = 0.5312; p = 0.8525$	
	F3	$F(9, 639) = 0.62842; p = 0.7732$		$F(9, 857) = 0.26448; p = 0.9838$	
	F4	$F(9, 639) = 1.0623; p = 0.3889$		$F(9, 857) = 1.1467; p = 0.3267$	
/e/	F1	$F(9, 2\,243) = 1.4601; p = 0.1571$		$F(9, 2\,853) = 1.5146; p = 0.1367$	
	F2	$F(9, 2\,243) = 1.6346; p = 0.09991$		$F(9, 2\,853) = 1.2357; p = 0.268$	
	F3	$F(9, 2\,243) = 1.5160; p = 0.1364$		$F(9, 2\,853) = 1.195; p = 0.2934$	
	F4	$F(9, 2\,243) = 1.6111; p = 0.1064$		$F(9, 2\,853) = 1.3742; p = 0.194$	
/ø/	F1	$F(9, 1\,994) = 1.4205; p = 0.1735$		$F(9, 2\,452) = 40863; p = 0.9312$	
	F2	$F(9, 1\,994) = 1.4909; p = 0.1455$		$F(9, 2\,452) = 0.55418; p = 0.8352$	
	F3	$F(9, 1\,994) = 1.3518; p = 0.2050$		$F(9, 2\,452) = 1.7097; p = 0.08148$	
	F4	$F(9, 1\,994) = 0.53136; p = 0.8526$		$F(9, 2\,452) = 1.4521; p = 0.1602$	
/i/	F1	$F(9, 2\,308) = 0.64206; p = 0.7617$		$F(9, 2\,929) = 1.6061; p = 0.1076$	
	F2	$F(9, 2\,308) = 1.5685; p = 0.119$		$F(9, 2\,929) = 0.42868; p = 0.9204$	
	F3	$F(9, 2\,308) = 3.181; p < 0.001$	0.14	$F(9, 2\,929) = 0.55936; p = 0.8312$	
	F4	$F(9, 2\,308) = 2.1043; p < 0.05$	0.09	$F(9, 2\,929) = 0.46384; p = 0.8994$	
/œ/	F1	$F(9, 286) = 0.97278; p = 0.5577$		$F(9, 396) = 0.97444; p = 0.4606$	
	F2	$F(9, 286) = 0.60016; p = 0.7966$		$F(9, 396) = 0.92243; p = 0.5054$	
	F3	$F(9, 286) = 0.70583; p = 0.7035$		$F(9, 396) = 0.8679; p = 0.5541$	
	F4	$F(9, 3\,819) = 2.0607; p < 0.05$	0.72	$F(9, 396) = 1.1872; p = 0.3016$	
/ɔ/	F1	$F(9, 897) = 0.31382; p = 0.9706$		$F(9, 1\,248) = 0.99636; p = 0.441$	
	F2	$F(9, 897) = 0.78726; p = 0.6283$		$F(9, 1\,248) = 1.7270; p = 0.07834$	
	F3	$F(9, 897) = 0.68493; p = 0.7231$		$F(9, 1\,248) = 0.90559; p = 0.5194$	
	F4	$F(9, 897) = 0.7483; p = 0.65695$		$F(9, 1\,248) = 0.90023; p = 0.5242$	
/u/	F1	$F(9, 619) = 1.2285; p = 0.2743$		$F(9, 839) = 1.3655; p = 0.1995$	

		Hommes		Femmes	
		effet	η^2 (%)	effet	η^2 (%)
	F2	$F(9, 619) = 1.1501; p = 0.3250$		$F(9, 839) = 0.96937; p = 0.464$	
	F3	$F(9, 619) = 0.60159; p = 0.7961$		$F(9, 839) = 1.8639; p = 0.05401$	
	F4	$F(9, 619) = 1.0391; p = 0.407$		$F(9, 839) = 0.78075; p = 0.6344$	
/y/	F1	$F(9, 882) = 0.84796; p = 0.5719$		$F(9, 1\,115) = 0.71882; p = 0.692$	
	F2	$F(9, 882) = 1.2826; p = 0.2422$		$F(9, 1\,115) = 0.98205; p = 0.453$	
	F3	$F(9, 882) = 1.5883; p = 0.1142$		$F(9, 1\,115) = 0.93774; p = 0.4911$	
	F4	$F(9, 882) = 1.7957; p = 0.06528$		$F(9, 1\,115) = 0.90057; p = 0.524$	

Annexe E

MFCC

E.1 η^2 dans BREF

L'effet du locuteur est significatif $p < 0.001$

Phonème	η^2(%)
t	2.22
k	2.77
p	3.35
l	3.37
d	3.75
ø	3.96
ʁ	4.01
ɛ	5.41
e	5.58
ɔ	5.69
pause	5.72
j	5.93
s	6.05
i	6.12
a	6.36
v	6.47
b	6.74

