HAL
open science

Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes

Clément Chastagnol

To cite this version:

Clément Chastagnol. Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes. Autre [cs.OH]. Université Paris Sud - Paris XI, 2013. Français. NNT : 2013PA112190 . tel-00923201

HAL Id: tel-00923201

https://theses.hal.science/tel-00923201v1

Submitted on 2 Jan 2014

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

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

UNIVERSITE PARIS-SUD

ÉCOLE DOCTORALE D'INFORMATIQUE DE PARIS-SUD (ED427)
Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur (LIMSI-CNRS)

***DISCIPLINE** : Informatique*

THÈSE DE DOCTORAT

soutenue le 04/10/2013

par

Clément CHASTAGNOL

Reconnaissance automatique des dimensions affectives dans l'interaction homme-machine pour des personnes dépendantes

Directeur de thèse :	Laurence DEVILLERS	LIMSI-CNRS, Université Paris-Sorbonne IV
Co-directeur de thèse :	Jean-Claude MARTIN	LIMSI-CNRS, Université Paris-Sud XI

Composition du jury :

Rapporteurs :	Mohamed CHETOUANI	ISIR, Université Paris VI
	Björn SCHULLER	Technische Universität München
Examinateurs :	Christophe LEROUX	CEA-LIST
	Nicolas SABOURET	LIMSI-CNRS, Université Paris-Sud XI

Université Paris-Sud 11

Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes

par
Clément Chastagnol

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)

Thèse présentée à l'École Doctorale d'Informatique de Paris-Sud
en vue de l'obtention du grade de Docteur
en informatique

octobre, 2013

© Clément Chastagnol, 2013.

Université Paris-Sud 11
École Doctorale d'Informatique de Paris-Sud

Cette thèse intitulée:

Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes

présentée par:

Clément Chastagnol

a été évaluée par un jury composé des personnes suivantes:

Laurence Devillers,	directeur de recherche
Jean-Claude Martin,	codirecteur
Christophe Leroux,	examinateur
Nicolas Sabouret,	examinateur
Mohammed Chetouani,	rapporteur
Björn Schuller,	rapporteur

Thèse acceptée le : . . . . . .

RÉSUMÉ

La majorité des systèmes de reconnaissance d'états affectifs est entrainée sur des données artificielles hors contexte applicatif et les évaluations sont effectuées sur des données pré-enregistrées de même qualité. Cette thèse porte sur les différents défis résultant de la confrontation de ces systèmes à des situations et des utilisateurs réels.

Pour disposer de données émotionnelles spontanées au plus proche de la réalité, un système de collecte simulant une interaction naturelle et mettant en oeuvre un agent virtuel expressif a été développé. Il a été mis en oeuvre pour recueillir deux corpus émotionnels, avec la participation de près de 80 patients de centres médicaux de la région de Montpellier, dans le cadre du projet ANR ARMEN.

Ces données ont été utilisées dans l'exploration d'approches pour la résolution du problème de la généralisation des performances des systèmes de détection des émotions à d'autres données. Dans cette optique, une grande partie des travaux menés a porté sur des stratégies cross-corpus ainsi que la sélection automatique des meilleurs paramètres. Un algorithme hybride combinant des techniques de sélection flottante avec des métriques de similitudes et des heuristiques multi-échelles a été proposé et appliqué notamment dans le cadre d'un challenge (InterSpeech 2012). Les résultats de l'application de cet algorithme offrent des pistes pour différencier des corpus émotionnels à partir des paramètres les plus pertinents pour les représenter.

Un prototype du système de dialogue complet, incluant le module de détection des émotions et l'agent virtuel a également été implémenté.

Mots clés : données émotionnelles naturelles, cross-corpus, sélection automatique de paramètres.

ABSTRACT

Most of the affective states recognition systems are trained on artificial data, without any realistic context. Moreover the evaluations are done with pre-recorded data of the same quality. This thesis seeks to tackle the various challenges resulting from the confrontation of these systems with real situations and users.

In order to obtain close-to-reality spontaneous emotional data, a data-collection system simulating a natural interaction was developed. It uses an expressive virtual character to conduct the interaction. Two emotional corpora where gathered with this system, with almost 80 patients from medical centers of the region of Montpellier, France, participating in. This work was carried out as part of the French ANR ARMEN collaborative project.

This data was used to explore approaches to solve the problem of performance generalization for emotion detection systems. Most of the work in this part deals with cross-corpus strategies and automatic selection of the best features. An hybrid algorithm combining floating selection techniques with similarity measures and multi-scale heuristics was proposed and used in the frame of the InterSpeech 2012 Emotino Challenge. The results and insights gained with the help of this algorithm suggest ways of distinguishing between emotional corpora using their most relevant features.

A prototype of the complete dialog system, including the emotion detection module and the virtual agent was also implemented.

Keywords: natural emotional data, cross-corpus, automatic feature selection.

TABLE DES MATIÈRES

RÉSUMÉ . . . . . **iii**

ABSTRACT . . . . . **iv**

TABLE DES MATIÈRES . . . . . **v**

LISTE DES TABLEAUX . . . . . **ix**

LISTE DES FIGURES . . . . . **xii**

REMERCIEMENTS . . . . . **xvii**

CHAPITRE 1 : INTRODUCTION GÉNÉRALE **1**

1.1 Besoins et défis actuels 2

1.2 Contexte - Le projet ARMEN 4

1.3 Contributions 5

1.4 Organisation du document 6

I État de l'art **7**

CHAPITRE 2 : THÉORIE DES ÉMOTIONS ET DE LEURS EXPRESSIONS **8**

2.1 L'émotion : une notion complexe 8

2.2 Théories modernes des émotions 9

2.2.1 Théories catégorielles 9

2.2.2 Théories dimensionnelles 9

2.2.3 Théories cognitives de l'évaluation 11

2.3 Expression vocale des émotions 13

2.4 Utilisation des théories des émotions dans l'*affective computing* 14

CHAPITRE 3 : SYSTÈMES AUTOMATIQUES DE RECONNAISSANCE DES ÉMOTIONS . . . 16

3.1 Composants d'un système de reconnaissance . . . 16

3.2 Problématiques relatives à l'apprentissage automatique . . . 19

3.2.1 Généralités . . . 19

3.2.2 Problème du sur-apprentissage . . . 20

3.2.3 Évaluation : métriques et méthodologies . . . 21

3.2.4 Algorithmes - Détails sur les SVM . . . 22

3.3 Corpus émotionnels : collecte et annotation, spontanéité des données . . . 24

3.4 Performances des systèmes actuels . . . 27

CHAPITRE 4 : INTERACTION ÉMOTIONNELLE AVEC DES MACHINES 29

4.1 Éléments théoriques de la communication non-verbale humaine . . . 29

4.2 Machines interactives . . . 31

4.2.1 Agents virtuels expressifs . . . 31

4.2.2 Robots . . . 33

4.2.3 Cas des robots assistants . . . 34

4.3 Aspects dialogiques . . . 36

4.4 Aspects perceptifs . . . 37

4.5 Nouveaux challenges de l'interaction homme-machine . . . 38

II Reconnaissance des émotions dans la parole 41

CHAPITRE 5 : COLLECTE DE DONNÉES . . . 42

5.1 Introduction - Motivation . . . 42

5.2 Protocoles et système de collecte de données . . . 42

5.2.1 Première collecte (ARMEN_1) . . . 43

5.2.2 Seconde collecte (ARMEN_2) . . . 49

5.2.3 Quelques remarques sur les collectes . . . 53

5.3 Segmentation et annotation . . . 55

5.3.1 Segmentation 55
5.3.2 Annotation 58
5.4 Corpus finaux 60
5.5 Discussion 61

CHAPITRE 6 : DÉTECTION DES ÉMOTIONS EN CROSS-CORPUS . . 66
6.1 État de l'art 66
6.1.1 Motivation 66
6.1.2 Difficultés 66
6.1.3 Stratégies 67
6.2 Expériences menées 68
6.2.1 Présentation des corpus 69
6.2.2 Expériences et résultats 71
6.3 Conclusion 80

CHAPITRE 7 : SÉLECTION AUTOMATIQUE DE PARAMÈTRES . . . 82
7.1 État de l'art 83
7.1.1 Algorithmes de sélection séquentielle 85
7.1.2 Sélection flottante 88
7.2 SFFS-SSH 91
7.3 H-SFFS 92
7.3.1 Fonctionnement de l'algorithme H-SFFS : pseudo-code 93
7.3.2 Résultats 93
7.4 Analyse des paramètres sélectionnés 108
7.5 Méthodologie : combattre le sur-apprentissage 115
7.6 Conclusion 118

III Système de dialogue émotionnel avec un personnage virtuel 119

CHAPITRE 8 : IMPLÉMENTATION 120
8.1 Fonctionnalités 120

8.2 Architecture et détail des modules 120
8.2.1 Gestion du flux audio 121
8.2.2 Segmentation de la voix 122
8.2.3 *Word-spotting*, grammaires et arbres de dialogue 123
8.2.4 Détection des émotions 124
8.2.5 AVE : contrôle et expressions 125
8.3 Conclusion 125

CHAPITRE 9 : VERS UNE MESURE OBJECTIVE DE L'ENGAGEMENT 130
9.1 Description de l'approche 130
9.2 Résultats 131
9.3 Discussion 135

CHAPITRE 10 : CONCLUSION 137
10.1 Contexte et rappel des objectifs de recherche 137
10.2 Résumé des contributions 138
10.3 Perspectives - Discussion 139
10.4 Conclusion générale 142

BIBLIOGRAPHIE 145

Annexes i

LISTE DES TABLEAUX

2.I Comparaison entre plusieurs listes d'émotions primaires 10

2.II Effets des émotions sur l'expression vocale, mesurés à l'aide de plusieurs paramètres acoustiques (d'après [222]). 14

3.I Résumé des paramètres primaires utilisés pour le Challenge Interspeech 2009 [229]. Couplés à des fonctionnelles, ils résultent en 384 paramètres finaux. 18

3.II Liste des principaux corpus émotionnels. 26

3.III Comparaison des performances entre quelques systèmes de reconnaissance des émotions. 28

4.I Détails de quelques robots interactifs. 34

5.I Détails des scénarios pour la collecte du corpus ARMEN_1. . . . 46

5.II Résumé des caractéristiques des sujets de la collecte du corpus ARMEN_1. 48

5.III Résumé des caractéristiques des sujets de la collecte du corpus ARMEN_2. 52

5.IV Code des phases pour le corpus ARMEN1 57

5.V Code locuteurs et correspondances pour le corpus ARMEN1 . . . 57

5.VI Résumé de l'étape de segmentation pour ARMEN_1 et ARMEN_2. 57

5.VII Informations sur le schéma d'annotation pour les corpus ARMEN_1 et ARMEN_2 59

5.VIII Évaluation quantitative de l'accord inter-annotateur 59

5.IX Résumé de l'étape d'annotation pour ARMEN_1 et ARMEN_2. . 62

5.X Récapitulatif pour les corpus ARMEN_1 et ARMEN_2. 62

6.I Regroupements des étiquettes émotionnelles fines en macro-classes pour le corpus CEMO. 70

6.II Regroupements des étiquettes émotionnelles fines en macro-classes pour le corpus EmoVox (d'après [258]). 70
6.III Répartition des segments entre les macro-classes pour le corpus CEMO. 72
6.IV Répartition des segments entre les macro-classes pour le corpus EmoVox. 72
6.V Résumé des résultats pour la première expérience cross-corpus. Pour chaque paire corpus d'apprentissage/de test, la première ligne donne les performances pour l'ensemble de 384 paramètres, la deuxième pour 988 paramètres. 74
6.VI Détails des ensembles d'apprentissage et de test pour les trois conditions expérimentales. 76
6.VII Résultats pour les trois conditions expérimentales. 77
6.VIII Gain de performance (en points de pourcentage du RR) pour la stratégie Pooling par rapport aux conditions Deux-à-deux et Intra. 77
6.IX Scores de similarité fondés sur un classement de paramètres. . . . 78
6.X Détails sur la composition des corpus. 79
6.XI Résultats UAR pour les conditions Intra et Pooling (niveau de chance : 25%). 81
6.XII Résultats UAR (niveau de chance : 25%) pour la condition Deux-à-deux avec les critères d'âge et de qualité vocale. 81

7.I Évolution du nombre de paramètres dans les systèmes de reconnaissance automatique des émotions dans la voix. 82
7.II Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement sur les meilleurs ensembles pour le corpus ARMEN_1. 97
7.III Comparaison des performances entre les algorithmes deux-à-deux pour le corpus ARMEN_1. 99

7.IV Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement sur les meilleurs ensembles pour le corpus ARMEN_2. 101
7.V Comparaison des performances entre les algorithmes deux-à-deux pour le corpus ARMEN_2. 101
7.VI Comparaison des performances entre les algorithmes deux-à-deux pour le corpus JEMO. 103
7.VII Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement d'une part et apprentissage et test d'autre part sur les meilleurs ensembles pour le corpus ARMEN + JEMO. 104
7.VIII Résumé des meilleurs résultats (et tailles d'ensembles correspondantes) atteints en test. 106
7.IX Scores de similarité fondés le meilleur ensemble à 20 paramètres pour chaque corpus. 111

9.I Mesures utilisées 132
9.II Matrice de corrélation croisée entre les différentes mesures relevées. 132

LISTE DES FIGURES

1.1 Évolution du nombre de résultats de recherche pour les termes "affective computing" sur le site Google Scholar (les chiffres avancés ne sont pas exacts ; on peut d'ailleurs remarquer une baisse du nombre de résultats pour certaines années ; cependant ils permettent d'avoir une bonne idée de la tendance générale). 2

2.1 Représentation graphique du modèle circumplexe de Plutchik (d'après [205]) 11

2.2 Capture d'écran de l'outil Feeltrace (d'après [61]) 12

3.1 Schéma de fonctionnement d'un système de reconnaissance des émotions dans la voix. 18

3.2 Illustration du phénomène de sur-apprentissage 21

3.3 Illustration du kernel trick 23

3.4 Illustration de la séparation avec marge maximale. L'hyperplan H_1 ne sépare pas les données, H_2 et H_3 les séparent mais H_3 a la marge maximale. 24

4.1 Illustration d'un modèle d'implication affective en interaction, d'après [258] 31

4.2 Quelques exemples de robots interactifs ayant une forme humanoïde, animaloïde ou autre. De gauche à droite et de haut en bas : PaPeRo, Kismet, AIBO, Care-o-Bot, Paro, Nexi-MDS, Nao, Jazz et Reeti. 35

4.3 Architecture modulaire d'un système de dialogue parlé (d'après [41]). 37

4.4 Illustration du phénomène de la "vallée de l'étrangeté", d'après [184]. 38

5.1 Dispositif du matériel pour la collecte de données ARMEN_1. . . 45

5.2 Répartition des sujets de la collecte ARMEN_1 en termes de centres médicaux, sexe et âge. 49
5.3 Dispositif du matériel pour la collecte de données ARMEN_2. . . 51
5.4 Répartition des sujets de la collecte ARMEN_2 en termes de centres médicaux, sexe et âge. 53
5.5 Illustration de l'interface de contrôle pour le système de collecte ARMEN_2 en *magicien d'Oz*. 54
5.6 Histogramme de la durée des segments (en secondes) pour les corpus ARMEN_1 et ARMEN_2 (segments des sujets uniquement). . 58
5.7 Comparaison des distributions des étiquettes entre les deux annotateurs pour le corpus ARMEN_1. 60
5.8 Matrice de confusion des annotations pour le corpus ARMEN_1. . 61
5.9 Comparaison des distributions des étiquettes entre les deux annotateurs pour le corpus ARMEN_2. 63
5.10 Matrice de confusion des annotations pour le corpus ARMEN_2. . 64

6.1 Conditions expérimentales pour la deuxième expérience cross-corpus. 75

7.1 Évolution du nombre de paramètres dans les systèmes de reconnaissance automatique des émotions dans la voix. 83
7.2 Catégorisation des approches de SAP (d'après [63]) 85
7.3 Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_1. 96
7.4 Comparaison des performances sur les algorithmes séquentiels pour le corpus ARMEN_1 (les points correspondent aux scores sur le corpus d'apprentissage et les croix aux scores sur le corpus de développement). 96

7.5 Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN_1 (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement et les triangles aux scores sur le corpus de test). 98
7.6 Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_1 98
7.7 Représentation des 5 meilleurs sets pour le corpus ARMEN_1 en fonction de leur taille et de leur performance en apprentissage (points), développement (croix), et test (triangles). Les lignes verticales correspondent à la taille maximale d'ensemble atteinte pour chaque algorithme. 100
7.8 Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_2. 100
7.9 Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN_2 (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement, et les triangles aux scores sur le corpus de test). 102
7.10 Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus JEMO (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement, et les triangles aux scores sur le corpus de test). 103
7.11 Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN + JEMO (zoom). . 104

7.12 Représentation des 5 meilleurs sets pour le corpus ARMEN + JEMO en fonction de leur taille et de leur performance en apprentissage (points), développement (croix) et test (triangles). Les lignes verticales correspondent à la taille maximale d'ensemble atteinte pour chaque algorithme. 105
7.13 Fonctionnement de l'algorithme SFFS. 107
7.14 Fonctionnement de l'algorithme SFFS-SSH. 107
7.15 Fonctionnement des étapes de l'algorithme H-SFFS. 109
7.16 Fonctionnement de l'algorithme H-SFFS. 109
7.17 Comparaison des proportions entre types de paramètres pour différents résultats de sélection (meilleur résultat choisi). 110
7.18 Comparaison des proportions entre types de paramètres pour différents résultats de sélection (meilleur résultat pour 20 paramètres). 110
7.19 Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus ARMEN_1. 112
7.20 Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus ARMEN_2. 113
7.21 Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus JEMO. 114
7.22 Illustration du phénomène de sur-apprentissage avec un modèle trop complexe et un modèle adapté à la complexité de la cible. . . 116
7.23 Illustration du phénomène de sur-apprentissage avec trop peu de données d'apprentissage. 116

8.1 Architecture du prototype. 121
8.2 Schéma de fonctionnement de l'algorithme de segmentation. . . . 122
8.3 Les différentes erreurs de segmentation possibles. 123

8.4 Exemple de règle pour la présentation de l'utilisateur. Les alternatives au sein d'un groupe délimité par des parenthèses sont séparées par le caractère "barre verticale". Le placement de gauche à droite donne l'ordre séquentiel et donc la forme de la phrase. Le jeton "[$GARBAGE]" est utilisé pour n'importe quel contenu que l'on ne cherche pas à reconnaître, un peu comme le *wildcard* "*" en informatique. Cela signifie ici que le système ne comprend pas le prénom de l'utilisateur. 124
8.5 Représentation graphique du scénario d'alerte. 127
8.6 Comportement du système en interaction. 128
8.7 Architecture du module de détection d'émotions. 128
8.8 Capture d'écran de l'agent Mary, utilisant la plate-forme MARC, arborant plusieurs expressions (au centre, expression neutre ; à partir du haut et dans le sens des aiguilles d'une montre : angoisse, doute, curiosité, intérêt, expression positive, agacement). 129

9.1 Diagramme de Hinton pour la matrice de corrélation. L'aire de chaque carré est proportionnelle à la valeur absolue du coefficient de corrélation correpondant dans la matrice et la couleur correspond au signe (blanc pour positif, noir pour négatif). 133
9.2 Évaluation de la qualité de l'interaction par les participants à l'expérience. 135
9.3 Répartition des participants à l'expérience selon leur âge. 136

10.1 Résumé des performances des participants du challenge de sélection de paramètres de la conférence NIPS 2003. 141

REMERCIEMENTS

Loin de l'image du chercheur isolé dans son bureau sombre, travaillant le dos courbé sous la lumière d'une petite lampe de bureau projetant aux murs les ombres vacillantes d'empilements de papiers à l'équilibre incertain, mon séjour au LIMSI s'est déroulé au sein d'un groupe accueillant et dans une ambiance chaleureuse. Je voudrais pour cela remercier les membres de l'équipe du thème *Dimensions Affectives et Sociales dans les Interactions Parlées* : Christophe, Julieta, Caroline, Marie, Agnès et Mariette, ainsi que les membres du groupe TLP dont certains sont devenus de très bons amis, sans ordre particulier : Nadi, Thiago, Nicolas F., Nicolas P., Penny, Thomas, Hervé, Claude, Bill, Éric et Artem.

Les voyages que j'ai eu la chance d'effectuer pendant cette thèse, en conférence et école d'été, resteront parmi mes meilleurs souvenirs avec de très belles rencontres : Vijay, Richard, Sandra, Aurore et Claudia, je pense à vous.

Le projet ARMEN, qui a financé ma thèse, a donné lieu à de belles collaborations. Parmi nos partenaires, je tiens à remercier Nicolas, Olivier L., Pedro, Olivier V. et Aymeric de Voxler pour leur aide technique, le soutien de Voxler quand j'en ai eu besoin et leur sympathie depuis maintenant plusieurs années ; Violaine d'APPROCHE pour sa bonne humeur et son aide précieuse dans les collecte de données ; M. Nguyen pour son accueil toujours affable à l'EHPAD Malbosc ; Christophe du CEA pour avoir mené à bien le projet et avoir présidé le jury de ma soutenance ; et l'équipe du LASMEA pour leur hospitalité à Clermont-Ferrand et leur impressionnante démonstration de véhicules automatiques. Je souhaite également remercier Rodolphe d'Aldebaran Robotics, avec qui j'ai pu travailler plusieurs fois à l'occasion de projets annexes, toujours compétent et serviable malgré son emploi du temps plus que chargé.

Je voudrais remercier sincèrement mes directeurs de thèse, Jean-Claude Martin et Laurence Devillers, pour leur suivi attentif, leurs suggestions, leurs encouragements et leur encadrement bienveillant. Cela a été très agréable de travailler avec vous et de vous connaître mieux au fil du temps, que ce soit en discutant autour d'un verre à l'étranger, ou pendant ces trajets du labo à Bourg-la-Reine.

Même si ce n'est pas une condition suffisante, le fait d'avoir une vie saine et épanouissante en dehors du labo est souvent une condition nécessaire à l'accomplissement heureux d'une thèse. Pour cela, je tiens à remercier mes parents, Marc et Fatima, et mon frère et ma soeur, Loïc et Camille, d'avoir toujours été là pour me soutenir et m'encourager dans tous les aspects de ma vie. Et lorsqu'il a fallu décompresser, les amis musiciens qui répondaient présents pour une jam, un concert ou un verre : Victoria Caffè, la bande de l'atelier jazz, Ruth et Xavier. Enfin, merci à celle qui a partagé au quotidien les bons moments et supporté les passages moins agréables de ces dernières années ; merci à toi Sarah.

CHAPITRE 1

INTRODUCTION GÉNÉRALE

The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.

— Marvin Minsky, *Society of Mind, 1988 [181]*

Permettre aux machines de comprendre non seulement nos explications, mais également nos intentions, nos émotions et intonations subtiles, tel est le but d'une communauté scientifique relativement jeune, aux confluents de la psychologie, de l'informatique, de la robotique et des mathématiques appliquées. Depuis la création du champ de recherche de l'*affective computing* en 1997 par le professeur Rosalind Picard du MIT Media Lab, des progrès impressionnants ont été observés, avec la réalisation de robots capables de percevoir et de réagir aux états affectifs de leurs utilisateurs, des programmes informatiques à qui l'on peut s'adresser via des personnages virtuels qui détectent le stress ou le doute.

Cependant, le domaine étant fondé sur des bases théoriques encore non consensuelles, de nombreuses questions restent ouvertes. Par exemple, on ne comprend pas encore bien comment transcrire les intonations particulières qui font que les humains perçoivent la joie ou la tristesse dans la voix de l'autre, ou comment un robot doit se comporter pour être facilement accepté et compris dans ses actions. De plus, une fois la question de la détection introduite, les problèmes de la compréhension et du retour sont soulevés : comment intégrer la détection d'émotions à un dialogue humain-machine et comment réagir à l'émotion exprimée par l'utilisateur ?

Cette thèse s'intéresse à ces problématiques et plus particulièrement à la mise en oeuvre de systèmes de reconnaissance automatique des émotions dans la voix parlée au sein de systèmes de dialogue complets, incluant une interface homme-machine de type personnage virtuel.

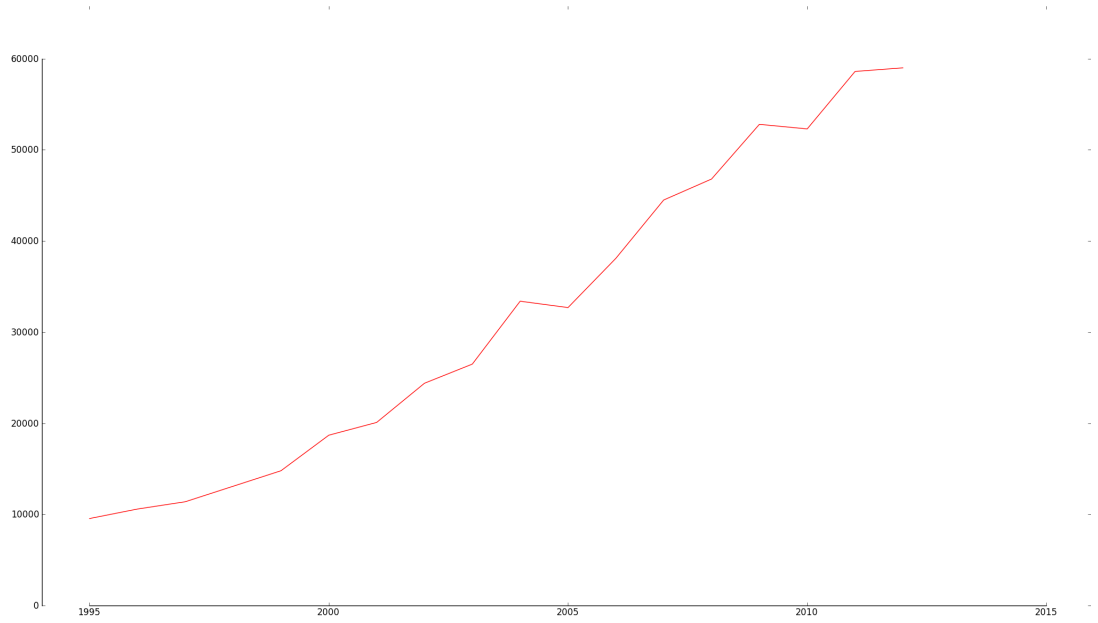

Figure 1.1 – Évolution du nombre de résultats de recherche pour les termes "affective computing" sur le site Google Scholar (les chiffres avancés ne sont pas exacts ; on peut d'ailleurs remarquer une baisse du nombre de résultats pour certaines années ; cependant ils permettent d'avoir une bonne idée de la tendance générale).

