Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile

Frederic Aman

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Frédéric AMAN

Thèse dirigée par **Michel VACHER**

préparée au sein **Laboratoire d'Informatique de Grenoble (LIG)**
et de **Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile

Thèse soutenue publiquement le **9 décembre 2014**,
devant le jury composé de :

Mme Christine VERDIER
Professeur des Universités, Université Joseph Fourier, Grenoble 1, LIG, Président

Mme Martine ADDA-DECKER
Directeur de Recherche, CNRS, Laboratoire de Phonétique et Phonologie, Paris, Rapporteur

M. Jean-François BONASTRE
Professeur des Universités, Université d'Avignon, LIA, Rapporteur

M. Vincent RIALLE
Maître de Conférences-Praticien Hospitalier, Université Joseph Fourier, Grenoble 1, AGIM, Examinateur

M. Jacques DUCHENE
Professeur des Universités, Université de Technologie de Troyes, Examinateur

M. Alain ANFOSSO
Ingénieur, CSTB, Nice, Examinateur

M. Michel VACHER
Ingénieur de Recherche CNRS HDR, LIG, Directeur de thèse

Mme Solange ROSSATO
Maître de Conférences, Université Stendhal, Grenoble 3, LIG, Co-Encadrant de thèse

Résumé

Environ un tiers de la population française aura plus de 65 ans à l'horizon 2050. Face au manque de places dans les institutions spécialisées pour personnes âgées, le maintien à domicile le plus longtemps possible est un enjeu sociétal et économique majeur qui gagnerait à bénéficier d'une assistance technologique pour soulager le travail des aidants. C'est le but poursuivi par la Maison Intelligente qui est une résidence équipée de technologie informatique pour assister ses habitants dans les situations diverses de la vie domestique aussi bien sur le plan du confort que celui de la sécurité. La reconnaissance automatique de la parole (RAP) pourrait être un apport essentiel dans la détection des situations anormales qui constitue un point essentiel d'un système de surveillance à domicile.

C'est pourquoi, le but des travaux de cette thèse est d'inclure dans le milieu de vie de la personne âgée dépendante et isolée à domicile un système de RAP capable de reconnaître des appels vers les proches ou les aidants et de détecter des appels de détresse prononcés par la personne âgée. Pour ce faire, il sera nécessaire d'adapter les techniques de RAP aux caractéristiques particulières de la voix de ces personnes.

En effet, des études ont montré que les performances des systèmes de RAP diminuent avec les voix âgées car les modèles des systèmes de RAP existants sont majoritairement appris avec des corpus de parole non âgée. De plus, les corpus d'apprentissage sont pour la plupart prononcés de façon neutre et enregistrés dans des conditions idéales. Cependant, en situation réelle, nous sommes loin des conditions idéales, et les appels de détresse seront aussi prononcées de manière expressive. Pourtant, si de nombreuses études portent sur la reconnaissance automatique des émotions, très peu d'études ont été réalisées pour évaluer les performances des systèmes de RAP dans le cas d'une parole exprimée avec des émotions fortes. L'apprentissage des modèles acoustiques et l'évaluation des systèmes de RAP nécessitent des corpus spécifiques adaptés à la tâche et aux locuteurs visés. Or nous constatons l'inexistence de corpus de parole âgée en français adapté au contexte applicatif, c'est-à-dire comprenant d'une part des appels à l'aide mais aussi vers les aidants.

C'est pourquoi, les recherches présentées dans ce manuscrit s'appuient sur trois corpus que nous avons enregistrés. Le premier, AD80, a été constitué à partir d'enregistrements de phrases adaptées à la situation, lues aussi bien par des personnes jeunes que des personnes âgées en institution. Le second est constitué d'entretiens avec des personnes âgées (parole spontanée) tandis que le troisième est composé de voix émues actées (détresse) enregistrées en laboratoire. Une étude phonémique et une étude prosodique des différences entre la voix jeune et la voix âgée ont montré une plus grande dispersion des résultats pour la voix âgée mais aussi la possibilité dans la plupart des cas d'améliorer les performances. Ces résultats nous ont ensuite permis de développer un système de RAP adapté à la tâche qui a été évalué sur corpus. Ce système a été évalué ensuite sur des données enregistrées pendant une expérimentation en situation réelle incluant des chutes jouées dans l'appartement de test DOMUS du LIG par des personnes jeunes et âgées.

Mots clés : Reconnaissance automatique de la parole, voix âgée, voix émue, dépendance, environnements perceptifs, Assistance à la Vie Autonome (AAL).

Abstract

About a third of the French population will be over 65 years old to 2050. Due to the lack of places in specialized institutions for the elderly, the long term domestic support is a social and economic challenge, which deserves to benefit from technological assistance to facilitate the work of caregivers. It is the aim of the Smart Home which is a residence equipped with computing technology to assist people in various situations of domestic life, in terms of comfort and safety. The automatic speech recognition (ASR) could be an essential contribution to the detection of abnormal situations, which is an essential point of a home monitoring system.

Therefore, the aim of this thesis work is to include into the living environment of the dependent and isolated elderly an ASR system. This system will recognize the calls to relatives or caregivers and detect distress calls uttered by the elderly. To do this, it will be necessary to adapt the ASR system to the particular characteristics of the elderly voice.

Furthermore, studies have shown that the performances of ASR systems decrease with aging voices. Indeed, models of existing systems of ASR are mostly learned with non-elderly speech. In addition, the training corpora are mostly spoken neutrally and recorded under ideal conditions. However, in real life, we are far from these ideal conditions, and distress calls will also be expressively pronounced. However, while many studies focus on the automatic recognition of emotions, very few studies have been conducted to evaluate the performance of automatic speech recognition systems in the case of speech spoken under stressful conditions. The learning of acoustic models and the evaluation of ASR systems require specific corpora adapted to the task and the speakers. Nevertheless, we see the lack of French speech corpora recorded by elderly adapted to the application context, with calls to relatives as well as distress calls.

Therefore, the research presented in this manuscript is based on three corpora that we recorded. The first corpus, AD80, has been formed from recordings of sentences appropriate to the situation, read by young and elderly people. The second one consists of interviews with seniors (spontaneous speech) while the third one is made up of emotional speech (distress) recorded in the laboratory. A phonemic analysis and a study of prosodic differences between the young and the elderly voices demonstrated a greater variability of results for the aging voice but also the possibility in most cases to improve the performance. These results allowed us to develop an ASR system adapted to the task. The system was then tested on data recorded during an experiment in a real situation, including falls played by young and elderly people, in the DOMUS test apartment.

Remerciements

Je tiens à remercier mon directeur de thèse Michel Vacher et ma co-encadrante Solange Rossato pour leurs conseils, leur soutien et leur disponibilité tout au long de ma thèse. Leur expérience et leur passion ont permis de guider de façon stimulante le déroulement de cette thèse. Je les remercie également pour leurs efforts de relecture du manuscrit.

Je remercie aussi François Portet pour ses nombreuses idées qui ont aidé à donner les directions scientifiques de ma thèse, et pour ses contributions dans le développement des logiciels GEOD et CirdoX.

Je remercie Véronique Aubergé pour ses encouragements, pour son aide dans la mise en place du protocole de recueil de corpus de voix de détresse, ainsi que pour ses contributions scientifiques dans la thèse rendues possibles par son immense culture dans les sciences du langage.

Je remercie également Laurent Besacier et tout le groupe de recherche GETALP de m'avoir accueilli dans leur équipe.

J'aimerais remercier Martine Adda-Decker et Jean-François Bonastre pour avoir accepté d'être les rapporteurs de ma thèse. Je remercie également Chritine Verdier qui a accepté d'être le président du jury, ainsi que Vincent Rialle, Jacque Duchêne et Alain Anfosso pour leur participation au jury comme examinateurs.

Je remercie ma sœur Caroline et tout le personnel de l'EHPAD Château de Labahou et du SSR Les Cadières de m'avoir accueilli dans leur établissement pour réaliser les enregistrements de corpus de parole.

Je remercie de plus tous les participants qui ont accepté d'être enregistrés dans le cadre de ma collecte de corpus.

Je remercie également Sarah Samson Juan et Benjamin Lecouteux pour leur soutien technique sur les systèmes de reconnaissance automatique de la parole. Merci également à Sarah pour ses relectures attentives de mes articles en anglais.

Je tiens aussi à remercier Elodie, qui, en me transférant l'offre d'emploi pour cette thèse, a été le point de départ de cette aventure.

Un grand merci aux doctorants, post-doctorants, stagiaires et ingénieurs de l'équipe GETALP et MRIM, restés plus ou moins longtemps au laboratoire, qui ont créé une bonne ambiance de travail et permis de partager des moments de convivialité très agréables : Sarah, Elodie, Nadia, Juline, Pathatai, Uyanga, Marion, Manuela, Yuko, Johann, David, Guillaume, Pedro, Remus, Marwen, Quang, Sohnoun, Andrew, Issam, Mateusz, Cheick et Shachar.

Merci à Vanildo pour le traditionnel café après le resto U.

Je remercie également mes parents, Cécile, Elodie, Cécilia, Pathatai, Sarah et Mateusz pour avoir organisé mon pot de thèse.

Enfin, un grand merci à ma famille et mes amis qui m'ont soutenu et encouragé durant ma thèse.

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

1.1 Motivation

Les pays développés sont engagés dans un processus de transition démographique marqué par un accroissement de la population âgée résultant de l'augmentation de l'espérance de vie. Le vieillissement de la population amène à une augmentation du nombre de personnes dépendantes, la notion de dépendance étant basée sur une déficience des fonctions physiques, sensorielles et cognitives, sur une incapacité d'effectuer certaines activités de la vie quotidiennes, et sur l'altération de la vie sociale. Il en résulte un besoin en aide et assistance par un tiers pour les soins et les tâches de la vie courante.

La prise en charge de la population vieillissante est un enjeu sociétal majeur. Cette question est d'ailleurs au cœur des préoccupations du gouvernement français actuel. Le projet de loi sur l'*adaptation de la société au vieillissement* devrait être votée avant la fin de l'année 2014. Ce projet de loi vise à couvrir toute les dimensions de la prise en compte du vieillissement par la société. Le site web du Ministère de Affaires Sociales et de la Santé [1] présente le projet ainsi : « *Il s'agit, pour le Gouvernement, de répondre à une demande forte de nos concitoyens et d'anticiper les conséquences du vieillissement de la population sur la vie sociale et les politiques publiques dans leur ensemble. En effet, en 2060, un tiers des Français aura plus de 60 ans. Les personnes âgées de plus de 85 ans seront près de 5 millions, contre 1,4 million aujourd'hui. Il est essentiel de rappeler que 83% des plus de 85 ans vieillissent sans perte d'autonomie. Aujourd'hui, il est nécessaire que notre société s'adapte pour garantir, au fur et à mesure, de l'avancée en âge la meilleure vie possible et soutenir la solidarité familiale.* »

Certains domaines de recherche du Laboratoire d'Informatique de Grenoble sont concernés par ce projet de loi, car parmi les mesures proposées figurent la création d'une aide publique permettant l'accès aux technologies nouvelles (domotique, numérique, téléassistance) pour les personnes âgées à faibles revenus, et l'organisation d'ateliers de prévention de chutes. Ce projet de loi cherche également à favoriser l'innovation technologique et la production en France d'équipements domotiques adaptés aux besoins des personnes âgées.

Le gouvernement actuel n'est pas le premier gouvernement à s'intéresser à la problématique du vieillissement. Au cours des dernière années, de nombreuses lois et des plans na-

1. http ://www.social-sante.gouv.fr/espaces,770/personnes-agees-autonomie,776/dossiers,758/adaptation-de-la-societe-au,2971/

tionaux ont été créés : journée de solidarité, plan « Vieillissement et solidarités » (2004-2007), plan « Solidarité grand âge » (2006), plan « Bien vieillir » (2007-2009), plan national « d'action concertée pour l'emploi des seniors » (2006-2010), plan « d'amélioration de la qualité de vie des personnes atteintes de maladies chroniques » (2007-2011), plan « Alzheimer » (2008-2012), création d'une allocation personnalisée d'autonomie, d'une allocation de solidarité aux personnes âgées et d'une allocation supplémentaire d'invalidité.

Ainsi, favoriser le bien-être, l'autonomie et la sécurité des personnes âgées est un concept qui a émergé avec l'augmentation de l'espérance de vie. Le maintien à domicile est souvent la préférence des personnes âgées et des familles, et se révèle généralement moins coûteux pour la collectivité. Le vieillissement de la population est également porteur de croissance, avec la génération d'un développement économique centré sur les besoins des personnes âgées, avec la création d'emplois et de services.

Dans ce contexte, de grands espoirs sont mis sur les TIC (Technologies de l'Information et de la Communication) pour améliorer l'autonomie, la santé, le confort, la sécurité et la vie sociale des personnes âgées tout en les maintenant à domicile. De nombreuses études et appels à projets ANR (Agence Nationale de la Recherche) et Européens promouvant des recherches dans ce domaine ont été initiés. Par ailleurs, des sociétés cherchent à répondre aux problèmes liés au vieillissement en commercialisant des produits tels que des téléalarmes, des visiophones, des robots compagnons, des jeux cognitifs, etc. L'habitat intelligent pourrait dans ce sens être une solution pour maintenir à domicile les personnes âgées.

Les maison intelligentes, ou *smart homes*, se reposent sur une intelligence ambiante permettant une analyse du comportement de l'utilisateur à partir de capteurs pour le déclenchement d'une assistance dans les diverses situations de la vie quotidienne. Cette assistance vise à améliorer le confort et le bien-être de l'habitant grâce à des interfaces naturelles permettant de contrôler les différents éléments de la maison tels que la lumière, la température, les stores, la télévision, la radio, etc. Pour les personnes âgées et les personnes à mobilité réduite, ces dispositifs doivent être prévus et personnalisés pour compenser au mieux les handicaps. Les maisons intelligentes permettent également la mise à disposition de dispositifs d'interaction sociale avec le monde extérieur grâce aux réseaux sociaux ou à la visioconférence, pour faciliter le contact avec les proches, et pour la mise en relation avec les aidants ou les cliniciens. Aussi, un but essentiel des maisons intelligentes est d'agir sur la sécurité des habitants en prévenant et détectant les situations de détresse et de risque, cela étant d'autant plus pertinent pour les personnes fragilisées, comme les personnes âgées. Du fait du manque de familiarité pour les TIC généralement constaté chez les personnes âgées, il est indispensable de concevoir des interfaces très faciles d'utilisation et naturelles pour la communication entre les utilisateurs et les appareils. La parole étant le moyen de communication le plus naturel, un dispositif de reconnaissance automatique de la parole dans la maison intelligente peut donc être pertinent pour la commande des appareils domotiques et pour les appareils permettant de faciliter et maintenir le lien social. La reconnaissance vocale peut également être utilisée pour détecter les appels de détresse lorsque la personne estime qu'elle est en situation de risque.

1.2 Le projet CIRDO

Le projet *CIRDO*[2] financé par l'ANR a donné un cadre à ce travail de thèse. Ce projet, selon ses concepteurs, vise à mettre au point un « Compagnon Intelligent Réagissant au Doigt et à l'Œil », qui constituera un produit de télélien social augmenté et automatisé par l'intégration de services innovants (reconnaissance automatique de la parole, analyse de situations ou de scènes dans un environnement complexe non contrôlé) dans le but de favoriser l'autonomie et la prise en charge par les aidants des patients âgés en perte d'autonomie, et d'améliorer la sensation de sécurité. Cette technologie ambiante installée à domicile permettra, grâce à des systèmes audio et vidéo, de détecter automatiquement les chutes et d'alerter les secours le cas échéant (Bouakaz et coll., 2014).

Le point fort de ce projet est que, par contraste avec les systèmes existants utilisant des capteurs tels que des gyroscopes (Nyan et coll., 2008) ou des accéléromètres, le système développé n'utilisera aucun capteur porté par la personne.

Ce projet a également pour but de permettre la validation de technologies génériques, une évaluation psychologique et ergonomique portant sur les usages des services développés (concernant l'utilité, l'utilisabilité, l'accessibilité, l'acceptation, les aspects éthiques, le modèle économique, etc.) ainsi que des enquêtes critiques des connaissances acquises par les professionnels des services à la personne.

Ainsi, différentes problématiques sont traitées par le projet *CIRDO* et sont décrites dans (Vacher et coll., 2014a) : l'analyse et l'identification de la chute en milieu domestique reste un défi important en raison de la variabilité des profils et des fragilités des sujets, et de la configuration des milieux de vie ; en outre, l'acceptation réelle des capteurs vidéo/audio est très variable selon le profil des utilisateurs, des proches et des professionnels de santé. Les autres points soulevés par le projet concernent le traitement vidéo rendu difficile par les conditions d'éclairage non contrôlées, la difficulté de traitement dans le cas des voix atypiques et la présence de bruit, la prise de décision à partir de données incertaines, et l'interaction homme-machine.

Pour développer le système *CIRDO* et répondre aux différentes problématiques évoquées, les étapes de l'étude réalisée par les différents partenaires du projet *CIRDO* sont les suivantes :

— modélisation du processus de chute des personnes âgées au domicile et enquête sur les conditions d'acceptation du système *CIRDO* par les différents acteurs du foyer,

— conception des modèles d'analyse informatique nécessaires sur la base de cette étude,

— élaboration d'un système de surveillance complet,

— validation avec les utilisateurs ciblés.

Ainsi, l'originalité du projet *CIRDO* réside dans la combinaison des analyses audio et vidéo ainsi qu'à une conception qui s'est appuyée sur une analyse de terrain.

2. http ://liris.cnrs.fr/cirdo/doku.php

Dans le projet *CIRDO*, notre part du travail dans l'équipe GETALP du Laboratoire d'Informatique de Grenoble porte sur l'inclusion dans le milieu de vie de la personne âgée dépendante d'un système de reconnaissance automatique de la parole capable de reconnaître des appels vers les aidants (ex. : « Appelle l'infirmière ! ») ou des appels de détresse (ex. : « Au secours ! »).

L'interfaçage entre les personnes âgées et les aidants, lors d'un appel, se fera au travers du système de télélien social *e-lio*, développé par la société Technosens [3], qui est partenaire du projet *CIRDO*.

1.3 Définition du problème

Des contraintes fortes sont induites à la fois par le public visé et aussi par les conditions d'utilisation au domicile de la personne. Le signal sonore enregistré par les microphones sera fréquemment perturbé par des bruits de fond (sonneries de téléphones, télévision, appareils électroménager, bruits de rue, mélange de voix, etc.). De plus, l'acoustique de l'habitat joue sur la qualité du signal avec les phénomènes d'écho et de réverbération. Les microphones portés donnent une bonne qualité de signal mais ne peuvent être utilisés au quotidien par la grande majorité des personnes âgées car ils sont perçus comme une forte contrainte par la personne. A l'inverse, les microphones placés à demeure posent le problème de la parole distante entraînant une dégradation du signal avec l'éloignement de la personne. Un système faisant l'acquisition de l'environnement sonore en continu sera confronté à l'analyse de la parole spontanée, et devra donc être suffisamment robuste pour ne pas déclencher une alarme en présence d'une conversation proche d'un appel de détresse (ex. : « Hier, mon voisin s'est blessé, il a appelé les pompiers »).

Le système considéré est destiné à être utilisé par des personnes âgées, or il a été montré que la voix âgée est moins bien reconnue que la voix jeune par les systèmes de reconnaissance automatique de la parole (RAP). En effet, la voix âgée subit des modifications du fait du vieillissement du conduit vocal et des dégradations cognitives et motrices, conduisant à des modifications prosodiques, fréquentielles, rythmiques, à une augmentation des bruits de bouche, etc. qui peuvent venir perturber la reconnaissance automatique de la voix âgée.

Les systèmes de RAP existants ont pour hypothèse que la parole est prononcée de façon neutre, et les modèles acoustiques existants sont appris sur des corpus enregistrés dans des conditions idéales : lectures d'articles de journaux, voix de journalistes ou de conférenciers, enregistrements dans des pièces insonorisées, évitement des erreurs de prononciation, etc. Cependant, en situation réelle, nous sommes bien loin de ces conditions idéales, et les commandes de détresse seront prononcées naturellement de manière expressive. En effet, plus les commandes vocales sont reliées à quelque chose d'important pour l'utilisateur, plus le signal de la parole sera expressif. Le système de RAP en environnement domotique, en plus de devoir être efficace sur les commandes usuelles d'appels vers les aidants et les proches, se doit dans notre contexte d'être particulièrement efficace pour les commandes de détresse.

3. http ://www.technosens.fr/

Pourtant, si de nombreuses études portent sur la reconnaissance automatique des émotions, très peu d'études ont été réalisées pour évaluer les performances des systèmes de reconnaissance automatique de la parole dans le cas d'une parole exprimée avec des émotions fortes, et aucune étude n'existe pour la langue française.

L'apprentissage des modèles acoustiques et l'évaluation des systèmes de RAP nécessitent des corpus spécifiques adaptés à la tâche et aux locuteurs visés. Or nous constatons l'inexistence de corpus en français de parole âgée adaptés au contexte applicatif, c'est-à-dire contenant des appels vers les aidants et des appels de détresse.

1.4 Objectifs du travail de thèse

Notre travail de recherche portera sur l'évaluation des systèmes de reconnaissance automatique de la parole selon deux axes : la voix âgée et la voix émue. Notre objectif est de mettre en évidence les difficultés pouvant être rencontrées lorsque la voix âgée et/ou la voix émue sont analysés par un système de reconnaissance automatique de la parole. Pour cela, nous ferons une étude sur corpus de l'influence de la voix âgée et de la voix émue sur les performances des système de RAP, et nous évaluerons les techniques existantes permettant d'adapter les modèles acoustiques aux signaux décodés. De plus, nous mettrons en évidence quelles sont les données devant être utilisées pour l'adaptation acoustique. Aussi, nous chercherons quels sont les paramètres pouvant être utilisés pour prédire avec un classifieur quelles seront les performances du système de RAP pour une personne donnée, afin d'évaluer *a priori* si l'installation d'un tel système chez la personne est pertinent ou non. En outre, nous analyserons les facteurs explicatifs de la baisse de performance des systèmes de RAP avec la voix âgée et émue.

Une part importante du travail portera sur la constitution de corpus spécifiques, et les enregistrements seront effectués auprès de personnes âgées volontaires que nous recruterons dans des établissements spécialisés de type EHPAD (établissement d'hébergement pour personnes âgées dépendantes). Des personnes adultes non âgées seront également enregistrées afin d'effectuer des comparaisons entre la voix âgée et la voix non âgée. A la suite des enregistrements, une tâche importante consistera à annoter les séquences audio, c'est-à-dire effectuer la transcription manuelle des phrases prononcées.

Enfin, nous développerons une application temps-réel implémentant un système de RAP afin de détecter les phrases cibles prononcées : phrases de détresse et appels aux aidants. Une évaluation de ce système en conditions simulées, où des volontaires joueront des scénarios de détresse, sera effectuée dans un appartement de test.

1.5 Plan du manuscrit

Un état de l'art des applications en maintien à domicile sera présenté lors du chapitre 2. Dans ce chapitre, nous aborderons le rôle et l'intérêt du maintien à domicile. Ensuite, nous définirons la notion d'intelligence ambiante et nous décrirons des projets existants repré-

sentatifs du domaine. Puis nous ferons un point sur l'acceptabilité des gérontechnologies par les personnes âgées et leur entourage. Nous présenterons ensuite des projets liés à l'analyse sonore pour la reconnaissance des situations problématiques et des chutes. Enfin, nous décrirons quelques systèmes de détection de chute utilisant des techniques autres que l'analyse sonore.

Le chapitre 3 sera consacré à l'état de l'art de la reconnaissance des voix âgées et émues. Dans un premier temps, nous définirons les spécificités de la voix âgées, puis nous décrirons les travaux les plus pertinents sur la reconnaissance automatique de la parole appliquée à la voix âgée. Enfin, nous présenterons l'état de l'art des travaux sur la RAP appliquée à la voix émue.

Le chapitre 4 détaillera la méthodologie utilisée pour construire notre système de détection de phrases cibles. Nous présenterons les outils intégrés dans ce système, et terminerons par la méthodologie utilisée pour l'acquisition de nos données d'apprentissage et de test.

Le chapitre 5 présentera notre acquisition de nouveaux corpus adaptés. Ce chapitre sera dans un premier temps consacré à la description des outils existants d'acquisition de corpus, puis nous présenterons le logiciel *GEOD* que nous avons développé pour l'acquisition de corpus auprès de personnes âgées. Enfin, nous décrirons les corpus enregistrés : le corpus *AD80*, le corpus *ERES38* et le corpus *Voix Détresse*.

Dans le chapitre 6, nous présenterons dans un premier temps les principes généraux des systèmes de RAP, puis nous détaillerons nos expérimentations réalisées avec les systèmes de RAP sur nos corpus de voix âgées.

Dans le chapitre 7, nous chercherons quels sont les facteurs influançant la dégradation des performances des systèmes de RAP avec la voix âgée.

Nous décrirons dans le chapitre 8 les expérimentations réalisées avec les systèmes de RAP sur la voix émue de détresse.

Nous consacrerons le chapitre 9 à la description de notre expérimentation en situation réaliste jouée. Dans ce chapitre, nous décrirons tout d'abord le projet *CIRDO*, puis nous ferons mention de l'étude de terrain réalisée par l'équipe GRePS, partenaire du projet *CIRDO*. Ensuite, nous ferons une présentation générale du système *CIRDO* de détection de chute et d'appels de détresse/aux aidants, composé d'un module de traitement vidéo, d'un module de traitement sonore et d'un module de fusion pour la décision. Le module de traitement sonore que nous avons développé sera lui expliqué en détail. Enfin, nous décrirons l'expérimentation réalisée dans l'appartement DOMUS qui a permis l'enregistrement d'un corpus expérimental que nous utiliserons pour valider notre système de reconnaissance.

Pour finir, nous présenterons les conclusions de ce travail de thèse et nos perspectives.

CHAPITRE 2

État de l'art des applications en maintien à domicile

Dans les pays développés, nous assistons à une augmentation de la taille de la population associée à un vieillissement croissant et une augmentation de la longévité.

Au 1er janvier 2012, 65,4 millions de personnes résident en France, dont 63,5 millions en métropole et 1,9 million dans les départements d'outre-mer (hors Mayotte), soit 10 millions d'habitants de plus qu'il y a 30 ans. En 2011, comme les années précédentes, la population a augmenté de 0,5% (+ 350 000 personnes environ) et cette hausse est davantage imputable au solde naturel (+ 272 000) qu'au solde migratoire (+ 77 000) (Beaumel et Bellamy, 2013; Insee, 2012). Le dynamisme démographique repose principalement sur des naissances nombreuses et un nombre de décès encore relativement faible. La forme de la pyramide des âges est marquée par un baby-boom particulièrement important et durable (1946-1974) (Beaumel et Bellamy, 2013). Au 1er janvier 2012, la génération « 1946 », première génération nombreuse du baby-boom, qui compte plus de 200 000 personnes de plus que les précédentes, a atteint 65 ans (Insee, 2012).

En 2012, plus de 17% de la population est âgée d'au moins 65 ans, contre 14% il y a 20 ans (Insee, 2012), et 9% de la population française a plus de 75 ans (Bellamy et Beaumel, 2013). La population des plus de 60 ans devrait augmenter de 80% environ jusqu'à l'horizon de 2060 : en chiffres bruts elle atteindrait 23,6 millions (contre 13,2 en 2007), à rapporter à environ 33 millions de personnes âgées de 20 à 59 ans (Blanchet et Gallo, 2013). Selon les dernières projections démographiques de l'Insee, plus d'un habitant sur quatre aura 65 ans ou plus en 2040, et le nombre de centenaire en France passera de 15 000 en 2010 à 200 000 en 2060 (Blanpain, 2010).

La hausse de l'espérance de vie serait donc à l'origine de l'ensemble de la tendance séculaire du vieillissement global (Blanchet et Gallo, 2013). Depuis 1994, les gains moyens d'espérance de vie sont de 3 mois par an pour les hommes et de 2 mois par an pour les femmes. En 2011, à la naissance, les femmes peuvent espérer vivre jusqu'à 85 ans en moyenne et les hommes 78,4 ans (Beaumel et Bellamy, 2013).

Ce vieillissement rapide de la population mène les politiques publiques à s'interroger sur les problématiques liées aux personnes âgées : retraites, dépenses de santé, mais aussi prise en charge des personnes dépendantes, dont le nombre augmentera dans les années à venir. En effet, 1,2 millions de personnes seront dépendantes en 2040, contre 800 000 en 1999 (Duée et Rebillard, 2006).

La dépendance est définie comme le besoin d'aide des personnes de 60 ans ou plus pour

accomplir certains actes essentiels de la vie quotidienne (Duée et Rebillard, 2006). Elle est liée non seulement à l'état de santé de l'individu, mais aussi à son environnement matériel : une personne âgée se déplaçant difficilement sera très dépendante, voire confinée chez elle, si elle habite en étage dans un immeuble sans ascenseur, mais plus autonome dans le cas contraire.

L'autonomie de la personne est évaluée par les gériatres à l'aide d'une échelle définissant l'aptitude ou l'inaptitude de la personne à s'assumer elle-même dans les activités de la vie quotidienne (Katz et Akpom, 1976). Ces activités représentent ce que chacun effectue naturellement dans la vie de tous les jours, tel que se nourrir, se reposer, se laver... Une autre échelle, définie par Lawton et Brody (1969), prend en compte les instruments de la vie quotidienne (les IADL ou *Intrumented Activities of Daily Living*) et la capacité de la personne à les utiliser. La grille AGGIR (Autonomie Gérontologie - Groupes Iso-Ressources) [1] est utilisée en France pour évaluer la perte de l'autonomie physique et psychique grâce à l'observation des activités effectuées par la personne âgée seule. Cette grille permet de répartir les degrés de la dépendance en 6 groupes, appelés GIR (Groupes Iso-Ressources). Elle fait office de référence dans le cadre de l'attribution de l'APA (Allocation Personnalisée d'Autonomie).

La prise en charge de la dépendance peut se faire par solidarité familiale ou par solidarité collective. La solidarité familiale se traduit par l'aide que les proches apportent pour les activités de la vie quotidienne, et la solidarité collective se concrétise par la mise en place de prestations spécifiques pour les personnes dépendantes (Duée et Rebillard, 2006). Fin 2010, 12 millions de personnes en France ont reçu l'APA ; 61% d'entre elles vivaient à domicile et 39% étaient en établissement spécialisé (Bellamy et Beaumel, 2013).

Ainsi, la vieillesse a des conséquences importantes sur la qualité de vie des personnes âgées en terme de maladies, dégradations physiques et incapacités diverses. D'après le Collège National des Enseignants de Gériatrie (2000), la dépendance survient lorsque la personne âgée arrive dans une situation qualifiée de décompensation fonctionnelle. La décompensation fonctionnelle est provoquée par la survenue de maladies chroniques et/ou aiguës sur un terrain plus ou moins fragilisé par le vieillissement. Le concept de décompensation peut être expliqué par la figure 2.1, prenant en considération 3 éléments qui souvent se cumulent :

— les effets du vieillissement qui réduisent progressivement les réserves fonctionnelles, sans jamais à eux seuls entraîner la décompensation,

— les affections chroniques surajoutées qui altèrent les fonctions,

— les facteurs de décompensation qui sont souvent multiples et associés chez un même patient : affections médicales aiguës, pathologies iatrogènes (trouble ou maladie provoqués par un acte médical ou par les médicaments, même en l'absence d'erreur du médecin), ou stress psychologique.

Il existe néanmoins des moyens de compensation pour éviter de tomber sous le seuil de décompensation, portant sur les domaines suivants : logement, mobilité, accessibilité,

1. http ://vosdroits.service-public.fr/particuliers/F1229.xhtml

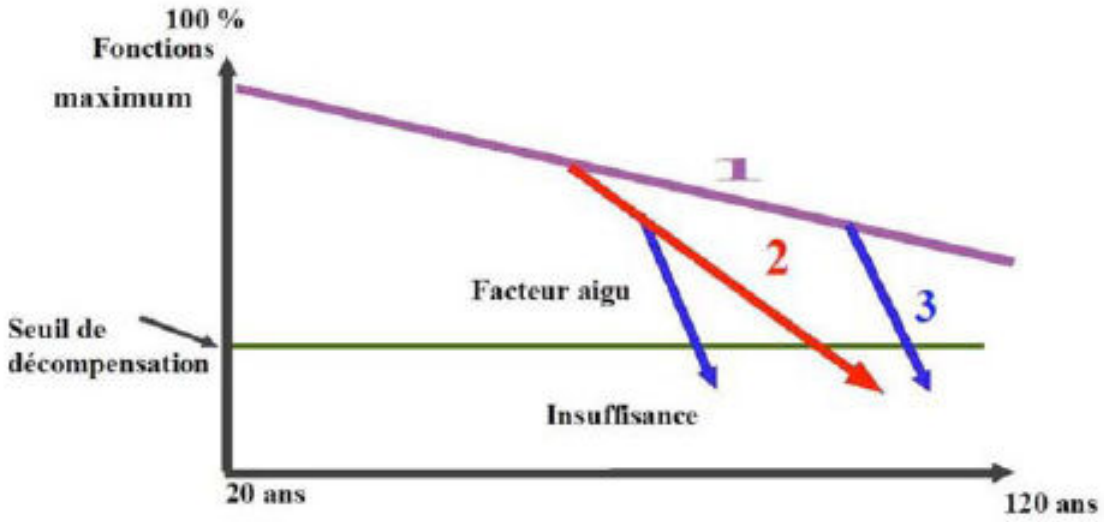


FIGURE 2.1: *« 1+2+3 » de J.P. Bouchon.*

sécurité, médicament, technologie, soin, emploi, formation et loisir (Franco, 2012).

Aussi, d'après Franco (2012), à âge constant, une personne de 80 ans est en meilleure santé en 2013 qu'en 1900. Vieillir en bonne santé est permis grâce aux progrès de la médecine avec la découverte de nouveaux traitements, médicaments et technologies médicales, et grâce aux innovations dans les sciences des matériaux, la génétique, la biotechnologie, la bioinformatique et la cybersanté.

Pour les personnes moins dépendantes, le maintien à domicile est la solution choisie dans plus de huit cas sur dix. Ce mode de prise en charge a la faveur des patients et de leurs familles. Il est également encouragé par les pouvoirs publics dans une logique de maîtrise des coûts. Le maintien à domicile des personnes âgées coûte en effet globalement moins cher que le placement en institution (Kubiak et Lorraine, 2012).

D'après Bobillier Chaumon et Oprea Ciobanu (2009), il devient nécessaire de réfléchir à de nouveaux moyens de prise en charge qui peuvent compléter les modes d'assistance traditionnels (aide à domicile, aide-soignante, infirmière...), notamment en raison de leur coût et d'un déploiement encore insuffisant (manque de personnel, formation insuffisante et/ou inadaptée, complexité des démarches...). Le développement des TIC (Technologies de l'Information et de la Communication) pour l'aide à la personne âgée représente une piste prometteuse. Elles laissent en effet augurer de nouvelles opportunités pour l'assistance et la prise en charge des personnes âgées à domicile ou en institutions spécialisées (EHPAD, hôpitaux...). Déjà présentes dans de nombreux secteurs médicaux et sanitaires (télémédecine, télédiagnostic...), ces technologies se répandent de plus en plus au domicile même des individus. Par l'intégration des TIC dans l'habitat, les maisons intelligentes fournissent des moyens de surveillance et d'assistance qui ont pour but d'améliorer le confort de vie des personnes âgées, et d'alerter au bon moment dans le cas des situations problématiques.

2.1 Intelligence ambiante

L'intelligence ambiante est définie comme un environnement numérique, qui, de manière proactive, aide les gens dans leur vie quotidienne (Augusto et coll., 2010). Les maisons intelligentes créent un environnement capable de réagir en anticipant, prédisant et prenant des décisions de manière autonome. L'idée qui anime l'intelligence ambiante est que, par l'enrichissement de l'environnement par la technologie (principalement des capteurs et appareils interconnectés via un réseau), d'après les informations recueillies en temps réel et le cumul des données historiques, les décisions puissent être prises au profit des utilisateurs de cet environnement. En effet, le système a pour objectif de déduire les situations et les besoins des utilisateurs à partir des activités enregistrées. Augusto et coll. (2010) comparent alors le système à un humain, passif, qui observerait les activités dans l'attente d'apporter de l'aide si, et seulement si, cela révèle être nécessaire. D'un autre côté, la diversité des utilisateurs nécessite une interaction directe avec le système pour indiquer les préférences, besoins, etc. Ainsi, l'intelligence ambiante a pour ambition de mettre en relation différents domaines des sciences informatiques, telles que l'IHM (Interactions Homme-Machine), l'IA (Intelligence Artificielle), les réseaux, les capteurs et l'informatique ubiquitaire, nécessitant des spécialistes en architecture, informatique, électricité et électronique, et les applications liées au secteur de la santé requièrent aussi une participation de professionnels des sciences sociales, de médecine, et d'ergothérapie (Augusto et Nugent, 2006).

Les fonctionnalités installées concernent généralement quatre catégories principales (Brush et coll., 2011) :

— connectivité (ex. : contrôle centralisé du système domotique),

— média (ex. : visualisation du contenu du PC sur la télévision),

— sécurité/monitoring (ex. : caméras de surveillance accessibles à distance),

— environnement (ex. : thermostat pour économiser l'énergie, ou système d'allumage/extinction des appareils en fonction de la présence dans la pièce).

Concernant les fonctionnalités ayant une perspective santé, les systèmes envisagent d'apporter une assistance selon les axes suivants (Vacher et coll., 2013c) :

— la santé, pour le suivi de l'évolution des constantes physiologiques, de l'activité, de l'autonomie,

— la sécurité, pour prévenir et détecter les situations problématiques et dangereuses,

— le confort, pour compenser les handicaps grâce à un accès plus aisé aux appareils domestiques,

— la communication avec l'entourage et l'extérieur, essentielle pour les personnes âgées seules à domicile.

Un tel environnement permet d'étendre la durée où la personne âgée peut encore se maintenir à domicile. En effet, l'introduction de soutien technologique dans la vie de ces personnes offre la possibilité pour elles d'entreprendre des activités quotidiennes dont elles

auraient auparavant nécessité un soutien externe. Augusto et Nugent (2006) montrent que la maison intelligente peut être utilisée pour le support de personnes souffrant de troubles cognitifs afin de leur fournir un environnement protecteur, avec plus d'indépendance et de vie privée que dans une institution spécialisée. C'est ainsi que la maison intelligente a aussi été utilisée pour l'assistance aux personnes atteintes de la maladie d'Alzheimer dans le projet CASAS (Seelye et coll., 2013).

De nombreux projets sont dédiés à l'intelligence ambiante pour les personnes âgées. Quelques exemples de projets intéressants sont donnés dans les sections suivantes.

2.1.1 e-lio

e-lio[2] est une plateforme de services développé par la société grenobloise *Technosens* – partenaire du projet *CIRDO* dans lequel s'inscrit cette thèse – permettant d'améliorer le lien entre les personnes âgées, leur famille et les professionnels de santé.

Le produit se compose de trois parties : une box, qui se branche sur le téléviseur, une caméra, et une télécommande. Cette dernière ne comprend que trois gros boutons avec des formes (rond, carré, triangle) et des couleurs distinctes. Elle remplace à la fois la télécommande de la télévision et le combiné téléphonique. La principale caractéristique d'*e-lio* est sa simplicité pour faciliter son utilisation par les personnes âgées. *e-lio* s'adresse aussi bien aux personnes en établissement qu'aux personnes à domicile.

En établissement, la personne animatrice de l'établissement peut partager avec les résidents et les familles le planning des activités de la semaine et les menus des repas. Les photos des activités peuvent ensuite être envoyées de façon ciblée aux familles et affichées sur les télévisions des résidents. Les familles et les résidents ont la possibilité de s'appeler en vidéo, de partager des photos, de s'envoyer des messages, etc. Enfin, les résidents utilisent *e-lio* au quotidien pour regarder la télévision, téléphoner (audio et vidéo), écouter la radio, visionner des photos, consulter le planning des activités, etc. Au domicile de la personne âgée, *e-lio* est un outil de lien social avec les proches et facilite la communication avec les professionnels de santé prenant soin des personnes âgées au quotidien. Il offre des services de visiophonie, téléphonie, de partage de photos, de messagerie, d'agenda, de télévision, radio, jeux, etc.

2.1.2 CompanionAble

CompanionAble[3] est un projet financé par l'Union Européenne qui a démarré en 2008. Son but est l'aide à la stimulation cognitive et la gestion de la thérapie des personnes âgées souffrant de démence ou de dépression et vivant à domicile. Ce support est donné par un robot « compagnon » fonctionnant de manière collaborative avec l'environnement intelligent de l'habitat. Dans ce projet, la reconnaissance des sons est utilisée pour la détection de situations de détresse (Rougui et coll., 2009), et la reconnaissance de la parole est utilisée pour le dialogue avec le robot (Caon et coll., 2010). Le système permet également de faciliter le lien

2. http ://www.technosens.fr
3. http ://www.companionable.net

entre la personne âgée et le personnel soignant et avec les proches. En 2009, les réalisations du projet ont été :

— la mise en place d'un système de planning intelligent pour la prise des médicaments et la gestion des rendez-vous,

— la génération des contenus du système de stimulation cognitive et leur délivrance à travers de plusieurs canaux (statiques et mobiles),

— la vidéo conférence entre personnes âgées et proches ou professionnels à travers le robot ou les interfaces domotiques,

— la reconnaissance des appels de détresse en français et la localisation de la source sonore,

— la détection par l'image des poses des personnes et l'analyse des émotions,

— le développement d'un module de dialogue naturel et multimodal (synthèse et reconnaissance vocale, écrans tactiles, reconnaissance de gestes).

2.1.3 GERHOME

GERHOME[4] (Zouba et coll., 2009) est un projet du CSTB (Centre Scientifique et Technique du Bâtiment) dont le but est d'appliquer les technologies de la domotique à l'aide au maintien à domicile des personnes âgées. *GERHOME* reconstitue un appartement type, meublé et instrumenté pour évaluer des solutions facilitant le maintien à domicile des personnes âgées. Ces solutions pour le bâtiment intelligent sont constituées principalement par des capteurs qui recueillent les données d'usage des équipements (par exemple le lit, les plaques de cuisson, le four), les mouvements de la personne, ou la présence de la personne dans la pièce, et des données environnementales, comme la température, l'humidité ou la luminosité. Grâce à l'analyse de ces données (Pomponio et coll., 2012), il est possible de reconnaître et de modéliser en continue par apprentissage en temps réel l'activité quotidienne de la personne âgée dans son logement. En comparant selon plusieurs niveaux d'abstraction les modèles produits l'on peut détecter des situations inhabituelles pouvant révéler une fragilité momentanée ou naissante, ou même un risque potentiel d'accident : l'occupant se lève plus tard depuis une semaine, prend ses repas de manière décalée, oublie d'éteindre les plaques de cuisson ou de s'alimenter, ne boit pas assez... Une nouvelle étude, débuté en novembre 2013, est en cours afin de démontrer l'applicabilité et l'opérabilité d'une telle démarche.

Plusieurs projets (GERHOME, GERHOMELABS, ADHORA) ont permis d'évaluer la technologie déployée en terme de coût, de facilité et de rapidité de déploiement, de maintenabilité et d'acceptabilité sur le plan socio-médical. Les projets SIGAAL et VIVRAUDOM permettent de créer un passerelle entre les services numériques facilitant le lien social et le monitoring bienveillant de la personne isolée.

4. http ://gerhome.cstb.fr

2.1.4 House_n

House_n[5] (Intille, 2002) est une projet du MIT (Massachusetts Institute of Technology). Dans ce projet, des recherches sont menées pour concevoir et construire des lieux de vie de type *living labs*, qui sont utilisés pour étudier les technologies en contexte. Une habitation, le *PlaceLab*, a été équipée afin d'étudier le comportement et les interactions de ses habitants avec les objets et l'environnement de l'habitat. Des volontaires ont accepté de vivre à l'intérieur pendant une période de quelques jours à une semaine. Des centaines de capteurs ont été installés pour enregistrer les activités des habitants. Ces capteurs sont utilisés pour développer des interfaces innovantes pour aider les personnes à facilement contrôler leur environnement, économiser les ressources, rester mentalement et physiquement actives, et rester en bonne santé. Les capteurs sont aussi utilisés par les chercheurs pour analyser l'activité des personnes et analyser comment elles réagissent aux nouveaux appareils et systèmes de l'habitat domotique. Parmi les capteurs utilisés, on compte :

— des capteurs d'état sans fil sur les objets utilisés ou manipulés par les personnes, comme les contacts de portes, les fenêtres, les récipients de cuisine, etc.,

— des dispositifs à fréquence radio pour localiser les habitants,

— des microphones pour capturer l'information audio,

— des caméras vidéos incluant des caméras infrarouges,

— des dispositifs PocketPC pour recevoir les retours des utilisateurs,

— des capteurs embarqués attachés sur le corps des volontaires pour la surveillance biométrique.

Les volontaires n'étaient pas spécifiquement des personnes âgées ou handicapées, mais des personnes de la population active.

2.1.5 RoboCare

RoboCare[6] (Bahadori et coll., 2004) est un projet de recherche italien, démarré en 2002 et financé par le ministère de l'éducation, de l'université et de la recherche italien. Il porte sur l'utilisation de robots autonomes et de technologies domotiques pour l'implémentation d'un système d'aide à la prise en charge des personnes âgées, dans un environnement domestique ou en institution. Le projet cherche notamment à trouver quels types de services aux personnes âgées peuvent être envisagés à partir d'un système état de l'art. Ce projet est organisé en 3 tâches : développement du cadre hardware et software du système (capteurs vidéos, avec détection des mouvements et des positions des personnes et robots), étude et implémentation d'un agent superviseur (monitoring du système, diagnostique de l'état des robots, planification et supervision de leurs tâches), intégration du système robotique (programmation des robots pour leur permettre de se repérer dans l'espace, de se déplacer de façon sécurisée, de manipuler des objets, de suivre des personnes ou d'autres robots, etc.).

5. http ://architecture.mit.edu/house_n
6. http ://robocare.istc.cnr.it

Une part de l'étude a également porté sur la problématique de l'acceptabilité du système par les utilisateurs.

2.1.6 Sweet-Home

Le projet de recherche industrielle *Sweet-Home*[7] (Vacher et coll., 2013a), financé par l'ANR, s'inscrit dans le contexte du développement des maisons intelligentes permettant un contrôle accru de l'environnement domestique pour faciliter la vie des personnes en perte d'autonomie ou atteintes de handicaps moteurs.

Le but de ce projet est de définir, à partir d'une étude d'usage conduite auprès d'utilisateurs finaux, les fonctionnalités et l'ergonomie d'un système domotique ubiquitaire et attentif, capable d'interagir naturellement avec l'utilisateur.

Cette approche est abordée à travers la mise en place d'un contrôleur intelligent communiquant avec les appareils domotiques par des protocoles réseau standards, dans le but de développer un système d'assistance domotique par commandes vocales, et d'apporter plus de sécurité par la détection de situations de détresse.

Dans ce projet, un partenariat associe d'une part des équipes de recherche dans les domaines de l'usage, du traitement automatique de la parole et du son, de l'intelligence artificielle, et d'autre part une entreprise du domaine de l'informatique embarquée et des réseaux informatiques, ainsi que des entreprises et associations spécialisées dans les systèmes de communication adaptés aux personnes âgées.

Le système développé permet à la fois de réagir à des commandes vocales (ex. : « Nestor allume la lumière ») pour actionner la lumière, les stores ou la télévision, mais aussi de détecter les situations particulières de risque (porte restée ouverte). Ce système a été testé avec des utilisateurs potentiels (Vacher et coll., 2013b).

2.2 Acceptabilité

La mise en œuvre de dispositifs issus des nouvelles technologies pose le problème de leur acceptabilité par les utilisateurs concernés.

L'acceptabilité des technologies est définie par Arning et Ziefle (2007) comme l'approbation, la réception favorable et l'utilisation continue des dispositifs et des systèmes nouvellement introduits. Chen et coll. (2012) étudient l'acceptation des produits de gérontechnologie par les personnes âgées de Hong Kong. Ils se basent sur le modèle TAM (*Technology Acceptance Model*) de Davis (Davis et coll., 1989), qui est un modèle de prédiction de l'acceptabilité d'une technologie. Le but de ce modèle est de prédire l'acceptabilité d'un outil et d'identifier les modifications qui doivent être apportées au système afin de le rendre acceptable aux utilisateurs. Ce modèle postule que l'acceptabilité d'un système d'information est déterminée par deux facteurs : la perception de l'utilité et la perception de la facilité d'utilisation. Le modèle de Davis est présenté figure 2.2.

7. http ://sweet-home.imag.fr/

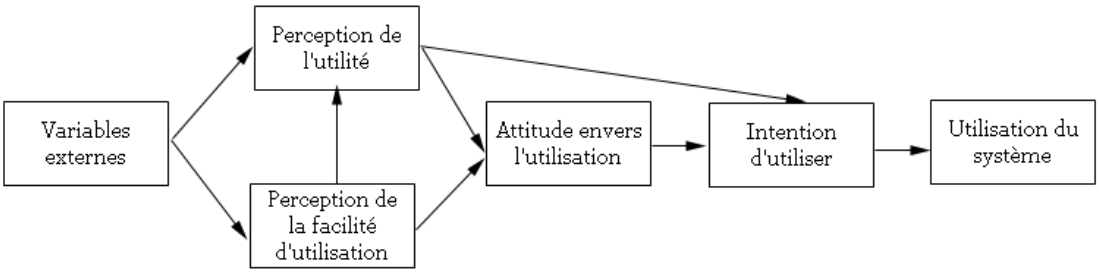


FIGURE 2.2: *Modèle d'acceptation de la technologie traduit du schéma de Davis et coll. (1989).*

En se basant sur un questionnaire, Chen et coll. (2012) ont interrogé 104 personnes âgées en institutions spécialisées et ont mesuré le taux d'adoption des technologies par ces personnes. Il en ressort que les produits électroniques grand public sont largement utilisés par les personnes âgées (télévision : 100%, téléphones mobiles : 77%, lecteurs DVD/CD : 55%, appareils photo numériques : 22%), du fait de leur usage déjà ancien dans la société. En revanche, les produits de haute technologie (systèmes domotiques, monitoring à domicile, smartphones, GPS, télémédecine, PDA, informatique) sont très peu utilisés par les personnes âgées (utilisation inexistante, sauf pour l'informatique avec un taux de 37%).

En effet, selon Chen et coll. (2012), les personnes âgées ont souvent des difficultés à adopter de nouvelles technologies et à apprendre de nouvelles compétences. Les personnes âgées utilisatrices, quant à elles, éprouvent généralement certaines difficultés dans leurs usages. En effet, elles se sentent souvent mal à l'aise lorsqu'elles utilisent des nouvelles technologies car elles peuvent avoir peur de faire des erreurs d'utilisation qu'elles ne pourraient pas corriger. De plus, le vieillissement entraîne une détérioration des capacités perceptives et cognitives, ce qui peut grandement influer sur leurs interactions avec les équipements.

En outre, les fabricants ne considèrent souvent pas les besoins et caractéristiques spécifiques des personnes âgées lors de la conception de leurs produits, et ont tendance à créer des produits ayant des performances techniques très élevées et le maximum de fonctionnalités, sans savoir si cela répond aux besoins effectifs des utilisateurs (Chen et coll., 2012). De nombreux systèmes, en particulier les détecteurs de chutes, se basent sur les caméras vidéos (Keshavarz, 2006; Zouba et coll., 2009), mais on en sait peu sur l'acceptabilité de tels capteurs par les utilisateurs ciblés, car les concepteurs n'intègrent pas toujours les utilisateurs dans la conception de ces systèmes. Pour les personnes âgées, l'équilibre doit être trouvé entre les bénéfices de cette surveillance et l'intrusion dans la vie privée. Une étude (Rialle et coll., 2008) montre que le degré d'acceptabilité d'une technologie intrusive varie selon la sévérité de la pathologie de la personne âgée, une personne avec de fortes incapacités étant plus à même d'accepter une technologie intrusive. Aussi, un critère important pour l'acceptation d'une technologie par les personnes âgées est le critère du coût, incluant le prix d'achat, et à long terme le coût de maintenance. Un service incluant un abonnement mensuel, tel qu'une téléalarme, peut être une barrière pour certaines personnes âgées.

Certains usages des technologies par les personnes âgées peuvent s'expliquer par la crainte d'être dépassé, d'être exclu de la société, de ne plus être « dans le coup ». Le cas

contraire peut aussi se manifester parmi les personnes âgées avec une moins bonne image d'elles-même, qui peuvent se considérer trop âgées pour pouvoir comprendre et maîtriser les technologies (Bobillier Chaumon et Oprea Ciobanu, 2009).

De plus, la domotique et l'assistance à domicile visant à optimiser le cadre de vie de la personne âgée, ainsi que les services web apportant à domicile des prestations et services, peuvent néanmoins, selon Bobillier Chaumon et Oprea Ciobanu (2009), jouer en la défaveur des personnes âgées car ces systèmes les stigmatisent en les identifiant comme une personne déficiente ou dépendante. Bobillier Chaumon et Oprea Ciobanu (2009) montrent également que les technologies se substituent à la personne âgée pour la réalisation de tâches élémentaires de sorte que la personne âgée n'a plus les ressources requises pour assumer seule ses activités, accentuant ses restrictions d'activité par une dépendance accrue à ces nouvelles modalités d'assistance.

En se basant sur le modèle TAM, Chen et coll. (2012) montrent néanmoins que les personnes âgées ont un point de vue positif sur l'utilité des technologies.

D'après Steele et coll. (2009) et Mahmood et coll. (2008), les personnes âgées acceptent et utilisent les nouvelles technologies si elles sont convaincues que ces technologies peuvent être utilisées pour améliorer leur vie et satisfaire leurs besoins, et, d'après McCloskey (2006) et Pan et Jordan-Marsh (2010), les personnes âgées sont susceptibles d'accepter les nouvelles technologie qui sont faciles à comprendre et ont une interface simple.

Chen et coll. (2012) montrent que la participation à des activités sociales peut faciliter l'utilisation des technologies. Les personnes âgées participant à des activités sociales peuvent ainsi recevoir, grâce aux produits connectés, des informations sur le thème de leurs activités, et partager leurs expériences avec leurs proches. Pour les personnes ayant des difficultés avec les nouvelles technologies, leurs proches peuvent ainsi leur fournir de l'aide et changer leur vision en positif envers les technologies. Ainsi, les réseaux sociaux et affectifs sont de puissants levier d'influence dans l'adoption des technologies, la famille et l'entourage intervenant souvent auprès des personnes âgées dans le processus d'achat et d'utilisation des objets techniques (Bobillier Chaumon et Oprea Ciobanu, 2009).

Chen et coll. (2012) montrent également que le niveau d'éducation et le revenu influencent positivement l'usage des technologies dédiées, et que les femmes sont plus susceptibles de les utiliser. Par ailleurs, l'expérience socioprofessionnelle des personnes âgées qui auraient fait l'usage de technologies durant leur carrière peut influencer favorablement leur compréhension et perception des TIC.

De plus, certaines personnes ayant perdu leur capacités de mobilités ont plus l'usage de technologies telles que les téléphones portables, ordinateurs et internet qui leur permettent de compenser leurs incapacités, par exemple grâce à l'utilisation de services en ligne pour les courses avec livraison à domicile (Chen et coll., 2012).

Un autre aspect sur la vie privée concerne ce qui est fait des données collectées. De par le fait que le système reçoive des informations d'une importance vitale, il doit être protégé contre les intrusions et doit s'assurer que les informations parviennent aux bonnes personnes.

2.3 Analyse sonore pour la reconnaissance des situations problématiques et des chutes

Dans le cadre du maintien à domicile des personnes âgées, la prévention et la détection des situations problématiques est un enjeu majeur pour la sécurité des personnes âgées. La chute d'une personne âgée et les conséquences sur sa santé (fractures) et sur ses émotions (peur que cela ne se reproduise) est souvent le facteur déclenchant pour sa mise en institution. Les systèmes de détection de chutes et de situations problématiques sont donc d'une part des systèmes rassurant à la fois pour les personnes âgées et pour les proches, favorisant de ce fait le maintien à domicile, et d'autre part permettent de sauver des vies grâce au déclenchement immédiat d'une alerte à destination des secours lors de l'occurrence d'un accident.

La parole est un moyen de communication naturelle pouvant être utilisée comme moyen d'interaction avec les outils technologiques. La reconnaissance automatique de la parole (RAP) est un moyen d'interaction avec l'habitat, et permet ainsi de réaliser des commandes vocales et de détecter des situations problématiques. La reconnaissance des sons de la vie courante pourrait apporter des éléments d'information pour assurer un suivi des activités quotidiennes de l'habitant.

Les travaux que nous décrivons dans cette thèse portent sur la détection des situations de détresse des personnes âgées grâce à des capteurs sonores. Nous allons donc présenter un bref état de l'art concernant cette détection selon 2 axes : la reconnaissance des sons, et la reconnaissance automatique de la parole.

2.3.1 Reconnaissance des sons dans le cadre de la détection des situations de détresse

Nous allons présenter quelques projets et systèmes utilisant la reconnaissance des sons dans la détection des situations de détresse.

L'analyse sonore s'appuie principalement sur des techniques utilisant des modèles probabilistes et des méthodes de l'intelligence artificielle (réseaux de Markov cachés, réseaux de neurones...).

Istrate et coll. (2006) présentent un système d'analyse sonore dont le but est de détecter les accidents sérieux tels qu'une chute ou un malaise, grâce à la reconnaissance en temps réel des sons de la vie quotidienne, en vue de déclencher une alarme. Le corpus utilisé pour l'apprentissage des modèles et les tests est composé de sons enregistrés par les auteurs (15%) et de sons issus de CD commerciaux utilisés pour le bruitage de films (85%). Les auteurs obtiennent comme résultat un taux d'erreur de classification de 26,82% sur des signaux artificiellement bruités selon un rapport signal sur bruit égal à +10dB. Un système plus complet, AuditHIS, permet cette fois de reconnaître aussi les paroles prononcées (Glasson, 2008). Dans ces 2 systèmes, une détection des événements sonores est effectuée pour déterminer à chaque instant s'il y a présence ou absence de signal dans le bruit de fond. Une discri-

mination son/parole détermine s'il s'agit d'un son de la vie courante ou de parole grâce à un algorithme de classification à base de GMM (*Gaussian Mixture Model*) utilisant des paramètres LFCC (*Linear Frequency Cepstral Coefficient*). Le taux d'erreur de la discrimination est de 5,1% à +10dB de RSB (Rapport Signal sur Bruit) (Fleury et coll., 2008). Puis une classification des sons produits lors de l'activité de la personne est réalisée avec une méthode GMM (Istrate et coll., 2006) ou HMM (Hidden Markov Model) (Vacher et coll., 2006). Le résultat de la classification est un taux d'erreur de 21,3% à 10 dB de RSB avec des modèles GMM (Fleury et coll., 2008). Ces performances ont été toutes deux évaluées sur des corpus de test.

Le système AuditHIS a été utilisé pour des expérimentations dans l'appartement de test de la Faculté de Médecine de Grenoble mettant en jeux une quinzaine de participants jouant des scénarios de la vie quotidienne (AVQ) incluant des appels d'urgence (Vacher et coll., 2011). Les auteurs ont observé une extrême variété des sons enregistrés dans l'appartement aussi bien produits à l'intérieur de l'appartement par le participant lui-même (sons non langagiers...), par des objets qu'il manipule (vêtements, papier, téléphone portable, appareillage électroménager...), ou à l'extérieur de l'appartement (hélicoptère, ascenseur, tonnerre, pelleteuse...). Pour beaucoup de ces sons, seuls un très petit nombre de données ont pu être enregistrées. Sur cet aspect, les auteurs en concluent que les méthodes purement statistiques ne sont pas adaptées à la reconnaissance de ces sons en milieu réel et préconisent des méthodes de classification hiérarchiques exploitant les caractéristiques physiques du signal enregistré (enveloppe temporelle et durée, caractéristiques spectrales, périodicité/non périodicité, etc.).

Le système PATSH (Vacher et coll., 2013b), permettant la détection d'ordres domotiques et d'appels de détresse, utilise également une discrimination son de la vie quotidienne/-parole par une classification utilisant des GMM. Lors de l'évaluation du système dans un appartement de test avec des participants jouant des scénarios de la vie quotidienne et utilisant une commande vocale de la domotique, les auteurs ont trouvé que 23,4% des signaux paroles étaient classifiés comme étant des sons de la vie quotidienne, et que 3,1% des sons de la vie quotidienne étaient classifiés comme étant de la parole, ce qui reste insuffisant pour un utilisation en milieu réel.

Dans les systèmes temps réel de commandes domotiques ou de détection d'appels de détresse, la discrimination son de la vie quotidienne/parole est une étape importante car elle permet de filtrer les signaux à envoyer au système de RAP (Reconnaissance Automatique de la Parole). Il est en effet indispensable que les sons de la vie courantes ne soient pas envoyés au système de RAP, auquel cas des phrases indésirables pourraient être reconnue par le décodeur. Les résultats obtenus sont bons mais devront être améliorés avant de pouvoir envisager une utilisation en réel. Par contre, la classification des sons eux-mêmes ne semble pas assez précise pour être utilisée pour une détection de détresse, sauf à combiner cette information avec d'autres de nature différente (par exemple de nature vidéo).

D'autres travaux visent à la détection de chutes par une analyse incluant la prise en compte de paramètres sonores.

Litvak et coll. (2008) utilisent un microphone et un accéléromètre placés sur le sol à un

des angles de la pièce dans le cadre d'un projet visant à détecter les chutes des personnes âgées, avec une classification capable de discriminer un évènement provoqué par une chute de la personne par rapport aux autres évènements. Leur classification se fait à partir de classifieurs utilisant des modèles basés sur une combinaison d'arbres de décisions, de kNN (*k-Nearest Neighbors*) et de SVM (*Support Vector Machine*) à partir de coefficients extraits pour le son grâce à une analyse cepstrale, et pour les vibrations de paramètres spectraux de réponse au choc et de la longueur d'onde de vibration. Le système proposé a été testé avec le mannequin articulé « Rescue Randy », les tests ont montré que ce système permet une détection des chutes de personnes avec une sensibilité de 95% et une spécificité de 95%. Popescu et coll. (2008) détectent les chutes grâce à deux microphones placés verticalement l'un au ras du sol et l'autre à environ 1,22m de haut. L'utilisation de ces 2 microphones permet de déterminer si le son a bien été émis au niveau du sol. Si c'est bien le cas et lorsque le niveau sonore est jugé suffisamment important, les auteurs utilisent un algorithme de reconnaissance utilisant les kNN sur des coefficients MFCC. Les auteurs ont montré grâce à des expérimentations impliquant un utilisateur chutant sur un tapis que l'utilisation du paramètre « hauteur » d'émission du signal est essentiel pour réduire le taux de fausses détections à 5%, la classification sonore conduisant à un taux très élevé de fausses alarmes.

Certains projets basés sur la reconnaissance des sons portent sur l'analyse des activités de la vie quotidienne. Par exemple, dans le cadre d'un projet visant à détecter le comportement des personnes souffrant de troubles cognitifs, Chen et coll. (2005) ont effectué le suivi de l'hygiène personnelle de celles-ci. Ils ont développé un système capable de reconnaître et classifier les activités se déroulant dans une salle de bain en se basant sur les sons. La classification est effectuée à partir de HMM utilisant des coefficients MFCC, et les tests présentent un taux de bonne classification de 84%. Ceci permet ensuite d'émettre un compte-rendu journalier au personnel soignant ou au médecin traitant pour faciliter son diagnostic.

Les développements actuels concernent surtout l'analyse de scènes sonores comme en témoigne le challenge IEEE ASP « Detection and classification of acoustic scenes and events » organisé en 2013 (Giannoulis et coll., 2013). L'analyse de scènes sonores y est fondée sur une approche à base de GMM et MFCC (Vuegen et coll., 2013), ou sur un modèle basé sur la perception humaine de la reconnaissance des sons (cochléogramme) (Krijnders et Gineke, 2013).

2.3.2 Reconnaissance automatique de la parole dans le cadre de la détection des situations de détresse

La reconnaissance automatique de la parole (RAP) est très étudiée et il existe de nombreuses applications grand public (Dragon NaturallySpeaking, Siri, etc.), mais, actuellement, très peu d'études portent sur son application à la détection des situations de détresse et de risque.

Dans un système temps réel de reconnaissance d'ordres domotiques et d'appels de détresse, les microphones font partie intégrante de l'habitat et ne sont pas portés par la per-

sonne. Le problème de la parole distante et du bruit complique fortement le reconnaissance par le système de RAP (Woelfel et McDonough, 2009).

Dans le système d'analyse sonore temps réel *AuditHIS*, si le signal sonore est reconnu comme étant de la parole suite à la discrimination son de la vie quotidienne/parole, une reconnaissance automatique de la parole utilisant le système Raphael (Vacher et coll., 2009a) permet de détecter l'occurrence de phrases de détresse parmi des phrases de la vie quotidienne. Les auteurs ont évalué le système sur la détection d'appels de détresse, prononcés par 10 personnes non âgées, avec comme résultat un taux de fausses alarmes de 4% et un taux d'alarmes manquées de 30%.

Vacher et coll. (2012a) utilisent le système de RAP *Speeral* (Linarès et coll., 2007) pour la reconnaissance d'ordres domotiques et d'appels de détresse à partir d'un corpus acquis en conditions réelles en environnement bruité. Ils ont utilisé un algorithme de décodage guidé pour améliorer les résultats à partir de la combinaison des plusieurs canaux sonores, et obtiennent comme résultats (tests sur la voix de 23 locuteurs non âgés) pour la détection d'ordres domotiques un taux de bonne détection de 83,5%, et pour la détection d'appels de détresse un taux de 81,2%.

Dans le système PATSH (Vacher et coll., 2013b), le décodeur *Speeral* est utilisé pour reconnaître des ordres domotiques et des appels de détresse en temps réel et dans des situations réelles. Lors de l'évaluation du système (Vacher et coll., 2013b) (16 locuteurs non âgés), les auteurs ont trouvé un Domotic Error Rate (incluant les effets de toutes les étapes : détection, discrimination son/parole et décodage) de 38%. Les mêmes auteurs parviennent à un DER=3,2% avec des méthodes plus élaborées (SGMM et décodage sur plusieurs canaux) sur des données enregistrées par des personnes âgées ou malvoyantes (Vacher et coll., 2014c).

Principi et coll. (2013) utilisent le décodeur *PocketSphinx* pour décoder des ordres domotiques et des phrases de détresse prononcées « normalement » ou de façon criée, et en situation proche ou distante du microphone (corpus de 20 locuteurs non âgés). Ils obtiennent pour les appels de détresse après adaptation acoustique au locuteur un taux de bonne reconnaissance moyen de 100% pour le cas non crié/non distant, de 95,33% dans le cas crié/non distant, 85% dans le cas non crié/distant et 72,67% dans le cas crié/distant.

Hamill et coll. (2009) présentent la phase initiale du projet de développement d'un système de détection de situations d'urgence au domicile des personnes âgées, interfacé avec un centre d'appels pour provoquer les secours si nécessaire. Ce projet vise à remplacer les boutons portés en médaillon par les personnes âgées en utilisant des techniques d'intelligence artificielle et de dialogue homme-machine pour reconnaître les situations à risque. Les travaux présentés visent à montrer la possibilité d'utiliser un réseau de microphones et un logiciel de RAP pour permettre la communication et le dialogue comme moyen d'interaction avec le centre d'appels. De plus, l'idée du projet est de réduire les fausses alarmes grâce à un système de dialogue homme-machine à questions fermées (réponse de l'utilisateur par « oui » ou « non ») permettant à l'utilisateur de choisir s'il a réellement besoin d'aide, et de déterminer quelle action entreprendre (appel des urgences, d'un contact pré-défini par l'utilisateur, tel que famille, voisin ou ami, ou connexion à un opérateur). Les auteurs

ont testé le système de RAP Sphinx4 auprès de 9 locuteurs non âgés (âge entre 20 et 30 ans) auxquels ils ont demandé de prononcer 12 fois « oui » et « non » selon 3 conditions dépendant de leur position dans la pièce et des interférences, et ont trouvé un taux de précision de 93%. En testant le système de dialogue selon différents scénarios (4 locuteurs âgés entre 20 et 30 ans), les auteurs ont obtenu un WER=21%.

2.4 Autres systèmes de détection de chute

Dans leur grande majorité, les systèmes de détection de la chute des personnes âgées utilisent des capteurs non invasifs autres que des microphones. Les systèmes vus à la section 2.3.1 doivent plutôt être considérés comme des tentatives.