Phonème	η^2(%)
z	7.17
f	7.49
u	8.54
ɑ̃	8.87
n	8.99
ʒ	9.08
m	9.18
u	9.58
g	9.62
ɔ̃	10.03
o	10.05
w	11.17
ɛ̃	13.16
œ	13.87
ʃ	14.35
ɥ	20.57
œ̃	30.41

E.2 η^2 dans NIST

L'effet du locuteur est significatif $p < 0.001$

Phonème	η^2(%)
ʌ	3.69
t	4.42
ɪ	7.65
d	7.77
ð	8.73
ɑ	8.76
ɛ	8.96
h	9.12
b	9.25
l	9.40

Phonème	η^2(%)
ɑɪ	9.61
k	9.63
ɹ	9.99
e	11.18
ʊ	11.18
v	11.54
oʊ	11.84
w	12.17
ɾ	12.50
i	12.74
æ	12.80
ɚ	12.82
p	12.95
u	13.13
n	13.57
j	13.67
g	14.70
pause	15.21
z	15.27
ɔ	15.31
m	15.72
s	17.43
dʒ	18.87
f	20.36
θ	20.99
rire	21.67
aʊ	24.82
ŋ	25.41
ʃ	26.80
ts	28.74
ʒ	75.97
ɔɪ	85.07

Liste des illustrations

Liste des tableaux

Bibliographie

(Alexander et al., 2004) A. Alexander, F. Botti, D. Dessimoz, et A. Drygajlo, 2004. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic science international 146*, 95–99.

(Amino et al., 2006) K. Amino, T. Sugawara, et T. Arai, 2006. Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Sciences and Technology. 27*, 233–235.

(Auckenthaler et al., 2000) R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. Score normalization for Text-Independent speaker verification systems. *Digital Signal Processing 10*(1-3), 42–54.

(Austin, 1970) J. Austin, 1970. *Quand dire, c'est faire, trad*. Paris : Seuil.

(Bimbot et al., 2004) F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, et D. A. Reynolds, 2004. A tutorial on Text-Independent speaker verification. *Eurasip Journal on applied signal processing 4*, 430–451.

(Blatchford, 2006) H. Blatchford, 2006. Identification of voices in shouting. *The International Journal of Speech, Langage and the Law 13*, 241–254.

(Boersma et Weenink, 2009) P. Boersma et D. Weenink, 2009. *Praat : doing phonetics by computer (Version 5.1.05) [Computer program].* à partir de http ://www.praat.org/.

(Bohm et Shattuck-Hufnagel, 2007) T. Bohm et S. Shattuck-Hufnagel, 2007. Listeners recognize speakers' habitual utterance-final voice quality. Dans les actes de *International workshop on Paralinguistic Speech - between models and data*, Saarbrucken, 29–34.

(Bolt et al., 1973) R. Bolt, F. Cooper, E. David, P. Denes, J. Pickett, et K. Stevens, 1973.

Speaker identification by speech spectrograms : some further observations. *Journal of the Acoustical Society of America, 54*, 531–534.

(Bonastre et al., 2003) J. Bonastre, F. Bimbot, L. Boë, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau, 2003. Person authentification by voice : A need for caution. Dans les actes de *EUROSPEECH*.

(Bonastre et al., 2008) J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, et N. Evans, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. Dans les actes de *ISCA-IEEE Speaker Odyssey*, Stellenbosch.

(Boë et al., 1999) L. Boë, F. Bimbot, J. Bonastre, et P. Dupont, 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.

(Bradlow et al., 1996) A. R. Bradlow, G. M. Torretta, et D. B. Pisoni, 1996. Intelligibility of normal speech I : Global and fine-grained acoustic-phonetic talker characteristics,. *Speech Communication 20*(3-4), 255–272.

(Bricker et al., 1976) P. Bricker, S. Pruzansky, et N. Lass, 1976. Speaker recognition. Dans N. Lass (Ed.), *Experimental Phonetics*, 295–326. London : Academic Press.

(Brockmann et al., 2011) M. Brockmann, M. Drinnan, ClaudioStorck, et P. Carding, 2011. Reliable jitter and shimmer measurements in voice clinics : the relevance of vowels, gender, vocal intensity and fundamental frequency effects in a typical clinical task. *Journal of Voice 25*, 44 – 53.

(Broeders et Amelsvoort, 1999) A. P. Broeders et A. G. V. Amelsvoort, 1999. Lineup construction for forensic earwitness identification : A practical approach. Dans les actes de *International Conference in Phonetic Sciences*, Leeds, 1373–1376.