1.1 Besoins et défis actuels

Malgré les progrès réalisés ces dernières années, les systèmes actuels de reconnaissance des émotions montrent leurs limites. Conçus avec une approche réductionniste, dans des contextes simples et contrôlés dans un premier temps, les performances qu'ils atteignent en laboratoire ne se reproduisent pas lorsqu'ils sont confrontés au monde réel. Énormément de difficultés surgissent alors : variabilités de tous types (locuteurs, qualité vocale, conditions acoustiques), problèmes de bruits, de superposition... Trois besoins essentiels se distinguent alors.

Tout d'abord, les systèmes étant fondés sur un paradigme d'apprentissage à partir de données exemplifiées et annotées, il y a un besoin important de données de meilleure qualité et en plus grande quantité.

- L'utilisation de données "actées", c'est-à-dire d'émotions interprétées par des ac-

teurs dans un contexte contrôlé (acoustique, contenu lexical), et en quantité insuffisante (peu de locuteurs, peu de données pour chaque catégorie d'émotion considérée) est encore la norme malgré une prise de conscience de la communauté. Le problème de la faible quantité des données spontanées est notamment dû à leur caractère privé, qui rend leur partage difficile voire impossible au sein de la communauté. L'organisation de challenges et de *benchmarks* pour la communauté a permis d'avancer de ce point de vue et de regrouper la communauté, mais il reste encore beaucoup de chemin à parcourir.

- La qualité des données dépend entre autres de la qualité du protocole de collecte, or de nombreuses questions restent encore en suspens à ce sujet, même si des protocoles de collectes dans un cadre écologique ou du type "Magicien d'Oz" sont progressivement adoptés.

- La plupart des études n'évaluent leurs systèmes que sur un seul corpus de données, perdant ainsi de vue la grande diversité des cas d'utilisation dans un usage réaliste. Même si quelques travaux explorent l'évaluation "cross-corpus" des systèmes, ils restent minoritaires aujourd'hui.

Ensuite, on assiste à une stagnation de la compréhension du phénomène d'expression et de perception des émotions dans la voix. Nous pensons que cela est en partie dû à une approche du type "force brute", adoptée de manière assez large dans la communauté, qui repose sur la génération combinatoire de paramètres à partir de descripteurs et de fonctionnelles. Il faut concéder que cette approche a conduit dans beaucoup de cas à l'amélioration significative des performances de systèmes de reconnaissance, mais trop souvent sur des données en quantité trop faible ou ne répondant pas aux exigences de spontanéité et de réalisme exposées plus haut. Par ailleurs, outre les problèmes d'interprétation que pose déjà l'augmentation du nombre de paramètres, on assiste à une délégation de la complexité qui en résulte à des algorithmes de classification de plus en plus sophistiqués, et de moins en moins analysables.

Enfin, concernant les systèmes de dialogue auxquels les fonctions de reconnaissance des émotions doivent ultimement se raccorder, on constate plusieurs challenges.

- Beaucoup de travail reste à accomplir sur la question de la stratégie optimale de comportement pour un personnage virtuel expressif dans la conduite d'une interaction homme-machine affective, en particulier sur la gestion de différents types d'utilisateurs (expert/novice, jeune/âgé, connu/inconnu).

- L'évaluation des systèmes de dialogue est assez bien comprise pour des tâches spécifiques (succès d'une réservation de billets d'avion par exemple). Cependant, beaucoup d'interactions entre humains n'ont pas d'objectif particulier pouvant être caractérisé par un succès ou un échec ; la conversation informelle en est un exemple important. Or il n'existe pas de consensus sur l'évaluation d'une interaction sans tâche particulière. Les mesures objectives habituellement utilisées pour des tâches simples (succès/échec, nombre de tours de paroles) ne suffisent pas. Il faut donc concevoir des mesures de plus haut niveau et portant plus de sens.

- La problématique du long terme dans l'interaction est très rarement abordée, généralement car elle pose de nombreux problèmes logistiques. La plupart des interactions sont actuellement très courtes (de l'ordre de la demi-heure au maximum) et ne se répètent pas dans le temps. La gestion de la mémoire de ces systèmes, afférente à la contrainte du long terme, constitue également comme un obstacle important.

Le déploiement de la reconnaissance des émotions dans le monde réel nécessite donc encore des efforts importants et par là même, présente de nombreux défis à relever.

1.2 Contexte - Le projet ARMEN

C'est dans le cadre du projet ARMEN que s'est déroulée cette thèse. Il a pour but la conception d'un robot assistant pour les personnes âgées et handicapées, dans l'optique d'un regain d'indépendance personnelle et de prolongement de la vie à domicile plutôt qu'en institution. C'est une problématique cruciale dans une perspective d'évolution profonde de la démographie à moyen terme sous la forme d'un renversement de la pyramide des âges et avec des conséquences très directes comme le manque de main d'oeuvre en

accompagnement et soin des personnes dépendantes. Un robot de ce type paraît également séduisant pour des utilisateurs qui ne souhaiteraient pas, par choix personnel, d'un(e) aide-soignant(e) à domicile ; pour ceux-là, un robot serait alors vu comme une extension d'eux-mêmes leur permettant de surmonter leur invalidité.

La diversité des utilisateurs potentiels apparaît immédiatement comme très large : de tous âges, étant éventuellement atteints de pathologies vocales... Il s'agit d'une population vraisemblablement peu habituée à utiliser des systèmes informatiques, n'étant de surcroît pas forcément à même de manipuler les interfaces usuelles du type souris, clavier ou joystick. L'interaction doit donc être naturelle et le système doit s'adapter aux utilisateurs en prenant compte leurs caractéristiques, plutôt que le contraire.

Il existe relativement peu de projets de ce type, c'est donc une chance de travailler dans un contexte réaliste avec des utilisateurs aussi intéressants.

1.3 Contributions

Les principaux résultats de cette thèse portent sur plusieurs aspects :

- Une approche de réduction du nombre de paramètres a été menée pour avancer sur le problème de leur compréhension avec l'application de techniques de sélection de paramètres sophistiquées sur des données émotionnelles. L'efficacité de cette approche a été montrée dans le cadre d'un challenge international (InterSpeech 2012) et sur d'autres expériences en termes de compacité des modèles et de performance. La pertinence de groupes de paramètres particuliers a été mise en évidence, des différences relatives à la reconnaissance des émotions entre différentes populations ont été investiguées et un nouvel algorithme de sélection a été développé.

- Pour répondre au problème de rareté des données, l'approche cross-corpus est prometteuse car elle permet de disposer de plus de données d'apprentissage et de données plus variées. Nous avons montré son intérêt en termes d'amélioration des performances de classification grâce à plusieurs expériences.

- Une approche de l'évaluation d'une interaction homme-machine de type conversation informelle en terme de satisfaction de l'utilisateur a été explorée à partir d'indices dialogiques et paralinguistiques de bas-niveau.

Parmi les réalisations de cette thèse figurent également la collecte complète de deux corpus émotionnels (conception du protocole, encadrement de la segmentation et de l'annotation), avec plus de 70 patients de centres médicaux interrogés entre 2010 et 2011 ; l'implémentation d'une interface de collecte par Wizard-of-Oz utilisant un personnage virtuel expressif et l'intégration d'un démonstrateur pour le projet ARMEN. Une liste des publications est disponible en annexe I.

1.4 Organisation du document

La teneur des contributions développées dans cette thèse se reflète dans l'organisation du document. Après une première partie présentant l'état de l'art, découpée en un chapitre portant sur la théorie des émotions et de leurs expressions (chapitre 2), un chapitre décrivant l'architecture et les performances des systèmes actuels de reconnaissance des émotions (chapitre 3) et un chapitre passant en revue les interfaces homme-machine pour l'interaction émotionnelle (chapitre 4), une deuxième partie présente mes travaux relatifs à la détection des émotions à partir d'indices paralinguistiques sur des données réalistes. Au sein de cette partie, le chapitre 5 résume la collecte de deux corpus émotionnels ; le chapitre 6 retrace les expériences cross-corpus que j'ai menées pour évaluer la robustesse des modèles et le chapitre 7 expose l'algorithme de sélection de paramètres que j'ai développé et les expériences effectuées pour l'évaluer. Une troisième partie, plus applicative, décrit les travaux réalisés dans le cadre du projet ARMEN sur l'évaluation d'un système de dialogue intégrant la prise en compte des émotions dans une interaction avec un agent virtuel expressif en lien avec un robot. Elle comprend un chapitre sur l'implémentation du système proprement dit (chapitre 8) et un chapitre sur l'exploration d'une mesure d'engagement de l'utilisateur dans l'interaction à partir d'indices dialogiques et paralinguistiques (chapitre 9). Enfin, une conclusion et des perspectives sont exposées dans le chapitre 10.

Première partie

État de l’art

CHAPITRE 2

THÉORIE DES ÉMOTIONS ET DE LEURS EXPRESSIONS

2.1 L'émotion : une notion complexe

Reconnues comme partie intégrante de l'humain, il est ainsi surprenant de constater qu'à ce jour, il n'existe pas de définition ou même de cadre théorique consensuel des émotions. C'est même le contraire, une étude recensait 92 définitions pour le concept d'émotion en 1981 [144]. Le problème n'est donc pas tant que le mot "émotion" n'a pas de signification précise, c'est qu'il en a beaucoup [128]. Nous avons choisi d'utiliser la définition d'émotion suivante dans cette thèse, car elle est suffisamment générale et recouvre la plupart des aspects descriptifs et explicatifs du phénomène :

> Les émotions sont le résultat de l'interaction de facteurs subjectifs et objectifs, réalisés par des systèmes neuronaux ou endocriniens, qui peuvent : a) induire des expériences telles que des sentiments d'éveil, de plaisir ou de déplaisir ; b) générer des processus cognitifs tels que des réorientations pertinentes sur le plan perceptif, des évaluations, des étiquetages ; c) activer des ajustements physiologiques globaux ; d) induire des comportements qui sont, le plus souvent, expressifs, dirigés vers un but et adaptatifs.
>
> *D'après Kleinginna & Kleinginna (1981) [144]*

On peut ajouter que les émotions ont des propriétés bien spécifiques :

- Elles sont brèves et marquées, c'est-à-dire distinctes d'un état habituel [59]. Elles ne s'arrêtent pas brusquement en temps normal mais décroissent lentement en intensité.
- Elles ont un caractère unique.

C'est l'induction de comportements par les émotions qui nous intéresse particulièrement ici, c'est-à-dire la manifestation extérieure des émotions, particulièrement dans la voix, et également leur perception par les humains.

2.2 Théories modernes des émotions

Cette partie présente les théories modernes des émotions les plus importantes. Elles sont notamment utilisées dans le domaine de l'affective computing comme fondation des systèmes de reconnaissance d'émotions. On pourra se référer à l'Annexe III pour une perspective historique sur les théories des émotions.

2.2.1 Théories catégorielles

Les théories catégorielles sont les plus simples et les plus naturelles : elles tentent d'établir des catégories d'émotions différentes et clairement reconnaissables, tout en répertoriant les signes extérieurs reliées à ces émotions (expressions faciales, variations de la prosodie, changements physiologiques...). En général, elles désignent un petit nombre d'émotions comme étant basiques ou canoniques, les autres émotions n'étant alors que des composés des premières. Cette démarche était déjà adoptée au milieu du 17ème siècle par Descartes dans son traité *"Les Passions de l'Âme"* [72].

En se fondant notamment sur les travaux de Darwin, un des chercheurs les plus influents de ce courant, Paul Ekman, a mis en évidence l'existence de six émotions basiques et de leurs expressions faciales reconnaissables universellement : la colère, la joie, la tristesse, la peur, la surprise et le dégoût [89]. Il est même suggéré que ces émotions de base correspondent à des circuits neuronaux spécifiques et qu'elles sont clairement mesurables et donc distinctes [90]. D'autres travaux proposent des listes d'émotions basiques. Quelques uns parmi les plus reconnus sont résumés dans le tableau ci-dessous.

Il est aisé de voir les limites de ces théories : aucune ne s'accorde complètement sur une liste d'émotions et il y a des différences de granularité, c'est-à-dire que certaines émotions peuvent en contenir d'autres (par exemple, la joie de Izard peut contenir la satisfaction de Kemper).

2.2.2 Théories dimensionnelles

Les théories dimensionnelles cherchent à définir des dimensions abstraites pour s'affranchir des descriptions et du vocabulaire des théories catégorielles et ainsi les repré-

Auteurs	Nombre d'émotions primaires	Liste
Descartes (1649) [72]	6	Admiration, amour, haine, désir, joie, tristesse
Tomkins (1962, 1963) [254, 255]	8	Surprise, intérêt, joie, rage, peur, dégoût, honte, angoisse
Izard (1977) [127]	11	Joie, surprise, colère, peur, tristesse, mépris, détresse, intérêt, culpabilité, honte, amour
Plutchik (1980) [204]	8	Acceptation, colère, anticipation, dégoût, joie, peur, tristesse, surprise
Kemper (1981) [136]	4	Peur, colère, dépression, satisfaction
Ekman (1992) [89]	6	Colère, peur, tristesse, joie, dégoût, surprise

Tableau 2.I – Comparaison entre plusieurs listes d'émotions primaires

senter sur un continuum. Si certains modèles unidimensionnels ont été proposés, en général centrés autour de la valence (évaluation positive ou négative de l'émotion en terme d'agréabilité) ou de l'activation, la plupart des modèles comportent au moins deux dimensions.

Parmi les plus importants, le modèle à deux dimensions de Russel allie la valence et l'activation pour représenter plusieurs catégories d'émotions [217]. Cependant cette approche est critiquée car une projection sur deux dimensions uniquement causerait une perte d'information trop importante : par exemple, la peur et la colère sont presque confondues dans un espace valence/activation [59, 106].

Déjà en 1874, Wundt argumentait pour une représentation des sentiments en trois dimensions (agréable/déplaisant, excité/calme et tension/relaxation) [274]. Plus tard, le modèle PAD utilise trois dimensions assez similaires (*Pleasure, Activation, Dominance*) [177].

Plutchick a développé un modèle que l'on peut qualifier d'hybride, car il mélange une notion d'intensité à des couples d'émotions opposées (anticipation/surprise, rage/terreur...) [204]. Une représentation graphique de ce modèle sous forme de cône existe, où l'axe vertical représente l'intensité ; les émotions contraires sont diamétralement opposées et les émotions proches sont placées de manière adjacente. (cf Figure 2.1).

Il existe d'autres représentations graphiques des correspondances entre les émotions

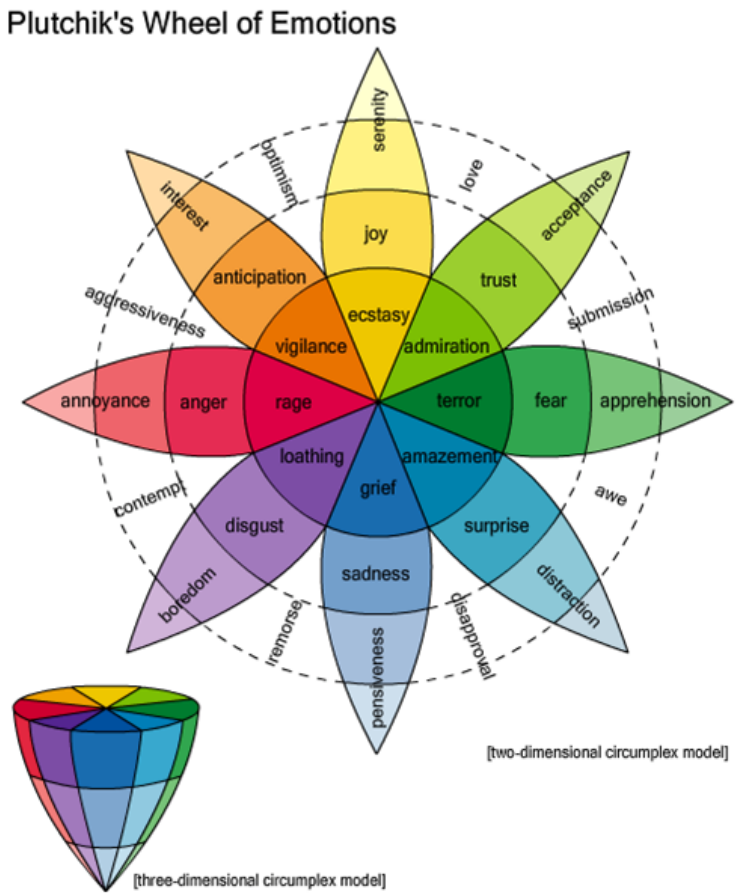

Figure 2.1 – Représentation graphique du modèle circumplexe de Plutchik (d'après [205])

catégorielles et un modèle dimensionnel. L'outil Feeltrace a notamment été développé dans cet objectif [61]. Il permet l'annotation de données audio et vidéo : les émotions doivent être placées, sous forme de pastilles colorées, dans un disque centré sur un état neutre et orienté par les axes valence et activation (cf Figure 2.2). Une composante temporelle est présente car la taille des pastilles varie avec le temps.

2.2.3 Théories cognitives de l'évaluation

Il s'agit d'un des modèles les plus sophistiqués, théorisé à partir des années 1980 [108, 152, 220]. Il postule que les émotions sont nées de l'évaluation d'évènements par rapport à des critères internes ; c'est donc un modèle génératif des émotions et pas seule-

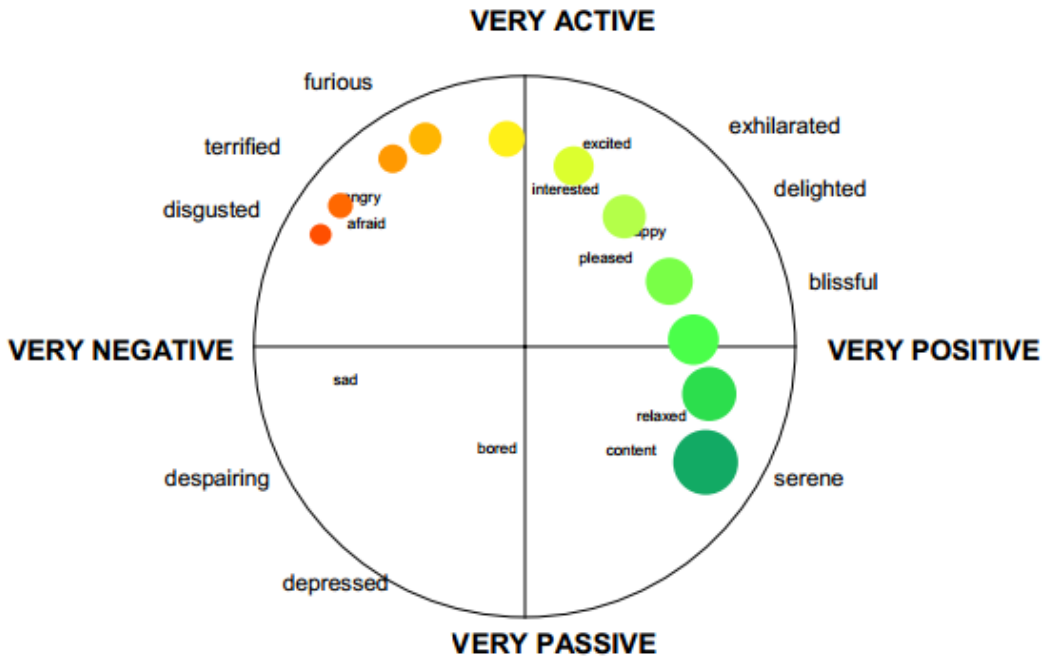

Figure 2.2 – Capture d'écran de l'outil Feeltrace (d'après [61])

ment perceptuel. Les mécanismes cognitifs (pas nécessairement conscients ou contrôlés) complexes et cachés qui permettent cette évaluation sont expliqués et pris en compte. La grande nouveauté est également d'intégrer une dynamique temporelle, en contraste avec les théories précédentes qui affectaient une étiquette à un état considéré comme statique. Les processus de vérification des critères (*checks*) s'effectuent séquentiellement de manière rapide [221]. Dans le *Componential Process Model* de Scherer, cinq critères (SEC - *Stimulus Evaluation Check*) sont vérifiés sur un total de dix-huit variables :

Nouveauté (Soudaineté, Familiarité, Prévisibilité) : caractère inattendu ou non de l'évènement ;

Agrément (Intrinsèque ou global, Désirabilité) : expérience plaisante ou déplaisante ;

Rapports aux causes et buts (Causalité interne, Causalité externe, Pertinence, Degré de certitude dans la prédiction des conséquences, Attentes, Opportunité, Urgence) ;

Potentiel de maîtrise (Contrôle de l'évènement, Contrôle des conséquences, Puissance, Ajustement) : possibilité de s'adapter ;

Accord avec les standards (Externes, Internes) : accords aux normes sociales et concepts de soi.

Cette théorie est très complète, mais elle n'est pas exempte de problèmes : certains concepts demeurent difficiles à expliquer, tels que l'amour ou le désir [93] ; elles restent contre-intuitives et donc difficiles à expliquer ; la plupart des résultats expérimentaux supportant la théorie ont été obtenus par auto-évaluation des sujets interrogés. De plus, elle est très difficile à appliquer dans un contexte de détection automatique d'émotions car la plupart des étapes se font de manière cachées, dans l'intimité des processus cognitifs du sujet, même si des corrélations entre la vérification des critères SEC et des réactions externes (expressions faciales, modification de la voix) ont été reportées dans la littérature [222]. Sa cohérence a par contre été évaluée avec succès par un modèle informatique [218] et elle peut être bien adaptée dans un contexte génératif, pour un cas de synthèse d'émotion pour un personnage virtuel par exemple [57].

On peut noter qu'il existe des pendants à ces théories dans le domaine du verbal et de la linguistique. La théorie de l'énonciation de Charaudeau [45] décrit un processus complexe d'évaluation en cinq modalités basées sur des critères subjectifs.

2.3 Expression vocale des émotions

Il est bien connu qu'une personne émue peut voir sa voix altérée de manière reconnaissable par d'autres. En fait, ce changement est si important que la voix est l'un des vecteurs d'émotion les plus importants [178]. Des études ont été menées pour tenter de comprendre l'influence des émotions sur l'expression vocale, trouver des mesures pertinentes de cette altération et inférer quelles modifications de la voix permettent la perception d'une émotion donnée chez les autres. Différents protocoles expérimentaux ont été utilisés à ces fins (études perceptives et corrélations statistiques, manipulation voire synthèse de signaux vocaux émotionnels) ; ils ont permis de montrer que de manière fiable et répétitive, certaines altérations de la voix étaient perçues comme relevant

d'une émotion particulière. Par exemple, une voix perçue comme exprimant de la tristesse est généralement de faible intensité et d'intonation plus grave qu'une voix dans un état neutre, avec un rythme de parole plus faible ; au contraire, une voix exprimant de la joie a un débit de parole plus élevé, une intonation plus aigue et est de plus forte intensité. Certaines des corrélations entre mesures acoustiques objectives et états émotionnels perçus sont présentées dans le tableau 2.II.

	Stress	Colère/rage	Peur/panique	Tristesse	Joie/euphorie	Ennui
Intensité	↗	↗	↗	↘	↗	
F0 (moyenne)	↗	↗	↗	↘	↗	
F0 (variabilité)		↗		↘	↗	↘
F0 (étendue)		↗	↗(↘)	↘	↗	↘
F0 (direction de la trajectoire)		↘		↘		
Énergie en haute fréquence		↗	↗	↘	(↗)	
Vitesse de parole et taux d'articulation		↗	↗	↘	(↗)	↘

Tableau 2.II – Effets des émotions sur l'expression vocale, mesurés à l'aide de plusieurs paramètres acoustiques (d'après [222]).

La description de ces mesures est intéressante pour le domaine de l'*affective computing* car elle suggère la possibilité une reconnaissance automatique des états émotionnels à partir de la prosodie uniquement, c'est-à-dire l'intonation de la voix, en apprenant les modifications de ces mesures objectives selon l'émotion exprimée. C'est l'un de champs de recherche du domaine, qui pourrait déboucher sur des systèmes complémentaires à la reconnaissance de la parole et à une compréhension bien meilleure de la communication humaine.

2.4 Utilisation des théories des émotions dans l'*affective computing*

L'approche imaginée par l'*affective computing* apporte de nouvelles méthodes et techniques, mais elle s'appuie essentiellement sur les travaux théoriques et expérimentaux menés auparavant par les psychologues. Cependant, malgré la diversité des théories existantes et leur sophistication, la plupart des études en reconnaissance automatique des émotions fait le choix d'utiliser de manière très simple des catégories d'émotions correspondant à la tâche à accomplir. Les raisons sont multiples mais il s'agit essentiellement d'un choix pragmatique car il n'existe pas de base d'émotions primaires consen-

suelles, que les descriptions dimensionnelles peuvent parfois être compliquées à mettre en oeuvre pour l'annotation des données[1], et que les théories de l'évaluation sont tout simplement trop compliquées à appliquer.