Alwan et coll. (2006) présentent un système basé sur un transducteur de force piézoélectrique placé sur le sol afin de détecter l'empreinte spécifique des vibrations engendrées par la chute d'une personne sur le sol. Les résultats obtenus pour la détection de chutes de personnes (représentées par un mannequin) – par rapport à la chute d'objets autres – est de 100% de vrais positifs et de 0% de fausses alarmes. Sixsmith et Johnson (2004) ont développé le système *Simbad* permettant la détection de chutes à partir de caméras thermiques infrarouges. Ils obtiennent un taux de bonne détection de chute de 35,7%.

Les quelques exemples de projets de détection de chute que nous allons maintenant décrire utilisent des capteurs portés par la personne. Ces capteurs s'avèrent plus contraignants car la personne est astreinte à les porter tout au long de la journée (ou de l'expérience).

Bourke et Lyons (2008) utilisent un capteur gyroscopique porté par les sujets pour distinguer les chutes par rapport aux autres activités de la vie quotidienne. Pour ménager les personnes âgées lors de la phase de test, les chutes ont été réalisées par des personnes jeunes, alors que les activités de la vie quotidienne ont été réalisées par des personnes âgées. Les résultats montrent que les chutes sont distinguées des autres activités avec un taux de 100% de bonnes détections pour 480 mouvements. Degen et coll. (2003) présentent un capteur de chute placé au poignet. Ils obtiennent comme résultat un taux de bonne détection de 100% lors de chutes en avant, de 58% lors de chutes en arrière, et de 45% lors de chutes sur le côté. Lindemann et coll. (2005) détectent les chutes grâce à un accéléromètre placé sur la tête et démontrent que ce placement sur la tête donne de meilleurs résultats que s'il est placé sur le poignet ou la hanche. Enfin, Bloch et coll. (2011) évaluent le système *Vigi'Fall*, composé de la combinaison de différents capteurs portés et non portés, tels des capteurs accélérométrique, de posture et d'activité. Ils obtiennent pour la détection des chutes une sensibilité et une spécificité de 32,5% et 99,5% respectivement (voir les définitions de la sensibilité et de la spécificité en annexe G).

2.5 Conclusion

La reconnaissance des sons, du fait de la diversité des sons possibles dans l'habitat, nous semble avoir une robustesse insuffisante pour la détection des chutes. Les systèmes propo-

sés utilisent toujours la dimension sonore en complément d'un autre paramètre physique. Cependant, la classification d'un signal sonore entre son de parole et son de la vie quotidienne est essentielle pour filtrer les signaux sonores qui doivent être envoyés au système de RAP. Par ailleurs, la reconnaissance de la parole semble très prometteuse pour les applications de détection de situations de détresse bien qu'elle soit encore peu envisagée dans des applications réelles. Quant à l'analyse vidéo, se pose le problème de son acceptation par les utilisateurs.

Nous allons aborder dans le chapitre suivant l'étude de voix âgée et de la voix émue par un état de l'art de ces domaines.

CHAPITRE 3

État de l'art de la reconnaissance des voix âgées et émues

3.1 Voix âgée

Dans le contexte de la reconnaissance automatique de la parole (RAP) appliquée à des utilisateurs âgés, nous nous sommes intéressés à l'état de l'art traitant de la nature et des spécificité de la voix âgées. Nous avons examiné les travaux existants portant sur la RAP de la voix âgée. Cet état de l'art a principalement mis à profit la synthèse particulièrement riche sur les caractéristiques de la voix âgée exposée dans l'ouvrage de Kreiman et Sidtis (2011), et s'appuie également sur la thèse de Vipperla (2011).

3.1.1 Spécificités de la voix âgée

3.1.1.1 Évolution physiologique de la voix âgée

D'après l'ouvrage de Kreiman et Sidtis (2011) (p.116), dans des conditions normales, le larynx, le système respiratoire et le système articulatoire restent assez stables tout au long de l'âge adulte. A partir de l'âge d'environ 60 ans, des changements commencent à affecter le système vocal (Linville et Rens, 2001; Kent et Vorperian, 1995). D'après Woo et coll. (1992) et Hagen et coll. (1996), les changements de la voix dus au vieillissement seul sont peu importants par rapport aux changements dus à des pathologies. A l'opposé, d'après Ramig et coll. (2001), le processus général du vieillissement de l'ensemble du corps amène à un vieillissement de la voix, avec une différenciation à faire entre facteurs génétiques et facteurs environnementaux dans le vieillissement.

Avec l'augmentation de l'âge, la respiration, et donc la phonation, deviennent plus difficiles. En effet, les cartilages entre les côtes s'ossifient, et la cage thoracique devient plus rigide. Les fibres de collagène des poumons deviennent interconnectées et les poumons deviennent progressivement moins flexibles, cela entraînant une diminution de la capacité vitale des poumons de 2,4-2,7 litres comparé aux 3,5-5,9 litres chez les personnes jeunes adultes (Zemlin, 1981; Linville et Rens, 2001). De plus, la dégénérescence des muscles et la diminution du contrôle neuromusculaire avec l'âge affecte le contrôle respiratoire et phonatoire.

Le larynx baisse dans le cou tout au long de la vie, en raison de sa croissance continue, d'une atrophie des muscles et d'un étirement des ligaments qui soutiennent le larynx dans

le cou. Les cartilages du larynx subissent un processus d'ossification tout au long de la vie et le sont complètement à l'âge de 65 ans. Cependant, cet effet sur la qualité de la voix reste encore peu clair (Kreiman et Sidtis (2011), p.117).

Par ailleurs, avec le vieillissement, en particulier chez les hommes, peut apparaître une atrophie des muscles du larynx ainsi qu'une diminution de leur force contractile (Ramig et coll., 2001). Une atrophie des muscles thyro-aryténoïdiens, dont l'action contribue à modifier la tension des cordes vocales, produit une fermeture incomplète des cordes vocales qui serait compensée par une augmentation de la tension musculaire dans le larynx, provoquant une perception de souffle dans la voix, surtout chez les hommes (Ryan et Burk, 1974). Cela s'accompagne généralement d'une atrophie du tissu conjonctif contribuant à une fermeture incomplète de la glotte.

Des changements surviennent au niveau des cordes vocales avec le vieillissement. Chez les femmes, après la ménopause, un œdème peut apparaître dans la muqueuse couvrant les cordes vocales, provoquant des perturbations dans leurs vibrations pouvant conduire à une voix rauque (Hirano et coll., 1983). D'après Honjo et Isshiki (1980), les changements de la voix chez la femme semblent être en partie dus à une augmentation de la masse des cordes vocales du fait du changement hormonal. D'après Hirano et coll. (1989), des altérations des propriétés mécaniques des cordes vocales aussi bien chez les hommes que chez les femmes âgés apparaissent avec la diminution des fibres élastiques et l'augmentation des fibres de collagène.

En outre, la perte des dents peut altérer la forme du conduit vocal supraglottique, provoquant une réduction de la précision de l'articulation et une modification des résonances vocales (Kreiman et Sidtis (2011), p. 117).

Des facteurs neurologiques liés au vieillissement viennent diminuer la capacité de la personne âgée à contrôler précisément sa phonation. D'après Tiago et coll. (2007), chez les personnes de plus de 60 ans, on observe une diminution du nombre de fibres nerveuses dans le nerf laryngé récurrent et le nerf laryngé supérieur, et aussi une baisse du nombre de fibres myélinisées dans le nerf laryngé récurrent. Cette diminution du nombre de fibres du nerf laryngé supérieur diminue le contrôle de la F0 et du volume sonore du fait d'une perte du contrôle fin des cordes vocales (Thomas et coll., 2008).

3.1.1.2 Évolution des caractéristiques acoustiques et aérodynamiques de la voix âgée

Les changements physiologiques de la voix avec l'âge s'accompagnent d'effets sur le plan acoustique et aérodynamique. D'une manière générale, la voix des personnes âgées montre une augmentation des hésitations, des cassures de la voix (Benjamin, 1981), des tremblements, des imprécisions dans la prononciation des consonnes, ainsi qu'une articulation plus lente (Ryan et Burk, 1974).

Kreiman et Sidtis (2011) ont résumé les changements acoustiques de la voix avec le vieillissement dans la table 3.1.

D'après Hoit et Hixon (1987), les locuteurs masculins de plus de 70 ans utilisent un plus grand pourcentage de leur volume pulmonaire pour chaque syllabe prononcée, et pro-

Enfants	Adultes < 60 ans	Adultes > 60 ans
F0 haute, émergence des différences filles/garçons	La F0 diminue avec l'âge	La F0 diminue avec l'âge pour les femmes, mais augmente pour les hommes
Large gamme de hauteurs de tons	Gamme de hauteurs de tons assez constante	Gamme de hauteur de tons assez constante, mais la fréquence centrale peut baisser
Fréquences des formants hautes, différences filles/garçons commencent à apparaître à 4 ans	Fréquences des formants plus basses avec de larges différences hommes/femmes	Les fréquences des formants continuent à baisser
Contrôle de la phonation pauvre, voix rauque	Phonation stable	La phonation devient moins stable, voix rauque, souffle
Contrôle du volume de la voix pauvre	Bon contrôle du volume	Le volume peut augmenter ou diminuer
Le débit de parole est initialement faible, mais augmente avec l'âge	Débit de parole rapide	Le débit de parole diminue (dû à des respirations plus fréquentes)

TABLE 3.1: *Changements acoustiques de la voix avec l'âge (Kreiman et Sidtis, 2011).*

duisent moins de syllabes par respiration, par rapport aux locuteurs de moins de 25 ans. En effet, la fermeture incomplète de la glotte induit une augmentation de la quantité d'air utilisé durant la phonation (Linville et Rens, 2001).

De plus, l'augmentation de la variabilité entre locuteurs âgés dans les mesures acoustiques, aérodynamiques et physiologiques de la voix est une caractéristique de la voix âgée (Kreiman et Sidtis (2011), p. 118).

Une étude longitudinale de Decoster et Debruyne (2000) montre une diminution moyenne de la fréquence de vibration des plis vocaux de 13 Hz en 30 ans chez les locuteurs femmes, diminution également observée dans l'étude longitudinale menée par Russell et coll. (1995) avec 45 Hz de moins sur 50 ans de mesures. Chez les hommes, la F0 diminue également jusqu'à l'âge de 50 ans, après quoi elle augmente (Hollien et Shipp, 1972). Cette diminution de la F0 chez la femme est probablement due à l'augmentation de la masse des plis vocaux causée par un œdème, à la perte du tonus musculaire, à l'ossification du larynx, et à un changement hormonal. L'augmentation de la F0 chez l'homme est encore mal expliquée, mais pourrait être due à une augmentation de la raideur des cordes vocales ou à l'atrophie de celles-ci. Concernant la baisse avec l'âge de la fréquence centrale de la gamme des hauteurs de tons, cela pourrait s'expliquer par la baisse de l'efficacité et de la flexibilité phonatoire (Linville et Rens, 2001). Cependant, concernant les études longitudinales, les différences observées sur un individu au cours du temps peuvent ne pas seulement être dues à l'âge, mais peuvent être aussi liées à des changements de la santé de l'individu, un change-

ment d'accent régional dû à un déménagement, des variations de l'état émotionnel, etc. En effet, la fréquence fondamentale de la voix des personnes âgées, que ce soit chez les hommes ou chez les femmes, est très variable (Xue et Deliyski, 2001; Morgan et Rastatter, 1986; Morris et Brown Jr, 1994b). Ainsi, la stabilité de la fréquence fondamentale est moindre dans le cas de la voix des personnes âgées et cela va de pair avec une plus grande variabilité de l'amplitude crête-à-crête du signal de parole.

Linville et Rens (2001) ont mesuré l'instabilité de la voix et ont observé une augmentation des niveaux de jitter (mesure la variation en fréquence) et de shimmer (mesure la variation en amplitude) attribuée à la diminution du contrôle moteur, à l'altération de la fonction respiratoire, et à la dégénérescence des tissus du larynx, en particulier chez les hommes. Cependant, l'instabilité de la voix peut être également liée à la condition physique du locuteur. Ainsi, Ramig et Ringel (1983) ont mesuré la F0, le jitter, le shimmer, et la gamme de F0 chez les locuteurs en bonne ou en mauvaise condition physique. Les auteurs de cette étude ont observé que la stabilité de la voix était autant reliée à la condition physique qu'à l'âge. Les sujets en bonne condition physique montraient des niveaux de jitter et shimmer bas, et avec une gamme de fréquence plus large que les locuteurs en mauvaise condition physique. Chez les personnes âgées, les différences observées entre les individus étaient plus prononcées.

Du fait de la descente du larynx dans le cou au cours de la vie, les fréquences des formants diminuent avec l'allongement du conduit vocal supraglottique. Dans (Rastatter et coll., 1997), les auteurs montrent des différences significatives des fréquences des formants entre les hommes jeunes et les hommes âgés, mais pas chez les femmes. Au contraire, Linville et Rens (2001) observent une baisse des fréquences des formants plus marquée chez les femmes que chez les hommes.

Concernant les changements de volume de la voix, le volume augmente avec l'âge chez les hommes de plus de 60 ans (Ryan et Burk, 1974), et il n'augmente pas chez les femmes (Morris et Brown Jr, 1994a). Le volume maximum que peut émettre un locuteur diminue avec l'âge aussi bien chez l'homme que chez la femme (Ptacek et coll., 1966). Les problèmes de contrôle du volume sonore de la voix sembleraient être plus liés à des pertes de l'audition qu'à des changements du larynx ou de la fonction respiratoire.

En ce qui concerne la diminution du débit, les personnes âgées prononcent moins de syllabes par secondes que les personnes jeunes, et reprennent plus de fois leur respiration durant la prononciation d'une phrase (Shipp et coll., 1992). D'après Jacewicz et coll. (2009), le débit d'articulation est 11% plus rapide chez les adultes âgés de 20 à 34 ans que chez les adultes âgés de 51 à 61 ans. L'étude de Linville et Rens (2001) l'explique, chez les locuteurs masculins, par la dégénérescence neuromusculaire entraînant des perturbations dans l'articulation. Weismer et Liss (1991) expliquent la diminution du débit de la parole par les capacités moindres de l'appareil respiratoire – il en résulte le besoin de faire de plus fréquentes respirations entre les mots (Shipp et coll., 1992). Un déclin général sur le plan cognitif peut conduire à des difficultés à trouver les mots et à formuler les phrases (Weismer et Liss, 1991), ce qui peut entraîner une baisse du débit de parole.

Par ailleurs, certaines pathologies peuvent avoir une incidence sur l'émission de parole.

Des hésitations et des halètements dans les voix pathologiques sont associés à une augmentation du bruit dans le signal vocal dû à des vibrations apériodiques des cordes vocales (Selby et coll., 2003). Des mesures du rapport de l'énergie des harmoniques sur le bruit ont quantifié ce phénomène en comparant la voix de personnes âgées et de personnes jeunes (Xue et Deliyski, 2001; Ferrand, 2002).

Ces études montrent donc que les voix âgées présentent une plus grande instabilité que les voix des personnes jeunes. La question se pose maintenant de savoir comment les systèmes de RAP se comportent avec ce type de voix.

3.1.2 Reconnaissance automatique de la parole sur la voix âgée

La RAP adaptée à la voix des personnes âgées est un domaine encore peu exploré. De ce fait, les langues étudiées sont principalement l'anglais (Vipperla et coll., 2008; Wilpon et Jacobsen, 1996) et les langues asiatiques telles que le japonais (Baba et coll., 2004) ou le mandarin (Su et coll., 2014), mais rarement le français (Privat et coll., 2004). Un des facteurs qui affecte les performances des systèmes de reconnaissance automatique de la parole est la discordance entre les propriétés acoustiques de la parole de l'utilisateur et le modèle acoustique. Nous avons vu dans la section précédente que les personnes âgées présentent souvent une articulation moins précise. Comme la parole des adultes plutôt jeunes est souvent utilisée pour construire les modèles acoustiques, il en résulte un décalage entre les propriétés acoustiques de la parole des personnes âgées et les modèles acoustiques utilisés, conduisant à une dégradation des performances des systèmes de RAP avec la voix âgée (Wilpon et Jacobsen, 1996; Vipperla et coll., 2008; Privat et coll., 2004; Gerosa et coll., 2009; Anderson et coll., 1999).

Wilpon et Jacobsen (1996) utilisent le corpus *Danish Rafael.0 telephon speech* qui contient la parole de locuteurs âgés entre 8 et plus de 80 ans. 487 locuteurs ont été sélectionnés pour l'apprentissage, et 460 autres pour le test. Les locuteurs ont été répartis en 5 groupes en fonction de l'âge : enfants, adolescents, jeunes adultes, adultes et personnes âgées. Les auteurs ont comparé le *Word Error Rate* (WER, ou taux d'erreurs de mots) en testant différents modèles acoustiques sur les différents groupes de locuteurs. Les modèles acoustiques ont été réalisés en combinant plus ou moins de groupes ensemble. Les résultats montrent un WER augmenté de 50% (différence relative) pour les personnes âgées par rapport aux groupes plus jeunes, et, au delà de 70 ans, le WER augmente encore plus. Il semble donc pour les auteurs que l'âge de 70 ans représente un seuil pour les performances.

Dans (Privat et coll., 2004), l'étude des auteurs a porté sur la question de savoir s'il est possible d'utiliser les systèmes de reconnaissance de dictée vocale dans le cadre d'un système de dialogue homme-machine. Les auteurs ont utilisé un corpus en français composé de 30 locuteurs âgés de 20 à 30 ans et 15 locuteurs âgés de plus de 60 ans. Les enregistrements étaient composés de différents modes d'élocution : dictées ponctuées, non ponctuées, jouées, et dialogue avec une machine à travers une plate-forme magicien d'Oz. Dans cette étude, les auteurs ont regardé les effets de l'âge et du mode d'élocution sur le système

Dragon NaturallySpeaking. Ils ont trouvé une dégradation des performances corrélée au niveau de contrôle de l'élocution : plus le mode d'élocution est libre, plus les performances sont mauvaises. De plus, les auteurs ont observé une dégradation des performances de reconnaissance liée à l'âge, avec un pourcentage de taux de bonne reconnaissance de 78,6% pour les personnes jeunes et de 67,8% pour les personnes âgées dans le cas de la dictée ponctuée, ce mode d'élocution donnant de meilleures performances que les autres modes. Les auteurs ont également réalisé une comparaison au niveau phonémique et ont trouvé que les voyelles antérieures et arrondies seraient plus susceptibles d'être mal reconnues par le système de RAP.

Baba et coll. (2004) ont comparé des modèles acoustiques « âgés » appris sur 301 locuteurs âgés de 60 à 91 ans avec des modèles acoustiques « jeunes » appris sur 260 locuteurs. Les modèles étaient dépendants du genre : hommes, femmes, ou mixtes. Les enregistrements portaient sur des lectures de phrases en japonais. Les tests ont été faits avec différents systèmes de RAP dont un à base de triphones. Les tests sur 46 locuteurs âgés ont montré en moyenne une différences absolue entre 6,4% (modèles acoustiques genre-dépendants) et 7,3% (modèles acoustiques mixtes) sur le taux de bonne reconnaissance en faveur du modèle appris sur de la voix âgée par rapport au modèle appris sur de la voix jeune, avec une amélioration plus forte chez les hommes que chez les femmes. Les tests comparant 46 locuteurs âgés avec 46 locuteurs jeunes ont montré que les taux de bonne reconnaissance pour les femmes âgées avec un modèle acoustique « âgé » sont très proches des taux de bonne reconnaissance pour les femmes jeunes avec un modèle acoustique « jeune ». En revanche, pour les hommes, les taux de bonne reconnaissance avec un modèle « âgé » pour les hommes âgés sont inférieurs aux taux de bonne reconnaissance avec un modèle « jeune » pour les hommes jeunes. Enfin, les auteurs ont réalisé une adaptation au locuteur avec la méthode MLLR (*Maximum Likelihood Linear Regression*) et ont trouvé une amélioration de 1,8 à 8,1% (différence absolue) après adaptation, avec de meilleurs résultats si 50 phrases sont utilisées au lieu de seulement 10, et si le modèle de base est un modèle acoustique de voix âgées.

Vipperla et coll. (2008) ont réalisé une étude longitudinale des performances de la RAP sur de la parole prononcée par les mêmes locuteurs à quelques années de distance, donc à des âges différents. Ils ont en effet utilisé le corpus *SCOTUS*, constitué d'enregistrements de plaidoiries des juges et avocats de la Court Suprême des États-Unis depuis les années 80 jusqu'à 2008. Les auteurs ont entraîné un modèle acoustique à base de triphones sur 90 heures de parole des avocats, majoritairement des hommes (77 heures pour les hommes, 13 heures pour les femmes). A partir du corpus *SCOTUS*, ils ont tout d'abord comparé le WER entre un groupe de locuteurs adultes (100 hommes et 25 femmes) ayant prononcé 8655 phrases, et un groupe de locuteurs âgés (10 juges, dont 5 hommes et 2 femmes) ayant prononcé environ 200 phrases par locuteur pour chaque année à partir de 1999. Ils ont trouvé un WER de 36,4% pour les adultes et de 47,8% pour les personnes âgées, soit une différence absolue de 11,4%, avec un WER plus grand chez les femmes. Concernant l'étude longitudinale, les auteurs ont observé de manière générale une augmentation graduelle du WER en fonction de l'augmentation de l'âge pour chacun des 7 locuteurs testés. Les auteurs ont également effectué une

adaptation au locuteur avec la méthode MLLR et ont trouvé une amélioration de 6% chez les personnes âgées (différence absolue), avec une différence entre la parole des adultes et des personnes âgées passant de 11,4% à 7,7% après adaptation.

Il ressort de ces différentes études que la discordance entre les propriétés acoustiques des voix des personnes âgées et celles plus jeunes utilisées pour élaborer les modèles acoustiques a un impact important sur les performances des systèmes de RAP et une adaptation des modèles acoustiques est nécessaire. Plusieurs études ont cherché à savoir quelles modifications acoustiques de la voix âgée entraînent cette dégradation des performances des systèmes.

Ainsi, Vipperla et coll. (2010) analysent l'effet des changements de certains paramètres acoustiques de la voix tels que la fréquence fondamentale (F0), le jitter et le shimmer sur les performances de la RAP. Ils comparent également les scores de vraisemblances des phonèmes avec le PER (*Phonem Error Rate*). L'étude porte sur l'analyse du phonème /a/, 2970 réalisations de ce phonèmes ont été extraites du corpus *SCOTUS* prononcés par 23 hommes adultes ainsi que 2105 réalisations prononcées par 10 hommes âgés. Il a été observé une F0 plus basse chez les hommes âgés (128 Hz) que chez les hommes adultes (144 Hz). En baissant artificiellement la valeur de F0 de 10%, pour voir les effets d'une baisse de la F0 sur le système de RAP, le WER augmente légèrement en passant de 32,1% à 33,2%. Les auteurs observent également une augmentation significative des valeurs du jitter et du shimmer chez les personnes âgées par rapport aux adultes, en revanche en augmentant artificiellement ces valeurs, ils ne trouvent pas de variations significatives du WER.

Enfin, Pellegrini et coll. (2013) présentent une étude sur l'impact de certains paramètres acoustiques sur la performance de la reconnaissance vocale sur des corpus de parole en portugais européen. Trois groupes de locuteurs sont comparés : le groupe « jeunes adultes » (100 locuteurs de 19 à 30 ans), soit 20h extraites du corpus *BD-PUBLICO*, et les groupes « seniors » (114 locuteurs de 60 à 75 ans) et « âgés » (94 locuteurs de 75 à 90 ans), avec respectivement 8,7 heures et 6,7 heures extraites du corpus EASR. Les auteurs ont utilisé le décodeur *Audimus*, un système hybride HMM (*Hidden Markov Models*) et perceptron multicouches. Les auteurs ont obtenu comme résultats des WER moyens de 13,55%, 27,25% et 34,15% pour respectivement les groupes « jeunes adultes », « seniors » et « âgés ». Ils ont calculé les corrélations entre les paramètres acoustiques et les WER, et trouvé que les WER sont corrélés avec l'utilisation des pauses, le débit, la durée des phonèmes et le pourcentage de phonèmes longs. Pour les hommes, ils ont observé une légère corrélation du WER avec le pourcentage de schwas et le rapport harmonique sur bruit. Pour le jitter et le shimmer, la corrélation est très basse pour les 2 genres.

3.2 Voix émue

Dans le contexte d'une reconnaissance automatique de la parole en vue d'un appel à l'aide ou d'un appel à un proche corrélé à une charge émotionnelle, la personne sera la plupart du temps sous l'emprise d'une émotion, et non dans le cas d'une lecture neutre de texte

(c'est à dire la situation de recueil de corpus que nous avons assimilé comme le plus analogue à la situation de commandes domotiques sans contexte émotionnel particulier).

En situation réelle, pour un système domotique détectant des commandes vocales, celles-ci peuvent être expressives globalement pour 2 raisons :

— si l'utilisateur se construit une représentation de l'appartement comme étant une entité communicative (paradigme de HAL (Clarcke, 1968)), faisant que la prosodie et la morpho-syntaxe peuvent mettre en œuvre des intentions, des attitudes et autres valeurs socio-affectifs interactionnelles,

— si la commande est motivée par une contexte fortement émotionnel (panique, joie de parler à un proche, etc.).

Ainsi, nous pouvons supposer que plus l'énoncé à reconnaître reflète un désir ou un besoin important pour le sujet, plus l'énoncé est modifié par l'expressivité. Or la caractéristique de la voie émue est précisément de perturber les structures acoustiques du continuum sonore, avec des modifications prosodiques incluant de fortes modifications de la qualité de la voix (Audibert, 2008; Vlasenko et coll., 2011) de façon à ce que ces émotions soient perçues par l'interlocuteur. En effet, d'après Scherer (2003), la prosodie (étendue à la qualité de voix) est le principal vecteur des émotions exprimées dans la parole, et plus généralement de tous les socio-affects (Campbell, 2009; Aubergé, 2002).

Dans leur utilisation classique avec le microphone proche du locuteur, les systèmes de RAP modernes ont des performances élevées pour la reconnaissance de parole prononcée de façon neutre, mais ne peuvent pas maintenir leurs performances pour la parole spontanée, la RAP pour la parole spontanée n'étant pas un problème résolu, même s'il a été largement étudié (Shinozaki et coll., 2001; Nanjo et Kawahara, 2002; Kawahara et coll., 2003; Furui, 2003; Furui et coll., 2005; Dufour et coll., 2009). Les principaux problèmes abordés sont d'ordre pragmatico-syntaxique (reformulation, fillers, etc.), donc ces études s'intéressent surtout à l'amélioration des modèles de langage et beaucoup moins souvent à l'amélioration des modèles acoustiques : les perturbations prosodiques du signal de la parole, dont une part est liée à l'état émotionnel de la personne, ne sont pas réellement abordées comme étant un problème pour les performances des systèmes de RAP. Or il est sûrement plus facile pour un utilisateur de contrôler une stricte formulation lexico-syntaxique de la commande que de contrôler l'expressivité prosodique, car les émotions sont souvent exprimées de façon involontaire.

Par ailleurs, beaucoup d'études sont liées à la reconnaissance automatique des émotions à partir de la voix dans des situations, en particulier, de détresse (Vidrascu, 2007; Schuller et coll., 2011; Vaudable, 2012; Chastagnol, 2013). De nombreux descripteurs extraits du signal audio peuvent être utilisés pour la classification des émotions. Dans (Soury et Devillers, 2013), les auteurs utilisent des descripteurs tels que l'énergie du signal, le pitch, la qualité de la voix (le jitter et le shimmer, qui ont été étudiés plutôt pour les voix pathologiques), et des paramètres spectraux et cepstraux pour la détection de la voix stressée, en comparant différentes tailles de fenêtres d'analyse.

Cependant, peu d'études portent sur l'évaluation de l'impact des émotions sur les performances des systèmes de RAP ; nous pouvons citer l'étude de Vlasenko et coll. (2012) qui porte sur l'adaptation des modèles acoustiques à la voix émue. Or le problème de l'utilisation du système de RAP dans le contexte d'ordres domotiques et d'appels à l'aide est qu'il doit être robuste tant pour les commandes vocales prononcées de façon neutre que celles prononcées de façon expressive.

3.3 Conclusion

Nous avons vu que la voix âgée a des caractéristiques spécifiques pouvant dégrader les performances des systèmes de RAP, et que peu d'études portent sur l'impact des émotions sur les performances de ces systèmes.

Dans le chapitre suivant, nous allons décrire la méthodologie que nous avons employée pour développer notre système de détection d'appels de détresse adapté aux personnes âgées.

CHAPITRE 4

Méthodologie

Notre but est de développer un système permettant la détection automatique d'appels de détresse et d'appel vers les aidants et les proches grâce à un système de RAP, ce service étant destiné aux personnes âgées vivant à domicile.

Dans cette section, nous décrirons quelles devront être les caractéristiques et les contraintes fondamentales de notre système, et chercherons dans la littérature quelles sont les approches existantes permettant de développer notre système en prenant en compte ces contraintes.

4.1 Un système ubiquitaire

Le système doit fonctionner en tâche de fond et de façon transparente pour être en permanence à l'écoute de l'environnement sonore du lieu de vie afin de reconnaître et filtrer toute survenue de parole, notamment émise lors d'une situation de détresse.

Pour cela, le système devra utiliser une détection d'activité sonore permettant de sélectionner automatiquement en continu les segments sonores contenant du signal sans intervention de l'utilisateur, c'est-à-dire sans qu'il ne soit nécessaire d'appuyer sur un bouton « marche/arrêt » pour lancer la reconnaissance vocale. Dans (Principi et coll., 2013), les auteurs ont développé un système de détection d'appels de détresse et d'ordres domotiques avec appels automatiques pour une assistance. Ils utilisent une détection d'activité grâce à un algorithme basé sur l'énergie du signal. Le système *AuditHIS* (Vacher et coll., 2009a) du projet *DESDHIS2* (reconnaissance des sons et de la parole pour une reconnaissance d'activité) utilise une détection qui se fait également par analyse de l'énergie du signal (décomposition en utilisant un arbre d'ondelettes de profondeur 3) avec seuil adaptatif. La même technique est utilisée par le système *PATSH* du projet *Sweet-Home* (Vacher et coll., 2013b).

Ces techniques détectent tout segment de parole. Or, il ne s'agit pas ici d'enregistrer toute conversation détectée à domicile. Tout ce qui n'est pas appel de détresse ou appel des aidants ne doit pas être reconnu pour préserver l'intimité de la personne âgée. Principi et coll. (2013) utilisent dans leur modèle de langage une grammaire réduite aux ordres domotiques et appels de détresse. De la même façon, dans le projet *Sweet-Home*, les auteurs utilisent un modèle de langage correspondant à la grammaire des ordres domotiques et des appels de détresse, et interpolent ce modèle avec un modèle plus large mais avec un poids réduit pour prendre en compte les petites variations. De plus, un premier filtrage est effectué pour ne

conserver que les commandes et non toutes les phrases prononcées : tous les événements sonores de durée inférieure à 150 ms ou supérieure à 2,2 s sont ignorés, de même que les sons de rapport signal sur bruit trop bas.

Ce système étant installé dans un environnement non contrôlé, le domicile de la personne âgée, il devra être robuste en toutes circonstances, même en condition de parole distante ou avec des bruits de fond ou échos. Dans le projet *Sweet-Home* (Vacher et coll., 2012b), les auteurs utilisent des microphones distants pour éviter à la personne la contrainte du port d'un microphone. Afin de minimiser l'influence des effets de l'écho relatifs à la parole distante, ils utilisent les microphones les plus proches du locuteur en choisissant les canaux avec les meilleurs rapport signal sur bruit et utilisent simultanément ces canaux pour améliorer la reconnaissance à la méthode du décodage guidé. Ce système intègre aussi une annulation de source de bruit connu avec la technique AEC (*Acoustic Echo Cancellation*) au niveau de l'acquisition du signal sonore. Dans (Principi et coll., 2013), les auteurs réalisent une annulation d'interférences en amont de la reconnaissance vocale ainsi qu'une AEC lors de l'appel téléphonique d'urgence.

Du fait de l'âge des personnes visées et de leurs éventuelles incapacités physiques, et de leur possible non familiarité avec les nouvelles technologies, le système doit être conçu de manière à fonctionner de façon totalement autonome au quotidien, sans que soit nécessaire une quelconque intervention de la part de la personne âgée ou d'un opérateur autre que lors de la mise en route, le réglage et la configuration initiale. Bien entendu, le traitement des données doit fonctionner en pseudo temps réel de telle manière à déclencher une alarme dans un temps le plus court possible après l'émission d'un appel de détresse. Par exemple, le système PATSH (Vacher et coll., 2013b) fonctionne en pseudo temps réel avec un temps de décodage de 1,47 fois la durée de la phrase, augmenté du temps nécessaire pour le traitement des événements simultanés sur les différents microphones (environ 1 s).

4.2 Un système adapté à l'application

Le système doit être adapté aux voix âgées, nécessitant que les modèles acoustiques soient adaptés à ce type de voix. Dans (Principi et coll., 2013), les auteurs étudient la différence de comportement d'un système de RAP pour la reconnaissance de commandes domotiques et d'appels de détresse prononcés de façon neutre ou criée. Le modèle acoustique est appris sur un corpus de parole lue, le corpus *APASCI*, qui est composé de 1310 phrases en italien prononcées par 30 hommes et 30 femmes. La robustesse à la différence entre les conditions d'apprentissage et les conditions de test est améliorée par une adaptation de ce modèle acoustique par la technique MLLR. Avant l'utilisation du système pour la première fois, il était demandé aux locuteurs de prononcer des phrases phonétiquement riches couvrant l'intégralité des phonèmes de la langue italienne. Une adaptation était faite pour chaque locuteur, chaque style de prononciation (parole normale ou criée), et chaque condition d'enregistrement (microphone proche ou distant). Dans le système *PATSH* (Vacher et coll., 2013b), le modèle acoustique est entraîné sur 80 heures de parole lue annotée, et

une adaptation MLLR est effectuée sur chaque locuteur à partir d'un texte lu dans les conditions d'enregistrement de l'appartement intelligent (microphones distants). Cependant, un critère important étant la facilité d'utilisation, nous souhaitons que le système soit livré directement adapté à la voix âgée en utilisant des données d'adaptation enregistrées auprès de locuteurs différents des utilisateurs du système. En effet, la lecture du texte et le lancement de l'adaptation s'avèrent contraignants en situation réelle.

Par ailleurs, l'application visée impose que le système soit tolérant à des commandes vocales émises en situation de stress ou à des appels de détresse de type « A l'aide ! » ou « Au secours ! ». Drosos et coll. (2012) utilisent une reconnaissance de mots-clé basée sur les algorithmes DTW (*Dynamic Time Warping*) et kNN (*k-nearest neighbors*) pour détecter les mots indiquant une situation d'urgence, associée à un classifieur de voix émue pour déterminer si la voix est affectée par la peur ou la colère. En revanche, le système de Vacher et coll. (2013b) détermine si la phrase est une phrase de détresse en se basant uniquement sur la grammaire de la phrase reconnue sans détection de présence d'émotions dans la voix.

4.3 Les outils

Dans un premier temps, nous avons donc procédé à la sélection des outils de RAP possédant des caractéristiques adaptées à la construction d'un tel système, ou du moins satisfaisant du mieux possible l'ensemble de ces besoins. Dans une première étape, nous avons sélectionné et regroupé en catégories les principaux outils de RAP existants :

— Applications : *Dragon NaturallySpeaking*[1], *S Voice*[2], *Siri*[3]

— Boîtes à outil pour la recherche : Sphinx[4] (*Sphinx3*, *Sphinx4*, *PocketSphinx*), *Speeral*[5], *RASR*[6], *KALDI*[7]

— API : *Google Web Speech API*[8], *Microsoft Speech Recognition API*[9]

Les systèmes définis comme des boîtes à outils pour la recherche peuvent être adaptés à une tâche spécifique par l'apprentissage et l'adaptation de leurs modèles acoustiques et de leurs modèles de langage, en étant capables de reconnaître des mots-clé spécifiques, pour un type de voix spécifique (par exemple la voix âgée). Ils représentent donc un grand intérêt pour nous. L'utilisation de petits modèles de langages permet par ailleurs d'augmenter la rapidité du décodage. Les modèles acoustiques peuvent être adaptés au locuteur ou à une famille de locuteurs. Ces systèmes sont bien documentés, et disposent d'une communauté d'entre-aide très active. Pour certains d'entre eux, il existe une expertise technique dans notre laboratoire.

1. http ://www.nuance.com/dragon/index.htm
2. http ://www.samsung.com/global/galaxys3/svoice.html
3. http ://www.apple.com/fr/ios/siri/
4. http ://cmusphinx.sourceforge.net
5. http ://speeral.univ-avignon.fr
6. http ://www-i6.informatik.rwth-aachen.de/rwth-asr
7. http ://kaldi.sourceforge.net
8. https ://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html
9. http ://msdn.microsoft.com/en-us/library/jj127860.aspx

Selon ces critères, nous avons choisi d'utiliser *Sphinx3* pour construire un système de RAP adapté à la voix âgée qui respecte assez bien les contraintes mentionnées précédemment. Cependant, le système *Google Web Speech API*, ayant pour avantage d'être acoustiquement entraîné sur une quantité très importante de données, a également été utilisé en tant que système état de l'art pour une comparaison avec *Sphinx3*. Nous avons utilisé, pour sa simplicité, la version beta de cette API (utilisée dans le navigateur *Google Chrome 11*), en sachant que ce système (boîte noire) est peu paramétrable et ne permet pas d'entraîner nos propres modèles.

Nous avons développé un système d'analyse sonore temps réel, *CirdoX*, qui intègre un système de RAP. *CirdoX* permet l'acquisition du signal audio, la détection de l'occurrence d'événements sonores (parole ou son), la discrimination de la parole et du son avec un appel du module de RAP s'il s'agit de parole ou un appel d'un classifieur de sons s'il s'agit d'un son, la mise en forme des événements sonores, et le filtrage des mots-clés correspondant à une détresse. Une grande partie de *CirdoX* intègre le savoir-faire acquis lors du développement des logiciels *AuditHIS* et *PATSH*. Une fonctionnalité supplémentaire a été ajoutée pour la reconnaissance de mots-clés par mesure de distance avec les mots-clés de référence. Ce système a été construit de manière modulaire, ce qui permet d'intégrer facilement différents systèmes de RAP (*Sphinx, KALDI, API Google*).

Une première version de ce système a été mise en œuvre en condition réalistes dans un appartement de test. Le recueil des données produites nous a permis de construire un corpus d'évaluation.

4.4 Les données

Bien entendu, pour qu'une adaptation soit possible aux locuteurs âgés, nous devons disposer de corpus de parole spécifiques à la tâche et à ce type de locuteurs. Nous avions à disposition au laboratoire un corpus contenant des appels de détresse lus par des personnes âgées, le corpus *Voice-Age*, enregistré par seulement 7 locuteurs et de taille insuffisante (4 heures) pour être utilisé pour les apprentissages et les tests ; cette taille réduite s'explique par les difficultés à enregistrer un corpus de voix de personnes âgées du fait de la fragilité et de la fatigabilité de ces personnes qui peuvent difficilement se déplacer et venir dans un studio d'enregistrement. Nous avions également à disposition d'autres corpus de voix de personnes âgées, le corpus *Mémorial de la Shoah* et le corpus *Phonologie du français contemporain*. Cependant, le contenu de ces corpus (témoignages sur la vie des personnes interrogées) était trop éloigné de l'application visée. Un autre corpus enregistré par le laboratoire, le corpus *Anodin-Détresse*, contenant des appels de détresse lus par 20 locuteurs, s'approchait de l'application visée, mais les participants n'étaient pas des personnes âgées. Pour le français, nous utilisons habituellement au laboratoire le corpus *BREF120* pour l'apprentissage des modèles acoustiques du fait de la longue durée de ce corpus (100 heures). Cependant, les phrases de ce corpus sont elles aussi hors domaine d'application et les personnes ayant participé à son enregistrement n'étaient pas des personnes âgées. N'ayant pas les ressources suffisantes

pour enregistrer une quantité très importante de voix de personnes âgées pour créer des modèles de voix âgées (une centaine d'heures aurait été nécessaire), nous avons utilisé comme modèle acoustique de base (modèle générique) un modèle appris sur le corpus *BREF120*, et nous avons enregistré un corpus contenant plusieurs heures de voix de parole âgée pour adapter le modèle générique à la voix des personnes âgées. En effet, certaines techniques permettent d'adapter les modèles acoustiques à des caractéristiques particulières à partir d'une quantité réduite de données contenant ces caractéristiques. Pour réaliser les modèles de langage, la question s'est posée de déterminer quelles sont les phrases prononcées pour appeler à l'aide lorsque la personne est en situation de détresse, pour cela nous avons bénéficié de l'aide du laboratoire GRePS, partenaire du projet *CIRDO* (Bobillier-Chaumon et coll., 2012). Les enregistrements de parole âgée ont nécessité de développer le logiciel d'enregistrement de parole lue *GEOD* spécifiquement adapté à cette frange de la population. Enfin, pour adapter les modèles acoustiques à la voix émue, nous avons enregistré un corpus de voix de détresse auprès de personnes jouant des situations définies dans des scènes représentées dans des images.

4.5 Conclusion

Dans ce chapitre, nous avons décrit les caractéristiques et les contraintes – ubiquité, adaptation à la voix âgée et émue – de notre système de détection d'appels de détresse, ainsi que les outils et les données que nous utiliserons.

Dans le chapitre suivant, nous décrirons les corpus de voix âgée et de voix émue que nous avons enregistrés en vue d'adapter et d'évaluer notre système.

CHAPITRE 5

Corpus

5.1 Outils d'enregistrement de parole lue

Pour enregistrer la voix des personnes âgées, il nous était nécessaire de disposer d'un outil d'enregistrement. Malgré quelques logiciels existants, nous avons fait le choix de développer notre propre application d'enregistrement de corpus, tel que décrit dans les sections suivantes.

5.1.1 Les logiciels existants

5.1.1.1 EMACOP

Le logiciel *EMACOP* (Environnement Multimedia pour l'Acquisition et la gestion de Corpus de Parole) a été développé en 1999 au sein de l'équipe GEOD du CLIPS pour l'enregistrement de corpus de parole lue (Vaufreydaz et coll., 1998). Il a notamment été utilisé pour l'enregistrement des corpus *BRAF100* (Vaufreydaz et coll., 2000) et *Anodin-Détresse* (Vacher et coll., 2006).

Ce logiciel a été conçu pour que l'enregistrement puisse être réalisé de manière autonome par la personne enregistrée grâce à un système de type client-serveur. L'ensemble du logiciel se présente donc sous la forme de deux applications distinctes qui communiquent via le réseau grâce au protocole TCP/IP. L'application serveur permet la définition de la base de données et est exécutée sur une machine maître. L'entrée des données se fait à l'aide d'un utilitaire qui connaît plusieurs types d'entités : les corpus, leurs éléments, les locuteurs, et les scénarios d'acquisition. L'application cliente a en charge toutes les fonctionnalités de présentation des items et d'acquisition du signal. Le locuteur peut choisir un scénario, le mener à son terme ou l'arrêter en cours de route, le reprendre ou l'abandonner. Le locuteur peut accepter le signal qu'il vient de prononcer et passer au suivant, écouter et visualiser le signal qu'il a produit. Le texte à prononcer lui est présenté tout au long du déroulement du scénario dans une police de taille réglable, et des images peuvent être également présentées. Un seuil de détection de la voix permet de déclencher les enregistrements et de passer aux phrases suivantes automatiquement.

5.1.1.2 ROCme!

ROCme! (Recording of Oral Corpora Made Easy) est un logiciel pour la gestion de l'enregistrement de corpus oraux (Ferragne et coll., 2013). Il a été développé par le laboratoire Dynamique Du Langage du CNRS - Université Lumière Lyon 2.

Le logiciel permet d'enregistrer la voix des locuteurs à partir de stimuli qui s'affichent à l'écran. Le locuteur fait défiler texte, images, vidéos ou sons à l'écran et enregistre sa voix de façon autonome, les locuteurs pouvant gérer par eux-même l'enregistrement audio, la lecture, la sauvegarde et le défilement des phrases. Le logiciel permet également de recueillir des métadonnées par le biais de questionnaires et de gérer l'acquisition des corpus à travers des projets d'enregistrement. L'interface pour le recueil de métadonnées sur les locuteurs est totalement personnalisable via des balises XML, permettant de recueillir des données telles que l'âge et certaines caractéristiques personnelles (gaucher ou droitier, accent, etc). Aussi, le logiciel permet une personnalisation de l'affichage des phrases du corpus avec balises HTML et style CSS. A la création d'un projet, le logiciel présente plusieurs options telles que la présentation des stimuli en ordre aléatoire, l'apparition d'un masque entre le déclenchement de l'enregistrement et l'apparition d'un stimulus, l'interdiction de la sauvegarde d'un signal écrêté, la possibilité d'enregistrer plusieurs fois chaque phrases, etc.

5.1.1.3 Limitation des logiciels existants

Notre protocole d'enregistrement de corpus prévoyait d'enregistrer certains locuteurs, notamment les personnes âgées, à leur domicile. Il est vraisemblable que pour un nombre important de locuteurs, aucune connexion de type internet n'aurait été disponible. La solution *EMACOP* avec client local et serveur distant n'était donc pas envisageable. De plus, nous avions connaissance de problèmes rencontrés lors de l'enregistrement du corpus *Voice-Age* enregistré au CHU de Grenoble où les expérimentateurs, afin de pouvoir utiliser *EMACOP* sans réseau à l'intérieur de l'hôpital, ont hébergé le client et le serveur sur la même machine (un PC portable) : le serveur étant prévu pour tourner en continu sur une machine autonome et ne jamais être arrêté, il y a eu systématiquement des pertes de données à chaque interruption pendant un scénario. De plus, le serveur devait être renseigné après chaque arrêt de l'ordinateur. Nous avons donc choisi de développer une nouvelle application d'enregistrement de corpus, *GEOD*, destinée à remplacer *EMACOP*, et fonctionnant de manière autonome.

Parallèlement, le laboratoire Dynamique Du Langage de Lyon 2 a développé l'application *ROCme!*, dont les fonctionnalités étaient proches de celles de *GEOD*. Cependant, au début de la thèse, nous n'avions pas connaissance du développement de *ROCme!*, jusqu'à la conférence *JEP-TALN 2012* où ce logiciel nous a été présenté.

5.1.2 Le logiciel GEOD

5.1.2.1 Cahier des charges

GEOD est un logiciel d'enregistrement de corpus de parole lue (figure 5.1). Ce logiciel permet l'acquisition du signal sonore, avec déclenchement automatique de l'enregistrement à la détection du signal de parole, et la génération de fichiers dans le format wav correspondant à l'enregistrement de chaque énoncé prononcé. Le logiciel peut être actionné par un opérateur en cas d'enregistrement de personnes âgées. La taille des caractères est réglable pour s'adapter à des locuteurs ayant une mauvaise vision.

Les fonctionnalités de *GEOD* sont :

— La gestion des corpus, avec la création d'un nouveau corpus ou le chargement d'un corpus existant.

— La gestion des locuteurs, avec la création d'un nouveau locuteur ou le chargement d'un locuteur existant.

— L'acquisition de parole avec détection automatique du début de l'énoncé, avec l'affichage des énoncés à lire par le locuteur sous forme de scénarios, l'enregistrement de signaux sonores, la détection de la voix du locuteur pour détecter le début et la fin d'un énoncé de parole, la possibilité d'être en mode de défilement des phrases automatique ou manuel, et la création de fichiers wav et de fichiers de transcription au format Sphinx pour chaque énoncé enregistré.

— La vérification, qui consiste en une vérification visuelle des signaux enregistrés par affichage du signal en fonction du temps, un suivi de l'avancement des enregistrements, et la lecture des signaux sonores enregistrés.

— Les réglages, qui consistent en un réglage des paramètres d'acquisition et de détection, et en un test du microphone et de la détection des instants de début et de fin des événements sonores.

5.1.2.2 Protocole d'enregistrement avec le logiciel GEOD

Un enregistrement commence toujours par la signature d'une fiche de consentement de participation à l'étude (voir annexe A).

Après le lancement du logiciel, l'expérimentateur commence par choisir l'emplacement du répertoire qui contiendra le corpus. Il saisit les informations concernant le locuteur (identifiant, âge, genre, accent, commentaires), puis procède au réglage du niveau sonore d'entrée du microphone, et vérifie les réglages des seuils de détection.

Puis vient la phase d'enregistrement. Les énoncés à prononcer sont répartis dans des scénarios, c'est-à-dire des listes d'énoncés qui peuvent être enregistrés les uns à la suite des autres. Il est demandé au locuteur de lire la liste des énoncés du scénario en cours. Les énoncés s'affichent à tour de rôle : une fois qu'un énoncé a été prononcé par le locuteur, le suivant

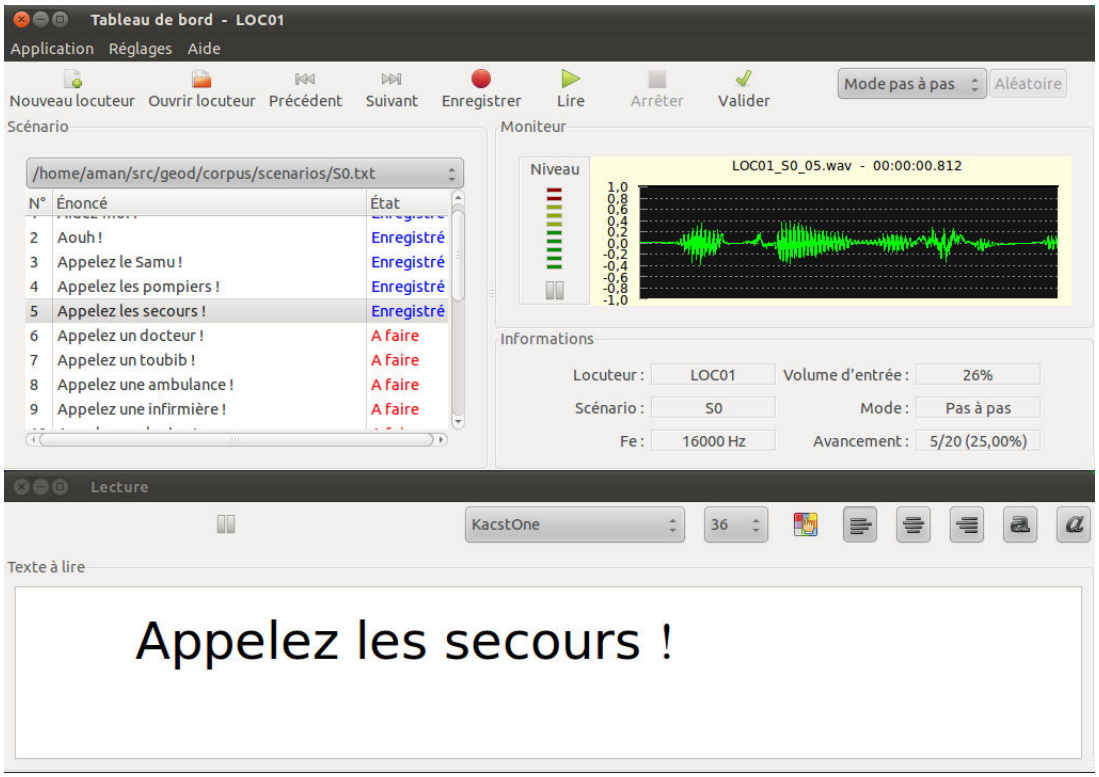


FIGURE 5.1: *Le logiciel GEOD.*

s'affiche grâce à la détection automatique de signal, jusqu'à ce que tous les énoncés du scénario aient été prononcés. Puis, après une pause de durée suffisante, on continue avec les scénarios suivants, jusqu'à ce que tous les scénarios aient été lus.

L'expérimentateur peut contrôler pendant l'acquisition que le microphone ne sature pas grâce à l'affichage du niveau sonore sous forme de barre, il peut aussi visualiser en temps réel la forme du signal. Une fois les énoncés prononcés par le locuteur, l'opérateur peut écouter le signal sonore enregistré d'un énoncé donné pour le vérifier, et éventuellement le ré-enregistrer par exemple si le signal est saturé ou que le locuteur a mal prononcé la phrase.

Le résultat de l'enregistrement du locuteur est une collection de fichiers wave (mono, 16 bits, PCM) accompagnés de leurs fichiers de transcriptions contenus dans le dossier affecté au locuteur donné, et l'ensemble des dossiers des locuteurs constitue le corpus.

5.2 Corpus AD80

5.2.1 Les corpus précurseurs au sein de l'équipe GETALP

5.2.1.1 Le corpus Anodin-Détresse

Le corpus *AD* (Anodin-Détresse) a été enregistré par l'équipe GEOD du CLIPS en 2004 lors du projet *DESDHIS* (Vacher et coll., 2006). Il est constitué de phrases courtes de type appels de détresse et de type anodines (conversation courante). Un exemple de phrases retenues est « Aidez-moi » pour un appel de détresse et « Il fait beau » pour une conversation anodine.

Ce corpus de parole a été enregistré dans le but d'étudier un système d'analyse sonore permettant notamment la détection de situations de détresse, utilisant la discrimination entre son et parole (Vacher et coll., 2006) et la détection de mots-clé (Vacher et coll., 2008).

21 locuteurs (11 femmes et 10 hommes) âgés entre 22 et 64 ans ont été enregistrés, la plupart (16 locuteurs sur 21) ayant un âge inférieur à 30 ans. Les phrases prononcées, des énoncés courts (entre 1 et 3 secondes), sont les mêmes pour tous les locuteurs, soit 60 phrases de détresse et 66 phrases anodines. Les enregistrements ont été effectués en studio au sein de l'équipe GETALP à l'aide du logiciel *EMACOP* (Vaufreydaz, 1998).

Au final, 2646 phrases ont été prononcées, ce qui représente 55 minutes de parole. L'ensemble des énoncés est présenté en annexe B.

5.2.1.2 Le corpus Voice-Age

Le corpus *AD* est vite apparu trop limité car restreint à une frange jeune de la population, alors que le domaine d'application visé concerne plus particulièrement des personnes âgées vivant seules à domicile. C'est pourquoi il a paru nécessaire de disposer de paroles adaptées prononcées par des personnes âgées.

Le corpus *VA* (Voice-Age) est constitué de l'assemblage de 2 corpus : le corpus *CHU* et le corpus *Normand*.

Le corpus *CHU* a été enregistré auprès de personnes âgées par l'équipe GETALP (Aynaud, 2009) en 2009 dans le service de gérontologie de l'Hôpital Sud du CHU de Grenoble. Il a consisté à faire lire à ces personnes des phrases courtes anodines et de détresse, similaires ou identiques à celles du corpus *AD*, et des phrases inspirées de phrases de journaux et de magazines du corpus *BRAF100* (Vaufreydaz et coll., 2000), de durée supérieure (jusqu'à 10 secondes) aux phrases *AD*. Les enregistrements ont été effectués à l'hôpital, à l'aide d'un microphone relié à un PC portable par un pré-ampli, l'acquisition se faisant grâce au logiciel *EMACOP*.

Les personnes capables de réaliser ces enregistrements étaient peu nombreuses, déjà fortement handicapées, et avait une courte période de disponibilité, ce qui explique le temps important nécessaire pour ces enregistrements (4 demi-journées par locuteur, avec seulement une demi-journée par semaine). Deux locuteurs (une femme et un homme) de 82 et 89 ans ont été enregistrés, pour un total de 2089 phrases lues (1 heure 48 minutes d'enregistrement).

En 2010, l'équipe GETALP a poursuivi ses travaux d'enregistrement de voix de personnes âgées en réalisant l'acquisition d'un corpus appelé corpus *Normand* (Lefol, 2010). Il a été demandé aux personnes de lire les mêmes phrases que lors de l'enregistrement du corpus *CHU*. Les enregistrements ont été effectués à domicile auprès de personnes âgées faisant partie d'une même famille et habitant en Normandie. Le même type de matériel a été utilisé, excepté le logiciel d'acquisition *EMACOP* qui a été remplacé par le logiciel *Audacity* suite aux problèmes évoqués en section 5.1.1.3.

Cinq personnes (3 femmes et 2 hommes) ont été enregistrées, âgées entre 70 et 79 ans, pour un total de 3352 phrases et 2 heures 21 minutes d'enregistrement.

Du fait de la fatigabilité des personnes âgées et de la longueur de certaines phrases, toutes les phrases n'ont pu être prononcées par l'ensemble des locuteurs : entre 407 et 1187 phrases ont été prononcées par chaque locuteur, et les enregistrements, pour chaque locuteur, se sont étalés sur plusieurs jours.

La liste complète des phrases se trouve en annexe D.

5.2.2 Nos enregistrements

5.2.2.1 Première étape

Toujours dans le cadre de nos travaux sur la détection de situations de détresse pour les personnes âgées isolées à domicile, du fait du faible nombre de locuteurs âgés (7 locuteurs) ayant été enregistrés lors des précédentes études évoquées ci-dessus, nous avons complété le corpus de voix de personnes âgées avec des énoncés adaptés à la tâche.

Pour faciliter l'acquisition de notre corpus, sachant que les personnes âgées peuvent rapidement se fatiguer, nous avions comme objectif que les sessions d'enregistrement soient courtes (moins de 20 minutes). Nous avons choisi de leur faire lire des phrases en nombre restreint, et courtes, cela permettant d'avoir le même nombre de phrases lues par chaque locuteur. Nous avons donc exclu les phrases de magazines et de journaux utilisées dans le corpus *VA*, du fait de leur longueur trop importante. Aussi, nous voulions que notre nouveau corpus de voix âgées puisse être comparé avec le corpus *AD* de voix jeunes dans notre évaluation des performances des systèmes de RAP. Nous avons donc repris les 126 énoncés de détresse et anodins utilisés dans le corpus *AD* enregistré en 2004 pour les faire lire aux personnes âgées.

Ainsi, en 2012, nous avons réalisé une session d'enregistrements dans un centre de réhabilitation pour personnes âgées (SSR Les Cadières) et une EHPAD (Château de Labahou) du sud-est de la France [1]. Les personnes visées étaient des personnes âgées de plus de 65 ans, capable de lire et sans troubles cognitifs connus au moment de l'enregistrement. Les enregistrements ont été effectués à partir d'un microphone positionné à environ 30 cm de la bouche du locuteur et relié à un PC portable par un pré-ampli, l'acquisition étant réalisée grâce au logiciel *GEOD* (voir section 5.1.2).

29 locuteurs (21 femmes et 8 hommes) âgés de 62 à 94 ans ont ainsi été enregistrés en 2012, et ont lu au total 3611 phrases de détresse et anodines, pour une durée de 1 heure et 10 minutes. La liste des phrases prononcées est celle de l'annexe B.

5.2.2.2 Deuxième étape

Les phrases de détresse envisagées précédemment ont été choisies *a priori* et peuvent ne pas correspondre à ce qu'une personne en difficulté prononce de manière spontanée.

C'est pourquoi, au début du projet CIRDO, une étude a été menée par le laboratoire GRePS (Bobillier-Chaumon et coll., 2012) pour analyser les différents cas de chutes afin

1. Ces établissements appartiennent à la Fondation Diaconesses de Reuilly http ://www.oidr.org

d'identifier les poses-clés (comportement, gestes et poses) et les phrases caractéristiques des situations à risque, qui pourront ensuite être utilisées pour construire le système. En effet, à partir de méthodes d'entretien et d'observation, le laboratoire GRePS a construit des scénarios de différentes chutes. Dans chaque scénario sont décrites les conditions de la chute (caractéristiques de la personne, activité réalisée, lieu, moment, causes et circonstances de l'incident...), les modalités de la chute (quels sont les différents membres du corps mobilisés, la direction et l'amplitude de chaque mouvement, la vélocité, ses réactions au sol, ainsi que les temps d'action ou d'inaction), ainsi que les phrases prononcées.

Grâce à cette étape, de nouvelles phrases associées à une situation de détresse et des phrases d'appels aux aidants ont été ajoutées à la liste des phrases existantes (de détresse et anodine), et nous avons également ajouté des phrases anodines proches des phrases de détresse (par exemple, « Le docteur a appelé. »). Les appels aux aidants (voir table 5.1) se caractérisent par la présence d'un mot-clé représentant le nom du système, suivi de la demande d'un appel à une personne donnée (ex : « e-lio appelle une infirmière »).

A partir de là, nous avons disposé d'une nouvelle liste de 178 phrases (126 précédentes + 52 nouvelles), et nous les avons utilisées pour continuer l'enregistrement du corpus, en enregistrant cette fois-ci aussi bien des personnes âgées que des personnes jeunes afin d'obtenir 2 groupes distincts : le groupe *locuteurs âgés* et le groupe *locuteurs jeunes*.

31 locuteurs jeunes (17 femmes et 14 hommes) âgés de 18 à 59 ans ont ainsi été enregistrés en 2013, et ont lu au total 5518 phrases de détresse, d'appels aux aidants et anodines, pour une durée de 1 heure 39 minutes.

De plus, 14 nouveaux locuteurs âgés (11 femmes et 3 hommes) ont été enregistrés dans le centre de réhabilitation pour personnes âgées où nous étions déjà intervenus en 2012. Ils étaient âgés de 68 à 90 ans. Ces personnes ont lu au total 2492 phrases, ce qui représente un total de 1 heure 4 minutes d'enregistrement.

La liste complète des phrases se trouve en annexe C.

5.2.2.3 Le corpus résultant

L'ensemble de ces enregistrements a été intégré dans le corpus *AD80*, qui regroupe donc le corpus *AD* initial de 2004 et des corpus réalisés en 2012 et 2013. Nous n'avons pas intégré le corpus *VA* dans le corpus *AD80* du fait de la trop grande disparité des énoncés en terme de quantité et de contenu par rapport aux énoncés du corpus *AD80*. Le corpus *AD80* a été utilisé pour l'évaluation des systèmes de RAP (voir figure 5.2).

Le corpus *AD80* est composé d'un total de 95 locuteurs (59 femmes et 36 hommes). Le groupe *locuteurs âgés* est composé de 43 locuteurs (32 femmes et 11 hommes) âgés de 62 à 94 ans, qui ont lu au total 2663 phrases de détresse (52 minutes) et 434 phrases d'appels aux aidants (16 minutes), et 3006 phrases anodines (1 heure et 6 minutes). Le groupe *locuteurs jeunes* est composé de 52 locuteurs (27 femmes et 25 hommes) âgés de 18 à 64 ans, qui ont lu au total 3306 phrases de détresse (56 minutes), 961 phrases d'appels aux aidants (26 minutes), et 3897 phrases anodines (1 heure et 12 minutes). Un locuteur de 62 ans, bien qu'il soit sous notre seuil de personnes âgées fixé à 65 ans, a été placé dans le groupe de personnes

Phrases de détresse	Appels aux aidants	Phrases anodines
Aidez-moi !	e-lio appelle le SAMU !	Bonjour madame !
Au secours !	e-lio appelle les pompiers !	Ça va très bien.
Je me sens mal !	e-lio appelle les secours !	Ce livre est intéressant.
Je suis tombé !	e-lio appelle un docteur !	Il fait soleil.
Du secours s'il vous plaît !	e-lio appelle une ambulance !	J'ai ouvert la porte.
Je ne peux plus bouger !	e-lio appelle une infirmière !	Je dois prendre mon médicament !
Je ne suis pas bien !	e-lio appelle ma fille !	J'allume la lumière !
Je suis blessé !	e-lio appelle mon fils !	Je me suis endormi tout de suite !
Je ne peux pas me relever !	e-lio tu peux téléphoner au samu ?	Le café est brûlant !
Ma jambe ne me porte plus !	e-lio il faut appeler les secours !	Où sont mes lunettes ?

TABLE 5.1: *Exemples de phrases du corpus AD80.*

âgées du fait de sa présence en maison de retraite, sa perte d'autonomie et son vieillissement physique avancé. Finalement, le corpus *AD80* est constitué de 14 267 phrases annotées, avec 4 heures et 49 minutes d'enregistrements.

10 exemples de phrases de détresse, d'appels aux aidants et de phrases anodines sont donnés table 5.1.

Par ailleurs, durant l'acquisition du corpus *AD80* en 2012 et 2013, il a également été demandé aux locuteurs jeunes et âgés de lire le texte utilisé pour le corpus *ERES38*, décrit dans la section suivante, afin d'effectuer une adaptation aux locuteurs des modèles acoustiques. Au total, 2 heures et 41 minutes de lectures ont été prononcées (1 heure par les locuteurs jeunes et 1 heure et 41 minutes par les locuteurs âgés).

5.3 Corpus ERES38

Dans le même temps, un corpus de parole spontanée a été enregistré auprès de personnes âgées en collaboration avec Sasa et Legrand (2011). Le corpus *ERES38* (Entretiens RESidences 38) a pour but d'être utilisé pour l'adaptation des modèles acoustiques à la voix des personnes âgées (voir figure 5.2) et pour étudier les caractéristiques de la voix des personnes âgées. Ce corpus a été enregistré en 2011 dans le lieu de vie des personnes âgées volontaires, qui étaient résidentes de structures spécifiques pour personnes âgées (foyers logements ou maisons de retraite) dans l'agglomération grenobloise.

Ce corpus est constitué principalement de parole spontanée : il a été demandé à chaque locuteur de parler librement de leur vie dans une interview semi-dirigée. Chaque entrevue s'est déroulée entre une personne âgée et deux expérimentateurs dont l'un s'est fait interlocuteur privilégié. Une première partie introductive permettait de récupérer les informations personnelles ainsi que les habitudes linguistiques du locuteur. Cette phase d'habituation avec le matériel d'enregistrement permettait d'établir le passage vers une parole un peu plus informelle et spontanée pour recueillir le récit de vie de la personne, incluant une description des activités quotidiennes et/ou de leur habitat, un récit d'accidents éventuels et des anecdotes choisis par la personne interrogée elle-même.

Une activité de lecture a également été proposée lors des entretiens. Le texte lu (voir le

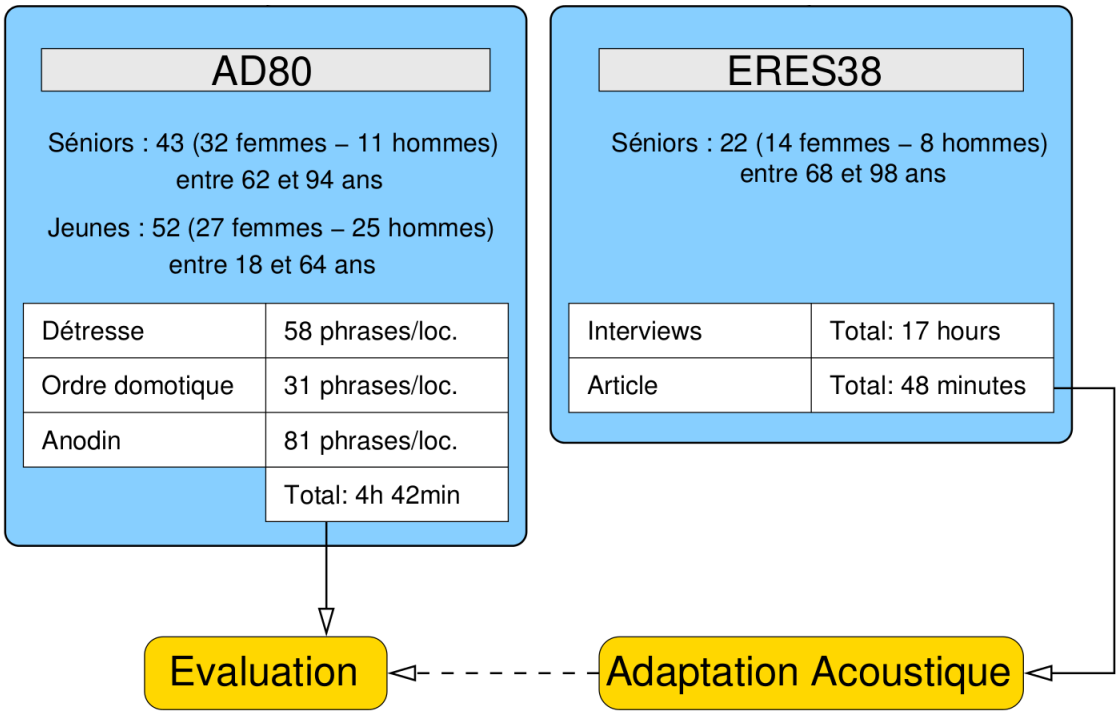


FIGURE 5.2: *Les corpus utilisés pour notre étude sur la voix âgée.*

texte complet en annexe E) visait à être cohérent et fluide avec un sujet attractif pour une personne âgée. Le support choisi est un article de jardinage créé par les expérimentateurs afin de faire en sorte qu'il soit phonétiquement riche dans l'optique de pouvoir étudier quels phonèmes sont les plus mal reconnus par les systèmes de RAP (Dugheanu, 2011).