(Butcher, 2002) A. Butcher, 2002. Forensic phonetics : Issues in speaker identification evidence. Dans les actes de *International Conference of the Institute of Forensic Studies :'Forensic Evidence : Proof and Presentation'*, Prato, 3–5.

(Calliope, 1989) Calliope, 1989. *La parole et son traitement automatique*. Paris : Masson & Cie édition.

(Campbell et Reynolds, 1999) J. Campbell et D. Reynolds, 1999. Corpora for the evaluation of speaker recognition systems. Dans les actes de *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Volume 2, 829–832.

(Campbell et Erickson, 2004) N. Campbell et D. Erickson, 2004. What do people hear ? A study of the perception of non-verbal affective information in conversational speech. *Journal of the Phonetic Society of Japan 7*(4), 9–28.

(Campbell et Mokhtari, 2003) N. Campbell et P. Mokhtari, 2003. Voice quality : the 4th prosodic dimension. Dans les actes de *International Conference of Phonetic Sciences (ICPhS)*, Barcelone, 2417–2420.

(Campbell et al., 2004) W. Campbell, J. Campbell, D. Reynolds, D. A. Jones, et T. Leek, 2004. Phonetic speaker recognition with support vector machines. *Advances in Neural Information Processing Systems 16*, 1377–1384.

(Campbell et al., 2007) W. Campbell, T. Gleason, D. Reynolds, et W. Shen, 2007. Speaker verification using support vector machines and High-Level features. *Transaction on Audio, Speech, and Language Processing 15*, 2085–2094.

(Campbell et al., 2005) W. Campbell, D. Reynolds, J. Campbell, et K. Brady, 2005. Estimating and evaluating confidence for forensic speaker recognition. Dans les actes de *International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP)*, Philadelphie, 717–720.

(Cappé, 1995) O. Cappé, 1995. Etat actuel de la recherche en reconnaissance du locuteur et des applications en criminalistique. http ://www.tsi.enst.fr/cappe.

(Carey et Parris, 1992) M. Carey et E. Parris, 1992. Speaker verification using connected words. *Institute of Acoustics 14*, 96–100.

(Chang et Lin, 2011) C. Chang et C. Lin, 2011. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*(3), 27.

(Cieri et al., 2004) C. Cieri, D. Miller, et K. Walker, 2004. The fisher corpus : A resource for the next generations of speech-to-text. Dans les actes de *International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, 69–71.

(Cornut, 2009) G. Cornut, 2009. *La voix*. Paris : Presses Universitaires de France.

(Cutler et al., 2011) A. Cutler, A. Andics, et Z. Fang, 2011. Inter-dependent categorization of voices and segments. Dans les actes de *International Conference of Phonetic Sciences (ICPhS)*, Hong Kong, 552–555.

(Davis et Mermelstein, 1980) S. Davis et P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing 28*(4), 357–366.

(de Jong et al., 2007) G. de Jong, K. McDougall, et F. Nolan, 2007. Sound change and speaker identity : an acoustic study. Dans C. Miller (Ed.), *Speaker Classification II*, 130–141. Berlin : Springer.

(de Looze et al., 2011) C. de Looze, C. Oertel, S. Rauzy, et N. Campbell, 2011. Measuring dynamics of mimicry by means of prosodic cues in consversational speech. Dans les actes de *International Conference of Phonetic Sciences (ICPhS)*, Hong Kong, 1294–1297.

(de Saussure, 1916) F. de Saussure, 1916. *Cours de linguistique générale*. Paris : Payot et Rivages.

(Dehak et al., 2009) N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, et P. Dumouchel, 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, Brighton, 1559–1562.

(Dehak et al., 2007) N. Dehak, P. Dumouchel, et P. Kenny, 2007. Modeling prosodic features with joint factor analysis for speaker verification. *Audio, Speech, and Language Processing 15*(7), 2095–2103.

(Doddington et al., 1998) G. Doddington, W. Liggett, A. Martin, M. Przybocki, et D. Reynolds, 1998. Sheep, goats, lambs and wolves : A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*, Sydney, 1351–1354.

(Doddington et al., 2000) G. R. Doddington, M. A. Przybocki, A. F. Martin, et D. A. Reynolds, 2000. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication 31*(2-3), 225–254.

(Ducrot et Schaeffer, 1995) O. Ducrot et J.-M. Schaeffer, 1995. *Nouveau dictionnaire encyclopédique des sciences du langage*. Paris : Seuil.

(Erikson, 2007) E. J. Erikson, 2007. *That voice sounds familiar : Factors in speaker recognition*. Thèse de Doctorat, Umea University.