C'est également le choix adopté dans cette thèse, où une approche résolument catégorielle est suivie.

Des efforts de standardisation sont menés pour proposer un modèle cohérent à la communauté, par exemple avec le standard EmotionML du W3C [225], applicable pour l'annotation des données, la reconnaissance automatique d'états émotionnels et la génération de comportements émotionnels[2].

Actuellement, les recherches évoluent vers la reconnaissance d'états autres qu'émotionnels à partir d'indices paralinguistiques : des caractéristiques propres au locuteur comme l'âge, le sexe [179, 230], la corpulence ou la taille [188] ; des états affectifs comme la frustration [278] ou le stress [151] ; des états physiques comme la fatigue ou l'intoxication alcoolique [162] ; ou encore des états cognitifs comme la certitude ou les tentatives de mensonge [115].

[1]Une proportion significative de travaux utilise cependant une description dimensionnelle, dans un but de régression plutôt que de classification

[2]La version 1.0 du standard est très récente, les dernières recommandations datant d'avril 2013. Il est disponible à l'adresse suivante : `http://www.w3.org/TR/emotionml/`.

CHAPITRE 3

SYSTÈMES AUTOMATIQUES DE RECONNAISSANCE DES ÉMOTIONS

Cette partie présente les aspects d'ingénierie et de recherche relatifs à la conception et la mise en place de systèmes automatiques de reconnaissance des émotions.

3.1 Composants d'un système de reconnaissance

Les systèmes complets se décomposent en plusieurs modules, selon le modèle qu'on peut trouver dans le domaine de la reconnaissance de la parole [15]. On se place dans la situation où un seul utilisateur interagit avec le système ; on ne considère donc pas les problèmes de superposition de voix ou de détection de locuteur. Les modules sont présentés ci-dessous et représentés sur le schéma de la figure 3.1.

Captation du son Dans un système en ligne, un microphone permet de capter la voix de l'utilisateur ; dans un système hors-ligne, des fichiers sons sont lus sur disque. La qualité de la captation est importante pour la suite des traitements ; autant que faire se peut, des microphones de bonne qualité sont utilisés.

Pré-traitement Ce module est optionnel ; il regroupe les étapes éventuelles de filtrage fréquentiel, adaptation du niveau sonore, annulation d'écho... Ces techniques proviennent des domaines de l'acoustique et du traitement du signal. Leur but est de nettoyer le signal d'éventuels bruits et d'améliorer le signal de parole.

Segmentation Le flux sonore est découpé de manière à séparer la parole de l'utilisateur du silence ou du bruit. Le but est d'obtenir des segments audio à l'échelle de la phrase, avec une unité sémantique et émotionnelle. Contrairement au domaine de la reconnaissance automatique de la parole où une transcription est souvent disponible [35], la segmentation est dite ici "aveugle", c'est-à-dire uniquement basée sur la prosodie. Plusieurs méthodes de segmentation ont été répertoriées [135] : certaines se fondent sur les variations d'énergie pour repérer les silences, d'autres

modélisent explicitement les différents contenus susceptibles d'être présents dans l'audio, d'autres encore examinent la différence entre deux fenêtres glissantes... Au delà de la segmentation de la parole uniquement, des évènements paralinguistiques peuvent également être segmentés, tels que des respirations, des rires, des hésitations, des pleurs... Une approche récente utilise la factorisation en matrices non-négatives du spectrogramme audio pour détecter de tels évènements [235].

Extraction de paramètres Une représentation numérique est calculée à partir des segments audio afin d'extraire les variations de prosodie et d'intonation pertinentes. Généralement des paramètres spectraux, rythmiques et prosodiques sont extraits, mais on peut ensuite leur appliquer des fonctionnelles (dérivée temporelle, extrema, moments statistiques...) [16]. L'ensemble des paramètres primaires (*Low-Level Descriptors*) extraits pour le Challenge Interspeech 2009 est décrit dans le tableau 3.I ci-dessous ; couplés à des fonctionnelles, ils résultent en 384 paramètres finaux (*features*). Il n'existe pas de consensus dans la littérature pour décider quel ensemble de paramètres est optimal pour une tâche de reconnaissance donnée ; cela est d'ailleurs parfois décrit comme le "Graal" du domaine [10]. Les paramètres sont décrits de manière plus détallée ci-dessous.

Classification À ce stade, des techniques d'apprentissage automatique (*machine learning*) sont mises en oeuvre pour distinguer entre les différentes émotions présentes dans l'audio. Si des systèmes-experts ont été utilisés dans les premières approches [132, 198], les techniques d'apprentissage supervisé et les classifieurs statistiques du type SVM (*Support Vector Machines*), GMM (*Gaussian Mixture Models*) ou réseau de neurones se sont ensuite imposés. Ces techniques nécessitent des données exemplifiées pour pouvoir apprendre les concepts à différencier, elles ont donc donné naissance à un besoin important de données émotionnelles annotées (cf section 3.3).

Interface utilisateur Ce module est optionnel et peut servir à signifier de manière plus explicite l'émotion détectée à l'utilisateur, par exemple en utilisant un personnage virtuel réagissant de manière adéquate en fonction du résultat de détection.

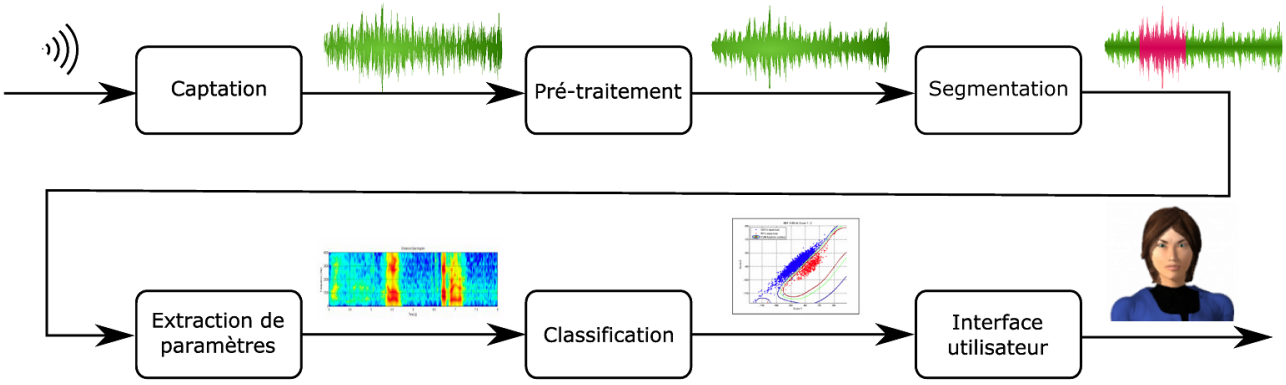

Figure 3.1 – Schéma de fonctionnement d'un système de reconnaissance des émotions dans la voix.

Paramètre de base	Type de paramètre
F0	Prosodie (timbre)
Énergie RMS	Prosodie (énergie)
MFCC 1-12	Spectre
HNR	Qualité vocale
ZCR (taux de passage par zéro)	Domaine temporel

Tableau 3.I – Résumé des paramètres primaires utilisés pour le Challenge Interspeech 2009 [229]. Couplés à des fonctionnelles, ils résultent en 384 paramètres finaux.

Les paramètres acoustiques utilisés servent à étudier précisément la prosodie de la voix. Elle caractérise les phénomènes vocaux supra-segmentaux, c'est-à-dire les propriétés attribuées aux segments de paroles qui sont plus grands que les phonèmes, tels que les syllabes, les mots, les phrases ou les tours de paroles complets [195]. Les caractéristiques perçues de la prosodie, comme le *pitch* (hauteur perçue), le débit de parole, ou le niveau sonore se répartissent alors dans les trois dimensions principales :

- le timbre, définie comme la "couleur" de la voix ;
- le rythme, mesurant la fréquence des évènements sonores (phonèmes, silences) ;
- l'intensité, qui représente l'énergie de la voix.

Ces caractéristiques n'ont pas toujours d'équivalent acoustique unique dans le signal de parole, mais il existe des paramètres acoustiques avec lesquels elles sont très corrélées, comme la fréquence fondamentale F0, très corrélée au pitch, ou l'énergie du signal à court terme, corrélée à la "bruyance" perçue. On peut distinguer quatre types de paramètres acoustiques : prosodiques, spectraux, relatifs à la qualité vocale et au domaine temporel.

Les paramètres prosodiques se divisent en paramètres relatifs à la F0 (ils décrivent ses changements de valeur dans le temps à l'échelle d'un mot, d'une phrase ou d'un tour), à l'énergie et à la durée. Les paramètres spectraux décrivent les caractéristiques du signal de parole dans le domaine fréquentiel hors F0, comme les harmoniques ou les formants, qui sont des résonnances de la fréquence fondamentale produites par le conduit vocal ; les MFCC (*Mel Frequency Cepstral Coefficients*), utilisant un filtre perceptif et issus du domaine de la reconnaissance de la parole, sont aussi utilisés de manière standard. Les paramètres de qualité vocale les plus utilisés sont le *jitter*, le *shimmer* et le rapport harmoniques/bruit (HNR) ; ils permettent de distinguer entre plusieurs types de voix (modale, c'est-à-dire neutre, soufflée, rauque) et également d'étudier des pathologies de la voix [250]. Enfin, les paramètres du domaine temporels comme le taux de passage par zéro (ZCR) ou le ratio voisé/non-voisé permettent notamment de détecter de présence de la parole.

Plusieurs logiciels permettent d'extraire ces paramètres du signal vocal. On peut citer entre autres Praat [27] et WinPitch [172] pour le calcul de la F0 et des formants et les librairies généralistes openEAR [98] et YAAFE [173]. Nous utilisons openEAR dans le cadre de cette thèse.

3.2 Problématiques relatives à l'apprentissage automatique

3.2.1 Généralités

À partir d'un corpus d'apprentissage, c'est-à-dire une base d'exemples représentés numériquement par plusieurs paramètres, les méthodes d'apprentissage automatique tentent de relier un concept aux valeurs des paramètres. Lorsque le concept est composé

de catégories discrètes, on parle de classification ; lorsqu'il s'agit d'une grandeur continue, on parle de régression. Si les valeurs du concept sont connues pour les exemples, on parle d'apprentissage supervisé.

Le modèle construit à partir des données d'apprentissage est appelé classifieur ou parfois prédicteur. Son but est d'être capable de prédire le concept correctement à partir des paramètres, à la fois sur le corpus donné, mais également sur des exemples non-vus pendant l'apprentissage. La possibilité de correctement classifier des exemples non-vus provient de l'hypothèse selon laquelle les données vues en apprentissage sont représentatives du reste des données possibles.

3.2.2 Problème du sur-apprentissage

Les deux objectifs (performance intra et extra-corpus) ne sont similaires que jusqu'à un certain point, déterminé par la quantité d'information apprise par le classifieur. Par exemple, considérons une tâche de classification d'arbres à partir de paramètres pertinents (taille, couleur de l'écorce, forme des feuilles) et d'autres moins (météo), sans que l'on puisse savoir quels sont les paramètres pertinents. En utilisant de plus en plus de paramètres pertinents, le classifieur va vraisemblablement voir sa performance de reconnaissance augmenter intra et extra-corpus, jusqu'au moment où les informations non-pertinentes vont être prises en compte. À ce moment, alors que la performance intra-corpus va continuer à augmenter, la performance extra-corpus va peu à peu décrocher (cf figure 3.2) ; le classifieur est alors en train d'apprendre du "bruit" plus que le concept désiré. C'est le problème de sur-apprentissage. Pour donner un autre exemple, il s'agit de toute la différence entre reconnaître les éléments constitutifs du style d'un musicien permettant d'être capable de l'identifier sur un enregistrement inconnu, et mémoriser un nombre limité de ses oeuvres ; on considérerait alors qu'une oeuvre inconnue a été interprétée par un autre artiste, puisqu'on ne la reconnaît pas.

Le but ultime de l'apprentissage automatique n'est donc pas tellement la bonne performance sur les données d'apprentissage, mais sur les données non-vues. Il s'agit du concept de généralisation.

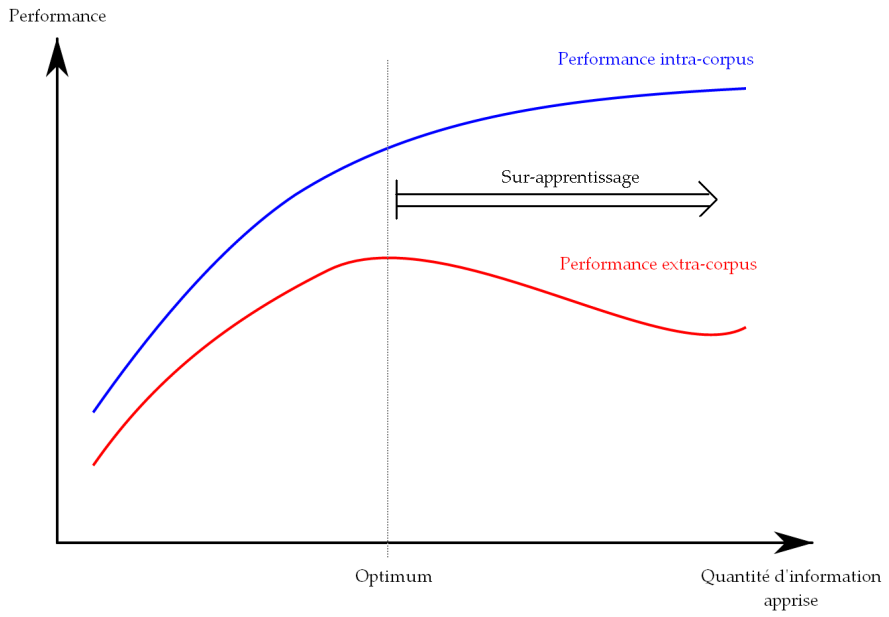

Figure 3.2 – Illustration du phénomène de sur-apprentissage

3.2.3 Évaluation : métriques et méthodologies

Il existe de nombreuses mesures pour évaluer la performance d'un classifieur. La plus naturelle est le taux de reconnaissance moyen (RR - *recognition rate*), c'est-à-dire le pourcentage d'exemples dont la classe est prédite correctement par le classifieur. De nombreuses autres mesures existent, qui permettent de dépasser les limitations du taux de reconnaissance moyen (sensibilité aux distributions déséquilibrées). Parmi elles, le taux de rappel moyen non-pondéré (UAR - *Unweighted Average Recall*) est très utilisée au sein de la communauté [15]. Ces deux métriques sont appliquées dans cette thèse.

Pour estimer le pouvoir de généralisation d'un classifieur, c'est-à-dire sa performance hors corpus d'apprentissage, la méthodologie la plus simple est de diviser les données en deux, d'en utiliser une partie pour l'apprentissage et de réserver l'autre pour le test du classifieur. Mais on sait que les classifieurs sont d'autant plus performants qu'ils disposent de beaucoup de données pour leur apprentissage. Il est donc dommage de "gâcher" des données, qui sont souvent coûteuses à collecter... Une solution à ce problème est d'entraîner un classifieur sur le premier corpus et de tester sur le deuxième,

puis de relancer une procédure d'apprentissage en échangeant les corpus : on peut ainsi utiliser la totalité des données pour l'apprentissage et en même temps évaluer le classifieur sur la totalité des données. Cette procédure élégante est appelée validation croisée (*cross-validation*). Si l'on découpe le corpus en plus de deux parties, on parle de validation croisée à N plis, avec N choisi souvent à 10 empiriquement.

La validation croisée est très utilisée, car c'est un estimateur non-biaisé de la performance hors corpus [2]. Cependant, elle n'est pas suffisante : la plupart des algorithmes de classification possèdent des réglages spécifiques, appelés hyper-paramètres, qui sont souvent déterminants dans la performance finale. Leur réglage n'a rien d'intuitif, il faut donc les régler petit à petit en essayant différentes configurations. Or on utilise la performance du classifieur pour évaluer chaque configuration et même en validation croisée, on court le risque du sur-apprentissage en multipliant les apprentissages. Il s'agit en fait d'une problématique de recherche complète (*parameter tuning*), qui a donné naissance à quantité d'heuristiques et de procédures (par exemple, des descentes de gradient dans l'espace des hyper-paramètres [133] ou des recherches aléatoires [21]). Parmi elles, la méthodologie *"train/develop/test"* permet de dépasser ces problèmes : les données sont séparées en trois ensembles (apprentissage, développement, test). Les hyper-paramètres sont réglés en entraînant autant de classifieurs que nécessaire sur l'ensemble d'apprentissage et en testant sur l'ensemble de développement. Une fois les meilleurs hyper-paramètres trouvés, ils sont utilisés pour entraîner un nouveau classifieur, cette fois-ci sur les ensembles d'apprentissage et de développement concaténés ; il sera évalué sur l'ensemble de test.

3.2.4 Algorithmes - Détails sur les SVM

Parmi la multitude d'algorithmes d'apprentissage existant, aucun n'a été identifié comme optimal pour le problème de reconnaissance des émotions. Beaucoup d'algorithmes différents sont donc utilisés [15], même si certains sont plus fréquents que d'autres, comme les Random Forests, les SVM ou les réseaux de neurones. Les réseaux de neurones bénéficient d'ailleurs depuis peu d'un regain de popularité avec les réseaux récurrents à base de *long short-term memory* [270] et le *deep learning* [121]. Dans cette

thèse, les SVM sont utilisés la plupart du temps, car c'est un algorithme qui s'est révélé efficace dans de nombreux domaines, que ses mécanismes sont relativement simples à comprendre et qu'il existe des implémentations *open-source* faciles à prendre en main, comme libSVM [44].

Les SVM (*Support Vector Machines*), formalisés par Vladimir Vapnik en 1995 [257], traitent un problème de classification binaire en minimisant le risque structurel : l'hyperplan de séparation entre les deux classes est choisi pour maximiser la marge, c'est-à-dire l'espace entre les instances les plus proches des deux classes (cf figure 3.4). Il est en effet intuitif que plus la marge est grande, meilleur sera le pouvoir de généralisation.

Il s'agit d'un algorithme utilisant la distance (produit scalaire) entre les instances pour résoudre la classification. À l'origine conçu pour traiter les problèmes linéairement séparables, il peut être étendu à des problèmes beaucoup plus complexes. En effet il est fréquent qu'un problème ne soit pas linéairement séparable, mais qu'en projetant les données dans un espace de dimension supérieure, on puisse trouver un hyperplan de séparation. Plutôt que de préciser explicitement cet espace et pour des raisons de facilités de calculs, on a recours au *"kernel trick"* : plutôt que de calculer explicitement la transformation, on remplace le produit scalaire entre deux vecteurs transformés par une fonction-noyau [37]. Une illustration d'un problème non-linéairement séparable résolu par projection dans un espace à l'aide d'un noyau gaussien est donnée figure 3.3.

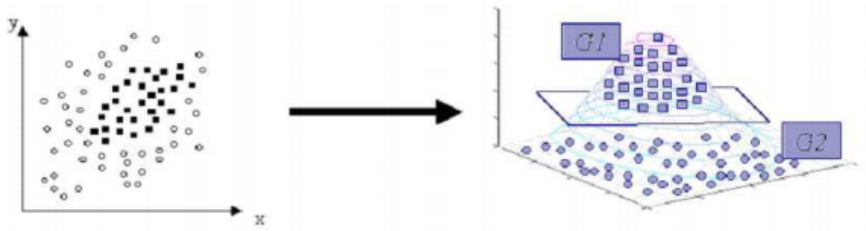

Figure 3.3 – Illustration du kernel trick

Le choix de l'hyperplan optimal s'appuie sur les données : comme on le voit dans la figure 3.4, il suffit de deux instances et de la contrainte de marge maximale pour déterminer complètement l'hyperplan H_3. Ces deux instances sont appelées "vecteurs supports" de H_3.

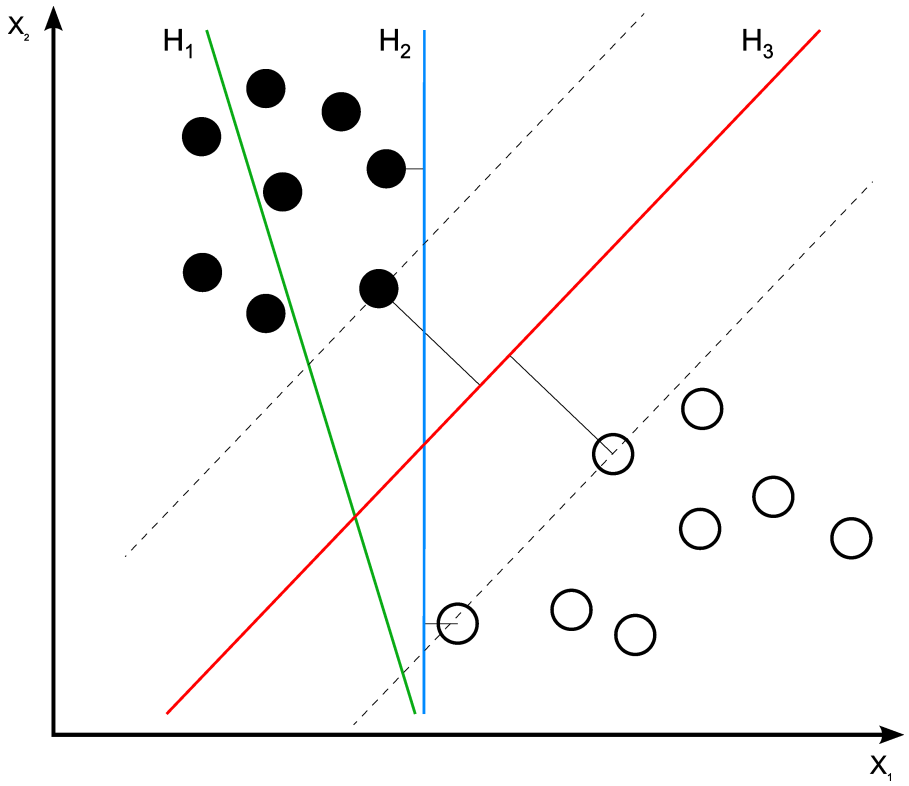

Figure 3.4 – Illustration de la séparation avec marge maximale. L'hyperplan H_1 ne sépare pas les données, H_2 et H_3 les séparent mais H_3 a la marge maximale.

Les SVM peuvent être étendus au cas multi-classe en combinant plusieurs SVM binaires (approches *"1 versus 1"* ou *"1 versus all"* [125]).

3.3 Corpus émotionnels : collecte et annotation, spontanéité des données

Les premières études s'appuyaient essentiellement sur des données en très petite quantité, avec peu de locuteurs et un contenu linguistique limité, dans un environnement contrôlé de type laboratoire [70]. La plupart du temps, des acteurs ont été enregistrés en train de jouer des émotions sur commande. Ainsi l'analyse d'un panel de 104 études sur l'expression vocale des émotions en 2003 a montré que 87% utilisaient des données actées [131]. Cependant il a rapidement été montré que cela n'était pas suffisant pour tenir compte de la variabilité des locuteurs, des voix, des situations... De plus les émotions

actées sont perceptivement et quantitativement différentes des émotions spontanées, à la fois dans leur expression vocale [223, 252, 269] et faciale (sourires de Duchenne). Enfin, les modèles entraînés sur des données actées ont de piètres performances s'ils sont testés sur des données spontanées [11, 12]. Le rôle central de bonnes bases de données a encore été souligné récemment, que ce soit pour l'apprentissage de systèmes de reconnaissance des émotions ou pour la validation d'agents affectifs compétents [60].

Pourquoi, dès lors, la communauté n'utilise pas uniquement des données spontanées ? Il s'avère qu'il est extrêmement difficile et coûteux de les collecter : les émotions sont rares en temps normal, représentant jusqu'à moins de 10% des données [15], il faut donc en recueillir d'importants volumes pour espérer extraire suffisamment de données émotionnelles. Ces données doivent être de plus annotées, ce qui est long et coûteux si des annotateurs professionnels sont impliqués (on estime que la segmentation et l'annotation des données audio nécessitent un temps environ 10 fois plus important que la durée des données). Un seul annotateur n'est pas suffisant car son choix est subjectif. Il faut donc l'avis de plusieurs personnes et opérer une sorte de vote ; il est ainsi recommandé d'utiliser au moins 2 annotateurs, la variabilité des annotations se stabilisant autour de 10 annotateurs [1, 81]). De plus les données spontanées posent souvent des problèmes de confidentialité ou de protection de la vie privée : c'est le cas notamment des données de centres d'appel, qui sont abondantes mais ne peuvent être partagées facilement avec la communauté. Un bon compromis à l'utilisation d'acteurs semble donc être l'élicitation d'émotions dans un cadre expérimental du type Magicien d'Oz [11]. Des bonnes pratiques pour la collecte de données ont été données dans [75].

Le LIMSI a depuis une dizaine d'années recueilli et annoté un grand nombre de corpus émotionnels [74]. Par exemple, le corpus BOURSE est mentionné dans le tableau 3.II ci-dessous, regroupant les principaux corpus disponibles pour la communauté et leurs caractéristiques.