Au final, le corpus *ERES38* a été acquis auprès de 23 personnes âgées (16 femmes et 7 hommes) âgées entre 68 et 98 ans. Le corpus inclut 48 minutes de lectures annotées et 16 heures et 56 minutes d'interviews, dont 4 heures et 53 minutes ont été annotées.

5.4 Corpus Voix Détresse

Le corpus *Voix Détresse* a été constitué afin d'étudier l'impact d'une voix émue sur un système de RAP en comparaison avec de la voix sans émotions (lecture). Ce corpus a été enregistré en 2013 et 2014 au laboratoire. Il a été demandé aux locuteurs de lire 20 phrases de détresse de façon neutre (voir la liste des phrases en annexe F). Puis, nous avons associé à chaque phrase une photo (voir les photos en annexe F) représentant une situation où un personnage est en détresse, et avons demandé aux locuteurs de se mettre dans la peau des personnages et d'énoncer les phrases de façon très expressive. Les émotions recherchées était principalement les émotions négatives telles que la peur, la colère et la tristesse.

Nous avons constitué deux groupes, le groupe *locuteurs jeunes* et le groupe *locuteurs âgés*.

Dans le groupe *locuteurs jeunes*, nous avons enregistré 20 locuteurs (12 femmes et 8 hommes), âgées entre 23 et 60 ans (moyenne : 32 ans). Les locuteurs ont été recrutés parmi le personnel du laboratoire. 421 phrases neutres ont été prononcées, soit 6 minutes d'enregistrement, et 1001 phrases émues ont été prononcées, soit 16 minutes d'enregistrement.

5 locuteurs (tous de sexe féminin) ont permis de constituer le groupe *locuteurs âgés*, recrutés dans des maisons de retraite de Grenoble et enregistrés au laboratoire. Ces locuteurs étaient âgés entre 67 et 85 ans. 220 phrases neutres ont été prononcées, soit 1 minutes et 30 secondes d'enregistrement, et 320 phrases émues ont été prononcées, soit 4 minutes et 30 secondes d'enregistrement.

Au total, 1742 phrases ont été prononcées (521 phrases neutres et 1221 phrases émues), soit 28 minutes d'enregistrement (7 minutes et 30 secondes pour les phrases neutres et 20 minutes et 30 secondes pour les phrases émues).

5.5 Bilan

Au final, 3 corpus, le corpus de parole lue *AD80* prononcé à la fois par des personnes âgées et des personnes jeunes, le corpus de parole lue et spontanée *ERES38* prononcé par des personnes âgées et le corpus *Voix Détresse* prononcé par des personnes jeunes et âgées, sont utilisés dans notre étude.

Le corpus *AD80* est constitué de phrases lues de façon neutre, de type appels de détresse (ex. : « Aidez-moi ! »), appels aux aidants (ex. : « e-lio appelle une infirmière ! ») ou phrases anodines (ex. : « Je dois prendre mon médicament. »). Le corpus a été enregistré d'une part par un groupe de personnes âgées avec 43 locuteurs d'un âge situé entre 62 et 94 ans, et, d'autre part, par un groupe de personnes jeunes, avec 52 locuteurs entre 18 et 64 ans, ce qui représente 4 heures et 49 minutes d'enregistrements annotés au total.

Le corpus *ERES38* est constitué de 17 heures d'interviews et 48 minutes de lecture d'un article de journal sur l'apiculture, enregistrées auprès de personnes âgées. 23 personnes âgées ont été enregistrées, leur âge étant compris entre 68 et 98 ans. Le texte sur l'apiculture a également été lu par les locuteurs du corpus *AD80* enregistrés en 2012 et 2013 pour effectuer une adaptation acoustique aux locuteurs, soit 72 locuteurs pour 2 heures et 41 minutes d'enregistrement.

Enfin, le corpus *Voix Détresse* est constitué de deux groupes, le groupe de personnes âgées (67 à 85 ans), avec 5 locuteurs, et le groupe de personnes jeunes (23 à 60 ans), avec 20 locuteurs. Ce corpus est constitué de phrases de détresse, qui sont soit lues de façon neutre, soit jouées de façon à contenir des émotions. Ce corpus représente au final 1742 énoncés, soit 28 minutes d'enregistrement.

Dans notre étude, le corpus *AD80* sera utilisé pour effectuer une comparaison des performances des systèmes de RAP entre le groupe *locuteurs âgées* et le groupe *locuteurs jeunes*. Le corpus *ERES38* sera utilisé pour étudier les caractéristiques acoustiques des voix des personnes âgées, et pour adapter les modèles acoustiques à ce type de voix. Enfin, le corpus *Voix Détresse* sera utilisé pour évaluer les performances des systèmes de RAP lorsque la voix contient des émotions.

CHAPITRE 6

Développement d'un système de RAP adapté à la tâche pour la voix âgée

Dans ce chapitre, nous allons décrire la structure des systèmes de RAP ainsi que les évaluations effectuées à partir des corpus de voix âgées par 2 systèmes différents. Les améliorations obtenues après les différents processus d'adaptation seront présentés. La dernière section présentera l'étape de détection des phrases cibles et les évaluations globales.

6.1 Principes généraux des systèmes de RAP

Dans cette section nous allons présenter les principes généraux théoriques des systèmes de reconnaissance automatique de la parole actuels. Les informations sont essentiellement extraites de l'ouvrage « Reconnaissance automatique de la parole : Du signal à son interprétation » de Haton et coll. (2006).

La reconnaissance automatique de la parole (RAP) fait appel à diverses données que l'on peut classer en 3 catégories : les modèles acoustiques décrivant les entités (les unités de base) à reconnaître, le lexique codant les mots du vocabulaire, et un modèle de langage décrivant la structure des phrases du langage, comme illustré figure 6.1. Au cours d'une phase préalable d'apprentissage, les différentes entités sont mémorisées par le système. Ces entités peuvent être de différentes formes selon le type de système : mots, phonèmes, syllabes, etc. Le phonème est aujourd'hui devenu l'unité de base de la plupart des systèmes de RAP, très souvent dans sa version « triphone » permettant une représentation du contexte.

La reconnaissance automatique de la parole demande une paramétrisation du signal vocal. La paramétrisation permet de réduire la redondance du signal vocal, et de diminuer le temps de traitement et l'encombrement en mémoire. Les paramètres extraits doivent être aussi invariants que possible et être pertinents pour la reconnaissance (caractéristiques des sons bruités, fréquences des formants, etc.). Généralement, ces paramètres sont estimés via des fenêtres glissantes sur le signal. Cette analyse par fenêtre permet d'estimer le signal sur une portion du signal jugée stationnaire : généralement 10 à 30 ms en limitant les effets de bord et les discontinuités du signal via une fenêtre de Hamming (Lecouteux, 2008). La méthode actuelle adoptée dans la plupart des systèmes de RAP est la méthode d'analyse spectrale, dans une version fournissant en ensemble de coefficients MFCC (*Mel Frequency Cepstrum Coefficients*). Ces coefficients sont intéressants car ils sont peu sensibles à la puis-

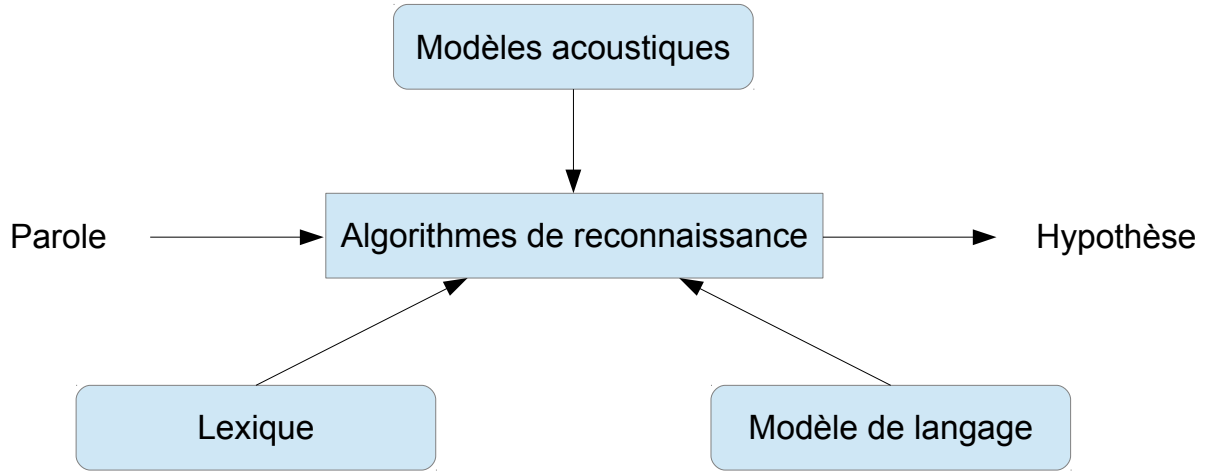


FIGURE 6.1: *Architecture d'un système de reconnaissance automatique de la parole (Haton et coll., 2006)*

sance du signal analysé. L'ajout des dérivées premières et secondes par rapport au temps des coefficients cepstraux rend ces derniers encore plus résistants aux fluctuations dues au locuteur ou à l'environnement.

Dans la plupart des systèmes actuels, les méthodes de reconnaissance sont des méthodes de modélisations stochastiques avec une approche de décision bayésienne. La reconnaissance revient à trouver, connaissant la suite de vecteurs (les paramètres du signal) en entrée X, la suite de mots W dont la probabilité conditionnelle $P(W|X)$ est maximale. D'après la formule de Bayes, $P(W|X)$ peut être calculée à partir du produit des deux quantités suivantes (voir figure 6.2) :

— $P(X|W)$: la probabilité d'observer la séquence de vecteurs X lorsque la suite de mots W est prononcée. Cette probabilité est donnée par le modèle acoustique,

— $P(W)$: la probabilité de la suite de mot W. Elle est fournie par le modèle de langage,

soit :

$$\tilde{W} = \arg\max_{W} P(X|W)P(W) \tag{6.1}$$

Les deux modèles, acoustiques et linguistiques, doivent être ajustés au cours d'une phase préalable d'apprentissage, nécessitant une très grande quantité de données acoustiques et linguistiques représentatives des conditions de l'application considérée.

Les modèles acoustiques représentant les entités (généralement les phonèmes) à reconnaître prennent couramment dans les systèmes actuels la forme de modèles HMM (*Hidden Markov Model*). Un HMM (figure 6.3) est un automate probabiliste contrôlé par deux processus stochastiques. Le premier débute à l'état initial et se déplace d'état en état, en respectant les transitions de l'automate. Le second génère une observation dans chaque état du HMM. A chaque état du modèle HMM est associée une distribution de probabilité (l'observation) modélisant la génération des vecteurs acoustiques via cet état.

On suppose ainsi que le signal de parole est produit par une suite d'états, chaque état étant gouverné par une loi statistique. Chaque unité de parole (phonème) est associée à un modèle HMM et la concaténation de tels modèles permet d'obtenir des mots et phrases. Les

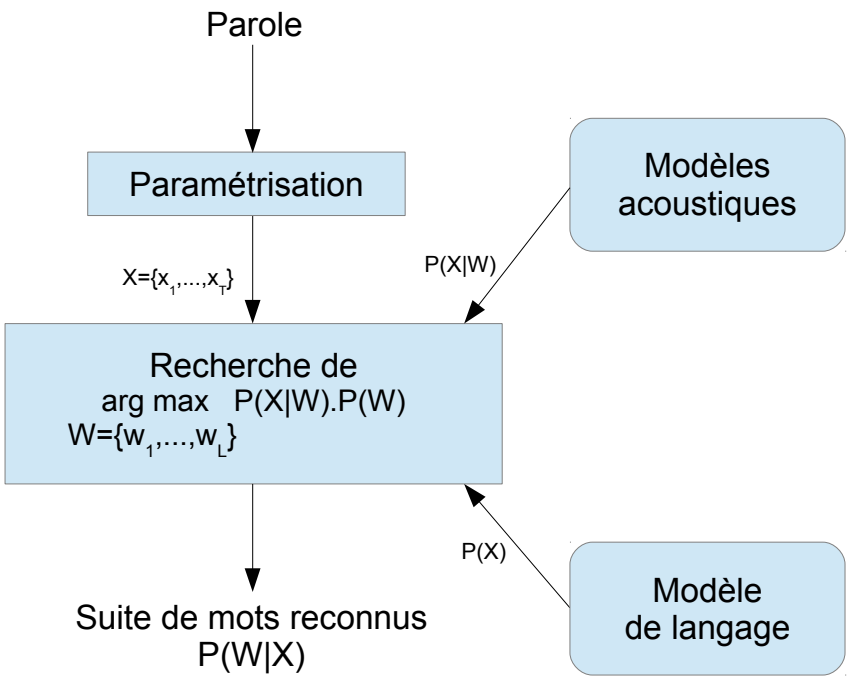


FIGURE 6.2: *Principe de la reconnaissance avec une approche bayésienne (Haton et coll., 2006)*

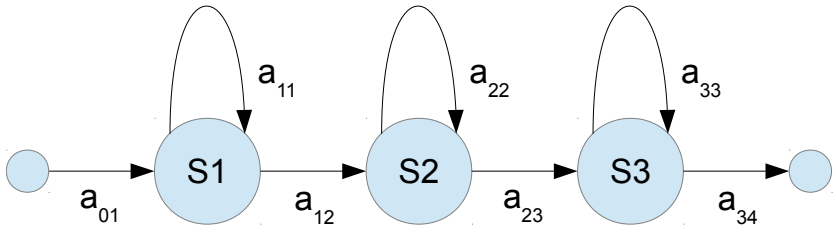


FIGURE 6.3: *Exemple de HMM*

modèles HMM sont généralement des modèles gauche-droite à 3 états. Un modèle à 3 états, appelé triphone, représentera l'état du début du phonème, l'état de milieu du phonème et l'état de fin du phonème, permettant de prendre en compte le phénomène de coarticulation (un même phonème peut être prononcé différemment selon son contexte). Ensuite, la reconnaissance revient à choisir le HMM ayant la plus grande probabilité d'avoir émis le signal en entrée.

L'apprentissage des modèles HMM utilise des algorithmes itératifs qui estiment les différents paramètres et les distributions de probabilités à partir de données acoustiques annotées représentatives de l'application envisagée. Différentes techniques d'apprentissage existent pour modéliser les distributions de probabilités des HMM (Lecouteux, 2008) : nous citerons par exemple l'algorithme EM (espérance-maximisation), qui est une méthode de maximisation de vraisemblance des paramètres proposée par Dempster et coll. (1977), et la méthode MMIE (*Maximum Mutual Information Estimation*), qui est une approche discriminante introduite par Bahl et coll. (1986). D'autres approches discriminantes fonctionnent sur une principe similaire à celui du MMIE : MPE (*Minimum Phone Error*) (Povey et Woodland, 2002) et MWE (*Minimum Word Error*) (Heigold et coll., 2005; Yan et coll., 2008).

Les méthodes d'adaptation des modèles acoustiques permettent de modifier les paramètres acoustiques d'un modèle générique pour rapprocher ce dernier du corpus de test,

permettant par exemple d'adapter un modèle acoustique générique à un locuteur spécifique (Lecouteux, 2008). Deux exemples de méthodes d'adaptation sont l'adaptation MAP (adaptation par maximum a posteriori) (Gauvain et Lee, 1994) et l'adaptation MLLR (*Maximum Likelihood Linear Regression*) (Gales, 1998).

D'un autre côté, le modèle de langage modélise les contraintes liées à une langue. Le rôle d'un modèle de langage statistique est d'estimer cette probabilité *a priori*, afin d'estimer la probabilité d'une suite de mots.

La probabilité $P(W)$ donnée par le modèle de langage est définie par la formule suivante :

$$P(w_1 ... w_n) = \prod_{i=1}^{n} P(w_i | w_1 ... w_{i-i}) \tag{6.2}$$

La suite de mots $w_1 ... w_{i-1}$ est appelée l'historique du mot w_i. En pratique, il est impossible de trouver suffisamment de textes pour exprimer correctement la probabilité de production d'un mot sachant un historique aussi long. On est donc amené à supposer que la probabilité d'observation du mot w_i dépend uniquement des k mots précédents :

$$P(w_1 ... w_n) = \prod_{i=1}^{n} P(w_i | w_{1-k+1} ... w_{i-i}) \tag{6.3}$$

Lorsque k vaut 2 ou 3, on parlera respectivement de modèles bigrammes ou trigrammes.

L'apprentissage des modèles de langage se fait à partir de grands corpus de texte. L'estimation des paramètres d'un modèle de langage s'effectue en combinant 2 modèles (Lecouteux, 2008) :

— le modèle de décompte : effectue un décompte des suites de mots observées pour en extraire une probabilité d'observation. Les méthodes utilisées sont le *linear discounting*, l'*absolute discounting*, ou le *good-turing*,

— le modèle de redistribution : permet un lissage des probabilités afin d'estimer les événements non observés. Les méthodes utilisées sont les méthodes d'interpolation et de repli.

Enfin, le moteur de reconnaissance permet le décodage du signal de parole. Un tour d'horizon des différentes techniques de décodage est présenté dans (Aubert, 2002). Les systèmes actuels fonctionnent généralement en 2 passes. Une première passe génère un treillis de mots en procédant de façon synchrone (trame par trame) à partir des observations et du modèle de langage. Puis une deuxième passe va générer les N meilleures phrases en explorant le treillis afin de trouver le chemin qui maximisera la fonction de coût, qui regroupe les hypothèses linguistiques et acoustiques. Cette deuxième passe génère les solutions possibles grâce à l'algorithme A^* (Lecouteux, 2008).

Pour plus de détails mathématiques et algorithmiques sur la reconnaissance automatique de la parole, nous vous invitons à vous référer à l'ouvrage de Haton et coll. (2006).

6.2 Référence pour la voix âgée

Afin de comparer les performances des systèmes de reconnaissance automatique de la parole entre les voix jeunes et les voix âgées, nous avons réalisé des décodages sur les phrases de détresse et les phrases d'appels aux aidants de notre corpus *AD80*. Nous avons réalisé l'étude sur 2 systèmes de RAP différents : *Sphinx3* et *Google Speech API*.

6.2.1 Décodage avec Sphinx3

Avec le décodeur de *Sphinx3*, nous avons utilisé un modèle acoustique de type HMM, contexte-dépendant, triphone, appris sur le corpus *BREF120* (Lamel et coll., 1991). *BREF120* est un large corpus oral de textes lus, contenant plus de 100 heures de parole produites par 120 locuteurs (65 femmes et 55 hommes), tous les textes enregistrés ont été extraits du journal *Le Monde*. Nous avons appelé ce modèle générique le modèle acoustique *BREF120*.

Un modèle de langage spécifique aux phrases à reconnaître a été estimé à partir des phrases de détresse et des phrase d'appels aux aidants du corpus *AD80*. Ce modèle est trigramme, avec 88 unigrammes, 193 bigrammes et 223 trigrammes. Un modèle de langage général a été appris à partir du corpus *Gigaword*[1] qui est une collection d'articles de journaux ayant été acquis sur plusieurs années par le *Linguistic Data Consortium* (LDC) de l'Université de Pennsylvanie. Ce modèle est unigramme et contient 11018 mots. Le modèle final est la combinaison du modèle spécifique avec le modèle général, en donnant un poids plus important (90%) sur les probabilités du modèle spécifique. La combinaison d'un modèle de langage général avec un modèle spécifique de poids plus important a montré qu'elle conduisait à un meilleur WER (voir la définition du WER en annexe G.0.1) pour les applications de RAP spécifique à un domaine donné (Lecouteux et coll., 2011). L'intérêt est de biaiser la reconnaissance vers le domaine spécifique, mais lorsque le locuteur dévie du domaine, le modèle général permet d'éviter de reconnaître des phrases hors-domaine comme appartenant au domaine.

Nous avons réalisé des décodages à partir des modèles précédemment décrits sur les phrases de détresse et d'appels aux aidants du corpus *AD80* (corpus décrit en section 5.2.2.3). Les voix ont été réparties en 4 groupes :

— *hommes jeunes* (25 locuteurs),

— *femmes jeunes* (27 locuteurs),

— *hommes âgés* (11 locuteurs),

— *femmes âgées* (32 locuteurs).

Les résultats des décodages sont présentés table 6.1.

Nous observons une dégradation du WER moyen pour les voix âgées avec une différence absolue de 34,7% (voir la définition de la différence absolue en annexe G.0.3), hommes et femmes confondus, par rapport aux voix jeunes. Aussi, nous observons une dégradation plus

1. http ://catalog.ldc.upenn.edu/LDC2006T17

	Jeunes	Âgés
Hommes	11,7%	61,3%
Femmes	10,4%	40,3%
Hommes+femmes	11,0%	45,7%

TABLE 6.1: *WER moyens issus du décodage avec Sphinx3 (modèle acoustique BREF120).*

importante pour les hommes (49,6%) que pour les femmes (29,9%). De façon générale, les voix des femmes ont un meilleur WER que les voix des hommes, cela étant probablement dû à la quantité plus importante de données de voix de femmes utilisées lors de l'apprentissage du modèle acoustique (65 femmes pour 55 hommes dans *BREF120*). Les résultats obtenus avec le modèle acoustique générique *BREF120* constituent ainsi notre niveau de référence ou *baseline*.

La figure 6.4 représente les WER de chaque locuteur en fonction de l'âge pour les groupes *hommes jeunes*, *femmes jeunes*, *hommes âgés* et *femmes âgées*. Nous pouvons également observer un WER plus important pour les voix âgées, comme cela a été montré lors de précédentes études : sans adaptation des modèles acoustiques, Vipperla et coll. (2008) obtiennent un WER de 36,4% pour les adultes et de 47,8% pour les personnes âgées, et Pellegrini et coll. (2012) obtiennent avec leur système de référence un WER de 29,1% pour les personnes de 60 à 65 ans et de 54,9% pour les personnes de 86 à 90 ans.

De plus, nous pouvons également remarquer que la variabilité entre locuteurs augmente avec l'âge. Par exemple, deux locuteurs du même âge, 81 ans, ont leur WER qui est égal pour l'un à 13,7%, et pour l'autre à 63,8%. L'écart-type entre les WER est de 6,4% pour les voix jeunes et de 16,8% pour les voix âgées. Ainsi, le WER est moins prévisible pour les voix âgées que pour les voix jeunes.

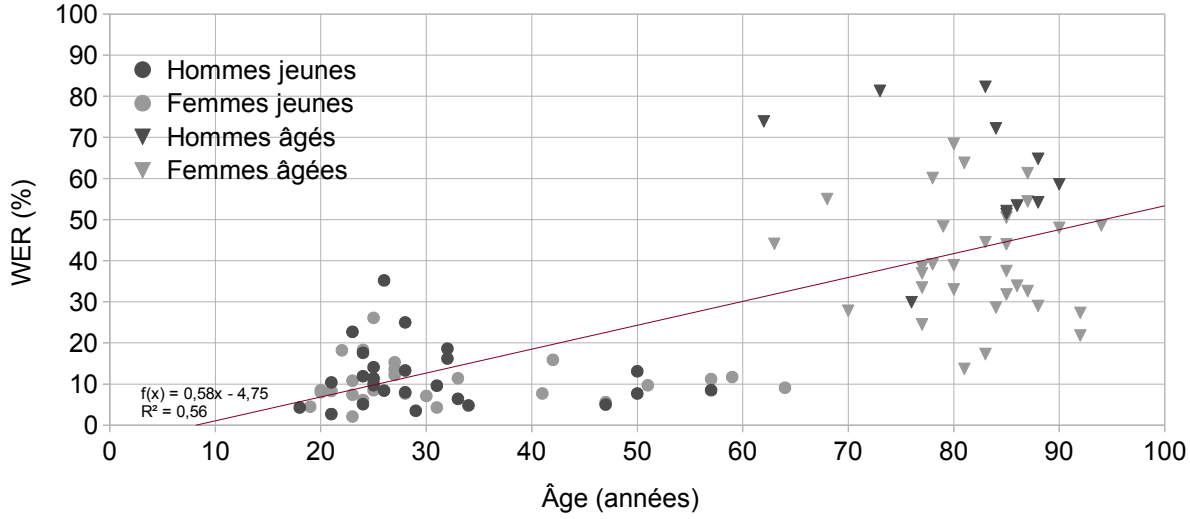


FIGURE 6.4: *WER en fonction de l'âge pour les différents groupes avec Sphinx3 (modèle acoustique BREF120), avec la droite de régression linéaire.*

6.2.2 Décodage avec Google Speech API

A l'aide du système *Google Speech API*, nous avons réalisé des décodages sur les phrases de détresse et d'appels aux aidants du corpus *AD80* (même données que précédemment). Les résultats des décodages sont présentés table 6.2.

Avec le système *Google Speech API*, de la même manière qu'avec *Sphinx3*, nous observons une dégradation des performances de la RAP pour les voix âgées par rapport aux voix jeunes. En effet, la différence absolue entre le WER moyen des voix âgées et des voix jeunes, hommes et femmes confondus, est de 21,2%.

De plus, la dégradation est plus importante pour les hommes (29,6%) que pour les femmes (18,5%). Pour les voix jeunes, la différence entre hommes et femmes est quasiment nulle. En revanche, pour les voix âgées, les femmes ont un meilleur WER que les hommes, la différence absolue étant de 11,5%. Nous n'avons pas accès ici aux données utilisées pour construire le modèle acoustique et cette différence entre hommes et femmes ne peut pas être attribuée au corpus d'apprentissage dans ce cas.

La figure 6.5 représente les WER de chaque locuteur en fonction de l'âge pour les groupes *hommes jeunes*, *femmes jeunes*, *hommes âgés* et *femmes âgées*. Nous pouvons remarquer que la variabilité des WER entre locuteurs augmente avec l'âge. L'écart-type est de 11,2% pour les voix jeunes et de 20,0% pour les voix âgées. De la même manière qu'avec *Sphinx3*, nous observons que le WER est moins prévisible pour les voix âgées que pour les voix jeunes.

	Jeunes	Âgés
Hommes	19,7%	49,3%
Femmes	19,3%	37,8%
Hommes+femmes	19,5%	40,7%

TABLE 6.2: *WER moyens issus du décodage avec Google Speech API.*

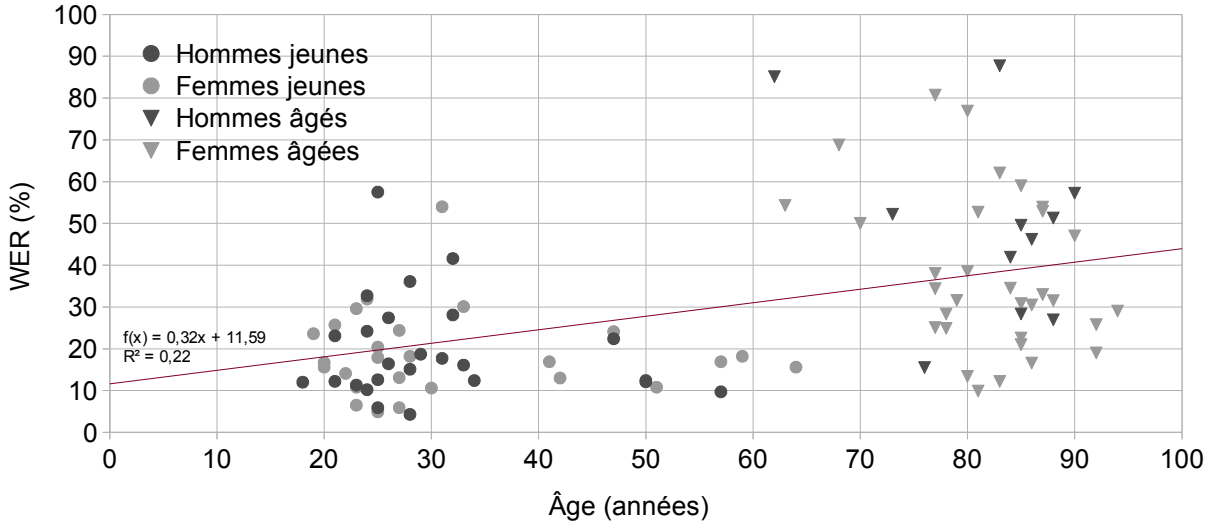


FIGURE 6.5: *WER en fonction de l'âge pour les différents groupes avec Google Speech API, avec la droite de régression linéaire.*

6.2.3 Bilan

Pour les 2 systèmes *Sphinx3* et *Google Speech API*, nous observons une importante dégradation des performances de la RAP pour les voix âgées par rapport aux voix jeunes. Avec *Sphinx3*, cette dégradation est de 34,7%, et, avec *Google Speech API*, elle est de 21,2%. Ainsi, nous observons une plus grande variabilité au sein des locuteurs âgés qu'au sein des locuteurs jeunes.

Dans la section suivante, nous allons analyser dans quelle mesure cette différence jeunes/âgés peut être diminuée à l'aide des techniques d'adaptation acoustique.

6.3 Adaptation des modèles acoustiques à la voix âgée

La méthode la plus courante pour améliorer les performances de la RAP est d'utiliser l'adaptation des modèles acoustiques aux locuteurs. L'adaptation consiste à générer un nouveau modèle acoustique à partir d'un modèle générique et à partir de données prononcées par les locuteurs qui sont annotées et en quantité limitée. Une des techniques les plus populaires est la méthode *Maximum-Likelihood Linear Regression* (MLLR). La méthode MLLR (Leggetter et Woodland, 1995) est une technique d'adaptation qui utilise de petites quantités de données pour effectuer une transformation linéaire qui déforme la gaussienne des moyennes de façon à maximiser la vraisemblance des données. Le principe est que les classes acoustiquement proches sont regroupées et transformées ensemble.

Différentes adaptations ont été réalisées :

— Adaptation globale à la voix âgée : le modèle acoustique générique *BREF120* est adapté à la voix âgée en général, selon 3 modalités (voix mixtes, féminines, masculines) présentées table 6.3. Les données pour réaliser l'adaptation sont les textes lus par les personnes âgées du corpus *ERES38*, soit 48 minutes, dont 33 minutes prononcées par les femmes et 15 minutes prononcées par les hommes. Les locuteurs pris en compte pour l'adaptation (corpus *ERES38*) ne sont pas les mêmes que pour le test (corpus *AD80*). Ainsi, pour chaque modalité, un seul modèle acoustique est utilisé pour tous les locuteurs testés.

— Adaptation au locuteur : le modèle acoustique *BREF120* est adapté à chaque locuteur testé. L'adaptation est faite à partir d'un texte que le locuteur a lu au préalable, soit en moyenne 2 minutes et 27 secondes de lectures par locuteur. Pour un locuteur donné, les données utilisées pour le test (corpus *AD80*) appartiennent au même locuteur que les données utilisées pour l'adaptation du modèle acoustique (voir table 6.3). Chaque locuteur testé possède ainsi son propre modèle acoustique adapté à sa voix en particulier.

Nous allons maintenant présenter les résultats obtenus avec ces différentes méthodes d'adaptation.

6.3.1 Adaptation globale à la voix âgée

Nous avons réalisé des décodages avec *Sphinx3* en utilisant les modèles adaptés à la voix âgée (table 6.3). Le modèle de langage utilisé est le modèle appris sur les phrases de détresse et d'appels aux aidants du corpus *AD80* combiné au modèle appris sur le corpus *Gigaword* comme pour la baseline (voir section 6.2.1). Les données testées sont les mêmes que celles utilisées en section 6.2.1. Il s'agit des phrases de détresse et d'appels aux aidants du corpus *AD80*.

La table 6.4 présente les résultats des décodages avec *Sphinx3*, avec le WER moyen pour chacun des groupes. La colonne « BREF120 » représente notre référence avec les résultats obtenus avec le modèle acoustique générique (voir section 6.2.1)

Nous observons que pour les voix âgées masculines, l'adaptation MLLR globale à la voix âgée (modèle acoustique *BREF120_MLLR_G*) améliore le WER de 35,6% (différence absolue). L'adaptation à la voix âgée masculine (modèle acoustique *BREF120_MLLR_H*) est encore meilleure, avec une amélioration de 39% du WER.

Pour les voix âgées féminines, l'adaptation MLLR globale à la voix âgée améliore le WER de 25,7%. L'adaptation à la voix âgée féminine (modèle acoustique *BREF120_MLLR_F*) améliore le WER de 24,9%, et est légèrement moins bonne que l'adaptation globale à la voix âgée. Les modèles adaptés à la voix âgée avec ou sans prise en compte du genre ont donc des résultats très proches entre eux.

Nous avons aussi voulu observer les conséquences de l'utilisation de modèles adaptés à la voix âgée par des locuteurs jeunes, le cas pouvant se présenter en situation réelle.

Pour le groupe *hommes jeunes*, l'utilisation d'un modèle adapté globalement à la voix âgée ne change quasiment pas le WER (diminution de 0,6%), de même que l'utilisation d'un modèle adapté à la voix âgée masculine (augmentation très faible de 0,1%).

Pour le groupe *femmes jeunes*, l'utilisation d'un modèle adapté globalement à la voix âgée améliore le WER de 2,8%. L'adaptation à la voix âgée féminine améliore de 3,1% le WER.

Modèle générique	Lectures d'adaptation	Type d'adaptation	Modèles adaptés
BREF120	Ensemble des voix du corpus ERES38	A la voix âgée	BREF120_MLLR_G
BREF120	Voix masculines du corpus ERES38	A la voix âgée masculine	BREF120_MLLR_H
BREF120	Voix féminines du corpus ERES38	A la voix âgée féminine	BREF120_MLLR_F
BREF120	Texte lu par le locuteur	Au locuteur	BREF120_MLLR_LOC

TABLE 6.3: *Modèles adaptés à la voix âgée à partir du modèle générique BREF120.*

Groupe	BREF120	BREF120_MLLR_G	BREF120_MLLR_H	BREF120_MLLR_F
Hommes jeunes	11,7%	11,1%	11,8%	-
Femmes jeunes	10,4%	7,6%	-	7,3%
Hommes âgés	61,3%	25,7%	22,3%	-
Femmes âgées	40,3%	14,6%	-	15,4%

TABLE 6.4: *WER moyens pour les différents groupes (95 locuteurs) en fonction des modèles acoustiques BREF120 et BREF120 adaptés.*

Au vu des résultats très proches obtenus pour chaque groupe après les adaptations du modèle *BREF120* à la voix âgée en général (modèle *BREF120_MLLR_G*) et à la voix âgée en fonction du genre (modèles *BREF120_MLLR_H* et *BREF120_MLLR_F*), nous avons réalisé une ANOVA sur les résultats de WER en comparant les modèles. Cette approche permettra de vérifier si une adaptation selon le genre se justifie par rapport à une adaptation globale sur les 2 genres.

Le test de l'ANOVA (*ANalysis Of VAriance*) permet d'étudier le comportement d'une variable continue à expliquer en fonction d'une variable explicative catégorielle. La variable explicative est une variable qualitative, elle est aussi appelée « facteur », et les différentes modalités du facteur sont appelées les « niveaux ». La variable continue est une variable quantitative. Un échantillon correspond aux valeurs prises par la variable explicative pour un niveau donné du facteur. L'objectif de l'ANOVA est de tester si les moyennes de différents échantillons ont des différences statistiquement significatives.

Les conditions de l'ANOVA sont les suivantes :

— Les distributions des échantillons suivent une loi normale. Cependant, l'ANOVA est peu sensible à cette condition, les résultats de l'ANOVA restant valables si les distributions s'écartent un peu de la loi normale. Le test de Shapiro-Wilk permet de tester si un échantillon suit une loi normale ou non. Les hypothèses du test de Shapiro-Wilk sont les suivantes :

  * H_0 : l'échantillon suit une loi normale,
  * H_1 : l'échantillon ne suit pas une loi normale.

  Si la valeur de la p-value est significative (p-value < 0,01), il faut alors rejeter l'hypothèse nulle selon laquelle la distribution de l'échantillon est normale. La normalité est donc supposée si p-value > 0,01.

— Les variances des échantillons sont homogènes (hypothèse d'homoscédasticité). L'ANOVA est très sensible à cette condition. Le test de Levene permet de tester si les variances entre les échantillons sont égales. Les hypothèses du test de Levene sont les suivantes :

  * H_0 : les variances entre les échantillons sont égales,
  * H_1 : les variances entre les échantillons ne sont pas égales.

  Si p-value < 0,05, l'hypothèse nulle est rejetée, cela signifiant qu'il y a une différence significative entre les variances des échantillons. L'homoscédasticité est donc suposée si p-value > 0,05.

Si les conditions du test de l'ANOVA sont respectée, celui-ci peut être effectué. Les hypothèse de l'ANOVA sont les suivantes :

— H_0 : toutes les moyennes sont identiques,

— H_1 : au moins une des moyennes est différente des autres.

L'hypothèse nulle selon laquelle toutes les moyennes sont identiques est rejetée lorsque p-value < 0,05.

Groupe	BREF120	BREF120_MLLR_G	BREF120_MLLR_H	BREF120_MLLR_F
Hommes jeunes	**p=0,007936**	**p=3,247e-06**	**p=2,936e-06**	-
Femmes jeunes	p=0,04187	**p=0,008146**	-	p=0,01398
Hommes âgés	p=0,4874	p=0,191	p=0,6275	-
Femmes âgées	p=0,9144	p=0,02548	-	p=0,02417

TABLE 6.5: *p-values issues du test de Shapiro-Wilk (hypothèse que l'échantillon suit une loi normale validée si p>0,01) pour chaque groupe en fonction du modèle acoustique.*

L'ANOVA permet simplement de savoir si tous les échantillons suivent une même loi normale. Si une différence significative est notée (hypothèse nulle rejetée) cette analyse ne permet pas de savoir quels sont les échantillons qui s'écartent de cette loi.

Pour identifier entre quels échantillons il y a une différence significative, nous pouvons utiliser différents tests « post-hoc », tels que le test de Tukey HSD (Honestly Significant Difference) ou le test de Bonferroni.

Nous avons réalisé une ANOVA sur chacun des groupes *hommes jeunes*, *femmes jeunes*, *hommes âgés* et *femmes âgées*. Pour chaque ANOVA, le facteur est le modèle acoustique, pouvant prendre les valeurs qualitatives *BREF120*, *BREF120_MLLR_G*, *BREF120_MLLR_H* et *BREF120_MLLR_F*, correspondant donc à 4 échantillons. La variable quantitative étudiée est le WER. Nous avons ainsi théoriquement 4 échantillons par groupes, soit 16 échantillons. Cependant, nous n'avons pas effectué les tests avec voix masculines sur le modèle adapté à la voix féminine, de même pour les voix féminines sur le modèle adapté à la voix masculine, car cela ne fait pas sens. Nous avons ainsi au final 12 échantillons.

Pour tester la normalité des échantillons, un test de Shapiro-Wilk a été réalisé. Un échantillon suit une loi normale si l'hypothèse nulle du test est validée, soit p-value > 0.01. La table 6.5 représente les p-values issues du test de Shapiro-Wilk pour les 12 échantillons.

Tous les échantillons suivent une loi normale à l'exception des 4 échantillons représentés en gras dans la table 6.5. Cependant, les 2 échantillons en gras p=0,007936 et p=0,008146 sont assez proche du seuil 0,01. Malgré les 2 échantillons du groupe *hommes jeunes* qui ont leurs p-values très inférieures à 0,01, et en sachant que l'ANOVA est peu sensible si la loi normale n'est pas tout à fait vraie, nous considérons que l'hypothèse de normalité est suffisamment respectée pour effectuer des ANOVA sur les différents échantillons.

Ensuite, un test de Levene a été réalisé pour tester l'égalité des variances entre les échantillons pour chacun des 4 groupes. Les p-values issues des tests de Levene sont présentées

Groupe	Test de Levene
Hommes jeunes	F(2 ;72)=0,3864 ; p=0,6809
Femmes jeunes	F(2 ;78)=0,1425 ; p=0,8674
Hommes âgés	F(2 ;30)=0,3838 ; p=0,6846
Femmes âgées	F(2 ;93)=8,0763 ; **p=0,0005833**

TABLE 6.6: *Résultats du test de Levene pour chaque groupe de locuteurs, obtenus avec 3 échantillons par groupe (BREF120 et BREF120_MLLR_G, et BREF120_MLLR_H ou BREF120_MLLR_F) (hypothèse que les échantillons ont leurs variances homogènes si p>0,05).*

Groupe	Test de l'ANOVA
Hommes jeunes	F(2 ;72)=0,023 ; p=0,9777
Femmes jeunes	F(2 ;78)=2,99 ; p=0,0561
Hommes âgés	F(2 ;30)=27,1 ; **p=0,0000**
Femmes âgées	F(2 ;88)=61,7 ; **p=0,0000**

TABLE 6.7: *Résultats du test de l'ANOVA pour chaque groupe de locuteurs, obtenus avec 3 échantillons par groupe (BREF120 et BREF120_MLLR_G, et BREF120_MLLR_H ou BREF120_MLLR_F) (hypothèse que les échantillons ont leurs moyennes différentes si p<0,05).*

table 6.6. Si p-value > 0,05, cela signifie que les variances des échantillons sont homogènes.

Les échantillons ont leurs variances homogènes au sein des groupes de locuteurs *hommes jeunes*, *femmes jeunes* et *hommes âgés* avec p > 0,05. Le groupe *femmes âgées* a une p-value de 0,0005833, donc inférieure à 0,05. Ce résultat est dû à des WER extrêmes obtenus avec le modèle acoustique baseline *BREF120* pour quelques locutrices. Ainsi 5 locutrices ont des WER de 68,4%, 63,8%, 61,3%, 60,1% et 55%. Or ces mêmes locutrices, après adaptation à la voix âgée (modèle *BREF120_MLLR_G*), obtiennent des WER de 16,1%, 29,3%, 16,5%, 23,4% et 29,3% respectivement, ce qui reste dans des ordres de grandeur comparables aux autres locutrices. Les valeur extrêmes de WER ne sont observées que pour le modèle baseline *BREF120*. En retirant les 5 valeurs extrêmes (WER>=55%) et en effectuant à nouveau le test, la p-value devient supérieure à 0,05 (F(2 ;88)=2,8993 ;p=0,06034). Pour la suite de l'ANOVA, les 5 WER extrêmes sont retirés du jeu de données.

En considérant que les conditions de loi normale et d'homoscédasticité sont vérifiées, nous avons réalisé une ANOVA et un test de Tukey HSD pour chacun des groupes. La table 6.7 présentent les résultats des ANOVA pour chaque groupe, et les tables 6.8a, 6.8b, 6.8c et 6.8d présentent les résultats des tests de significativité issus du test de Tukey HSD portant sur les différences entre paires de modèles acoustiques pour chaque groupe, avec un rappel des WER.

Pour le groupe *hommes jeunes* et *femmes jeunes*, la p-value pour chaque paire de modèles acoustiques est supérieure à 0,05, donc il n'y a pas de différence significative entre les WER moyens issus des différents modèles acoustiques.

Pour les groupes *hommes âgés* et *femmes âgées*, les p-values pour les paires BREF120/BREF120_MLLR_G, BREF120/BREF120_MLLR_H et BREF120/BREF120_MLLR_F sont largement inférieures à 0,05, cela signifiant que les différences de WER moyens entre l'utilisation du modèle acoustique générique *BREF120* et l'utilisation des modèles adaptés à la voix âgée sont significatives. En revanche, il n'y a pas de différence significative entre le modèle acoustique adapté de façon globale à la voix âgée et les modèles acoustiques adaptés à la voix âgée selon le genre. Ainsi, il n'est pas nécessaire de différencier le genre pour l'adaptation des modèles acoustiques à la voix âgée.

En comparant les groupes entre eux, nous observons, pour les femmes, que la différence absolue entre le WER moyen du groupe *femmes âgés* issu du modèle *BREF120_MLLR_G* et le WER moyen du groupe *femmes jeunes* issu du modèle *BREF120* est de 4,2% : les résultats après adaptation pour les femmes âgées sont donc très proches des résultats pour les

	BREF120	BREF120_MLLR_G	BREF120_MLLR_H	BREF120_MLLR_F
BREF120	-	p=0,9821	p=1,0000	-
BREF120_MLLR_G	-	-	p=0,9806	-
BREF120_MLLR_H	-	-	-	-
BREF120_MLLR_F	-	-	-	-
WER (%)	11,7	11,1	11,8	-

(a) Hommes jeunes

	BREF120	BREF120_MLLR	BREF120_MLLR_H	BREF120_MLLR_F
BREF120	-	p=0,1242	-	p=0,0718
BREF120_MLLR	-	-	-	p=0,9648
BREF120_MLLR_H	-	-	-	-
BREF120_MLLR_F	-	-	-	-
WER (%)	10,4	7,6	-	7,3

(b) Femmes jeunes

	BREF120	BREF120_MLLR_G	BREF120_MLLR_H	BREF120_MLLR_F
BREF120	-	**p=0,0000**	**p=0,0000**	-
BREF120_MLLR_G	-	-	p=0,8279	-
BREF120_MLLR_H	-	-	-	-
BREF120_MLLR_F	-	-	-	-
WER (%)	61,3	25,7	22,3	-

(c) Hommes âgés

	BREF120	BREF120_MLLR_G	BREF120_MLLR_H	BREF120_MLLR_F
BREF120	-	**p=0,0000**	-	**p=0,0000**
BREF120_MLLR_G	-	-	-	p=0,9176
BREF120_MLLR_H	-	-	-	-
BREF120_MLLR_F	-	-	-	-
WER (%)	40,3	14,6	-	15,4

(d) Femmes âgées

TABLE 6.8: *WER et p-value du test de Tukey HSD résultants des décodages sur les modèles BREF120 et BREF120 adaptés pour chacun des groupes (hypothèse de différence entre les groupes validée si p<0,05).*

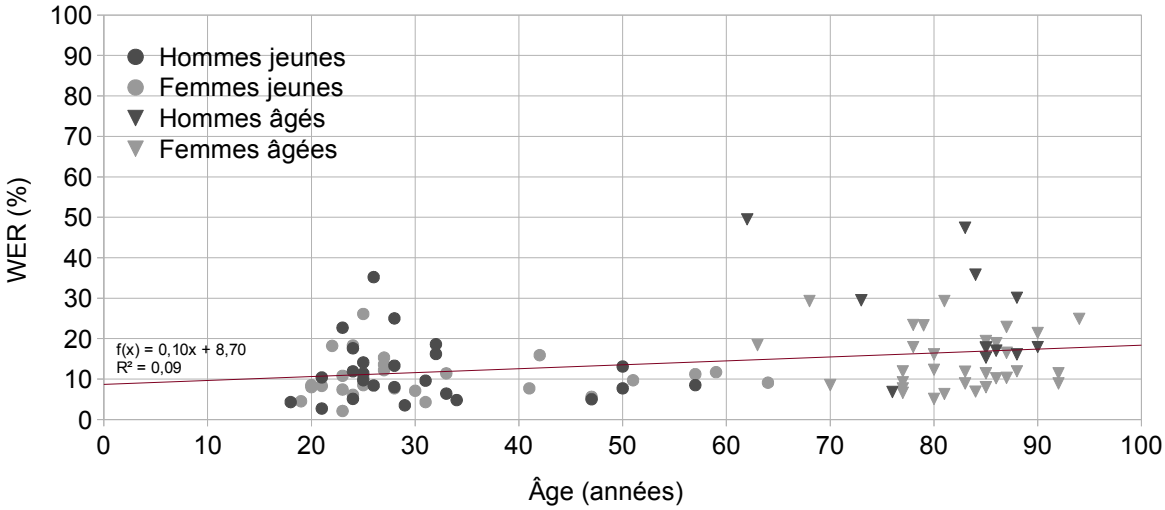


FIGURE 6.6: *WER en fonction de l'âge pour les différents groupes (modèle acoustique BREF120 pour les voix jeunes, modèle acoustique BREF120_MLLR_G pour les voix âgées), avec la droite de régression linéaire.*

femmes jeunes. Pour les hommes, la différence est de 14%, ce qui est un peu moins bon, mais il faut tenir compte du fait que nous possédons moins d'enregistrements de voix d'hommes âgés dans le corpus *AD80* (11 hommes âgés et 32 femmes âgées). L'adaptation acoustique tendrait donc à gommer l'effet de l'âge : la figure 6.6, représentant le WER en fonction de l'âge avec le modèle *BREF120* pour les voix jeunes et le modèle adapté *BREF120_MLLR_G* pour les voix âgées, montre que la droite de régression linéaire a une pente moins importante que lors de l'utilisation du modèle *BREF120* pour les voix âgées (pente de 0,58 avant adaptation (figure 6.4) et de 0,10 après adaptation (figure 6.6)).

6.3.2 Adaptation au locuteur

Nous avons réalisé un décodage avec *Sphinx3* en utilisant des modèles acoustiques adaptés aux locuteurs. Le modèle acoustique *BREF120* a été utilisé comme modèle de base ; nous avons adapté celui-ci à chaque locuteur testé. Pour permettre l'adaptation acoustique (technique MLLR), les locuteurs du corpus *AD80* enregistrés en 2012 et 2013, soit 41 personnes âgées et 31 personnes jeunes, ont eu la tâche de lire un article (texte en annexe E). Au moment du décodage, l'annotation des enregistrements n'avait été réalisée que pour 29 locuteurs, qui étaient des personnes âgées. Les adaptations au locuteur des modèles acoustiques et le décodage ont donc été réalisées sur la voix de ces 29 locuteurs (8 hommes et 21 femmes), âgés de 62 à 94 ans. Pour chaque locuteur, la durée moyenne de l'enregistrement du texte utilisé pour l'adaptation acoustique était de 2 minutes et 27 secondes. Les phrases utilisées pour le décodage étaient les phrases de détresse, soit un total de 1739 phases (31 minutes) prononcées par les 29 locuteurs. Le modèle de langage utilisé est le même que précédemment (voir section 6.2.1).

Les WER obtenus sur ce sous-ensemble de locuteurs à partir du modèle générique *BREF120*, du modèle adapté de façon globale à la voix âgées *BREF120_MLLR_G* et les modèles adaptés aux locuteurs sont donnés table 6.9. Les WER avec *BREF120* et *BREF120_MLLR_G* diffèrent très légèrement des valeurs moyennes données table 6.4 du fait du nombre plus restreint de locuteurs.

Les WER issus des décodages avec les modèles adaptés aux locuteurs sont présentés figure 6.7 pour chaque locuteur. Les WER obtenus avec le modèle générique *BREF120* et le modèle adapté de façon globale à la voix âgées sont également présentés figure 6.7.

D'après la table 6.9, pour les hommes, le WER moyen obtenu avec l'adaptation au locuteur est de 6,13% plus bas (différence absolue) que l'adaptation globale à la voix âgée. Pour les femmes, en revanche, le WER moyen obtenu avec l'adaptation au locuteur est plus haut

Groupe	BREF120	BREF120_MLLR_G	BREF120_MLLR_LOC
Voix hommes âgés (8 loc.)	63,65%	28,93%	22,80%
Voix femmes âgées (21 loc.)	36,74%	13,44%	15,21%
Moyenne (29 loc.)	44,17%	17,71%	17,30%

TABLE 6.9: *WER moyens pour les différents groupes (29 locuteurs au total) en fonction des modèles acoustiques BREF120, BREF120_MLLR_G et BREF120_MLLR_LOC.*

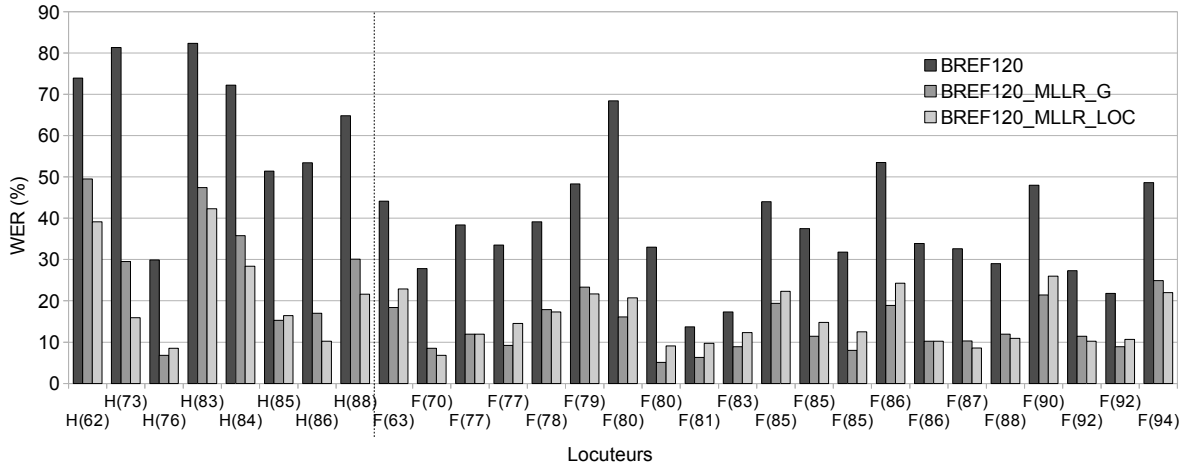


FIGURE 6.7: *WER par locuteur pour le modèle générique (BREF120), le modèle avec adaptation globale à la voix âgée (BREF120_MLLR_G) et les modèles avec adaptation au locuteur (BREF120_MLLR_LOC). L'axe des abscisses est représenté en fonction du genre (homme : H ou femme : F) et de l'âge (entre parenthèses).*

de 1,77% que le WER avec adaptation globale à la voix âgée. Les comparaisons hommes-femmes sont cependant à prendre avec précaution car il y a peu de locuteurs masculins parmi l'échantillon testé. Hommes et femmes confondus, le WER moyen obtenu est quasiment identique entre l'adaptation au locuteur et l'adaptation globale.

Nous observons, d'après la figure 6.7, une grande variabilité du WER entre les locuteurs. Aussi, il ne semble pas y avoir de relation linéaire entre l'âge et le WER au sein des locuteurs âgés. Pour tous les locuteurs âgés, l'adaptation a permis de diminuer de façon importante le WER. En revanche, quant à savoir quelle type d'adaptation – globale à la voix âgée, ou au locuteur – est la meilleure, les résultats sont partagés : pour 13 locuteurs, l'adaptation au locuteur donne un meilleur WER que l'adaptation globale à la voix âgée, et inversement pour les 14 autres locuteurs.

Aussi, une ANOVA suivi d'un test post-hoc de Tukey HSD entre les 3 échantillons correspondant aux 3 modèles *BREF120*, *BREF120_MLLR_G* et *BREF120_MLLR_LOC* ont été réalisés. L'ANOVA montre qu'il existe une différence significative entre certains échantillons : $F(2;84)=37,11$; $p=2,83e\text{-}12$. Le test de Tukey HSD montre une différence significative entre les modèles *BREF120* et *BREF120_MLLR_G* ($p<0,01$) et entre les modèles *BREF120* et *BREF120_MLLR_LOC* ($p<0,01$) ; en revanche, le test de Tukey HSD montre qu'il n'y a pas de différence significative entre les modèles *BREF120_MLLR_G* et *BREF120_MLLR_LOC* ($p=0,99$).

6.4 Détection des phrases cibles

6.4.1 Élaboration du modèle de langage

Une fois notre système de RAP acoustiquement adapté à la voix âgée, nous avons dû déterminer comment le système reconnaît des phrases de type « appels de détresse » et des

ordres de type « appels aux aidants » et les distingue des phrases de type « anodines ». Le problème qui se pose pour la reconnaissance d'appels de détresse est qu'il est extrêmement difficile, sinon impossible, de prévoir toutes les situations de détresse dans lesquelles pourraient se trouver les personnes âgées, et de déterminer quelles phrases seraient spontanément énoncées dans de telles situations. L'utilisation de gros modèles de langage, tels que les modèles utilisés pour reconnaître de la parole spontanée, permettrait une couverture exhaustive des phrases que pourrait prononcer une personne en détresse. Cependant, plus le modèle de langage est gros, plus le système a de chances de se tromper, et donc d'augmenter les erreurs de reconnaissance des mots. Un modèle de langage spécifique à la tâche permet de réduire le WER pour les phrases ciblées, mais la contrepartie est que la liste des phrases ciblées ne peut être exhaustive.

Nous avons donc pour objectif que notre système soit capable de reconnaître :

— des phrases de type appels aux aidants, structurées syntaxiquement de la façon suivante : « <mot-clé> <action> <contact> », avec <mot-clé> le nom du système, <action> l'action de lancer un appel téléphonique, et <contact> le nom ou la fonction du contact à appeler (par exemple : « e-lio appelle une infirmière »). Ces phrases auraient un rôle d' « interrupteur » pour lancer une conversation téléphonique à distance. L'utilisateur aurait donc à mémoriser la structure des phrases d'appels aux aidants, et un paramétrage permettrait d'entrer la liste des contacts (nom, fonction et numéro de téléphone),

— une liste non exhaustive de phrases de détresse, sans structure syntaxique particulière. Un paramétrage permettrait la mise à jour de la liste (par exemple pour la compléter avec de nouvelles phrases). L'objectif serait de lancer automatiquement un appel téléphonique à un contact référent lorsque ces phrases seraient reconnues et que les autres informations (visuelles, capteurs, historiques...) les conforteraient.

Pour tester notre système, nous avons utilisé le corpus *AD80* (voir section 5.2), qui contient une liste (non exhaustive) de phrases de détresse et d'appels aux aidants, mais également une liste de phrases « anodines » qui est un petit échantillon de phrases rencontrées dans la vie quotidienne (par exemple : « Dehors il fait beau. »). Nous avons réalisé un modèle de langage spécifique en utilisant l'ensemble de la liste des phrases que nous avions à notre disposition – la liste des phrases de détresse et d'appels aux aidants du corpus *AD80* (voir annexe C). Nous nous sommes ainsi mis dans le cas idéal où toutes les phrases de détresse et d'appels aux aidants décodées sont connues dans le modèle de langage, simulant ainsi l'exhaustivité du modèle de langage, afin d'avoir une référence vers laquelle un système exhaustif peut tendre.

Les phrases testées contenant également des phrases anodines, la question s'est posée de savoir comment prendre en compte ces phrases au niveau du modèle de langage. Avec un modèle trop spécifique, il existe le risque important qu'une phrase anodine, surtout si elle est courte, soit décodée comme la phrase de détresse la plus proche acoustiquement, créant ainsi un « faux positif ». Nous avons donc combiné notre modèle acoustique spécifique à un modèle générique appris sur le corpus *Gigaword*, contenant un nombre important de phrases extraites d'articles de journaux, en mettant un poids de 90% sur le modèle spéci-

fique. Ainsi, les phrases de détresse et d'appels aux aidants seront bien reconnues en tant que telles, le modèle de langage étant biaisé vers ce type de phrases ; les phrases anodines seront certes mal reconnues du fait de la taille importante du modèle générique mais auront moins de chance d'être reconnues en tant que phrases de détresse ou d'appels aux aidants du fait du large choix possible offert par le modèle de langage générique.

Le modèle de langage spécifique appris sur les phrases de détresse et d'appels aux aidants est un modèle trigramme. Nous avons choisi de créer le modèle générique appris sur le corpus *Gigaword* comme un modèle unigramme plutôt que trigramme afin de réduire le temps de décodage. Le modèle final combinant les deux modèles est un modèle trigramme (65077 unigrammes, 191 bigrammes et 225 trigrammes). C'est ce modèle que a été utilisé dans les sections précédentes.

6.4.2 Filtrage

Une fois le décodage réalisé par le système de RAP, nous avons complété notre système avec un filtre de détection de phrases de détresse et d'appels aux aidants. En effet, de nombreuses phrases peuvent être reconnues par le système de RAP, mais seules les deux catégories « phrases de détresse » et « appels aux aidants » doivent être prises en compte. Aussi, pour préserver la vie privée des utilisateurs, les phrases prononcées lors de discussions de la vie courante ne doivent pas être traitées par le système. Il est donc nécessaire de déterminer si les hypothèses de sortie du système de RAP appartiennent à une des 2 catégories « phrases de détresse » et « appels aux aidants », et dans le cas contraire les rejeter. Ainsi, notre filtre calcule une distance entre l'hypothèse de sortie et la liste déterminée de phrases appartenant aux 2 catégories « phrases de détresse » et « appels aux aidants ». Si cette distance est petite, l'hypothèse est traitée (le traitement étant le déclenchement d'un processus d'appel téléphonique) car considérée comme appartenant à une des catégories « phrases de détresse » et « appels aux aidants », sinon l'hypothèse n'est pas traitée.

La distance utilisée dans le filtre est la distance de Levenshtein (Levenshtein, 1966). La distance de Levenshtein est une métrique mesurant la différence entre 2 chaînes de caractères. Elle est définie comme étant le nombre minimum de modifications de caractères (insertions, délétions, substitutions) nécessaires pour changer une chaîne vers l'autre. Le filtre calcule la distance entre l'hypothèse de sortie du système de RAP et chaque phrase de la liste de phrases (appels de détresse et appels aux aidants) à reconnaître. La phrase de la liste avec la meilleure distance de Levenshtein par rapport à l'hypothèse est alors sélectionnée. Plus la distance est petite (score de 0 à l'infini), meilleure est la correspondance entre les 2 phrases. Pour ne pas être biaisé par l'orthographe (par exemple « appelez » décodé en « appeler »), la distance est calculée sur des phrases transformées en suites phonétiques (par exemple « aa pp ee ll ei ll ee ss aa mm uu » pour « Appelez le SAMU ! »). Cette approche tient compte de certaines erreurs de reconnaissance telle que les légères variations, les erreurs d'accord, etc. En outre, dans de nombreux cas, un mot mal décodé est phonétiquement proche du mot correct.

Puis le filtre normalise la distance de Levenshtein par le nombre de phonèmes de la phrase sélectionnée dans le filtre. Un seuil est appliqué sur la distance normalisée : si la distance normalisé est au-dessous du seuil, la phrase est traitée car classifiée comme appartenant à la catégorie « phrases de détresse » ou « appels aux aidants », sinon, elle est classifiée comme appartenant à la catégorie « phrases anodine » et elle est rejetée.

La valeur du seuil est choisie grâce à une courbe ROC. La courbe ROC est construite à partir du taux de vrais positifs (TVP) en fonction du taux de faux positifs (TFP) en faisant varier la valeur du seuil, ou de manière équivalente à partir de la sensibilité (Se) et de la spécificité (Sp) par $Se = f(1 - Sp)$. Le seuil optimal est donné par la valeur du seuil à l'intersection de la courbe avec la diagonale $f(x) = -x + 1$ (voir figure 6.8). Si une phrase de détresse ou d'appel aux aidants a une distance normalisée située sous le seuil, elle sera considérée comme un vrai positif (VP). Une phrase de détresse ou d'appels aux aidants avec une distance au-dessus du seuil sera un faux négatif (FN). Enfin, une phrase anodine sera un faux positif (FP) si au-dessous du seuil, et un vrai négatif (VN) si au-dessus du seuil. Les définitions de la sensibilité, de la spécificité, du TVP et du TFP sont données en annexe G.0.2.

6.4.3 Évaluation et résultats

Nous avons réalisé le décodage et le filtrage sur 2 groupes :

— *locuteurs jeunes* : 52 locuteurs jeunes issus du corpus *AD80*, soit 3306 phrases de détresse, 961 appels aux aidants et 3897 phrases anodines (2 heures et 34 minutes d'enregistrement),

— *locuteurs âgés* : 43 locuteurs âgés issus du corpus *AD80*, soit 2663 phrases de détresse, 434 appels aux aidants et 3006 phrases anodines (2 heures et 15 minutes d'enregistrement).

Les WER obtenus avec *Sphinx3*, à partir du modèle générique *BREF120* pour les locuteurs jeunes, et à partir du modèle adapté à la voix âgée *BREF120_MLLR_G* pour les locuteurs âgés pour les phrases de détresse/appels aux aidants et pour les phrases anodines, sont présentés table 6.10.

Les courbes ROC pour les 2 groupes sont données figure 6.8. Pour les seuils donnés par l'intersection des courbes ROC avec leur diagonale (seuil de 0,812 pour les locuteurs jeunes et de 0,75 pour les locuteurs âgés), les matrices de confusion des tests positifs et négatifs sont données table 6.11.

	Détresse/appels aux aidants	Anodines
Locuteurs jeunes	11,0%	67,9%
Locuteurs âgés	17,4%	37,8%

TABLE 6.10: *WER moyens pour les phrases de détresse/appels aux aidants et les phrases anodines en fonction des groupes « locuteurs jeunes » (modèle acoustique BREF120) et « locuteurs âgés » (modèle acoustique BREF120_MLLR_G).*

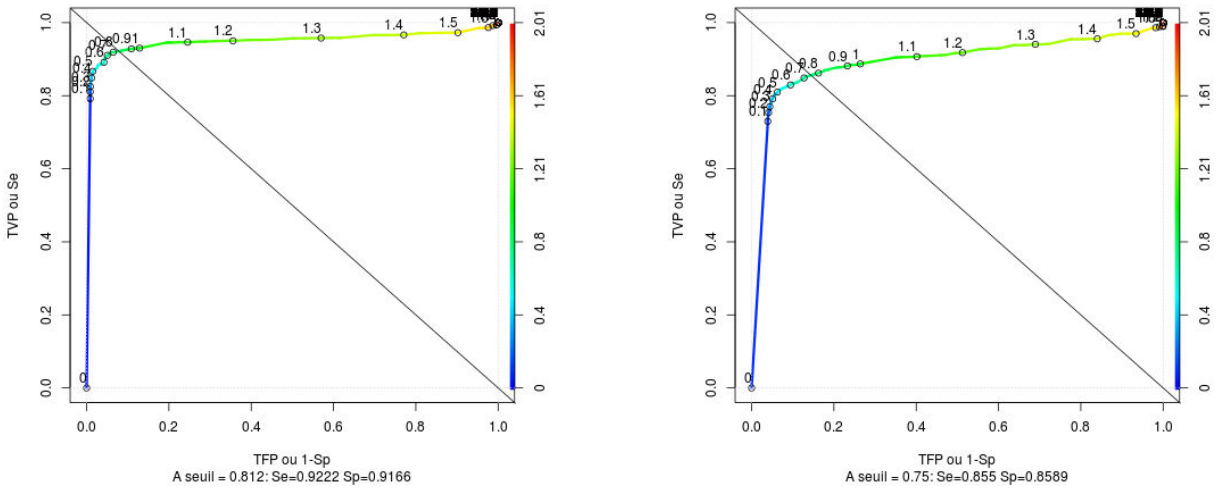
seuil = 0,812	d < seuil	d >= seuil
Phrases cibles	$VP = 3935$	$FN = 332$
Perturbateurs	$FP = 325$	$VN = 3572$

(a) Locuteurs jeunes

seuil = 0,75	d < seuil	d >= seuil
Phrases cibles	$VP = 2648$	$FN = 449$
Perturbateurs	$FP = 424$	$VN = 2582$

(b) Locuteurs âgés

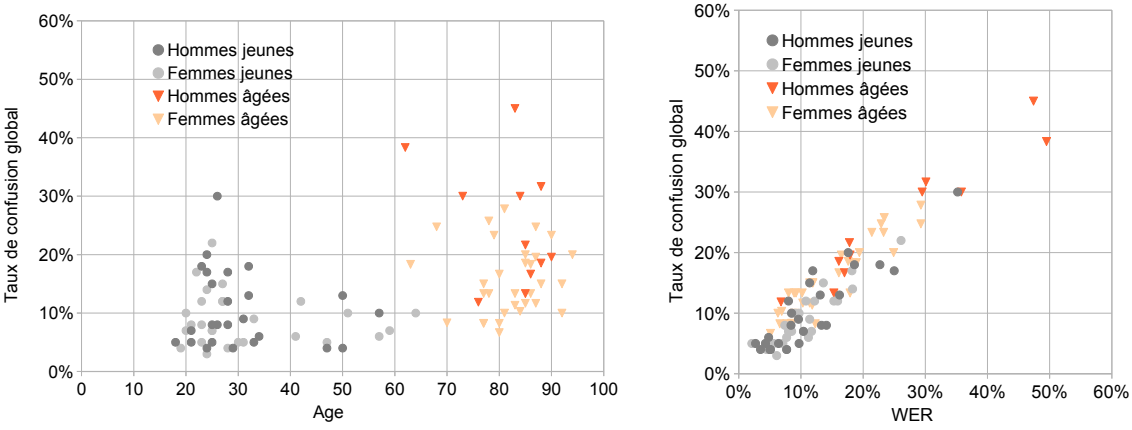
TABLE 6.11: *Matrices de confusion du filtrage des phrases de détresse/appels aux aidants pour les locuteurs jeunes et les locuteurs âgés.*



(a) Locuteurs jeunes

(b) Locuteurs âgés

FIGURE 6.8: *Courbes ROC représentant TVP en fonction de TFP pour le filtrage des appels de détresse et des appels aux aidants pour les locuteurs jeunes et les locuteurs âgés.*



(a) En fonction de l'âge

(b) En fonction du WER

FIGURE 6.9: *Taux de confusion global pour chaque locuteur des différents groupes, en fonction de l'âge et en fonction du WER.*

Pour les locuteurs jeunes, au seuil égal à 0,812, la sensibilité (Se) et la spécificité (Sp) sont les suivantes : Se = 92,22% et Sp = 91,66%. Aussi, nous avons trouvé un rappel, précision et F-mesure dont les valeurs sont respectivement égales à 92,22%, 92,37% et 92,30% (voir les définitions du rappel, précision et F-mesure en annexe G.0.2). La sensibilité nous indique que parmi les phrases de détresse et d'appels aux aidants prononcées, 92,22% ont été reconnues comme appartenant à la catégorie « détresse/appels aux aidants » et ont donc été traitées (distance au-dessous du seuil). La spécificité montre que parmi les phrases anodines prononcées, 91,66% ont été rejetées par le système car reconnues comme appartenant à la catégorie « anodine » (distance au-dessus du seuil). Les 7,63% de phrases anodines reconnues comme « détresse/appels aux aidants » (distance au-dessous du seuil) donnent lieu à des fausses alarmes.