(Eriksson, 2006) A. Eriksson, 2006. Charlatanry and fraud–an increasing problem for forensic phonetics ? Dans les actes de *Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, Göteborg, 10–11.

(Fant, 1970) G. Fant, 1970. *Acoustic theory of speech production*. Berlin : Mouton De Gruyter.

(Fauve et al., 2008) B. Fauve, N. Evans, et J. Mason, 2008. Improving the performance of text-independent short duration SVM-and GMM-based speaker verification. Dans les actes de *IEEE/ISCA Speaker Odyssey*, Stellenbosch.

(Fauve et al., 2007) B. Fauve, N. Evans, N. Pearson, J. F. Bonastre, et J. Mason, 2007. Influence of task duration in text-independent speaker verification. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, Antwerp, 794–797.

(Ferrer et al., 2010) L. Ferrer, N. Scheffer, et E. Shriberg, 2010. A comparison of approaches for modeling prosodic features in speaker recognition. Dans les actes de *International Conference of Acoustics Speech and Signal Processings*, Dallas, 4414–4417.

(French et Harrison, 2007) P. French et P. Harrison, 2007. Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by peter french & philip harrison. *International Journal of Speech Language and the Law 14*(1), 137–144.

(Furui, 1981) S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. *Acoustic, Speech and Signal processings 29*, 254 – 272.

(Galliano et al., 2005) S. Galliano, E. Geoffroy, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier, 2005. The ester phase II evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Lisboa, 1149–1152.

(Galton, 1892) F. Galton, 1892. *Finger Prints*. Londres : Macmillan.

(Garde, 1954) E. Garde, 1954. *La voix*. Paris : Presses Universitaires de France.

(Garrido et al., 2009) L. Garrido, F. Eisner, C. McGettigan, L. Stewart, D. Sauter, J. R. Hanley, S. R. Schweinberger, J. D. Warren, et B. Duchaine, 2009. Developmental phonagnosia : A selective deficit of vocal identity recognition. *Neuropsychologia 47*(1), 123–131.

(Gauvain et Lee, 1994) J.-L. Gauvain et C.-H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. Dans les actes de *International Conference in Acoustic Speech and Signal Processings (ICASSP)*, Adelaide, 291–298.

(Gendrot et Adda-Decker, 2007) C. Gendrot et M. Adda-Decker, 2007. Impact of duration and vowel inventory size on formant values of oral vowels : an automated formant analysis from eight languages. Dans les actes de *International Conference in Phonetic Sciences (ICPhS)*, Saalzburg.

(Goggin et al., 1991) J. Goggin, C. Thompson, G. Strube, et L. Simental, 1991. The role of language familiarity in voice identification. *Memory and Cognition 19*, 448–458.

(González et al., 2011) I. Q. González, M. A. B. León, P. Belin, Y. Martínez-Quintana, L. G. García, et M. S. Castillo, 2011. Person identification through faces and voices : An ERP study. *Brain Research 1407*, 13 – 26.

(Gordon et al., 2002) M. Gordon, P. Barthmaier, et K. Sands, 2002. A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association 32*(2), 141–174.

(Goren et al., 1975) C. Goren, M. Sarty, et P. Wu, 1975. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics 56*(4), 544.

(Gravier, 2011) G. Gravier, 2011. Spro, a speech signal processing toolkit. https ://gforge.inria.fr/projects/spro.

(Gravier et al., 2004) G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffroy, K. Mc Tait, et K. Choukri, 2004. The ESTER evaluation campaign of rich transcription of french broadcast news. Dans les actes de *International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, 885–888.

(Greenberg et al., 2011a) C. Greenberg, A. Martin, B. Barr, et G. Doddington, 2011a. Report on performance results in the nist 2010 speaker recognition evaluation. Dans les

actes de *International Conference of the International Speech Communication Association (Interspeech)*, Florence, 261–264.

(Greenberg et al., 2011b) C. Greenberg, A. Martin, G. Doddington, et J. Godfrey, 2011b. Including human expertise in speaker recognition systems : report on a pilot evaluation. Dans les actes de *International Conference of Acoustic, Speech and Signal Processing (ICASSP)*, Prague, 5896–5899.

(Haton et al., 2006) J. Haton, C. Cerisara, D. Fohr, Y. Laprie, et K. Smaïli, 2006. *Reconnaissance automatique de la parole, Du signal à son interprétation*. Paris : Dunod.

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustic Society of America 51*, 1738–1752.

(Hermansky et al., 1991) H. Hermansky, N. Morgan, A. Bayya, et P.Kohn, 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Gènes, 329–332.

(Higgins et al., 1991) A. Higgins, L. Bahler, et J. Porter, 1991. Speaker verification using randomized phrase prompting. *Digital Signal Processing 1*, 89–106.