Identifiant et publication associée	Contenu émotionnel	Méthode d'élicitation	Taille	Notes	Langue
Leeds [216]	Ensemble d'émotions intenses	Naturelle : interviews non-scriptées, remémoration d'expériences émotionnelles intenses	Environ 4h et demi	Discours interactif non-scripté	Anglais
Depression [107]	Dépression, état suicidaire, neutre	Naturelle : sessions thérapeutiques, conversations téléphoniques et évaluations post-thérapie	115 sujets (48 femmes et 67 hommes)	Discours interactif non-scripté	Anglais
CREST [85]	Ensemble large d'états et d'attitudes émotionnels	Naturelle : interactinos sociales et domestiques parlées, enregistrées par des volontaires sur de longues périodes	Cible : 1000 heures en 5 ans	Discours interactif non-scripté	Anglais, mandarin et japonais
BOURSE [77]	Principalement négatif (peur, colère, stress)	Naturelle : appels à un standard téléphonique de service financier	100 dialogues	Discours interactif non-scripté	Français
SYMPAFLY [14]	9 classes	Interactions Humain-Machine avec un système de dialogue	110 dialogues, environ 29000 mots	Utilisateurs réservant des billets d'avion avec une machine	Allemand
AIBO [246]	10 classes	Interactions Humain-Machine entre des enfants et un robot	51 enfants, environ 56000 mots	Enfants donnant des ordres à un robot	Allemand et anglais
Danish [94]	5 classes (colère, joie, tristesse, surprise, neutre)	Acté	Quatre sujets lisant une dizaine de phrases neutres en jouant des émotions	Scripté et acté	Danois
Berlin [140]	7 classes	Acté	Dix sujets (5 hommes et 5 femmes) lisant 10 phrases neutres en jouant des émotions	Scripté	Allemand

Tableau 3.II – Liste des principaux corpus émotionnels.

Concernant l'annotation des données, c'est une problématique beaucoup plus compliquée qu'initialement considéré, avec par exemple l'apparition fréquente de mélanges d'émotions qui rend difficile les annotations et les accords interjuges. Un point important concerne l'évaluation de l'annotation : il existe une multitude de mesures d'agrément inter-annotateurs, la plus connue et la plus utilisée étant le Kappa de Cohen [52]. Ses défauts ont été pointés très tôt [104] et de nombreuses modifications ont été proposées, par exemple pour tenir compte de plus de deux annotateurs [73, 247]. Cependant elle reste la mesure de choix pour reporter un agrément inter-annotateurs, en faisant référence de plus aux "valeurs seuils", qui ont pourtant été montrées comme inadaptées [39]. Nous utiliserons tout de même le Kappa pour évaluer nos annotations, mais en le complétant de l'accord inter-annotateur brut et de représentations graphiques (cf section 5.3.2.2).

Il existe des logiciels pour l'annotation de données. Parmi ceux-là, on peut citer ANVIL [142], plutôt adapté à la vidéo, et Transcriber [9], pour l'audio uniquement. Transcriber a été utilisé dans cette thèse car nous ne travaillons pas sur des données visuelles. De plus, son écriture dans le langage de script Tcl/Tk le rend aisé à modifier et adapter.

3.4 Performances des systèmes actuels

La performance des systèmes actuels est très difficile a évaluer : la plupart des études utilisent des données différentes et ne testent pas leurs modèles sur d'autres données, notamment sur des données réalistes. De plus les données sont annotées différemment : les étiquettes émotionnelles n'ont pas de définition standard et peuvent donc se recouvrir, leur nombre varie, parfois des schémas dimensionnels sont utilisés... Les algorithmes d'apprentissage varient également beaucoup, les paramètres utilisés pour représenter les données ne sont pas les mêmes et les mesures utilisées pour évaluer la performance diffèrent d'une étude à l'autre. Cette variété est illustrée avec des exemples de systèmes connus dans le tableau 3.III.

Il est donc extrêmement difficile de comparer les systèmes entre eux. À cet égard, l'organisation des challenges InterSpeech a permis d'avancer et de regrouper la com-

Auteurs	Performance	Informations
Petrushin (1999) [199]	~77% RR	Deux états émotionnels (agitation et calme) ; données représentées par 8 paramètres
McGilloway et al. (2000) [174]	55% RR	Cinq classes (peur, joie, neutre, tristesse et colères) ; 40 locuteurs (20h/20f), 197 phrases lues
Lee et al. (2001) [155]	79% RR	Deux classes : négatif (colère ou frustration) et non-négatif (neutre, positif, joie) ; usagers d'un système de dialogue de centre d'appel ; 142 phrases
Busso et al. (2004) [38]	71% RR	Quatre classes ; une actrice lisant 258 phrases en jouant des émotions
Lee & Narayanan (2005) [156]	81% RR	Deux classes, usagers d'un système de dialogue de centre d'appel ; 1367 phrases (591h/776f)
Neiberg et al. (2006) [190]	90%	Trois classes, 7 619 phrases tirées d'un corpus de conversations téléphoniques de centre d'appel
Kim et al. (2007) [141]	91.9% RR	Deux classes (colère et neutre), corpus acté de 1 964 phrases avec 4 locuteurs
Schuller et al. (2007) [228]	51.3% RR	Quatre classes, environ 6 000 mots
Polzehl et al. (2009) [206]	67.6% UAR	Deux classes, environ 10 000 tours de paroles provenant d'interactions en allemand entre des enfants et un robot
Yildirim et al. (2011) [276]	~62% UAR	Trois classes (politesse, neutre, frustration), 15 585 tours de parole spontanée collectés avec un dispositif "Magicien d'Oz", avec 103 enfants de 6 à 14 ans
Kotti (2012) [148]	87.7% RR	7 classes, 535 phrases prononcées par 10 acteurs

Tableau 3.III – Comparaison des performances entre quelques systèmes de reconnaissance des émotions.

munauté en proposant un cadre clair et des *benchmarks* permettant de comparer rigoureusement les performances des participants. Il reste cependant beaucoup de chemin à parcourir, par exemple en mettant en place des évaluations cross-corpus systématiques sur des données spontanées partagées par la communauté.

CHAPITRE 4

INTERACTION ÉMOTIONNELLE AVEC DES MACHINES

Ce chapitre s'intéresse principalement à l'interaction entre les machines et les utilisateurs. En particulier, il décrit plusieurs types de dispositifs conçus pour interagir de manière naturelle et émotionnelle.

4.1 Éléments théoriques de la communication non-verbale humaine

Après avoir présenté les émotions et leurs expressions au chapitre 2, nous allons maintenant nous intéresser à d'autres aspects et phénomènes de la communication non-verbale humaine. Une étude fameuse et controversée a consacré son importance : 7% seulement de la communication humain serait verbal [178]. Mais la notion est assez large, regroupant plusieurs concepts .

Tout d'abord, le rôle important des aspects non-verbaux dans la communication des humains (mais également d'autres espèces comme certains mammifères) peut s'expliquer pour partie par le concept d'empathie. Beaucoup de définitions existent [66], certaines très larges, englobant des phénomènes connexes comme la contagion émotionnelle ou la sympathie, d'autres plus précises. Comme le terme n'est pas central dans cette thèse, nous adopterons une définition large de l'empathie, comme une *compréhension des sentiments de l'autre et un partage des affects* [209]. Historiquement, c'est un concept très lié à l'action motrice : le psychologue Theodor Lipps suggérait déjà au début du 20ème siècle qu'en simulant mentalement une expression faciale, c'est-à-dire en adoptant la disposition mais sans véritablement bouger un muscle, on pouvait directement faire l'expérience de l'émotion correspondante [201]. La découverte des neurones mirroirs chez les singes [215] puis récemment chez les humains [137] est à cet égard vue comme une preuve directe de cette théorie. Il s'agit de neurones correspondant à une certaine action motrice s'activant en observant simplement cette action chez un autre, comme si le cerveau voulait imiter les actions sans véritablement les exécuter.

Différents phénomènes ont été étudiés dans la littérature. La CAT (*Communication accommodation theory*), développée à partir des années 70, décrit des comportements d'imitation et d'adaptation entre des interlocuteurs, à la fois pour minimiser ou faire ressortir leurs différences. Ces comportements concernent des modalités aussi différentes que l'accent [112], la structure de la parole en termes de débit [249], de fréquence et durée des pauses [26], les postures [54], les expressions faciales [267]. Le mécanisme du feedback interactionnel par les *back-channels* [88], incluant des "continueurs" ("hum", "aha") et des gestes (hochements de têtes, rires) est également considéré comme important dans l'interaction, permettant de montrer au locuteur qu'il a l'attention du public. On peut aussi citer d'autres micro-expressions telles que les *affect bursts* [8].

L'évolution dynamique de ces phénomènes interactionnels, leur cooccurence et l'adaptation des locuteurs dans leurs comportements est désignée par le terme de synchronie [69]. L'étude de la synchronie inclut notamment la coordination interactionnelle, exemplifiée par le fait que des interlocuteurs sont capables de prédire de manière précise le début et la fin des phrases de la conversation, phénomènes multimodaux marqués par la syntaxe, la morphologie et l'intonation [51]. Elle se fait maintenant de manière automatique en bénéficiant des avancées récentes en analyse de l'image, du son et en aprentissage automatique, par exemple, en analysant de manière multimodale la coordination entre les gestes et la parole dans une expérience de jeu de construction en coopération [68]. Une bonne analyse des dernières avancées sur l'analyse automatique des comportements non-verbaux dans les interactions sociales en petits groupes est disponible ici : [110].

Des modèles d'implication affective dans l'interaction ont été développés, envisageant l'expression d'émotions dans le contexte de la conversation, avec des déclencheurs et des réactions des autres locuteurs [114]. Un schéma d'annotation de cette implication selon un axe prospectif et un axe réactif, tous deux gradués sur une échelle d'intérêt, a été décrit récemment et appliqué dans le cas de données issues d'un centre d'appels téléphoniques [260]. Une illustration en est faite figure 4.1.

Pour avoir des machines efficaces dans la communication avec des humains, il ne suffit donc pas de détecter leurs signaux non-verbaux, il faut encore les comprendre

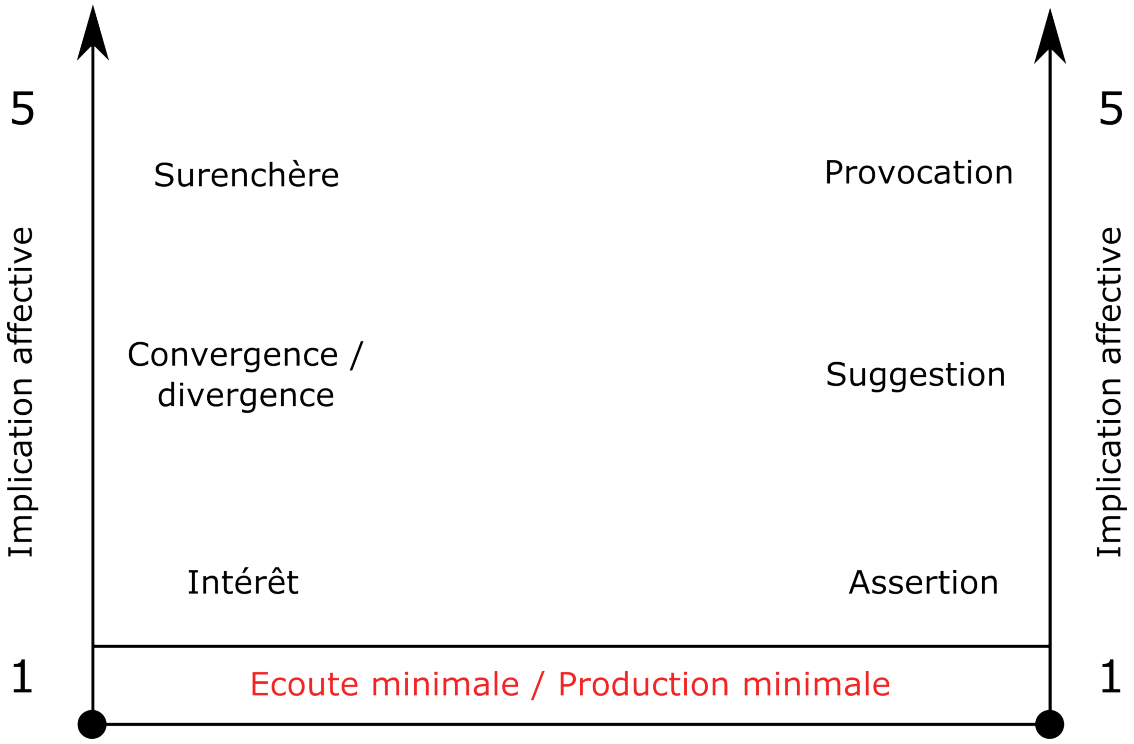

Figure 4.1 – Illustration d'un modèle d'implication affective en interaction, d'après [258]

dans leur contexte et savoir y réagir. Les phénomènes interactionnels décrits plus hauts commencent donc à être appliqués à des personnages virtuels dans le but explicite de fluidifier la communication avec eux [116]. L'utilisation de signaux non-verbaux pour améliorer l'interaction sociale homme-robot est également explorée [67].

4.2 Machines interactives

4.2.1 Agents virtuels expressifs

Le besoin d'interfaces naturelles avec les systèmes, incluant un aspect émotionnel, a donné lieu à la conception de personnages virtuels, c'est-à-dire des représentations réalistes ou non de personnages constituant le point focal de communication avec le système dans une métaphore de "conversation en face-à-face" [42], plutôt qu'une interface basée sur la métaphore du bureau et les skeuomorphismes [62]. Ils sont en effet décrits

comme idéaux pour explorer les interactions homme-machine grâce à leur apparence anthropomorphique [23]. Ces personnages apparaîssent donc rétrospectivement comme naturels avec le développement des jeux vidéos dès les années 70, mais leur utilisation dans d'autres types de systèmes n'émerge qu'à partir des années 90 [43].

Les recherches sur les personnages virtuels, plus fréquemment appelés agents virtuels expressifs (AVE), s'orientent dans plusieurs directions. Une partie des chercheurs s'attachent ainsi à augmenter le réalisme de leur apparence, car il a été montré qu'il augmente l'engagement de l'utilisateur [275]. Profitant des progrès du matériel informatique et des techniques d'infographie, des personnages au rendu quasi photo-réaliste existent désormais [71]. D'autres chercheurs travaillent plutôt sur la notion d'expressivité et en particulier sur les expressions faciales. Leur modélisation est ainsi un axe de recherche important. Un exemple important de modèle est le système de codage FACS (*Facial Action Coding System*) [91] : il décompose toute expression en "unités d'action" élémentaires mettant en jeu un ou plusieurs muscles. Il a été adapté et est très utilisé pour animer des personnages virtuels [5, 193]. Un autre exemple est la norme de codage MPEG-4, qui présente un volet utilisable pour la modélisation des expressions faciales d'émotion [197] ; le système GRETA est basé sur ce standard [192]. On peut également parler du langage de balise BML (*Behavior Markup Language*), plus large, permettant de contrôler le comportement à la fois verbal et non-verbal des personnages virtuels [264] ; il a été utilisé dans cette thèse pour contrôler l'agent virtuel MARC [58] (cf chapitre 8). Un autre axe de recherche est l'étude de phénomènes interactionnels à l'aide d'AVE, comme la synchronie d'expressions ou de sourires entre deux agents en interaction [207].

Les AVE ont été utilisés dans des applications très diverses, comme un recruteur virtuel pour les simulations d'entretien avec un modèle affectif [124, 130], tuteur virtuel pour l'*e-learning* [175], ou encore pour un jeu de go interactif, avec un AVE muni d'un modèle émotionnel de type évaluatif (cf section 2.2.3), qui lui permet d'adapter ses expressions faciales en fonction du déroulement du jeu [56, 57].

4.2.2 Robots

Les robots peuvent être vus comme le pendant physique et concret des AVE. Après s'être longtemps concentré sur les applications industrielles et les problématiques du type manipulation ou navigation, le domaine de la robotique s'intéresse de plus en plus aux robots humanoïdes et utilisables dans un contexte social. Plusieurs concepts ont ainsi été formalisés récemment ; leurs définitions sont parfois assez larges et se recoupent souvent, mais tous ont en commun une conception du robot centrée sur l'humain. On peut ainsi présenter :

Les robots compagnons [65, 248] : ils sont supposés afficher un comportement sociable et réaliser différentes tâches en coopération avec un utilisateur humain ; ils sont ainsi censés assister de manière proactive leurs utilisateurs dans les tâches quotidiennes et interagir de manière intuitive, expression et affective. Plusieurs rôles sont ainsi fréquemment imaginés : babysitter, assistant ou domestique.

Les robots assistants [99] : désigne des robots qui assistent des personnes handicapées ou âgées par une interaction physique. Cette définition n'est cependant pas assez large.

Les robots sociaux [105] : décrit les robots dont la tâche principale est l'interaction.

Les robots cognitifs [266] : un robot cognitif est défini comme étant autonome et capable d'inférence, de perception et d'apprentissage.

Les robots sociaux assistants [99] : reprend les définitions de robot assistant et robot social : un tel robot doit porter assistance aux utilisateurs humains, mais cette assistance doit se faire à travers une interaction sociale ; le but du robot est alors de créer une relation proche et efficace pour que l'utilisateur puisse réaliser des progrès mesurables dans sa tâche (convalescence, rééducation, apprentissage, etc).

Le nombre de robots conçus ne cesse d'augmenter : un site internet en liste plus de 500 pour les dix dernières années[1]. Pour se rendre compte de la diversité des robots dans

[1] `http://www.plasticpals.com/?page_id=26736`

leur forme, leurs fonctionnalités et leurs utilisations prévues, quelques exemples ont été regroupés dans le tableau 4.I. Ils sont représentés sur la figure 4.2.

Nom	Type	Année de lancement	Utilisation
PaPeRo	Créature	1997	Robot de la firme NEC, utilisé dans une étude comme coach pour la perte de poids [138]
Kismet	Tête expressive	1999	Utilisé dans la recherche pour l'interaction [29]
AIBO	Robot animal (chien)	1999	Commercialisé par Sony, utilisé en recherche pour l'interaction [13]
Care-o-Bot	Plate-forme mobile	1999	Robot de l'institut Fraunhofer, conçu pour l'assistance aux personnes âgées et handicapées [219]
Paro	Robot animal (bébé phoque)	2001	Utilisé en recherche pour le traitement de la démence [265]
Nexi-MDS	Tête expressive	2007	Développé par le MIT, utilisé dans la recherche pour l'interaction [28]
Nao	Robot humanoïde	2007	Développé par Aldebaran Robotics, utilisé comme plate-forme de recherche notamment au LIMSI pour l'étude de la détection des émotions spontanées [251]
Jazz	Plate-forme mobile	2010	Robot de téléprésence de la société Gostai. Utilisé en recherche dans le cadre d'un jeu affectif [7]
Reeti	Robot expressif	2012	Développé par l'entreprise française Robopec, destiné à être utilisé comme plate-forme de recherche

Tableau 4.I – Détails de quelques robots interactifs.

4.2.3 Cas des robots assistants

Parmi les robots interactifs, les robots assistants sont particulièrement intéressants car ils combinent beaucoup de problématiques comme le travail au plus près des utilisateurs, avec donc des contraintes de sécurité importantes, la coopération avec le personnel soignant, le besoin d'interaction naturelle. Ils sont destinés à jouer un grand rôle dans la soin des personnes handicapées et âgées dans les prochaines années [33] et sont sujets à controverse d'un point de vue éthique et professionnel [166]. Leur perception par de potentiels utilisateurs a été étudiée par des questionnaires et il a été montré qu'ils étaient favorables à leur utilisation dans un cadre hospitalier ou pour des applications de sécurité [95].

Plusieurs projets robotiques français récents s'intéressent au cas des robots assistants. Les projets Romeo 1 puis 2[2] ont pour objectif de concevoir un robot humanoïde d'assistance pour les personnes en perte d'autonomie, développé par Aldebaran Robo-

[2] `http://projetromeo.com/`

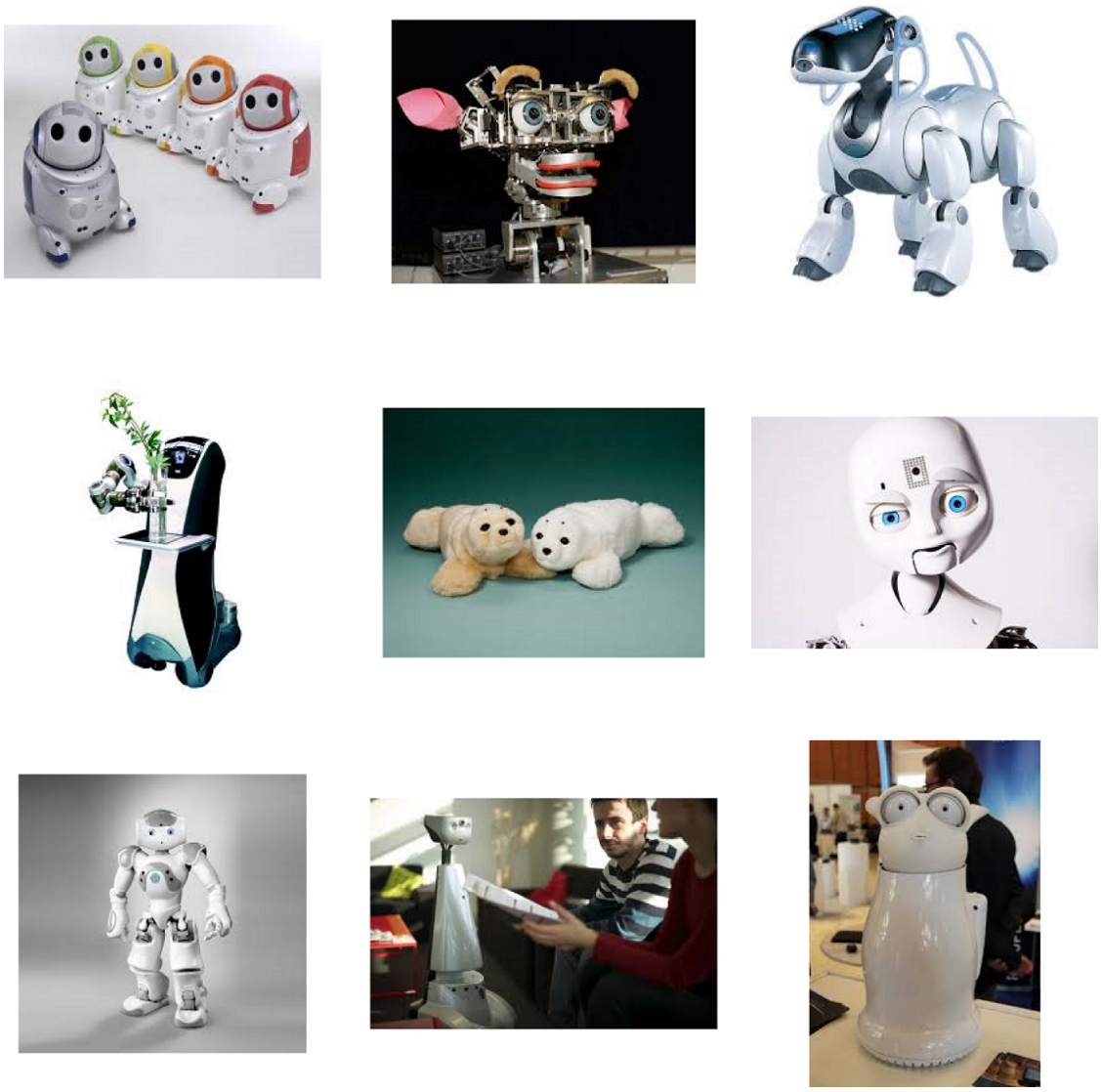

Figure 4.2 – Quelques exemples de robots interactifs ayant une forme humanoïde, animaloïde ou autre. De gauche à droite et de haut en bas : PaPeRo, Kismet, AIBO, Care-o-Bot, Paro, Nexi-MDS, Nao, Jazz et Reeti.

tics sur le modèle du robot Nao. Le projet ARMEN[3], qui finance cette thèse, développe un prototype d'assistant robotique, sous la forme d'une plate-forme mobile commandée par un AVE, pour le maintien des personnes âgées dans leur lieu de vie [154, 161] ; d'autres études décrivent des dispositifs hybrides robots-AVE similaires [149]. Le projet

[3] `http://projet_armen.byethost4.com/`

ROBADOM[4] étudie l'impact d'un robot majordome à domicile sur l'état psychoaffectif et cognitif de personnes âgées ayant des troubles cognitifs légers.

L'évaluation de ces robots consistue encore une question ouverte ; un benchmark comprenant des mesures issues de la psychologie, de l'anthropologie, de la médecine et de l'interaction homme-robot a été proposé pour évaluer l'impact des robots assistants sur l'utilisateur et sur la population en général [100].

4.3 Aspects dialogiques

Diverses problématiques liées aux systèmes de dialogue sont brièvement introduites ici. Un système de dialogue parlé peut être en général défini comme acceptant une entrée sous forme de voix humaine et retournant une sortie sous forme de synthèse vocale. Il est structuré en plusieurs modules, accomplissant chacun une fonctionnalité donnée [176] :

Reconnaissance automatique de la parole : transformation du signal audio en texte.

Compréhension du langage : extraction du sens pour pouvoir interprêter le message.

Gestion du dialogue : mise à jour du contexte, coordination avec les autres modules, plannification des réponses.

Génération des réponses : organisation du contenu qui doit être communiqué sous forme de phrases intelligibles.

Synthèse vocale : à partir d'un texte, ce module synthétise un message vocal artificiel.