Pour les locuteurs âgés, au seuil égal à 0,75, la sensibilité est de 85,50%, et la spécificité est de 85,89%. Le rappel, la précision et la F-mesure ont des valeurs respectivement égales à 85,50%, 86,20% et 85,85%. Enfin, le taux de fausses alarmes est de 13,80%.

Cependant, une phrase de détresse ou d'appel aux aidants peut être correctement classifiée dans la catégorie « détresse/appels aux aidants » sans que pour autant la phrase sélectionnée par le filtre ne soit la bonne. Il peut y avoir des confusions entre les différentes phrases de détresse et d'appels aux aidants à cause d'un mauvais décodage par le système de RAP. Il y aura confusion lorsque par exemple la phrase prononcée aura été « e-lio appelle ma fille », que le filtre l'aura bien classifié dans la catégorie « détresse/appels aux aidants » (distance de levenshtein normalisée inférieure au seuil), mais que la phrase sélectionnée aura été « e-lio appelle les pompiers ».

Ainsi, parmi les phrases de détresse et d'appels aux aidants correctement classifiées dans la catégorie « détresse/appels aux aidants », nous avons calculé que 1,45% (locuteurs jeunes) et 4,04% (locuteurs âgés) des phrases de détresse/appels aux aidants ont été sélectionnées avec confusion par le filtre (ce qui a été prononcé ne correspond pas à ce qui a été sélectionné par le filtre). Parmi l'ensemble des phrases de la catégorie « détresse/appels aux aidants », 8,97%(locuteurs jeunes) et 17,29% (locuteurs âgés) ont été traitées avec confusion. De façon globale, 9,12% (locuteurs jeunes) et 17,95% (locuteurs âgés) des cas d'appels aux aidants ou phrases de détresse ont été reconnues incorrectement par le système (taux de confusion globale).

La figure 6.9 représente le taux de confusion global pour chaque locuteur en fonction de l'âge et du WER, et nous pouvons ainsi observer que ce taux est peu corrélé à l'âge et est fortement corrélé au WER.

6.4.4 Bilan

Dans ce chapitre, nous avons comparé l'utilisation de différentes données pour l'adaptation des modèles acoustiques à la voix âgée. Dans un premier temps, nous avons adapté le modèle générique *BREF120* à la voix âgée en général en utilisant la voix de locuteurs âgés comme données d'adaptation, soit au total 48 minutes de lectures par 22 locuteurs du corpus

ERES38, pour obtenir le modèle *BREF120_MLLR_G*. Nous avons testé ce modèle adapté de façon globale à la voix âgée sur les locuteurs du corpus *AD80*, et observé une réduction significative du WER par rapport à l'utilisation du modèle générique. En réalisant l'adaptation en fonction du genre masculin ou féminin des locuteurs du corpus *ERES38*, les modèles obtenus, *BREF120_MLLR_M* et *BREF120_MLLR_F*, n'ont pas donné de résultats significativement différents des résultats obtenus avec le modèle adapté de façon global *BREF120_MLLR_G*. De plus, l'utilisation des modèles adaptés à la voix âgée n'a pas d'impact significatif sur le WER des locuteurs jeunes.

Nous avons également réalisé une adaptation au locuteur à partir de lectures prononcées par chaque locuteur testé. Les résultats montrent qu'il n'y a pas de différence significative entre les WER obtenus suite à une adaptation au locuteur et ceux obtenus suite à une adaptation globale à la voix âgée. D'un point de vu applicatif, il est donc intéressant d'utiliser un modèle adapté de façon globale à la voix âgée, les deux genres confondus, ceci évitant à l'utilisateur l'étape fastidieuse de lecture de phrases pour l'adaptation à sa voix.

Nous avons ensuite présenté notre évaluation de la recherche de phrases cibles parmi les hypothèses de reconnaissance. Cette recherche est réalisée par un filtre qui calcule au niveau phonétique la distance de Levenshtein entre chaque hypothèse de sortie du système de RAP et chacun des éléments de la liste des phrases cibles.

Dans le chapitre suivant, nous chercherons à déterminer quels sont les facteurs pouvant expliquer les dégradations des performances des systèmes de RAP avec la voix âgée. En effet, une grande disparité des performances du système de RAP est observée en fonction des sujets âgés. Comprendre les facteurs qui affectent ces performances permettrait d'anticiper si la mise en place d'un tel système est envisageable pour un individu spécifique, le taux de confusion global des phrases étant corrélé au WER.

CHAPITRE 7

Les facteurs explicatifs des dégradations de performances pour la voix âgée

Dans ce chapitre, nous allons chercher quels sont les facteurs pouvant influencer la dégradation des performances des systèmes de RAP avec la voix âgée. Nous allons étudier l'influence de l'âge, de la dépendance évaluée grâce à la grille AGGIR, des caractéristiques phonétiques et des paramètres prosodiques étendus à la qualité de voix.

7.1 Influence de l'âge sur le WER

Nous avons vu qu'il existe une différence de performance des systèmes de RAP entre le groupe *locuteurs jeunes* et le groupe *locuteurs âgées*. En comparant les figures 6.4 (WER en fonction de l'âge avec le modèle *BREF120*) et 6.6 (WER en fonction de l'âge avec le modèle *BREF120* pour les voix jeunes et le modèle adapté *BREF120_MLLR_G* pour les voix âgées), nous observons une droite de régression linéaire montante entre les deux groupes. En revanche, à l'intérieur des groupes, il semble ne pas y avoir de relation linéaire.

Pour vérifier ces observations, nous avons donc calculé les corrélations de Pearson (corrélation linéaire) entre l'âge et les WER obtenus avec les modèles *BREF120* et *BREF120_MLLR_G* pour les groupes *locuteurs jeunes*, *locuteurs âgées* et *tous les locuteurs*.

Nous observons dans la table 7.1 que les corrélations entre l'âge des locuteurs et les WER à l'intérieur du groupe *locuteurs jeunes* et du groupe *locuteurs âgés* sont proches de 0 et non significatives (p-values largement supérieures à 0,05%). En revanche, si nous considérons le groupe *tous les locuteurs*, les corrélations entre âge et WER sont significatives : la corrélation est forte lors de l'utilisation du modèle *BREF120* car ce modèle n'est pas adapté à la voix âgée, donc le WER moyen est plus important pour les voix âgées, et la corrélation est modérée lorsque l'on applique le modèle *BREF120_MLLR_G* pour les voix âgées car celui-ci permet

Groupe	BREF120	BREF120_MLLR_G	BREF120(jeunes)+BREF120_MLLR_G(âgés)
Locuteurs jeunes	-0,081(p=0,568)	–	–
Locuteurs âgées	-0,116(p=0,459)	-0,206(p=0,184)	–
Tous les locuteurs	0,747(p<0,05)	–	0,295(p<0,05)

TABLE 7.1: *Corrélations de Pearson entre l'âge et les WER moyens obtenus avec les modèles BREF120 et BREF120_MLLR_G (si p<0,05, on rejette l'hypothèse H0 qu'il n'y a pas de corrélation entre les deux variables).*

de réduire son WER moyen en s'approchant du WER moyen obtenu pour les voix jeunes avec le modèle *BREF120*. Le signe positif des corrélations montre que les variables évoluent dans le même sens : le WER augmente lorsque l'âge augmente.

7.2 Influence de la dépendance sur le WER

Nous avons vu dans la section 3.1 que les caractéristiques de la voix changent avec le vieillissement. Par conséquent, les performances de la RAP baissent pour les personnes âgées, sans que l'adaptation des modèles acoustiques ne rattrape totalement l'écart avec les performances obtenues pour les voix jeunes.

Nous observons des écarts importants entre les WER des différents locuteurs âgés, et cela même entre les personnes âgées du même âge (voir figure 6.4). Nous avons vu que l'âge à l'intérieur du groupe *locuteurs âgés* n'est pas un facteur corrélé avec le WER. Par conséquent, nous avons étudié si un autre critère que l'âge pouvait expliquer cette disparité. Le critère du niveau de dépendance des personnes âgées nous a semblé être un critère pouvant être un bon indicateur pour expliquer cette variabilité du WER. En effet, les personnes ne vieillissent pas toutes de la même façon, certaines peuvent vieillir moins vite et garder une voix « jeune » : cette baisse des performances des systèmes de RAP peut donc être très variable selon les personnes âgées. Comme référence, nous avons utilisé la grille AGGIR (Autonomie Gérontologie Groupes Iso-Ressources) afin de déterminer le niveau de dépendance des personnes âgées.

7.2.1 Grille AGGIR

La grille AGGIR [1] est un test national utilisé pour évaluer le degré de perte d'autonomie physique ou psychique d'une personne âgée ou handicapée. Elle sert de support pour déterminer le montant de l'aide financière pouvant lui être versé, l'APA (Allocation Personnalisée d'Autonomie). L'évaluation est effectuée en utilisant 17 variables.

10 variables font référence à la perte d'autonomie physique et cognitive :

— cohérence : converser et se comporter de façon sensée,

— orientation : se repérer dans le temps, les moments de la journée et les lieux,

— toilette du haut du corps (rasage, coiffage, tronc, membres supérieurs et main) et du bas du corps (régions intimes, membres inférieurs, pieds),

— habillage du haut (passer des vêtements par la tête et les bras), moyen (fermeture éclair, boutons, ceinture ou bretelles), bas (passage de vêtements par le bas, chaussettes, bas et chaussures),

— alimentation : se servir (couper la viande, ouvrir un pot, peler un fruit, remplir un verre), manger (porter les aliments et les boissons à la bouche et les avaler),

— élimination : assurer l'hygiène de l'élimination urinaire et fécale,

1. http ://vosdroits.service-public.fr/particuliers/F1229.xhtml

— transferts : se lever, se coucher, s'asseoir,

— déplacements à l'intérieur avec ou sans canne, déambulateur, fauteuil roulant,

— déplacements à l'extérieur : à partir de la porte d'entrée sans moyen de transport,

— communication à distance : utiliser les moyens de communication, téléphone, sonnette, alarme.

7 variables se rapportent à la perte d'autonomie domestique et sociale :

— gestion : gérer ses propres affaires, son budget, ses biens,

— cuisine : préparer des repas et les conditionner pour être servis,

— ménage : effectuer l'ensemble des travaux ménagers courants,

— transport : prendre ou commander un moyen de transport collectif ou individuel,

— achats : acquisition directe ou par correspondance,

— suivi du traitement : se conformer à l'ordonnance du médecin,

— activités de temps libre : activités sportives, culturelles, sociales, de loisir ou de passe-temps.

Chaque variable est codée en fonction du degré de dépendance :

— A : fait seul, totalement, habituellement, correctement,

— B : fait partiellement, non habituellement, non correctement,

— C : ne fait pas.

L'évaluation se fait régulièrement par un médecin ou une équipe médico-sociale. En fonction de son degré de dépendance, la personne âgée est rattachée à l'un des Groupes Iso-Ressources (GIR). Il existe 6 GIR :

— GIR 1 : personne confinée au lit ou au fauteuil, dont les fonctions mentales sont gravement altérées et qui nécessite une présence indispensable et continue d'intervenants. Ou personne en fin de vie,

— GIR 2 : personne confinée au lit ou au fauteuil, dont les fonctions mentales ne sont pas totalement altérées et dont l'état exige une prise en charge pour la plupart des activités de la vie courante. Ou personne dont les fonctions mentales sont altérées, mais qui est capable de se déplacer et qui nécessite une surveillance permanente,

— GIR 3 : personne ayant conservé son autonomie mentale, partiellement son autonomie locomotrice, mais qui a besoin quotidiennement et plusieurs fois par jour d'une aide pour les soins corporels,

— GIR 4 : personne n'assumant pas seule ses transferts mais qui, une fois levée, peut se déplacer à l'intérieur de son logement, et qui a besoin d'aides pour la toilette et l'habillage. Ou personne n'ayant pas de problèmes locomoteurs mais qui doit être aidée pour les soins corporels et les repas,

— GIR 5 : personne ayant seulement besoin d'une aide ponctuelle pour la toilette, la préparation des repas et le ménage,

— GIR 6 : personne encore autonome pour les actes essentiels de la vie courante.

Seuls les GIR 1 à 4 donnent droit à l'APA. Les personnes relevant des GIR 5 ou 6 peuvent demander une aide ménagère.

7.2.2 Relations entre dépendance et WER

Pour chacune des 43 personnes âgées du corpus *AD80*, nous avons fait remplir une grille AGGIR par le personnel paramédical des établissements visités. Nous avons ainsi pu caractériser chaque locuteur âgé par un score GIR. Quatre locuteurs étaient *GIR 2*, deux locuteurs étaient *GIR 3*, vingt et un locuteurs étaient *GIR 4*, un locuteur était *GIR 5* et quinze locuteurs étaient *GIR 6*. Aucun locuteur n'était représenté dans *GIR 1*. Du fait du faible nombre de locuteurs *GIR 2*, *GIR 3* et *GIR 5*, nous avons fusionné les groupes *GIR 2* avec *GIR 3*, et *GIR 5* avec *GIR 6*, soit six locuteurs dans le groupe *GIR 2-3*, vingt-deux locuteurs dans le groupe *GIR 4-5* et quinze locuteurs dans le groupe *GIR 6*.

Nous avons réalisé un décodage sur les phrases de détresse et d'appels aux aidants des locuteurs âgés du corpus *AD80* avec le modèle acoustique adapté à la voix âgée *BREF120_MLLR_G* et le modèle de langage appris sur les phrases de détresse et d'appels aux aidants combiné au modèle appris sur le corpus *Gigaword*.

Les moyennes et écarts-types des WER pour les groupes *GIR 2-3*, *GIR 4-5* et *GIR 6* sont présentés table 7.2.

Nous observons que les moyennes et écarts-type du WER augmentent avec la diminution de l'autonomie.

Nous avons réalisé une ANOVA entre les 3 échantillons correspondant aux groupes *GIR 2-3*, *GIR 4-5* et *GIR 6*. Les 3 échantillons suivaient une loi normale et vérifiaient l'homogénéité des variances. L'ANOVA montre que le score GIR a un effet significatif sur le WER, avec comme résultat $F(2;40)=5{,}164$; $p=0{,}0101$. Un test post-hoc de Tukey HSD nous a permis de caractériser quels groupes sont significativement différents de quels autres groupes. Le test a montré qu'il y a une différence significative entre les groupes *GIR 2-3* et *GIR 4-5*, ainsi qu'entre *GIR 2-3* et *GIR 6*, et qu'il n'y a pas de différence significative entre les groupes *GIR 4-5* et *GIR 6*.

Ainsi, au sein des locuteurs âgés, le WER n'est pas directement corrélé avec l'âge mais a une relation avec le niveau de dépendance. En effet, la corrélation de Pearson entre les scores GIR et les WER pour les personnes âgées est de -0,318 ($p=0{,}038$) avec le modèle *BREF120_MLLR_G*. Cette corrélation est modérée, et statistiquement significative ($p<0{,}05$),

Groupe	Moyenne	Écart-type
GIR 6	14,58%	7,56%
GIR 4-5	16,32%	7,93%
GIR 2-3	28,68%	16,72%
Tous	17,44%	10,27%

TABLE 7.2: *Moyennes et écarts-types des WER en fonction des GIR pour les personnes âgées.*

le signe négatif signifiant que les variables évoluent dans le sens opposé : le WER augmente lorsque le score GIR diminue.

Ainsi, le WER est corrélé dans une certaine mesure avec le niveau de dépendance, car le niveau de dépendance reflète le niveau réel de vieillissement physique et psychique de la personne, ce vieillissement influençant la production de la parole.

Dans la section suivante, nous allons analyser quelles sont les caractéristiques phonétiques permettant de différencier la voix âgée de la voix jeune, et quelles sont leurs corrélations avec le WER.

7.3 Étude phonétique

7.3.1 Comparaison des scores d'alignement forcé entre voix jeunes et voix âgées

Les différences dans les scores d'alignement forcé pour chaque catégorie de phonèmes entre les voix âgées et les voix jeunes permettent d'identifier les phonèmes qui sont les plus difficiles à reconnaître par le système de RAP pour les voix âgées.

L'alignement forcé consiste à trouver les limites des phonèmes (les segments) dans un énoncé en connaissant la phrase prononcée. Durant l'alignement forcé, les phrases sont phonétisées en utilisant un dictionnaire pour contraindre un alignement optimal entre le modèle acoustique et l'énoncé de parole. Pour chaque segment de la phrase, les scores d'alignement forcé sont calculés, et correspondent à la probabilité d'appartenance du segment à un modèle de phonème. Ce score peut être interprété comme une proximité à la prononciation « standard » définit par le modèle acoustique. Le score est une probabilité à l'échelle logarithmique, les valeurs étant comprises entre -∞ (faible proximité avec le modèle) et la limite théorique de 0 (forte proximité).

Pour cette étude, nous avons utilisé l'algorithme d'alignement forcé de Sphinx3. Les vecteurs acoustiques utilisés étaient composés de 13 coefficients MFCC, et du delta et delta-delta de chaque coefficient. Le modèle acoustique utilisé est le modèle acoustique *BREF120* décrit section 6.2.1.

Les scores d'alignement forcé sur les groupes *locuteurs âgés* et *locuteurs jeunes* (corpus *AD80*) sont présentés figure 7.1 selon les catégories phonémiques. Ces dernières sont présentées table 7.3.

Pour les voix jeunes, les 3 catégories de phonèmes avec les meilleurs scores sont les fricatives sourdes (f, s, ʃ), les consonnes nasales (m, n, ŋ) et les voyelles mi-ouvertes (ɛ, œ, ɔ).

Les 3 catégories de phonèmes avec les scores les plus bas chez les personnes jeunes sont les plosives sourdes (p, t, k), les consonnes liquides (l) et les voyelles ouvertes (a, ɑ).

Pour les voix âgées, nous pouvons observer que chaque catégorie de phonèmes a un score d'alignement forcé plus faible que pour les voix jeunes.

Les 3 catégories de phonèmes avec les scores les plus hauts pour les voix âgées sont les fricatives sourdes (f, s, ʃ), les consonnes nasales (m, n, ŋ) et les voyelles fermées (i, y, u).

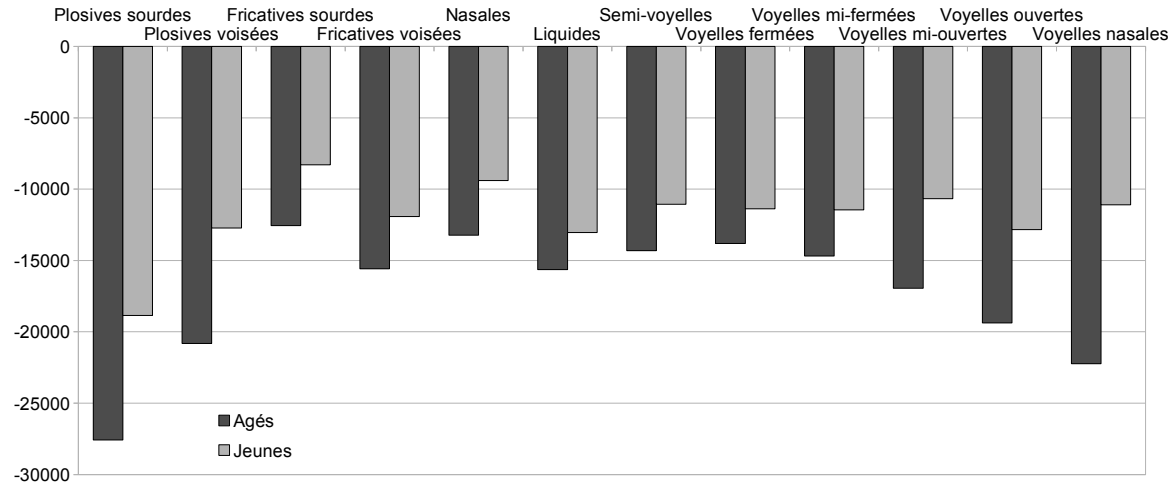


FIGURE 7.1: *Scores d'alignement forcé en fonction des catégories de phonèmes pour les voix jeunes et âgées.*

Les catégories de phonèmes avec les scores les plus bas pour les voix âgées sont les plosives sourdes (p, t, k), les voyelles nasales (ɛ̃, ɑ̃, ɔ̃, œ̃) et les plosives voisée (b, d, g).

Pour évaluer quelles catégories de phonèmes sont les plus dégradées chez les personnes âgées par rapport aux personnes jeunes, nous avons calculé les différences relatives (voir la définition en annexe G.0.3) entre les scores d'alignement forcé.

Les différences relatives des scores d'alignement forcé entre les voix âgées et les voix jeunes ont été classées par catégorie de phonèmes dans l'ordre décroissant dans la table 7.4.

Pour les consonnes, les scores des plosives voisées, des fricatives sourdes et des plosives sourdes subissent plus de dégradation que ceux des consonnes nasales, des fricatives voisées, des semi-voyelles et des consonnes liquides.

Pour les voyelles, les voyelles nasales, les voyelles mi-ouvertes et les voyelles ouvertes subissent plus de dégradation que les voyelles mi-fermées et les voyelles fermées.

Ainsi, les consonnes demandant lors de leur production une pression intra-orale plus importante (Lisker, 1970; Arkebauer et coll., 1967) sont celles qui sont le plus dégradées chez les voix âgées par rapport aux voix jeunes. En effet, les personnes âgées utilisent un plus

Catégorie	Symboles
Plosives sourdes	p, t, k
Plosives voisées	b, d, g
Fricatives sourdes	f, s, ʃ
Fricatives voisées	v, z, ʒ, ʁ
Nasales	m, n, ŋ
Liquides	l
Semi-voyelles	ɥ, j, w
Voyelles fermées	i, y, u
Voyelles mi-fermées	e, ø, o, ə
Voyelles mi-ouvertes	ɛ, œ, ɔ
Voyelles ouvertes	a, ɑ
Voyelles nasales	ɛ̃, ɑ̃, ɔ̃, œ̃

TABLE 7.3: *Catégories de phonèmes (symboles IPA).*

Catégorie	Score
Plosives voisées	-63,56
Fricatives sourdes	-51,22
Plosives sourdes	-46,34
Nasales	-40,71
Fricatives voisées	-30,77
Semi-voyelles	-29,25
Liquides	-19,86

(a) Consonnes

Catégorie	Score
Voyelles nasales	-100,3
Voyelles mi-ouvertes	-58,89
Voyelles ouvertes	-50,93
Voyelles mi-fermées	-28,3
Voyelles fermées	-21,37

(b) Voyelles

TABLE 7.4: *Différences relatives des scores d'alignement forcé entre voix jeunes et âgées pour les consonnes et les voyelles.*

grand pourcentage de leur volume pulmonaire sur chaque syllabe (Lass, 2012; Hoit et Hixon, 1987), et la fermeture déficiente de la glotte augmente la quantité d'air utilisé pour la phonation (Linville et Rens, 2001) : la pression intra-orale diminue rapidement et perturbe donc la production de certaines consonnes.

7.3.2 Relations entre scores d'alignement forcé et WER

Nous cherchons maintenant à déterminer la corrélation entre les WER obtenus à partir des décodages avec le modèle acoustique *BREF120* et les scores d'alignement forcé pour l'ensemble des locuteurs jeunes et âgés. Les résultats sont présentés table 7.5.

Nous observons table 7.5 que les scores d'alignement forcé sont fortement corrélés avec les WER (corrélations négatives), et que les corrélations sont significatives pour toutes les

Catégorie	Corrélation	p-value
Plosives sourdes	-0,653	7,796e-13
Plosives voisées	-0,769	2,2e-16
Fricatives sourdes	-0,680	3,444e-14
Fricatives voisées	-0,559	3,895e-09
Nasales	-0,682	2,781e-14
Liquides	-0,595	2,043e-10
Semi-voyelles	-0,638	3,645e-12
Voyelles fermées	-0,600	1,336e-10
Voyelles mi-fermées	-0,707	1,119e-15
Voyelles mi-ouvertes	-0,775	2,2e-16
Voyelles ouvertes	-0,600	1,322e-10
Voyelles nasales	-0,788	2,2e-16
Ensemble des phonèmes	-0,810	2,2e-16

TABLE 7.5: *Corrélation de Pearson entre les scores d'alignement forcé et les WER, locuteurs jeunes et âgés confondus (corrélations significatives si p-value<0,05).*

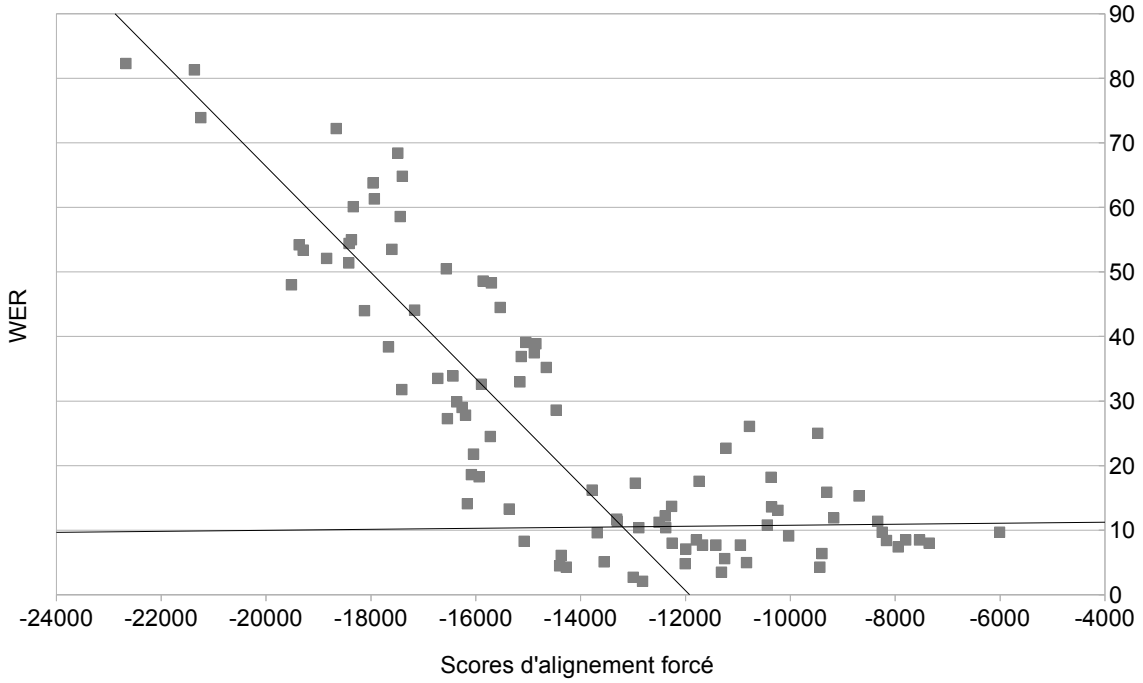


FIGURE 7.2: *WER en fonction des scores d'alignement forcé (locuteurs jeunes et âgés confondus)*

catégories phonémiques (p-values<0,05). Toutes catégories phonémiques confondues, la moyenne des scores d'alignement forcé est corrélée au WER avec un score de -0,81. Les voyelles nasales sont les phonèmes avec la corrélation la plus élevée (-0,788), suivies des voyelles mi-ouvertes (-0,775), des plosives voisées (-0,769) et des voyelles mi-fermées (-0,707).

La figure 7.2 représente le WER en fonction des scores d'alignement forcé pour chaque locuteur. Nous observons que dans les hautes valeurs de scores d'alignement forcé, jusqu'à un score de valeur -14000 (partie droite du nuage de points), le WER reste autour d'une droite de régression quasiment horizontale, autour d'une valeur de WER de 10%. Pour les valeurs basses du score d'alignement forcé, en deçà de -14000 (partie gauche du nuage de points), le WER augmente autour d'une droite de régression de pente relativement importante.

Étant donné ces résultats, nous allons chercher dans la section suivante si les scores d'alignement forcé permettent de prédire le WER en utilisant un classifieur.

7.3.3 Prédiction du WER à partir des scores d'alignement forcé

Nous avons entraîné un classifieur à régression logistique, à l'aide de l'outil *Weka*[2], dans le but de prédire à quelle classe de WER appartiennent les locuteurs.

L'apprentissage et le test ont été réalisés par validation croisée par la méthode du *k-fold*. L'échantillon de n observations est divisé k fois, un des k sous-échantillons constitue l'ensemble de validation, et les k-1 autres sous-échantillons constituent l'ensemble d'apprentissage. L'opération est répétée en sélectionnant un autre sous-échantillon de validation parmi les k-1 sous-échantillons qui n'ont pas encore été utilisés pour la validation du

2. http ://www.cs.waikato.ac.nz

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,907	0,171	0,875	0,907	0,891	0,850	WER=[0-26[%
	0,829	0,093	0,872	0,829	0,850	0,850	WER=[26-100]%
Moyenne pondérée :	0,874	0,137	0,874	0,874	0,873	0,850	

TABLE 7.6: *Validation croisée sur les classes WER=[0-26[% et WER=[26-100]% (modèle BREF120).*

WER=[0-26[%	WER=[26-100]%	<– classifié comme
49	5	WER=[0-26[%
7	34	WER=[26-100]%

TABLE 7.7: *Matrice de confusion pour la validation croisée sur les classes WER=[0-25[% et WER=[25-100]% (modèle BREF120).*

modèle, l'opération se répétant k fois jusqu'à ce que chaque sous-échantillon ait été utilisé une fois comme ensemble de validation.

Nous appliquons cette méthode de la manière suivante : les 95 locuteurs du corpus *AD80* représentent les 95 observations (n=95), et nous prenons comme valeur de k : k=n=95. Ainsi, un locuteur est retiré du groupe de locuteurs et constitue le locuteur de test, tandis que les 94 autres locuteurs servent à l'apprentissage. L'opération est répétée 95 fois jusqu'à ce que tous les locuteurs aient été testés.

Nous avons d'abord considéré 2 classes : la classe *WER=[0-26[%* (bon WER) et la classe *WER=[26-100]%* (mauvais WER). Les résultats de la validation croisée sont donnés tables 7.6 et 7.7.

Nous obtenons les scores suivants : précision=0,884, rappel=0,884 et F-mesure=0,884 (moyenne pondérée). La prédiction du WER avec un classifieur grâce aux scores d'alignement forcé par catégorie phonémique est donc relativement bonne si nous considérons une répartition en 2 classes de WER.

Nous avons ensuite augmenté le nombre de classes pour affiner notre analyse. En considérant les 4 classes *WER=[0-13[%*, *WER=[13-26[%*, *WER=[26-39[%* et *WER=[29-100[%*, nous obtenons les résultats présentés tables 7.8 et 7.9.

Les scores obtenus ne sont pas très bons, avec précision=0,632, rappel=0,632 et F-mesure=0,630 (moyenne pondérée). Les scores d'alignement forcé ne permettent globalement pas une prédiction très précise du WER avec un classifieur. Cependant, pour les classes

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,730	0,207	0,692	0,730	0,711	0,859	WER=[0-13[%
	0,353	0,103	0,429	0,353	0,387	0,747	WER=[13-26[%
	0,563	0,127	0,474	0,563	0,514	0,859	WER=[26-39[%
	0,720	0,071	0,783	0,720	0,750	0,941	WER=[39-100]%
Moyenne pondérée :	0,632	0,139	0,632	0,632	0,630	0,861	

TABLE 7.8: *Validation croisée sur les classes WER=[0-13[%, WER=[13-26[%, WER=[26-39[% et WER=[39-100]% (modèle BREF120).*

WER=[0-13[%	WER=[13-26[%	WER=[26-39[%	WER=[39-100]%	<– classifié comme
27	7	3	0	WER=[0-13[%
8	6	2	1	WER=[13-26[%
2	1	9	1	WER=[26-39[%
2	0	5	18	WER=[39-100]%

TABLE 7.9: *Matrice de confusion pour la validation croisée sur les classes WER=[0-13[%, WER=[13-26[%, WER=[26-39[% et WER=[39-100]% (modèle BREF120).*

extrêmes *WER=[0-13[%* et *WER=[39-100]%*, nous observons que celles-ci sont mieux reconnues que les classes intermédiaires, avec précision=0,692, rappel=0,730 et F-mesure=0,711 pour la classe *WER=[0-13[%*, et précision=0,783, rappel=0,720 et F-mesure=0,750 pour la classe *WER=[39-100[%*.

7.4 Étude prosodique

Pour mettre en évidence les différences prosodiques existant entre les voix âgées et les voix jeunes, nous avons mesuré pour les 95 locuteurs du corpus *AD80* différents paramètres vocaux. Les paramètres mesurés, qui sont des paramètres prosodiques et des paramètres de la qualité de la voix, sont les suivants :

— débit moyen : mesure le rapport entre le nombre de phonèmes et la durée du signal,

— fréquence fondamentale (F0) moyenne : mesure globalement la hauteur de la voix (aiguë, grave ...),

— jitter moyen : mesure l'instabilité à court terme de la F0, se traduisant par des variations de fréquence entre chaque cycle d'oscillation, et se calcule par la moyenne de la différence de F0 entre deux cycles d'oscillation consécutifs,

— shimmer moyen : mesure l'instabilité à court terme de l'amplitude de l'oscillation, se traduisant par des variations d'amplitude entre chaque cycle d'oscillation. Il se calcule par la moyenne des rapports d'amplitudes entre deux cycles d'oscillation consécutifs.

— rapport harmonique sur bruit moyen (HNR ou *Harmonic to Noise Ratio*) : mesure en dB le rapport entre la partie périodique du signal et la partie non périodique. Les voix rauques ont un HNR bas, et les voix aiguës un HNR élevé.

Par la suite, ces 5 variables seront notées *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR*.

La variable *Débit* a été calculée à partir du résultat de l'alignement forcé (alignement phonémique) obtenu avec *Sphinx3*. Les variables *F0*, *Jitter*, *Shimmer* et *HNR* ont été calculées à l'aide du logiciel Praat (Boersma, 2002), et ont été mesurées pour chacun des 95 locuteurs sur une fenêtre de durée de 2 minutes en moyenne, correspondant, pour chaque locuteur, à l'ensemble des phrases de détresse et d'appels aux aidants. Nous avons ainsi 475 mesures (95 locuteurs x 5 variables).

Paramètres	Voix jeune	Voix âgée	Test t de Welch
Débit (Phonèmes/s)	12,38943	10,56711	p<0,05 (p=3,52e-08)
F0 (hz)	192,6216	186,6780	p>0,05 (p=0,5479)
Jitter (%)	2,374596	3,315884	p<0,05 (p=7,9e-10)
Shimmer (%)	10,80915	14,83337	p<0,05 (p=1,215e-09)
HNR (dB)	12,98104	11,49695	p<0,05 (p=0,0007667)

TABLE 7.10: *Moyennes des paramètres Débit, F0, Jitter, Shimmer et HNR en fonction du type de voix jeunes ou âgées, et p-values du test t de Welch (différence significative si p<0,05).*

7.4.1 Comparaison des moyennes des variables prosodiques entre voix jeunes et âgées

Dans un premier temps, nous avons calculé les moyennes des 5 variables précédemment décrites pour les groupes *voix âgées* et *voix jeunes*, et nous avons vérifié si les moyennes sont significativement différentes entre elles. Les variables suivent une loi normale, nous avons donc utilisé le test t de Welch. Les résultats sont donnés table 7.10.

Nous observons des différences significatives pour toutes les variables sauf pour la variable *F0*. Pour les personnes âgées, par rapport aux personnes jeunes, nous observons une diminution du débit, une augmentation du jitter et du shimmer, et une diminution du HNR.

7.4.2 Clustering

Nous avons réalisé un clustering par la méthode de Ward afin de tester dans quelle mesures les 5 variables permettent de partitionner les données en groupes *voix jeunes* et *voix âgées*, comme illustré figure 7.3. Le clustering vise à diviser un ensemble de données en différents « paquets » homogènes, les données de chaque sous-ensemble partageant des caractéristiques communes.

Chacun des locuteurs est repéré sur la figure 7.3 par un code identifiant (ex : *A-H-90-4-58,6-13l33*). Ce code doit être interprété de la façon suivantes : GROUPE[jeune (J) ou âgé (A)]-GENRE[homme (H) ou femme (F)]-AGE-GIR-WER-ANNEE-ID. Par exemple, le locuteur J-H-28-6-8-04l05 est un locuteur du groupe *voix jeunes*, homme, 28 ans, avec un score GIR de 6, un WER avec le modèle *BREF120* de 8%, enregistré en 2004, avec comme identifiant *L05*.

Nous observons sur le dendogramme (figure 7.3) que les clusters se sont principalement formés d'une part selon le groupe *voix jeunes* ou *voix âgées* auquel appartiennent les locuteurs, et d'autre part selon le genre des locuteurs :

— les clusters 1 et 4 regroupent 79% des locuteurs jeunes de sexe masculin,

— les clusters 2 et 3 regroupent 88% des locuteurs jeunes de sexe féminin,

— les clusters 6 et 8 regroupent 57% des locuteurs âgés de sexe masculin,

— les clusters 5, 7, 8 et 9 regroupent 73% des locuteurs âgés de sexe féminin.

Ainsi, la méthode de clutering de Ward appliquée sur les paramètres prosodiques des locuteurs a mis en évidence que les groupes (les clusters) résultants sont une classification des

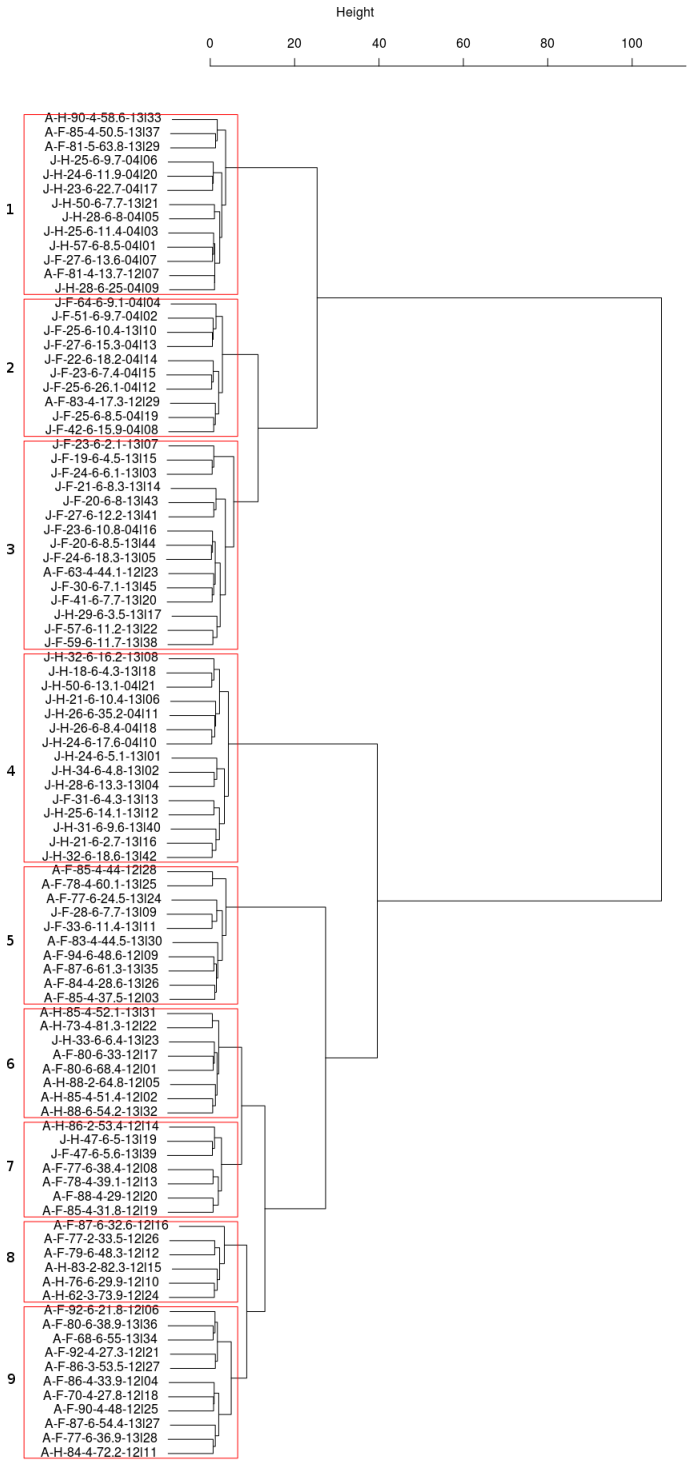


FIGURE 7.3: *Clustering des locuteurs AD80 à partir des paramètres prosodiques.*

locuteurs selon leur type de voix, jeune ou âgé, et selon leur genre, homme ou femme. En revanche, sur le dendogramme, nous ne visualisons pas quels ont été les critères sous-jacents ayant permis la formation des différents clusters. La section suivante propose une approche permettant de visualiser ces critères grâce à l'utilisation d'une analyse en composante principale.

7.4.3 Analyse en composantes principales

Afin de caractériser les relations existant entre les types de voix (jeunes ou âgées, masculines ou féminines) et les différents paramètres prosodiques, nous avons réalisé une analyse en composantes principales (ACP).

Classiquement, la description de la structure d'une corrélation entre 2 variables se fait par une représentation des observations en nuage de points ou par une régression, sur 2 dimensions. Pour 3 variables, une troisième dimension est ajoutée. Au delà de 3 variables, la question se pose de déterminer comment représenter la structure des corrélations des observations sur plus de 3 dimensions.

L'analyse en composantes principales est une méthode permettant de projeter les observations depuis l'espace à p dimensions des p variables vers un espace de dimensions k, avec k<p, tel qu'un maximum d'information soit conservée sur les premières dimensions. Si l'information associée aux 2 ou 3 premiers axes représente un pourcentage suffisant de la variabilité totale du nuage de points, on pourra représenter les observations sur un graphique à 2 ou 3 dimensions (2 ou 3 axes), afin d'identifier des groupes homogènes d'observations. Il s'agit donc d'effectuer un changement de base pour représenter le nuage de points sur 2 ou 3 nouveaux axes représentatifs des variables étudiées. Ces axes sont nommés « composantes principales », et représentent des « facteurs », qui sont mathématiquement calculés à partir des vecteurs propres de la matrice des corrélations. Ces axes sont indépendants et représentent au mieux la variabilité des données.

L'ACP se calcule sur des données centrées et réduites. Nous mesurons dans un premier temps la matrice des corrélations entre les 5 variables *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR*. Les résultats sont donnés table 7.11.

Puis nous calculons les valeurs propres de la matrice des corrélations, représentées table 7.12. Chaque ligne de la table 7.12 correspond à un facteur, dont la colonne « Valeur propre » fournit la variance du facteur, la somme des valeurs propres étant 5 (car 5 variables étu-

	Débit	F0	Jitter	Shimmer	HNR
Débit	1,000	-0,263	0,018	-0,0189	-0,073
F0	-0,263	1,000	-0,403	-0,587	0,650
Jitter	0,018	-0,403	1,000	0,867	-0,658
Shimmer	-0,019	-0,587	0,867	1,0007	-0,837
HNR	-0,073	0,650	-0,6587	-0,837	1,0007

TABLE 7.11: *Matrice des corrélations.*

Facteur	Valeur propre	% variance	Cumul % variance
1	3,000425	60,01	60,01
2	1,13994779	22,80	82,81
3	0,54575133	10,91	93,72
4	0,24645966	4,93	98,65
5	0,06741627	1,35	100,00

TABLE 7.12: *Valeurs propres.*

diées). La colonne « Cumul % variance » traduit la quantité d'information contenue dans les facteurs.

Les deux premiers facteurs (les 2 premiers axes) permettent de représenter 82,81% de l'information. Nous calculons donc les vecteurs propres sur les deux premiers axes.

Les vecteurs propres obtenus à partir de la matrice des corrélations (table 7.13) représentent l'inertie des variables initiales sur les facteurs. En effet, un facteur est une combinaison linéaire des variables initiales dans laquelle les coefficients sont donnés par les coordonnées des vecteurs propres, avec :

$F1 = -0,136.Débit + 0,720.F0 - 0,862.Jitter - 0,946.Shimmer + 0,909.HNR$, et

$F2 = 0,924.Débit - 0,423.F0 - 0,240.Jitter - 0,221.Shimmer + 0,016.HNR$.

Les corrélations des variables avec les axes factoriels peuvent être représentées par le cercle des corrélations présenté figure 7.4.

Une variable est d'autant plus corrélée à un axe que l'angle formé avec cet axe est faible. En outre, plus l'inertie d'une variable est absorbée par les 2 premiers axes, plus elle est proche du cercle : si l'inertie de la variable n'est presque pas absorbée par les 2 premiers

Facteurs –>	F1	F2
Débit	-0,136	0,924
F0	0,720	-0,423
Jitter	-0,862	-0,240
Shimmer	-0,946	-0,221
HNR	0,909	0,016

TABLE 7.13: *Vecteurs propres.*

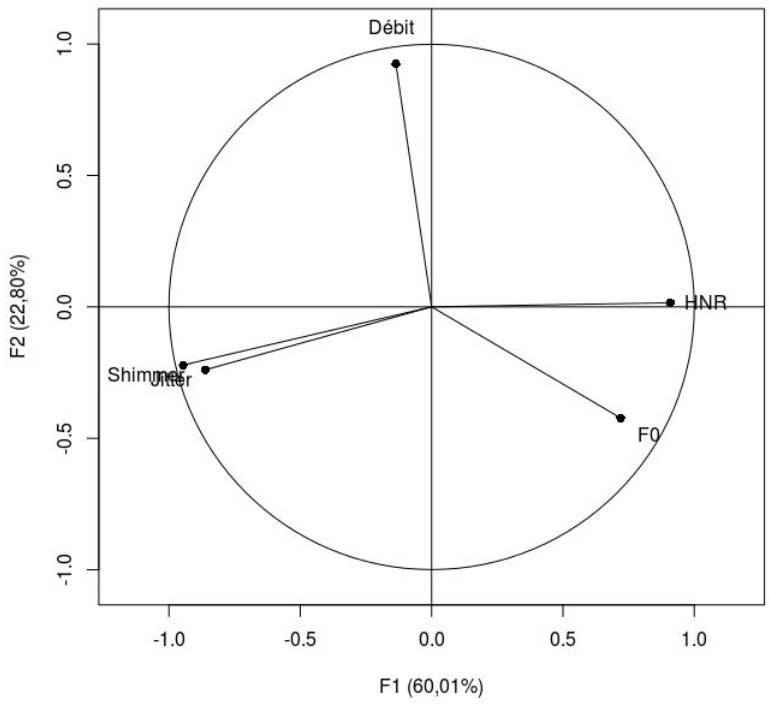


FIGURE 7.4: *ACP : projection des variables sur les axes principaux F1 et F2.*

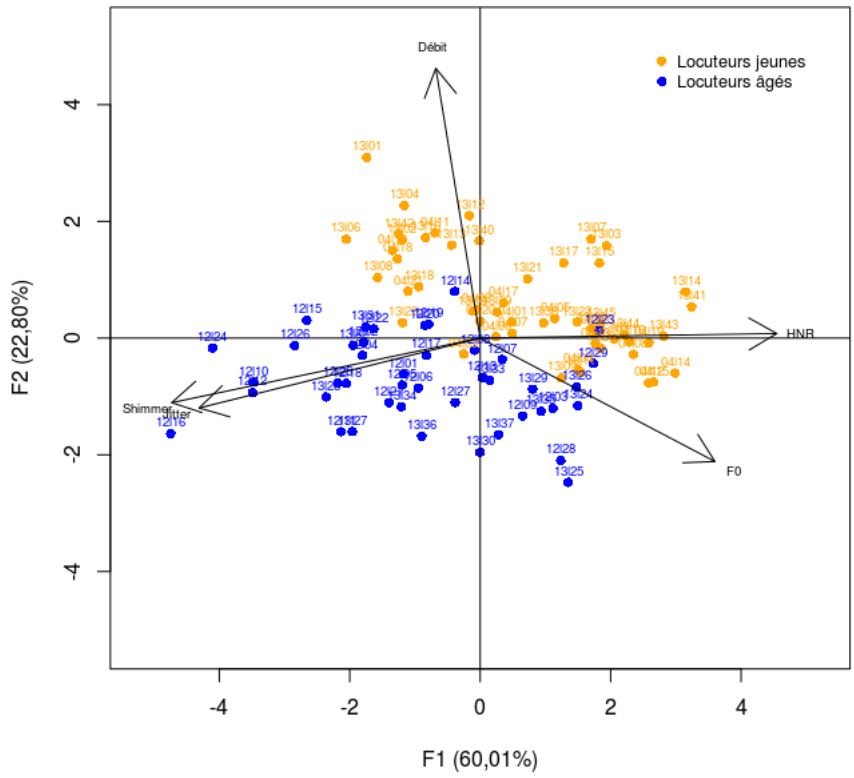


FIGURE 7.5: *ACP : projection des locuteurs sur les axes principaux F1 et F2 avec représentation des groupes « locuteurs jeunes » et « locuteurs âgés »*

axes mais par les axes suivants non représentés, son point représentatif est assez loin du cercle.

En observant la proximité entre les variables, nous remarquons la formation des 3 groupes suivants :

— *Jitter* et *Shimmer* : ces variables sont très proches l'une de l'autre car très corrélées entre elles, et leur inertie est principalement absorbée par l'axe *F1* dans la partie négative de cet axe. Ces variables sont également absorbées modérément par l'axe *F2* dans sa partie négative.

— *F0* et *HNR* : la variable *F0* a son inertie aussi bien absorbée par l'axe *F1* que par l'axe *F2*, dans la partie positive de l'axe *F1* et dans la partie négative de l'axe *F2*. *HNR* a une inertie totalement absorbée par l'axe *F1* dans sa partie positive.

— *Débit* : son inertie est principalement absorbée par l'axe *F2* dans sa partie positive, et très modérément par l'axe *F1* dans sa partie négative.

Toutes les variables sont proches du cercle et sont donc bien représentées sur le plan factoriel *F1-F2*.

La projection des individus sur les axes factoriels *F1* et F2 est donnée figures 7.5 et 7.6. On dit que les axes maximisent l'inertie de la projection du nuage de points. L'axe *F1* est l'axe absorbant le plus d'inertie des variables (60,019%), c'est donc l'axe représentant au mieux les

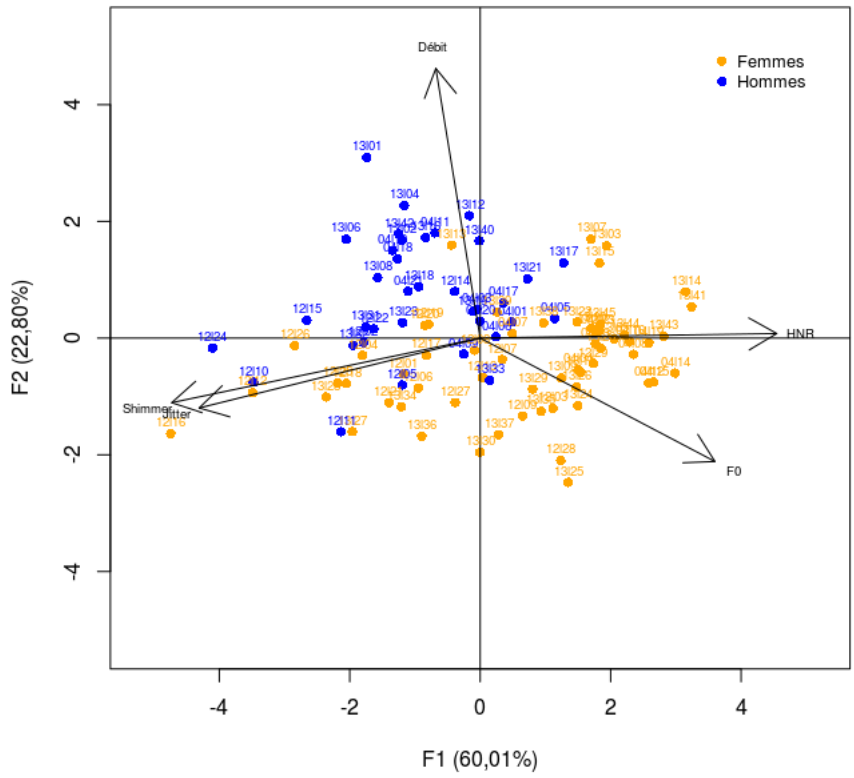


FIGURE 7.6: *ACP : projection des locuteurs sur les axes principaux F1 et F2 avec représentation des groupes « femmes » et « hommes »*

données.

La figure 7.5 met en évidence la distinction entre les groupes *locuteurs jeunes* et *locuteurs âgés*, avec une séparation bien marquée entre les 2 groupes. Nous observons que la discrimination entre les 2 groupes se fait principalement dans le sens des variables *Jitter* et *Shimmer* (valeurs plus grandes pour les locuteurs âgés), ainsi que dans le sens de la variable *Débit* (valeurs plus grandes pour les locuteurs jeunes).

La figure 7.6 met en évidence les groupes *hommes* et *femmes*, avec une séparation entre les 2 groupes plus floue. Cette fois-ci, la discrimination entre ces groupes se fait principalement dans le sens des variables *F0* (valeurs plus grandes pour les femmes).

La variable HNR semble ne pas jouer un rôle important dans la discrimination hommes/-femmes ou jeunes/âgés.

La discrimination entre les locuteurs jeunes et âgés se faisant avec les variables *Jitter*, *Shimmer* et *Débit*, la corrélation entre le WER et ces variables devrait être assez importante, ce que nous allons vérifier dans la section suivante.

7.4.4 Corrélation entre les paramètres prosodiques et le WER

La table 7.14 présente les corrélations entre les WER obtenus avec le modèle *BREF120* et les paramètres prosodiques *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR*.

	Corrélation avec WER
Débit	-0,484 (p=6,685e-07)
F0	-0,172 (p=0,096)
Jitter	0,510 (p=1,274e-07)
Shimmer	0,540 (p=1,669e-08)
HNR	-0,362 (p=3,100e-04)

TABLE 7.14: *Corrélation entre le WER et les paramètres prosodiques.*

Nous observons que le WER a une corrélation en valeur absolue d'environ 0,5 avec les variables *Jitter*, *Shimmer* et *Débit*, la corrélation étant négative pour *Débit*, et positive pour *Jitter*, *Shimmer*. Enfin, le WER est faiblement corrélé en négatif avec la variable *HNR*, et la corrélation avec la variable *F0* est non significative (p-value>0,05).

7.4.5 Application de classifieurs sur les paramètres prosodiques

7.4.5.1 Prédiction du type de voix « jeune » ou « âgée »

Il peut être intéressant d'être en mesure de déterminer automatiquement si une voix est de type *jeune* ou *âgé*, en vue par exemple d'une sélection automatique d'un modèle acoustique approprié et adapté. Aussi, une prédiction des performances de la RAP sur un futur utilisateur peut être intéressante pour déterminer si l'installation du système de RAP à son domicile fait sens.

Nous avons entraîné un classifieur à régression logistique, à l'aide de l'outil Weka, dans le but de prédire si une voix appartient à la classe *jeune* ou *âgés* à partir des 5 variables *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR*.

L'apprentissage et le test ont été réalisés par validation croisée par la méthode du *k-fold*.

Pour les 95 locuteurs du corpus *AD80* (nous avons donc 95 observations), les résultats de la validation croisée avec k=n=95 sont donnés tables 7.15 et 7.16.

Les variables prosodiques *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR* permettent d'obtenir une bonne prédiction pour les classes *jeunes* ou *âgés*, avec précision=0,905, rappel=0,905 et F-mesure=0,905 (moyenne pondérée, table 7.15). Cependant, pour une application temps-réel, il faudrait réaliser l'apprentissage et la classification à partir de mesures réalisées sur

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,884	0,077	0,905	0,884	0,894	0,966	Âgés
	0,923	0,116	0,906	0,923	0,914	0,966	Jeunes
Moyenne pondérée :	0,905	0,098	0,905	0,905	0,905	0,966	

TABLE 7.15: *Validation croisée sur les classes « âgés » et « jeunes ».*

Âgés	Jeunes	<– classifié comme
38	5	Âgés
4	48	Jeunes

TABLE 7.16: *Matrice de confusion pour la validation croisée sur les classes « âgés » et « jeunes ».*

une fenêtre temporelle plus petite, de l'ordre de quelques secondes. En effet, les mesures utilisées ici ont été calculées sur une fenêtre temporelle assez longue, d'environ 2 minutes.

7.4.5.2 Prédiction du WER

Afin d'analyser si une prédiction du WER est possible à partir des variables prosodiques, un nouvel apprentissage a été réalisé afin de prédire les 2 classes *WER=[0-26[%* (bon WER) ou *WER=[26-100]%* (mauvais WER), les WER étant obtenus avec un décodage à partir du modèle acoustique *BREF120*. Les résultats de la validation croisée avec k=n=95 sont donnés tables 7.17 et 7.18.

Les scores obtenus sont : précision=0,884, rappel=0,884 et F-mesure=0,884 (moyenne pondérée). La prédiction avec un classifieur du WER à partir des paramètres prosodiques est donc relativement bonne en considérant une répartition du WER sur 2 classes de WER.

Une validation croisée a été effectuée avec un nombre plus important de classes de WER. Sur les classes de WER définies comme suit : *WER=[0-13[%*, *WER=[13-26[%*, *WER=[26-39[%*

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,907	0,146	0,891	0,907	0,899	0,913	WER=[0-26[%
	0,854	0,093	0,875	0,854	0,864	0,913	WER=[26-100]%
Moyenne pondérée :	0,884	0,123	0,884	0,884	0,884	0,913	

TABLE 7.17: *Validation croisée sur les classes WER=[0-26[% et WER=[26-100]% (modèle BREF120).*

WER=[0-26[%	WER=[26-100]%	<– classifié comme
49	5	WER=[0-26[%
6	35	WER=[26-100]%

TABLE 7.18: *Matrice de confusion pour la validation croisée sur les classes WER=[0-26[% et WER=[26-100]% (modèle BREF120).*

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,811	0,293	0,638	0,811	0,714	0,826	WER=[0-13[%
	0,118	0,064	0,286	0,118	0,167	0,618	WER=[13-26[%
	0,375	0,114	0,400	0,375	0,387	0,703	WER=[26-39[%
	0,600	0,157	0,577	0,600	0,588	0,845	WER=[39-100]%
Moyenne pondérée :	0,558	0,186	0,519	0,558	0,528	0,773	

TABLE 7.19: *Validation croisée sur les classes WER=[0-13[%, WER=[13-26[%, WER=[26-39[% et WER=[39-100]% (modèle BREF120).*

WER=[0-13[%	WER=[13-26[%	WER=[26-39[%	WER=[39-100]%	<– classifié comme
30	5	1	1	WER=[0-13[%
12	2	1	2	WER=[13-26[%
2	0	6	8	WER=[26-39[%
3	0	7	15	WER=[39-100]%

TABLE 7.20: *Matrice de confusion pour la validation croisée sur les classes WER=[0-13[%, WER=[13-26[%, WER=[26-39[% et WER=[39-100]% (modèle BREF120).*

k=95	TVP	TFP	Précision	Rappel	F-Mesure	Aire ROC	Classe
	0,817	0,629	0,690	0,817	0,748	0,653	WER=[0-15[%
	0,371	0,183	0,542	0,371	0,441	0,653	WER=[15-50]%
Moyenne pondérée :	0,653	0,465	0,635	0,653	0,635	0,653	

TABLE 7.21: *Validation croisée sur les classes WER=[0-15[% et WER=[15-50]% (modèle BREF120 pour les locuteurs jeunes, et modèle BREF120_MLLR_G pour les locuteurs âgés).*

WER=[0-15[%	WER=[15-50]%	<– classifié comme
49	11	WER=[0-15[%
22	13	WER=[15-50]%

TABLE 7.22: *Matrice de confusion pour la validation croisée sur les classes WER=[0-15[% et WER=[15-50]% (modèle BREF120 pour les locuteurs jeunes, et modèle BREF120_MLLR_G pour les locuteurs âgés).*

et *WER=[39-100]%*, les résultats sont donnés tables 7.19 et 7.20.

Les scores obtenus sont médiocres, avec précision=0,519, rappel=0,558 et F-mesure=0,528 (moyenne pondérée). Les paramètres prosodiques ne permettent globalement pas une prédiction très précise du WER avec un classifieur. Malgré tout, nous pouvons observer que les classes extrêmes *WER=[0-13[%* et *WER=[39-100]%* sont mieux reconnues que les classes intermédiaires, avec précision=0,638, rappel= 0,811 et F-mesure=0,714 pour la classe *WER=[0-13[%*, et précision=0,577, rappel= 0,600 et F-mesure=0,588 pour la classe *WER=[39-100[%*.

Enfin, en utilisant le WER obtenu avec le modèle *BREF120* pour les locuteurs jeunes et le modèle adapté*BREF120_MLLR_G* pour les locuteurs âgés, nous avons effectué une validation croisée sur les classes *WER=[0-15[%* et *WER=[15-50]%*. Les résultats sont donnés tables 7.21 et 7.22.

Les scores pour les 2 classes *WER=[0-15[%* et *WER=[15-50]%* sont les suivants : précision=0,635, rappel=0,653 et F-mesure=0,635 (moyenne pondérée). Suite à l'adaptation acoustique à la voix âgée, les écarts entre les WER des locuteurs sont plus faibles, aboutissant à une prédiction du WER peu performante.

L'utilisation de ce type de paramètres prosodiques apparaît donc peu probante pour la prédiction des performances des systèmes de RAP adapté à la voix âgée.

7.5 Bilan

Dans ce chapitre, nous avons étudié comment différents paramètres, dont l'âge, peuvent expliquer la dégradation du WER. Nous avons ensuite recherché s'il était possible de prédire le WER d'un locuteur à partir de ces paramètres.

Nous avons vu qu'il existe une corrélation élevée statistiquement significative (corrélation=0,747 ($p<0,05$)) entre l'âge et le WER lorsque l'ensemble des locuteurs est considéré (jeunes et âgés). En revanche, à l'intérieur du groupe *jeunes* ou du groupe *âgés*, cette corrélation est très faible (autour de -0,1, avec $p>0,05$).

Aussi, nous avons vu qu'il existe une différence significative de WER entre les locuteurs âgés très dépendants et les locuteurs âgés autonomes, et que la corrélation entre le score GIR et le WER est significative, mais faible (corrélation=-0,318 ($p<0,05$)).

Nous observons que les scores d'alignement forcé sont fortement corrélés avec le WER (corrélation=-0,81 ($p<0,05$)). En essayant de prédire le WER avec un classifieur à partir des scores d'alignement forcé, les résultats sont assez bons lorsque 2 classes de WER sont considérées (WER compris entre 0 et 25%, et entre 26 et 100%), avec F-mesure=0,873. En revanche, en augmentant le nombre de classe, soit 4 classes de WER, les résultats diminuent sensiblement, avec F-mesure=0,63. Il est donc très difficile d'avoir une prédiction précise du WER avec cette méthode.

Nous avons ensuite étudié les différences prosodiques entre les locuteurs jeunes et les locuteurs âgés en considérant 5 paramètres : *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR*. Ces paramètres semblent pertinents car en réalisant un clustering, ils permettent une séparation et regroupement des locuteurs en fonction de leur âge et de leur genre. Nous avons aussi réalisé une ACP, qui a montré que la distinction jeunes/âgés se fait grâce aux paramètres *Jitter*, *Shimmer* et *Débit*, et qu'une distinction hommes/femmes se fait grâce au paramètre *F0*. Pour les locuteurs âgés par rapport aux jeunes, le jitter et le shimmer augmentent, et le débit diminue, et la F0 est évidemment plus grande pour les femmes que pour les hommes. Le paramètre *HNR* semble ne pas intervenir dans les distinctions jeunes/âgés et hommes/femmes. Nous avons aussi calculé les corrélations entre ces paramètres et le WER. Les corrélations sont statistiquement significative ($p<0,05$) pour les paramètres *Débit*, *Jitter*, *Shimmer* et *HNR* mais restent modérées, entre 0,36 et 0,54 en valeur absolue, et la corrélation est non significative ($p>0,05$) pour le paramètre *F0*. Nous avons appliqué un classifieur à partir des paramètres prosodique pour prédire la catégorie des locuteurs : jeune ou âgés. La classification donne de bons résultats, avec F-mesure=90,5%. Nous avons également réalisé avec le classifieur une prédiction du WER à partir des paramètres prosodiques. En considérant 2 classes de WER, la F-mesure est assez bonne (F-mesure=0,884). En revanche, avec 4 classes de WER, la F-mesure chute à 0,028. Les paramètres prosodiques ne permettent donc pas de prédire avec précision le WER.

Après avoir étudié l'influence de ces différents paramètres, nous allons étudier dans le chapitre suivant les effets de la voix émue en situation de détresse sur les performances des systèmes de RAP.

CHAPITRE 8

Etude des performances des système de RAP avec la voix émue en situation de détresse

Notre objectif est de développer un système capable de détecter des appels de détresses et des appels aux aidants prononcés par des personnes âgées. Dans les chapitres précédents, 6 et 7, nous avons traité la problématique de la dégradation des performances des systèmes de RAP avec la voix des personnes âgées. Les phrases de test utilisées, des phrases lues, étaient certes des phrases dont le contenu concernait le thème de la détresse, mais elles étaient prononcées de façon « neutre » [1] et lues, sans émotion particulière.

En situation réelle, une phrase d'appel aux aidants peut être prononcée sur un ton « neutre » (par exemple pour commander un plateau repas) ou sur un ton fortement perturbé par l'émotion (par exemple sur le ton de la peur pour appeler une infirmière suite à une difficulté). Dans une situation de détresse, une phrase d'appel au secours (par exemple : « Au secours, je suis tombé ! ») sera vraisemblablement fortement chargée en émotions et perturbée.

Or, d'après Vlasenko et coll. (2011), il existe des différences acoustiques entre la voix émue et la voix neutre. Les modèles acoustiques étant appris sur de la voix neutre, Vlasenko et coll. (2012) ont mis en évidence une dégradation des performances des systèmes de RAP avec la voix émue.

Comme nous l'avons vu précédemment, beaucoup d'études sont consacrées à la reconnaissance des émotions (détresse, peur, stress, etc.). Mais le vrai enjeu immédiat, peut-être plus facilement accessible, est de rendre robuste aux perturbateurs émotionnels la RAP plutôt que de reconnaître l'état de la détresse lui-même.

L'étude que nous proposons ici vise à mettre en évidence le manque de robustesse des systèmes de RAP avec de la voix prononcée de façon émue. Dans une étude préliminaire, nous mettrons en évidence les différences de performance des systèmes de RAP entre la voix émue actée et la voix émue spontanée à partir d'un corpus de voix émues enregistré grâce à la technique du magicien d'Oz, et nous comparerons ces résultats avec la voix neutre. Puis nous poursuivrons par une évaluation à partir d'un corpus de phrases de détresse, le corpus *Voix Détresse* décrit en section 5.4, que nous avons enregistré au laboratoire, et qui est un corpus de phrases prononcées de façon neutre et émue dans un protocole d'élicitation actée.

1. Nous utiliserons la qualification « neutre » pour désigner des élocutions non perturbées fortement par des facteurs expressifs (intentions, attitudes, émotions, etc.) même s'il est souvent rappelé dans la littérature traitant de la parole spontanée que la neutralité n'est pas une réalité écologique.