(Hollien, 1990) H. Hollien, 1990. *The acoustics of Crime-the New Science of Forensic Phonetics*. Dordrecht : Springer.

(Hollien et Doherty, 1982) H. Hollien et E. Doherty, 1982. Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics 10*, 139–148.

(Kahn et Rossato, 2009) J. Kahn et S. Rossato, 2009. Do human and automatic systems use the same information to identify speaker ? Dans les actes de *Conference of the International Speech Communication Association (Interspeech)*, Brigthon, 2375–2378.

(Kenny et al., 2005) P. Kenny, G. Boulianne, P. Ouellet, et P. Dumouchel, 2005. Factor analysis simplified. Dans les actes de *International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP)*, Philadelphie, 637–640.

(Kenny et Dumouchel, 2004) P. Kenny et P. Dumouchel, 2004. Experiments in speaker verification using factor analysis likelihood ratios. Dans les actes de *ODYSSEY04-The Speaker and Language Recognition Workshop*, Toledo, 219–226.

(Kersta, 1962) L. G. Kersta, 1962. Voiceprint identification. *Nature 196*, 1253–1257.

(Klusacek et al., 2003) D. Klusacek, J. Navratil, D. A. Reynolds, et J. P. Campbell, 2003. Conditional pronunciation modeling in speaker detection. Dans les actes de *International Conference in Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 804–808.

(Kockmann et al., 2010) M. Kockmann, L. Burget, et J. Cernocky, 2010. Investigations into prosodic syllabe contour features for speaker recognition. Dans les actes de *International Conference in Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, 4418–4421.

(Kunzel, 1994) H. J. Kunzel, 1994. Current approaches to forensic speaker recognition. Dans les actes de *Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 135–141.

(Labov, 1972) W. Labov, 1972. *Sociolinguistic patterns*. Philadelphia : University of Pennsylvania Press.

(Ladefoged, 1980) P. Ladefoged, 1980. The ability of listeners to identify voices. *Phonetics 49*, 43–55.

(Ladefoged, 2005) P. Ladefoged, 2005. *Vowels and consonants an Introduction to the sounds of Languages*. Oxford : Wiley-Blackwell.

(Laird, 1993) N. Laird, 1993. The em algorithm. Dans C. Rao (Ed.), *Handbook of statistics*. Oxford : Elsevier.

(Lamel et al., 1991) L. Lamel, J. Gauvain, et E. M., 1991. BREF, a large vocabulary spoken corpus for French. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Gènes, 505–508.

(Lancker et Canter, 1982) D. R. V. Lancker et G. J. Canter, 1982. Impairment of voice and face recognition in patients with hemispheric damage. *Brain and cognition 1*(2), 185–195.

(Larcher, 2009) A. Larcher, 2009. *Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

(Larcher et al., 2010) A. Larcher, C. Lévy, D. Matrouf, et J.-F. Bonastre, 2010. LIA NIST-SRE'10 systems. Dans les actes de *NIST-SRE Workshop*, Brno.

(Laver, 1980) J. Laver, 1980. *The Phonetic Description of Voice Quality.* Cambridge : Cambridge University Press.

(Lavner et al., 2000) Y. Lavner, I. Gath, et J. Rosenhouse, 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication 30*, 9–26.

(Lavner et al., 2001) Y. Lavner, J. Rosenhouse, et I. Gath, 2001. The prototype model in speaker identification by human listeners. *International Journal of Speech Technology 4*(1), 63–74.

(Legge et al., 1984) G. E. Legge, C. Grosmann, et C. M. Pieper, 1984. Learning unfamiliar voices. *Journal of Experimental Psychology : Learning, Memory, and Cognition 10*, 298–303.

(Levine et Hullett, 2002) T. Levine et C. Hullett, 2002. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research 28*(4), 612–625.

(Linares et al., 2007) G. Linares, P. Nocera, D. Massonie, et D. Matrouf, 2007. The LIA speech recognition system : from 10xRT to 1xRT. *Lecture Notes in Computer Science 4629*, 302–308.

(Macmillan et Creelman, 1991) N. Macmillan et C. Creelman, 1991. Signal detection theory : A user's guide.

(Magrin-Chagnolleau et al., 1995) I. Magrin-Chagnolleau, J. Bonastre, et F. Bimbot, 1995. Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Madrid, 337–340.

(Mami, 2003) Y. Mami, 2003. *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence*. Thèse de Doctorat, Ecole supérieure des télécommunications.

(Marchal, 2007) A. Marchal, 2007. *La production de la parole*. Paris : Hermès.

(Markel et Gray, 1976) J. Markel et A. H. Gray, 1976. *Linear prediction of speech*. Dordrecht : Springer.