Un système de dialogue naturel, émotionnel et interactif rajoute plusieurs modules à cette description. La compréhension des émotions peut se faire en parallèle à la reconnaissance de la parole pour les indices para-linguistiques et en complément de la compréhension du langage pour les indices linguistiques, venant ainsi étoffer la sémantique ; ces indices peuvent en outre être intégrés à la gestion du dialogue. De plus, la gestion du dialogue peut également être augmentée : des travaux récents ont ainsi exposé la possibilité de donner une dimension affective à la narration d'histoires simples [180] ; d'autres

[4] `http://www.robadom.vermeil.org/`

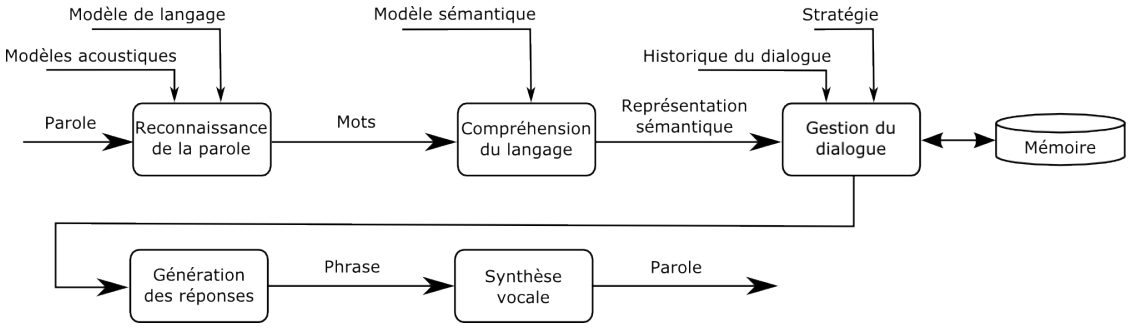

Figure 4.3 – Architecture modulaire d'un système de dialogue parlé (d'après [41]).

ont étudiés l'implémentation des effets verbaux décrits par la CAT comme la sélection d'un vocabulaire commun [208]. La génération des réponses peut prendre en compte un AVE en ajoutant une dimension expressive appropriée (expression faciale, gestuelle). La synthèse vocale doit être couplée à l'AVE pour réaliser la synchronisation des lèvres avec l'audio.

4.4 Aspects perceptifs

L'évaluation des robots et particulièrement des AVE utilise des protocoles perceptifs, c'est-à-dire que des sujets vont évaluer des traits d'un robot ou d'un AVE en exprimant leur ressenti au travers d'un questionnaire. Par exemple, la contribution de rides rendues de manière photo-réaliste à l'expressivité d'AVE a été étudiée par ce biais [55]. En ce qui concerne l'évaluation de la perception des robots, une échelle mesurant les attitudes négatives à l'égard des robots a été développée [194].

Par ailleurs, il a été montré que l'appréciation des utilisateurs varie de manière non-linéaire en fonction du degré de réalisme. Il s'agit du phénomène bien connu de la "vallée de l'étrangeté" ou *uncanny valley* [184] : une réaction négative face à un système humanoïde à la fois très réaliste mais qui n'est pas complètement humain. Ce phénomène est illustré par la figure 4.4. Cependant, plusieurs études contestent la simplicité de cette relation [169] et mettent en évidence de nombreuses causes comme une incohérence dans les niveaux de réalisme des diverses modalités d'expression du système, qu'elles soient

statiques ou dynamiques (apparence, voix, mouvements...) [182].

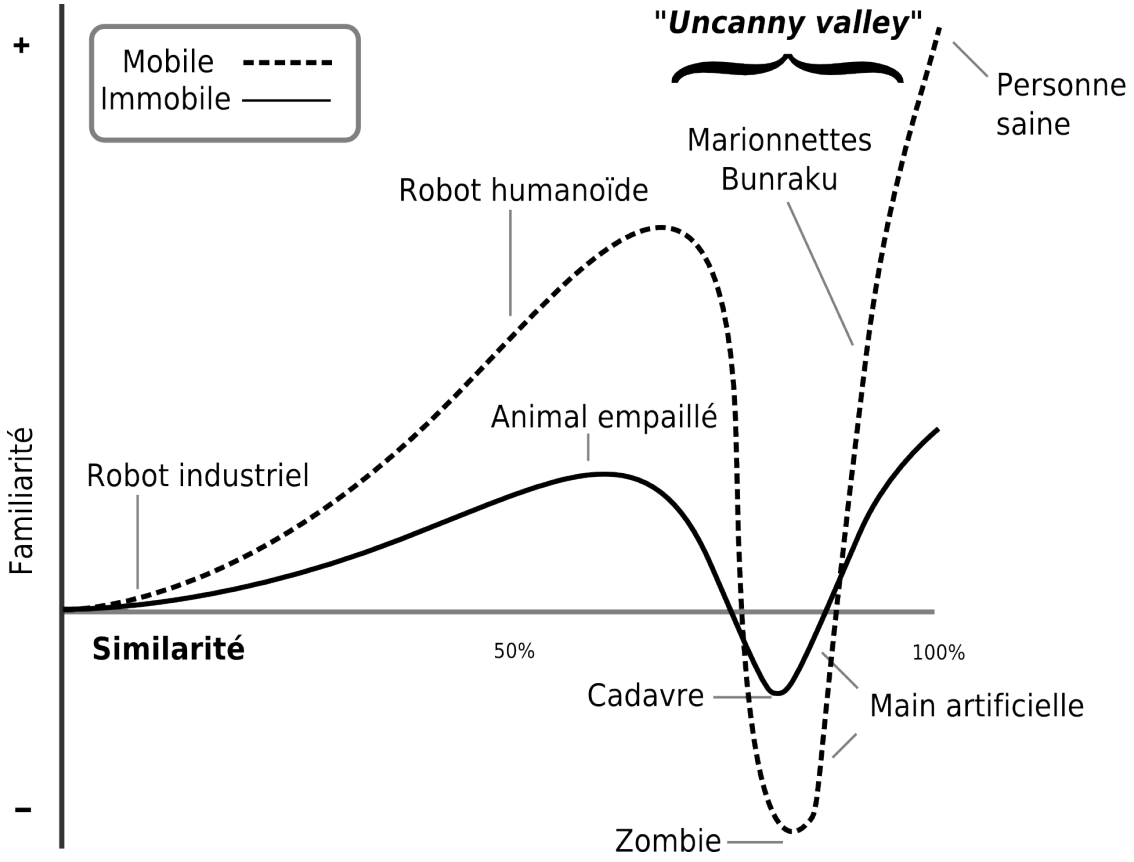

Figure 4.4 – Illustration du phénomène de la "vallée de l'étrangeté", d'après [184].

4.5 Nouveaux challenges de l'interaction homme-machine

Une caractéristique essentielle des machines interactives telles qu'elles sont imaginées depuis maintenant plusieurs années est leur utilisation quotidienne. Leur conception en prenant en compte toutes les dimensions d'une relation sur le long-terme est donc nécessaire, d'autant plus qu'il a été montré que l'effet de nouveauté se dissipe très rapidement et que les utilisateurs se lassent alors et changent leur attitude [113]. Plusieurs aspects ont été mis en évidence pour l'établissement et le maintien d'une relation homme-machine à long-terme [103] :

- Un sentiment d'appartenance : pour une relation forte et stable.
- De la compréhension : pour un partage du sens et une prédictibilité des comportements.
- De la confiance : indispensable pour être perçu comme inoffensif.
- Du contrôle : pour percevoir le lien entre comportement et conséquences.
- Une capacité d'amélioration.

Or ces caractéristiques reposent toutes sur un composant essentiel des futurs systèmes : la mémoire. C'est actuellement un problème souvent laissé de côté [164]. Les premières approches étaient orientées vers la construction de bases de données importantes et l'optimisation de la recherche dans ces bases mais plus récemment, des approches s'inspirant du fonctionnement de la mémoire chez les êtres vivants et notamment les humains sont plébiscitées [273]. Différents axes sont explorés, comme par exemple la mémoire autobiographique, dont il a été montré qu'elle permettait d'améliorer la perception des AVE en donnant l'impression d'un "soi" et d'une expérience interactive plus profonde [24] ; la mise à jour de la mémoire pour éviter les répétitions et apprendre des précédentes interactions [122] ; ou encore l'enregistrement des expériences de l'utilisateur pour pouvoir donner plus de contexte aux interactions [40]. Des études ont également intégré une mémoire émotionnelle à des AVE, c'est-à-dire la capacité d'enregistrer des expériences émotionnelles et de pouvoir les éprouver à nouveau dans un contexte similaire [163]. Quelques modèles complets de mémoires ont été proposés à ce jour [34, 123].

Pour revenir à la problématique de la relation à long-terme, plusieurs facteurs expliquent le petit nombre d'études à ce sujet. Sa mise en oeuvre expérimentale présente des difficultés de logistique (en termes de matériel, de coûts et de temps), au niveau du recrutement, de l'assiduité et de l'éventuelle rémunération des sujets expérimentaux ; les prototypes (matériel et logiciel) doivent être robustes car il faut éviter les anomalies qui pourraient endommager la relation et donc biaiser l'expérimentation [23]. Pour tenter de

contourner ces difficultés, Bickmore a mis au point un "laboratoire virtuel", permettant de faciliter les protocoles d'études longitudinales [23].

Les études de ce type sont très rares et concernent essentiellement les domaines de la santé et du bien-être [160]. On peut citer notamment un robot-réceptionniste [113], un AVE coach sportif pour les personnes âgées [25], un robot destiné à aider les personnes en surpoids en les motivant à suivre leur régime [139], un jouet robotique à la forme de dinosaure [101] ou encore un robot-livreur de snacks [157]. Cependant, plusieurs projets européens récents explorent les problématiques liées aux relations à long-terme : LIREC[5], Companions[6], SERA[7], CompanionAble[8] et ALIZ-E[9].

[5] `http://lirec.eu`
[6] `http://www.companions-project.org`
[7] `http://project-sera.eu`
[8] `http://companionable.net`
[9] `http://aliz-e.org`

Deuxième partie

Reconnaissance des émotions dans la parole

CHAPITRE 5

COLLECTE DE DONNÉES

5.1 Introduction - Motivation

En mettant l'accent sur des utilisateurs potentiels spécifiques (personnes âgées, handicapées, dépendantes...), le projet ARMEN propose de relever un premier challenge car à ce jour, il n'existe pas de corpus de données émotionnelles publiquement disponibles concernant cette population. Deux collectes ont été réalisées au cours de cette thèse, en collaboration avec l'association APPROCHE[1]. Au total, ce sont près de 80 patients de centres médicaux (centre de rééducation fonctionnelle, maison de retraite médicalisée, centre de vie pour jeunes handicapés) de la région de Montpellier qui ont été interviewés, pour un total d'environ 26 heures d'enregistrements audio et vidéo [49].

Ce chapitre présente les protocoles de collecte de ces données, avec la conception d'une interface du type "Magicien d'Oz", le travail sur la segmentation et l'annotation des données ainsi que les corpus finaux.

5.2 Protocoles et système de collecte de données

Nous avons choisi de travailler directement avec de potentiels utilisateurs finaux, dans des situations similaires à la réalité, plutôt qu'avec des acteurs ou dans un contexte de laboratoire. En effet, malgré les coûts plus importants en termes de temps et toutes les difficultés de logistique, de recrutement d'utilisateurs, d'annotation, il a été montré que les données actées ne sont pas satisfaisantes dans l'élaboration de systèmes destinés à être déployés sur le terrain [11].

Pour se rapprocher le plus possible de conditions réalistes, nous avons opté pour un protocole de collecte de type "Magicien d'Oz" [6, 134, 202]. Cette technique permet

[1]Nous tenons à remercier chaleureusement APPROCHE pour son aide dans l'organisation des collectes de données, notamment Violaine Leynaert et les professeurs Isabelle Laffont et Charles Fattal. Leur accueil et leur assistance tout au long des expériences ont été déterminants dans le succès de celles-ci.

de sortir du cercle vicieux *"la conception du système nécessite des tests avec des utilisateurs, mais un système fonctionnel doit être présenté aux utilisateurs pour que leurs réactions soient pertinentes"*. Le comportement souhaité du système est conçu à l'avance et les utilisateurs ne sont pas mis en interaction avec le système final, mais plutôt avec une maquette dont le fonctionnement est piloté par un opérateur caché, selon le comportement défini. Ainsi, le déroulement de l'interaction permet de se faire une bonne idée des cas d'usages typiques dans la réalité (au biais d'expérimentation près).

Avec ce protocole, nous avons choisi d'éliciter des émotions des utilisateurs au travers de scénarios d'interaction fortement inspirés de situations quotidiennes vécues. En collaboration avec une équipe de médecins et d'ergothérapeutes du Centre Mutualiste Neurologique Propara, membres de l'association APPROCHE, plusieurs trames de scénarios ont été conçues, écrites et validées. Les scénarios devaient respecter plusieurs contraintes : correspondre aux tests des fonctionnalités du robot, s'approcher de l'expérience-utilisateur finale, offrir une certaine diversité dans l'interaction mais en restant dans le cadre d'un dialogue limité pour des objectifs de robustesse.

Le protocole expérimental exact diffère légèrement pour les deux collectes, elles sont donc présentées séparément ci-dessous.

5.2.1 Première collecte (ARMEN_1)

La première collecte avait plusieurs objectifs : rassembler le maximum de données aussi naturelles que possibles pour l'apprentissage du système de détection des émotions, fournir des éléments permettant de guider la conception de l'AVE. Nous avons essayé de voir les sujets les plus divers possibles en termes d'âge, de sexe et de pathologies éventuelles pour avoir une idée de l'étendue des cas possibles en utilisation réelle.

5.2.1.1 Déroulement

La première collecte (cinq jours en juin 2010) faisait intervenir trois phases pour chaque sujet, qui était au préalable conduit par un membre du personnel médical dans

une salle où était installé le matériel[2].

Dans la première phase, un premier expérimentateur (*interviewer* par la suite) présentait le projet au sujet et lui expliquait le but de l'expérience pendant qu'un deuxième expérimentateur (*technicien* par la suite, rôle assumé par l'auteur dans les deux collectes) installait un micro-cravate et le calibrait. À la manière d'un jeu et en guise d'entraînement, le sujet était ensuite invité à prononcer la phrase "Ma voix exprime des émotions" en exprimant tour à tour de la colère, de la joie, de la tristesse... Dans cette phase actée, destinée à faire rentrer progressivement le sujet dans l'expérience, l'interviewer insistait pour que le sujet n'hésite pas à surjouer les émotions demandées.

Dans la deuxième phase, le sujet interagissait librement avec un système de dialogue simple dans le cadre de sept courts scénarios (des situations quotidiennes avec un potentiel émotionnel). Chaque scénario était d'abord présenté par l'interviewer, qui demandait au sujet de s'imaginer dans la situation décrite et de faire comprendre au système l'émotion ressenti ; il n'intervenait ensuite plus. Le système de dialogue était piloté par le magicien d'Oz (technicien) à l'insu du sujet, qui pensait avoir une conversation avec un système automatique qui comprenait ses réponses. Différentes stratégies de réponses avaient été pré-établies pour le technicien : comprendre, comprendre en montrant de l'empathie, ne pas comprendre, se tromper. Chaque scénario donnait lieu à un dialogue durant en moyenne 4 à 5 tours de parole pour le sujet.

Dans la troisième phase, l'interviewer avait un questionnaire à présenter au sujet. Certaines questions concernaient la qualité de l'interaction (compréhension, agréabilité), d'autres les souhaits des sujets quant à l'apparence future de l'AVE (humanoïde, réaliste ou pas, masculin ou féminin...) tandis que les dernières étaient plus ouvertes ("Souhaiteriez-vous disposer d'un tel système chez vous ?", "Lui donneriez-vous un nom ? Si oui, lequel ?"...).

Le matériel utilisé était composé d'un micro-cravate AKG sans-fil avec une carte d'acquisition audio externe M-Audio connectée à un ordinateur portable pour l'enregistrement du son (qualité 32 bits, échantillonné à 44.1kHz et sauvegardé au format WAV en mono). Un deuxième ordinateur portable permettait de contrôler le système de collecte

[2]Un petit nombre de sujets ont été interviewés dans leur chambre car ils ne pouvaient se déplacer.

et déclenchait les réponses audio sur le premier, auquel étaient connectées des enceintes. Une caméra sur un trépied, opérée par le technicien, filmait le sujet de face, en plan américain. Un schéma de l'installation pour la collecte est donné sur la figure 5.1. Il faut noter que, comme la collecte se déroulait sur site et que nous ne maîtrisions pas les locaux, les sujets n'étaient pas seuls dans une salle avec le système.

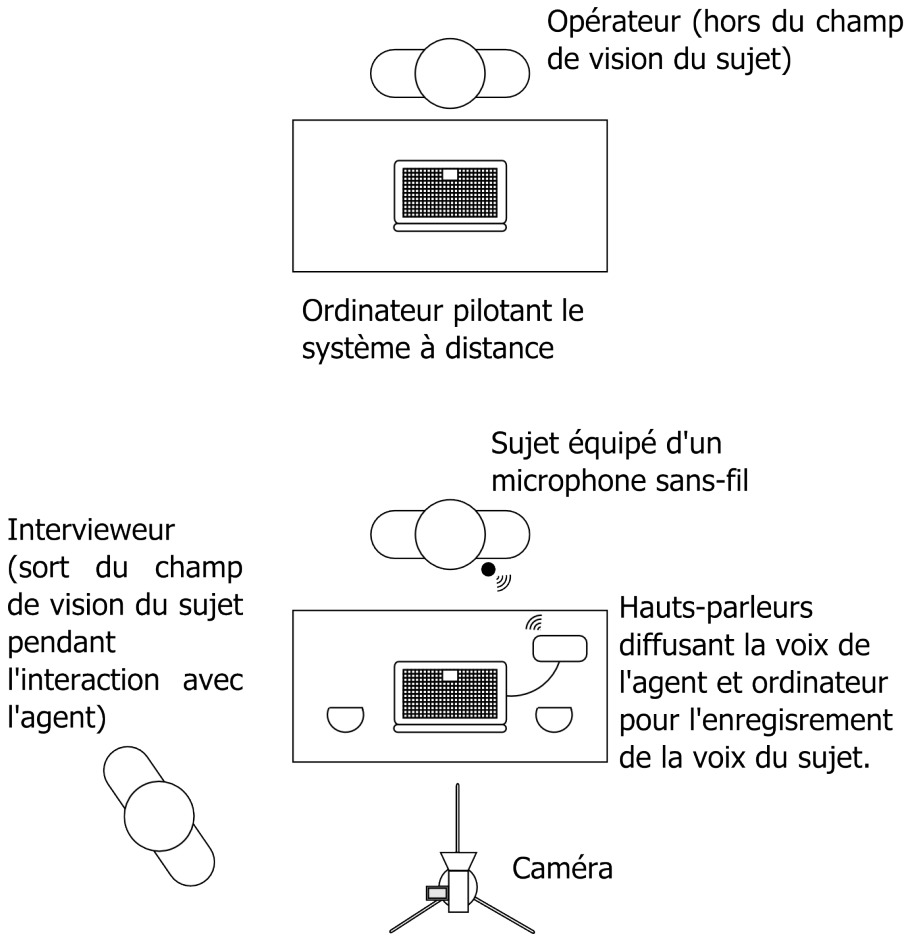

Figure 5.1 – Dispositif du matériel pour la collecte de données ARMEN_1.

5.2.1.2 Détail des scénarios

Les scénarios pour ARMEN_1 ont été conçus avec l'association APPROCHE. Plusieurs situations quotidiennement vécues par les patients des centres médicaux ont été

proposées par APPROCHE et analysées, par exemple l'oubli fréquent du pilulier et de la prise de médicaments, la difficulté de récupérer un objet comme une télécommande ou des lunettes s'il est tombé par terre ou sous un meuble pour les personnes en fauteuil roulant ; d'autres contextes ont été rajoutés. Au total, sept situations ont été retenues et scénarisées. Le détail des scénarios est donné dans le tableau 5.I.

Nom du scénario	Indications
Joie	Vous avez la visite de quelqu'un qui vous est proche. Vous êtes très heureux et joyeux et vous voulez le faire comprendre au robot.
Peur/stress	Vous n'aimez pas les piqures. Vous en avez même très peur et vous attendez l'infirmière qui doit vous en faire une. Vous faites partager votre stress au robot.
Douleur	Vous avez très mal, vous souhaitez que quelqu'un vienne d'urgence, vous criez au robot de prévenir quelqu'un, vous appelez à l'aide... la douleur est très forte.
Colère	Vous avez attendu depuis votre réveil, quelqu'un qui doit vous aider vous êtes furieux, en colère contre lui et vous en partagez votre colère avec le robot et vous vous plaignez.
Déception/Tristesse	On vient de vous apprendre que votre visiteur ne viendrait pas. Vous dîtes au robot combien vous êtes déçu et triste.
Soulagement	Vous aviez perdu vos papiers, on vient de vous les ramener. Vous expliquez votre soulagement au robot.
Fierté	Vous êtes très fier des progrès que vous avez faits, vous le dites au robot.

Tableau 5.I – Détails des scénarios pour la collecte du corpus ARMEN_1.

À chaque scénario correspondait un ensemble de 3 à 5 répliques, plus des répliques neutres pour demander la répétition d'une phrase par exemple. Les répliques ont été conçues selon plusieurs stratégies de relance d'après la littérature sur les techniques psychologiques de remédiation. Ces stratégies sont détaillées et illustrées ci-dessous par des répliques du système :

- Réaction sympathique : "Je suis content pour toi".
- Réaction empathique : "Tu as l'air heureux"
- Réaction compassionnelle : "Je suis triste pour toi, je vois que ça te fait de la peine".

- Question fermée sur l'état émotionnel : "je crois que tu es joyeux ? c'est ça ?".
- Relance conversationnelle : "Tu as l'air déçu. Est-ce que tu crois que ton ami pourra venir plus tard ?".

5.2.1.3 Détail des questionnaires

Le questionnaire présenté aux participants de l'expérience comprenait 17 questions, dont 4 questions ouvertes, 8 questions à choix multiples (dont échelles de Likert) et 6 questions binaires. Des questions concernaient la personnalité des sujets ("Quelles émotions sont pour vous les plus fréquentes ?"), d'autres les caractéristiques du futur AVE ("Préféreriez-vous un personnage homme ou femme ?"), d'autres encore étaient plus générales et concernaient le rôle ou l'utilité d'un robot personnel.

Des caractéristiques des sujets ont également été relevées (âge, sexe, pathologie, familiarité avec la technologie et l'informatique).

5.2.1.4 Détail sur les sujets

Au total, 52 patients ont été recrutés pour la première collecte. Ils provenaient de trois centres médicaux :

- Un centre de rééducation fonctionnelle (Centre Mutualiste Neurologique Propara). Les patients de ce centre sont généralement des adultes en cours de rééducation après un accident grave (typiquement un accident de la route) mais également des personnes atteintes de maladies dégénératives (sclérose en plaques par exemple). La durée de séjour est de quelques mois en moyenne. Les pathologies les plus courantes sont des paraplégies ou tétraplégies (paralysie des membres inférieurs ou des quatre membres) ; ce type de pathologie a un impact non-négligeable sur la production de la voix (pose de canule ou de valve durant l'hospitalisation en cas de coma, perte de contrôle des muscles abdominaux et donc volume sonore plus faible de la voix, paralysies faciales qui affectent l'articulation...), ce qui entraîne une grande variabilité de la qualité vocale dans cette population.

- Un centre de vie pour personnes lourdement handicapées (Centre APIGHREM). Il s'agit d'un hébergement à long terme. La plupart des résidents sont atteints de pathologies très lourdes. Au niveau des dégradations de la voix, on retrouve les mêmes que dans le centre Propara, ainsi que des bruits causés par les dispositifs d'assistance à la respiration dans certains cas.

- Une maison de retraite médicalisée (EHPAD Malbosc). Ce type de structure accueille les personnes âgées dépendantes pour une longue durée. Une large diversité de pathologies (maladie de Parkinson, maladie d'Alzheimer, démence sénile...) peut être trouvée chez les résidents, dont la plupart sont des femmes. Les qualités vocales sont également très variées (dévoisement chez certaines personnes âgées, lenteur et difficulté d'articulation dues à la prise de médicaments ou à la maladie) et on trouve quelques problèmes d'audition et cognitifs qui peuvent parfois rendre difficile l'interaction avec un système automatique de dialogue...

La répartition des patients en termes de sexe, âge et centres médicaux est illustrée sur la figure 5.2 ; les données sont résumées dans le tableau 5.II. Il faut noter la grande diversité des âges.

Centre	Nombre de sujets	Répartition hommes/femmes	Âge (min/médian/max)
Propara	29	18 / 11	16 / 46 / 72
APIGHREM	15	12 / 3	21 / 34 / 55
Malbosc	8	5 / 3	52 / 83 / 90
Total	52	35 / 17	16 / 45 / 90

Tableau 5.II – Résumé des caractéristiques des sujets de la collecte du corpus ARMEN_1.

5.2.1.5 Système de collecte

Le système de collecte pour ARMEN_1 était très simple, consistant en une interface web, dans laquelle chaque scénario correspondait à une page, comprenant une liste de "boutons-radio" permettant de déclencher une des répliques du scénario courant et une

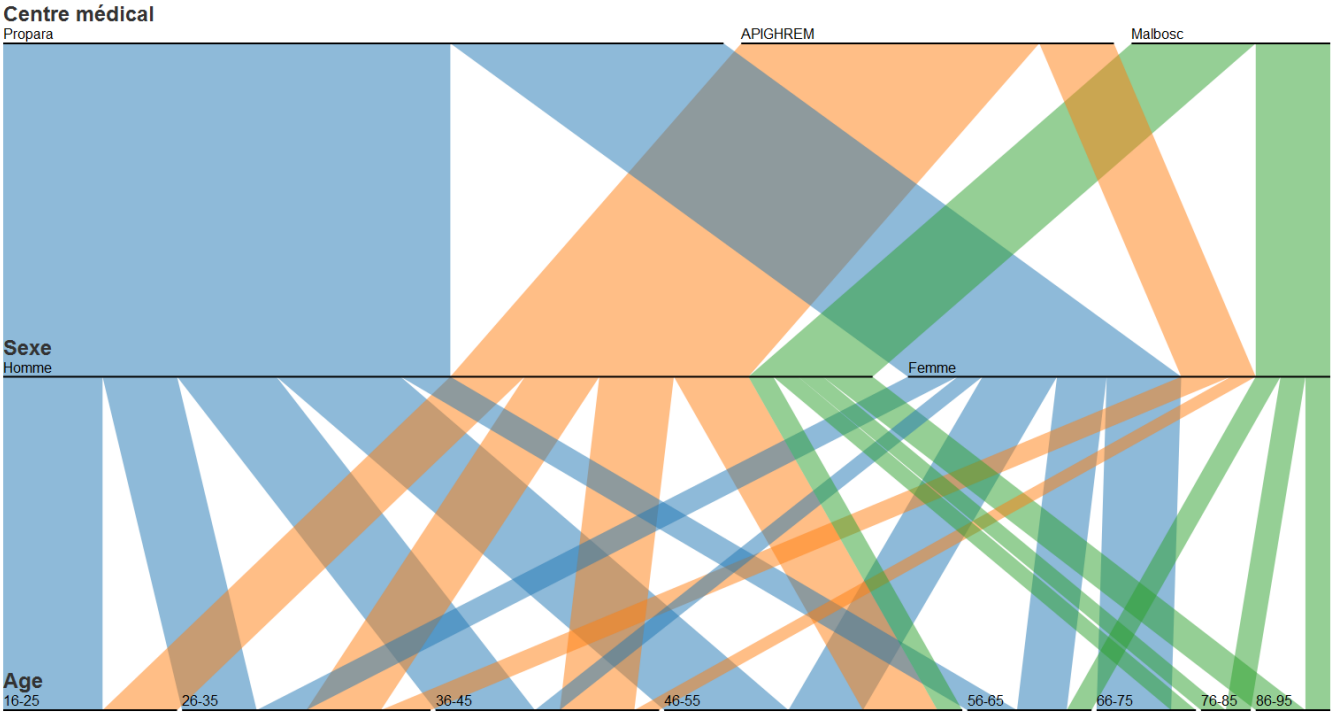

Figure 5.2 – Répartition des sujets de la collecte ARMEN_1 en termes de centres médicaux, sexe et âge.

liste additionnelle pour les répliques neutres, présente sur toutes les pages. Les répliques consistaient en des fichiers sons générés par un programme de synthèse vocale.