8.1 Étude préliminaire

Dans une première étude, nous avons étudié l'impact des émotions sur les systèmes de reconnaissance automatique de la parole en comparant 3 modalités d'expression :

— parole neutre,

— parole émue actée,

— parole émue spontanée.

Le corpus de test utilisé pour cette étude est le corpus *E-Wiz* (Aubergé et coll., 2004), qui est un corpus de parole émue. Le point fort de ce corpus est qu'il possède des enregistrements de parole émue prononcés de façon spontanée. Ce corpus contient également des énoncés de parole émue actée. Concernant la voix neutre, nous avons complété le corpus *E-Wiz* avec nos propres enregistrements de parole neutre. Les données de 6 locuteurs du corpus *E-Wiz*, enregistrés en 2003 par le laboratoire GIPSA ont été utilisées pour la voix émue. Pour compléter ces enregistrements pour la voix neutre, nous avons nous-même enregistré 7 locuteurs (différents des locuteurs du corpus *E-Wiz*) en 2012.

8.1.1 Le corpus E-Wiz

Le corpus *E-Wiz* a été enregistré à l'aide de la technique du magicien d'Oz, dans laquelle le sujet est convaincu d'être en interaction avec une interface homme-machine complexe, alors que le comportement apparent de l'application est contrôlé à distance par l'expérimentateur. Les 17 sujets ont été recrutés avec le prétexte de participer aux derniers essais de pré-commercialisation d'une nouvelle application d'apprentissage des langues à base de reconnaissance vocale. L'application était présentée comme agissant directement sur la plasticité cérébrale des sujets, permettant un apprentissage rapide et facile de la prononciation de voyelles étrangères. La tâche consistait en la discrimination perceptive de paires de voyelles, présentées visuellement dans le triangle acoustique. Les interactions des sujets avec le système étaient restreintes à des commandes vocales, composées des noms des couleurs monosyllabiques « brique », « jaune », « rouge », « sable » et « vert », et de la commande « page suivante », permettant la collecte d'au moins 20 déclarations de chaque stimulus par sujet.

Les performances d'apprentissage des 17 sujets étaient en réalité manipulées selon un scénario prédéfini. Les habilités de perception de sujets leur étaient d'abord présentées par *feedback* comme étant excellente, puis, au fur et à mesure, il leur était fait croire que leur compétence perceptive se dégradait, et cela de pire en pire. Dans la dernière étape du scénario, des stimuli audios modifiés étaient présentés aux sujets pour leur induire des choix de réponse aléatoires, en prétendant que l'application avait endommagé leur habilité de perception. Ce scénario permettait ainsi l'induction d'émotions aussi bien positives que négatives. Les émotions ont ensuite été annotées par les sujets eux-même à partir des enregistrements vidéos.

De plus, un protocole adapté a été présenté à 7 sujets acteurs, pratiquant le théâtre d'improvisation et/ou le théâtre de rue. Immédiatement après l'expérience magicien d'Oz, ces

sujets ont eu la tâche de prononcer les mêmes énoncés avec les émotions qu'ils avaient vécues durant l'expérience, et également avec les émotions les plus fréquemment étudiées (tristesse, colère, peur, dégoût, surprise et joie) en utilisant leurs techniques d'acteur.

Le corpus *E-Wiz* a été utilisé par Laukka et coll. (2007) dans leur étude sur l'expression vocale des émotions. Ils ont retenu les productions de 6 locuteurs (3 hommes, 3 femmes), tous acteurs, et ont classé les expressions en 3 catégories, désignées comme *spontané*, *acté non-prototypique* et *acté prototypiques*. Trois grandes catégories d'émotions, pour lesquelles il y avait un nombre suffisant d'énoncés, ont été choisis : la colère, la peur et la joie. Dans un pré-test, ils ont conservé 193 productions prononcés par les 6 locuteurs, et ils ont fait évaluer ces productions par 15 locuteurs francophones « naïfs ». Ils ont été autorisés à écouter les stimuli autant de fois qu'ils le voulaient, et devaient sélectionner pour chacun une classe d'émotion entre les 3 proposées (colère, peur ou joie), ou la classe *autre émotion*. Parmi les stimuli identifiés avec une précision au-dessus de hasard, les auteurs ont retenus 12 stimuli pour chaque émotion et dans chaque catégorie, soit un total de 108 stimuli.

8.1.2 Notre étude à partir du corpus E-Wiz

Dans notre étude, nous avons gardé les mêmes 108 productions, mais nous avons fusionné les catégories *acté non-prototypique* et *acté prototypique* dans une catégorie que nous avons appelé *acté*.

Pour comparer la parole émue avec la parole neutre, nous avons recueilli des données pour construire la troisième catégorie que nous avons qualifiée *neutre*. 7 sujets (3 hommes, 4 femmes) ont été enregistrés. Nous leur avons demandé de lire les énoncés de la manière la moins expressive possible, en utilisant une application permettant d'afficher les stimuli sur un moniteur (le logiciel *GEOD*). Les stimuli étaient les mêmes que pour la parole émue avec les 5 noms de couleurs monosyllabiques « brique », « jaune », « rouge », « vert » et « sable », et la commande « page suivante ». Chaque locuteur a lu près de 10 occurrences de chaque type de stimuli, pour un total de 415 énoncés.

Tous les locuteurs testés étaient des locuteurs non âgés.

Nous avons lancé le décodage le décodeur *Google Speech API* sur les 3 catégories de parole : *neutre*, *émue actée* et *émue spontanée*.

Le WER moyen est égal à 37,1% pour la parole neutre, 65,2% pour la parole émue actée et 75,5% pour la parole émue spontanée. Pour la parole émue en général (moyenne de actée et spontanée), le WER est égal à 69,0%. Nous pouvons observer une différence absolue de 31,9% entre la parole neutre et la parole émue. Cela démontre l'influence des émotions sur les performances des systèmes de RAP. Nous pouvons également observer que la parole expressive spontanée est moins bien reconnue que la parole expressive actée, avec une différence absolue de 10,3%.

A cause des énoncés mono-mot, un grand nombre d'entre eux sont décodés par un homonyme ou un mot phonétiquement très proche (par exemple « vert » décodé « verre », ou « jaune » décodé « John »). En appliquant le filtre décrit 6.4.2 basé sur la distance de Levensh-

tein entre les hypothèses et les énoncés cibles phonétisés, nous obtenons des taux de confusion de 2,41%, 4,17% et 8,61% (en passant au niveau phonétique, le problème des homonymes n'intervient plus), pour, respectivement, la parole neutre, la parole émue actée et la parole émue spontanée.

8.2 Utilisation du corpus Voix Détresse

Afin de comparer les performances du système Sphinx3 entre le ton neutre et le ton ému (détresse) pour les voix jeunes et les voix âgées, nous avons découpé notre corpus *Voix Détresse* (neutre vs. émotion élicitée) – prononcé par 20 locuteurs jeunes et 5 locuteurs âgées – en 4 groupes. Ce corpus que nous avons enregistré a été décrit dans la section 5.4. Pour chaque groupe, nous avons réservé la moitié des phrases pour l'adaptation des modèles acoustiques aux locuteurs et l'autre moitié pour les tests, obtenant une sous-partie du corpus nommée *adaptation* (742 phrases) et une autre nommée *test* (739 phrases). Les 4 groupes sont les suivants :

— voix neutres jeunes : 400 phrases, dont 200 pour l'adaptation et 200 pour le test,

— voix émue jeunes : 782 phrases, dont 393 pour l'adaptation et 389 pour le test,

— voix neutres âgées : 100 phrases, dont 50 pour l'adaptation et 50 pour le test,

— voix émue âgées : 199 phrases, dont 99 pour l'adaptation et 100 pour le test.

Étant donné que 25 locuteurs ont participé aux enregistrements, cela représente environ 10 phrases par locuteur réservées à l'adaptation à la voix neutre et 10 autres phrases par locuteur pour les décodages sur la voix neutre, et environ 20 phrases par locuteur réservées à l'adaptation à la voix émue, ainsi que 20 phrases par locuteur pour les décodages sur la voix émue.

8.3 Système de référence pour la voix émue

8.3.1 Décodage avec Sphinx3

Le modèle acoustique utilisé pour le décodage avec Sphinx3 est le modèle acoustique générique *BREF120*. Le même modèle de langage que celui des sections précédentes (modèle présenté en section 6.2.1) a été utilisé (modèle appris sur les phrases de détresse et d'appels aux aidants combiné au modèle *Gigaword*).

Le décodage a été effectué sur les phrases neutres et les phrases émues prononcées par les locuteurs jeunes et âgés de la sous-partie *test* du corpus *Voix Détresse* (739 phrases de test au total). Les résultats par locuteur sont présentés figure 8.1, et les WER moyens pour chaque groupe sont présentés table 8.1.

Nous observons une dégradation importante du WER entre les voix neutres et les voix émue avec une différence absolue de 29,95% pour les locuteurs jeunes, et de 19,6% pour les locuteurs âgées.

Nous avons réalisé un test statistique pour déterminer si les différences de WER entre les voix neutres et les voix émue sont significatives. Les échantillons suivent une loi normale, nous avons donc utilisé le test t de Welch. Les résultats du test montrent une différence significative entre les WER des voix neutres et des voix émues, avec t=-7,2026, df=21,365, p<0,05 (p=3,834e-07) pour les locuteurs jeunes, et t=-2,5165, df=7,874, p<0,05 (p=0,03645) pour les locuteurs âgés.

Sur la figure 8.1, nous observons que le WER a augmenté pour la voix émue par rapport à la voix neutre pour l'ensemble des locuteurs (à l'exception du locuteur *J10F25*). En observant la variation de WER entre voix neutre et voix émue pour chaque locuteur, nous remarquons que ces variations sont plus ou moins importantes selon les locuteurs. Par exemple, pour le locuteur *J16F53*, le WER sur voix neutre est de 8,8%, et, sur voix émue, le WER est de 72,7% (différence absolue de 63,9%) ; pour le locuteur *J01M31*, le WER est de 11,8% sur voix neutre, et de 25,0% sur voix émue (différence absolue de 13,2%). Ces différences de variations entre locuteurs s'expliquent en partie par le fait que le corpus de parole émue utilisé était un corpus de parole actée, enregistré par des non professionnels (nous avons recruté les volontaires non âgés parmi le personnel du laboratoire, et les sujets âgés (entre 67 et 87 ans) ont été enregistrés à la suite d'une autre expérience sur le maintien à domicile). En effet, les personnes, selon leur caractère et leurs compétences à simuler, exprimaient plus ou moins bien les émotions de détresse à travers leur voix, elles étaient plus ou moins à l'aise pour

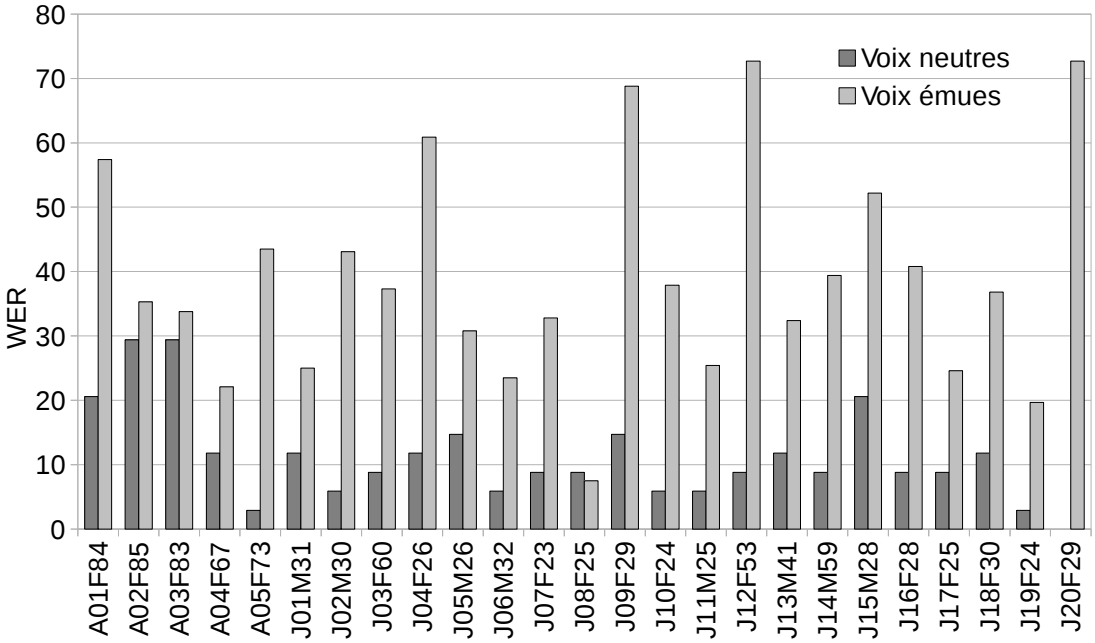


FIGURE 8.1: *WER issus du décodage avec Sphinx3 (modèle acoustique BREF120) en fonction du ton – neutre ou ému.*

	Locuteurs jeunes	Locuteurs âgés
Voix neutres	9,27%	18,82%
Voix émues	39,22%	38,42%

TABLE 8.1: *WER moyens issus du décodage avec Sphinx3 (modèle acoustique BREF120) pour les voix neutres et émues parmi les locuteurs jeunes et âgés.*

« jouer le jeu » de la détresse. Globalement, les personnes ressenties comme étant les plus à l'aise pour jouer les situations de détresse sont les personnes ayant un WER le plus élevé pour les phrases émues. Nous pouvons donc craindre qu'en situation réelle, une situation de détresse exprimée de façon extrêmement émue sera très mal reconnue par un tel système de RAP, alors que c'est justement sur ce type de situation qu'il est primordial de pouvoir agir.

8.3.2 Décodage avec Google Speech API

Nous avons réalisé avec le décodeur *Google Speech API* un décodage sur les 739 phrases neutres et émues de la sous-partie *test* du corpus *Voix Détresse*. Les résultats sont donnés table 8.2.

Nous observons que le WER avec *Google Speech API* est globalement élevé. Une part importante du WER est due à l'absence de résultat (hypothèse vide) retourné par Google pour un nombre important de phrases à décoder. En effet, Google ne renvoye pas de résultat si le score de confiance calculé est trop bas. La proportion de phrases sans résultat est indiquée entre parenthèses dans la table 8.2. Au total, Google a échoué (hypothèses vides) à décoder 46% des phrases soumises au décodage. Ces phrases ont été maintenues dans le calcul du WER, elles représentent une grande part des délétions.

Nous observons une dégradation des performances avec la parole émue par rapport à la parole neutre, avec une différence absolue de 18,8% pour les locuteurs jeunes, et de 4,3% chez les locuteurs âgés. Ainsi, au vu du taux plus important d'hypothèses vides (dans le cas des locuteurs jeunes) et du WER plus important pour la parole émue, il semblerait que les modèles acoustiques de Google soient moins adaptés à la voix émue qu'à la voix neutre.

Cependant, pour certains locuteurs, la voix émue présente un meilleur WER que la voix neutre. La figure 8.2 représente les WER de chaque locuteur pour la parole neutre et la parole émue. Pour 14 locuteurs jeunes et 4 locuteurs âgés, le WER est plus haut dans le cas de la parole émue par rapport à la parole neutre, mais pour les 6 locuteurs jeunes et un locuteur âgé, le WER est plus haut dans le cas de la voix émue.

Dans les cas où le WER de la voix émue est plus important que le WER de la voix neutre, la différence en valeur absolue est globalement largement plus grande que dans les cas où le WER de la voix émue est plus faible que le celui de la voix neutre.

Nous avons testé la significativité de ces différences. Les échantillons suivent une loi normale, nous avons donc utilisé, comme dans la section 8.3.1, le test t de Welch. Pour les locuteurs jeunes, nous obtenons une différence significative entre les WER des voix neutres et les WER des voix âgées, avec t=-3,0484, df=33,996, p<0,05 (p=0,004432). En revanche, pour les lo-

	Locuteurs jeunes	Locuteurs âgés
Voix neutres	34,2% (30%)	47,0% (46%)
Voix émues	53,0% (41,4%)	51,3% (44%)

TABLE 8.2: *WER moyens (et proportion d'hypothèses vides) pour le décodage avec Google Speech API pour les voix neutres et émues parmi les locuteurs jeunes et âgés.*

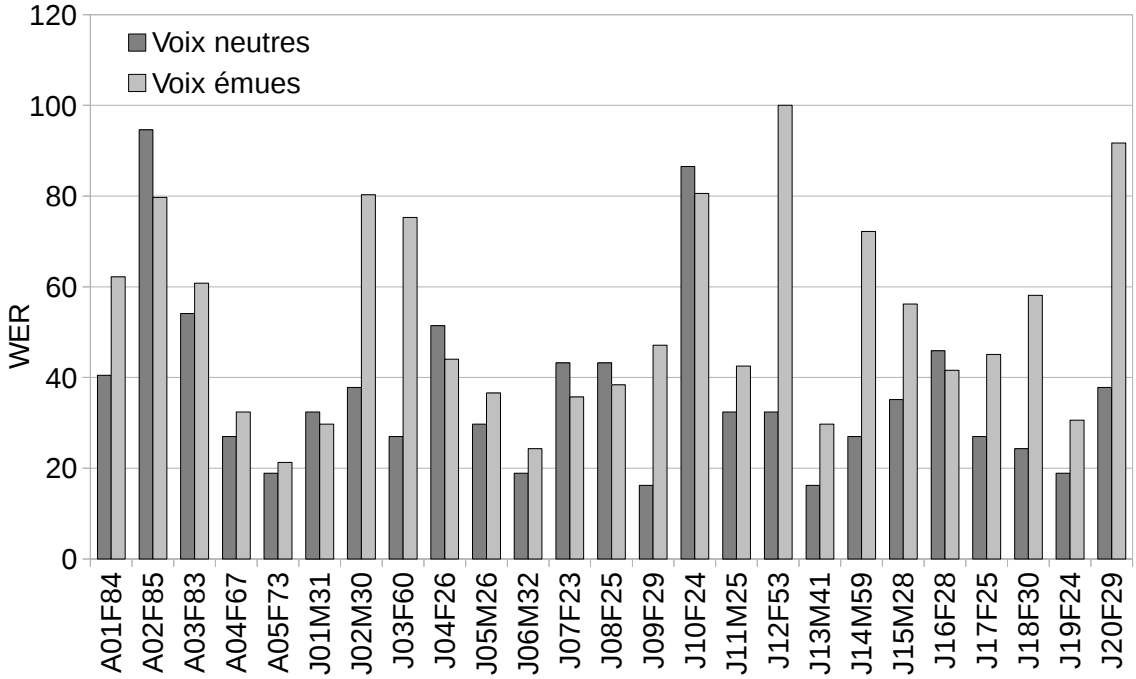


FIGURE 8.2: *WER issus du décodage avec Google Speech API en fonction du ton – neutre ou ému.*

cuteurs âgés, la différence n'est pas significative, avec t=-0,2497, df=7,634, p>0,05 (p=0,8094), mais cela étant peut-être lié au faible nombre de personnes âgées enregistrés (5 personnes).

8.4 Adaptation des modèles acoustiques au locuteur pour la voix émue

Afin de déterminer s'il est possible d'améliorer les résultats avec *Sphinx3*, nous avons réalisé une adaptation MLLR au locuteur sur le modèle acoustique générique *BREF120* à partir des phrases de la sous-partie *adaptation* du corpus *Voix Détresse*. Trois types d'adaptations au locuteur ont été réalisées :

— adaptation à la voix neutre : pour chaque locuteur, l'adaptation a été réalisée à partir de 10 phrases neutres prononcées par ce même locuteur. Les modèles adaptés obtenus sont les modèles nommés *BREF120_MLLR_LOC_N*,

— adaptation à la voix émue : pour chaque locuteur, l'adaptation a été réalisée à partir de

Groupe	BREF120	BREF120_MLLR_LOC_N	BREF120_MLLR_LOC_E	BREF120_MLLR_LOC_N+E
JN	9,27	7,80	11,03	7,21
JE	39,22	28,86	22,45	20,23
AN	18,82	17,06	16,48	15,88
AE	38,42	35,48	31,64	30,50
Moy. N	11,18	9,65	12,12	8,94
Moy. E	39,06	30,18	24,28	22,28

TABLE 8.3: *WER (%) en fonction des différents modèles acoustiques pour la voix des locuteurs jeunes (J) et âgés (A), avec une intonation neutre (N) ou émue (E).*

20 phrases neutres prononcées par ce même locuteur. Les modèles adaptés obtenus sont les modèles *BREF120_MLLR_LOC_E*,

— adaptation sans distinction neutre ou émue : pour chaque locuteur, l'adaptation a été réalisée à partir de 10 phrases neutres et 20 phrases émues prononcées par ce même locuteur. Les modèles adaptés obtenus sont les modèles *BREF120_MLLR_LOC_N+E*.

Chaque locuteur possède ainsi 3 modèles différents adaptés spécifiquement à sa voix.

Le décodage a été réalisé, comme dans les sections 8.3.1 et 8.3.2, sur les 739 phrases neutres et émues de la sous-partie *test* du corpus *Voix Détresse*, avec les différents modèles acoustiques adaptés aux locuteurs, et nous avons effectué une comparaison avec le modèle *BREF120*.

Les résultats des WER moyens pour les groupes *voix neutres jeunes*, *voix émues jeunes*, *voix neutres âgées* et *voix émues âgées* sont donnés table 8.3.

Une ANOVA suivie d'un test de Tukey HSD ont été réalisés pour chaque groupe (les échantillons suivent une loi normale et vérifient l'homogénéité des variances). Nous avons fusionné le groupe *voix neutres âgées* avec le groupe *voix neutres jeunes*, ainsi que le groupe *voix émues âgées* avec le groupe *voix émues jeunes*, car nous n'avons pas suffisamment de locuteurs âgés pour réaliser une ANOVA sur les groupes correspondant. Les ANOVA montrent qu'il n'y a pas de différence significative entre les échantillons du groupe *voix neutres* (F(3 ;96)=1,293 ; p=0,281), et qu'il existe une différence significative (p-value < 0,05) entre les échantillons du groupe *voix émues* (F(3 ;96)=7,828 ; p=9,96e-05). Les résultats du test de Tukey HSD sont donnés table 8.4.

Pour les voix neutres, nous n'observons aucune différence significative entre les WER obtenus avec les différents modèles acoustiques, avec p>0,05 pour chaque paire de modèles acoustiques.

	BREF120	BREF120..._N	BREF120..._E	BREF120..._N+E
BREF120	-	p=0,8298	p=0,9529	p=0,5977
BREF120_MLLR_LOC_N	-	-	p=0,5174	p=0,9786
BREF120_MLLR_LOC_E	-	-	-	p=0,2923
BREF120_MLLR_LOC_N+E	-	-	-	-
WER (%)	11,18	9,65	12,12	8,94

(a) *Groupe « voix neutres ».*

	BREF120	BREF120..._N	BREF120..._E	BREF120..._N+E
BREF120	-	p=0,0976	**p=0,0011**	**p=0,0002**
BREF120_MLLR_LOC_N	-	-	p=0,4120	p=0,1683
BREF120_MLLR_LOC_E	-	-	-	p=0,9526
BREF120_MLLR_LOC_N+E	-	-	-	-
WER (%)	39,06	30,18	24,28	22,28

(b) *Groupe « voix émues ».*

TABLE 8.4: *WER et p-value du test de Tukey HSD résultants des décodages sur les modèles BREF120 et BREF120 adaptés pour chacun des groupes (hypothèse de différence entre les groupes validée si p<0,05)).*

Pour les voix émues, il existe une différence de WER significative entre le modèle générique *BREF120* et les modèles adaptés à la voix émues *BREF120_MLLR_LOC_E*, avec p<0,05 (p=0,0011), et une différence absolue entre les WER moyens de 14,78%. De même entre les modèle *BREF120* et *BREF120_MLLR_LOC_N+E*, avec p<0,05 (p=0,0002), et une différence absolue de 16,78%. En revanche, il n'y a pas de différence significative entre les autres modèles acoustiques. Ainsi, lors du décodage des phrases émues, nous voyons que l'adaptation au locuteur à partir de phrases neutres n'est pas suffisamment efficace pour améliorer significativement le WER. Il est donc nécessaire d'utiliser des modèles adaptés à la voix émue. L'utilisation de modèles adaptés à partir de phrases aussi bien neutres qu'émues permet d'obtenir un WER similaire au cas où seules les phrases émues sont utilisées pour l'adaptation, les différences de WER entre les modèles *BREF120_MLLR_LOC_E* et *BREF120_MLLR_LOC_N+E* n'étant pas significatives. Aussi, il est intéressant de noter que l'usage de modèles adaptés à la voix émue (*BREF120_MLLR_LOC_*E et *BREF120_MLLR_LOC_N+E*) ne dégradent pas significativement les performances avec la voix neutre.

8.5 Détection des phrases cibles

Nous avons appliqué le filtre décrit en section 6.4.2 sur les hypothèses de sortie de Sphinx3 (modèles *BREF120_MLLR_LOC_N+E*) pour détecter quelles sont les phrases de détresse prononcées (les phrases cibles). Le filtre utilisé contenait les phrases que nous avions demandé de prononcer aux locuteurs du corpus *Voix Détresse*. Toutes les phrases testées appartenant à une seule classe (phrases de détresse), nous n'avons pas pu évaluer la capacité du filtre à discriminer les phrases cibles des autres phrases. Nous avons néanmoins calculé le taux de confusion pour les phrases neutres et pour les phrases émues : pour les phrases prononcées de façon neutre, 2% des phrases ont été mal reconnues par le filtre, et pour les phrases prononcées de façon émue, 9,41% des phrases ont été mal reconnues. Cette différence de 7,41% montre que les émotions ont un impact sur le système de détection des phrases cibles.

8.6 Caractérisation de la voix émue

Nous avons mesuré les valeurs moyennes des paramètres prosodiques étendus à la qualité de la voix à travers les valeurs de *Débit*, *F0*, *Jitter*, *Shimmer* et *HNR* et les avons comparé entre la voix neutre et la voix émue (c'est-à-dire détresse actée). Les données utilisées sont les 739 phrases neutres et émues des locuteurs jeunes et âgés du sous-corpus *test* du corpus *Voix Détresse*. Les mesures par locuteur sont présentées figures 8.4, 8.5, 8.6, 8.7 et 8.8.

Sur ces figures, les locuteurs sont représentés par un identifiant. Par exemple le locuteur *A01F84* est le locuteur du groupe *locuteurs âgés* numéro 01, de sexe féminin, âgé de 84 ans ; ou par exemple le locuteur J02M31 est le locuteur du groupe *locuteurs jeunes* numéro 02, de sexe masculin, âgé de 31 ans.

Nous observons pour la voix émue, par rapport à la voix neutre, en moyenne :

— une diminution du débit,

— une augmentation de la fréquence fondamentale,

— une diminution du jitter,

— une diminution du shimmer,

— une augmentation du rapport harmonique sur bruit.

Les moyennes par variable sur l'ensemble des locuteurs, avec les tests de significativité sur les différences entre voix neutres et voix émues, sont présentées table 8.5. Nous observons que ces différences sont statistiquement significatives pour toutes les variables testées.

Scherer et coll. (2003) comparent les effets de différentes émotions sur leurs paramètres acoustiques par rapport à la voix « normale » (dans des corpus principalement actés). Les auteurs ont synthétisé les résultats des différentes études de la communauté dans ce domaine, présentés figure 8.3. Les émotions étudiées par Scherer et coll. (2003) s'approchant

Acoustic Parameters	Arousal/Stress	Happiness/ Elation	Anger/Rage	Sadness	Fear/Panic	Boredom
Speech Rate and Fluency						
Number of syllables per second	>	>=	<>	<	>	<
Syllable duration	<	<=	<>	>	<	>
Duration of accented vowels	>=	>=	>	>=	<	>=
Number and duration of pauses	<	<	<	>	<>	>
Relative duration of voiced segments			>		<>	
Relative duration of unvoiced segments			<		<>	
Voice Source—F0 and Prosody						
F0 mean[3]	>	>	>	<	>	<=
F0: 5th percentile[3]	>	>	=	<=	>	<=
F0 deviation[3]	>	>	>	<	>	<
F0 range[3]	>	>	>	<	<>	<=
Frequency of accented syllables	>	>=	>	<		
Gradient of F0 rising and falling[3,6]	>	>	>	<	<>	<=
F0 final fall: range and gradient[3,4,7]	>	>	>	<	<>	<=
Voice Source—Vocal Effort and Type of Phonation						
Intensity (dB) mean[5]	>	>=	>	<=		<=
Intensity (dB) deviation[5]	>	>	>	<		<
Gradient of intensity rising and falling[2]	>	>=	>	<		<=
Relative spectral energy in higher bands[4]	>	>	>	<	<>	<=
Spectral slope[1]	<	<	<	>	<>	>
Laryngealization		=	=	>	>	=
Jitter[3]		>=	>=		>	=
Shimmer[3]		>=	>=		>	=
Harmonics/Noise Ratio[1,3]		>	>	<	<	<=
Articulation—Speed and Precision						
Formants—precision of location	?	=	>	<	<=	<=
Formant bandwidth	<		<	>		>=

Notes:
1. depends on phoneme combinations, articulation precision or tension of the vocal tract
2. depends on prosodic features like accent realization, rhythm, etc.
3. depends on speaker-specific factors like age, gender, health, etc.
4. depends on sentence mode
5. depends on microphone distance and amplification
6. for accented segments
7. for final portion of sentences

In specific phonemes, < "smaller," "lower," "slower," "less," "flatter," or "narrower"; = equal to "neutral"; > "bigger," "higher," "faster," "more," "steeper," or "broader"; <= smaller or equal, >=bigger or equal; <> both smaller and bigger have been reported

FIGURE 8.3: *Examen synthétique des résultats empiriques concernant l'effet de l'émotion sur les paramètres vocaux (Scherer et coll., 2003).*

le plus de l'émotion « détresse » que nous étudions pourraient être la peur/panique (*fear/panic*) et l'excitation/stress (*arousal/stress*). Cependant, les paramètres acoustiques que nous avons mesuré (débit, F0, jitter, shimmer et HNR) pour la détresse, en les comparant à la voix « neutre », ne suivent pas totalement les tendances observées par Scherer et coll. (2003). Pour le débit (*number of syllabes per second*), Scherer et coll. (2003) observent une augmentation pour la peur/panique et l'excitation/stress, alors que nous observons une diminution pour la détresse. Pour la F0 moyenne (*F0 mean*), nous observons pour la détresse une augmentation, et Scherer et coll. (2003) observent également une augmentation pour la peur/panique et l'excitation/stress. Pour le jitter et le shimmer (paramètres qui caractérisent certains traits de la qualité de voix), nous observons une diminution pour la détresse, alors que Scherer et coll. (2003) observent une augmentation pour la peur/panique (pas d'observation pour l'excitation/stress). Enfin, concernant le HNR (*harmonics/noise ratio*), nous observons une augmentation pour la détresse, alors que Scherer et coll. (2003) observent au contraire une diminution pour la peur/panique (pas d'observation pour l'excitation/stress).

Les analyses que nous avons faites considéraient que le corpus de détresse était homogène quant à l'expressivité de l'émotion actée (éventuellement avec une intensité expressive variable), et ne variait éventuellement qu'avec le locuteur (différence d'interprétation ou variation idiosyncratique physiologique). Cependant, surpris par ces résultats divergents de la littérature, en réécoutant plus précisément les énoncés, phrase par phrase, nous avons pu supposer que la nature même de la détresse est variable selon la situation suggérée par l'image d'élicitation, et que cette nature pouvait être homogène par situation (c'est-à-dire par phrase) sans variabilité dominante du sujet.

Paramètres	Voix neutre	Voix émue	Test t de Welch
Débit (Phonèmes/s)	13,58	11,71	p<0,05 (p=0,001875)
F0 (Hz)	162,74	260,43	p<0,05 (p=1,161e-06)
Jitter (%)	3,07	2,30	p<0,05 (p=0,0001526)
Shimmer (%)	13,63	9,07	p<0,05 (p=1,41e-07)
HNR (dB)	12,44	14,76	p<0,05 (p=0,0007667)

TABLE 8.5: *Moyennes des paramètres Débit, F0, Jitter, Shimmer et HNR en fonction du type de voix neutre ou ému, et p-values du test t de Welch (différence significative si p<0,05).*

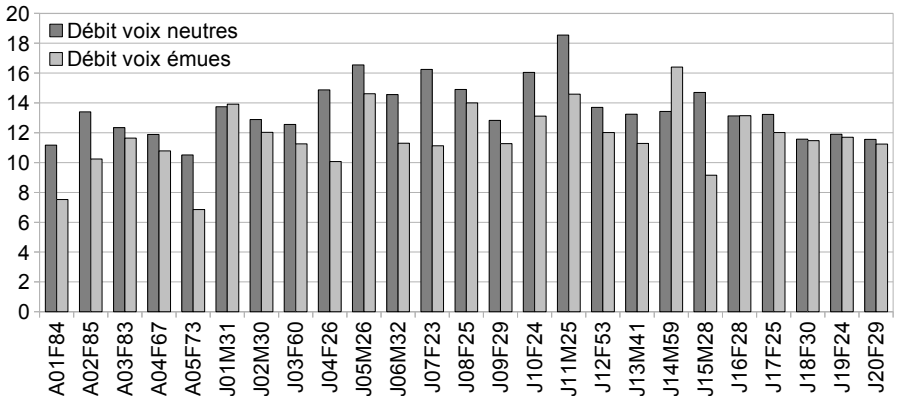


FIGURE 8.4: *Débit moyen par locuteur (nombre de phonèmes par seconde) pour la parole neutre et la parole émue.*

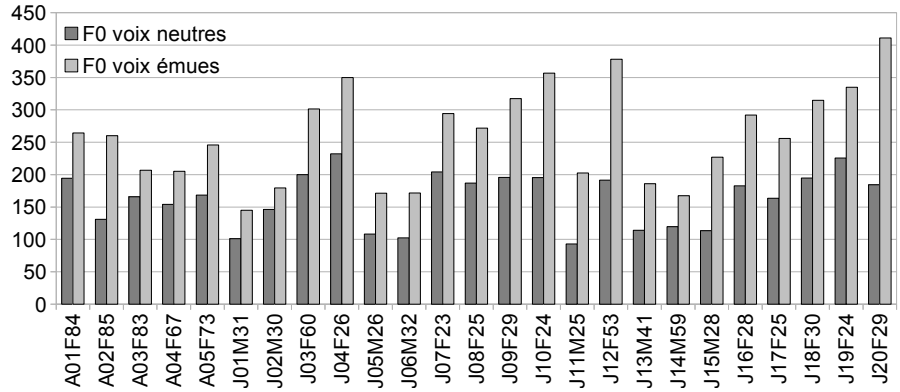


FIGURE 8.5: *F0 moyen par locuteur pour la parole neutre et la parole émue.*

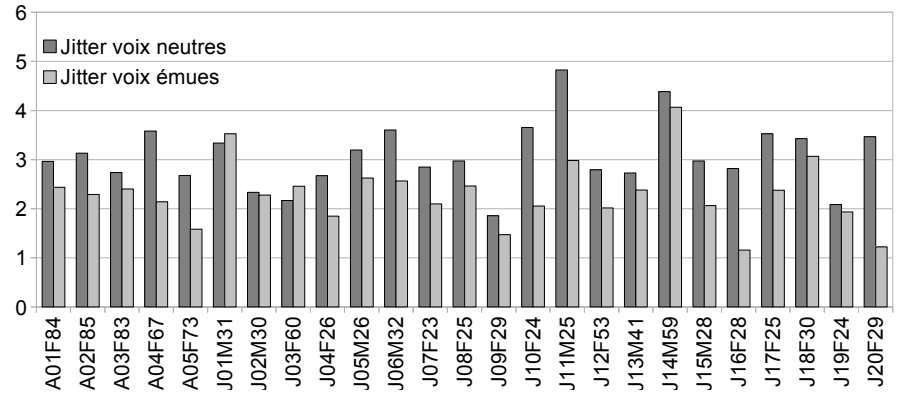


FIGURE 8.6: *Jitter moyen par locuteur pour la parole neutre et la parole émue.*

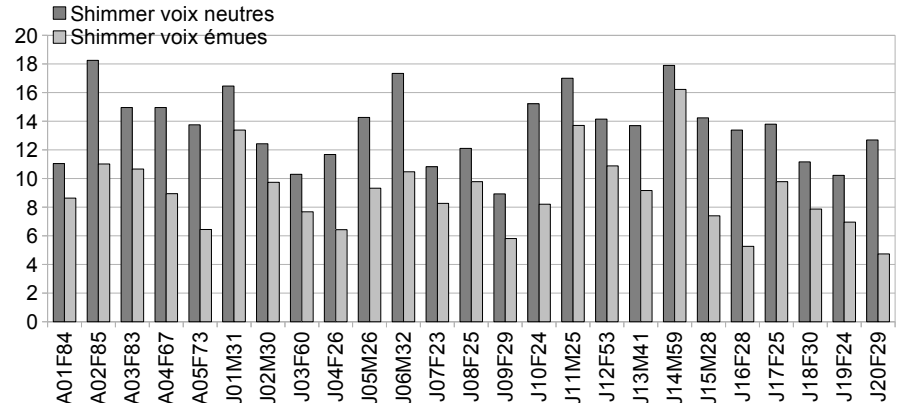


FIGURE 8.7: *Shimmer moyen par locuteur pour la parole neutre et la parole émue.*

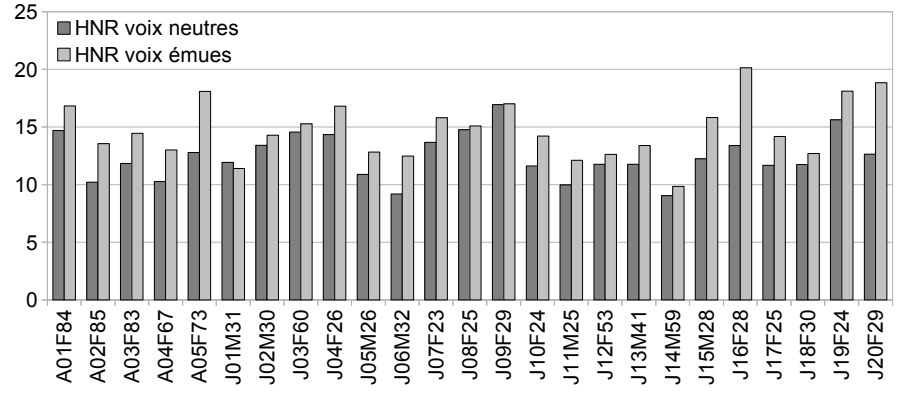


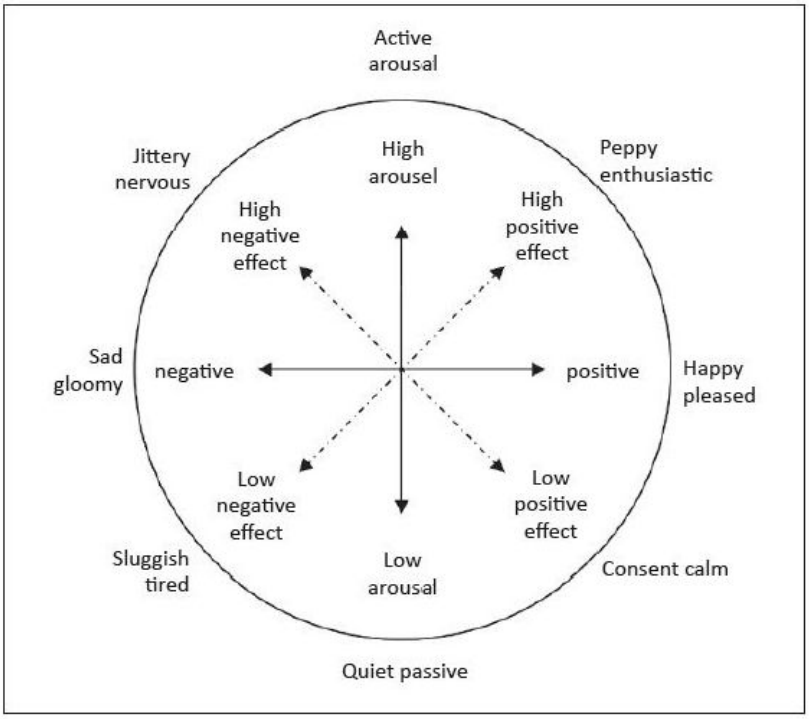
FIGURE 8.8: *HNR moyen par locuteur pour la parole neutre et la parole émue.*

Par exemple, la phrase « Je ne me sens pas bien ! » était associée à une image représentant un homme très abattu se tenant le bras, soutenu par une autre personne, donc très proche physiquement et déjà en train de secourir la personne en détresse. Par contre la phrase « A moi ! » était associée à l'appel d'une personne seule désespérée en train de se noyer : il s'agit là d'une demande de secours d'une personne isolée en danger vital. Les 2 situations ne sont donc pas « dialogiquement » similaires et sont même les 2 extrêmes opposés des situations de notre corpus : personne déjà prise en charge qui s'adresse à son secoureur vs. personne en danger vital qui demande du secours sans interlocuteur défini.

C'est ce que nous avons voulu vérifier en mesurant les paramètres acoustiques non plus par sujet mais par contexte (489 phrases émues, reparties en 10 contextes). Les paramètres jitter, shimmer, F0 et intensité ont été mesurés, et les résultats sont donnés table 8.6. Les images associées aux types de phrases sont données en annexe F. On voit que les 2 situations extrêmes (« A moi ! » et « Je ne me sens pas bien ! ») sont précisément caractérisées, tous locuteurs confondus, par des valeurs extrêmes de F0 moyenne, de jitter et de shimmer (voir table 8.6), ce qui confirmerait notre hypothèse sur la variabilité de la nature de la détresse.

Selon le modèle 2D de Russell (voir figure 8.9), classique en psychologie cognitive, les émotions peuvent être représentées selon 2 axes : l'axe de valence (positif/négatif) et l'axe d'*arousal* (actif/passif).

Ainsi, avec les situations « A moi ! », on serait dans une détresse très active, qu'on pourrait nommer détresse « chaude » (par analogie à la colère chaude proposée par Scherer), alors que les situations « Je ne me sens pas bien ! » seraient plutôt passives, donc une détresse « froide » si on garde l'analogie avec la colère froide de Scherer.



Source: Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

FIGURE 8.9: *Modèle 2D des émotions de Russell.*

Type	Jitter	Shimmer	F0	Intensité
A moi !	1,028%	5,962%	317Hz	70,27dB
Oh la la !	1,367%	6,383%	224Hz	65,72dB
J'ai un malaise !	1,411%	6,778%	232Hz	63,53dB
Aidez-moi !	1,423%	6,801%	279Hz	66,46dB
Du secours s'il vous plaît !	1,550%	6,791%	276Hz	64,80dB
Me laissez pas tout seul !	1,550%	7,131%	276Hz	64,27dB
Qu'est-ce qu'il m'arrive !	1,654%	7,470%	239Hz	62,21dB
Au secours !	1,679%	5,414%	294Hz	69,79dB
Je ne peux plus bouger !	1,708%	8,181%	282Hz	64,72dB
Je ne me sens pas bien !	1,718%	8,372%	220Hz	58,08dB

TABLE 8.6: *Paramètres acoustiques par types de phrases.*

Les contextes « A moi ! » ont été perçues à l'écoute comme étant prononcés de façon la plus tendue, et les contextes « Je ne me sens pas bien ! » ont été perçus comme étant les moins tendus. Les mesures corroborent ces observations : en effet, nous observons table 8.6 que les contextes « A moi ! » ont leur jitter le plus faible et leur F0 la plus haute par rapport aux 9 autres contextes, et les contextes « Je ne me sens pas bien ! » ont leur jitter le plus haut et leur F0 la plus basse.

Nous n'allons pas étudier en détail les autres contextes, mais nous pouvons voir qu'il existe des situations analogues d'expressions qui sont caractérisées par les mêmes valeurs de paramètres, telles que « Du secours s'il vous plaît ! » et « Ne me laissez pas tout seul ! », avec jitter=1,550% et F0=276Hz pour ces 2 contextes. Cependant, nous pouvons voir pour les autres contextes que le jitter et la F0 ne sont pas systématiquement reliés.

8.7 Bilan

Nous avons vu dans ce chapitre qu'il existe une dégradation des performances des systèmes de RAP avec la voix émue. Avec *Google Speech API*, dans notre étude préliminaire, la dégradation est de 31,9% entre la voix neutre et la voix émue. En différenciant la voix émue spontanée de la voix émue actée, le WER est plus important pour la voix émue spontanée, avec une différence de 10,3%.

Avec le corpus *Voix Détresse*, en utilisant le décodeur *Sphinx3*, une dégradation avec la voix émue est également observée, avec une différence de WER de 29,95% pour les locuteurs jeunes et de 19,6% pour les locuteurs âgés entre la voix neutre et la voix émue. Cette dégradation est également observée en décodant les phrases du corpus *Voix Détresse* avec le système *Google Speech API*.

Avec *Sphinx3*, en adaptant le modèle acoustique *BREF120* au locuteur avec des phrases émues qu'ils ont prononcées, nous observons une diminution statistiquement significative du WER (différence de 14,78% entre le modèle générique *BREF120* et les modèles adaptés *BREF120_MLLR_LOC_E*), et pas de diminution significative lorsque l'adaptation au locuteur est réalisée avec des phrases prononcées de façon neutre. Nous voyons donc l'importance d'adapter les modèles acoustiques à partir de données prononcées de façon émue (détresse)

pour une application de détection de phrases de détresse.

Après adaptation, le WER des voix émues avec les modèles *BREF120_MLLR_LOC_E* reste plus important que le WER des voix neutres avec le modèle générique *BREF120*, la différence étant de 13,1%. En outre, le corpus *Voix Détresse* que nous avons enregistré contient de la voix émue actée. En effet, la parole émue spontanée est beaucoup plus difficile à acquérir. En situation réelle de détresse, il faudra donc s'attendre à des résultats encore plus dégradés.

Quant aux caractéristiques acoustiques de la détresse, celles-ci sont subtiles. Nous pouvons retenir que pour une application réelle et robuste dans les situations qui seront rencontrées en habitat intelligent, il faudra être très vigilent sur le contexte physiologique, fonctionnel mais aussi communicatif des énoncés à reconnaître : il faudra donc cumuler suffisamment de données pour rendre le système robuste, mais avant tout que ces dernières soient suffisamment finement caractérisées quant à la nature de leur expressivité.

Dans le chapitre suivant, nous présenterons une expérimentation réalisée en situation réaliste et concernant à la fois la voix âgée et la voix émue de détresse.

CHAPITRE 9

Expérimentation en situation réaliste jouée

Dans ce chapitre, nous allons présenter l'expérimentation conduite en situation réaliste dans l'appartement Domus dans le but de tester le système *CIRDO*. Nous présenterons dans un premier temps le système *CIRDO*, puis nous décrirons l'expérimentation, avec l'élaboration des scénarios, le protocole expérimental et le corpus enregistré. Enfin, nous présenterons les résultats de nos tests de détection des phrases cibles en utilisant le logiciel *CirdoX*.

9.1 Le système CIRDO

Le système *CIRDO* est le système de détection de chute ou d'appel de détresse développé par les partenaires du projet *CIRDO* (Vacher et coll., 2014a). Dans ce contexte, l'équipe SAARA du laboratoire LIRIS de Lyon est en charge du développement du module de traitement vidéo, l'équipe GETALP du LIG est quant à elle en charge du développement du module d'analyse sonore. Le module de fusion de données est développé conjointement par les deux équipes et ne fait pas partie de ce travail de thèse.

L'interconnexion des différents composants du système *CIRDO* est décrite sur la figure 9.1. Le flux de données issu des caméras est analysé en continu par le module vidéo afin d'en déduire les postures et mouvements et détecter les chutes. Le module audio, que nous avons appelé *CirdoX*, analyse quant à lui en permanence le flux de données issu des microphones

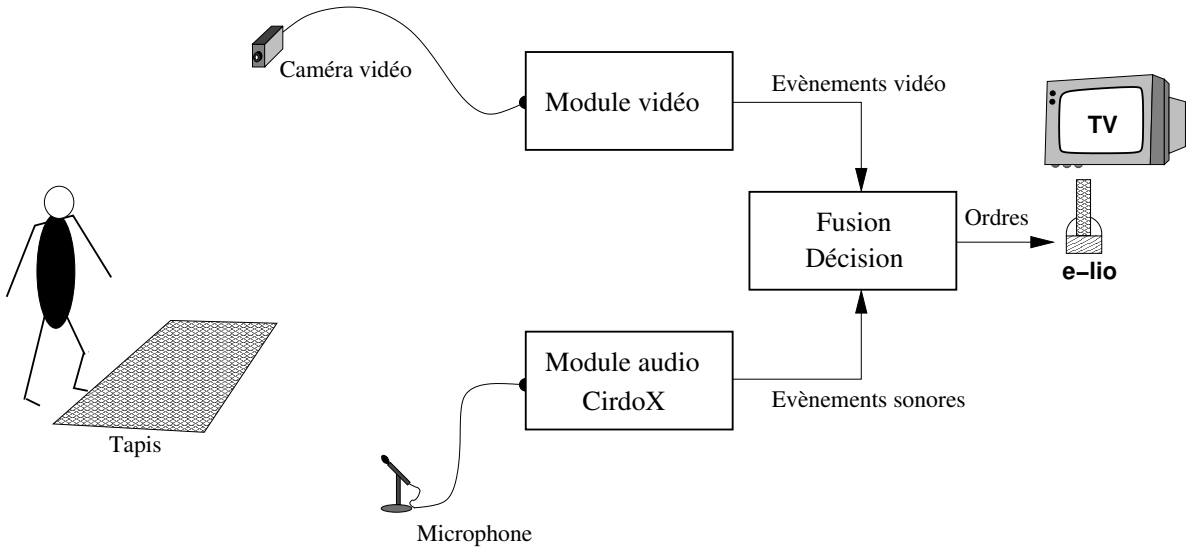


FIGURE 9.1: *Le système CIRDO.*

afin d'en extraire les segments de sons de type parole ou de sons de la vie quotidienne produits au domicile de l'utilisateur. Les résultats de l'analyse de ces deux types d'événements, vidéos et audios, sont envoyés ensuite au module de fusion et décision en charge de détecter les situations à risque qui décide s'il y a lieu d'émettre un appel d'urgence par l'intermédiaire de système *e-lio*[1].

Le système CIRDO est, à l'heure actuelle, en état de prototype installé sur PC. A terme, il est prévu qu'il soit intégré dans *e-lio* par la société Technosens après avoir été industrialisé.

9.1.1 Les fonctionnalités du logiciel CirdoX

Nous avons développé une première version du logiciel *CirdoX* dont le but est, dans le cadre du projet *CIRDO*, la détection d'appels volontaires de la part de personnes âgées en situation de détresse. Cette première version a été améliorée par Duclot (2014) en ce qui concerne la communication entre processus pour le rendre plus efficace lorsque plusieurs microphones sont utilisés, ce qui permettra à terme d'utiliser le logiciel dans un cadre plus général que le projet *CIRDO*.

CirdoX est un logiciel non graphique. Il est écrit en C, du fait de l' « héritage » venant des logiciels précédents *AuditHIS* (Glasson, 2008) et *StreamHIS* développés par l'équipe GETALP, et pour garantir la meilleure performance possible, indispensable pour un logiciel temps-réel. Ce logiciel est conçu pour fonctionner sur les systèmes d'exploitation GNU/Linux. *CirdoX* opère en tâche de fond sans que l'utilisateur n'ait à intervenir pour le faire fonctionner.

Le traitement se décompose en différentes tâches comme illustré sur la figure 9.2 :

- acquisition multicanal du flux audio,
- détection de l'événement sonore dans le bruit de fond,
- discrimination entre parole et sons de la vie courante,
- si une parole est détectée :
  - lancement du système de reconnaissance automatique de la parole prononcée,
  - filtrage des phrases cibles,
- si un son est détecté : classification des sons de la vie courante,
- envoi des données au système de fusion vidéo/audio par le module de mise en forme des données.

Notons que les modèles utilisés par le logiciel peuvent être adaptés à n'importe quel usage nécessitant la reconnaissance d'ordres vocaux. Le logiciel peut donc tout à fait être employé à d'autres usages que ceux prévus par le projet *CIRDO* – moyennant apprentissage de nouveaux modèles –, par exemple pour la reconnaissance d'ordres domotiques, la commande à la voix d'un robot compagnon, etc.

1. http ://www.technosens.fr/

9.1.2 L'architecture du logiciel

9.1.2.1 Les modules et plug-ins

Le logiciel *CirdoX* se veut modulaire grâce à l'utilisation de modules qui sont indépendants entre eux. Chaque module utilise un plug-in pour réaliser une tâche donnée, ce plugin-utilisant une technique particulière. Pour utiliser une autre technique, il devra être remplacé par un autre plug-in spécialement développé. Par exemple, dans le cas du module classification, le plug-in de classification basé sur les GMM pourra si nécessaire être remplacé par un plug-in utilisant les HMM. Au moment de l'installation du logiciel, l'opérateur définit pour chaque module quel plug-in sera affecté à son exécution.

Les paramètres du logiciel (seuils de détection, sélections de plug-ins, fréquence d'échantillonnage, sélection de la carte son, etc.) sont paramétrables à l'aide de fichiers de configuration éditables.

9.1.2.2 Les processus

L'acquisition et la détection d'une part, la discrimination, la classification, le lancement de la RAP et le filtrage d'autre part sont répartis en deux processus différents – *Processus 1* et *Processus 2* (voir figure 9.2) – pour permettre la parallélisation de leur exécution. En effet, pendant le traitement des événements sonores par le *Processus 2*, le système continue en parallèle l'acquisition/détection de nouveaux événements grâce au *Processus 1*.

9.1.2.3 Les événements sonores

Toute phrase prononcée dans la pièce ou tout son émis génère un signal acoustique qui est capté par le microphone. Nous appellerons événement sonore le signal correspondant recueilli par le microphone.

Les événements sonores sont incarnés dans *CirdoX* par un objet descriptif nommé *Sound Object* contenant les informations telles que l'identifiant de l'événement, le nom du fichier

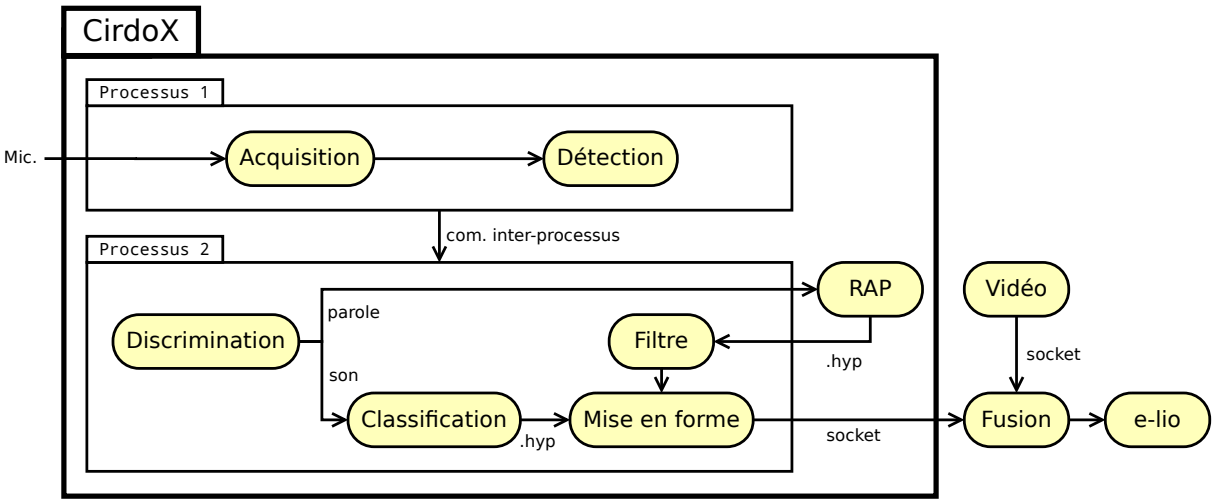


FIGURE 9.2: *Chaîne de traitement audio du logiciel CirdoX.*

associé au format wav, sa durée, le canal enregistré, le rapport signal sur bruit, l'hypothèse de parole ou la classe de son détectée, etc. Le *Sound Object* d'un événement est transmis d'un module à l'autre tout au long du traitement de cet événement, chaque module inscrivant le résultat de son traitement dans le *Sound Object*.

Les événements détectés par le *Processus 1* sont transmis au *Processus 2* par l'intermédiaire d'un *pipe* (tube), un *pipe* étant un système de communication inter-processus d'UNIX. Le *pipe* est d'une capacité suffisamment importante pour éviter son engorgement. Pour éviter leur perte, les événements transmis sont mis dans une file d'attente (une liste chaînée) pour être traités les uns après les autres par le *Processus 2*. Lors de la détection d'événements simultanés, seul l'événement avec le meilleur RSB (Rapport signal sur Bruit) est conservé dans la file d'attente pour être traité par le *Processus 2*.

En fin de traitement, les informations contenues dans le *Sound Object* sont envoyées sur le réseau par le module de mise en forme, sous la forme d'une socket, pour être traité par le module de fusion vidéo/son qui sera exécuté sous la forme d'un web service.

9.1.3 Modules d'acquisition et de détection des événements sonores

9.1.3.1 Module d'acquisition

Quatre plug-ins différents ont été développés pour l'acquisition du signal audio :

— Le plug-in *Portaudio* : l'acquisition audio est réalisée à l'aide de microphones connectés à la carte son du PC en se basant sur la bibliothèque *PortAudio*, qui permet de gérer les entrées et sorties audio de la carte son. *PortAudio* est une librairie audio GNU/Linux, multi-plateforme et open-source. L'acquisition est multicanal pour permettre le traitement simultané d'événements provenant de différents microphones.

— Le plug-in *Kinect* : l'acquisition audio est effectuée à partir du microphone de la Kinect grâce à la librairie *Freenect*, qui permet de récupérer les flux (audio/vidéo) provenant de la Kinect. Cette librairie est libre, multi-plateformes et open-source. Dans ce plugin, l'acquisition est effectuée sur un seul canal audio.

— Le plug-in *Reading wav* : la source du signal est non pas un microphone, mais un fichier wav (mono, 16-bit PCM, 16kHz). Ce plug-in peut être utilisé pour des tests ou à des fins de recherche (travail sur corpus audio). L'acquisition se fait en temps réel, par synchronisation sur la fréquence d'échantillonnage, et le fichier peut être écouté en même temps qu'il est traité. Ce plug-in utilise aussi la bibliothèque *PortAudio*.

— Le plug-in *Fast reading wav* : ce plug-in donne le même résultat que le plug-in *Reading wav*, mais permet d'effectuer le traitement très rapidement (environ 19 fois plus que le temps réel).

9.1.3.2 Module de détection

Dans le contexte spécifique de l'Habitat Intelligent, la parole aussi bien que les sons émis par la personne sont des signaux sporadiques. Il y a donc une grande difficulté à déterminer

le début et la fin d'un événement sonore au milieu du pseudo-silence constitué par le bruit de fond dans l'appartement. Par ailleurs, il s'agit de signaux hétérogènes avec peu de parole, la présence de beaucoup de sons inutiles et de bruits gênants.

Il peut aussi s'agir de signaux mélangés, cas par exemple de la parole en présence de bruit, mais ceci n'est pas traité dans le cas du projet *CIRDO* car cela nécessite des traitements intensifs du signal demandant une puissance de calcul importante incompatible avec un système embarqué à faible coût.

La détection consiste à identifier le début et la fin des événements sonores dans un environnement bruité. Les deux hypothèses de la détection sont :

$$\begin{cases} H_0 : & o(t) = b(t) \\ H_1 : & o(t) = s(t) + b(t) \end{cases} \tag{9.1}$$

où $o(t)$ est le signal analysé, $b(t)$ le bruit et $s(t)$ le signal à détecter. Le principe de base de la détection est l'extraction d'un (ou plusieurs) paramètre(s) du signal d'entrée par une fonction suivie de la comparaison entre la séquence des valeurs obtenues et un seuil.

Comparée à la transformée de Fourier, la transformée en ondelettes est mieux adaptée aux signaux ayant des caractéristiques bien localisées dans l'espace temps-fréquence et permet de développer une méthode efficace de détection (Vacher et coll., 2004).

L'algorithme proposé figure 9.3 a été utilisé pour réaliser la détection. Il repose sur le calcul de la transformée en ondelette discrète (ondelettes de Daubechies) appliquée sur une fenêtre du signal. La détection sera déterminée grâce à un seuil appliqué sur la somme des

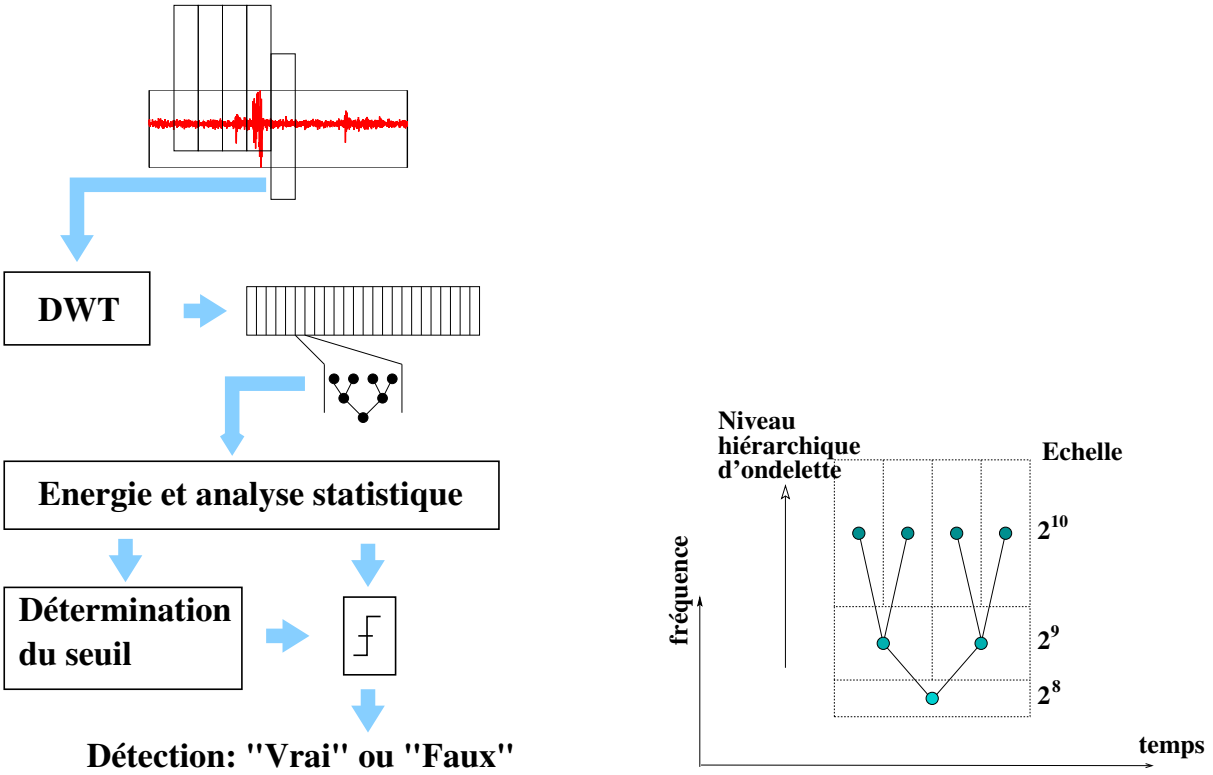


FIGURE 9.3: *Algorithme de détection du signal audio et arbre hiérarchique d'ondelettes pour une trame de 512 échantillons (Vacher et coll., 2004).*

énergies des trois ondelettes les plus hautes de la fenêtre d'analyse. Le seuil S est adaptatif (équation 9.2) et il dépend de la moyenne μ_E de N valeurs de l'énergie (dans notre cas, 40 valeurs sont utilisées). Une phase d'apprentissage est effectuée pour le réglage de α qui est un coefficient expérimental dépendant du niveau du signal acquis.

$$S = \alpha + 1,2\mu_E \tag{9.2}$$

La procédure de détection de la fin du signal débute au moment de la détection du début avec la mémorisation de la valeur du seuil adaptatif. On considère que la fin du signal est détectée à partir de l'instant où l'énergie se trouve en dessous de ce seuil pour un nombre fixé M de fenêtres d'analyse. Pour M, une valeur de 16 trames (0,256 s) a été choisie en vue de s'affranchir des silences existant entre les mots, et de permettre l'utilisation de l'algorithme dans le cas de la parole. La valeur de M a été obtenue après l'étude statistique du corpus de parole française *BRAF100* qui contient l'enregistrement de 100 locuteurs, 10 000 phrases, 20 000 mots et une durée de 28 heures. Dans ce corpus nous avons en moyenne 3 mots/s, avec une durée moyenne d'un mot de 0,33 s donc un silence moyen entre les mots de 0,2 s.

Une présentation complète de l'algorithme ainsi que son évaluation pour la détection de sons dans le bruit sont accessibles dans (Istrate et coll., 2006).

Dans le logiciel CirdoX, à chaque détection, l'identifiant – composé de la date, de l'heure et du canal –, la durée et le rapport signal sur bruit de l'événement est écrit dans le *Sound Object* décrivant celui-ci. Le signal sonore détecté est écrit dans un fichier wav, dont le nom correspond à l'identifiant de l'événement.

9.1.4 Processus de discrimination et filtrage des commandes vocales

9.1.4.1 Module de discrimination son de la vie quotidienne/parole

Une fois les événements sonores détectés, il est nécessaire de les identifier, c'est-à-dire de leur attribuer une signification. Dans la littérature, cette activité est généralement confiée à un classifieur dont le rôle est d'associer une étiquette (par exemple : toux, cri, musique, choc, voiture, etc.) à un événement sonore en entrée. Dans notre étude, nous distinguons deux tâches de classifications.

1. La discrimination son de la vie quotidienne/parole dont le but est de pouvoir extraire les événements de parole afin de les faire parvenir à un système de reconnaissance automatique de la parole.
2. La classification des sons de la vie quotidienne qui traite tous les autres événements n'ayant pas été reconnus comme de la parole.

La première étape avant la classification des événements acoustiques est leur description par un vecteur de paramètres. L'approche consiste à découper l'événement en trames et de calculer des paramètres acoustiques pour chaque trame. Dans les travaux précédents, les coefficients MFCC (*Mel-Frequency Cepstral Coefficients*) ont prouvé leur efficacité pour

la segmentation parole/non-parole (Istrate, 2003). L'utilisation des dérivées premières et secondes permet d'améliorer les performances. Nous avons choisi donc d'utiliser les coefficients MFCC avec les dérivées premières et secondes dans notre système de classification.

D'après (Istrate, 2003) de nombreuses techniques peuvent être utilisées pour la classification des sons, telles que les GMM (*Gaussian Mixture Models*), les HMM (*Hidden Markov Models*), les réseaux de neurones, etc. La méthode GMM permet une bonne performance pour la reconnaissance de locuteurs ou de sons, ainsi que que l'a démontré Reynolds (1995). La méthode HMM est plus complexe, avec des temps de calcul plus longs. Les résultats obtenus par Vacher et coll. (2007) avec des HMM sont similaires à ceux obtenus avec une classification GMM.

Dans le logiciel *CirdoX*, nous avons choisi d'effectuer la classification à l'aide d'un modèle GMM, qui est une méthode éprouvée par l'équipe GETALP (Vacher et coll., 2009b). D'autres plug-in pourraient être envisagés implémentant les autres techniques. Cependant, nous n'avons pas effectué de comparaison entre les différentes méthodes de classification car la classification des sons n'est pas le cœur de notre projet.

Les deux étapes de la classification GMM sont :

— Apprentissage : à partir d'un corpus de sons, pour chaque classe de sons, une étape d'apprentissage est effectuée pour obtenir un modèle contenant les caractéristiques de chaque distribution gaussienne de la classe : le poids de la gaussienne, le vecteur moyen et la matrice de covariance. Ces valeurs sont calculées après M itérations (dans notre cas, M=24) de l'algorithme EM (*Expectation-Maximisation*) (Dempster et coll., 1977).

— Reconnaissance : pour chaque vecteur acoustique, on calcule sa vraisemblance avec le modèle GMM de la classe testée. Ensuite, la moyenne des vraisemblances pour l'ensemble des vecteurs du signal permet d'obtenir un score de vraisemblance entre le signal et le modèle GMM de la classe. L'ensemble de ces opérations est réalisé pour chaque classe de son. Le signal est considéré comme appartenant à la classe pour laquelle le score de vraisemblance est le plus haut.