(Martin et al., 1997) A. Martin, G. Doddington, T. Kamm, M. Ordowski, et M. A. Przybocki, 1997. The det curve in assessment of detection task performance. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, 1895–1898.

(Martin et Greenberg, 2009) A. F. Martin et C. Greenberg, 2009. NIST 2008 speaker recognition evaluation : Performance across telephone and room microphone channels. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, Brighton, 2579–2582.

(Matrouf et al., 2008b) D. Matrouf, J. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLarren, et F. Huenupan, 2008b. GMM-SVM system description : NIST SRE. Montréal.

(Matrouf et al., 2008a) D. Matrouf, J. Bonastre, et S. Mezaache, 2008a. Factor analysis multi-session training constraint in session compensation for speaker verification. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, 1421–1424.

(Matrouf et al., 2007) D. Matrouf, N. Scheffer, B. Fauve, et J. F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, Antwerp, 1242–1245.

(Mazaira et al., 2010) L. M. Mazaira, A. Álvarez, P. Gómez, R. Martínez, et C. Muñoz, 2010. Glottal Source Cepstrum Coefficients applied to NIST-SRE 2010. Dans les actes de *Jornadas de Reconocimiento Biometrico de Personas*, Huesca. http ://oa.upm.es/7905/.

(McDougall, 2006) K. McDougall, 2006. Dynamic features of speech and characterization of speakers : towards a new approach using formant frequencies. *Speech, Language and the Law 13*, p89–126.

(McGehee, 1937) F. McGehee, 1937. The reliability of the identification of voice. *Journal of General Psychology 17*, 249–271.

(Mehler et al., 1976) J. Mehler, M. Barrière, et J.-G. D., 1976. La reconnaissance de la voix maternelle par le nourrisson. *La Recherche 70*, 786–788.

(Mehler et Dupoux, 1990) J. Mehler et E. Dupoux, 1990. *Naître humain*. Paris : Odile Jacob.

(Miles et Meluish, 1974) M. Miles et E. Meluish, 1974. Recognition of mother's voice in early infancy. *Nature 252*, 123–124.

(Mirghafori et al., 2005) N. Mirghafori, A. Hatch, S. Stafford, K. Boakye, D. Gillick, et B. Peskin, 2005. ICSI's 2005 speaker recognition system. Dans les actes de *IEEE Speech Recognition and Understanding Workshop (ASRU)*, San Juan, 23–28.

(Mitchell et Delbridge, 1965) A. Mitchell et A. Delbridge, 1965. *The pronunciation of English in Australia*. Sydney : Angus and Robertson.

(Moeschler et Reboul, 1994) J. Moeschler et A. Reboul, 1994. *Dictionnaire Encyclopédique de pragmatique*. Paris : Seuil.

(Morgan, 2010) N. Morgan, 2010. ISCI system. Dans les actes de *NIST Speaker Recognition Evaluation Workshop*, Brno.

(NIST, 2004) NIST, 2004. The NIST year 2004 Speaker Recognition Evaluation plan. à partir de www.itl.nist.gov/iad/mig/tests/sre/2004/sre04/.

(NIST, 2005) NIST, 2005. The NIST year 2005 Speaker Recognition Evaluation plan. à partir de www.itl.nist.gov/iad/mig/tests/sre/2005/.

(NIST, 2006) NIST, 2006. The NIST year 2006 Speaker Recognition Evaluation plan. à partir de www.itl.nist.gov/iad/mig/tests/sre/2006/.

(NIST, 2008) NIST, 2008. The NIST year 2008 Speaker Recognition Evaluation plan. à partir de www.itl.nist.gov/iad/mig/tests/sre/2008/.

(NIST, 2010) NIST, 2010. The NIST year 2010 Speaker Recognition Evaluation plan. à partir de www.itl.nist.gov/iad/mig/tests/spk/2010/.

(NIST, 2011) NIST, 2011. NIST-SRE presentation. à partir de http ://www.itl.nist.gov/iad/mig/tests/sre/.

(Nolan, 2001) F. Nolan, 2001. Speaker identification evidence : its forms, limitations, and roles". Dans les actes de *Law and Language, Prospect and Retrospect*, Levi, 12–15.

(Nolan et Grigoras, 2005) F. Nolan et C. Grigoras, 2005. A case for formant analysis in forensic speaker identification. *Speech, Language and the Law 12*, p142–173.

(Nolan et al., 2006) F. Nolan, K. McDougall, G. D. Jong, et T. Hudson, 2006. A forensic phonetic study of 'Dynamic' sources of variability in speech : The DyViS project. Dans les actes de *Australian International Conference on speech Science and Technology*, Auckland, 13–17.