5.2.2 Seconde collecte (ARMEN_2)

Ayant exploré le périmètre des sujets avec la collecte ARMEN_1 et disposant de l'AVE, la collecte ARMEN_2 avait pour objectif de collecter des données dans une interaction naturelle avec un système très similaire au système final. Moins de sujets, plus ciblés, ont été interviewés.

5.2.2.1 Déroulement

Le protocole de la seconde collecte (trois jours en juin 2011) était très similaire à celui de la première : il suivait trois phases avec deux expérimentateurs. Quelques points

changeaient néanmoins :

- Les sept scénarios ont été remplacés par trois scénarios plus développés (jusqu'à vingt tours de paroles du sujet au lieu de cinq en moyenne), plus un scénario de présentation de l'AVE.
- Le système de collecte était beaucoup plus abouti et faisait cette fois intervenir un AVE.
- L'interviewer intervenait beaucoup moins car l'AVE avait une présence plus importante dans l'interaction.
- L'ordinateur portable qui ne faisait que jouer des sons dans la première collecte servait aussi à afficher l'AVE et le montrer au sujet.
- Le questionnaire a été mis à jour pour porter plus spécifiquement sur la qualité de l'interaction et l'AVE.

Un schéma de l'installation, légèrement modifié pour refléter les modifications du protocole, est donné sur la figure 5.3

5.2.2.2 Détail des scénarios

Il y avait en tout quatre scénarios, conçus cette fois par APPROCHE en se basant sur les scénarios les plus intéressants et ayant déclenché le plus de réactions dans la collecte ARMEN_1, et en les étendant.

Le premier scénario (*présentation*) était conçu pour mettre les sujets à l'aise et construire un début de proximité avec l'AVE. Chacun devait se présenter et l'AVE demandait ensuite au sujet de s'entraîner en prononçant la phrase "Ma voix exprime des émotions" en exprimant diverses émotions. Ensuite, l'AVE présentait des photos des membres de l'équipe soignante et de patients dans différentes situations (jeu, repas, anniversaire...) et demandait au sujet de nommer les personnes, d'expliquer les situations et s'il en avait un bon souvenir.
Dans le deuxième scénario (*pilulier*), l'AVE rappelait au sujet qu'il ne devait pas oublier

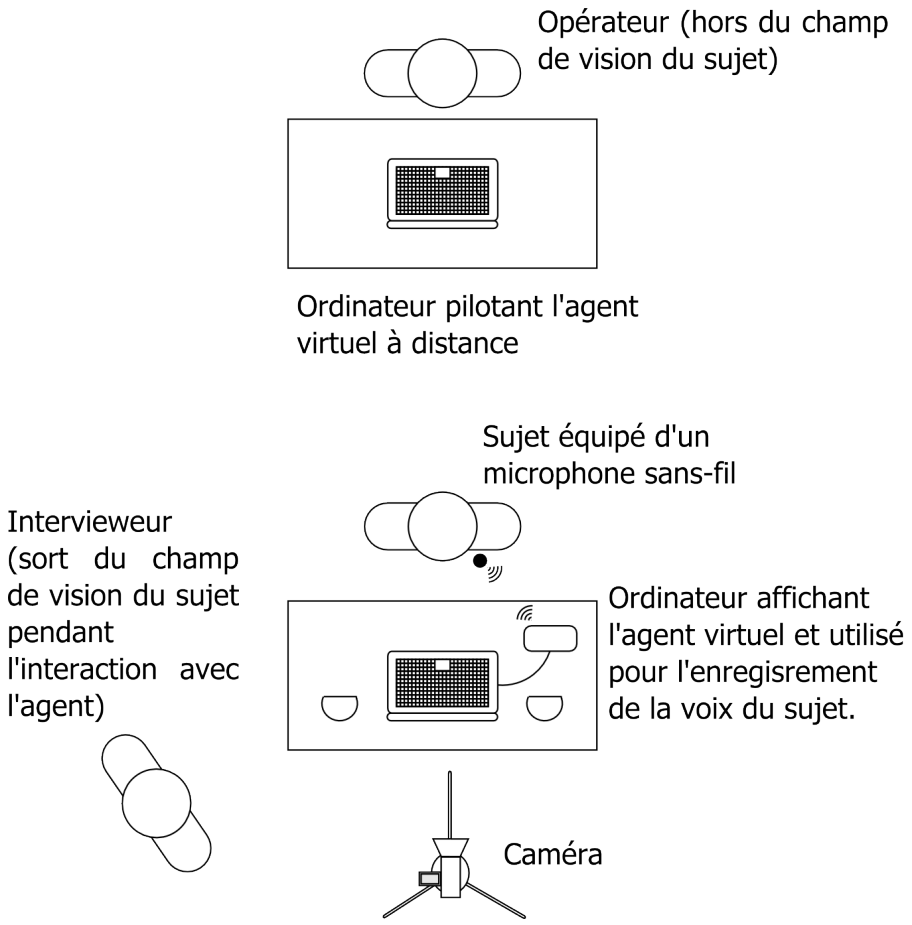

Figure 5.3 – Dispositif du matériel pour la collecte de données ARMEN_2.

de prendre ses médicaments. Il agissait ensuite comme si le robot partait à la recherche du pilulier et en attendant son retour, bavardait avec le sujet en demandant comment il se sentait, quels étaient les évènements régulièrements organisés au centre médical et lesquels il appréciait.

Dans le troisième scénario (*alerte*), le sujet devait imaginer qu'il ne se sentait pas bien et appeler l'AVE à l'aide, qui devait alors simuler un appel et tenter de rassurer le sujet.

Dans le quatrième scénario (*télécommande*), le sujet devait demander à l'AVE de trouver la télécommande car il voulait regarder la télévision. L'AVE faisait alors comme si le robot était parti chercher la télécommande et demandait au sujet ce qu'il voulait regarder, quel programme l'intéressait et bavardait sur ce sujet en attendant le retour du robot.

Nous avons écrit les répliques du système et organisé le contenu des scénarios, pour que l'interaction soit dirigée par le système. Ce choix résulte du besoin de robustesse : en restant dans un cadre délimité à l'avance et en donnant la main au système, notamment en posant des questions fermées, on réduit le risque d'incompréhension. Dans cette perspectives, les répliques ont été organisées dans un arbre de dialogue, choix courant pour les systèmes de dialogue [200]. Pour un exemple d'arbre de dialogue avec le scénario *alerte*, voir la figure 8.5 du chapitre 8.

5.2.2.3 Détail des questionnaires

Ce questionnaire cherchait à évaluer à la fois l'interaction et l'AVE en proposant une liste d'adjectifs dont le sujet devait dire s'ils correspondaient ou pas en les notant sur une échelle de Likert à cinq niveaux. Des adjectifs redondants et d'autres contradictoires ont été placés dans la liste pour vérifier la cohérence des réponses des sujets. Des questions supplémentaires étaient posées pour savoir si le sujet voudrait interagir à nouveau avec le système dans le futur et s'il aimerait posséder un tel système.

5.2.2.4 Détail sur les sujets

Cette collecte a impliqué 25 patients issus du centre Propara et de l'EHPAD Malbosc.

Centre	Nombre de sujets	Répartition hommes/femmes	Âge (min/médian/max)
Propara	9	6 / 3	24 / 48 / 74
Malbosc	16	7 / 9	54 / 84 / 91
Total	25	13 / 12	24 / 77 / 91

Tableau 5.III – Résumé des caractéristiques des sujets de la collecte du corpus ARMEN_2.

5.2.2.5 Système de collecte

Le système de collecte pour ARMEN_2 est assez complexe : il fait intervenir une interface de contrôle et la plate-forme MARC pour l'affichage et l'animation de l'AVE. L'interface de contrôle, codée sous le langage Python, utilise l'arbre de dialogue du

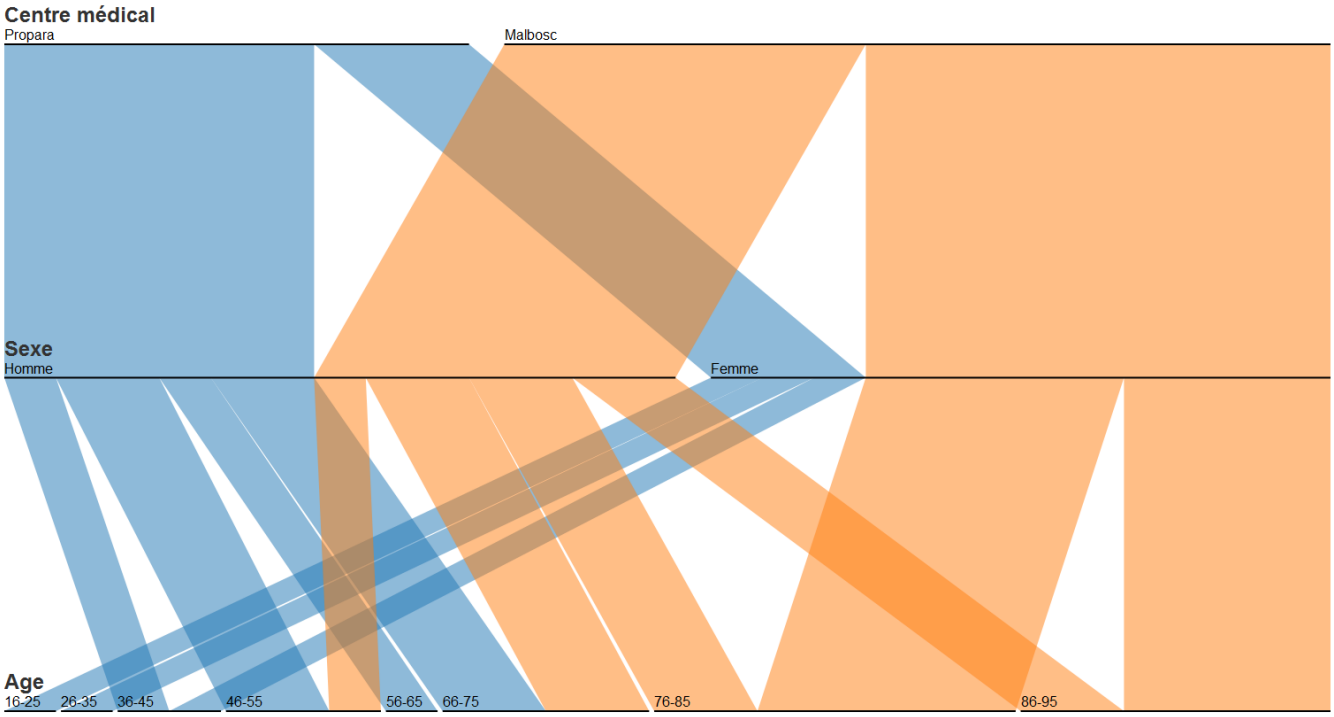

Figure 5.4 – Répartition des sujets de la collecte ARMEN_2 en termes de centres médicaux, sexe et âge.

scénario courant pour générer dynamiquement une représentation de l'arbre de dialogue, permettant à l'opérateur de facilement suivre l'interaction. Une illustration en est donnée sur la figure 5.5.

Les scénarios sont encodés sous forme d'arbre de dialogue, sous un format XML. L'interface de contrôle communique avec la plate-forme MARC via un serveur UDP qui envoie des messages BML pour déclencher les répliques et les animations.

5.2.3 Quelques remarques sur les collectes

Les collectes sur site ont été impressionnantes par l'intensité des réactions des sujets et par leur engagement très fort dans l'interaction avec un système pourtant très simple, chose à laquelle nous n'étions pas forcément préparés. Pour preuve, on peut citer quelques réponses au questionnaire de la collecte ARMEN_1 et des remarques libres

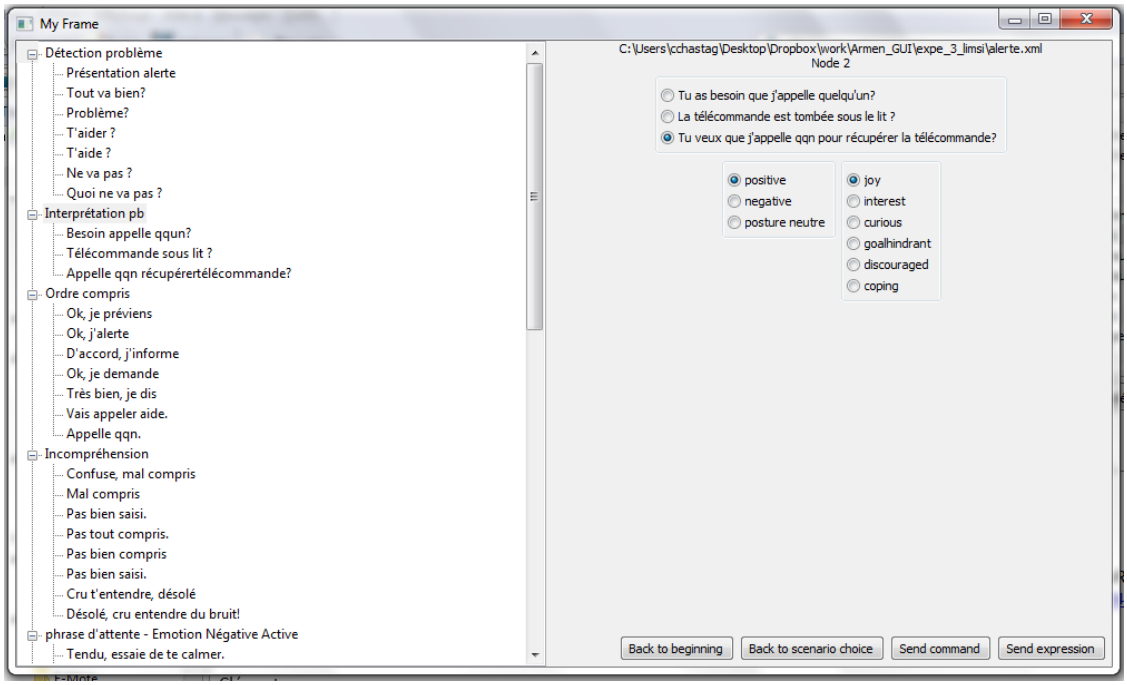

Figure 5.5 – Illustration de l'interface de contrôle pour le système de collecte ARMEN_2 en *magicien d'Oz*.

faites par les sujets :

- Plusieurs sujets ont déclaré qu'ils donneraient un nom à leur robot s'ils en possédaient un. Nous avons relevé des noms qui pourraient être donnés à des animaux domestiques (Bobby, Charlie), des noms évocateurs ((*"Marcheur, parce que c'est mon rêve"*, "Amour") et des noms de membres de la famille (mère pour deux sujets, fils pour un autre).

- Sur les 52 sujets d'ARMEN_1, 43 ont indiqués vouloir être tutoyé par le robot, 8 ont indiqué que ça leur était égal et 1 a insisté pour être vouvoyé, mais peut-être par jeu. Il est intéressant de noter que les sujets s'adressaient aux expérimentateurs par le vouvoiement. Le tutoiement avec le robot était cependant justifié par plusieurs sujets par la proximité nécessaire et souhaitée avec le robot s'ils devaient vivre avec.

- Certains sujets ont évoqué une relation très personnelle avec le robot, qu'ils imaginaient pouvoir devenir un "ami", un "confident" qui les "comprenait", voire un "membre de la famille". Ces réactions ont eu lieu alors que seule la voix synthétique interagissait avec les sujets.

L'engagement immédiat et profond d'une majorité de sujets avec le système, particulièrement lors de la collecte ARMEN_1 peut probablement s'expliquer par plusieurs facteurs : effet de nouveauté, isolement pour certains sujets... Cependant il nous semble important de devoir mener une réflexion éthique sur tous les aspects de la relation qui peut se développer avec un robot personnel.

5.3 Segmentation et annotation

Les données audio récupérées à l'issue des deux collectes ont été segmentées et étiquetées par deux annotateurs experts. La conception et la mise en application de ces étapes est expliquée ci-dessous.

5.3.1 Segmentation

5.3.1.1 Pré-traitement des données audio

Les données audio ont tout d'abord été normalisées avec un pic à -1dB pour pallier les différences de gain à l'enregistrement.

5.3.1.2 Protocole de segmentation

L'étape de segmentation permet de découper l'enregistrement selon trois niveaux hiérarchiques : de l'unité la plus grossière à la plus fine, les phases de l'expérience, les tours de paroles et les segments émotionnels.

Le protocole décrit tout d'abord comment séparer chaque enregistrement en phases correspondant aux différentes phases de l'expérience. Par exemple, pour ARMEN_1, les enregistrements étaient découpés en douze phases correspondant à la présentation

de l'expérience, l'entraînement sur la phrase "Ma voix exprime des émotions", les sept scénarios, le questionnaire, plus une éventuelle phase non-prévue (cf 5.IV).

L'audio a ensuite été découpé en tours de paroles, définis comme une période ininterrompue de discours d'un seul ou de plusieurs locuteurs s'ils sont indiscernables et parlent en même temps (superposition ou *overlap*). Toujours selon cette définition, les tours de paroles sont séparés par des silences d'au moins 500 ms. Un code a été attribué à chaque tour de parole, correspondant au(x) locuteur(s) identifiés ou à un bruit le cas échéant. Les codes pour la collecte du corpus ARMEN_1 sont listés dans le tableau 5.V.

Enfin, un découpage plus fin en segments émotionnels a été effectué. Un segment émotionnel a été défini dans le protocole comme ayant une durée minimale (500 ms) et maximale (5 s) et une unité émotionnelle ; un tour de parole peut comprendre plusieurs segments émotionnels. Un silence d'au moins 500 ms définit une frontière pour un segment émotionnel.

Le protocole était similaire pour les deux collectes (hormis quelques différences sur les codes de phases et de locuteurs). Toute l'étape de segmentation a été effectuée à l'aide du logiciel Transcriber [9].

5.3.1.3 Validation de la segmentation

Pour toutes ces étapes, les enregistrements audio ont été partagés entre les deux annotateurs. Une vérification croisée a ensuite été réalisée pour homogénéiser la segmentation. Un outil a également été développé pour pouvoir plus facilement visualiser les différences de segmentation entre deux annotateurs et quantifier les éventuelles différences (nombre de segments, pourcentage de recouvrement).

5.3.1.4 Résumé de la segmentation

Le résultat de l'étape de segmentation est détaillé dans le tableau 5.VI ci-dessous. L'histogramme des durées de segment pour les corpus ARMEN_1 et ARMEN_2 est également donné sur la figure 5.6 ; seuls les segments des sujets sont considérés, tous les autres segments (interviewer, voix du robot, superposition) ayant été éliminés.

Numéro	Thématique
1	Introduction - Explications
2	Entraînement
3	Scénario "joie"
4	Scénario "peur/stress"
5	Scénario "douleur"
6	Scénario "colère"
7	Scénario "déception/tristesse"
8	Scénario "soulagement"
9	Scénario "fierté"
10	Scénario libre
11	Questionnaire
12	Phase supplémentaire/non-prévue

Tableau 5.IV – Code des phases pour le corpus ARMEN1

Qui parle ?	Code locuteur
Sujet	*NuméroSujet*
Voix du robot	armen
Interviewer	itvr
Plusieurs locuteurs	ovl
Autre (bruit, autre locuteur non-identifié...)	aut

Tableau 5.V – Code locuteurs et correspondances pour le corpus ARMEN1

	ARMEN_1	ARMEN_2
Total		
Nombre de segments	23879	6714
Durée totale	14h45m42s	6h00m27s
Sujets seulement		
Nombre de segments	10533	3590
Durée totale	5h54m59s	2h27m16s
Durée minimale	0.23s	0.25s
Durée médiane	1.85s	2.27s
Durée maximale	7.43s	7.96s

Tableau 5.VI – Résumé de l'étape de segmentation pour ARMEN_1 et ARMEN_2.

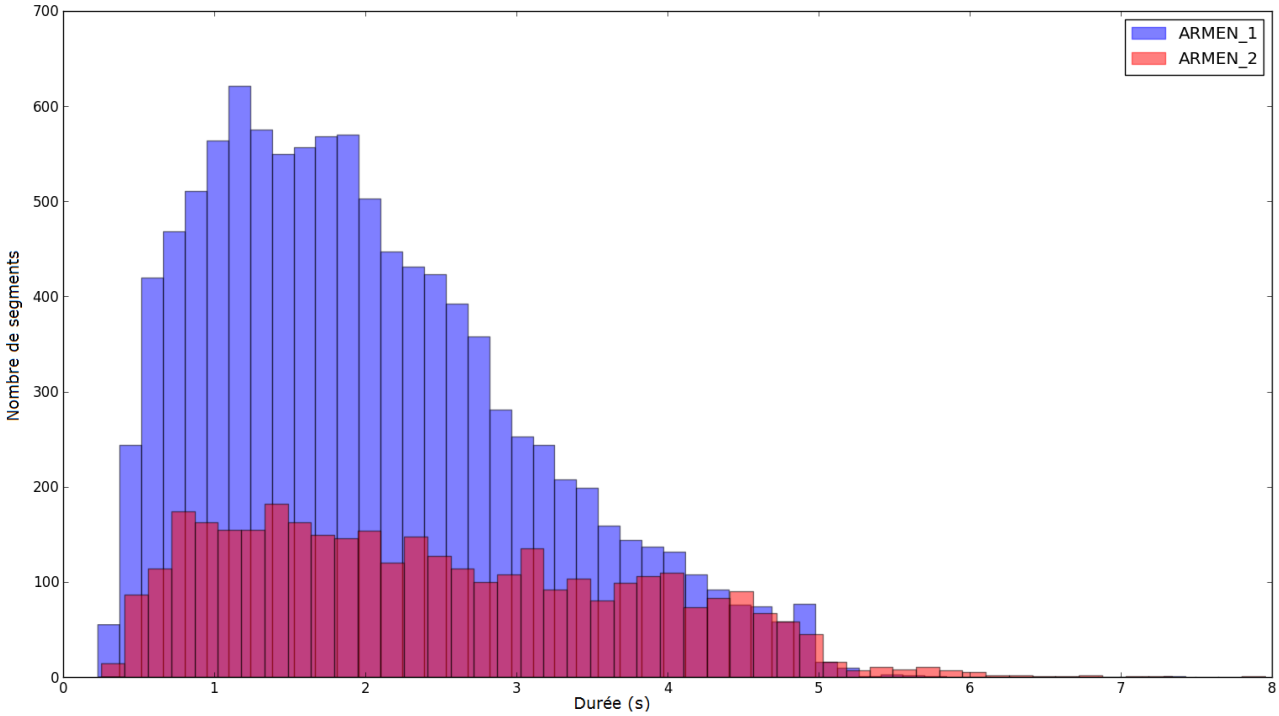

Figure 5.6 – Histogramme de la durée des segments (en secondes) pour les corpus ARMEN_1 et ARMEN_2 (segments des sujets uniquement).

5.3.2 Annotation

L'étape d'annotation est cruciale pour construire un corpus de données d'apprentissage. Nous avons pour objectif un prototype avec une base d'utilisateurs large et un besoin de robustesse. Nous avons donc opté pour un schéma catégorielle simple. Uniquement les segments des sujets pendant la phase "scénario" ont été annotés car ce sont les réactions des sujets en interaction avec le système de collecte, simulant le système final, qui nous intéressent.

5.3.2.1 Protocole

L'annotation a utilisé un schéma d'annotation relativement simple : cinq catégories émotionnelles ont été choisies (ainsi qu'une catégorie "Poubelle") et l'activation a été représentée par une échelle de Likert à cinq niveaux (de -2 à +2). Chaque annotateur a

ensuite étiqueté tous les segments des sujets pendant la phase "scénario" dans un ordre aléatoire à l'aide d'un outil d'annotation développé au LIMSI, très simple d'utilisation et de configuration[3].

Le protocole et le schéma d'annotation était le même pour les deux corpus.

Champ d'annotation	Valeurs possibles
Étiquette émotion	Colère, Joie, Neutre, Peur, Tristesse, Poubelle
Échelle d'activation	De -2 à +2

Tableau 5.VII – Informations sur le schéma d'annotation pour les corpus ARMEN_1 et ARMEN_2

5.3.2.2 Évaluation de l'accord inter-annotateur

L'accord inter-annotateur a été évalué à la fois quantitativement et qualitativement à l'aide de graphiques, sur les annotations catégorielles d'ARMEN_1 et ARMEN_2. Les mesures choisies sont l'accord inter-annotateur brut et la mesure Kappa, car elle reste très utilisée dans le domaine malgré ses défauts. Les résultats sont présentés dans le tableau 5.VIII.