Dans *CirdoX*, à chaque transmission d'événement depuis le *Processus 1* vers le *Processus 2*, le *Processus 2* lit le fichier wav correspondant à l'événement et exécute le module de discrimination son/parole. Ce dernier calcule des scores de probabilité d'appartenance à la classe *son* ou à la classe *parole* et les probabilités sont écrites dans le *Sound Object*. Si la probabilité est en faveur de la classe *parole*, le traitement de l'événement par le système de RAP est ensuite exécuté, et, dans le cas contraire, si la probabilité est en faveur de la classe *son*, l'événement est envoyé au module de classification des sons.

9.1.4.2 RAP et filtrage

La reconnaissance automatique de la parole est réalisée par un logiciel annexe, *Sphinx3*, qui est géré par un script Tcl/tk lancé en parallèle de *CirdoX*. Le script réalise l'exécution des différentes commandes *Sphinx3* à la demande de *CirdoX*. La synchronisation entre *Cir-*

doX et le script gérant *Sphinx3* se fait grâce à la manipulation de « fichiers de verrous », qui sont créés, lus et détruits mutuellement par le script et *CirdoX*. A chaque événement sonore classifié comme étant de la parole, une copie du fichier wav correspondant à l'événement parole est transmise au script, qui sort alors d'une boucle d'attente et lance l'exécution des commandes *Sphinx3*, les commandes opérant la création des paramètres MFCC et le décodage. *CirdoX* lit alors le fichier d'hypothèse généré par *Sphinx3* et intègre l'hypothèse dans le *Sound Object*.

Un filtre, décrit dans la section 6.4.2, est ensuite appliqué à l'hypothèse pour déterminer s'il s'agit d'une phrase de type *appel de détresse/appel aux aidants*, et le résultat est écrit dans le *Sound Object*.

9.1.5 Module vidéo

Le module vidéo du système *CIRDO* permettant la détection automatique des chutes a été développé par le laboratoire LIRIS de Lyon, partenaire du projet *CIRDO*.

L'acquisition vidéo se fait à l'aide de caméras fixes placées dans l'habitat. A l'instar de *CirdoX*, l'acquisition vidéo se fait en permanence, le module vidéo opérant en tâche de fond, et l'interprétation est réalisée de façon automatique et en temps réel sans l'intervention d'un opérateur. Le fonctionnement de ce module est illustré figure 9.4.

La détection des événements de type chute se base sur l'extraction de la silhouette de la personne, en la séparant de l'arrière plan. La silhouette extraite, segmentée en 6 parties (la tête, le tronc et les quatre membres), est utilisée pour obtenir des caractéristiques de mouvements utiles à la détection d'événements « à risque » et à l'identification des cas de chute.

Le traitement vidéo se décompose en 4 étapes :

— apprentissage de l'arrière-plan,

— extraction de la silhouette,

— analyse de la posture et suivi du mouvement,

— reconnaissance d'une posture de détresse/chute pour envoi au système de fusion vidéo/audio.

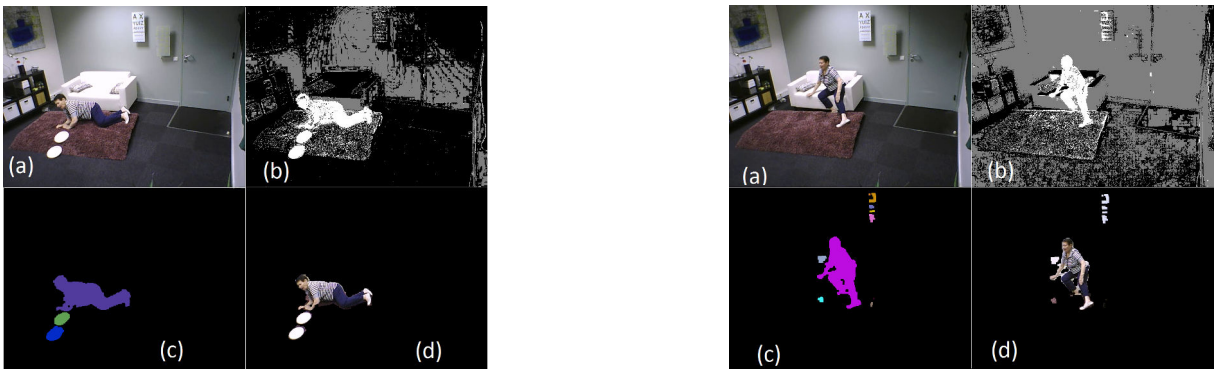


FIGURE 9.4: *Illustration de l'analyse vidéo : détection d'une chute, situation de détresse (difficulté de se lever) a) acquisition initiale, b) discrimination fond/premier plan, c) extraction de la silhouette, d) identification de la situation (Bouakaz et coll., 2014).*

Les scènes qui ne sont pas relatives à une situation de détresse sont effacées.

Certains événements, conditions ou éléments de l'environnement peuvent augmenter les difficultés de traitements et d'interprétation des situation, tels que :

— le changement brusque d'éclairage et de luminosité,

— la proximité des couleurs et des motifs ainsi que le manque de contrastes entre les éléments du décor, et l'habillement du sujet d'intérêt induisant des phénomènes de camouflage,

— la diversité des sources d'éclairage pouvant générer des ombres,

— la diversité des objets présents et dont certains sont susceptibles d'être déplacés.

Les modèles et méthodes mis en place par le LIRIS pour l'apprentissage de l'arrière-plan, l'extraction de la silhouette et la détection des chutes sont décrits dans (Vacher et coll., 2014a).

9.2 Expérimentation dans DOMUS

Une expérience a été menée dans une salle de la plateforme DOMUS du laboratoire LIG, configurée pour ressembler à un petit salon standard (chaises, tapis, table basse,...), dans lequel est installé un système *e-lio* en face d'un canapé. Elle visait à la collecte d'un échantillon significatif de données pour apprendre et tester les modèles utilisés dans le traitement vidéo, et pour tester le système de traitement audio. Il était demandé aux sujets de jouer certains scénarios choisis, définis lors d'une étude de terrain.

9.2.1 Élaboration des scénarios

Le laboratoire GRePS de Lyon, partenaire du projet *CIRDO*, a réalisé une étude de terrain dans l'objectif d'analyser en profondeur les différents cas de chutes afin d'identifier les poses clés (comportement, gestes et poses) et les différentes phrases prononcées lors d'une chute (Bobillier-Chaumon et coll., 2012). A partir de méthodes d'entretien et d'observation, le laboratoire GRePS a construit des scénarios de différentes chutes en décrivant d'une part les conditions de la chute (caractéristiques de la personne, activité réalisée, lieu, moment, causes et circonstances de l'incident...), et, d'autre part, les modalités de la chute : quels sont les différents membres du corps mobilisés, la direction et l'amplitude de chaque mouvement, la vélocité des mouvements, les réactions du sujet au sol, ainsi que les temps d'action ou d'inaction. Des phrases ou mots d'alerte ont également été identifiés à différents moments de la chute : « Aahhh, zuuut, qu'est-ce qu'il m'arrive ! Oh merde, merde... ». Un exemple de script de chute est présenté figure 9.5.

La méthode des *Personas* a ensuite permis au laboratoire GRePS de décrire finement la réalisation et le déroulement des chutes, afin d'avoir une représentation claire et précise des différentes chutes identifiées. Les différentes chutes ont été regroupées dans les 3 grandes classes suivantes :

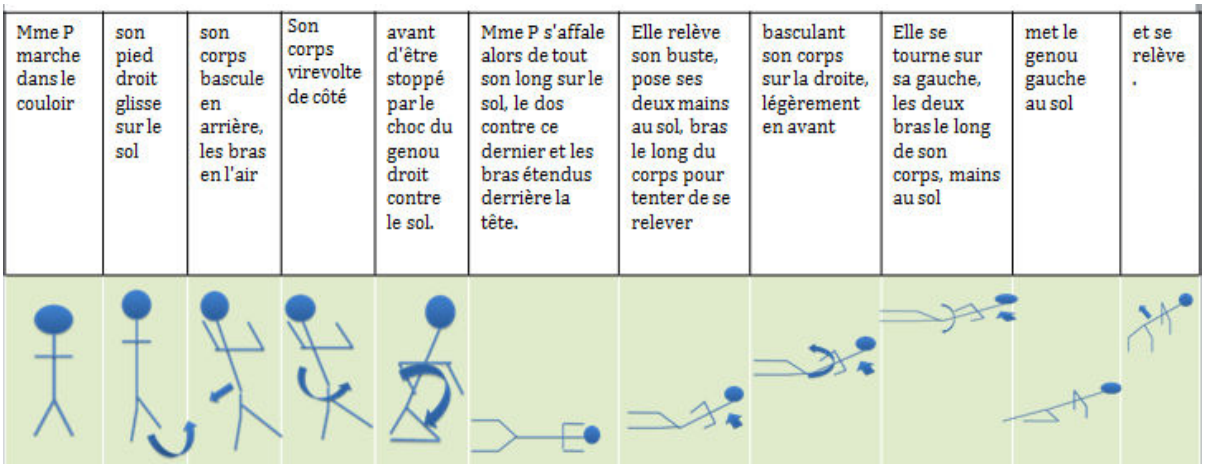


FIGURE 9.5: *Différentes phases d'une chute (Bouakaz et coll., 2014).*

— tomber : s'affaler au sol, depuis une posture statique,

— glisser : déséquilibre du corps dans la continuité d'un déplacement,

— trébucher : perte de verticalité induite par un choc contre un obstacle.

Les résultats de cette étude de terrain ont ensuite permis de définir des scénarios comportant chacun un type de chute identifiée. Ceux-ci ont ensuite été joués dans le contexte d'un expérimentation de type *living lab*. Cette expérimentation, réalisée auprès de 17 sujets, est décrite en section 9.2, et a eu pour objectif de valider le système de détection de chutes.

9.2.2 Protocole expérimental

Chaque participant était introduit au contexte de la recherche et a été invité à signer un formulaire de consentement. Il a ensuite été invité à jouer chacun des différents scénarios. Cinq scénarios permettaient de jouer quatre types de chutes (identifiées de *C1* à *C4*) et une situation où la personne est bloquée (identifiée par *B*), choisis parmi les 28 situations à risque identifiées :

— *C1* : « trébucher contre un tapis »,

— *C2* : « glisser dans le salon »,

— *C3* : « tomber de son canapé »,

— *C4* : « tomber en arrière »,

— *B* : « rester bloqué, assis sur le canapé ».

Les scénarios complets sont donnés en annexe H.

La figure 9.4 montre un participant simulant une chute. Ces situations ont été choisies parce qu'elles étaient représentatives des chutes à la maison et parce qu'elles pouvaient être jouées en toute sécurité par les participants.

Deux autres scénarios (identifiés *F1* et *F2*), appelés « Vrais Négatifs » (VN), ont été ajoutés pour tester la détection automatique des chutes :

Scénario	Phrases
C1	*Merde !* *Qu'est-ce qu'il m'arrive !* *Appelle quelqu'un e-lio !*
C2	*Oh la !* *e-lio, appelle quelqu'un !*
C3	*Oh la !* *Je saigne !* *Je suis blessé !* *e-lio appelle les secours !*
C4	*Aïe, j'ai mal !* *e-lio, appelle ma fille !*
B	*Ah !* *Aïe, j'ai mal !* *e-lio, appelle du secours !* *Je ne peux pas me relever !*
F1	-
F2	-

TABLE 9.1: *Phrases prévues pour chaque scénario de chute et de blocage du corpus Cirdo-Set.*

— *F1* : le premier VN consistait à essayer d'attraper une télécommande sur une table basse lorsque la personne est assise sur le canapé (situation proche de celle dans laquelle la personne a une hanche bloquée),

— *F2* : le second VN consistait à ramasser rapidement un magazine posé sur le sol (proche d'une situation de chute).

Pour les scénarios *C1* à *C4* et le scénarios *B*, les sujets devaient prononcer des phrases de détresse et d'appels aux secours, présentées table 9.1. Les scénarios *F1* et *F2* ne comportaient pas de phrases à prononcer. Les phrases choisies l'ont été en s'appuyant sur une étude du GRePS qui a été présentée en section 9.2.1 et déjà utilisée lors de la constitution du corpus *AD80*.

Avant de jouer une scène, l'expérimentateur précisait le contexte (lieu, moment de la journée...), l'activité (ce que la personne était en train de faire et ce qu'elle allait faire avant que l'incident ne se produise), ainsi que le geste et les mots à produire avant, pendant et après la chute. Ensuite, la personne devait répéter plusieurs fois la scène avant de commencer les enregistrements. Plusieurs prises pouvaient être effectuées si le sujet commettait des erreurs, jusqu'à ce que le scénario soit interprété conformément au script.

9.2.3 Domus et matériel d'enregistrement

La pièce expérimentale de la plateforme DOMUS a été équipée de plusieurs caméras et microphones, et dispose d'une salle technique adjacente (voir figure 9.6). Cette pièce a été configurée de telle façon a ce que la personne assise sur le canapé puisse être en visio-conférence avec un proche par l'intermédiaire d'*e-lio*.

En ce qui concerne l'analyse de l'audio, nous avons utilisé des microphones sans fil Sennheiser : un EW-300-G2 ME2 fixé au plafond de la pièce (« mic 1 » sur la figure 9.6), et un SKM-300-G2 posé sur un meuble à proximité du participant (« mic 2 »). Un enregistrement a été

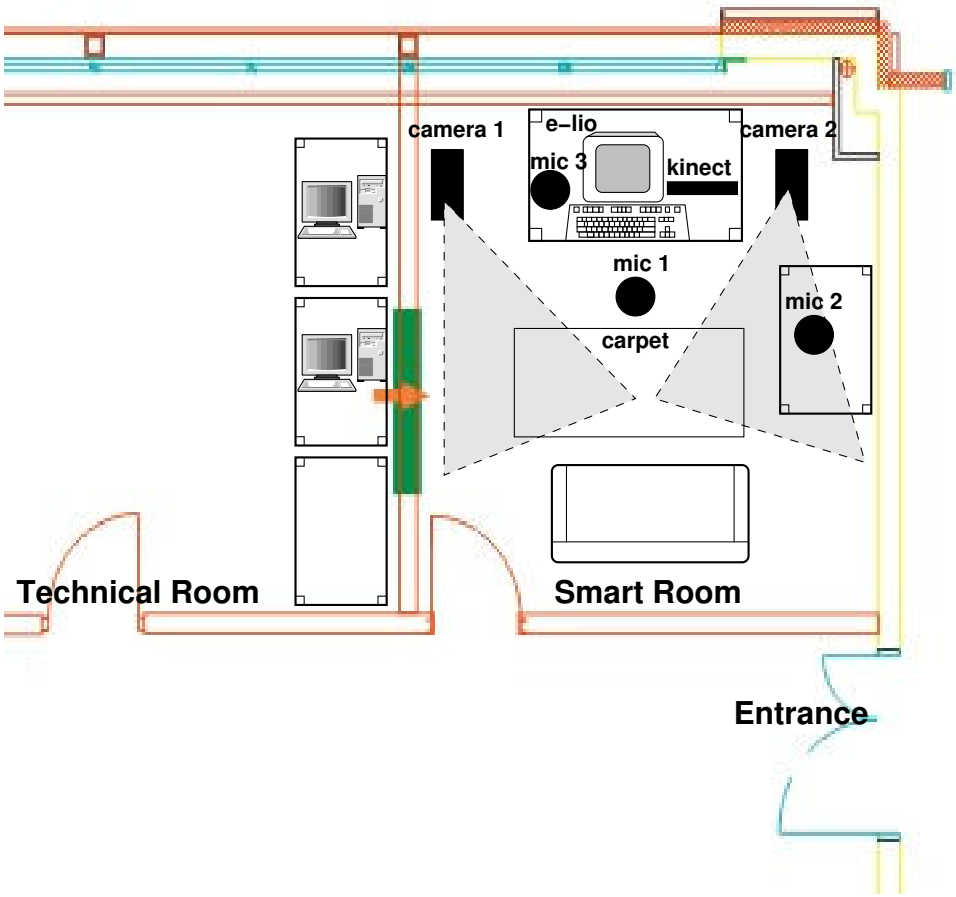


FIGURE 9.6: *Le local d'expérimentation et ses équipements.*

réalisé sur la durée de chaque expérience par un ordinateur PC équipé du logiciel *Stream-HIS* et d'une carte National Instruments PCI-6220 à 8 canaux. Un haut-parleur dans la salle de contrôle permettait aux expérimentateurs de suivre la progression de la scène. Le logiciel *CirdoX* a été installé sur un ordinateur situé dans la salle d'expérimentation et connecté à un microphone unidirectionnel à condensateur Sennheiser (« mic 3 »). Il était d'autre part relié au système *e-lio* placé à côté. En outre, l'acquisition audio était également faite par le microphone d'un système Kinect.

Pour l'analyse vidéo, nous avons utilisé deux caméras de type webcam (Sony PSeye : acquisition en 640x480 60Hz et connectées en USB 2.0) fixées sur le mur face au canapé ainsi qu'une caméra à profondeur (Microsoft Kinect). La synchronisation et l'enregistrement des flux vidéos ont été réalisés sur un ordinateur PC (processeur Intel i3 3,2 GHz, 4 Go Ram, carte graphique NVidia GeForce GTS 450 512Mo). L'affichage des flux vidéos en temps réel sur le pupitre de l'opérateur a permis la validation de chaque scène filmée du point de vue du scénario établi et du jeu des acteurs.

9.2.4 Le corpus enregistré Cirdo-Set

Les participants ciblés étaient des personnes âgées qui étaient encore en mesure de jouer chacun des scénarios élaborés à la section 9.2.1. Le recrutement d'une telle population s'avérant difficile et étant donné que le but était surtout de recueillir des données réalistes en si-

No	Âge	Scénarios (nombres de prises)							Total	Durée
		C1	C2	C3	C4	B	F1	F2		
J01	30	1	2	1	1	1	1	1	8	8min 40s
J02	24	1	1	1	1	1	1	1	7	4min 35s
J03	29	1	1	1	3	1	1	2	10	6min 30s
J04	44	1	1	1	2	1	1	1	8	5min 54s
J05	16	1	1	1	3	2	1	2	11	8min 50s
J06	16	1	1	1	1	1	1	1	7	5min 07s
J07	52	1	1	1	1	1	1	2	8	5min 17s
J08	28	1	1	2	1	1	1	2	9	7min 04s
J09	52	3	1	1	1	1	1	2	10	6min 48s
J10	23	1	1	2	3	1	1	1	10	5min 50s
J11	40	2	1	2	3	1	1	1	11	7min 31s
J12	40	1	1	1	2	2	1	2	10	8min 01s
J13	25	2	2	1	1	1	1	1	9	5min 54s
A01	83	1	1	3	2	3	2	2	14	9min 07s
A02	64	1	2	1	2	1	1	2	10	6min 31s
A03	61	1	1	1	3	1	1	1	9	6min 00s
A04	66	1	1	2	1	2	2	2	11	7min 16s

TABLE 9.2: *Composition du corpus Cirdo-Set.*

tuation, mais pas nécessairement réelles, des personnes de moins de 60 ont aussi participé en portant un équipement (simulateur de vieillissement GERT [2]) permettant de réduire leur mobilité, leur vision et leur audition afin de simuler les conditions dans lesquelles se trouve une personne âgée (voir figure 9.7). Au total, ce sont 13 personnes jeunes (de 16 à 52 ans, 32 ans en moyenne, 7 hommes, 6 femmes) et 4 personnes âgées (de 61 à 83 ans, 68,5 ans en moyenne, 2 hommes, 2 femmes) qui ont participé.

Le détail du corpus résultant, *Cirdo-Set*, est donné dans la table 9.2. *C1* à *C4* sont les scénarios de chute et *B* celui de la hanche bloquée, tandis que *F1* et *F2* représentent les « Vrais Négatifs ». La durée totale des enregistrements est égale à 1 heure 55 minutes, avec un total de 162 scénarios joués.

L'évaluation du logiciel *CirdoX* sur le corpus *Cirdo-Set* a été effectuée *off-line*, à partir des fichiers audios. Nous avons utilisé les enregistrements sonores acquis durant le jeu des différents scénarios par les 17 sujets. Pour les scénarios joués plusieurs fois, nous n'avons gardé que leur dernière prise. L'évaluation a été ainsi effectuée sur 119 scénarios (7 scénarios par sujet), soit 1 heure et 26 minutes d'enregistrements.

9.3 Adaptation et évaluation de CirdoX

9.3.1 Nécessaires adaptations à la tâche

De nouveaux modèles GMM de discrimination son/parole ont été appris à l'aide de la librairie *ALIZE* (Bonastre et coll., 2005) (format *SPRO4*) à partir de l'ensemble des données de type *son de la vie quotidienne* et *parole* du corpus *Sweet-Home* (Vacher et coll., 2014b), constitué d'enregistrements sonores effectués en conditions distantes. 9147 fichiers de pa-

2. www.simulateur-du-vieillissement.com/

FIGURE 9.7: *Sujet portant la combinaison de simulation de vieillesse (Bouakaz et coll., 2014).*

role (chaque fichier contenant une phrase de type ordre domotique ou appel de détresse), soit 4 heures et 53 minutes d'enregistrement, ont été utilisés pour l'apprentissage du modèle GMM *parole*, et 13448 fichiers de sons, soit 5 heures et 10 minutes d'enregistrement, ont été utilisés pour l'apprentissage du modèle GMM *son*.

A l'aide du module de *CirdoX* de discrimination son/parole intégrant ces nouveaux modèles GMM appris, une classification automatique dans la classe *son* ou la classe *parole* a été réalisée sur les 1950 événements sonores détectés par le module de détection (voir section 9.3.2.1), et les événements classés comme étant de la parole ont été présentés au module de *CirdoX* de reconnaissance automatique de la parole se basant sur *Sphinx3* (voir section 9.3.2.2).

Pour le modèle de langage de *Sphinx3*, nous avons utilisé un modèle général unigramme contenant 13304 mots appris à partir du corpus *Gigaword*[3], que nous avons interpolé avec un modèle de langage spécifique trigramme, contenant 99 mots, appris sur les phrases de détresse du corpus *AD80* et sur les phrases du corpus *Cirdo-Set* présentées table 9.1. Le mo-

3. http ://catalog.ldc.upenn.edu/LDC2006T17

dèle final, issu de l'interpolation entre le modèle général (poids de 10%) et le modèle spécifique (poids de 90%) possède 13316 unigrammes, 225 bigrammes et 273 trigrammes.

A partir du modèle acoustique générique *BREF120*, nous avons réalisé différentes adaptations avec la méthode MLLR pour obtenir les modèles acoustiques suivants :

— *BREF120_EMOTION_G* : adaptation MLLR globale du modèle *BREF120* à la voix émue (détresse) à partir de l'ensemble des phrases émues (1221 phrases, soit 20 minutes et 30 secondes d'enregistrement) du corpus Voix Détresse présenté en section 5.4,

— *BREF120_SWEETHOME_G* : adaptation MLLR globale du modèle *BREF120* à la voix âgée et aux conditions de parole distante dans un appartement. Les données utilisées pour cette adaptation sont des phrases domotiques et des phrases de détresse extraites du sous-ensemble *User specific interaction* du corpus *Sweet-Home* (Vacher et coll., 2014b) enregistré dans l'appartement de test Domus, soit 337 phrases représentant 9 min 30 s d'enregistrements, prononcées par des locuteurs âgés ou mal-voyants âgés de 49 à 91 ans (moyenne d'âge : 72 ans),

— *BREF120_SWEETHOME+EMOTION_G* : au lieu de partir du modèle *BREF120*, nous partons du modèle *BREF120_SWEETHOME_G* que nous adaptons à la voix émue (même adaptation que pour le modèle *BREF120_EMOTION_G*).

9.3.2 Évaluation de CirdoX

9.3.2.1 Résultats de la détection et de la discrimination automatique

La figure 9.8 présente le schéma global de l'évaluation de CirdoX. Les 119 fichiers wav correspondant aux enregistrements sonores des 119 scénarios joués par les sujets du corpus *Cirdo-Set* ont été présentés en entrée du module d'acquisition du système *CirdoX* (plug-in *Fast reading wav*). La détection des événements sonores dans les scénarios, réalisée par le module de détection de *CirdoX* (seuil statique de détection fixé à 10^{-7}), nous a permis le découpage automatique des événements sonores. 1950 événements sonores ont ainsi été détectés, soit 25 minutes et 28 secondes d'enregistrements. La totalité des 303 phrases prononcées par les locuteurs ont donné lieu à la création d'un (ou plusieurs) événements par le module de détection de *CirdoX*. Certaines phrases ont été coupées : 26 ont eu le début ou la fin coupés à cause du niveau sonore de la voix trop faible, et 19 ont été coupées en 2 suite à une pause entre 2 mots (par exemple, la phrase « *e-lio* appelle quelqu'un ! » a donné lieu aux événements sonores « *e-lio* » et « appelle quelqu'un », du fait d'une trop longue pause après l'énonciation de « *e-lio* »).

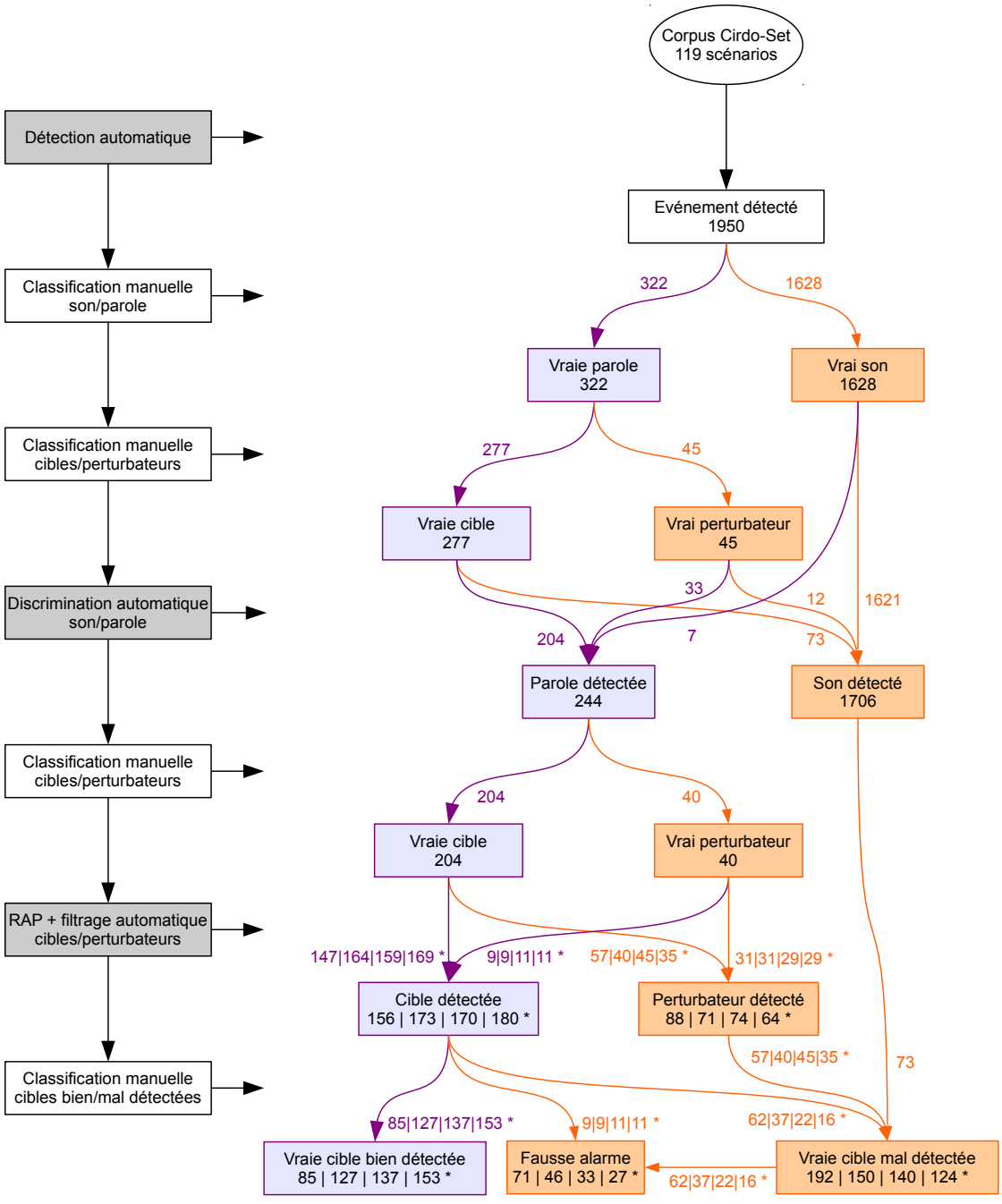
Par commodité de langage, nous conviendrons de désigner maintenant les sons de la vie quotidienne par *son* et les sons de type parole par *parole*.

Nous avons classifié manuellement dans les classes *parole* ou *son* les 1950 événements sonores détectés par le module de détection de *CirdoX* (voir figure 9.8). 322 événements ont été classifiés dans la classe *parole* (soit 16,51% des événements), et les 1628 restants ont été

classifiés dans la classe *son* (soit 83,49%). Nous avons ensuite classé manuellement les événements en deux catégories :

— catégorie *phrase cible* : contient les phrases présentées table 9.1, que les participants devaient prononcer durant le jeu des scénarios,

— catégorie *perturbateur* : correspond aux sons détectés comme de la parole et à toutes les phrases prononcées spontanément par les sujets et les expérimentateurs au cours des enregistrements des scénarios, mais ne faisant pas partie de ce qu'il était prévu de prononcer.



* Modèles acoustiques BREF120 | BREF120_EMOTION_G | BREF120_SWEETHOME_G | BREF120_SWEETHOME+EMOTION_G

FIGURE 9.8: *Schéma global de l'évaluation de CirdoX, affichant le nombre d'événements traités à chaque étape.*

	Parole	Son	<– classifié comme
Parole	$VP = 237$	$FN = 85$	
Son	$FP = 7$	$VN = 1621$	

TABLE 9.3: *Matrice de confusion de la classification GMM des événements sonores dans les classes « son » et « parole ».*

Parmi les 322 événements de type *paroles*, 277 phrases appartiennent à la catégorie *phrases cibles* (86%), et les 45 restantes appartiennent à la catégorie *perturbateurs* (14%).

Les résultats de la discrimination automatique dans les classes *son* et *parole* par un classifieur GMM des 1950 événements sonores sont donnés dans la matrice de confusion présentée table 9.3. 244 événements ont été classifiés dans la classe *parole*, et 1706 événements ont été classifiés dans la classe *son*.

Comme illustré sur la figure 9.8 et table 9.3, 244 événements ont été détectés comme appartenant à la classe *parole*, dont, en réalité, 7 événements sont des sons et 237 sont de la parole (204 phrases cibles et 33 perturbateurs). 1706 événements ont été détectés comme étant de la classe *son*, dont, en réalité, 1621 sont des sons et 85 sont de la parole (73 phrases cibles et 12 perturbateurs).

La sensibilité, représentant le taux d'événements *parole* classifiés dans la classe *parole*, est de 73,6%, ce qui représente une valeur assez faible. En revanche, la spécificité, représentant le taux d'événements *son* classifiés dans la classe *son*, est très haute, 99,6%.

9.3.2.2 Décodage avec Sphinx3

Les décodages avec *Sphinx3* ont été réalisés à partir du modèle générique *BREF120*, des modèles adaptés *BREF120_SWEETHOME_G*, *BREF120_EMOTION_G* et *BREF120_SWEET-HOME+EMOTION_G*, et du modèle de langage décrits en section 9.3.1.

Les données utilisées sont les 244 événements sonores classifiés par le classifieur GMM comme étant de type *parole*. Étant donné que le système de discrimination n'est pas parfait, 7 sons ont été décodés par Sphinx3, et 85 phrases n'ont pas été décodées (73 phrases de type *phrase cible* et 12 phrases de type *perturbateur*). Parmi les 244 événements sonores considérés, nous avons identifié manuellement que 204 sont des phrases de la catégorie *phrase cible*, et 40 appartiennent à la catégorie *perturbateur* dont 7 sont des sons et 33 sont des phrases non prévues dans les scénarios (voir figure 9.8).

Les WER résultant des décodages des catégories *phrase cible* et *perturbateur* avec les différents modèles acoustiques sont donnés table 9.4.

	BREF120	BREF120_ EMOTION_G	BREF120_ SWEETHOME_G	BREF120_ SWEETHOME+EMOTION_G
WER phrases cibles	80,46%	61,07%	49,32%	38,52%
WER pertubateurs	107,14%	113,27%	102,04%	115,31%

TABLE 9.4: *WER pour les phrases cibles et les perturbateurs pour chaque modèle acoustique.*

Le contenu des 33 phrases appartenant à la catégorie *perturbateur* étant très éloigné des phrases utilisées dans le modèle de langage, le WER pour la catégorie *perturbateur* est très élevé (supérieur à 100%) quel que soit le modèle acoustique utilisé. Ceci correspond à l'effet recherché, ces phrases ne devant pas être reconnues pour garantir l'intimité de la personne. De plus, les 7 sons décodés participent à la dégradation du WER.

Pour les 204 phrases de la catégorie *phrase cible*, le WER avec le modèle générique *BREF120* est assez élevé, 80,46%. Nous émettons l'hypothèse que cela est dû aux conditions d'enregistrement différentes de celles du corpus *BREF120*, qui est un corpus de parole lue prononcé de façon neutre, et les locuteurs étant proches du microphone. En effet, les phrases du corpus *Cirdo-Set* ont été prononcées dans l'appartement de test et enregistrées avec des microphones placés au plafond donc éloignés du locuteur. De plus, les phrases du corpus *Cirdo-Set* sont des phrases de détresse prononcées avec de l'émotion, les personnes étant placées dans des situations inhabituelles (chutes jouées), et certains locuteurs étaient des personnes âgées.

En adaptant le modèle acoustique par la méthode MLLR avec le corpus *Sweet-Home*, qui est un corpus enregistré dans des conditions similaires au corpus *Cirdo-Set* (mêmes microphones et parole distante), nous observons avec le modèle *BREF120_SWEETHOME_G* une amélioration de 31,14% (différence absolue). Aussi, le test avec le modèle adapté à la voix émue *BREF120_EMOTION_G*, qui, de surcroît, est un modèle dont les données d'adaptation proviennent d'un corpus (le corpus *Voix Détresse*) enregistré dans la même pièce que le corpus *Cirdo-Set*, permet une amélioration de 19,39% par rapport au modèle *BREF120*. Avec la combinaison de l'adaptation aux conditions de l'appartement et à la voix émue (modèle *BREF120_SWEETHOME+EMOTION_G*), nous obtenons un WER de 38,52%, soit une amélioration de 41,94% par rapport au modèle *BREF120* initial.

9.3.2.3 Détection des phrases cibles

Le système *CirdoX* applique un filtre sur les hypothèses de sortie de *Sphinx3* (voir figure 9.2). Comme décrit en section 6.4.2, le filtre calcule la distance de Levenshtein entre l'hypothèse de *Sphinx3* phonétisée et la liste des phrases cibles phonétisées à reconnaître. Cette liste est constituée des phrases présentées dans la table 9.1. La distance est normalisée par le nombre de phonèmes de l'hypothèse, et un seuil est appliqué sur la distance normalisée : si cette dernière est inférieure au seuil, l'hypothèse est considérée comme appartenant à la catégorie *phrase cible*, et si la distance normalisée est supérieure au seuil, l'hypothèse est considérée comme appartenant à la catégorie *perturbateur*.

Nous avons appliqué le filtre en sortie du système de RAP pour chacun des modèles acoustiques *BREF120*, *BREF120_SWEETHOME_G*, *BREF120_EMOTION_G* et *BREF120_SWEETHOME+EMOTION_G*.

Les courbes ROC, présentées figure 9.9 pour chaque modèle acoustique, nous ont permis de fixer les seuils présentés table 9.5. Nous observons que, pour un modèle acoustique donné, plus le WER des phrases cibles est élevé, plus la valeur du seuil trouvée à l'intersection entre la courbe ROC et la diagonale est haute. Les seuils ont été fixés à une valeur proche

de l'intersection de la courbe ROC avec la diagonale.

La table 9.5 présente les matrices de confusion des classifications par le filtre pour les différents modèles acoustiques. Pour les événements classés manuellement comme appartenant à la catégorie *phrase cible*, soit 204 événements, les vrais positifs (VP) sont les événements dont la distance normalisée est au-dessous du seuil, et les faux négatifs (FN) sont ceux dont la distance est au-dessus du seuil. Pour les événements classés manuellement dans la catégorie *perturbateur*, soit 40 événements (7 bruits et 33 paroles), les faux positifs (FP) sont les événements dont la distance normalisée est au-dessous du seuil, et les vrais négatifs (VN) sont ceux au-dessus du seuil. La figure 9.8 illustre les résultats en terme de nombre de phrases cibles et de perturbateurs détectés par le filtre pour les 4 modèles acoustiques.

La table 9.6 présente la sensibilité, la spécificité, le taux de fausses alarmes ainsi que le WER des phrases de type *phrase cible* en fonction de chaque modèle acoustique pour les seuils donnés table 9.5. Dans le cas de notre expérimentation :

— la sensibilité (Se) représente le taux des événements identifiés manuellement comme étant de la catégorie *phrase cible* et qui sont ensuite classés par le filtre comme étant de la catégorie *phrase cible*,

seuil = 1,364	d < seuil	d >= seuil
Phrases cibles	$VP = 147$	$FN = 57$
Perturbateurs	$FP = 9$	$VN = 31$

(a) *Modèle acoustique BREF120.*

seuil = 1,200	d < seuil	d >= seuil
Phrases cibles	$VP = 164$	$FN = 40$
Perturbateurs	$FP = 9$	$VN = 31$

(b) *Modèle acoustique BREF120_EMOTION_G.*

seuil = 1,125	d < seuil	d >= seuil
Phrases cibles	$VP = 159$	$FN = 45$
Perturbateurs	$FP = 11$	$VN = 29$

(c) *Modèle acoustique BREF120_SWEETHOME_G.*

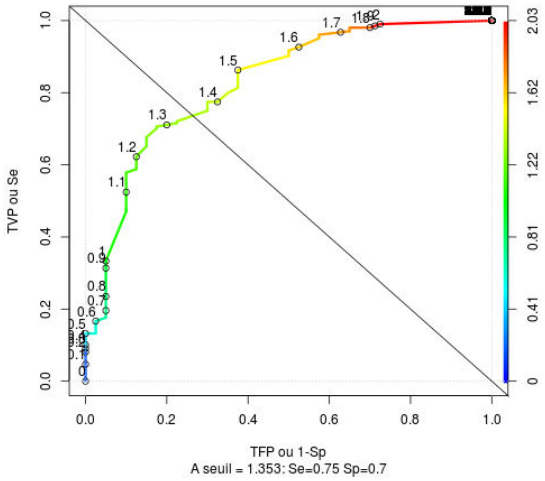
seuil = 1,000	d < seuil	d >= seuil
Phrases cibles	$VP = 169$	$FN = 35$
Perturbateurs	$FP = 11$	$VN = 29$

(d) *Modèle acoustique BREF120_SWEETHOME+EMOTION_G.*

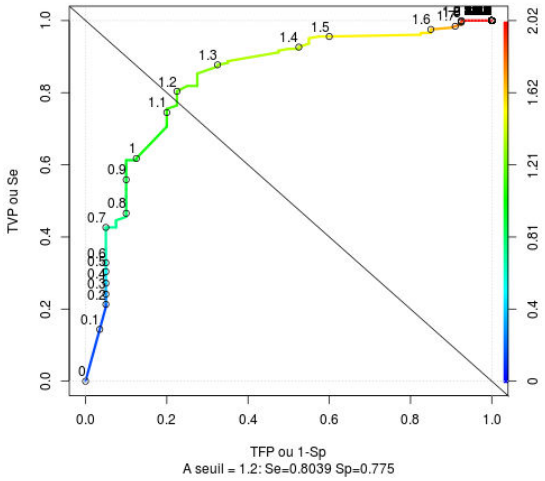
TABLE 9.5: *Matrice de confusion du filtrage des phrases cibles pour les différentes modèles acoustiques après discrimination son/parole automatique.*

	BREF120	BREF120- _EMOTION_G	BREF120- _SWEETHOME_G	BREF120- _SWEETHOME+EMOTION_G
Se	72,06%	80,39%	77,94%	82,84%
Sp	77,80%	77,50%	72,50%	72,50%
TFA	5,77%	5,20%	6,47%	6,11%
WER	80,46%	61,07%	49,32%	38,52%

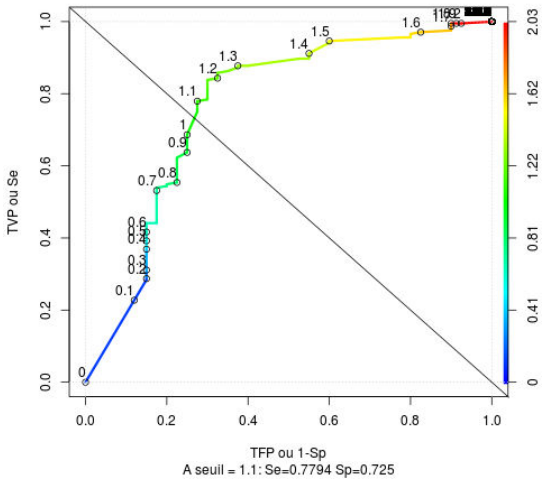
TABLE 9.6: *Sensibilité, spécificité, taux de fausses alarmes et WER des phrases cibles en fonction des modèles acoustiques.*



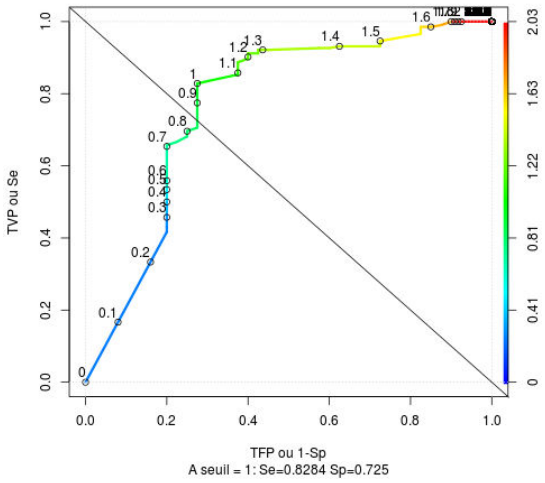
(a) Modèle BREF120



(b) Modèle BREF120_EMOTION_G



(c) Modèle BREF120_SWEETHOME_G



(d) Modèle BREF120_SWEETHOME+EMOTION_G

FIGURE 9.9: *Courbes ROC représentant le TVP en fonction du TFP pour le filtrage des phrases cibles en fonction des modèles acoustiques après discrimination son/parole automatique.*

— la spécificité (Sp) représente le taux des événements identifiés manuellement comme étant de la catégorie *perturbateur* classés par le filtre comme étant de la catégorie *perturbateur*,

— le taux de fausses alarmes (TFA) représente le taux d'événements classés par le filtre comme étant de la catégorie *phrase cible* alors qu'ils sont identifiés manuellement comme étant de la catégorie *perturbateur*.

Nous observons que les aires sous les courbes ROC sont très proches pour les 3 modèles adaptés *BREF120_SWEETHOME_G*, *BREF120_EMOTION_G* et *BREF120_SWEETHOME+EMOTION_G*, et donc que les points d'égale erreur (le point d'égal erreur est l'intersection entre la courbe ROC et la diagonale) sont équivalents pour ces 3 modèles adaptés. Aux seuils que nous avons fixés, les sensibilités sont comprises entre 77,9 et 82,8%, et les spécificités sont comprises entre 72,5 et 77,5% pour les modèles adaptés. Nous observons donc que la variation du WER des phrases cibles a peu d'influence sur la sensibilité et la spécifi-

	BREF120	BREF120- _EMOTION_G	BREF120- _SWEETHOME_G	BREF120- _SWEETHOME+EMOTION_G
TBR	41,67% (85)	62,25% (127)	67,16% (137)	75,00% (153)
WER	80,46%	61,07%	49,32%	38,52%

TABLE 9.7: *TBR (avec le nombre de phrases correspondant) et WER en fonction des modèles acoustiques.*

cité. Pour le modèle *BREF120*, l'aire sous la courbe ROC est plus faible que pour les modèles adaptés, Se = 75% et Sp = 70%. Ainsi, pour le meilleur modèle acoustique (*BREF120_SWEET-HOME+EMOTION_G*), 82,84% des phrases de type *phrase cible* sont classées par le filtre comme étant *phrase cible*, et 72,50% des phrases de type *perturbateur* sont classées par le filtre comme étant *perturbateur*.

Une phrase identifiée manuellement comme appartenant à la catégorie *phrase cible* et classée par le filtre comme faisant partie de la catégorie *phrase cible* (distance sous le seuil) ne garantit pas pour autant que la phrase soit bien reconnue par le filtre. Ainsi, il y a une « confusion » lorsque le filtre commet une erreur dans la sélection d'une phrase (par exemple si le filtre sélectionne « *e-lio*, appelle quelqu'un ! » au lieu de « Aïe, j'ai mal ! »).

Nous avons donc calculé, par chaque modèle acoustique, le taux de bonne reconnaissance (TBR), qui représente le taux de phrases cibles (taux calculé sur les 204 phrases cibles totales) qui sont filtrées sans confusion. Dans la table 9.7, le TBR, avec le nombre de phrases correspondant entre parenthèses, et le WER des phrases cibles sont donnés pour chaque modèle acoustique.

Nous observons que les taux de bonne reconnaissance sont dépendants du WER : le TBR augmente lorsque le WER diminue, d'où l'importance du soin apporté à l'étape d'adaptation des modèles acoustiques. Pour le meilleur modèle acoustique (*BREF120_SWEET-HOME+EMOTION_G*), le taux de bonne reconnaissance est de 75% : 75% des 204 événements de type *phrase cible* (soit 153 événements) sont reconnus par le filtre comme appartenant à la catégorie *phrase cible* sans confusion.

9.3.2.4 Performances globales du système

Au final, pour le modèle *BREF120_SWEETHOME+EMOTION_G*, en suivant la chaîne de traitement *CirdoX* (détection, discrimination son/parole, RAP et filtrage des phrases cibles), et en prenant comme référence l'ensemble des événements identifiés manuellement comme appartenant à la catégorie *phrase cible*, soit 277 phrases, 204 phrases cibles sont classifiées par le classifieur GMM comme appartenant à la classe *parole* (soit 73,6%). Parmi ces 204 phrases, 169 sont identifiées par le filtre comme appartenant à la catégorie *phrase cible* (61,01%), dont 153 sont identifiées par le filtre comme appartenant à la catégorie *phrase cible* sans confusion (55,23%).

Ainsi, comme illustré figure 9.8, pour le modèle *BREF120_SWEETHOME+EMOTION_G*, 73 phrases de type *phrase cible*, soit 26,3% des phrases cibles, sont mal classifiées par la discrimination son/parole. Au total, 124 phrases de type *phrase cible*, soit 44,7% des phrases

cibles, ne sont pas ou sont mal détectées par le système : 73 phrases cibles sont classifiées par la discrimination son/parole comme étant de la classe *sons*, 35 phrases cibles sont filtrées comme étant de la catégorie *perturbateur*, et 16 phrases cibles sont filtrées comme étant de la catégorie *phrase cible* mais avec confusion. 27 événements sonores donnent lieu à des fausses alarmes : 16 phrases cibles sont détectées par le filtre mais sont mal reconnues (confusion), et 11 perturbateurs sont filtrés comme étant des phrases cibles.

Nous allons maintenant réaliser une analyse plus fine pour essayer d'identifier les principales raisons de non repérage des phrases cibles.

9.3.3 Analyse par module de la non détection des phrases cibles

9.3.3.1 Discrimination son/parole

Dans l'étape de la classification automatique son/parole, 85 événements sur les 322 de type *parole* sont classifiés de façon erronée comme appartenant à la classe *son*. Un facteur pouvant expliquer cette mauvaise classification est la présence de bruits dans la parole.

Ainsi, nous avons trouvé que 95 événements sur les 322 de type *parole* sont bruités, notamment du fait de la superposition de bruits (pour la plupart liés à la chute du sujet) sur la voix prononçant les appels de détresse. La table 9.8 compare les résultats de la classification son/parole pour l'ensemble des phrases, pour les phrases bruitées et pour les phrases non bruitées.

Pour les événements *parole* non bruités, 86,3% d'entre eux sont bien classifiés dans la classe *parole*. En revanche, pour les événements *parole* bruités, le taux des événements bien classifiés est beaucoup plus bas, il tombe à 43,2%.

Parmi les 85 événements *parole* classés comme étant des sons, 54 sont bruités. Ainsi, la présence de bruit dans la parole explique une bonne partie (63,5%) de la mauvaise classification.

9.3.3.2 Décodage avec Sphinx3

Suite à la discrimination son/parole automatique, un certain nombre d'événements de type *parole*, classés comme *son*, sont écartés du décodage. Dans cette section nous allons voir quels auraient été les WER obtenus dans le cas d'une discrimination parfaite, c'est-à-dire si 100% des événements de type *parole* avaient été correctement classifiés dans la classe *parole*.

Nous avons vu section 9.3.2.1 que 322 événements appartiennent à la classe *parole* et 1628 à la classe *son* (classification manuelle).

Parole	Classé parole	Classé son
Total : 322	237 (73,6%)	85 (26,4%)
Non bruité : 227	196 (86,3%)	31 (13,7%)
Bruité : 95	41 (43,2%)	54 (56,8%)

TABLE 9.8: *Discrimination son/parole des événements de type parole, bruités et non bruités.*

	BREF120	BREF120_ EMOTION_G	BREF120_ SWEETHOME_G	BREF120_ SWEETHOME+EMOTION_G
WER phrases cibles	87,35%	70,41%	59,41%	50,94%
WER perturbateurs	101,59%	107,14%	100,79%	111,90%

TABLE 9.9: *WER pour les phrases cibles et les perturbateurs en fonction des modèles acoustiques.*

Nous avons fourni en entrée du système de RAP *Sphinx3* les 322 fichiers wav de la classe *parole*, dont 277 appartiennent à la catégorie *phrases cibles* (86%), et les 45 autres appartiennent à la catégorie *perturbateurs* (14%).

Les décodages avec Sphinx3 ont été réalisés à partir des modèles décrits section 9.3.1.

Les résultats sont donnés table 9.9.

Les WER pour les phrases cibles après discrimination son/parole parfaite sont globalement moins bons que ceux obtenus après discrimination automatique (voir table 9.4). Cette fois-ci, avec le modèle donnant les meilleurs résultats – le modèle *BREF120_SWEETHOME+EMOTION_G*–, l'amélioration est de 36,41% par rapport au modèle *BREF120* initial pour les phrases cibles.

Nous pouvons émettre l'hypothèse que cette dégradation du WER avec l'étape de discrimination son/parole parfaite est due à l'ajout d'événements *parole* de « mauvaise qualité » dans les phrases décodées. En effet, le décodage inclut cette fois-ci les 53 phrases bruitées classifiées par la discrimination automatique comme étant de la classe *son*.

Les phrases de la catégorie *perturbateurs* sont laissées de côté dans la suite de cette section car nous avons vu table 9.9 que leurs WER sont supérieurs à 100%, cela étant dû au modèle de langage spécifiquement adapté aux phrases cibles.

La figure 9.10 présente la répartitions des phrases cibles dans les classes *son* et *parole* suite à la discrimination automatique, ainsi que le nombre de phrases bruitées et non bruitées dans les phrases reconnues comme du son ou comme de la parole. Parmi les 204 phrases cibles reconnues comme appartenant à la classe *parole*, 29 sont bruitées et 175 sont non bruitées, et, parmi les 73 phrases classifiées comme *son*, 45 sont bruitées et 28 sont non bruitées. Soit, au total, parmi les 277 phrases, 74 sont bruitées et 203 sont non bruitées.

Nous allons à présent comparer les WER obtenus pour chacun des modèles acoustiques avec les phrases cibles selon les caractéristiques suivantes :

— bruitées (B)

— non bruitées (NB)

— classifiées son (S)

— classifiées parole (P)

— classifiées son + bruitées (S+B)

— classifiées parole + bruitées (P+B)

— classifiées son + non bruitées (S+NB)

— classifiées parole + non bruitées (P+NB)

Phrases cibles	BREF120	BREF120_ EMOTION_G	BREF120_ SWEETHOME_G	BREF120_ SWEETHOME+EMOTION_G
Toutes	87,35%	70,41%	59,41%	50,94%
B	106,57%	104,5%	95,45%	97,47%
NB	81,43%	60,90%	49,37%	37,83%
S	113,56%	109,04%	101,13%	101,69%
P	80,46%	61,07%	49,32%	38,52%
S+B	116,96%	116,96%	108,93%	112,50%
P+B	93,02%	88,37%	77,91%	77,91%
S+NB	107,69%	95,38%	87,69%	83,08%
P+NB	78,79%	57,43%	45,51%	33,28%

TABLE 9.10: *WER pour les différents cas de parole bruitée on non bruité, classifié parole ou son, pour les différents modèles acoustiques.*

Pour le meilleur modèle acoustique (le modèle BREF120_SWEETHOME+EMOTION_G), nous voyons que les phrases bruitées ont un WER proche de 100% (voir table 9.10). Le bruit a donc un fort impact sur la dégradation des performances. Les phrases reconnues par la discrimination son/parole comme appartenant à la classe *son* sont également très mal reconnues (WER=101,69%), qu'elles soient bruitées (WER=112,50%) ou non bruitées (WER=83,03%). L'amélioration de la discrimination automatique en vu de classifier 100% des phrases dans la classe *parole* semble donc avoir peu d'intérêt pour le décodage car les phrases classifiées *son* sont de toute façon très mal décodées. En comparant le WER des phrases classifiées *son* (WER=101,69%) et *parole* (WER=38,52%), nous constatons que la discrimination son/parole de *CirdoX* est un très bon filtre pour rejeter les événements *parole* de

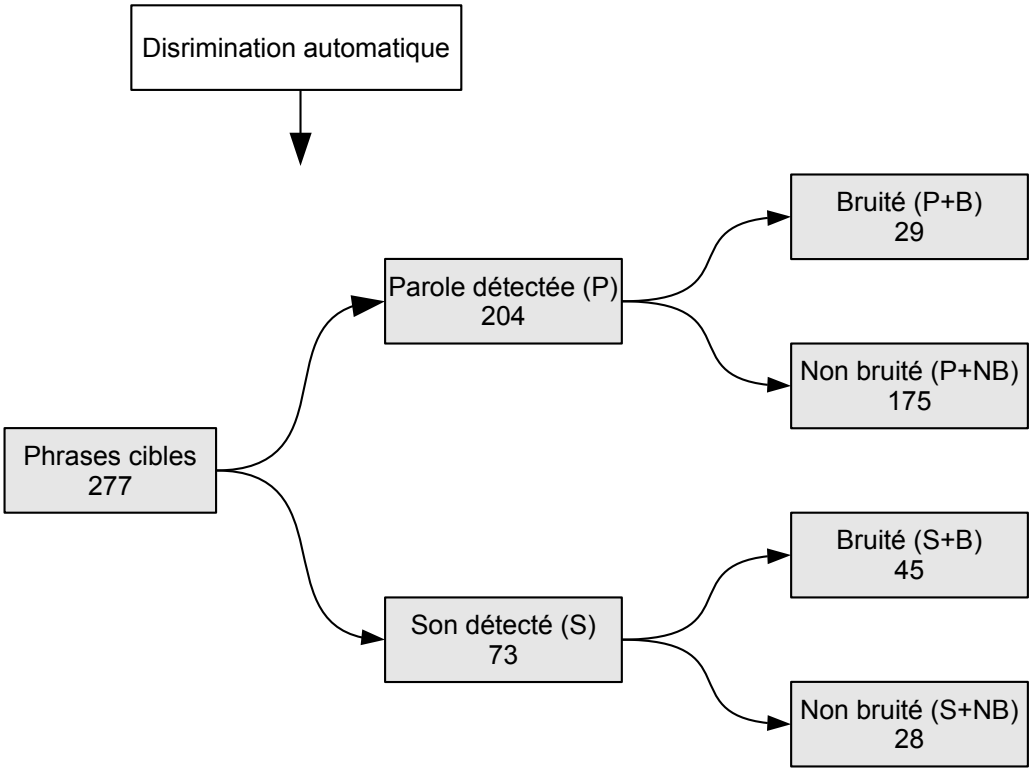


FIGURE 9.10: *Répartition des phrases cibles bruitées et non bruitées dans les classes « parole détectée » et « son détecté »*

« mauvaise qualité ». Pour les phrases classifiées *parole*, le bruit a un impact important : pour celles non bruitées, le WER est de 33,28%, alors que pour les phrases bruitées, il augmente à 77,91%.

Nous allons donc chercher la cause de la mauvaise discrimination et du mauvais décodage (WER=83,08%) des phrases classifiées *son* non bruitées (S+NB).

En écoutant les 28 phrases cibles *S+NB*, il est difficile de mettre en évidence, sur le plan perspectif, un facteur pouvant expliquer leur mauvaise discrimination car, subjectivement, elles ne semblent pas différentes des phrases *P+NB*. En revanche, leur mauvais décodage par le système de RAP pourrait s'expliquer par le fait qu'elles soient principalement – pour 25 d'entre elles – des énoncés courts. En effet, le modèle de langage est face à une plus grande indécision avec ce type de phrases. Les énoncés courts rencontrés – des interjections et phrases courtes – sont « Oh la ! », « Aïe ! », « e-lio ! » et « J'ai mal ! ».

9.3.3.3 Détection des phrases cibles

Les décodages présentés dans la section précédente 9.3.3.2 ont été réalisés sur l'ensemble des 322 événements classés manuellement comme étant de la parole (cas d'une discrimination son/parole dite « parfaite »). A la suite de la discrimination son/parole parfaite et de ces décodages, nous avons appliqué le filtre de détection des phrases cibles du système *CirdoX* sur l'ensembles des 322 hypothèses de sortie de *Sphinx3* (277 phrases cibles et 45 perturbateurs), obtenues à partir des différents modèles acoustiques (modèle générique et modèles adaptés).

Les courbes ROC de la classification *phrases cibles*/*perturbateurs* par le filtre sont données à la figure 9.11. A partir des courbes ROC, nous avons fixé les seuils permettant le

seuil = 1,364	d < seuil	d >= seuil
Phrases cibles	$VP = 201$	$FN = 76$
Perturbateurs	$FP = 17$	$VN = 28$

(a) *Modèle acoustique BREF120.*

seuil = 1,200	d < seuil	d >= seuil
Phrases cibles	$VP = 201$	$FN = 76$
Perturbateurs	$FP = 11$	$VN = 34$

(b) *Modèle acoustique BREF120_EMOTION_G.*

seuil = 1,125	d < seuil	>= seuil
Phrases cibles	$VP = 199$	$FN = 78$
Perturbateurs	$FP = 12$	$VN = 33$

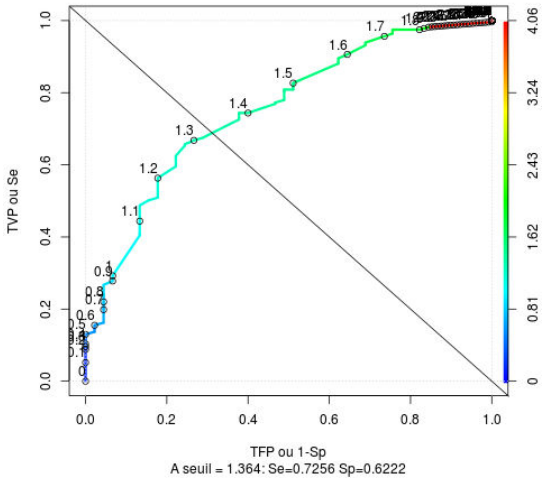
(c) *Modèle acoustique BREF120_SWEETHOME_G.*

seuil = 1,000	d < seuil	d >= seuil
Phrases cibles	$VP = 200$	$FN = 77$
Perturbateurs	$FP = 11$	$VN = 34$

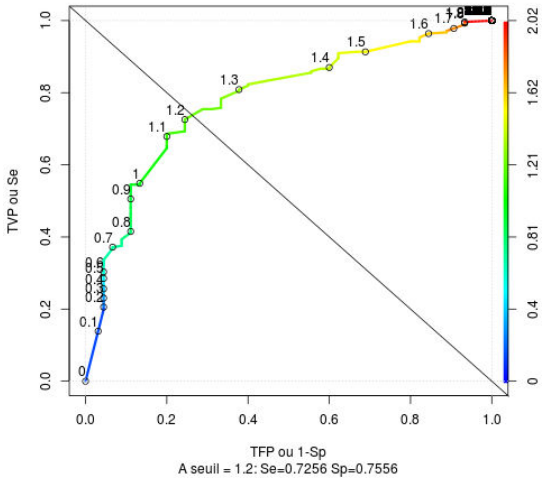
(d) *Modèle acoustique BREF120_SWEETHOME+EMOTION_G.*

TABLE 9.11: *Matrice de confusion du filtrage des phrases cibles pour les différentes modèles acoustiques.*

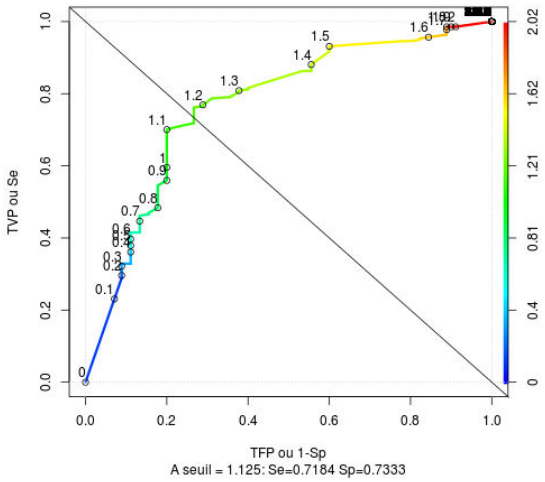
meilleur compromis entre sensibilité et spécificité. Pour chaque modèle acoustique et pour les seuils fixés, les résultats de la classification *phrases cibles*/ *perturbateurs* sont donnés table 9.11, qui présente les matrices de confusion, et par la table 9.12, qui présente la sensibilité, la spécificité et le taux de fausses alarmes, ainsi que le WER des phrases cibles.



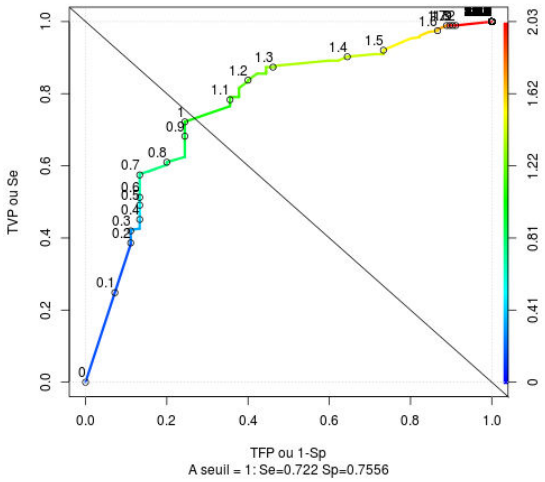
(a) Modèle BREF120



(b) Modèle BREF120_EMOTION_G



(c) Modèle BREF120_SWEETHOME_G



(d) Modèle BREF120_SWEETHOME+EMOTION_G

FIGURE 9.11: *Courbes ROC représentant le TVP en fonction du TFP pour le filtrage des phrases cibles en fonction des modèles acoustiques.*

	BREF120	BREF120- _EMOTION_G	BREF120- _SWEETHOME_G	BREF120- _SWEETHOME+EMOTION_G
Se	72,56%	72,56%	71,84%	72,20%
Sp	62,22%	75,56%	73,33%	75,56%
TFA	7,80%	5,19%	5,58%	5,21%
WER	87,35%	70,41%	59,41%	50,94%

TABLE 9.12: *Sensibilité, spécificité, taux de fausses alarmes et WER des phrases cibles en fonction des modèles acoustiques.*

	BREF120	BREF120- _EMOTION_G	BREF120- _SWEETHOME_G	BREF120- _SWEETHOME+EMOTION_G
TBR	42,24% (117)	54,15% (150)	59,21% (164)	63,18% (175)
WER	87,35%	70,41%	59,41%	50,94%

TABLE 9.13: *TBR (avec le nombre de phrases correspondant) et WER en fonction des modèles acoustiques.*

Comme illustré table 9.12, pour le modèle *BREF120_SWEETHOME+EMOTION_G*, 72,20% des phrases cibles (Se=72,2%) sont bien classifiées dans la catégorie *phrase cible*. De plus, nous observons que les performances de la détection sont moins bonnes qu'en section 9.3.2.3 (table 9.6), avec Se=72,20% table 9.12 et Se=82,84% table 9.6 pour le modèle *BREF120_SWEETHOME+EMOTION_G*. Cela est la conséquence du WER plus élevé (WER=50,94%) par rapport à celui trouvé en section 9.3.2.3 où WER=38,52%. En effet, dans le cas présent (discrimination son/parole manuelle), toutes les phrases ont été présentées au filtre, alors qu'en section 9.3.2.3 (table 9.6), les phrases classifiées dans la classe *son* par la discrimination son/parole automatique – pour la plupart des phrases bruitées donnant lieu à un WER élevé (voir table 9.4) – avaient été rejetées, donc non décodées.

La Table 9.13 présente les taux de bonne reconnaissance (TBR) pour chacun des modèles acoustique. Le TBR représente le taux de phrases cibles (taux calculé ici sur les 277 phrases cibles totales) qui sont filtrées sans confusion. Pour le meilleur modèle acoustique (*BREF120_SWEETHOME+EMOTION_G*), le TBR est de 63,18% : 175 des 277 événements de type *phrase cible* sont reconnus par le filtre comme appartenant à la catégorie *phrase cible* sans confusion. Dans le cas réel (discrimination son/parole automatique, voir section 9.3.2.3), le TBR était de 75% (153 événements sur 204 bien reconnus par le filtre). Nous voyons donc que le bruit dans la parole a un impact important sur le filtrage des phrases cibles.

9.3.3.4 Détection des phrases cibles pour la parole bruitée et non bruitée

Nous avons refait le filtrage réalisé dans la section précédente (voir section 9.3.3.3) avec le modèle acoustique BREF120_SWEETHOME+EMOTION_G sur les phrases non bruitées seules, puis sur les phrases bruitées seules.

La figure 9.12 illustre la répartition des phrases bruitées et non bruitées dans les phrases cibles et dans les perturbateurs. Les matrices de confusion de la détection phrases cibles/-perturbateurs sont présentées table 9.14 pour la parole bruitée et non bruitée. Les seuils ont été fixés à partir des courbes ROC présentées figure 9.13. De plus, la table 9.15 montre la sensibilité, la spécificité et le taux de fausses alarmes pour la parole non bruitée et la parole bruitée. Enfin, la table 9.16 montre les TBR pour la parole non bruitée et pour la parole bruitée.

Nous pouvons observer que les résultats pour la parole bruitée sont fortement dégradés par rapport aux résultats pour la parole non bruitée. Avec la parole bruitée, par rapport à la parole non bruitée, l'aire sous la courbe ROC est fortement diminuée avec une sensibilité

dégradée de 10,13% (différence absolue). De plus, la dégradation du TBR est de 31% pour la parole bruitée.