(Papcun et al., 1989) G. Papcun, J. Kreiman, et A. Davis, 1989. Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America 85*, 913–925.

(Reich et al., 1976) A. Reich, K. Moll, et J. F. Curtis, 1976. Effects of selected vocal disguises upon spectrographic speaker identification. *Journal of the Acoustical Society of America, 59*, p919–1976.

(Remez et al., 2004) R. E. Remez, S. C. Wissig, D. F. Ferro, K. Liberman, et C. Landau, 2004. A search for listener differences in the perception of talker identity. *Journal of the Acoustical Society of America 116*, 2544.

(Rey, 1964) A. Rey, 1964. *L'examen clinique en psychologie*. Paris : Presses Universitaires de France.

(Reynolds et al., 2000) D. Reynolds, T. F. Quatieri, et D. R. B, 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing 10*, p19–41.

(Reynolds, 1995) D. A. Reynolds, 1995. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication 17*(1-2), 91–108.

(Reynolds, 1997) D. A. Reynolds, 1997. Comparison of background normalization methods for text-independent speaker verification. Dans les actes de *European Conference on Speech Communication and Technology Association (Eurospeech)*, Rhodes, 963–966.

(Rogers, 1998) H. Rogers, 1998. Foreign accent in voice discrimination : a case study. *Forensic Linguistics 5*, p203–208.

(Rose, 2011) P. Rose, 2011. Forensic voice comparison with secular shibboleths- a hybrid fused gmm-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra. Dans les actes de *International Conference on Acoustic Speech and Signal Processings (ICASSP)*, Brno, 5900–5903.

(Rosenberg et al., 1992) A. E. Rosenberg, C.-H. DeLong, B.-H. Juang, et F. Soong, 1992. The use of cohort normalized scores for speaker verification. Dans les actes de *International Conference on Spoken Language Processing (ICSLP)*, Alberta, 599–602.

(Sangrigoli et al., 2005) S. Sangrigoli, C. Pallier, A. Argenti, V. Ventureyra, et S. de Schonen, 2005. Reversibility of the Other-Race effect in face recognition during childhood. *Psychological Science 16*(6), 440 –444.

(Sapir, 1927) E. Sapir, 1927. Speech as a personality trait. *The American Journal of Sociology 32*, 892–905.

(Saslove et Yarmey, 1980) H. Saslove et A. D. Yarmey, 1980. Long-term auditory memory : Speaker identification. *Journal of Applied Psychology 65*, 111–116.

(Scheffer, 2006) N. Scheffer, 2006. *Structuration de l'espace acoustique par le modèle générique pour la vérification du locuteur*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

(Scheffer et al., 2011) N. Scheffer, L. Ferrer, M. Graciarena, S. Kajarekar, E. Shriberg, et A. Stolcke, 2011. The SRI NIST 2010 speaker recognition evaluation system. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Brno, 5292–5295.

(Scherer, 1984) K. Scherer, 1984. Les émotions : fonctions et composantes. *Cahiers de psychologie cognitive 4*(1), 9–39.

(Scherer et al., 1998) K. R. Scherer, T. Johnstone, et J. Sangsue, 1998. L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole. Dans les actes de *Journée d'Etude de la Parole (JEP)*, Martigny, 249–257.

(Schiller et Köster, 1998) N. O. Schiller et O. Köster, 1998. The ability of expert witnesses to identify voices : a comparison between trained and untrained listeners. *Forensic Linguistics 5*, 1–9.

(Schlichting et Sullivan, 1996) F. Schlichting et K. Sullivan, 1996. Discrimination of imitated voices. Dans les actes de *Australian International Conference on Speech Science and Technology*, Adelaide, 103–108.

(Schlichting et Sullivan, 1998) F. Schlichting et K. P. Sullivan, 1998. Can voice imitation be detected in voice line-ups in a language unknown by the listeners ? Rapport technique, Department of Phonetics, Umea University, Sweden.

(Schmid, 2011) S. Schmid, 2011. An acoustic analysis of palatal obstruents in two romance varieties. Dans les actes de *International Conference of Phonetic Sciences (ICPhS)*, Hong Kong, 1762–1765.

(Schwartz et al., 2009) R. Schwartz, W. Shen, J. Campbell, et R. Granville, 2009. Measuring typicality of speech features in american english dialects : Towards likelihood ratios in speaker recognition casework. Dans les actes de *European Academy of Forensics Science*, Glasgow.

(Shriberg et Stolcke, 2008) E. Shriberg et A. Stolcke, 2008. The case for automatic Higher-Level features in forensic speaker recognition. Dans les actes de *International Conference on Speech Communication and Technology (Interspeech)*, Brisbane, 1509–1512.