Les outils graphiques choisis permettant de mieux comprendre les différences d'annotation sont les matrices de confusion et les représentations des distributions d'annotation sous forme d'histogrammes. Ils révèlent les biais des annotateurs et les confusions, par exemple entre les classes Neutre, Peur et Tristesse dans le cas d'ARMEN_1.

Corpus / Mesure	ARMEN_1	ARMEN_2
Kappa	0.33	0.37
Accord brut	46%	63%

Tableau 5.VIII – Évaluation quantitative de l'accord inter-annotateur

[3]L'outil est développé en Java, utilise la bibliothèque JMF pour la lecture des fichiers audios. Les schémas d'annotation sont configurables en XML et les annotations sont également enregistrées au format XML.

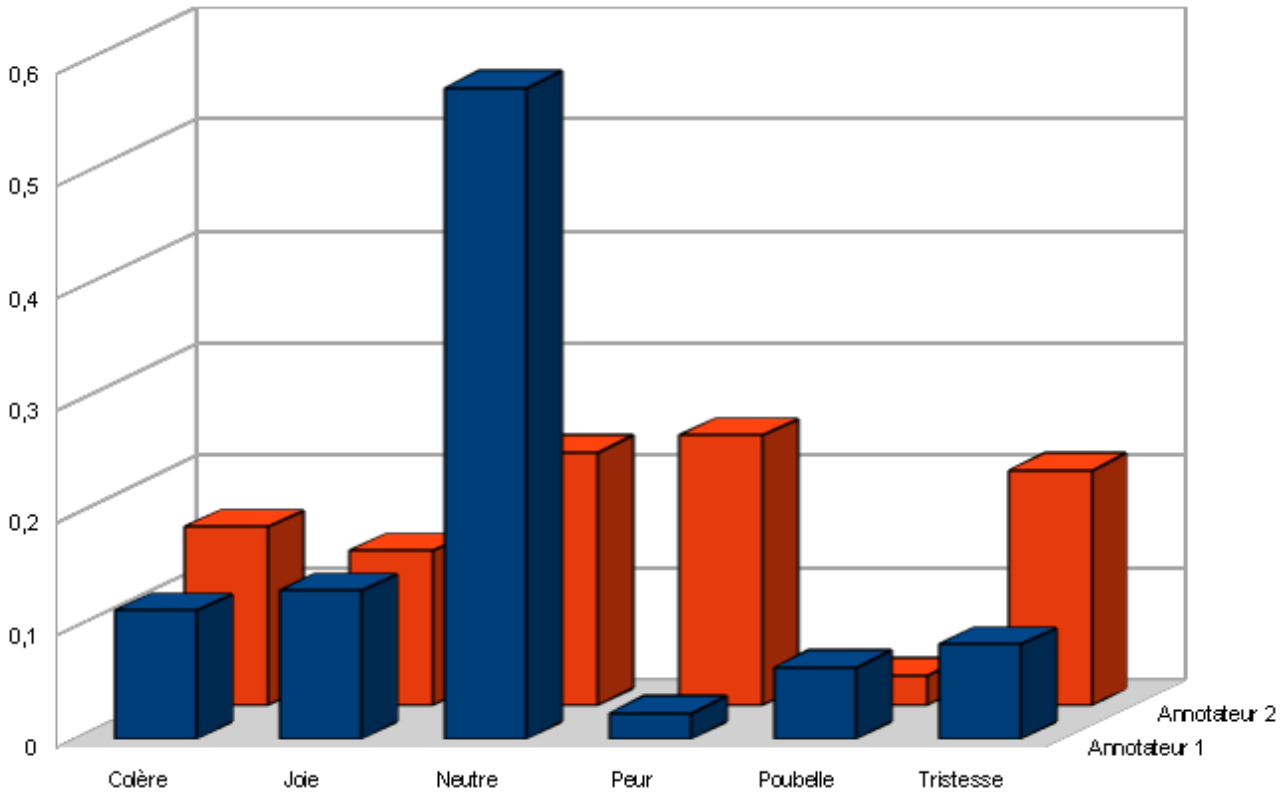

Figure 5.7 – Comparaison des distributions des étiquettes entre les deux annotateurs pour le corpus ARMEN_1.

5.3.2.3 Résumé de l'annotation

Le résultat de l'étape d'annotation est détaillé dans le tableau 5.IX ci-dessous. On peut voir que la collecte du corpus ARMEN_2 a donné lieu à une proportion beaucoup plus importante de segments annotés avec la classe Neutre. Il est cependant difficile de dire si cela est dû au protocole utilisé ou aux sujets.

5.4 Corpus finaux

Un récapitulatif de toute l'étape de collecte de données pour les corpus ARMEN_1 et ARMEN_2 est donné dans le tableau 5.X ci-dessous. On peut se rendre compte de

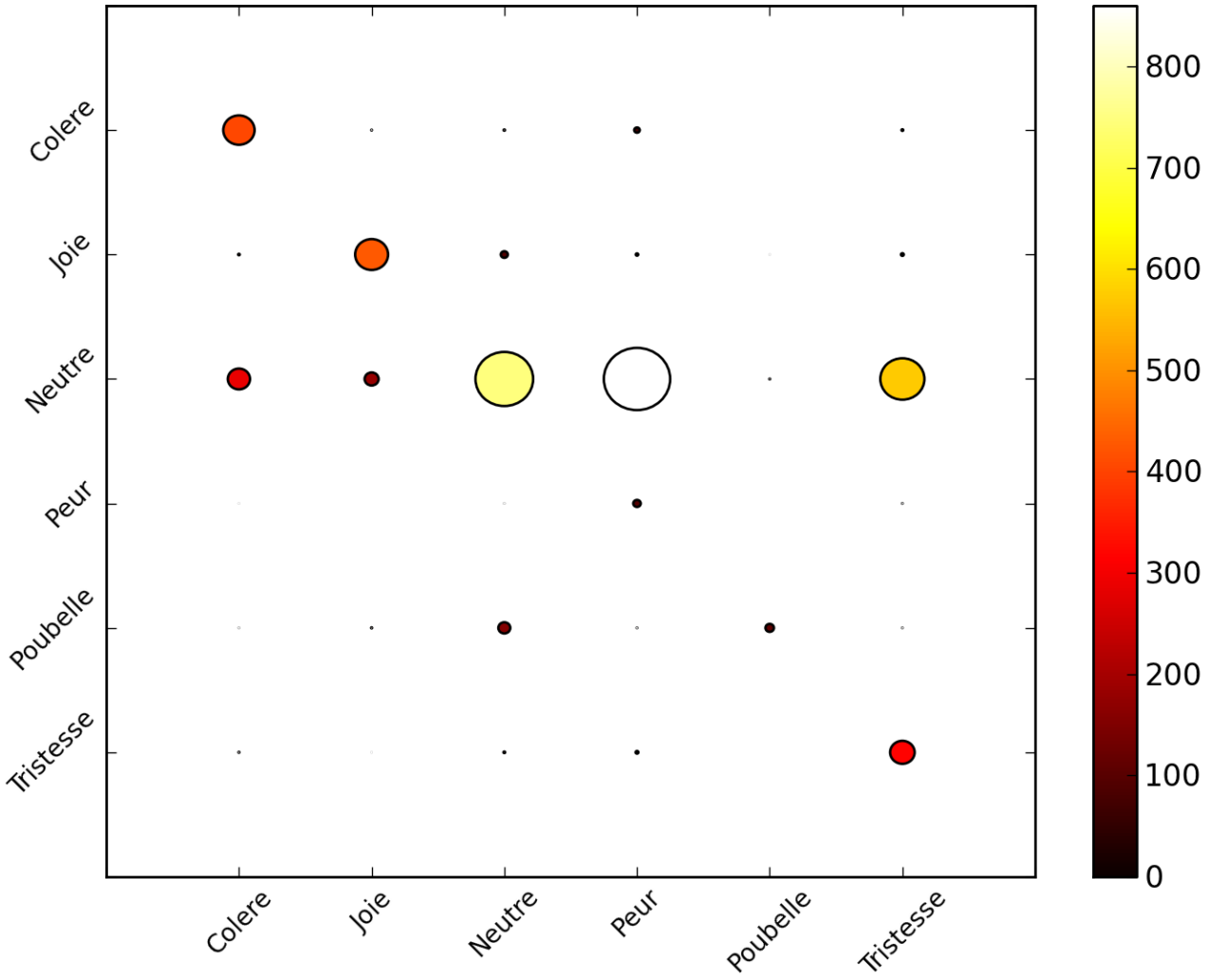

Figure 5.8 – Matrice de confusion des annotations pour le corpus ARMEN_1.

la difficulté de collecter des données émotionnelles spontanées, vu leur rareté : elles représentent environ 5% du total des enregistrements pour ARMEN_1 et 4% pour ARMEN_2.

5.5 Discussion

Les approches pour la segmentation et l'annotation des données exposées plus haut ont des limites. Les principaux problèmes sont le temps et le coût nécessaires à la réa-

	ARMEN_1	ARMEN_2
Nombre de segments annotés	4582	3581
Nombre de segments consensuels	2106	2309
Répartition (segments consensuels)		
Colère	406 (19%)	110 (5%)
Joie	427 (20%)	297 (13%)
Neutre	748 (36%)	1771 (77%)
Peur	97 (5%)	21 (1%)
"Poubelle"	110 (5%)	11 (<1%)
Tristesse	318 (15%)	99 (4%)

Tableau 5.IX – Résumé de l'étape d'annotation pour ARMEN_1 et ARMEN_2.

	ARMEN_1	ARMEN_2
Nombre de sujets	52	25
Répartition homme/femme	35 / 17	13 / 12
Date de collecte	juin 2010	juin 2011
Lieux	3 centres de Montpellier, France (Propara, APIGHREM et Malbosc)	2 centres de Montpellier, France (Propara et Malbosc)
Protocole	Woz avec une voix robotique, sept courts scénarios plus questionnaire	Woz avec AVE, quatre scénarios développés plus questionnaire
Type d'enregistrement	Audio et vidéo	Audio et vidéo
Durée d'enregistrement	18h02m	8h40m
Durée des segments "sujets"	5h54m59s	2h27m16s
Durée des segments annotés ("sujets" en phase "scénario")	2h46m	1h53m
Durée totale des segments annotés consensuellement (hors classe "Poubelle")	1h16m	1h10m
Durée totale des segments consensuels émotionnels (hors classe "Neutre")	49m	21m

Tableau 5.X – Récapitulatif pour les corpus ARMEN_1 et ARMEN_2.

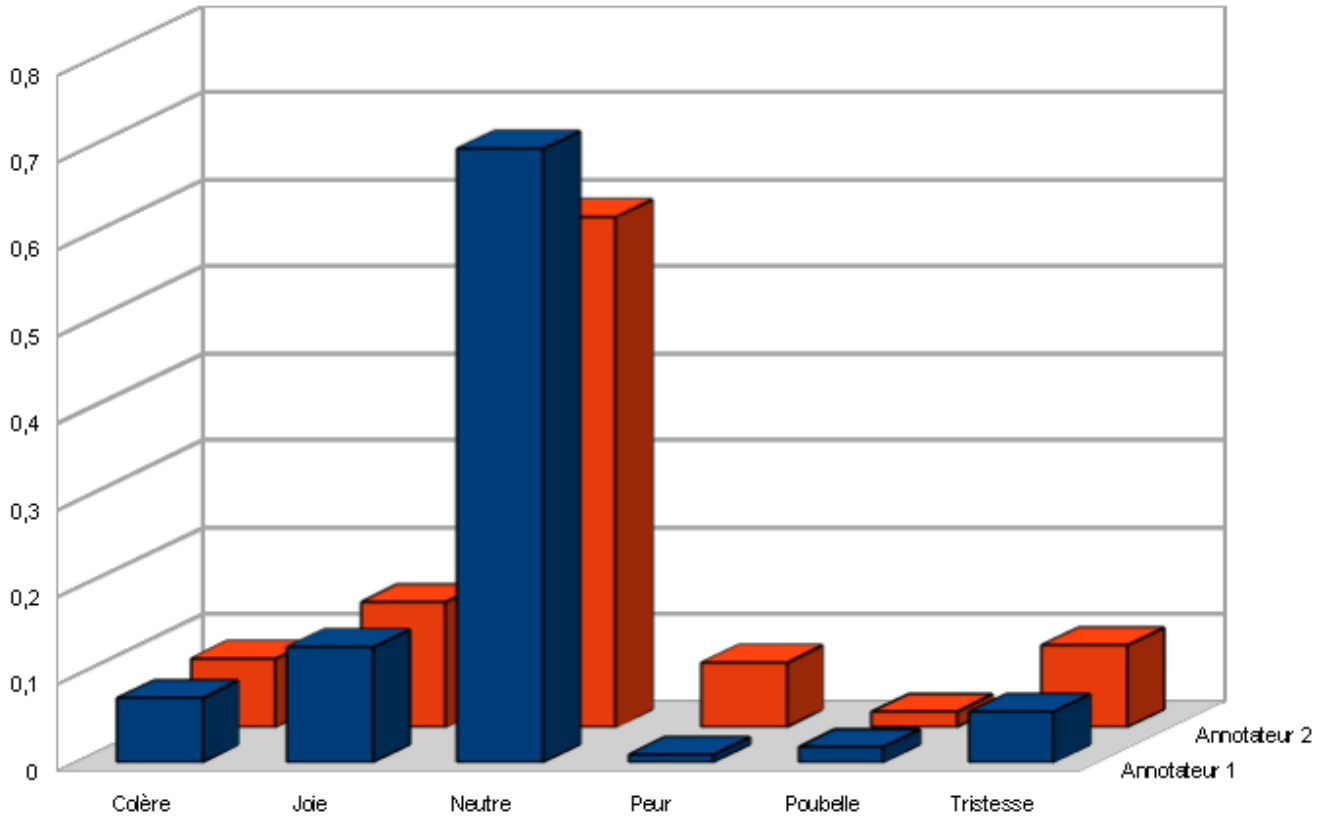

Figure 5.9 – Comparaison des distributions des étiquettes entre les deux annotateurs pour le corpus ARMEN_2.

lisation de ces étapes en employant des annotateurs "experts". Concernant le schéma d'annotation, on peut critiquer le choix d'un schéma catégorielle avec peu d'étiquettes émotionnelles, conduisant à une grande proportion de la classe "Neutre".

D'autres approches sont possibles pour obtenir une segmentation et une annotation satisfaisantes. Récemment, l'utilisation du *crowdsourcing* pour ce type de tâche a en effet pris de l'ampleur. Il s'agit de découper le travail d'annotation ou de segmentation en micro-tâches rémunérées et déléguées à un grand nombre de personnes. Cette méthode a déjà été explorée dans divers domaines grands consommateurs d'études utilisateur comme le marketing ou le test d'ergonomie de sites internet [143], utilisant dans la grande majorité des cas la plate-forme *Mechanical Turk* d'Amazon. Le *crowdsourcing*

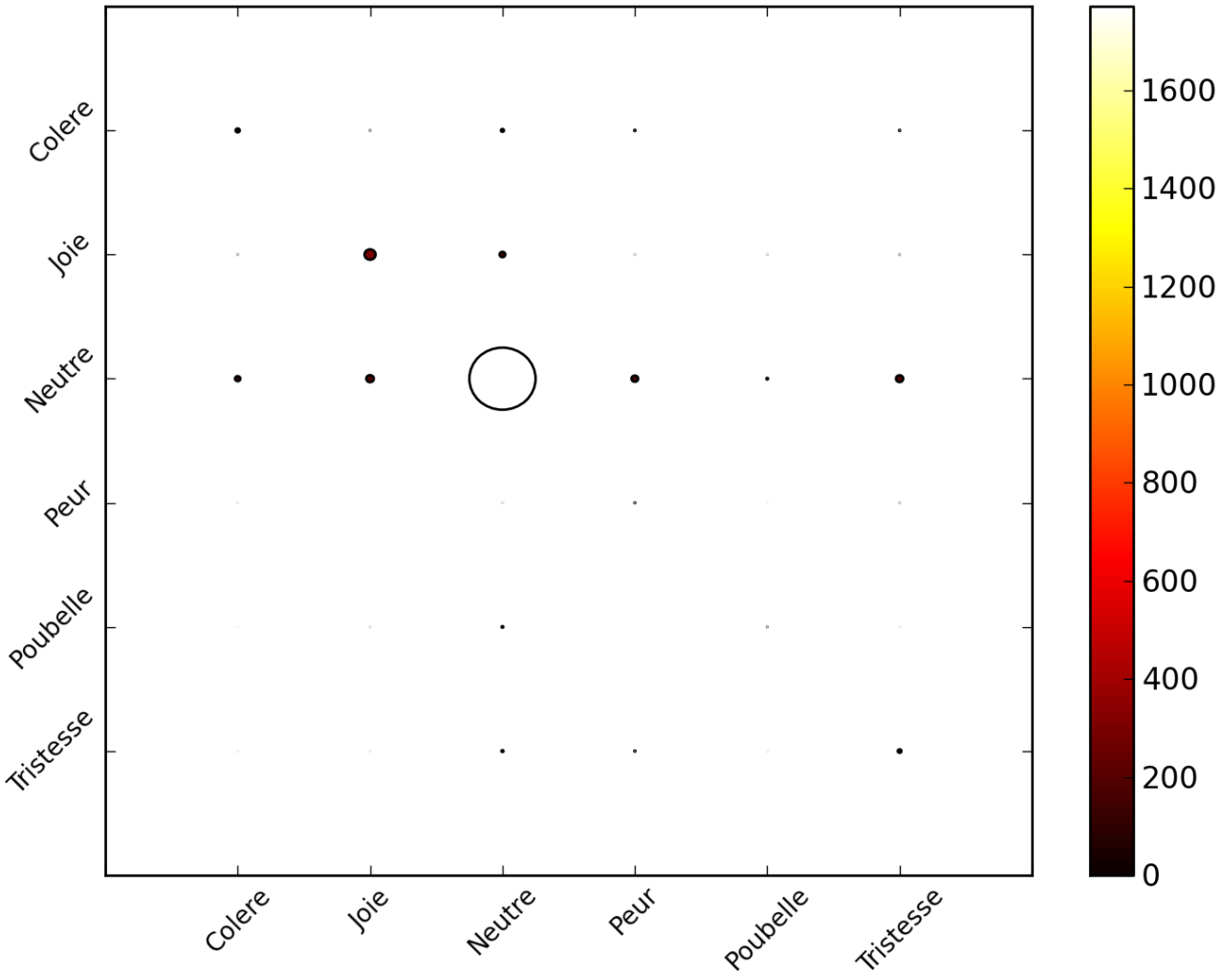

Figure 5.10 – Matrice de confusion des annotations pour le corpus ARMEN_2.

paraît compétitif à la fois en termes économiques et de temps, et il a de plus été suggéré qu'en assez grande quantité, la qualité des annotations est proche de celle obtenue avec des annotateurs "experts" ou par des méthodes traditionnelles (études en laboratoire) [4, 36]. Des travaux récents ont montré la complémentarité du *crowdsourcing* avec le domaine de l'*affective computing* [185] ; il a par exemple été appliqué pour l'annotation de corpus de voix émotionnelle [243, 253] ou de vidéo contenant de l'ennui [244].

On peut également citer les récentes avancées en apprentissage semi-supervisé, dans

lequel seule une partie des données d'apprentissage est annotée par des humains, l'autre partie étant annotée automatiquement par comparaison [279]. Une approche de segmentation automatique aurait également pu être mise en place, comme c'est quelquefois le cas sur des données trop importantes pour être traitées à la main. Cependant nos données sont encore d'une taille raisonnable.

CHAPITRE 6

DÉTECTION DES ÉMOTIONS EN CROSS-CORPUS

Ce chapitre étudie l'intérêt des approches cross-corpus dans l'apprentissage de systèmes automatiques de reconnaissance des émotions. Il s'appuie à la fois sur les données collectées au cours du projet ARMEN et présentées dans le chapitre précédent et sur d'autres données bien adaptées pour étudier ces approches.

6.1 État de l'art

6.1.1 Motivation

L'approche cross-corpus désigne le fait d'utiliser plusieurs corpus dans un contexte d'apprentissage. Elle naît vraisemblablement de la rareté des données disponibles due aux coûts de collecte et d'annotation. De plus, elle permet de vérifier le pouvoir de généralisation des modèles entraînés, c'est-à-dire leur capacité à donner de bonnes performances sur des données non précédemment vues, mais ne provenant pas du même contexte (différences de conditions acoustiques, de locuteurs, de contenu linguistiques).

6.1.2 Difficultés

Plusieurs obstacles apparaissent lors de la mise en oeuvre d'une telle approche, notamment :

La cohérence des annotations entre corpus : en effet vu le manque d'un cadre théorique et de définitions claires pour les émotions, la plupart des corpus sont annotés avec une tâche en tête, résultant en des schémas d'annotation parfois très différents (granularité, annotations catégorielles ou dimensionnelles). Il faut donc étudier les similarités entre les différents schémas et si nécessaire procéder à une adaptation.

L'adaptation des schémas d'annotation : il s'agit de trouver des concepts communs entre les corpus. Cette étape est très empirique pour l'instant et dépendante des

données utilisées, mais elle implique généralement une simplification des schémas pour aller vers un dénominateur commun. Par exemple, pour faire correspondre deux schémas catégoriels avec des granularités différentes, il est possible de rassembler les catégories plus fines du schéma plus détaillé en "macro-classes" correspondant au schéma plus grossier. Dans le cas d'un schéma catégoriel et d'un schéma dimensionnel, il est possible de quantifier empiriquement les dimensions d'après les distributions des dimensions et de les faire correspondre aux catégories. Certaines de ces approches ont été mises en oeuvre dans les travaux menés dans cette thèse [48, 78] ainsi que dans les travaux d'autres auteurs [96, 239]. Des correspondances hybrides entre schéma dimensionnel et schéma catégoriel peuvent également être conçues, sur le modèle de l'outil Feeltrace [61].

La différence des contextes : notamment acoustiques, différences entre groupes de locuteurs, différences des contenus linguistiques, des langues, des tâches menées dans les corpus... La différence de contextes acoustiques peut être très large, par exemple entre des données enregistrées dans une salle sourde ou très réverbérante, entre un micro de bonne qualité et le combiné d'un téléphone, entre des conditions de type laboratoire (avec très peu de bruit ambiant) et des enregistrements effectués dans la rue. Ces différences commencent à être étudiées dans le cadre de la reconnaissance des émotions [97, 252, 271].

6.1.3 Stratégies

Différentes stratégies cross-corpus ont été décrites et testées dans la littérature [159, 234, 238]. Le *pooling*, ou mise en commun des corpus, est la plus simple : tous les corpus sont rassemblés pour créer un plus gros corpus. Celui peut ensuite être découpé en ensemble d'apprentissage et de test par exemple ou pour entraîner un modèle en validation croisée. La stratégie dite de vote majoritaire (majority voting) se rapproche plus d'une stratégie de fusion multi-classifieurs : un modèle est entraîné par corpus ; lors de la phase de classification, chaque modèle donne une prédiction pour l'instance considérée et on procède à un vote majoritaire sur toutes les prédictions. Enfin, dans la

classification deux-à-deux (pairwise cross-corpus classification), pour chaque paire de corpus possible, un modèle est entraîné sur le premier corpus et testé sur le second ; il peut alors être intéressant de comparer, pour un corpus donné, la performance "intra" en validation croisée à la performance moyenne sur ce corpus en classification deux-à-deux avec des modèles entraînés sur les autres corpus. Une illustration de ces stratégies est proposée sur la figure 6.1 plus bas.

Certains travaux suggèrent qu'il est plus avantageux d'utiliser le pooling en termes de performances et donc de pouvoir de généralisation [238], suggérant que le principal problème à surmonter dans le domaine de la reconnaissance automatique des émotions est le manque de données et confirmant ainsi l'adage bien connu : "There is no data like more data". Cependant cette conclusion dépend de l'algorithme d'apprentissage choisi (dans ce cas les SVM) et n'est donc pas immédiatement généralisable. Une stratégie également décrite dans ces travaux pour surmonter cette difficulté et lorsqu'on ne sait pas quel est l'algorithme d'apprentissage optimal est celle du "vote à deux étages" (two-stage voting), où un système est construit par algorithme et on procède à un vote des systèmes, s'inspirant donc très fortement d'une technique de fusion tardive.

Au delà de l'amélioration des performances et de l'obtention espérée de modèles plus robustes, les approches cross-corpus peuvent aussi être appliquées pour d'autres buts : vérifier la validité d'un choix de conception (par exemple le fait d'avoir des modèles spécifiques à différents groupes de locuteurs) ou évaluer la pertinence d'une sélection ou d'un classement de paramètres sur plusieurs corpus [237, 251].

6.2 Expériences menées

Les expériences menées sur l'approche cross-corpus l'ont été au début de cette thèse. Une première expérience s'est intéressée uniquement à l'approche "deux-à-deux" avec les corpus EmoVox et CEMO, en utilisant trois classes [78]. Une seconde expérience a comparé les approches "deux-à-deux" et pooling aux scores intra avec trois corpus (CEMO, EmoVox et Bourse) [48] ; une mesure de similarité entre corpus a également été proposée, s'inspirant de travaux précédents [32]. Ces expériences suggèrent un apport

intéressant des approches cross-corpus.