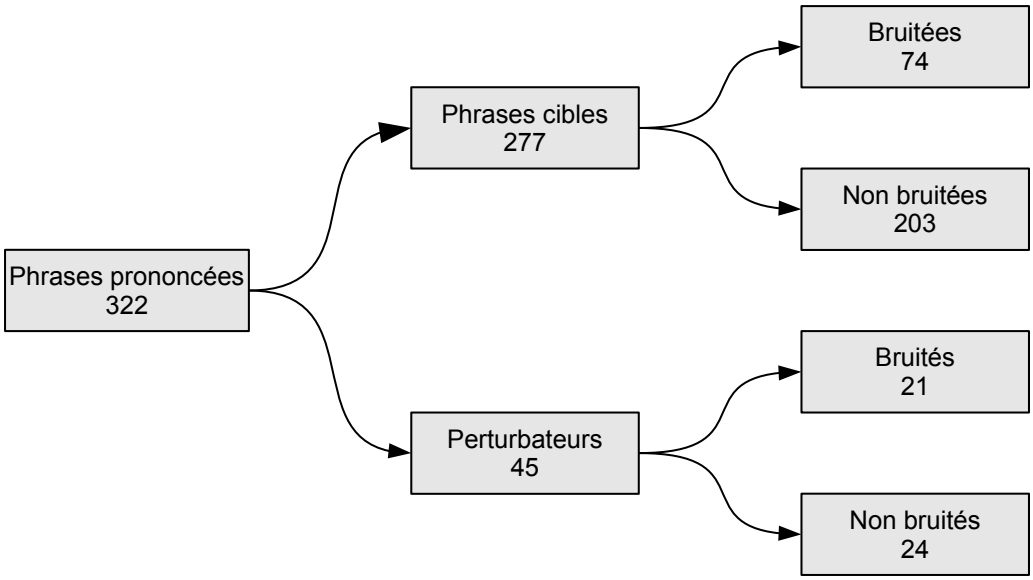


FIGURE 9.12: *Répartition des phrases bruitées et non bruitées dans les catégories « phrases cibles » et « perturbateurs »*

seuil = 0,778	d < seuil	d >= seuil
Phrases cibles	$VP = 144$	$FN = 59$
Perturbateurs	$FP = 8$	$VN = 16$

(a) *Phrases non bruitées.*

seuil = 1,167	d < seuil	d >= seuil
Phrases cibles	$VP = 45$	$FN = 29$
Perturbateurs	$FP = 8$	$VN = 13$

(b) *Phrases bruitées.*

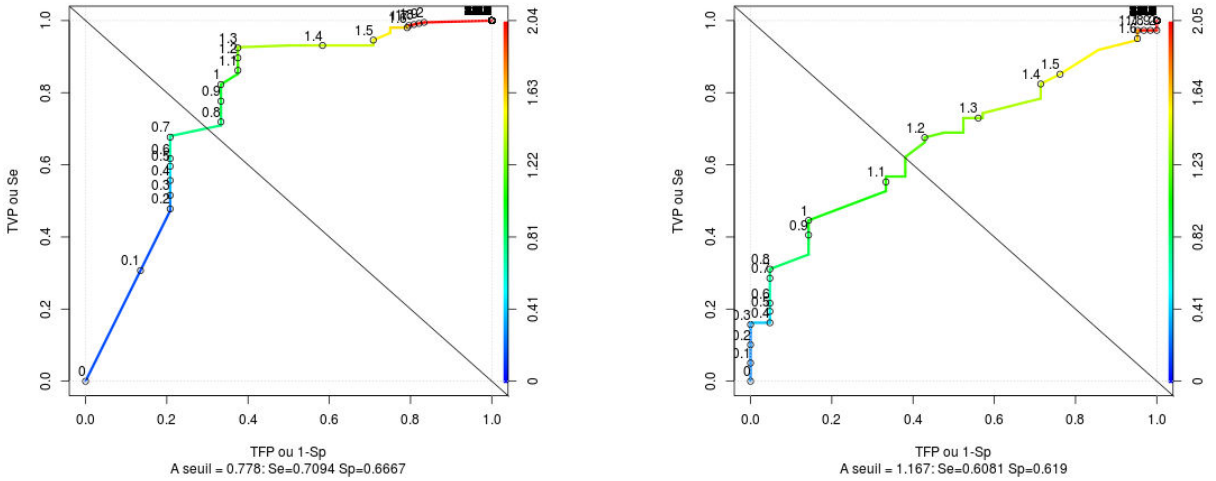
TABLE 9.14: *Comparaison des matrices de confusion du filtrage des phrases cibles pour la parole non bruitée et bruitée.*

	Phrases non bruitées	Phrases bruitées
Se	70,94%	60,81%
Sp	66,67%	61,90%
TFA	5,29%	15,09%

TABLE 9.15: *Sensibilité, spécificité et taux de fausses alarmes pour les phrases non bruitées et bruitées.*

	Phrases non bruitées	Phrases bruitées
TBR	67,49% (137)	36,49% (27)

TABLE 9.16: *TBR (avec le nombre de phrases correspondant) pour les phrases non bruitées et bruitées.*



(a) Phrases non bruitées (b) Phrases bruitées

FIGURE 9.13: *Courbes ROC représentant le TVP en fonction du TFP pour le filtrage des phrases cibles en fonction de la présence ou non de bruit dans la parole.*

9.3.3.5 Performances globales du système sans les phrases bruitées

Pour le modèle *BREF120_SWEETHOME+EMOTION_G*, en suivant la chaîne de traitement *CirdoX* (détection, discrimination son/parole automatique, RAP et filtrage des phrases cibles), et en ne prenant pas en compte les phrases bruitées, soit 203 phrases de référence (identifiées manuellement comme appartenant à la classe *parole* et à la catégorie *phrases cibles*), les résultats sont les suivants : 175 événements sont classifiés par le classifieur GMM comme appartenant à la classe *parole* (86,21%), 151 sont identifiés par le filtre comme appartenant à la catégorie *phrases cibles* (74,38%), et 139 sont identifiés par le filtre comme appartenant à la catégorie *phrases cibles* sans confusion, soit 68,47%.

9.4 Bilan

Dans ce chapitre, nous avons présenté l'expérimentation réalisée en situation réaliste dans l'appartement Domus, où il a été demandé à 17 sujets, pour la plupart jeunes, de réaliser 4 scénarios de chutes, un scénario de blocage et 2 scénarios « faux positifs », pendant que les expérimentateurs enregistraient les scènes en vue de leur analyse sonore et audio par le système *CIRDO*.

Le système *CIRDO* est composé d'un module vidéo, développé par le laboratoire LIRIS partenaire du projet *CIRDO*, d'un module audio (le logiciel *CirdoX*) que nous avons développé, et d'un module de fusion audio/vidéo pour la prise de décision (envoi d'une alerte), développé conjointement par le laboratoire LIRIS et notre laboratoire.

Nous avons présenté la chaîne de traitement audio du logiciel *CirdoX*, composé d'un module d'acquisition du flux sonore, d'un module de détection du signal, d'un module de discrimination son/parole, d'un module de classification du son, d'un module de reconnaissance automatique de la parole et de filtrage des phrases cibles, et d'un module de mise en

forme des données.

Les scénarios de chute et de blocage ont donné lieu à l'énoncé de phrases de détresse et de phrases anodines, qui nous ont permis de tester toute la chaîne de traitement du logiciel *CirdoX*. Au préalable, nous avons adapté le modèle de langage et le filtre aux phrases cibles recherchées, et nous avons adapté les modèles acoustiques à l'environnement sonore et à la voix émue.

Nous avons vu que la discrimination son/parole a bien classifié dans la classe *parole* 73,6% des événements de type *parole*, et a bien classifié dans la classe *son* 99,6% des événements de type *son*. Le décodage avec Sphinx3 des événements sonores classifiés comme étant de la classe *parole* a donné lieu à un WER=38,52% pour les phrases cibles avec le modèle de langage adapté à l'environnement et à la voix émue. A la suite du décodage Avec le filtrage des phrases cibles parmi les autres phrases appelées les « perturbateurs », 82,84% des événements de type *phrase cible* ont été reconnues comme *phrases cibles*. Parmi les phrases cibles décodées, 75% sont correctement reconnues par le filtre. En considérant l'ensemble de la chaîne de traitement (détection, discrimination son/parole, RAP et filtrage), 55,23% des phrases cibles prononcées par les sujets ont au final été correctement reconnues par le système (taux de performances globales du système).

Nous avons ensuite analysé quelle pouvait être la cause de la mauvaise reconnaissance des 44,77% de phrases cibles restantes. Nous avons vu que le bruit dans la parole était un facteur important de la dégradation des performances du système. En effet, pour le module de discrimination son/parole, la parole non bruitée est correctement classifiée dans la classe *parole* à 86,3%, alors que la parole bruitée l'est à 43,2%. Lors du décodage, le WER=37,83% pour la parole non bruitée, alors qu'il est de 97,47% pour la parole bruitée. Suite au filtrage, la sensibilité (taux des phrases cibles reconnues comme phrases cibles) est de 70,94% pour les phrases non bruitées, et de 60,81% pour la parole bruitée, et le taux de bonne reconnaissance et de 67,49% pour les phrases non bruitées et de 36,49% pour les phrases bruitées. Enfin, en retirant les phrases bruitées, le taux de performances globales du système passe de 55,23% à 68,47%.

CHAPITRE 10

Conclusion et perspectives

10.1 Conclusion

Dans le contexte de l'habitat intelligent et du maintien à domicile des personnes âgées, notre travail de thèse a eu pour objectif d'inclure dans le milieu de vie de la personne âgée dépendante un système de reconnaissance automatique de la parole capable de reconnaître des appels vers les aidants et détecter les appels de détresse prononcés pour lancer une alerte.

Nous avions ainsi à étudier le comportement d'un système de RAP sur la voix rendue « atypique » par le vieillissement du locuteur ou l'émotion que celui-ci ressent au moment de parler. Nous avons donc séparé les problèmes, en étudiant dans un premier temps les effets de la voix âgée sur les systèmes de RAP, puis en second temps les effets de la voix émue sur la RAP. En dernier lieu, nous avons évalué notre système au cours d'une expérimentation en situation réaliste jouée.

Une importante contribution de la thèse est la mise à disposition de nouveaux corpus annotés de voix âgées et de voix émues. Nous avons enregistré 3 corpus :

— Le corpus *AD80* : ce corpus est composé de phrases lues de type détresse ou anodines. Nous avons enregistré 43 locuteurs âgés dans une maison de retraite et un centre de réhabilitation, et 31 locuteurs jeunes au laboratoire. Ce corpus a été complété par un corpus existant, le corpus *AD*, enregistré en 2004 auprès de 21 locuteurs jeunes. Au total, *AD80* représente 14 267 énoncés annotés, soit 4 heures et 49 minutes d'enregistrements. Un texte prévu pour l'adaptation acoustique a également été lu par tous les locuteurs (à l'exception des locuteurs de 2004), représentant 2 heures et 41 minutes d'enregistrement.

— Le corpus *ERES38* : ce corpus de parole spontanée a été enregistré auprès de personnes âgées en collaboration avec Sasa et Legrand (2011). Ce corpus a été utilisé pour l'adaptation des modèles acoustiques. Il a été enregistré auprès de 23 personnes âgées dans leur lieu de vie (foyers logement, maison de retraite). Il est constitué d'interviews et de lectures d'un article. Le corpus inclut 48 minutes de lectures annotées et 16 heures et 56 minutes d'interviews, dont 4 heures et 53 minutes ont été annotées.

— Le corpus *Voix Détresse* a été enregistré auprès de 25 locuteurs, essentiellement jeunes (20 locuteurs jeunes et 5 locuteurs âgés). Il a été demandé aux locuteurs de lire 20

phrases de détresse de façon neutre, puis d'énoncer ces mêmes phrases de façon très expressive. 521 phrases neutres et 1221 phrases émues ont été prononcées, soit 28 minutes d'enregistrement (7 minutes et 30 secondes pour les phrases neutres et 20 minutes et 30 secondes pour les phrases émues).

L'enregistrement du corpus *AD80* a été réalisé à partir du logiciel GEOD que nous avons développé.

Comme cela a été déjà été observé dans plusieurs études, nos résultats du décodage sur un corpus de parole lue (corpus *AD80*) montre une importante dégradation des performances moyennes de la RAP avec la voix âgée par rapport à la voix jeune en utilisant un modèle acoustique enregistré dans des conditions idéales (phrases d'articles de journaux lues par des locuteurs jeunes en studio). Cette dégradation est observée pour tous les décodeurs étudiés. En regardant le WER individuel de chaque locuteur, nous observons une très grande variabilité de WER au sein des locuteurs âgés, l'écart-type des WER au sein des locuteurs âgés est de 1,8 à 2,6 fois plus important que l'écart-type des locuteurs jeunes.

Nous avons donc adapté acoustiquement les modèles du système *Sphinx3* afin d'améliorer les performances de la RAP avec la voix âgée. Nous avons comparé plusieurs adaptations : adaptation à la voix âgée de façon globale, adaptation à la voix âgée globale en fonction du genre, et adaptation au locuteur. Ces adaptations permettent de diminuer le WER, cependant l'écart entre les WER moyens des locuteurs âgés avec les modèles adaptés restent plus élevés que le WER moyen des locuteurs jeunes avec le modèle générique : pour les voix âgées, l'adaptation n'est pas un moyen suffisant pour égaler les performances du système avec les voix jeunes. Aussi, pour les locuteurs âgés, la variabilité du WER après adaptation reste plus importante que la variabilité des locuteurs jeunes, avec un écart-type pour les voix âgées 1,6 fois plus important que l'écart-type des locuteurs jeunes. De plus, en comparant les différentes adaptations (adaptation à la voix âgée de façon globale, adaptation à la voix âgée genre-dépendant, et adaptation au locuteur), nous observons que les performances obtenues sont similaires, il n'y a pas de différences significatives entre les WER obtenus. L'adaptation qui nous semble la plus pertinente est donc l'adaptation à la voix âgée de façon globale car cela évite à l'utilisateur l'étape fastidieuse de lecture de phrases pour l'adaptation à sa voix, l'adaptation étant faite en amont à partir de la voix de locuteurs différents des locuteurs utilisateurs.

A partir du constat de l'importante variabilité des performances de la RAP entre les locuteurs âgés, nous avons cherché quels pouvaient être les facteurs explicatifs de la dégradation du WER et cherché si ces facteurs pouvaient être utilisés pour prédire le WER pour un locuteur âgé donné.

Les facteurs suivants ont été pris en considération : l'âge, le niveau de dépendance, les scores d'alignement forcé et les paramètres prosodiques.

En considérant l'ensemble des locuteurs jeunes et âgés, l'âge est assez fortement corrélé au WER (corrélation de 0,747). En revanche, lorsque l'on regarde la corrélation pour le seul groupe de locuteurs âgés, la corrélation entre l'âge et le WER est proche de 0. En effet, il existe une différence entre l'âge physiologique et l'âge réel de la personne en fonction de

son vécu. Un âge physiologique avancé, correspondant à un état physique dégradé, sera lié à une perte de l'autonomie de la personne. Cet état général dégradé aura également un impact sur la voix. Le niveau de dépendance pouvant être facilement évalué (grille AGGIR), nous avons donc cherché s'il existe une corrélation entre la dépendance et le WER. La corrélation calculée est au bout du compte assez modérée (corrélation de -0,318).

Les scores d'alignement forcé permettent d'avoir une bonne estimation du niveau d'éloignement entre un phonème décodé et sa représentation dans le modèle acoustique. La corrélation calculée entre les scores d'alignement forcé et le WER est élevée (corrélation de -0,81). Aussi, les scores d'alignement forcé permettent d'observer quels sont les phonèmes prononcés par les locuteurs âgés étant les plus mal reconnus par le système. Pour les consonnes, les catégories les plus mal reconnues sont celles demandant une pression intra-orale plus importante (plosives et fricatives).

Nous avons mesuré le débit, la F0, le jitter, le shimmer et le rapport harmoniques sur bruit à l'aide de l'outil *Praat* pour l'ensemble des voix jeunes et âgées. Nous constatons grâce à un clustering et une analyse en composantes principale que ces paramètres permettent de bien séparer les locuteurs jeunes des locuteurs âgés. L'ACP montre que la discrimination jeunes/âgés est liée au jitter, au shimmer et au débit : pour les voix âgées par rapport aux jeunes, le jitter et le shimmer augmentent, et le débit diminue. La F0 permet quant à elle de distinguer les hommes des femmes (F0 augmente pour les femmes), et le rapport harmoniques sur bruit n'est pertinent ni pour la discrimination jeunes/âgés, ni pour la discrimination hommes/femmes.

La corrélation avec le WER est modérée pour le débit, le jitter et le shimmer (corrélations aux alentours de 0,5 en valeur absolue), faible pour le rapport harmoniques sur bruit (corrélation de -0,36), et non significative pour la F0.

A l'aide du classifeur *Weka*, nous avons cherché s'il est possible de prédire le WER à partir des scores d'alignement forcé et des paramètres prosodiques, ceux-ci ayant le plus de corrélation avec le WER. Nous avons donc rangé les WER dans 2 classes, puis dans 4 classes, chaque classe correspondant à une plage de WER. Pour 2 classes, le résultat de la classification avec les scores d'alignement forcé est assez bon, avec F-mesure=0,873. En revanche, avec 4 classes de WER, le résultat est beaucoup moins bon, avec F-mesure=0,663. Les résultats avec les paramètres prosodiques sont similaires : à partir du modèle acoustique générique, avec 2 classes de WER, F-mesure=0,884, et avec 4 classes, F-mesure=0,528. A partir du modèle adapté à la voix âgée, il est encore plus difficile de prédire le WER car la variabilité entre locuteurs est plus basse : avec 2 classes de WER, F-mesure=0,635. L'utilisation de ce type de paramètres apparaît donc peu probante pour la prédiction des performances des systèmes de RAP adapté à la voix âgée.

En revanche, la prédiction du type de voix jeunes ou âgés à partir des paramètres prosodiques donne de bons résultats, avec F-mesure=0,905.

Nous avons ensuite étudié l'impact de la voix émue sur les systèmes de RAP. La dégradation des performances est conséquente avec la voix émue par rapport à la voix neutre, quel que soit le système de RAP utilisé. Nous avons également montré qu'il y a une diffé-

rence entre la voix émue actée et la voix émue spontanée : la voix émue spontanée est moins bien reconnue que la voix émue actée. De plus, le type de voix utilisé pour l'adaptation des modèles acoustiques est très importante. Pour décoder la voix émue, une adaptation au locuteur à partir de phrases lues de façon neutre sera beaucoup moins performante qu'une adaptation à partir de voix émue.

En analysant les mesures des paramètres prosodiques débit, F0, jitter shimmer et rapport harmoniques sur bruit, nous observons pour la voix émue, par rapport à la voix neutre, une diminution du débit, une augmentation de la F0, une diminution du jitter et du shimmer et une augmentation du HNR. L'apparente opposition de certaines caractéristiques acoustiques (le jitter et le shimmer) de la détresse avec l'état de l'art est à nuancer car nous avons vu qu'il existe différents types d'émotions dans la détresse.

Cette thèse à également donné lieu au développement du logiciel *CirdoX*, mettant en application la chaîne de traitement audio issue du travail de l'équipe GETALP au cours des dernières années. Le logiciel *CirdoX*, dans le cadre du projet *CIRDO*, a pour but la détection d'appels volontaires, que ce soient des appels de détresse ou des appels aux aidants. Ce logiciel est un module du système *CIRDO*, qui est également composé d'un module vidéo et d'un module de fusion son/vidéo pour la prise de décision en vue du déclenchement d'une alerte vers le système *e-lio*. Le logiciel *CirdoX* est composé de la chaîne de traitement suivante : acquisition audio, détection du signal et du silence, discrimination son/parole, RAP si parole détectée, classification des sons si sons détectés, filtrage des phrases cibles, mise en forme des données et envoie des données sous forme de sockets pour leur traitement par le module de fusion son/vidéo.

Le système *CirdoX* a été évalué à partir d'une expérimentation réalisée en situation réaliste dans un appartement de test, DOMUS. Il a été demandé aux volontaires de jouer différents scénarios de chute et de blocage, et des faux positifs (situations ressemblant à une chute ou un blocage, mais n'en étant pas). 17 volontaires ont ainsi réalisé 7 scénarios en étant filmés, et leur voix étant enregistrée. La discrimination son/parole avec un classifieur GMM donne des résultats pouvant être améliorés. Si 99,6% des sons sont correctement classifiés comme son, 73,6% des signaux de parole sont reconnus comme de la parole. Après adaptation des modèles acoustiques à l'environnement (microphones distants) et aux émotions, le WER obtenu pour les phrases cibles est de 38,52%. A la suite de toutes les étapes de la chaîne de traitement, nous obtenons comme résultat que 55,23% des phrases de détresse/appels aux aidants prononcées sont correctement détectées. Il est donc important de prendre en considération l'écart important entre une évaluation sur corpus et les conditions réelles. En effet, viennent s'ajouter des phénomènes d'échos et de réverbérations ; par ailleurs, l'environnement dans un appartement est très bruité, et le bruit des mouvements des sujets perturbent également les signaux de parole, tout ceci expliquant la dégradation du système.

10.2 Perspectives

Les résultats obtenus montrent le pas à franchir pour envisager une utilisation en situation réelle. Pour améliorer ces résultats, plusieurs pistes peuvent être envisagées.

Pour la discrimination son/parole, un classifieur à base de modèles SVM (*Support Vector Machines*) pourrait être utilisé. Cette approche a été utilisée par Sehili et coll. (2012) dans leurs travaux sur la reconnaissance de sons de la vie courante dans un contexte domotique dans le cadre du projet ANR *Sweet-Home*. Les tests comparant un modèle classique GMM avec un modèle SVM (super vecteur UBM de 1024 distributions gaussiennes) ont été réalisés en variant progressivement le rapport signal sur bruit (RSB) du signal. Les résultats ont montré que, pour tous les tests, la méthode SVM donne de meilleures performances : par exemple, avec des signaux non-bruités, 69% et 75% de bonnes classifications ont été obtenues pour, respectivement, le modèle GMM et le modèle SVM, et, pour des signaux avec un RSB de 10dB, 52% et 62% de bonnes classifications ont été obtenues pour les modèles GMM et SVM respectivement.

Une deuxième piste d'amélioration de la discrimination son/parole pourrait être l'utilisation d'autres types de paramètres que les paramètres classiques MFCC. Pinquier et coll. (2003) utilisent par exemple comme paramètre, pour leurs travaux sur la classification automatique parole/musique, la modulation de l'énergie à 4Hz (Houtgast et Steeneken, 1985). En effet, le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4Hz. Ces modulations correspondent au rythme syllabique de l'élocution humaine. Ils obtiennent, avec une évaluation sur une feuilleton télévisuel de 20 minutes (« Chapeau melon et bottes de cuir »), 84,1% de bonne classification parole/musique.

La reconnaissance automatique de la parole, pour la voix âgée et la voix émue, pourrait être améliorée en utilisant le système *Kaldi* (Povey et coll., 2011) qui permet l'utilisation de modèles acoustiques basés sur les SGMM (*Subspace Gaussian Mixture Model*) et une adaptation acoustique CMLLR (*Constrained Maximum Likelihood Linear Regression*) ou fMLLR (*Feature space Maximum Likelihood Linear Regression*). Dans (Vu et coll., 2012), les auteurs expérimentent différents types de modèles acoustiques en fonctions du genre du locuteur. Leur évaluation sur le corpus *GlobalPhone* montre, pour la langue française, une amélioration de 15,1% (différence relative) entre un modèle classique triphone HMM (WER=26,5%) et un modèle SGMM (WER=22,5%). Avec un modèle SGMM couplé à une adaptation CMLLR, l'amélioration par rapport au modèle HMM est de 16,6% (WER=22,1%). Par ailleurs, Vacher et coll. (2014c) proposent un système basé sur *Kaldi* pour la détection d'ordres domotiques, l'expérimentation étant réalisée sur un corpus de locuteurs âgés ou malvoyants. Avec le modèle HMM triphone baseline, le WER est de 51,9%. Après adaptation fMLLR, l'amélioration est de 26,6% WER=39%), et, avec le modèle fMLLR+SGMM, l'amélioration est de 30,4% (WER=36,1%).

Concernant la reconnaissance en conditions distantes, l'intégration du décodeur *Kaldi* dans le logiciel *CirdoX* permettrait une utilisation des modèles SGMM et une adapta-

tion CMLLR ou fMLLR. En conditions faiblement bruitées, l'utilisation de canaux supplémentaires pourrait permettre une amélioration des résultats. Lecouteux et coll. (2013) obtiennent, avec la méthode DDA (*Driven Decoding Algorithm*) (Lecouteux et coll., 2007), une réduction de 3% du WER lors de leur expérimentation sur la combinaison de différents systèmes de RAP, l'évaluation étant réalisée à partir du corpus *ESTER* (corpus d'émissions de radio).

En cas de bruits de source connue, Principi et coll. (2013) utilisent, dans leur système de reconnaissance d'appel de détresse et de commandes commandes domotique en italien, l'algorithme du codec *Speex* (Valin, 2006) basé sur un filtre adaptatif MDF (*Multidelay Block Frequency Domain*) pour réduire les interférences de sources connues (télévision, radio, etc.). Le même algorithme a été utilisé pour réduire les effets de l'écho. Les performances de l'annulation d'écho sont mesurées par le ERLE (*Echo Return Loss Enhancement*), qui est mesuré à 30dB par les auteurs. L'étude du débruitage, en cas de sources de bruit inconnues, fait l'objet d'efforts importants de la communauté, comme en témoigne l'organisation du *CHiME Speech Separation and Recognition Challenge* (Barker et coll., 2013). Les techniques envisagées s'appuient sur l'algorithme de débruitage BSS (*Blind Audio Source Separation*) (Liutkus et coll., 2013; Vincent et coll., 2012).

Enfin, la phase finale de décision n'a pas été abordée. Le module de décision devra prendre en compte l'historique, les habitudes de la personne et les informations issues de la vidéo afin de décider si un événement détecté doit donner lieu au déclenchement d'une alerte. La prise de décision pourrait se baser sur des modèles de représentation de connaissance. Une approche se basant sur les réseaux logiques de Markov (MLN) a par exemple été implémentée par Chahuara et coll. (2013) dans le cadre du projet *Sweet-Home*[1] lors du développement d'un contrôleur intelligent en environnement domotique. Ce type de réseau présente l'avantage d'être un modèle statistique combinant la logique de premier ordre et les réseaux de Markov. Ce contrôleur permet la prise de décision à partir de sources multimodales, l'appartement étant ainsi capable de réagir en fonction des ordres domotiques, des sons, de la localisation et de l'activité de l'habitant, inférés à partir des différents capteurs de l'appartement.

1. http ://sweet-home.imag.fr/

Bibliographie

ALWAN, M., RAJENDRAN, P., KELL, S., MACK, D., DALAL, S., WOLFE, M. et FELDER, R. (2006). A smart and passive floor-vibration based fall detector for elderly. Dans *Information and Communication Technologies, 2006. ICTTA '06. 2nd*, volume 1, pages 1003–1007.

ANDERSON, S., LIBERMAN, N., BERNSTEIN, E., FOSTER, S., CATE, E., LEVIN, B. et HUDSON, R. (1999). Recognition of elderly speech and voice-driven document retrieval. Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '99*, volume 1, pages 145–148.

ARKEBAUER, H. J., HIXON, T. J. et HARDY, J. C. (1967). Peak intraoral air pressures during speech. *Journal of Speech, Language, and Hearing Research*, 10(2):196–208.

ARNING, K. et ZIEFLE, M. (2007). Understanding age differences in PDA acceptance and performance. *Computers in Human Behavior*, 23(6):2904–2927.

AUBERGÉ, V. (2002). Prosodie et émotion. *Actes des deuxiemes assises nationales du GdR I3 (Information Interaction Intelligence), Cépaduès-Editions*, pages 263–274.

AUBERGÉ, V., AUDIBERT, N. et RILLIARD, A. (2004). E-wiz : A trapper protocol for hunting the expressive speech corpora in lab. Dans *LREC 2004, Lisbon, Portugal*, pages 179–182.

AUBERT, X. L. (2002). An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech & Language*, 16(1):89–114.

AUDIBERT, N. (2008). *Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés*. Thèse de doctorat, Institut National Polytechnique de Grenoble-INPG, Ecole Doctorale « Ingénierie pour la Santé, la Cognition et l'Environnement ».

AUGUSTO, J., NAKASHIMA, H. et AGHAJAN, H. (2010). Ambient intelligence and smart environments : A state of the art. Dans NAKASHIMA, H., AGHAJAN, H. et AUGUSTO, J., éditeurs : *Handbook of Ambient Intelligence and Smart Environments*, pages 3–31. Springer US.

AUGUSTO, J. C. et NUGENT, C. D. (2006). Smart homes can be smarter. Dans *Designing Smart Homes*, pages 1–15.

AYNAUD, C. (2009). Reconnaissance de la parole chez les personnes âgées, adaptation d'un système de reconnaissance. Rapport de Master M1, Institut National Polytechnique de Grenoble, PHELMA, filière SICOM.

BABA, A., YOSHIZAWA, S., YAMADA, M., LEE, A. et SHIKANO, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2*, 87:49–57.

BAHADORI, S., CESTA, A., GRISETTI, G., IOCCHI, L., LEONE, R., NARDI, D., ODDI, A., PECORA, F. et RASCONI, R. (2004). Robocare : pervasive intelligence for the domestic care of the elderly. *Intelligenza Artificiale*, 1(1):16–21.

BAHL, L. B., de SOUZA, P. et P MERCER, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Dans *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, pages 49–52.

BARKER, J., VINCENT, E., MA, N., CHRISTENSEN, H. et GREEN, P. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633.

BEAUMEL, C. et BELLAMY, V. (2013). La situation démographique en 2011 - mouvement de la population. *INSEE, N° 145 Société*.

BELLAMY, V. et BEAUMEL, C. (2013). Bilan démographique 2012. la population croît, mais plus modérément. *INSEE Première*, 1429:1–4.

BENJAMIN, B. (1981). Frequency variability in the aged voice. *Journal of Gerontechnology*, 36:722–726.

BLANCHET, D. et GALLO, F. L. (2013). Baby-boom et allongement de la durée de vie : quelles contributions au vieillissement ? *INSEE Analyses N°12*.

BLANPAIN, N. (2010). 15 000 centenaires en 2010 en france, 200 000 en 2060 ? *INSEE Première*, 1319:1–4.

BLOCH, F., GAUTIER, V., NOURY, N., LUNDY, J.-E., POUJAUD, J., CLAESSENS, Y.-E. et RIGAUD, A.-S. (2011). Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects. *Annals of Physical and Rehabilitation Medicine*, 54(6):391 – 398.

BOBILLIER-CHAUMON, M.-E., CUVILLIER, B., BOUAKAZ, S. et VACHER, M. (2012). Démarche de développement de technologies ambiantes pour le maintien à domicile des personnes dépendantes : vers une triangulation des méthodes et des approches. Dans *Actes du 1er Congrès Européen de Stimulation Cognitive*, pages 121–122, Dijon, France.

BOBILLIER CHAUMON, M.-E. et OPREA CIOBANU, R. (2009). Les nouvelles technologies au service des personnes âgées : entre promesses et interrogations – une revue de questions. *Psychologie Française*, 54(3):271 – 285.

BOERSMA, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

BONASTRE, J.-F., WILS, F. et MEIGNIER, S. (2005). Alize, a free toolkit for speaker recognition. Dans *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 737–740.

BOUAKAZ, S., VACHER, M., BOBILLIER-CHAUMON, M.-E., AMAN, F., BEKKADJA, S., PORTET, F., GUILLOU, E., ROSSATO, S., DESSERÉE, E., TRAINEAU, P., VIMON, J.-P. et CHEVALIER, T. (2014). CIRDO : Smart companion for helping elderly to live at home for longer. *IRBM*, 35(2):101–108.

BOURKE, A. et LYONS, G. (2008). A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor. *Medical Engineering and Physics*, 30(1):84 – 90.

BRUSH, A. B., LEE, B., MAHAJAN, R., AGARWAL, S., SAROIU, S. et DIXON, C. (2011). Home automation in the wild : Challenges and opportunities. Dans *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2115–2124, New York, NY, USA. ACM.

CAMPBELL, N. (2009). The expanding role of prosody in speech communication technology. *DiaHolmia*, page 17. Stockholm, Sweden, keynote talk.

CAON, D. R., AMEHRAYE, A., RAZIK, J., CHOLLET, G., ANDREAO, R. V. et MOKBEL, C. (2010). Experiments on acoustic model supervised adaptation and evaluation by k-fold cross validation technique. Dans *I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on*, pages 1–4. IEEE.

CHAHUARA, P., PORTET, F. et VACHER, M. (2013). Making context aware decision from uncertain information in a smart home : A Markov logic network approach. Dans *Ambient Intelligence*, pages 78–93. Springer.

CHASTAGNOL, C. (2013). *Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes*. Thèse de doctorat, Université Paris Sud-Paris XI.

CHEN, J., KAM, A., ZHANG, J., LIU, N. et SHUE, L. (2005). Bathroom activity monitoring based on sound. Dans GELLERSEN, H.-W., WANT, R. et SCHMIDT, A., éditeurs : *Pervasive Computing*, volume 3468 de *Lecture Notes in Computer Science*, pages 47–61. Springer Berlin Heidelberg.

CHEN, K., CHAN, A. et CHAN, S. (2012). Gerontechnology acceptance by older Hong Kong people. *Gerontechnology*, 11(2).

CLARCKE, A. C. (1968). *2001 : l'Odyssée de l'espace*. J'ai lu. Traduit par M. Demuth.

COLLÈGE NATIONAL DES ENSEIGNANTS DE GÉRIATRIE (2000). *Corpus de gériatrie*, chapitre 3 - La personne âgée malade, pages 33–39. ISBN :2-909710-12-12.

DAVIS, F. D., BAGOZZI, R. P. et WARSHAW, P. R. (1989). User acceptance of computer technology : A comparison of two theoretical models. *Management Science*, 35(8):982–1003.

DECOSTER, W. et DEBRUYNE, F. (2000). Longitudinal voice changes : facts and interpretation. *Journal of Voice*, 14(2):184–193.

DEGEN, T., JAECKEL, H., RUFER, M. et WYSS, S. (2003). Speedy :a fall detector in a wrist watch. Dans *Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on*, pages 184–187.

DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

DROSOS, K., FLOROS, A., AGAVANAKIS, K., TATLAS, N.-A. et KANELLOPOULOS, N.-G. (2012). Emergency voice/stress-level combined recognition for intelligent house applications. Dans *Audio Engineering Society Convention 132*.

DUCLOT, W. (2014). Système autonome d'acquisition et de traitement temps-réel multivoie de signaux sonores. Rapport de stage, IUT2 de Grenoble, Université Pierre-Mendès-France.

DUÉE, M. et REBILLARD, C. (2006). La dépendance des personnes âgées : une projection en 2040. *Données sociales - La société française*, pages 613–619.

DUFOUR, R., JOUSSE, V., ESTÈVE, Y., BÉCHET, F. et LINARÈS, G. (2009). Spontaneous speech characterization and detection in large audio database. Dans *13-th International Conference on Speech and Computer (SPECOM 2009)*, St Petersburg (Russia).

DUGHEANU, R.-C. (2011). Acquisition de corpus pour la génération de modèles acoustiques adaptés à la voix des personnes âgées. Rapport de stage, Université Stendhal Grenoble 3, Master 2 IDL.

FERRAGNE, E., FLAVIER, S. et FRESSARD, C. (2013). Rocme ! software for the recording and management of speech corpora. Dans *Proceedings of Interspeech 2013*, pages 1864–1865. ISCA.

FERRAND, C. T. (2002). Harmonics-to-noise ratio : an index of vocal aging. *Journal of Voice*, 16(4):480–487.

FLEURY, A., NOURY, N., VACHER, M., GLASSON, H. et SERIGNAT, J.-F. (2008). Sound and speech detection and classification in a health smart home. Dans *30th IEEE EMBS Annual International Conference*, pages 4644–4647, Vancouver, British Columbia, Canada.

FRANCO, A. (2012). Conférence invitée : Nouveaux paradigmes et technologies pour la santé et l'autonomie (invited conference : New paradigms and technologies for health and autonomy) [in french]. Dans *JEP-TALN-RECITAL 2012, Workshop ILADI 2012 : Interactions Langagières pour personnes Agées Dans les habitats Intelligents (ILADI 2012 : Language Interaction for Elderly in Smart Homes)*, pages 1–2, Grenoble, France. ATALA/AFCP.

FURUI, S. (2003). Recent advances in spontaneous speech recognition and understanding. Dans *In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–6.

FURUI, S., NAKAMURA, M., ICHIBA, T. et IWANO, K. (2005). Why is the recognition of spontaneous speech so hard ? Dans MATOUŠEK, V., MAUTNER, P. et PAVELKA, T., éditeurs : *Text, Speech and Dialogue*, volume 3658 de *Lecture Notes in Computer Science*, pages 9–22. Springer Berlin Heidelberg.

GALES, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

GAUVAIN, J.-L. et LEE, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *Speech and audio processing, IEEE transactions on*, 2(2):291–298.

GEROSA, M., GIULIANI et D., Brugnara, F. (2009). Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499–509.

GIANNOULIS, D., STOWELL, D., BENETOS, E., ROSSIGNOL, M., LAGRANGE, M. et PLUMBLEY, M. D. (2013). A database and challenge for acoustic scene classification and event detection. *European Signal Processing Conf.*

GLASSON, H. (2008). Système autonome d'acquisition temps-réel multivoie de signaux sonores dans un Habitat Intelligent Santé (HIS). Mémoire de fin d'études, CNAM Rhône-Alpes, CUEFA - Grenoble Universités.

HAGEN, P., LYONS, G. D. et NUSS, D. W. (1996). Dysphonia in the elderly : diagnosis and management of age-related voice changes. *Southern medical journal*, 89(2):204–207.

HAMILL, M., YOUNG, V., BOGER, J. et MIHAILIDIS, A. (2009). Development of an automated speech recognition interface for personal emergency response systems. *Journal of NeuroEngineering and Rehabilitation*, 6(1):26.

HATON, J.-P., CERISARA, C., FOHR, D., LAPRIE, Y. et SMAÏLI, K. (2006). *Reconnaissance automatique de la parole : Du signal à son interprétation*. Dunod.

HEIGOLD, G., MACHEREY, W., SCHLUTER, R. et NEY, H. (2005). Minimum exact word error training. Dans *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 186–190. IEEE.

HIRANO, M., KURITA, S., NAKASHIMA, T. et coll. (1983). Growth, development and aging of human vocal folds. *Vocal fold physiology : Contemporary research and clinical issues*, pages 22–43.

HIRANO, M., KURITA, S., YUKIZANE, K. et HIBI, S. (1989). Asymmetry of the laryngeal framework : a morphologic study of cadaver larynges. *The Annals of otology, rhinology, and laryngology*, 98(2):135–140.

HOIT, J. D. et HIXON, T. J. (1987). Age and speech breathing. *Journal of Speech, Language, and Hearing Research*, 30(3):351–366.

HOLLIEN, H. et SHIPP, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech and Hearing Research*, 15:155–159.

HONJO, I. et ISSHIKI, N. (1980). Laryngoscopic and voice characteristics of aged persons. *Archives of Otolaryngology*, 106(3):149–150.

HOUTGAST, T. et STEENEKEN, H. J. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077.

INSEE (2012). France, portrait social. *Insee Références - Édition 2012*.

INTILLE, S. S. (2002). Designing a home of the future. *IEEE pervasive computing*, 1(2):76–82.

ISTRATE, D. (2003). *Détection et Reconnaissance des Sons pour la Surveillance Médicale*. Thèse de doctorat, INP Grenoble, École Doctorale « Électronique, Électrotechnique, Automatique, Télécommunications, Signal ».

ISTRATE, D., CASTELLI, E., VACHER, M., BESACIER, L. et SERIGNAT, J.-F. (2006). Information extraction from sound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions*, 10:264–274.

JACEWICZ, E., FOX, R. A., O'NEILL, C. et SALMONS, J. (2009). Articulation rate across dialect, age, and gender. *Language variation and change*, 21(2):233–256.

KATZ, S. et AKPOM, C. A. (1976). A measure of primary sociobiological functions. *International journal of health services*, 6(3):493–508.

KAWAHARA, T., NANJO, H., SHINOZAKI, T. et FURUI, S. (2003). Benchmark test for speech recognition using the corpus of spontaneous japanese. Dans *In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 135–138.

KENT, R. D. et VORPERIAN, H. K. (1995). *Development of the craniofacial-oral-laryngeal anatomy*. Singular Publishing Group San Diego.

KESHAVARZ, A. (2006). Distributed vision-based reasoning for smart home care. Dans *ACM SenSys Workshop on Distributed Smart Cameras (DSC'06)*.

KREIMAN, J. et SIDTIS, D. (2011). *Foundations of voice studies : An interdisciplinary approach to voice production and perception*. Wiley-Blackwell.

KRIJNDERS, J. D. et GINEKE, A. (2013). Tone-fit and mfcc scene classification compared to human recognition. *Energy [dB]*, 400(450):500.

KUBIAK, Y. et LORRAINE, A. (2012). Personnes âgées dépendantes : le maintien à domiciles, solution privilégiée mais exigeante. *Économie Lorraine, Insee*, 289:1–5.

LAMEL, L., GAUVAIN, J. et ESKENAZI, M. (1991). BREF, a large vocabulary spoken corpus for french. Dans *Proceedings of EUROSPEECH 91*, volume 2, pages 505–508, Geneva, Switzerland.

LASS, N. (2012). *Contemporary issues in experimental phonetics*. Elsevier.

LAUKKA, P., AUDIBERT, N. et AUBERGÉ, V. (2007). Graded structure in vocal expression of emotion : What is meant by "prototypical expressions". Dans *1st International Workshop on Paralinguistic and Speech–Between Models and Data*.

LAWTON, M. P. et BRODY, E. M. (1969). Assessment of older people : self-maintaining and instrumental activities of daily living. *Gerontologist*, 9:179–186.

LECOUTEUX, B. (2008). *Reconnaissance automatique de la parole guidée par des transcriptions a priori*. Thèse de doctorat, Université D'Avignon et des Pays de Vaucluse, Ecole Doctorale « Mathématique et Informatique ».

LECOUTEUX, B., LINARES, G., ESTEVE, Y. et GRAVIER, G. (2013). Dynamic combination of automatic speech recognition systems by driven decoding. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(6):1251–1260.

LECOUTEUX, B., LINARES, G., ESTEVE, Y. et MAUCLAIR, J. (2007). System combination by driven decoding. Dans *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–341. IEEE.

LECOUTEUX, B., VACHER, M. et PORTET, F. (2011). Distant speech recognition in a smart home : Comparison of several multisource ASRs in realistic conditions. Dans *12th International Conference on Speech Science and Speech Technology (InterSpeech 2011)*, pages 2273–2276, Florence, Italy.

LEFOL, Q. (2010). Acquisition de corpus pour la génération de modèles acoustiques adaptés à la voix des personnes âgées. Mémoire de stage M1, Institut National Polytechnique de Grenoble, PHELMA, filière SEI.

LEGGETTER, C. et WOODLAND, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171 – 185.

LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady*, 10:707–710.

LINARÈS, G., MASSONIÉ, D., NOCÉRA, P. et LÉVY, C. (2007). A scalable system for embeded large vocabulary continuous speech recognition. Dans *IEEE Workshop on DSP in Mobile and vehicular systems*, Istanbul, Turkey.

LINDEMANN, U., HOCK, A., STUBER, M., KECK, W. et BECKER, C. (2005). Evaluation of a fall detector based on accelerometers : A pilot study. *Medical and Biological Engineering and Computing*, 43(5):548–551.

LINVILLE, S. E. et RENS, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice*, 15(3):323–330.

LISKER, L. (1970). Supraglottal air pressure in the production of english stops. *Language and speech*, 13(4):215–230.

LITVAK, D., ZIGEL, Y. et GANNOT, I. (2008). Fall detection of elderly through floor vibrations and sound. Dans *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4632–4635.

LIUTKUS, A., DURRIEU, J.-L., DAUDET, L. et RICHARD, G. (2013). An overview of informed audio source separation. Dans *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pages 1–4. IEEE.

MAHMOOD, A., YAMAMOTO, T., LEE, M. et STEGGELL, C. (2008). Perceptions and Use of Gerotechnology : Implications for Aging in Place. *Journal of Housing For the Elderly*, 22(1-2):104–126.

MCCLOSKEY, D. W. (2006). The importance of ease of use, usefulness, and trust to online consumers : An examination of the technology acceptance model with older customers. *JOEUC*, 18(3):47–65.

MORGAN, E. E. et RASTATTER, M. (1986). Variability of voice fundamental frequency in elderly female speakers. *Perceptual and motor skills*, 63(1):215–218.

MORRIS, R. J. et BROWN JR, W. (1994a). Age-related differences in speech intensity among adult females. *Folia Phoniatrica et Logopaedica*, 46(2):64–69.

MORRIS, R. J. et BROWN JR, W. (1994b). Age-related differences in speech variability among women. *Journal of Communication Disorders*, 27(1):49–64.

NANJO, H. et KAWAHARA, T. (2002). Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. Dans *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–725–I–728.

NYAN, M., TAY, F. E. et MURUGASU, E. (2008). A wearable system for pre-impact fall detection. *Journal of Biomechanics*, 41(16):3475 – 3481.

PAN, S. et JORDAN-MARSH, M. (2010). Internet use intention and adoption among Chinese older adults : From the expanded technology acceptance model perspective. *Computers in Human Behavior*, 26(5):1111 – 1119.

PELLEGRINI, T., HÄMÄLÄINEN, A., de MAREÜIL, P. B., TJALVE, M., TRANCOSO, I., CANDEIAS, S., DIAS, M. S. et BRAGA, D. (2013). A corpus-based study of elderly and young speakers of european portuguese : acoustic correlates and their impact on speech recognition performance. Dans *Proceedings of Interspeech 2013*, pages 852–856, Lyon, France.

PELLEGRINI, T., TRANCOSO, I., HÄMÄLÄINEN, A., CALADO, A., DIAS, M. S. et BRAGA, D. (2012). Impact of Age in ASR for the Elderly : Preliminary Experiments in European Portuguese. Dans *Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, pages 139–147.

PINQUIER, J., ROUAS, J.-L. et ANDRÉ-OBRECHT, R. (2003). Fusion de paramètres pour une classification automatique parole/musique robuste. *Technique et science informatiques (TSI) : Fusion numérique/symbolique*, 8:831–852.

POMPONIO, L., LE GOC, M., ANFOSSO, A. et PASCUAL, E. (2012). Levels of abstraction for behavior modeling in the gerhome project. *Int. J. E-Health Med. Commun.*, 3(3):12–28.

POPESCU, M., LI, Y., SKUBIC, M. et RANTZ, M. (2008). An acoustic fall detector system that uses sound height information to reduce the false alarm rate. Dans *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4628–4631.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P. et coll. (2011). The KALDI speech recognition toolkit. Dans *Proc. ASRU*, pages 1–4.

POVEY, D. et WOODLAND, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. Dans *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–105. IEEE.

PRINCIPI, E., SQUARTINI, S., PIAZZA, F., FUSELLI, D. et MAURIZIO, B. (2013). A distributed system for recognizing home automation commands and distress calls in the Italian language. Dans *Proceedings of Interspeech 2013*, pages 2049–2053, Lyon, France.

PRIVAT, R., VIGOUROUX, N. et TRUILLET, P. (2004). Etude de l'effet du vieillissement sur les productions langagières et sur les performances en reconnaissance automatique de la parole. *Revue Parole*, 31-32:281–318.

PTACEK, P. H., SANDER, E. K., MALONEY, W. H. et JACKSON, C. R. (1966). Phonatory and related changes with advanced age. *Journal of Speech, Language, and Hearing Research*, 9(3):353–360.

RAMIG, L. A. et RINGEL, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech, Language, and Hearing Research*, 26(1):22–30.

RAMIG, L. O., GRAY, S., BAKER, K., CORBIN-LEWIS, K., BUDER, E., LUSCHEI, E., COON, H. et SMITH, M. (2001). The aging voice : a review, treatment data and familial and genetic perspectives. *Folia Phoniatrica et Logopaedica*, 53(5):252–265.

RASTATTER, M. P., MCGUIRE, R. A., KALINOWSKI, J. et STUART, A. (1997). Formant frequency characteristics of elderly speakers in contextual speech. *Folia Phoniatrica et Logopaedica*, 49(1):1–8.

REYNOLDS, D. A. (1995). Speaker identification and verification using gaussian mixture speaker models. *Speech communication*, 17(1):91–108.

RIALLE, V., OLLIVET, C., GUIGUI, C. et HERVÉ, C. (2008). What do family caregivers of alzheimer's disease patients desire in smart home technologies ? contrasted results of a wide survey. *Methods of Information in Medicine*, 47(1):63–69.

ROUGUI, J., ISTRATE, D. et SOUIDENE, W. (2009). Audio sound event identification for distress situations and context awareness. Dans *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 3501–3504. IEEE.

RUSSELL, A., PENNY, L. et PEMBERTON, C. (1995). Speaking fundamental frequency changes over time in women : A longitudinal study. *Journal of Speech, Language, and Hearing Research*, 38(1):101–109.

RYAN, W. et BURK, K. (1974). Perceptual and acoustic correlates in the speech of males. *Journal of Communication Disorders*, 7:181–192.

SASA, Y. et LEGRAND, J. (2011). Corpus parole de personnes Âgées grenoble 2011, explications, légendes et observations des transcriptions et dispositifs utilisés pour les entretiens. Mémoire de stage, UFR Sciences Du Langage, Master 1 IDL, parcours Traitement Automatique du Langage.

SCHERER, K. R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech Communication*, 40(1–2):227 – 256.

SCHERER, K. R., JOHNSTONE, T. et KLASMEYER, G. (2003). Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456.

SCHULLER, B., BATLINER, A., STEIDL, S. et SEPPI, D. (2011). Recognising realistic emotions and affect in speech : State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087.

SEELYE, A. M., SCHMITTER-EDGECOMBE, M., COOK, D. J. et CRANDALL, A. (2013). Naturalistic assessment of everyday activities and prompting technologies in mild cognitive impairment. *Journal of the International Neuropsychological Society*, 19:442–452.

SEHILI, M., ISTRATE, D., DORIZZI, B. et BOUDY, J. (2012). Daily sound recognition using a combination of gmm and svm for home automation. Dans *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1673–1677. IEEE.

SELBY, J. C., GILBERT, H. R. et LERMAN, J. (2003). Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre-and post-treatment. *Journal of Voice*, 17(4):557–570.

SHINOZAKI, T., HORI, C. et FURUI, S. (2001). Towards automatic transcription of spontaneous presentations. Dans *Eurospeech 2001*, volume 1, pages 491–494.

SHIPP, T., QI, Y., HUNTLEY, R. et HOLLIEN, H. (1992). Acoustic and temporal correlates of perceived age. *Journal of voice*, 6(3):211–216.

SIXSMITH, A. et JOHNSON, N. (2004). A smart sensor to detect the falls of the elderly. *Pervasive Computing, IEEE*, 3(2):42–47.

SOURY, M. et DEVILLERS, L. (2013). Stress detection from audio on multiple window analysis size in a public speaking task. Dans *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 529–533.

STEELE, R., LO, A., SECOMBE, C. et WONG, Y. K. (2009). Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. *International Journal of Medical Informatics*, 78(12):788 – 801. Mining of Clinical and Biomedical Text and Data Special Issue.

SU, B.-H., KUAN, T.-W., WANG, J.-F., FU, P.-W. et CHEN, J.-M. (2014). Mandarin elderly speech corpus with variabilities analysis in speech and language. Dans *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment / CASLRE (Conference on Asian Spoken Language Research and Evaluation*, pages 297–302, Phuket, Thaïland.

THOMAS, L. B., HARRISON, A. L. et STEMPLE, J. C. (2008). Aging thyroarytenoid and limb skeletal muscle : lessons in contrast. *Journal of Voice*, 22(4):430–450.

TIAGO, R., PONTES, P. et do BRASIL, O. C. (2007). Age-related changes in human laryngeal nerves. *Otolaryngology-Head and Neck Surgery*, 136(5):747–751.

VACHER, M., BOUAKAZ, S., BOBILLIER-CHAUMON, M.-E., PORTET, F., AMAN, F., ROSSATO, S. et GUILLOU, E. (2014a). Le projet CIRDO d'assistance aux personnes âgées isolées à domicile. Dans BOURHIS, C. B.-V. J. L. K. G., éditeur : *Congrès Handicap 2014 - 8ème Édition - Les technologies d'assistance : de la compensation à l'autonomie*, pages 229–235, Paris, France.

VACHER, M., CHAHUARA, P., LECOUTEUX, B., ISTRATE, D., PORTET, F., JOUBERT, T., SEHILI, M. E. A., MEILLON, B., BONNEFOND, N., FABRE, S., ROUX, C. et CAFFIAU, S. (2013a). The SWEET-HOME project : Audio technology in smart homes to improve well-being and reliance. Dans *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, pages 7298–7301, Osaka, Japon.

VACHER, M., FLEURY, A., PORTET, F., SERIGNAT, J.-F. et NOURY, N. (2009a). Reconnaissance des sons et de la parole dans un habitat intelligent pour la santé : expérimentations en situation non contrôlée. Dans *GRETSI Traitement du signal et des images*, pages 1–4, Dijon, France.

VACHER, M., FLEURY, A., SERIGNAT, J., NOURY, N. et GLASSON, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. Dans *9th International Conference on Speech Science and Speech Technology (InterSpeech 2008)*, volume 1, pages 496–499, Brisbane, Australia.

VACHER, M., GUIRAND, N., SERIGNAT, J.-F., FLEURY, A. et NOURY, N. (2009b). Speech recognition in a smart home : some experiments for telemonitoring. Dans *SPED 2009*, pages 171–179, Constanta, Romania.

VACHER, M., ISTRATE, D. et SERIGNAT, J. (2004). Sound detection and classification through transient models using wavelet coefficient trees. Dans SUVISOFT LTD, éditeur : *Proc. 12th European Signal Processing Conference (EUSIPCO)*, pages 1171–1174, Vienna, Austria.

VACHER, M., LECOUTEUX, B., CHAHUARA, P., PORTET, F., MEILLON, B. et BONNEFOND, N. (2014b). The Sweet-Home speech and multimodal corpus for home automation interaction. Dans *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

VACHER, M., LECOUTEUX, B., ISTRATE, D., JOUBERT, T., PORTET, F., SEHILI, M. et CHAHUARA, P. (2013b). Experimental evaluation of speech recognition technologies for voice-based home automation control in a smart home. Dans *4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT2013)*, pages 99–105, Grenoble, France.

VACHER, M., LECOUTEUX, B. et PORTET, F. (2012a). Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment. Dans *European Signal Processing Conference (EUSIPCO)*, pages 1663–1667, Bucarest, Romania.

VACHER, M., LECOUTEUX, B., PORTET, F. et coll. (2014c). Multichannel automatic recognition of voice command in a multi-room smart home : an experiment involving seniors and users with visual impairment. *Proceedings of Interspeech 2014*, pages 1–5.

VACHER, M., PORTET, F., FLEURY, A. et NOURY, N. (2011). Development of audio sensing technology for ambient assisted living : Applications and challenges. *International Journal of E-Health and Medical Communications*, 2(1):35 – 54.

VACHER, M., PORTET, F., LECOUTEUX, B. et GOLANSKI, C. (2013c). *Telehealthcare Computing and Engineering : Principles and Design*, chapitre Speech Analysis for Ambient Assisted Living : Technical and User Design of a Vocal Order System, pages 607–638. Numéro 21. CRC Press, Taylor and Francis Group. ISBN : ISBN-978-1-57808-802-7.

VACHER, M., PORTET, F., ROSSATO, S., AMAN, F., GOLANSKI, C. et DUGHEANU, R. (2012b). Speech-based interaction in an AAL context. *Gerontechnology*, 11:310.

VACHER, M., SERIGNAT, J.-F. et CHAILLOL, S. (2007). Sound Classification in a Smart Room Environment : an Approach using GMM and HMM Methods. Dans *The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD 2007), Publishing House of the Romanian Academy (Bucharest)*, volume 1 de *Advances in Spoken Language Technology*, pages 135–146, Iasi, Romania.

VACHER, M., SERIGNAT, J.-F., CHAILLOL, S., ISTRATE, D. et POPESCU, V. (2006). Speech and sound use in a remote monitoring system for healthcare. Dans SOJKA, P., KOPEČEK, I. et PALA, K., éditeurs : *Text, Speech and Dialogue*, volume 4188 de *Lecture Notes in Computer Science*, pages 711–718. Springer Berlin Heidelberg.

VALIN, J.-M. (2006). Speex : a free codec for free speech. Dans *Australian National Linux Conference, Dunedin, New Zealand*.

VAUDABLE, C. (2012). *Analyse et reconnaissance des émotions lors de conversations de centres d'appels*. Thèse de doctorat, Université Paris Sud-Paris XI.

VAUFREYDAZ, D. (1998). EMACOP, un environnement informatique pour l'acquisition et la gestion de grands corpus de parole. Mémoire de D.E.A., DEA d'Informatique : systèmes et communication. Université Joseph Fourier.

VAUFREYDAZ, D., AKBAR, M., CAELEN, J. et SERIGNAT, J. (1998). EMACOP : Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole. Dans *Journées d'Etude sur la Parole JEP'98*, pages 175–178, Martigny, Suisse.

VAUFREYDAZ, D., BERGAMINI, C., SERIGNAT, J.-F., BESACIER, L., AKBAR, M. et coll. (2000). A new methodology for speech corpora definition from internet documents. Dans *LREC'2000 (Language Resources & Evaluation international Conference)*.

VIDRASCU, L. (2007). *Analyse et détection des émotions verbales dans les interactions orales*. Thèse de doctorat, Université Paris Sud-Paris XI, Discipline : Informatique.

VINCENT, E., ARAKI, S., THEIS, F., NOLTE, G., BOFILL, P., SAWADA, H., OZEROV, A., GOWREESUNKER, V., LUTTER, D. et DUONG, N. Q. K. (2012). The signal separation evaluation campaign (2007-2010) : Achievements and remaining challenges. *Signal Processing*, 92(8): 1928–1936.

VIPPERLA, R. (2011). *Automatic Speech Recognition for ageing voices*. Thèse de doctorat, University of Edinburgh.

VIPPERLA, R., RENALS, S. et FRANKEL, J. (2008). Longitudinal study of ASR performance on ageing voices. Dans *Proceedings of Interspeech 2008*, pages 2550–2553, Brisbane, Australia.

VIPPERLA, R., RENALS, S. et FRANKEL, J. (2010). Ageing voices : The effect of changes in voice parameters on ASR performance. *EURASIP Journal of Audio, Speech and Music Processing*, 2010:5 :1–5 :10.

VLASENKO, B., PRYLIPKO, D., PHILIPPOU-HÜBNER, D. et WENDEMUTH, A. (2011). Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. Dans *Proceedings of Interspeech 2011*, pages 1577–1580.

VLASENKO, B., PRYLIPKO, D. et WENDEMUTH, A. (2012). Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. *Proc. of the KI 2012*.

VU, N. T., SCHULTZ, T. et POVEY, D. (2012). Modeling gender dependency in the subspace gmm framework. Dans *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4345–4348. IEEE.

VUEGEN, L., BROECK, B. V. D., KARSMAKERS, P., GEMMEKE, J., VANRUMSTE, B. et HAMME, H. V. (2013). An mfcc-gmm approach for event detection and classification. *AASP Challange : Detection and Classification of Acoustic Scenes and Events*.

WEISMER, G. et LISS, J. (1991). Reductionism is a dead-end in speech research : perspectives on a new direction. *Dysarthria and Apraxia of Speech : Perspectives On Management*, pages 15–27.

WILPON, J. et JACOBSEN, C. (1996). A study of speech recognition for children and the elderly. Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 349–352.

WOELFEL, M. et MCDONOUGH, J. (2009). *Distant Speech Recognition*. Wiley.

WOO, P., CASPER, J., COLTON, R. et BREWER, D. (1992). Dysphonia in the aging : physiology versus disease. *The Laryngoscope*, 102(2):139–144.

XUE, S. A. et DELIYSKI, D. (2001). Effects of aging on selected acoustic voice parameters : Preliminary normative data and educational implications. *Educational Gerontology*, 27(2):159–168.

YAN, Z.-J., ZHU, B., HU, Y. et WANG, R.-H. (2008). Minimum word classification error training of HMMs for automatic speech recognition. Dans *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4521–4524. IEEE.

ZEMLIN, W. R. (1981). *Speech and Hearing Science, Anatomy and Physiology*. Englewood Cliffs, N.J. : Prentice-Hall. 2nd ed.

ZOUBA, N., BREMOND, F., THONNAT, M., ANFOSSO, A., PASCUAL, È., MALLÉA, P., MAILLAND, V. et GUERIN, O. (2009). A computer system to monitor older adults at home : Preliminary results. *Gerontechnology*, 8(3).

Bibliographie personnelle

Articles dans des revues avec comité de lecture

[1] BOUAKAZ S., VACHER M., BOBILLIER-CHAUMON M.-E., AMAN F., BEKKADJA S., PORTET F., GUILLOU E., ROSSATO S., DESSERÉE E. ET TRAINEAU P. (2014). CIRDO : Smart companion for helping elderly to live at home for longer. Dans *IRBM*, 35,(2) :101–108.

[2] VACHER, M., PORTET, F., ROSSATO, S., AMAN, F., GOLANSKI, C. ET DUGHEANU, R. (2012). Speech-based interaction in an AAL context. Dans *Gerontechnology*, 11(2) :310.

Conférences internationales avec comité de lecture

[3] AMAN F., AUBERGE, V. ET VACHER M. (2013). How affects can perturbe the automatic speech recognition of domotic interactions. Dans *Workshop on Affective Social Speech Signals*, pages 1–5, Grenoble, France.

[4] AMAN, F., VACHER, M., ROSSATO, S. ET PORTET, F. (2013). Analysing the Performance of Automatic Speech Recognition for Ageing Voice : Does it Correlate with Dependency Level ? Dans *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 9–15, Grenoble, France.

[5] AMAN, F., VACHER, M., ROSSATO, S. ET PORTET, F. (2013). In-home detection of distress calls : the case of aged users. Dans *the 14rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2013*, pages 2065–2067, Lyon, France.

[6] AMAN, F., VACHER, M., ROSSATO, S. ET PORTET, F. (2013). Speech Recognition of Aged Voices in the AAL Context : Detection of Distress Sentences. Dans *The 7th International Conference on Speech Technology and Human-Computer Dialogue, IEEE, SpeD 2013*, pages 177–184, Cluj-Napoca, Romania.

[7] AMAN, F., VACHER, M., ROSSATO, S., DUGHEANU, R., PORTET, F., LEGRAND, J. ET SASA, Y. (2012). Etude de la performance des modèles acoustiques pour des voix de personnes âgées en vue de l'adaptation des systèmes de RAP (Assessment of the acoustic models performance in the ageing voice case for ASR system adaptation) [in French]. Dans *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 1 : JEP, pages 707–714, Grenoble, France. ATALA/AFCP.

[8] VACHER, M., PORTET, F., ROSSATO, S., AMAN, F., GOLANSKI, C. ET DUGHEANU, R. (2012). Speech-based interaction in an AAL context. Dans *ISG*ISARC2012*, pages 1–7, Eindhoven, The Netherlands.

Conférences nationales avec comité de lecture

[9] VACHER M., BOUAKAZ S., BOBILLIER-CHAUMON M.-E., PORTET F., AMAN F., ROSSATO S., GUILLOU E. (2014). Le projet CIRDO d'assistance aux personnes âgées isolées à do-

micile. Dans *Congrès Handicap 2014 - 8ème Édition - Les technologies d'assistance : de la compensation à l'autonomie*, pages 1–6, France.

[10] AMAN, F., VACHER, M., ROSSATO, S. ET PORTET, F. (2012). Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole (Contribution to the study of elderly people's voice variability in automatic speech recognition) [in French]. Dans *JEP-TALN-RECITAL 2012, Atelier ILADI 2012 : Interactions Langagières pour personnes Agées Dans les habitats Intelligents*, pages 49–59, Grenoble, France. ATALA/AFCP.

[11] LEGRAND, J., AMAN, F., VACHER, M., ROSSATO, S. ET PORTET, F. (2012). Utilisation de la Reconnaissance Automatique de la Parole pour l'aide à l'autonomie des personnes âgées. Dans *Actes de l'Université d'été de la E-santé*, pages 19–21, Castres-Mazamet, France.

Index

Glossaire

A

A* L'algorithme A* est un algorithme de recherche de chemin dans un graphe entre un nœud initial et un nœud final. Il utilise une évaluation heuristique sur chaque nœud pour estimer le meilleur chemin le traversant, et visite ensuite les nœuds selon l'ordre de cette évaluation heuristique.

AD80 (Anodin Détresse 80). Corpus de parole lue de personnes jeunes et âgées contenant des phrases de détresse, des appels aux aidants et des phrases anodines.

AGGIR Voir Grille AGGIR.

ALIZE ALIZE est un outil libre et ouvert de reconnaissance du locuteur, il est développé par le Laboratoire d'Informatique d'Avignon et se présente sous la forme d'une librairie logicielle utilisable à partir de programmes écrits en C++ et disponible sous licence LGPL. Elle contient un grand nombre de fonctionnalités relatives au domaine de la reconnaissance de la parole et du locuteur, notamment une implémentation de l'algorithme de Viterbi utilisé lors de la classification par HMM.

ANOVA (ANalysis Of VAriance ou analyse de la variance). Test statistique permettant de vérifier s'il existe des différences statistiquement significatives entre les moyennes de différents échantillons.

API (Application Programming Interface ou interface de programmation). Ensemble normalisé de classes, de méthodes ou de fonctions qui sert d'interface par laquelle un logiciel offre des services à d'autres logiciels.

AuditHIS Logiciel d'analyse sonore fonctionnant en temps-réel sur le flux audio fourni par plusieurs microphones, il a été développé pendant le projet DesdHIS. Il peut acquérir le signal sonore sur 8 canaux simultanément et délivre la meilleure hypothèse de reconnaissance de parole ainsi que la classe de son de la vie courante reconnue.

B

BREF120 Large corpus oral de textes lus, contenant plus de 100 heures de parole produites par 120 locuteurs, tous les textes enregistrés ayant été extraits du journal *Le Monde*.

C

Cepstre Le cepstre est le résultat de la transformée de Fourier inverse appliquée au logarithme naturel du module de la transformée de Fourier d'un signal.

CIRDO (Compagnon Intelligent Réagissant au doigt et à l'œil). Le projet CIRDO Recherche industrielle financé par l'ANR (ANR-2010-TECS-012) est un projet visant à mettre au point un compagnon, ce dernier représente un produit de télélien social augmenté et automatisé par l'intégration de services innovants (reconnaissance automatique de la parole, analyse de situations (scènes) dans un environnement complexe non contrôlé) visant à favoriser l'autonomie et la prise en charge par les aidants, des patients atteints de maladies chroniques ou de la maladie Alzheimer ou apparentées

CirdoX Logiciel d'analyse sonore permettant la détection des phrases cibles dans un flux sonore. Ce logiciel étend les capacités du logiciel AuditHIS.

Clustering (Regroupement hiérarchique). Méthode de classification automatique de données.

CTL Bâtiment du Centre des Technologies du Logiciel, situé sur le Domaine Universitaire de Grenoble.

D

DOMUS Appartement intelligent appartenant à la plateforme d'expérimentation du Laboratoire d'Informatique de Grenoble, situé dans le bâtiment du Centre des Technologies du Logiciel (CTL) sur le Domaine Universitaire de Grenoble.

DTW (Dynamic Time Warping ou déformation temporelle dynamique). Algorithme permettant de mesurer la similarité entre deux suites temporelles qui peuvent varier au cours du temps.

DWT (Discrete Wavelet Transforms ou transformée en ondelettes discrète). Transformée en ondelettes d'un signal qui est basée sur la décomposition de ce même signal en une somme d'ondelettes élémentaires.

E

EHPAD Établissement d'Hébergement pour Personnes Âgées Dépendantes.

EM (Expectation-maximisation algorithm ou espérance-maximisation). Classe d'algorithmes permettant de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables.

ERES38 (Entretiens RESidences 38). Corpus de parole spontanée et de lectures d'un article, enregistré auprès de personnes âgées dans des EHPAD du département de l'Isère (38).

F

F0 Fréquence fondamentale du signal de parole, mesure la hauteur de la voix.

G

GEOD Logiciel d'enregistrement de corpus de parole lue. Est aussi le nom d'une des équipes de recherche du laboratoire CLIPS qui a fusionné avec GETA pour créer l'équipe GETALP. (voir GETALP)

GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole). L'équipe GETALP, résultant de la fusion de l'équipe GEOD et de l'équipe GETA du laboratoire laboratoire CLIPS, est née en 2007 lors de la création du LIG. Le GETALP est une équipe pluridisciplinaire (informaticiens, linguistes, phonéticiens, traducteurs et traiteurs de signaux, ...) dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale).

GIR (Groupe Iso-Ressources). Voir aussi Grille AGGIR. Groupe dans lequel est positionnée la personne évaluée à partir de la grille AGGIR. Il existe 6 groupes, de GIR 1 (perte totale d'autonomie) à GIR 6 (absence de perte d'autonomie).

GMM Gaussian Mixture Model ou modèle de mélange gaussien : modèle statistique exprimé selon une densité mélange de plusieurs distributions qui suivent la loi normale.

Google Speech API Système de reconnaissance automatique de la parole du navigateur Google Chrome.