(Shriberg et Stolcke, 2011) E. Shriberg et A. Stolcke, 2011. Language-independent constrained cepstral features for speaker recognition. Dans les actes de *International Conference of Acoustic Speech and Signal Processing (ICASSP)*, Prague, 5296–5299.

(Siqueland et DeLucia, 1969) E. Siqueland et C. DeLucia, 1969. Visual reinforcement of nonnutritive sucking in humans infants. *Science 165*, 1144–1146.

(Sjostrom et al., 2006) M. Sjostrom, E. Eriksson, E. Zetterholm, et K. Sullivan, 2006. A switch of dialectas disguise. *Working Papers Departement of Linguistics and Phonetics, Lund University 52*, p113–116.

(Spence, 1996) M. Spence, 1996. Young infants Long-Term auditory memory : Evidence for changes in preference as a function of delay. *Developmental Psychobiology 29*, p685–695.

(Spence et Freeman, 1996) M. Spence et M. Freeman, 1996. Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice* 1. *Infant Behavior and Development 19*(2), 199–212.

(Stevens et al., 1957) K. N. Stevens, C. E. Williams, J. R. Carbonnel, et B. Woods, 1957. Speaker authentification and identification : a comparison of graphic and auditory presentations speech material. *Journal of the Acoustical Society of America 44*, 1596–1607.

(Stevens et al., 1937) S. Stevens, J. Volkmann, et E. Newman, 1937. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *Journal of Acoustic Society of America 8*(3), 185–190.

(Stoll et Doddington, 2010) L. Stoll et G. Doddington, 2010. Hunting for wolves in speaker recognition. Dans les actes de *ISCA-IEEE Speaker Odyssey*, Brno, 159–162.

(Sullivan et Schlichting, 2000) K. P. H. Sullivan et F. Schlichting, 2000. Speaker discrimination in a foreign language : First language environment, second language learners. *Forensic Linguistics : The International Journal of Speech, Language and the Law 7*, 1350 – 1771.

(Sussman et al., 1992) H. M. Sussman, K. A. Hoemeke, et H. A. McCaffrey, 1992. Locus equations as an index of coarticulation for place of articulation distinctions in children. *Journal of speech and hearing research 35*(4), 769.

(Teston, 2004) B. Teston, 2004. L'évaluation instrumentale des dysphonies : État actuel et perspectives. Dans A. Giovanni (Ed.), *Le bilan d'une dysphonie : état actuel et perspectives*, 244. Marseille : Solal.

(Toda, 2009) M. Toda, 2009. *Étude articulatoire et acoustique des fricatives sibilantes*. Thèse de Doctorat, Université Paris III.

(Vaissière, 2006) J. Vaissière, 2006. *La phonétique*. Paris : Presses Universitaires de France.

(Valentine et al., 1995) T. Valentine, P. Chiroro, et R. Dixon, 1995. An account of the own-race bias and the contact hypothesis based on a" face space" model of face recognition. Dans les actes de *Cognitive and computational aspects of face recognition : Explorations in face space*, 69–94. Londres : Routledge.

(Van Dommelen, 1987) W. Van Dommelen, 1987. The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech 30*, 325–338.

(Van Lancker et Kreiman, 1987) D. Van Lancker et J. Kreiman, 1987. Voice discrimination and recognition are separate abilities. *Neuropsychologia 25*(5), 829–834.

(Van Lancker et al., 1985) D. Van Lancker, J. Kreiman, et K. Emmorey, 1985. Familiar voice recognition : patterns and parameters part I : Recognition of backward voices. *Journal of Phonetics 13*, p19–38.

(Wan et Campbell, 2000) V. Wan et W. Campbell, 2000. Support vector machines for speaker verification and identification. Dans les actes de *Workshop Neural Networks for Signal Processing*, Sydney, 775–784.

(Wonnacott et Wonnacott, 1991) T. Wonnacott et R. Wonnacott, 1991. *Statistique Economie, Gestion, Sciences, Médecine*. Oxford : John Willey and Sons.

(Yu et al., 2001) P. Yu, M. Ouaknine, J. Revis, et A. Giovanni, 2001. Objective voice analysis for dysphonic patients : : A multiparametric protocol including acoustic and aerodynamic measurements. *Journal of voice 15*(4), 529–542.

(Zetterholm, 2006) E. Zetterholm, 2006. Same speaker : different voices : A study of one impersonator and some of his different imitations. Dans les actes de *Australian International Conference on Speech Science and Technology*, Auckland, 70–75.

(Zwicker et Feldtkeller, 1981) E. Zwicker et R. Feldtkeller, 1981. *Psychoacoustique : l'oreille, récepteur d'information*. Paris : Masson.