GRePS (Groupe de Recherche en Psychologie Sociale). Le laboratoire GRePS est une unité de recherche fondée en 2007. Il regroupe des psychologues sociaux et du travail de l'Institut de Psychologie et de l'Institut des sciences de l'éducation et de formation de l'Université Lyon 2.

Grille AGGIR La grille nationale AGGIR (Autonomie Gérontologie Groupes Iso-Ressources) permet d'évaluer le degré de perte d'autonomie ou le degré de dépendance physique ou psychique d'une personne âgée. Cette grille est utilisée pour l'attribution de l'Allocation Personnalisée d'Autonomie (APA).

GTK+ (The GIMP Toolkit). GTK+ est un ensemble de bibliothèques logicielles permettant de réaliser des interfaces graphiques. Cette bibliothèque a été développée originellement pour les besoins du logiciel de traitement et de retouche d'images GIMP. GTK+ est maintenant utilisé dans de nombreux projets, dont les environnements de bureau GNOME, Xfce et ROX. GTK+ est un projet libre (sous license GNU LGPL) et multi plateforme.

H

HIS (Habitat Intelligent pour la Santé). Le concept d'Habitat Intelligent pour la Santé vise au maintien à domicile de personnes handicapées, malades ou âgées. Des systèmes de télémédecine offrent une réponse graduée, donc adaptée à chaque cas, pour sécuriser le patient sans envahir inutilement sa vie quotidienne.

HMM (Hidden Markov Model ou modèle de Markov caché). Un HMM est un cas particulier des modèles stochastiques graphiques. Un HMM est caractérisé par un double processus stochastique : un processus interne non observable $X(t)$ et un processus externe $Y(t)$ qui génère une observation pour chaque état. Ces 2 chaînes se combinent pour former le processus stochastique.

HNR (Harmonic to Noise Ratio, ou rapport harmonique sur bruit). Rapport entre la partie périodique de l'énergie d'un signal et la partie non périodique, caractérise l'harmonicité d'un signal sonore.

J

Jitter Mesure l'instabilité à court terme de la voix, instabilité se traduisant par des variations de fréquence entre chaque cycle d'oscillation.

K

k-fold ("k-fold cross-validation" ou "leave-one-out cross-validation" ou LOOCV ou "validation croisée"). Technique permettant d'estimer la performance d'un classifieur. Par exemple, (n-1) observations sont utilisées pour l'apprentissage, puis le modèle est validé sur la énième observation, ensuite cette opération est répétée n fois jusqu'à explorer l'ensemble de la base des observations.

kNN (k-nearest neighbor ou k plus proches voisins). Méthode de classification non paramétrique.

L

LIG (Laboratoire d'Informatique de Grenoble). Laboratoire créé en 2007. Le LIG, regroupant environ 600 personnes, est sous la tutelle conjointe de l'Université Joseph-Fourier - Grenoble 1, de l'Université Pierre-Mendès-France - Grenoble II, de l'Institut Polytechnique de Grenoble, et du CNRS. Il est associé à l'Université Stendhal - Grenoble III et est partenaire de l'INRIA. Les activités de recherche du Laboratoire d'Informatique de Grenoble couvrent de nombreux domaines de l'informatique.

LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information). Le LIRIS, regroupant 320 membres, est une unité mixte de recherche dont les tutelles sont le CNRS, l'INSA de Lyon, l'Université Claude Bernard Lyon 1, l'Université Lumière Lyon 2 et l'Ecole Centrale de Lyon. Le champ scientifique de l'unité est l'Informatique et plus généralement les Sciences et Technologies de l'Information.

M

Mel Échelle fréquentielle non-linéaire qui prend en compte la manière dont l'oreille humaine perçoit les fréquences : moins sensible aux variations de fréquences pour les sons aigus (haute fréquence) que pour les sons graves.

MFCC (Mel-frequency cepstral coefficients). Les coefficients cepstraux de fréquence en échelle Mel (MFCC) sont très souvent utilisés en reconnaissance automatique de la parole. Le calcul des paramètres MFCC utilise une échelle fréquentielle non- linéaire (échelle Mel) qui prend en compte la manière dont l'oreille humaine perçoit les fréquences. En effet, nous sommes moins sensibles à une évolution de fréquence pour les sons aigus, c'est à dire à haute fréquence, que pour les sons graves.

MLLR (Maximum Likelihood Linear Regression). Méthode d'adaptation des modèles acoustiques permettant de réduire les différences entre conditions d'apprentissage et conditions de test, effectuant une transformation linéaire des moyennes des gaussiennes des modèles HMM.

MMIE (Maximum Mutual Information Estimation). Méthode d'apprentissage des modèles acoustiques cherchant à maximiser la probabilité postérieure de la séquence correcte d'un mot sachant toutes les séquences de mots possibles.

P

Praat Praat est un logiciel libre scientifique gratuit (GNU General Public License) conçu pour la manipulation, le traitement et la synthèse de sons vocaux (phonétique). Il a été conçu à l'Institut de Sciences Phonétiques de l'Université d'Amsterdam par Paul Boersma et David Weenink.

p-value Dans un test statistique, probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie.

R

RAP Reconnaissance Automatique de la Parole ou Automatic Speech Recognition (ASR).

ROC (Receiver Operating Characteristic). Mesure de la performance d'un classificateur binaire. Graphiquement, la courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs.

RSB (Rapport Signal sur Bruit). Le rapport signal sur bruit est un rapport de puissance exprimé en dB (déciBel) selon une échelle logarithmique. $RSB = 10\log\frac{E_{Signal}}{E_{Bruit}}$

S

Shapiro-Wilk Le test statistique de Shapiro-Wilk permet de tester si un échantillon suit une loi normale.

Shimmer Mesure l'instabilité à court terme de l'amplitude, instabilité se traduisant par des variations d'amplitude entre chaque cycle d'oscillation.

Sphinx3 Système de reconnaissance automatique de la parole développé par la *Carnegie Mellon University* (CMU).

T

Test de Tukey HSD (Honest Significant Difference). Test statistique pouvant être utilisé en conjonction avec une ANOVA (analyse post-hoc) afin de déterminer quelles moyennes sont statistiquement différentes l'une de l'autre.

Test t de Welch Test statistique sur l'hypothèse d'égalité de deux moyennes avec deux échantillons de variances inégales.

W

Ward En clustering, la méthode de Ward consiste à regrouper les classes de façon à ce que l'augmentation de l'inertie interclasse soit maximale et que l'augmentation de l'inertie intraclasse soit minimale.

WAV ou WAVE (Contraction de WAVEform audio format). Standard pour enregistrer un signal audio numérique dans un fichier numérique, il a été mis au point par les sociétés Microsoft et IBM. C'est le format le plus courant pour l'audio non compressé sur les plates-formes de Microsoft ; il est également courant sur les systèmes Unix tels que MacOS ou Linux.

Weka (Waikato Environment for Knowledge Analysis ou Environnement Waikato pour l'analyse de connaissances). Suite populaire de logiciels d'apprentissage automatique. Écrite en Java, développée à l'Université de Waikato, Nouvelle-Zélande. Weka est un logiciel libre disponible sous la Licence publique générale GNU (GPL).

WER (Word Error Rate). Métrique d'évaluation des systèmes de RAP calculant le taux d'erreur de mots. $WER = \frac{S+D+I}{N}$, où S = nombre de substitutions, D = nombre d'effacements, I = nombre d'insertions et N = nombre de mots dans la référence.

ANNEXE A

Formulaire de consentement du corpus AD80

Titre de l'étude : Enregistrement corpus de parole

Promoteur de l'étude : l'Université Joseph Fourier de Grenoble représentant l'équipe GETALP du laboratoire LIG

Investigateur Coordinateur de l'étude : Michel Vacher, ingénieur de Recherche CNRS
Téléphone : 04 76 63 57 95 - Courrier électronique : Michel.Vacher@imag.fr

Informations personnelles (au jour de l'enregistrement)

Nom :
Prénom :
Âge :
Sexe :
Langue maternelle :
Accent régional (si nécessaire) :

Consentement de participation pour le participant (à remplir par le participant)

Je soussigné(e), Nom . . . . . Prénom . . . . . , déclare avoir compris le but et les modalités de l'étude qui m'ont été pleinement expliqués par Frédéric Aman.

Les informations relatives au principe de l'étude, et son intérêt m'ont bien été communiquées. Des réponses ont été apportées à toutes mes questions. J'ai disposé d'un délai de réflexion avant de prendre ma décision.

J'accepte de participer volontairement à cette étude. Il m'a bien été précisé que je pouvais refuser d'y participer et que dans le cas d'une participation à celle-ci, je pouvais revenir sur ma décision à tout moment. On m'a expliqué également que j'ai la possibilité de contacter l'investigateur principal de l'étude pour poser des questions à tout moment avant et en cours d'étude.

J'ai bien noté que les données me concernant resteront strictement confidentielles. Je n'autorise leur consultation que par des personnes qui collaborent à la recherche, désignées par Michel Vacher.

J'ai été informé(e) :

- de la nature des informations transmises (âge, sexe, ...) et de la finalité du traitement des données.
- que pour permettre de garantir l'anonymat des données dès leur enregistrement, je ne pourrai pas demander ultérieurement l'effacement des données me concernant.
- que toutes les données seront anonymisées.
- que, conformément à la loi du 09 août 2004, à la fin de l'étude je peux demander à l'investigateur une synthèse des résultats globaux de la recherche.

Je certifie par ailleurs avoir souscrit une assurance de responsabilité civile.

Mon consentement ne décharge pas les organisateurs de la recherche de leurs responsabilités. Je conserve tous mes droits garantis par la loi.

Le	Le .....
Signature de l'intéressé(e) précédée de	Signature de l'investigateur précédée de
la mention "lu et approuvé"	la mention "lu et approuvé"

ANNEXE B

Texte corpus Anodin-Détresse 2004 et AD80 2012

Phrases détresse et appels aux aidants :
aidez-moi
aouh
appelez le samu
appelez les pompiers
appelez les secours
appelez un docteur
appelez un toubib
appelez une ambulance
appelez une infirmière
appelez quelqu'un
au secours
aïe
aïe aïe aïe
docteur
du secours s'il-vous-plaît
faîtes vite
help
infirmière
j'ai besoin d'aide
j'ai besoin d'un docteur
j'ai besoin d'une infirmière
j'ai besoin d'un toubib
j'ai besoin de secours
j'ai la tête qui tourne
j'ai mal
j'ai mal à la tête
j'ai très mal
j'ai un malaise
je me sens mal
je me sens très mal
je ne me sens pas bien
je ne peux plus bouger
je ne suis pas bien
je suis blessé
je suis tombé
ne me laissez pas tout seul
ne me laissez pas toute seule
oh
oh oh
ouh là là
ouh ouh
ouille
s'il-vous-plaît
sos
un docteur
un docteur s'il-vous-plaît
un docteur vite
un toubib
un toubib s'il-vous-plaît
un toubib vite
une infirmière
une infirmière s'il-vous-plaît
une infirmière vite
venez vite
vite
à moi
à l'aide
ça va pas bien
ça va pas bien du tout
ça va pas du tout

Phrases anodines :
avez-vous lu le journal
bonjour madame
bonjour monsieur

ça va bien
ça va très bien
ce n'est pas assez salé
ce livre est intéressant
comment allez vous
dehors il fait beau
il fait beau
il fait nuit
j'ai bien dormi
j'ai mal dormi
comment ça va
ça ira mieux demain
dehors il pleut
entrez
il est tard
il fait soleil
j'ai chaud
j'ai lu le journal
j'ai ouvert la porte
j'ai sommeil
j'allume la lumière
je dois prendre mon médicament
je me suis réveillé tôt
je me suis endormi tout de suite
je n'entend rien
j'ai faim
tu as fermé la porte
j'ai froid
j'ai soif
j'écoute la radio
je ferme la fenêtre
je ferme la porte
je me suis endormi tard
je n'ai pas faim
je n'ai plus faim
je n'ai vu personne
j'ai éteint la lumière
j'ouvre la porte
la chaise est tombée
la lumière est éteinte
la porte est fermée
la porte est ouverte
le café est brûlant
les pâtes sont cuites
le sucre est sur la table
où est la boîte de cachets
j'ouvre la fenêtre
je vais bien
je vais fermer la porte
où est mon verre
où est le sel
où est le sucre
le médecin est venu hier
l'infirmière va venir
où sont mes lunettes
quel temps fait-il dehors
je vais faire la vaisselle
je vais prendre une douche
les pommes de terre sont cuites
je vais prendre un bain
je vais faire couler l'eau
les carottes sont cuites
les patates sont cuites

ANNEXE C

Texte corpus AD80 2013

Phrases détresse et appels aux aidants :
aidez-moi
aouh
appeler le samu
appeler les pompiers
appeler les secours
appeler un docteur
appeler un toubib
appeler une ambulance
appeler une infirmière
appeler quelqu'un
au secours
aïe
aïe aïe aïe
docteur
du secours s'il-vous-plaît
faites vite
help
infirmière
j'ai besoin d'aide
j'ai besoin d'un docteur
j'ai besoin d'une infirmière
j'ai besoin d'un toubib
j'ai besoin de secours
j'ai la tête qui tourne
j'ai mal
j'ai mal à la tête
j'ai très mal
j'ai un malaise
je me sens mal
je me sens très mal
je ne me sens pas bien
je ne peux plus bouger
je ne suis pas bien
je suis blessé
je suis tombé
ne me laissez pas tout seul
ne me laissez pas toute seule
oh
oh oh
ouh là là
ouh ouh
ouille
s'il-vous-plaît
sos
un docteur
un docteur s'il-vous-plaît
un docteur vite
un toubib
un toubib s'il-vous-plaît
un toubib vite
une infirmière
une infirmière s'il-vous-plaît
une infirmière vite
venez vite
vite
à moi
à l'aide
ça va pas bien
ça va pas bien du tout
ça va pas du tout
elio appelle le samu
elio appelle les pompiers
elio appelle les secours
elio appelle un docteur
elio appelle un toubib

elio appelle une ambulance
elio appelle une infirmière
elio appelle quelqu'un
elio appelle ma fille
elio appelle mon fils
qu'est-ce qu'il m'arrive
je ne peux pas me relever
je ne peux pas me relever du tout
elio il faut appeler le samu
elio il faut appeler les pompiers
elio il faut appeler les secours
elio il faut appeler un docteur
elio il faut appeler un toubib
elio il faut appeler une ambulance
elio il faut appeler une infirmière
elio il faut appeler quelqu'un
elio il faut appeler mon fils
elio il faut appeler ma fille
mon dieu mon dieu
ma jambe ne me porte plus
ma jambe ne me porte plus du tout
elio tu peux l'appeler
elio tu peux appeler au samu
elio tu peux appeler aux pompiers
elio tu peux appeler aux secours
elio tu peux téléphoner au docteur
elio tu peux téléphoner au toubib
elio tu peux téléphoner à l'ambulance
elio tu peux téléphoner à l'infirmière
elio tu peux téléphoner à quelqu'un
elio tu peux téléphoner à mon fils
elio tu peux téléphoner à ma fille

Phrases anodines :
avez-vous lu le journal
bonjour madame
bonjour monsieur
ça va bien
ça va très bien
ce n'est pas assez salé
ce livre est intéressant
comment allez vous
dehors il fait beau
il fait beau
il fait nuit
j'ai bien dormi
j'ai mal dormi
comment ça va
ça ira mieux demain
dehors il pleut
entrez
il est tard
il fait soleil
j'ai chaud
j'ai lu le journal
j'ai ouvert la porte
j'ai sommeil
j'allume la lumière
je dois prendre mon médicament
je me suis réveillé tôt
je me suis endormi tout de suite
je n'entends rien
j'ai faim
j'ai fermé la porte
j'ai froid
j'ai soif
j'écoute la radio
je ferme la fenêtre
je ferme la porte
je me suis endormi tard
je n'ai pas faim
je n'ai plus faim
je n'ai vu personne
j'ai éteint la lumière
j'ouvre la porte
la chaise est tombée
la lumière est éteinte
la porte est fermée
la porte est ouverte
le café est brûlant
les pâtes sont cuites
le sucre est sur la table

où est la boîte de cachets
j'ouvre la fenêtre
je vais bien
je vais fermer la porte
où est mon verre
où est le sel
où est le sucre
le médecin est venu hier
l'infirmière va venir
où sont mes lunettes
quel temps fait-il dehors
je vais faire la vaisselle
je vais prendre une douche
les pommes de terre sont cuites
je vais prendre un bain
je vais faire couler l'eau
les carottes sont cuites
les patates sont cuites
oui
c'est ça
non
non ça va pas bien
elle l'a appelé à huit heures
elle lui a téléphoné
il l'a appelé à huit heures
il lui a téléphoné
tu peux l'appeler
l'ambulance est venue
le docteur a appelé
ma fille a appelé
le docteur a téléphoné
l'infirmière a appelé
mon fils a appelé

ANNEXE D

Texte corpus Voice-Age

Henri Poincaré.
ah là là
ah bon
Le sol joue pour les arbres un rôle de réservoir en eau et en sels minéraux.
ah non
ah oui
La qualité de la relation doit exister entre le malade et le médecin quand on est soigné.
aidez moi
aidez-moi
Nous sommes jeudi, le jury est là, au grand complet.
allô
allô c'est bien moi
Comment on appelle un chien qui n'a pas de pattes.
allô c'est moi
allô c'est qui
Quarante huit mille trois cent trente
allô c'est ça
allô heu
Le metteur en scène est un médiateur entre le texte et le public.
allô non
allô oui
Ce département assure les études et recherches sur le poste.
allô qui c'est
allô-allô
Vous savez ce qu'est le jeu des fonctionnaires le lundi matin.
allume la lumière
allume la lumière s'il-te-plaît
Au dos de chaque figure, coller une étiquette.
allumez la lumière
allumez la lumière s'il-vous-plaît
Le serveur apache est le grand vainqueur
allumer la lumière
aouh
Le gardien met alors ses gants car il a froid.
appelez le 18
appelez le samu
Quarante et un mille six cent soixante
appelez les pompiers
appelez les secours
Leur production n'est pas plus étonnante que celle des parfums et des couleurs dans la plante.
appelez quelqu'un
appelez le docteur
Pourquoi les fleuves débordent-ils, mais pas la mer.
appelez un docteur
appelez un toubib
En un mot on reprend la législation qui a été en vigueur
appelez une ambulance
appelez l'infirmière
Pour une révolution lente, nous ne pensons pas que le principal se décide dans le champ électoral.
appelez une infirmière
attention
Vu dans un journal de petites annonces : Aime les enfants.

augmente le chauffage
augmente la température
C'est souvent la courbure de la cornée qui est en cause.
augmente un peu le chauffage
augmente un peu la température
Un système de balises ponctuelles assurera un même niveau.
augmentez le chauffage
augmentez la température
Il faut avoir une capacité créative.
augmentez un peu le chauffage
augmentez un peu la température
ça veut dire que je ne me lève pas trop tôt le matin
au feu
au feu vite
Un forum permet de participer au débat
au mois prochain
au revoir
Les cannibales sont en train de manger deux missionnaires.
au-secours
avez-vous bu
Le centre de la mer et des eaux vient de rééditer son guide
avez-vous lu le journal
aïe
La conception du service public a été bouleversé
aïe aïe
aïe aïe aïe
C'est un marqueur que l'on peut doser dans le sang
baisse le chauffage
baisse la température
Après avoir obtenu leur BTS, les diplômés entrent dans la vie active.
baisse un peu le chauffage
baisse un peu la température
Les théories de la zone monétaire optimale.
baissez le chauffage
baissez la température
Des temps de réponses inacceptables.
baissez un peu le chauffage
baissez un peu la température
Une femme de soixante ans consulte pour des lombalgies basses.
bof
bon
C'est un petit nuage et sa maman.
bah
ben non
J'ai toujours eu peur d'être fusillé comme un chien au détour d'un bois
ben oui
ben si
L'étudiant doit remettre son projet de thèse définitif.
bon après-midi
bonjour
Le syndicat joue son rôle en exerçant sa fonction.
bonjour madame
bonjour mademoiselle
Cette cuisson est adaptée aux grosses pièces.
bonjour monsieur
bonsoir
Tableau deux : exemples de descriptions verbales par dix neuf sujets naïfs.
bonsoir madame
bonsoir mademoiselle
Les modèles de dissipation présentés ici selon deux approches.
bonsoir monsieur
bonne soirée
La comédie est un genre théâtral qui présente des personnages de la vie ordinaire.
c'est bien
c'est bien ça
Cet ouvrage est un essai d'analyse du progrès technique.

c’est bien vrai
c’est bon
Ce délai va de la date du dépôt à la date de priorité.
c’est pas bien
c’est pas bon
Les salariés ne sont donc pas pris en compte.
c’est pas bien du-tout
c’est pas bon du-tout
Les commissaires de police appartiennent au corps de la police nationale.
c’est très bien
c’est chaud
Confucius a dit : L’homme sage apprend de ses erreurs.
c’est très chaud
c’est d’accord
Les réunions du groupe questions théologiques.
c’est entendu
c’est froid
La science change la face de la civilisation par le chemin de fer.
c’est très froid
c’est faux
Le choix est large dans ce magasin de trois étages.
c’est grave
c’est ici
Elle suit le torrent des yeux.
c’est là
c’est là bas
Entre le début et la fin du passage de la lune.
c’est le feu
c’est pas ça
Le droit au travail pour tous est une idée fondamentale.
c’est pas ça du tout
c’est pas d’accord
Il est communément admis que les tests sont la pierre angulaire.
c’est qui à l’appareil
c’est urgent
La mise en oeuvre des ordinateurs selon trois grandes étapes.
c’est vrai
c’est pas vrai
La largeur du lobe de détection et sa hauteur.
c’est vraiment pas bon
c’est ça
Le rapport de recherche est publié.
c’était hier
ça brûle
Les articles ont été soumis par les auteurs.
ça chauffe
ça fait mal
Ce document réunit un ensemble de notes.
ça ira mieux demain
ça ira
La charte de la terre avant-projet de référence date du dix-huit mars.
ça saigne
ça va
Il ne doit être utilisé qu’au cas où.
ça va bien
ça va pas
La vie de groupe n’est plus alors vécue sur le mode de la promiscuité.
ça ne va pas
ça-ne-va pas
Le stage modulaire dans ses grandes lignes.
ça va pas bien
ça ne va pas bien
C’est un point facile a repérer si l’on est en hauteur.
ça va pas bien du-tout
ça ne va pas bien du-tout
On me répondra que le virtuel fait partie du monde réel.
ça va pas-du-tout
ça ne va pas-du-tout
Ils sont reportés sans blancs ni ratures sur un

registre coté.
ça va pas du tout
ça ne va pas du tout
La valorisation optimale des bois de petites dimensions.
ça ne va pas très bien
ça va pas très bien
Les progrès accomplis dans la maîtrise des matériaux.
ça va très bien
ce livre est intéressant
Car ceci est analysé, sans caricature, sur la base des faits.
ce n'est pas assez salé
comment
Une femme de cinquante huit ans consulte.
comment allez vous
comment ça va
Le lion tint conseil, et dit : mes chers amis.
d'accord
de rien
La surface au sol minimale de la cage pour une lapine.
dehors il fait beau
dehors il fait très beau
En ce sens elle se place sous le signe de la disqualification.
dehors il fait chaud
dehors il fait très chaud
Nous avons ouvert la boîte de pandore.
dehors il fait froid
dehors il fait très froid
Le choix rationnel sur les projets.
dehors il pleut
dehors il neige
Dans les grands espaces du nord.
docteur
du secours s'il-vous-plaît
Les membres du bureau doivent être âgés de plus de dix huit ans.
eh non
eh oui
Le haut comité de la santé publique est présidé par le ministre chargé de la santé.
entendu
entrez
Aucun fichier n'est passé en paramètre.
est-ce que vous êtes sûr
euh
Le bureau des affaires générales assure la gestion.
euh comment
euh euh
Tout service requis dans le cas de crises ou de calamités.
euh bon
euh non
Elle est une description sous forme numérique du territoire.
euh oui
éteins la lumière
Enfin je débouchai dans une plaine morne.
éteins la lumière s'il-te-plaît
éteignez la lumière
Le plus court chemin suit le tracé des arcs du réseau.
faites vite
faites le 18
Le centre de documentation est spécialisé dans les domaines de la conservation et de la restauration des livres.
ferme la porte
ferme le verrou
La rémunération de votre épargne est comptabilisée au jour le jour.
ferme les volets
fermez la porte
Le groupe de traitement automatique des langues.
fermez la porte s'il-vous-plaît
fermez le verrou
Il faut favoriser une réflexion générale sur la

mobilité des agents au sein du réseau.
fermez le verrou s'il-vous-plaît
fermez les volets
Ainsi a été contrôlé le respect.
fermez les volets s'il-vous-plaît
help
C'est en fait de l'eau qui s'est évaporée, de la mer par exemple.
ici ça va
il devrait pleuvoir demain
Le titre de la thèse doit refléter et annoncer le contenu.
il-ne devrait pas pleuvoir demain
il pourrait pleuvoir demain
La carte marine se présente sous la même forme que la carte.
demain il pourrait pleuvoir
il est tard
Il peut se déplacer au domicile du témoin.
il se fait tard
il fait beau
Le médecin généraliste traitant était en relation quasi permanente avec le professeur.
il fait chaud
il fait très chaud
On lui attacha les mains derrière le dos avec un fil de fer.
il fait froid
il fait très froid
En basse saison, vous pouvez annuler votre réservation.
il fait nuit
il fait soleil
Nous entrons dans la période de la science.
il fait très beau
il fait pas chaud
Il est plus facile de revendre les parts ou les actions de ces sociétés.
il fait pas très chaud
il faut faire attention
Deux cent quarante deux éléments identifiés.
il-ne fait pas chaud
il-ne fait pas très chaud
Sur la place du capitole on voit la croix occitane.
il neige
il neige dehors
Un sujet original pour ce film.
il neige beaucoup
il neige souvent
Les côtés du stade et les bordures des voies diverses ont été libérés.
il fait beau dehors
il fait très beau dehors
Une très grande place est donnée aux travaux pratiques.
il fait chaud dehors
il fait très chaud dehors
Il me semble parfois être si loin de mon point de départ.
il fait froid dehors
il fait très froid dehors
Le coût de l'appel entrant est le sujet hypersensible.
il pleut
il pleut dehors
à peine si l'on voit dans toute la croisée une vitre sur trois qui ne soit pas brisée.
il pleut beaucoup
il pleut souvent
De manière générale, le bleu est plus dévié que le vert.
il y a le feu
infirmière
Les repas sont tous faits par la cuisinière.
j'ai besoin d'aide
j'ai besoin d'un docteur
Le réseau hertzien est constitué de plus de mille deux cents émetteurs.
j'ai besoin d'un toubib
j'ai besoin d'une infirmière
Dans le cas où le défendeur est une personne

morale.
j'ai besoin de secours
j'ai bien dormi
La renverse de flot a lieu au voisinage de la pleine mer.
j'ai bu de l'eau
j'ai bu de-la bière
Ce qui est en jeu ici, c'est la culture accumulée.
j'ai bu de-la tisane
j'ai bu du café
Mon sujet de thèse porte sur une méthode de conception de systèmes d'information.
j'ai bu du thé
j'ai bu du jus de fruit
Son objectif est de les rassembler.
j'ai bu du vin
j'ai bu ma bière
Je ne sais pas ce que je serai dans deux ou trois ans.
j'ai bu ma tisane
j'ai bu mon café
Il y avait déjà eu des alertes.
j'ai bu mon eau
j'ai bu mon thé
La machine initiatrice de la prise de main distance est appelée le preneur.
j'ai bu mon vin
j'ai chaud
La demande doit en être formulée un mois avant.
j'ai très chaud
j'ai de la fièvre
Ce est en effet dans la perspective d'un système démocratique.
j'ai faim
j'ai fait attention
Le reste, c'est : du pétrole, du gaz, du bois, un peu de charbon et un tout petit peu de solaire et de géothermie.
j'ai pas faim
j'ai pas très faim
Le présent accord entre en vigueur le quinze février pour une durée de un an.
j'ai plus faim
j'ai fermé la fenêtre
Mais il a bien voulu reconnaître la validité de ma démarche.
j'ai fermé la porte
j'ai fermé le verrou
Pour la beauté et le soin de votre regard.
j'ai fermé les volets
j'ai froid
Le passant chagrin, que tu frôles.
j'ai très froid
j'ai la tête qui tourne
Un regard critique doit être porté sur la nature.
j'ai lu le journal
j'ai mal
Il existe beaucoup de situations intermédiaires.
j'ai mal à la tête
j'ai mal à-la-tête
Il est hors des stades et des salles, dans la rue.
j'ai mal au bras
j'ai mal aux dents
Le plus connu de ces thés parfumés est le thé au jasmin.
j'ai mal dormi
j'ai mal au poignet
Le premier numéro du magazine contient un dossier complet sur les personnels enseignants du supérieur.
j'ai mal au ventre
j'ai mal de partout
Je serai riche avec deux gobelets.
j'ai mal entendu
j'ai mal partout
Ce système sera adapté pour traiter le cas général du problème.
j'ai ouvert la fenêtre

j'ai ouvert la porte
La science est aujourd'hui parmi les moteurs du changement.
j'ai ouvert le verrou
j'ai ouvert les volets
Un premier contrôle est réalisé automatiquement
j'ai soif
j'ai sommeil
Le combat s'est élevé dans le coeur de votre femme.
j'ai très mal
j'ai un malaise
Nombre de clic sur un bandeau publicitaire.
j'ai éteint la lumière
j'ai vu personne
Dans tous les cas, ne vous laissez pas prendre au piège de la facilité.
j'ai vraiment mal
j'allume la lumière
Ce malheur s' explique assez par la nature du sol.
j'en peux plus
j'en peux vraiment plus
Les quatre préliminaires sont les quatre pratiques.
j'en ai marre
j'entends mal
Le contrat de prestation de service concerne exclusivement du savoir-faire existant.
j'entends pas bien
j'entends rien
La prostitution des enfants est au coeur du débat de la prostitution.
j'entends rien du tout
j'ouvre la fenêtre
La cellule cent cinquante neuf a été isolée.
j'ouvre la porte
j'écoute la radio
Situé au centre de paris, au coeur du marais.
je bois de l'eau
je bois de-la bière
La pelote basque est issue de la province voisine dont elle porte le nom.
je bois de-la tisane
je bois du café
Si la base du nez est étroite, le pont est fin et haut.
je bois du thé
je bois du jus de fruit
Le monde a été fait par les hommes pour les hommes.
je bois du vin
je bois ma bière
Ce kit pédagogique a été conçu.
je bois ma tisane
je bois mon café
Les cancers de la peau augmentent le plus rapidement.
je bois mon eau
je bois mon thé
La connaissance des modalités de prise en charge.
je bois mon vin
je dois prendre mon médicament
Le nez est assez fruité avec des notes de pomme et de poire.
je ferme la fenêtre
je ferme la porte
C'est la différence de vitesse entre vélos et autos.
je ferme le verrou
je ferme les volets
La bibliothèque des arts décoratifs possède plus de cent vingt volumes.
je lis le journal
je me sens mal
Elle est une madone de décadence.
je me sens pas bien
je me sens pas bien du-tout
On retrouve à peu de chose près le même écart.

je me sens très mal
je me suis cassé quelque chose
Voilà un exemple : on part du sommet de gauche, on va vers le haut.
je me suis coupé
je me suis endormi tard
Nul ne peut être lésé, dans son travail ou son emploi.
je me suis endormi très tard
je me suis endormi tout-de-suite
Il faudra donc prévoir une somme de six mille francs.
je me suis réveillé tôt
je me suis réveillé très tôt
Vous avez retrouvé le numéro.
je me suis tordu la cheville
je n'ai pas faim
La perception du public est un fait qui doit être pris comme tel.
je n'ai pas très faim
je n'ai plus faim
Les stations et terminaux sont utilisés par les élèves lors des enseignements.
je n'ai vu personne
je n'en peux plus
Le monde du jeu de la vie est un plan infini quadrillé.
je n'en peux vraiment plus
je n'entends rien
Lors de son assemblée générale, le bureau a donné sa démission en bloc.
je ne peux pas bouger
je ne veux pas
Le joueur a mis le plus fort atout.
je ne me sens pas bien
je ne me sens pas bien du-tout
La requête est portée devant le président du tribunal.
je ne me sens pas très bien
je ne peux plus bouger
Les logements : des chambres, des studios, des deux pièces.
je ne suis pas bien
je ne suis pas bien du-tout
C'est la base de ce même principe.
je regarde la tv
je regarde la télévision
On aurait pu mettre un immeuble, ou même deux.
je saigne
je sais pas
Le sol de cette plaine était d'un blanc d'ivoire.
je suis blessé
je suis coincé
On code tout d'abord le texte original avec la première clé.
je suis tombé
je vais bien
Or deux jours après la bataille, il ne reste plus trace des corps.
je vais pas bien
je veux bien
Dans cette volonté, la culture joue par ailleurs un rôle primordial.
je ne vais pas bien
je vais pas très bien
Son objectif est de former des spécialistes de haut niveau en informatique.
je vais pas-très-bien
je ne vais pas très bien
Il aura lieu début novembre.
je ne vais pas-très-bien
je vais très bien
Les informations ci-dessous sont éventuellement incomplètes ou erronées.
je vais éteindre la lumière
je vais faire couler l'eau
à la barre, un vieux loup de mers qui a bourlingué de ports en ports.
je vais faire du café
je vais faire du thé
Mon cher monsieur, cette étude n'a pas été le

but de ma vie.
je vais faire de la tisane
je vais faire la vaisselle
La première a été évoquée dès le premier jour.
je vais fermer la fenêtre
je vais fermer la porte
La charte des services bancaires de base a été établie.
je vais fermer le verrou
je vais fermer les volets
Le montage tente lui aussi de faire dans le novateur.
je vais me coucher
je vais prendre un bain
Le prix de transaction est égal au VRAI prix.
je vais prendre une douche
l'infirmière va venir
Le nom de la section est le nom du service.
la chaise est tombée
l'assiette est cassée
Mais en entendant les moqueries de sa femme, le baron la prit par le bras.
l'assiette est tombée
la lumière est éteinte
Les prix ont été établis en francs français.
la lumière est allumée
la porte est fermée
Au cours des vingt dernières années.
la porte est ouverte
la soupe est brûlante
Le module optionnel calcul formel est accessible.
la soupe est chaude
la soupe est très chaude
Elle prête des lits à ceux qui font leur ronde.
la soupe est froide
la soupe est prête
Son goût pour les voyages a sans doute été favorisé par son père.
le café est brûlant
le café est chaud
Le courant de dérive en surface est sensiblement orienté dans le lit du vent.
le café est très chaud
le café est froid
Il existe une multitude de types de code barre.
le café est prêt
le couteau est tombé
La mise en demeure ne peut concerner que les cotisations.
le docteur va venir
le docteur est venu hier
Le cépage, c'est la variété de raisin.
le médecin va venir
le téléphone a sonné
Une volonté de fer dans un corps de fée.
le téléphone sonne
le thé est brûlant
Au fil des années on a vu se créer des différents ministères.
le thé est chaud
le thé est très chaud
Un centre de documentation technique est en cours de mise en place.
le thé est froid
le thé est prêt
Le divin infuse une civilisation.
la tasse est tombée
la tisane est brûlante
Les premières décisions ont été prises
la tisane est chaude
la tisane est très chaude
On a beau faire, le froid est une glu.
la tisane est froide
la tisane est prête
Les lettres et les dessins sont exposés dans cette pièce
le médecin est venu hier
le verre est tombé
Ce séminaire de cinq jours se déroule en

deux modules.
le sucre est sur la table
les carottes sont cuites
Le bloc de téflon sur le fond du socle garantit une plus grande concentration de la lumière.
les patates sont cuites
les pommes de terre sont cuites
Il s'agit d'un vertige visuel du tableau.
les pâtes sont cuites
les volets sont fermés
Ces antigènes peuvent être libérés dans le sang.
le verre est cassé
l'eau est chaude
Frappe ta tête contre une cruche.
l'eau est froide
l'eau est trop chaude
Recherche du patient par les premières lettres du nom.
l'eau est trop froide
là
Le candidat sera rayé des listes des écoles où il était admissible.
monte le chauffage
monte la température
Or, le décalage en temps est approximativement constant.
monte un peu le chauffage
monte un peu la température
Les principes de recherche généraux dans les banques de données.
montez le chauffage
montez la température
Au trot, vus de devant et de derrière, les membres se portent en avant.
montez un peu le chauffage
montez un peu la température
Autisme infantile : syndrome existant dès la naissance.
ne me laissez pas tout seul
ne me laissez pas toute seule
D'autres études ont mesuré le taux de mercure des cheveux.
non
non c'est ça
Marqué par la médecine, il entend soustraire le progrès humain au hasard.
non c'est pas ça
non merci
Mais ces droits de l'homme seront vraiment respectés.
non merci bien
non ça va
Il est moins souvent utile.
non ça va pas
non ça ne va pas
Les élections du vingt et un septembre ont vu émerger le parti démocratique de gauche.
oh
oh là là
Il faut agir contre la violence.
oh oh
ouh là là
En ce qui concerne les repas, il existe des restaurants universitaires.
ouh ouh
oui
C'est la seule voie.
oui c'est ça
oui entrez
La synergie est récente, sans doute trop récente.
oui et non
oui j'arrive
C'est que les nuits sont bien plus longues.
oui merci
oui merci bien
Un nouveau code pénal est en cours de rédaction.
oui ou non
oui sans doute
Par la suite, la détermination de la date opti-

male a été décisive.
oui ça va
ouille
Les temps de calcul ont encore été réduits.
ouille ouille ouille
oui peut-être
Vous voulez faire le point.
oui tout à fait
oui tout-à-fait
Le lecteur pourra donc suivre les tours.
où
où est la boîte de cachets
Visitez notre fameux musée sonore de la synthèse de la parole en français.
où est le sel
où est le sucre
Pour cette création, il a voulu une danse.
où est ma chemise
où est ma culotte
Parmi eux, le mur occidental plus connu sous le nom de mur des lamentations.
où est ma tasse
où est ma tasse à café
Le nom de domaine permet d'être identifié.
où est ma tasse de café
où est ma tasse à thé
Que dire de la difficulté de maintenance de ces filtres.
où est ma tasse de thé
où est ma veste
Je me coucherai dans les citernes et dans les vaisseaux creux.
où est mon assiette
où est mon bonnet
Elle est au coeur de la création pour les arts.
où est mon manteau
où est mon pantalon
La surface derrière le monument.
où est mon pull
où est mon slip
En chêne fendu de vingt sept millimètres d'épaisseur.
où est mon verre
où sont les assiettes
Il pense en effet mentionner le nom et l'adresse du président de commission.
où sont les couteaux
où sont les cuillères
Un développement récent du modèle permet de gérer les bords du tissu.
où sont les flûtes à champagne
où sont les fourchettes
Cela dit, je ne considère pas que la vie est une maladie.
où sont les petites assiettes
où sont les petites cuillères
La gélinotte et le tétras lyre.
où sont les petits couteaux
où sont les verres
Mais je garde une autre image en moi.
où sont les verres à apéritif
où sont les verres à champagne
La ville sainte est la ville du salut par excellence.
où sont les verres à liqueur
où sont mes chaussettes
Donc c'est difficile de gérer le temps de travail des cadres par exemple.
où sont mes chaussures
où sont mes lunettes
Les demandes divisionnaires bénéficient de la date de dépôt.
où sont mes pantoufles
où çà
La simple visite de ces sites représentait un trajet d'une durée de deux jours.
ouvre les volets
ouvrez les volets
Les deux étoiles sont assez rapprochées.
ouvrir les volets
parlez plus fort
Au moyen de ces concepts, nous donnons

une preuve concise.
pas bien
pas très bien
Le système de production est formulé ici dans sa définition la plus large.
pas bien du-tout
pas-du-tout
Le candidat peut utiliser la note obtenue dans ces matières.
pas mal
peut-être
Or, c'est un sujet qui mérite la plus grande attention.
peut-être pas
peut-être que non
Le rôle de délégué est de représenter la diversité.
peut-être que oui
pour l'instant ça va
Prévoyez des repas froids et des boissons.
quel temps fait-il dehors
qui c'est
De même, il ne sera infligé aucune peine plus forte.
qui êtes vous
qui est à l'appareil
En outre, ils doivent disposer de moyens de subsistance suffisants.
qu'est-ce qui va pas
regarde la télévision
Je lui ai fermé les yeux.
s'il-te-plaît
s'il-vous-plaît
Merci pour cet acte de justice.
sos
tout est tombé
Le logiciel dispose également de plusieurs autres fonctions.
tout va bien
un docteur
Il y a aussi des moments de détente et de rencontre.
un docteur s'il-vous-plaît
un docteur vite
Nos travaux ont également fait apparaître une nouvelle classe.
un toubib
un toubib s'il-vous-plaît
Cette demande doit être présentée dans le délai de deux jours.
un toubib vite
une ambulance
Cette mission est menée en liaison étroite.
une ambulance s'il-vous-plaît
une ambulance vite
L'inauguration de la fontaine est alors l'occasion de fêter la république.
une infirmière
une infirmière s'il-vous-plaît
Il faut affirmer que le patient ne peut pas devenir un objet.
une infirmière vite
venez vite
Elle montre un effet des réseaux de transport rapide sur l'espace.
vite
vous êtes sûr
Ce dernier a été élaboré dans le cadre d'un contrat de recherche.
y'a le feu
à ce soir
La manette sur la chaudière gère la montée et la descente de la grille.
à demain
à dimanche
Par exemple, un lancer de six et cinq permet de retirer le jeton onze.
à jeudi
à l'aide
On lui donne la valeur un.
à la semaine prochaine
à lundi

Ainsi, c'est la longueur des jupes.
La sobriété des coupes des vestes en font des produits durables dans le temps.
à mardi
à mercredi
Il doit poser le ballon sur une ligne perpendiculaire.
La ligne de but.
La distance de son choix de cette ligne de but.
à midi
à moi
Ce bâti de dépôt sous vide a été entièrement réalisé.
Il a été développé et mis au point au laboratoire, ainsi que le logiciel de pilotage.
à samedi
à vendredi
Le rôle des départements est reconnu.
Ils sont des partenaires des contrats de plan.
Même si les régions sont chefs de file.
çà et là
çà
Allume la radio
Allume la télé
Allume la télévision
Allumez la musique
Allumez la radio
Allumez la télévision
Appelle ma femme
Appelle ma fille
Appelle mon fils
Appelle mon mari
Arrête
Augmentez la luminosité
Augmentez le son
Augmentez le volume
Baisse le son
Baisse le store
Baissez les volets
Changez de chaîne
Changez de station
Continue
Décroche !
Décroche le téléphone
Diminuez la luminosité
Diminuez le son
Diminuez le volume
Encore
éteignez la musique
éteignez la radio
éteignez la télévision
éteins l'alarme
éteins la radio
éteins la télé
éteins la télévision
Fermez la fenêtre
Moins bas
Moins fort
Moins haut
Moins vite
Monte le son
Monte le store
Ouvre la fenêtre
Ouvrez la porte
Plus bas
Plus fort
Plus haut
Plus vite
Raccroche !
Raccroche le téléphone
Stop
Il est reconnu d'une manière unique et universelle au même titre qu'un nom de marque.
Leur gestion au jour le jour ainsi que leurs évolutions.
En tous lieux vains et fades où gît le goût de la grandeur.
Le théâtre, le conte, la littérature orale et l'écriture ou la chanson.
Des rails de chemin de fer relient le monument au wagon.

C'est le chemin des martyrs.
Il mesure un virgule quatre mètre de haut et a huit cercles galvanisés.
Les spécialistes ne suffisent pas.
La prolifération cellulaire et tissulaire dynamique.
La vie c'est la santé et quelque chose de merveilleux.
Le lagopède alpin est un gallinacé.
Dans la famille c'est lui qui monte le plus haut.
Une image que les gens ont bien vue.
C'est quand je soulève cette coupe du monde.
Là où le christ a marché, a souffert, où il est mort et ressuscité.
Il faut répondre aux exigences du code du travail.
Le cas échéant.
La date de priorité de la demande initiale.
La route fait quatre cents kilomètres environ.
La matière est située près de la pointe de ce lobe.
Elle est attirée vers la naine blanche.
La validité de la conjecture pour le cas de trois machines.
Concept central du livre.
Le secteur rural.
Le cadre des examens de licence et de maîtrise en droit.
Car le numérique change la nature même de la copie.
Vous le savez.
La prise de position de son groupe local, et non sa seule conviction.
Un sac de couchage et une petite tente.
Vous passez la nuit dans les gorges.
N'en déduis pas que c'est forcément la cruche qui est vide.
Le prénom ou la date de naissance ou le nom de jeune fille.
Sa réponse ne nous parvient pas dans les huit jours.
La banque de données est organisée en champs et en index.
Des plans parallèles au plan médian du corps.
Cela commence presque toujours au cours des premiers trente mois.
Les ongles, le sang et les urines du personnel dentaire.
Il le place dans le sens de la révélation divine.
La mise en place d'une véritable culture de la tolérance.
La réponse est exacte ou erronée.
La voie empruntée est la bonne.
Il est désormais composante de la majorité.
Dans le même temps et avec la même résolution.
Il faut agir sur les causes de cette violence.
Le passage de la lune au méridien du lieu et la basse mer.
Elle permet de concilier la qualité de la vie et le développement économique.
Les zones sensibles.
La lutte contre la pauvreté et la promotion des droits de l'homme.
Le ciel bien plus noir.
Le ballet des astres est plus fascinant que jamais.
La peine de mort ne pourra être envisagée comme une mesure de répression.
En fonction de la maturité de chaque cépage.
La dynamique de la balle et les textures ont été améliorées.
Un rendu visuel inégalé.
Connaître la position de tous vos comptes entre deux relevés.
Plus besoin de vous déplacer.
Les détours de cette légende dorée de la canonisation de saint Urbain.
Essayez la synthèse du français en temps

réel.
Elle est comme une lecture de la musique ou une lecture musicale de sa danse.
Lieu saint par excellence de la foi juive.
Celle-ci était applicable au moment où l'acte délictueux a été commis.
Je lui ai crié dans les oreilles.
Je lui ai frappé les joues.
Ainsi je lui ai redonné des sens.
Tout le monde a le droit de profiter du génie.
Il trouve ici son juste lieu d'exposition.
Ils devraient satisfaire même les amateurs les plus passionnés.
Tous les soirs sous le chapiteau avec des concerts ouverts au public.
Nous désirons étudier des fautes non maîtrisées actuellement.
Suivant la distribution du projet ou de la proposition de loi.
La direction du livre et le ministère de la culture et de la communication.
Honorer le notable qui a pris l'initiative.
Au nom des droits de l'homme.
Il est et demeure un acteur.
Elle ne montre pas les distances entre les lieux.
Elle porte sur le développement des réseaux de l'arc atlantique.
La première manette est sur le tuyau de droite.
Le jeton neuf ou les deux jetons huit et deux.
Il tolère les mots qui diffèrent du critère.
Une lettre ou par une inversion de lettre.
En ce moment, ce sont les soldes. Elle est partie faire du shopping.
Sa voiture est garée dans le parking souterrain.
Elle a été victime d'un kidnapping.
Elle habite en face du pressing.
Cet été, nous faisons du camping.
ne le dérangeons pas dans son travail.
On ne l'appelle pas, on va le chercher.
six foix quatre.
vingt dix sept mille cent vingt six.
Entre le théâtre et le monde.
le poste de travail du contrôleur en route.
Le premier qui bouge a perdu.
préciser son numéro et le nom du premier auteur.
il affirme mois après mois sa position sur le marché.
et là, il ne peut plus sentir le métal de la clef.
huit foix quatre vingt
douze mille sept cent dix sept.
Parce que dans la mer, il y a des éponges.
avant mille sept cent quatre-vingt neuf.
à vendre chien, mange n'importe quoi.
Forme légèrement ovale plutôt que ronde.
La ponctualité, la fiabilité et la sécurité.
Une curiosité pour capter le cadeau que la réalité vous donne.
Je ne me couche pas trop tard le soir.
Les thèmes proposés dans le guide envoyé dans les collèges.
L'un d'eux déclare entre deux bouchées.
Les formations mer, eau, environnement.
des demandes nouvelles dans la dernière décennie.
Dont le taux traduit le volume de la tumeur.
ou ils peuvent poursuivre des études.
et les critères de convergence seront exposés par Pierre René.
Des erreurs sont apparues lors du démarrage en octobre dernier.
Des douleurs des membres inférieurs.
Le petit nuage dit : Maman, maman, j'ai envie de faire pluie pluie.
On vous crie : qui vive.
une semaine après la remise du rapport des stages.
La représentation et la défense des intérêts des salariés.

Les rôtis, les volailles ou les poissons entiers.
L'odeur de deux corps purs.
Ils rendent compte de la ligne d'irréversibilité.
Ce qui provoque le rire.
Les conséquences dans les économies africaines.
La demande de brevet si une priorité a été revendiquée.
Le contrat de travail a pris fin avant la fin du mois.
L'homme plus sage apprend des erreurs des autres.
Cette page rassemble les résumés de chaque réunion du groupe.
l' argot l' a déjà nommé le roulant vif.
les prix sont petits et il y en a pour tous les goûts.
Elle change les brins de paille et de joncs en mâts de misaine et de beaupré.
Entre terre et soleil.
Cette ombre se déplace sur la terre.
Il ne faut abandonner ni en fait, ni en droit.
La réussite du programme an deux mille.
La conduite de processus s'est effectuée.
Des tiers peuvent faire des observations pour la délivrance du brevet.
La sélection a été faite par le comité de programme du colloque.
la fin des options ne coïncide pas avec le début du champ de données.
mille neuf cent quatre-vingt dix-sept.
Au fil de l'apprentissage et de la pratique du langage java.
Elles dépendent de la portée et du réglage de sensibilité du récepteur.
les êtres y puisent chaleur et lumière.
le stage est conçu par modules organisés par quinzaine.
La tour de télécommunication qui y est fichée.
C'est le cas, on ne peut le dissocier de la réalité.
Paraphé par le préfet de la Seine ou son délégué.
Cela passe par une réduction de leur coût de mobilisation.
Ils sont l'un des principaux moteurs du progrès technologique.
C'est l'un des problèmes majeurs de ce siècle.
Elle a découvert lors de la toilette, un nodule du sein gauche.
Je crois que le ciel a permis pour nos péchés cette infortune.
Sa portée inclut la surface au sol de la boite nids.
La nullité impose un travail de deuil.
Tous les maux du monde s'en échappent.
Ils y sont entrés.
Les chiffres prévisionnels de dépenses et de gains sont exacts.
le bruit de l'hiver est celui de l'eau qui gèle, de la glace et de la neige.
Jouir de leur droits civiques et politiques.
Il est composé de vingt-six membres.
C'est le contenu de son entrée standard qui est envoyé sur la sortie.
Les moyens budgétaires, matériels et humains de la direction.
Ils menacent la vie ou le bien-être de la communauté.
pour des échelles allant de un à plus de cinq cents.
Un ciel en feu fermait l'horizon sans borne d'un cercle de carmin.
Elle est distribuée le trente décembre de chaque année.
Il a présenté son calendrier des réunions et des thèmes de recherche.
Le législateur organique.
Le principe de non discrimination posé par

le traité.
Elle s'est condensée pour former les nuages.
du texte avec le maximum de précision et de concision.
Mais elle est imprimée sur un papier classique.
à tout autre endroit.
ou convoquer le témoin au siège de la juridiction.
Il avait établi le protocole.
On lui banda les yeux pour le monter au maquis.
Vous perdez alors tous les arrhes que vous avez versé.
Il faut un air doctoral, un ton d'autorité pour s'imposer au public.
Revendre par soi-même un immeuble.
Quarante quatre mille six cent trente deux articles principaux.
Les douze boules de la croix sont les signes du zodiaque.
Il a obtenu le prix de la mise en scène au festival de cannes quatre vingt sept.
Côté sud il est bordé d'un autre stade,
La maîtrise des techniques industrielles et de laboratoire.
Des années se sont écoulées depuis que je me suis levé.
On en juge par le nombre et par le ton des messages.
La porte ne ferme plus.
Il est plus dévié que le jaune.
Il est lui même plus dévié que le rouge.
Le menu est le même pour tout le monde, sauf pour les deux femmes.
Il y a deux cent quinze émetteurs principaux.
C'est le tribunal du lieu du siège social qui doit être saisi.
La renverse de jusant au voisinage de la basse mer.
Le produit est basé sur la réutilisation.
Pour échanger souvenirs, photos ou histoires.
Pour se documenter et pour agir.
Pour cette raison, je vis pratiquement au jour le jour.
Le rapport du commissariat général au plan et le rapport de la cour des comptes.
La machine prise en main est le pris.
la date présumée de la naissance ou de l'arrivée au foyer de l'enfant.
On peut enseigner les droits de l'homme.
Mille neuf cent quatre vingt dix huit.
Dans un de ses livres, il lui a consacré un chapitre.
Pour vous démaquiller.
Posez un coton imbibé de lotion sur la paupière.
Il est ébloui par la santé.
Cela jaillit comme une clarté de tes bras et de tes épaules.
Le lien existe entre ce que nous disons et ce que nous savons.
Entre régie et gestion déléguée.
Ceci montre la souplesse du système.
Dans les salles de concert, sur les podiums des défilés de mode.
Il se prépare avec des fleurs qui ne se sont ouvertes que la nuit.
Je répudie et la raison et la doctrine.
Je veux épouser la fille de la vigne.
Dans la vision dynamique, les caméras ne sont plus fixes.
Il est un des plus importants et peut-être le moins compris.
La qualité des données lors de la saisie téléinformatique.
Il devient devant l'ennemi plus réel et plus périlleux que jadis.
Cela ne dure pas au-delà du vingt et un.
Des ravins, des torrents, des lacs et des marais.

Ils permettent de retrouver en soi la source du bouddhisme tibétain.
La prostitution des enfants est un fléau mondial.
Elle est séparée de la cent soixante et un par une paroi rigide, et du couloir par un sas.
Notre hôtel a été entièrement rénové en mille neuf cent quatre-vingt seize.
Elle est aussi beaucoup appréciée dans les landes.
Sur une base plutôt large, le pont sera aussi large et bas.
Il est déjà dans le domaine public.
Nombre de fois où la page avec publicité est vue.
De plus en plus il est fait par des hommes riches pour des hommes riches.
Le dosage est réalisable par les élèves en une seule séance de travaux pratiques.
La fréquence double tous les dix ans.
Le sevrage tabagique est le plus souvent inconnue du grand public.
La finale est marquée par des arômes de miel et de cire.
Elle crée les risques les plus grands pour les cyclistes.
C'est un lieu de consultation et de conservation.
En dépit ou plutôt en raison de toute sa joliesse.
Son gai sourire de soubrette.
Il sépare les hommes et les femmes.
Cela concerne la formation du couple.
On trace le toit.
Ensuite on repart vers le bas.
En raison de ses origines, de ses opinions ou de ses croyances.
Soixante-dix centimes.
Les salaires de la personne recrutée.
Vous ne savez plus si c'est celui de la grande brune ou de la petite blonde.
On pense que cette perception est erronée
Ils sont situés en cent dix sept et cent dix neuf.
Chaque case est soit occupée par une cellule, soit vide.
Il est revenu sur sa décision en fin de réunion.
La carte la plus forte dans la couleur demandée remporte le pli et rejoue.
Les affaires de sécurité sociale du lieu de la demeure du débiteur.
Face à la mer, ils ouvrent sur un balcon ou en rez de jardin.
Le maire doit se retirer de la séance lors du vote du compte administratif.
Un très grand avec des lignes et des déliés, des plans inclinés.
Un fleuve la coupait comme un ruban de moire du rouge le plus vif.
Puis le texte est codé avec une seconde clé et ainsi de suite.
Le vingt et un mai mille six cent quarante trois.
Elle s'exprime dans les lieux d'art ou de fête.
Les sciences de la vie et de la santé.
Début mai mille neuf cent quatre vingt dix huit.
Merci de nous signaler inexactitudes, oublis.
Il nous fait partager ses visions multicolores.
Mille neuf cent quatre vingt dix sept.
La tâche que je me suis tardivement donnée est accidentelle.
C'est la tension entre le court terme et le long terme, tension budgétaire.
Elle a été adoptée par le comité consultatif.
Le conseil national du crédit.
Mais là, c'est tout de suite moins bien, même si ça reste correct.
Il est augmenté de la part de la fourchette.
Elle est due aux coûts de passage des ordres.
Les paramètres de cette section définissent les attributs de ce service.

Il l'emmena dans l'embrasure d'une fenêtre.
C'est en fonction du nombre de nuits et non au nombre de jours passés sur place.
la communication audiovisuelle a connu de profondes transformations.
Le plein droit des étudiants titulaires de la licence de mathématique.
Il avait mis au point le baguage des oiseaux migrateurs.
Il est sur une ou deux dimensions, pour le codage numérique ou alphanumérique.
Elles sont échues dans les trois années qui précèdent la date de son envoi.
Il va donner au vin son bouquet et son goût.
Les caractères du terroir.
Cela pourrait être la devise de cette nouvelle venue dans la cour des grandes.
Les emplois viennent recouvrir la notion de grades.
Les différents domaines de la productique.
Il façonne tous les aspects.
Il fonde au long des millénaires, la pérennité.
Nuit et jour.
Ils n'ont jamais dormi.
Elles ont fait leur effet, engageant dans tous les pays le processus de développement.
Elle les colle les uns aux autres.
Elle finit progressivement par fermer les yeux.
Ils ont été réalisés par les enfants, pendant leur séjour dans cette maison.
Trois et deux jours espacés de quelques semaines.
Une plus grande longévité.
Miroir de la réalité, il crée une cavité factice où vivent des regards immobiles.
On les dose comme marqueurs.
Le taux reflète le volume de la tumeur.
Tu obtiens un son creux.

ANNEXE E

Corpus ERES38 et AD80 : texte d'adaptation

ECOLOGIE ET BIODIVERSITE

Rencontre de la semaine

Pour la rubrique « Rencontre de la semaine » du journal « Ecologie et Biodiversité », nous sommes allés à la rencontre de Raymond Macé, apiculteur dans les Bouches-du-Rhône. Nous l'avions déjà rencontré à l'occasion du Salon de l'Agriculture, où il nous avait parlé de son miel bio.

Aujourd'hui, il nous présente son jardin potager. Il cultive sur un hectare plusieurs variétés de salades, citrouilles, courgettes, tomates, carottes ou encore des choux, mais également des fleurs comme des pivoines, des soucis, des acacias ou de la lavande pour ses abeilles. Ainsi, les abeilles sont tout près du lieu de butinage et n'ont pas besoin d'aller vers d'autres cultures bien souvent traitées avec des produits toxiques. Grâce à une palissade autour du jardin, ainsi que plusieurs épouvantails et des filets, il réduit les attaques des nuisibles, et cela se révèle assez efficace.

Au fond du jardin, il y a un petit cabanon où il range ses outils : taille-haie, cisaille, fourche, de nombreux paniers, sécateurs, pelles, râteaux et brouette. Quand il ne s'occupe pas des abeilles, Raymond passe beaucoup de temps à retourner la surface de la terre, enfouir, découper, planter, arroser, toujours dans l'idée de l'équilibre naturel, n'utilisant aucun produit de traitement, mais ses « petits trucs » bien à lui.

Filtrant l'eau des marécages, l'arrosage est plus facile. Dans les campagnes des Bouches-du-Rhône, il y a souvent de très fortes chaleurs, c'est pourquoi Raymond irrigue pour éviter le desséchage. L'an passé, ses premières tentatives de culture avaient échoué à cause du givre et une bonne partie des plantations avait pourri. Deux fois par semaine, il fait les marchés avec son chien Cachou. Il charge dans son camion des kilos de légumes et de miel très appréciés des clients.

ANNEXE F

Texte corpus Voix Détresse

A l'aide
Je suis blessé
Appelez les secours
Je suis tombé
J'ai mal à la tête
Un toubib, vite
Aïe, aïe, aïe
J'ai très mal
Appelez le SAMU
Je ne peux pas me relever
Ne me laissez pas tout seul
Au secours
Je ne peux plus bouger
Je ne me sens pas bien
J'ai un malaise
Aidez-moi
Du secours s'il-vous-plaît
Qu'est-ce qu'il m'arrive
Ouh là là
A moi

FIGURE F.1: *Images utilisées pour l'enregistrement du corpus Emotions*

ANNEXE G

Métrique

G.0.1 WER

L'évaluation des performances d'un système de RAP est faite de manière classique par le taux d'erreur de mots, ou *Word Error Rate* (WER). Celui-ci indique le taux de mots incorrectement reconnus par rapport à un texte de référence. Plus le taux est faible (minimum 0%) plus la reconnaissance est bonne. Le taux maximum n'est pas borné et peut dépasser 100% en cas de très mauvaise reconnaissance, s'il y a beaucoup d'insertions. Le WER est calculé par un alignement de la référence avec l'hypothèse décodée. Il est donné par :

$$WER = \frac{I + D + S}{N}.100 \tag{G.1}$$

où :

— I est le nombre d'insertions,

— D est le nombre de déletions,

— S est le nombre de substitutions,

— N est le nombre de mots de la référence.

G.0.2 Sensibilité, spécificité, rappel, précision, F-mesure

Pour mesurer la capacité du système de RAP à décoder certains mots-clés, nous avons utilisé des métriques de validité de tests statistiques telles que la sensibilité, la spécificité, le rappel, la précision et la F-mesure.

La Table G.1 représente les résultats possibles lors de la mesure de la validité d'un test, avec :

— Hyp 1 : l'hypothèse est vérifiée

— Hyp 2 : l'hypothèse n'est pas vérifiée

— VP : les vrais positifs,

— FP : les faux positifs,

— FN : les faux négatifs,

— VN : les vrais négatifs.

	Hyp 1	Hyp 2
Test positif	VP	FP
Test négatif	FN	VN

TABLE G.1: Matrice de confusion des tests positifs et négatifs

La sensibilité d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée :

$$Se = \frac{VP}{VP+FN} \tag{G.2}$$

Elle s'oppose à la spécificité, qui mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée :

$$Sp = \frac{VN}{VN+FP} \tag{G.3}$$

La Précision est la proportion de résultats positifs qui sont pertinents. Elle mesure la capacité du système à refuser les résultats non-pertinents :

$$P = \frac{VP}{VP+FP} \tag{G.4}$$

Le rappel est la proportion de résultats positifs qui sont trouvés. Il mesure la capacité du système à donner tous les résultats pertinents. Il est égal à la sensibilité :

$$R = \frac{VP}{VP+FN} \tag{G.5}$$

La F-mesure mesure la capacité du système à donner tous les résultats pertinents et à refuser les autres. Elle correspond à un compromis de la précision et du rappel donnant la performance du système :

$$F-mesure = \frac{2P.R}{P+R} \tag{G.6}$$

Le taux de vrais positifs (TVP) est égal au rappel et à la sensibilité, il est donné par la relation suivante :

$$TVP = R = Se = \frac{VP}{Positifs} \tag{G.7}$$

Le taux de faux positifs (TFP) est donné par la relation suivante :

$$TFP = 1–Sp = \frac{FP}{Négatifs} \tag{G.8}$$

G.0.3 Différences

La différence absolue entre les 2 scores A et B est donnée par la relation suivante :

$$Différence\ absolue_{(A,B)} = A - B \tag{G.9}$$

La différence relative entre les 2 scores A et B, B étant le score de référence, est donnée par la relation suivante :

$$Différence\ relative_{(A,B)} = \frac{A-B}{B}.100 \tag{G.10}$$

ANNEXE H

Scénarios du corpus Cirdo-Set

« Trébucher contre un tapis »

Vous portez une pile d'assiettes (de famille, donc de valeur) pour les déplacer d'un endroit à un autre.

Vous avancez d'un pas rapide et vous trébuchez contre le tapis sur votre chemin.

Vous perdez donc votre équilibre et tentez donc de ne pas tomber en avant (sans y parvenir), veillant surtout à ne pas faire tomber les assiettes.

Sur votre chemin se trouvent donc le tapis et un canapé. Vous tentez de rétablir votre équilibre en essayant de vous appuyer sur le canapé. Malheureusement, vous vous cognez contre celui ci est êtes projeté(e) au sol.

Ne voulant absolument pas lâcher les assiettes, vous essayez d'amortir votre chute avec votre genou gauche. À ce moment là, vous vous écriez « meeeeeeerde !!!».

Vous vous retrouvez donc avec le genou gauche à terre et la jambe droite pliée, vous maintenant en équilibre.

Dans un deuxième temps, vous vous mettez donc sur les deux genoux, pendant une dizaine de secondes.

Vous posez les assiettes sur votre gauche. Vous posez alors vos mains, maintenant qu'elles sont libérées, sur le sol, devant vous. Vous dites à ce moment « qu'est ce qui m'arrive ».

Vous basculez sur votre droite et vous êtes maintenant en position assise, les jambes légèrement repliées, les bras derrière vous.

Lorsque vous reprenez vos esprits vous dites alors « Appelle quelqu'un E-Lio».

« Glisser dans le salon »

C'est la nuit. Vous avez soif et vous vous levez pour aller boire un verre d'eau dans la cuisine.

Vous êtes dans le noir et en arrivant dans le salon votre pied glisse.

Vous perdez l'équilibre et votre corps bascule en arrière. Surpris(e), vous vous écriez « ohhhhh laaaaaaa !!!! »

Vous vous retrouvez allongé(e) sur le dos, les bras étendus derrière la tête.

Vous ramenez vos bras le long de votre corps et tentez en les pliant et en prenant appui sur le sol, de vous redresser. Vous tentez de basculer vers la droite. En vain. Vous essayez sur le côté gauche. Vous pliez votre jambe afin de prendre appui sur le genou. En poussant sur vos bras, vous vous retrouvez sur les genoux.

Vous vous relevez et abasourdi(e), vous allez vous asseoir sur le canapé.

Vous dites alors « Elio, appelle quelqu'un ».

« Tomber de son canapé »

Vous êtes en train de faire une sieste sur votre canapé, allongé(e) sur le dos.
Vous vous tournez brusquement vers l'extérieur et chutez dans le vide.
Vous vous retrouvez alors à plat ventre, face contre terre, les bras allongés contre le terre : votre nez est cassé et vous saignez.
Vous vous écriez alors « Oh laaa ! je saigne !!! je me suis blessé(e) !!! »
Abasourdie par cette chute, vous tentez, en rampant à l'aide de vos bras d'atteindre une sonnette d'alarme. Mais vous n'y arrivez pas.
Vous dites alors « E-lio appelle les secours »

« Tomber en arrière »

La nuit tombe et vous voulez fermer les volets.
Vous ouvrez la fenêtre et vous vous penchez pour attraper un des volets.
Vous tirez un peu fort car cela semble bloqué. Votre main glisse et dans un mouvement en arrière vous êtes projeté(e), perdant votre équilibre.
Vous tombez donc vers l'arrière et ce sont d'abord les fesses qui heurtent le sol.
La chute vous choque et malgré la douleur causée par celle-ci, vous tentez, en vous traînant/dandinant (vers l'avant) sur votre postérieur, aidé(e) de vos mains et jambes, d'atteindre le téléphone pour appeler votre fille. En chemin, vous vous écriez « aïe !! j'ai mal !!! » : La douleur est trop grande, et ne pouvant davantage vous déplacer, vous dites « E-lio appelle ma fille »

« Rester bloqué(e), assis(e) sur le canapé »

Vous êtes en train de lire un magazine, confortablement installé (e) dans votre canapé.

Vous tentez de vous relever.

Vous posez votre magazine à côté de vous et tentez une nouvelle fois de vous relever.

Vous prenez appui sur vos deux bras mais vous n'y arrivez toujours pas.

Vous retombez en arrière en vous écriant « ahhhhhhh !!!! »

Vous recommencez deux autres fois mais la douleur augmente et vous ne tentez plus de vous relever : la douleur est trop grande. Vous vous écriez « aïe j'ai mal !!!!! ».

Vous basculez sur le côté et vous vous allongez.

Vous dites alors « E-lio, appelle du secours, je peux pas me relever »

Scenarii alternatifs

« <u>Prendre la télécommande</u> »

Vous regardez la TV, installé(e) dans votre canapé. Vous souhaitez changer de chaîne : la télécommande est sur la table basse devant vous. En prenant appui sur vos bras, *au bout de deux essais*, vous vous relevez, vous atteignez la table basse et saisissez la télécommande.

« <u>Se baisser pour ramasser un objet au sol</u> »

Vous avez les bras chargés de magazines (ou autres) et lors d'un déplacement, vous faites tomber un magazine par terre. **Vous vous baissez pour le ramasser. Vous avez des difficultés pour le ramasser (ça glisse), les autres magazines tombent et vous devez donc ramasser le tas.**