6.2.1 Présentation des corpus

En plus des corpus ARMEN, trois autres corpus ont été utilisés lors de ces expériences. Ils ont été collectés par l'équipe du thème "Reconnaissance d'émotions" du LIMSI en collaboration avec plusieurs organismes et entreprises. Leur particularité et intérêt réside dans le fait qu'ils proviennent tous de centres d'appels (call-centers) et que leur taille est importante (plusieurs dizaines d'heures). Comme ils contiennent des informations personnelles, ils ne sont pas disponibles pour la communauté. Les centres d'appels sont une source de données intéressante car abondante, malgré la faible qualité audio des enregistrements (qualité téléphonique). Pour cette raison, ils sont assez étudiés dans la communauté, notamment pour de la détection de stress [158].

Le premier, CEMO, contient des enregistrements d'un centre d'appel d'urgences médicales obtenus grâce à une convention entre l'Assistance Publique - Hôpitaux de Paris (APHP) et le LIMSI [80]. Le but de ce service est d'offrir des conseils médicaux et dans le cas d'urgences, d'évaluer le degré de gravité de la situation afin d'envoyer une équipe d'intervention le cas échéant. Les appelants, qui peuvent être la personne malade ou un tiers (membre de la famille, ami, collègue), expriment très souvent du stress, de la douleur, de la peur et parfois un véritable panique. Le corpus annoté contient environ 20 heures de données audio téléphoniques provenant 874 locuteurs uniques (7 agents et 867 appelants). Il y a bien entendu une asymétrie dans la répartition des temps de parole par locuteurs, avec une sur-représentation des agents. Le schéma d'annotation de CEMO est assez sophistiqué, il comprend 21 étiquettes émotionnelles fines regroupées en 7 catégories plus grossières (voir tableau 6.I). Chaque segment audio a été annoté par deux annotateurs, qui choisissaient une étiquette émotionnelle "majeure" et une "mineure", résultant en un *soft vector* d'annotation [80], intéressant pour exprimer les émotions complexes comme un mélange de plusieurs émotions plus simples. Il s'agit d'un corpus riche en émotions spontanées, qui a donné lieu à plusieurs études [79, 81, 263].

Le deuxième, EmoVox, est un sous-ensemble du corpus CallSurf, riche d'environ 1 000 heures de données audio téléphoniques correspondant à plus de 10 000 appels

Macro-classe	Étiquettes fines
Fear	Fear, Anxiety, Stress, Panic, Embarrassment, Dismay
Anger	Annoyance, Impatience, HotAnger, ColdAnger
Sadness	Disappointment, Sadness, Despair, Resignation
Positive	Interest, Compassion, Amusement, Relief
Pain	Pain
Surprise	Surprise
Neutral	Neutral

Tableau 6.I – Regroupements des étiquettes émotionnelles fines en macro-classes pour le corpus CEMO.

[109]. Il a été collecté dans l'un des centres d'appel et de réclamation d'EDF. Sur les 150 heures d'audio transcrites, 15 heures ont été sélectionnées pour annotation, pour leur contenu émotionnel ; ces 15 heures correspondent à 77 appels. EmoVox est composé de 243 extraits de ces appels, pour une durée totale de 2 heures et 42 minutes. L'annotation a utilisé un schéma complexe à niveaux multiples, dont une partie catégorielle avec 16 étiquettes fines regroupées en 5 macro-classes (voir tableau 6.II) ; des *soft vector* d'annotation sont utilisés de manière similaire au corpus CEMO. Une étude explorant l'annotation automatique par *bootstrap* a utilisé ce corpus [259].

Enfin, le corpus Bourse est issu de dialogues agents-clients provenant du service client d'un site de trading en ligne [80]. Le corpus est constitué de 100 dialogues en français, transcrits et annotés. Il a permis de mener les premiers travaux sur les indices

Macro-classe	Étiquettes fines
Positif	Satisfaction, Soulagement, Joie, Excitation, Surprise positive, Amusement, Espoir
Colère	Irritation, Surprise négative
Peur	Doute, Inquiétude, Stress
Tristesse	Déception, Désespoir, Impuissance
Neutre	Neutre

Tableau 6.II – Regroupements des étiquettes émotionnelles fines en macro-classes pour le corpus EmoVox (d'après [258]).

prosodiques et linguistiques pour la détection des émotions [76].

Ces corpus sont uniques car ils partagent un mode de communication commun (centre d'appel téléphonique), mais dans des domaines différents. En cela, ils sont donc particulièrement intéressants pour étudier des approches cross-corpus. Il faut également noter qu'ils partagent tous une asymétrie dans la répartition du temps de parole des locuteurs, puisque les agents du centre d'appel représentent un petit nombre de locuteurs avec une forte proportion du temps de parole (généralement supérieur à 50%) alors que le reste est réparti entre un grand nombre d'appelants (désignés par le terme "clients" dans la suite).

6.2.2 Expériences et résultats

6.2.2.1 Première expérience

La première expérience a porté sur l'application de l'approche deux-à-deux avec deux corpus (EmoVox et CEMO) [78]. Trois "macro-classes" émotionnelles ont été considérées (colère, positif, neutre).

Une étape préalable à l'établissement des macro-classes a été l'étude des schémas d'annotation des deux corpus et la distribution des étiquettes dans chaque cas. Pour CEMO, seuls les segments pour lesquels les deux annotateurs s'accordaient sur la macro-classe à la fois de l'émotion majeure et mineure ont été gardés. Par exemple, un segment annoté Interest et Amusement pour l'émotion majeure et Neutral et Neutral pour l'émotion mineure était gardé, alors qu'un segment annoté ColdAnger et Sadness pour l'émotion majeure et Anxiety et Neutral pour l'émotion mineure ne l'était pas. La macro-classe de l'émotion majeure a alors été utilisée pour créer les sets Anger1, Positive et Neutral. Un ensemble plus large dénommé Anger2 regroupant à la fois les segments du ensemble Anger1 ainsi que d'autres segments annotés comme des mélanges de colère avec d'autres émotions négatives provenant des macro-classes Fear (Fear, Anxiety, Stress, Panic, Embarrassment, Dismay) et Sadness (Sadness, Disappointment, Resignation, Despair) a également été créé. Au total près de 17 000 segments ont été gardés ; leur répartition entre les macro-classes est indiquée dans le tableau 6.III ci-dessous.

La même procédure a été appliquée au corpus EmoVox pour que les macro-classes

Macro-classe	Agents	Clients	Total agents + clients
Anger1	523	197	720
Anger2	600	482	1082
Positive	1101	54	1155
Neutral	7542	5564	13106
Total	9766	6297	16603

Tableau 6.III – Répartition des segments entre les macro-classes pour le corpus CEMO.

soient adaptées avec CEMO : l'ensemble Anger1 contient les segments pour lesquels les annotateurs s'accordaient sur l'émotion majeure et l'annotaient en utilisant la macro-classes "Colère" uniquement. L'ensemble Anger2 contient les segments pour lesquels les annotateurs ont utilisé les macro-classes à valence négative "Colère", "Peur" et "Tristesse" ; il contient donc Anger1 et est plus varié au niveau du contenu. Près de 2 000 segments ont été retenus avec ces critères ; leur répartition entre les macro-classes est indiquée dans le tableau 6.IV ci-dessous.

Quatre ensembles d'apprentissage ont été construits à partir de ces données, correspondant aux deux corpus avec deux conditions : Anger1 et Anger2. Les données ont été sous-échantillonnées pour obtenir des corpus d'apprentissage équilibrés : pour le corpus d'apprentissage "CEMO - Anger1", la macro-classe contenant le moins de segments est Anger1 avec 720 segments, les macro-classes ont donc été sous-échantillonnées à 720 segments chacune. La taille des ensembles d'apprentissage est donc de 2 160 segments (720×3) pour le corpus "CEMO - Anger1", de 3 246 segments pour "CEMO - Anger2", et de 1 026 segments pour "EmoVox - Anger1" et "EmoVox - Anger2".

Classes	Agents	Clients	Total agents + clients
Anger1	38	348	386
Anger2	120	493	613
Positive	120	276	396
Neutral	167	175	342
Total	445	1292	1737

Tableau 6.IV – Répartition des segments entre les macro-classes pour le corpus EmoVox.

La bibliothèque openEAR [98] a été utilisée pour extraire deux ensembles de paramètres différents pour représenter les données : un ensemble de 384 paramètres, correspondant à la configuration du challenge InterSpeech 2009 [229], et un ensemble de 988 paramètres, correspondant à la configuration de base. Des SVM à noyau à base radiale ont été entraînés en validation croisée avec la bibliothèque libSVM [44], en utilisant une procédure de type *grid search* pour l'optimisation des hyper-paramètres.

Les résultats de cette exploration de la stratégie pairwise sont donnés dans le tableau 6.V ci-dessous. Comme on pouvait s'y attendre, ils sont assez faibles, reproduisant des résultats d'autres études [159] ; ils restent cependant significativement au-dessus du niveau de chance. Le fait que dans la plupart des cas quand CEMO est utilisé en apprentissage et EmoVox en test, très peu voire aucune instance n'est classifiée Positive tend à suggérer que les deux corpus sont très différents à cet égard. En examinant les répartitions des segments par rôle (agent/client) dans les tableaux 6.III et 6.IV, on peut d'ailleurs voir qu'il y a très peu de segments positifs chez les clients dans CEMO alors que la répartition est plus équilibrée pour EmoVox. On peut remarquer que le fait d'utiliser beaucoup de paramètres n'apporte pas vraiment de gain de performance dans ce cas, la moyenne du taux de reconnaissance (RR) étant de 43.6% pour 988 paramètres et de 43.2% pour 384 paramètres (le niveau du hasard est de 33.3% avec trois classes équilibrées).

En conclusion, on peut dire qu'il est possible de généraliser des classes émotionnelles apprises sur un corpus à un autre, mais que les facteurs influençant les performances sont difficiles à comprendre, même si on peut en suggérer quelques uns : nature des tâches, différences au niveau des schémas d'annotation, taille de corpus variable...

6.2.2.2 Deuxième expérience

La deuxième expérience a porté sur la comparaison des approches "deux-à-deux" et pooling par rapport aux scores "intra-corpus", c'est-à-dire la performance en utilisant un ensemble d'apprentissage et un ensemble de test, tous deux issus du même corpus. Trois corpus (CEMO, EmoVox et Bourse) ont été utilisés pour cette expérience, sur deux classes émotionnelles (colère et neutre) [48].

Ensembles d'apprentissage/test	RR (hasard : 33.3%)	F-scores par classe (Anger/Positive/Neutral)
CEMO Ang1 / EmoVox Ang1	47.9%	0.59 / 0.01 / 0.55
	43.9%	0.49 / 0 / 0.55
CEMO Ang1 / EmoVox Ang2	46.7%	0.58 / 0.06 / 0.55
	43.1%	0.47 / 0 / 0.55
CEMO Ang2 / EmoVox Ang1	43.2%	0.50 / 0 / 0.53
	47,2%	0.59 / 0.03 / 0.54
CEMO Ang2 / EmoVox Ang2	43.1%	0.50 / 0 /0.53
	45.2%	0.55 / 0.03 / 0.52
EmoVox Ang1 / CEMO Ang1	41.8%	0.34 / 0.44 / 0.47
	42.9%	0.27 / 0.44 / 0.55
EmoVox Ang1 / CEMO Ang 2	41.1%	0.30 / 0.45 / 0.47
	42.8%	0.25 / 0.44 / 0.56
EmoVox Ang2 / CEMO Ang1	41.4%	0.35 / 0.44 / 0.46
	42.6%	0.29 / 0.43 / 0.55
EmoVox Ang2 / CEMO Ang2	40.7%	0.32 / 0.44 / 0.47
	40.7%	0.25 / 0.43 / 0.52

Tableau 6.V – Résumé des résultats pour la première expérience cross-corpus. Pour chaque paire corpus d'apprentissage/de test, la première ligne donne les performances pour l'ensemble de 384 paramètres, la deuxième pour 988 paramètres.

De manière similaire à la première expérience décrite plus haut, les schémas d'annotation des trois corpus ont été adaptés pour n'utiliser que deux macro-classes, Colère et Neutre. Nous n'avons utilisé que des segments issus de clients, ne tenant pas compte de ceux issus des agents. Le même nombre de segments "clients" a alors été utilisé pour chaque émotion dans chaque corpus. Un ensemble d'apprentissage et de test a ensuite été construit pour chaque corpus avec des proportions 90%/10%, en veillant ce que les locuteurs présents en apprentissage ne le soient pas en test et en équilibrant les classes. Dans la condition expérimentale "Intra", un classifieur était entraîné sur l'ensemble d'apprentissage pour chaque corpus et testé sur l'ensemble de test issu du même corpus. Dans la condition "Deux-à-deux", un classifieur était entraîné sur l'ensemble d'apprentissage pour chaque corpus et testé sur les ensembles de test issus des deux autres corpus. Dans la condition "Pooling", les ensembles d'apprentissage de deux ou trois corpus était agrégés pour entraîner un classifieur, qui était ensuite testé sur tous les ensembles de test. Les trois conditions sont illustrées sur la figure 6.1 ci-dessous ; le détail des ensembles

d'apprentissage et de test est donné dans le tableau 6.VI.

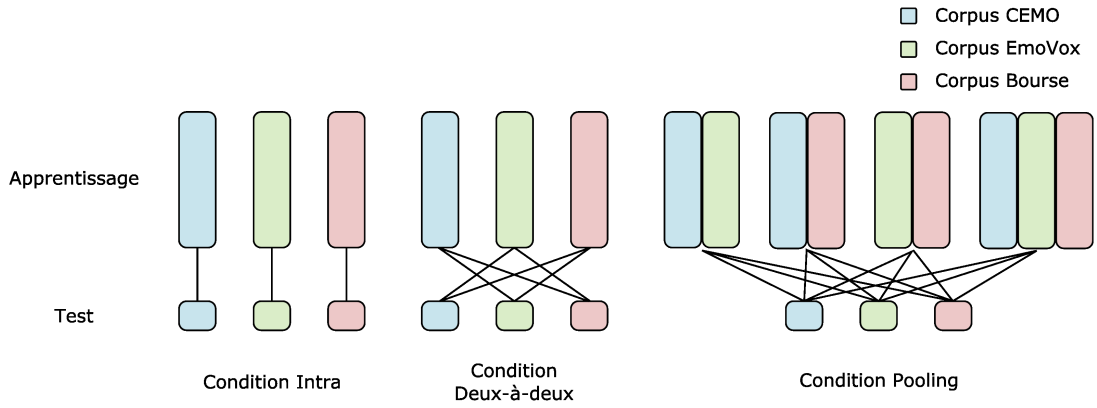

Figure 6.1 – Conditions expérimentales pour la deuxième expérience cross-corpus.

De manière similaire à la première expérience, 384 paramètres acoustiques ont été extraits pour représenter les données et des SVM ont été entraînés en cross-validation, avec optimisation des hyper-paramètres à l'aide d'une procédure grid search. Les performances brutes (le taux de reconnaissance a été utilisé) sont listées dans le tableau 6.VII, ainsi que le nombre de vecteurs supports en proportion des instances d'apprentissage, qui donne une estimation immédiate de la "difficulté" d'un corpus [196]. De ce point on peut voir que Bourse est un corpus plus difficile que CEMO ou EmoVox, cependant il faudrait contrôler la taille du corpus, qui peut jouer. On observe également que la stratégie Deux-à-deux donne ici des résultats très encourageants, avec 79.4% et 70.6% de reconnaissance sur EmoVox, à comparer avec la performance Intra, de 73.5%, ou encore 76.0% et 75.0% de reconnaissance sur CEMO (Intra 77.1%). Ce n'est pas le cas sur le corpus Bourse, avec 54.4% et 50.0% de reconnaissance, soit tout juste le niveau de chance (Intra 63.0%). Les résultats sont également très intéressants avec la stratégie Pooling, puisqu'on obtient des meilleurs sur EmoVox dans tous les cas (jusqu'à 85.3% par rapport à 73.5% en Intra), et dans trois cas sur quatre pour CEMO (jusqu'à 80.2% par rapport à 77.1% en Intra). Sur Bourse, la stratégie Pooling n'atteint pas la performance de la stratégie Intra, mais les résultats sont tout de même meilleurs qu'avec la stratégie Deux-à-deux.

Nom de l'ensemble	Nombre de segments	Nombre de locuteurs
Bourse Test	46	4
CEMO Test	96	22
EmoVox Test	34	5
Bourse Intra Apprentissage	422	56
CEMO Intra Apprentissage	868	264
EmoVox Intra Apprentissage	314	50
Bourse-CEMO Pooling Apprentissage	844	320
Bourse-EmoVox Pooling Apprentissage	628	314
CEMO-EmoVox Pooling Apprentissage	628	106
Bourse-CEMO-Emovox Pooling Apprentissage	942	370

Tableau 6.VI – Détails des ensembles d'apprentissage et de test pour les trois conditions expérimentales.

Il est également très intéressant de noter que la stratégie Pooling permet de diminuer le nombre de vecteurs supports utilisé pour les mélanges Bourse-CEMO et Bourse-EmoVox par rapport aux corpus individuels, suggérant une véritable complémentarité de ces corpus et qu'une partie redondante de l'information est éliminée.

Le fait que le corpus Bourse soit plus difficile est peut-être dû à son contenu : il s'agit d'un centre d'appel pour des services financiers en ligne et les clients expriment rarement de la colère franche et plus fréquemment de l'inquiétude. Ce n'est pas le cas pour le corpus CEMO d'urgences médicales dans lequel les appelants crient au téléphone (souvent parce qu'ils sont très stressés ou parce qu'ils se disputent avec quelqu'un d'autre au sujet de l'état de santé d'une personne blessée) ni pour le corpus EmoVox, dont les enregistrements sont issus d'un centre de réclamation.

La performance de la stratégie Pooling est analysée plus en détail dans le tableau 6.VIII : pour chaque ensemble d'apprentissage de cette condition, on compare la performance sur les trois ensembles de tests avec les corpus individuels composant l'ensemble. Par exemple, l'ensemble Bourse-CEMO Pooling est comparé à la performance en apprenant sur Bourse d'une part et CEMO d'autre part (donc comparaison avec les stratégie Intra et Deux-à-deux simultanément). À l'exception du corpus Bourse, la stratégie Pooling montre donc une amélioration des performances dans la majorité des cas. Cela est probablement dû à l'augmentation des données disponibles pour l'apprentissage.

Nom de l'ensemble d'apprentissage	Nombre de vecteurs supports	RR sur Bourse Test	RR sur CEMO Test	RR sur EmoVox Test
Bourse Intra	85.6%	63.0%	76.0%	70.6%
CEMO Intra	71.2%	54.4%	77.1%	79.4%
EmoVox Intra	75.5%	50.0%	75.0%	73.5%
Bourse-CEMO Pooling	65.6%	52.2%	78.1%	85.3%
Bourse-EmoVox Pooling	65.1%	59.0%	75.0%	79.4%
CEMO-EmoVox Pooling	72.3%	52.2%	79.2%	82.4%
Bourse-CEMO-EmoVox Pooling	74.0%	58.7%	80.2%	82.4%

Tableau 6.VII – Résultats pour les trois conditions expérimentales.

Nom de l'ensemble d'apprentissage	Nom de l'ensemble de test	Comparaison avec Bourse Intra	Comparaison avec CEMO Intra	Comparaison avec EmoVox Intra
Bourse-CEMO Pooling	Bourse	-10.9	-2.2	N.A.
	CEMO	**+2.1**	**+1.1**	N.A.
	EmoVox	**+14.7**	**+5.9**	N.A.
Bourse-EmoVox Pooling	Bourse	-4.0	N.A.	**+9.0**
	CEMO	-1.0	N.A.	+0.0
	EmoVox	**+8.8**	N.A.	**+5.9**
CEMO-EmoVox Pooling	Bourse	N.A.	-2.2	**+2.2**
	CEMO	N.A.	**+2.1**	**+4.2**
	EmoVox	N.A.	**+2.9**	**+8.8**
Bourse-CEMO-EmoVox Pooling	Bourse	-4.3	**+4.4**	**+8.7**
	CEMO	**+4.2**	**+3.1**	**+5.2**
	EmoVox	**+11.8**	**+2.9**	**+8.8**

Tableau 6.VIII – Gain de performance (en points de pourcentage du RR) pour la stratégie Pooling par rapport aux conditions Deux-à-deux et Intra.

En observant que les corpus CEMO et EmoVox tendaient à se comporter de la même manière dans cette expérience tandis que le corpus Bourse semblait différent, nous avons proposé une mesure de similarité entre corpus. Sachant que la sélection de paramètres est assez dépendante du corpus d'apprentissage utilisé [159], nous avons voulu utiliser les meilleurs paramètres comme une "empreinte" du corpus, en s'inspirant de travaux menés précédemment [32]. D'autres approches pour mesurer la similarité entre corpus se fondent plutôt sur un clustering des exemples d'apprentissage [239] ou sur la construction d'un profil des émotions présentes dans le corpus [186].

La procédure suivie était la suivante : pour chaque ensemble d'apprentissage de la condition Intra, un classement des paramètres individuels a été obtenu très simplement en entraînant un classifieur pour chaque paramètre et en utilisant sa performance en termes de taux de reconnaissance comme critère de classement. Les 25 meilleurs paramètres pour chaque corpus ont ensuite été sélectionnés pour constituer son "empreinte". Nous évaluons la similarité entre les corpus en calculant la similarité de Jaccard des empreintes (voir équation 7.1 pour la définition). Cela permet d'avoir une mesure entre 0 et 1 qui se comporte plutôt correctement (symétrie, similarité pour deux corpus identiques égale à 1). Les résultats pour nos trois corpus sont présentés dans le tableau 6.IX, ainsi que la performance obtenue en entraînant un modèle avec ces 25 meilleurs paramètres. On voit que selon cette mesure de similarité, CEMO et EmoVox sont plus similaires que Bourse. On peut également remarquer que la performance en utilisant 25 paramètres est meilleure qu'en utilisant 384 paramètres pour Bourse et EmoVox.

Cette expérience a donc permis de montrer que l'approche cross-corpus est prometteuse pour surmonter les problèmes de rareté de données et pour augmenter la robustesse des classifieurs. En particulier, la stratégie Pooling paraît à mieux adaptée à ces objectifs.

	Bourse vs CEMO	Bourse vs EmoVox	CEMO vs EmoVox
Similarité de Jaccard	0.140	0.163	0.191
	Bourse	CEMO	EmoVox
RR avec les meilleurs paramètres	66.6%	68.4%	73.6%

Tableau 6.IX – Scores de similarité fondés sur un classement de paramètres.

6.2.2.3 Troisième expérience

Une troisième expérience a porté sur l'étude des corpus ARMEN_1 et ARMEN_2. Une évaluation de la stratégie Pooling par rapport aux scores Intra a été menée, ainsi qu'une évaluation de la stratégie Deux-à-deux sur les deux corpus divisés selon deux critères : âge des locuteurs et qualité vocale. Ces critères ont déjà été sujets de travaux, par exemple pour reconnaître les sujets âgés et déclencher un comportement spécial pour un système de navigation [189].

Tout d'abord, le corpus ARMEN_1+2 a été construit en concaténant ARMEN_1 et ARMEN_2. La classe Neutre étant trop prépondérante, le déséquilibre a été réduit en la sous-échantillonnant pour que son nombre d'instances soit égal à la moyenne du nombre d'instances des autres classes. La classe Peur a été supprimée car trop faible en nombre d'instances. L'ensemble ARMEN_1+2 a été divisé en deux paires de sous-ensembles selon deux critères : l'âge des locuteurs (plus ou moins de 60 ans) et la qualité vocale (normale ou dégradée), selon des informations fournies par des orthophonistes concernant la qualité vocale (volume faible, sauts de volume, timbre de voix altéré, dévoisement...), l'articulation et selon la présence de bruits parasites (respirateurs, valves...). L'idée est que l'âge et la qualité vocale ont une influence sur la voix et donc sur l'expression des émotions, nous voulons donc comprendre si l'expression émotionnelle est comparable selon les critères. Le détail sur la composition de tous ces ensembles est donné dans le tableau 6.X.

Des paramètres acoustiques (les 384 paramètres du challenge Interspeech 2009) ont ensuite été extraits des segments audio par la librairie openEAR. Une optimisation "grid

	ARMEN_1	ARMEN_2	ARMEN_1+2	Plus de 60 ans	Moins de 60 ans	Voix "normales"	Voix dégradées
Nombre de segments	1288	1588	2080	658	997	978	677
Nombre de locuteurs	52	25	77	31	37	33	35
Répartition des classes							
Colère	406 (20%)	92 (6%)	498 (24%)	108 (16%)	260 (26%)	247 (25%)	121 (18%)
Joie	427 (21%)	236 (15%)	663 (32%)	231 (35%)	309 (31%)	383 (39%)	157 (23%)
Neutre	748 (38%)	1158 (73%)	520 (25%)	164 (25%)	249 (25%)	244 (25%)	169 (25%)
Peur	97 (5%)	21 (1%)	0	0	0	0	0
Tristesse	318 (16%)	81 (5%)	399 (19%)	155 (24%)	179 (18%)	104 (11%)	230 (34%)

Tableau 6.X – Détails sur la composition des corpus.

search" à deux dimensions a été réalisée sur le paramètre de coût C et le paramètre Gamma d'un classifieur SVM avec un noyau à base radiale. Pour chaque couple de paramètres (C, Gamma), une évaluation Leave One Speaker Out a été réalisée, pour s'assurer que le classifieur n'apprenait pas les voix des locuteurs, ce qui est à prendre particulièrement en compte dans le cas de données avec des voix très spécifiques et très différentes. La moyenne non-pondérée des rappels par classe (*Unweighted Average Recall* - UAR) a été utilisée pour quantifier la performance du classifieur, vu le déséquilibre persistant entre les classes.

Les premiers résultats, donnés dans le tableau 6.XI, montrent que le système réalise une meilleure performance que les systèmes entraînés de manière spécifiques sur un corpus ou une catégorie d'âge ou de qualité vocale donnée. Notamment pour la reconnaissance de la Tristesse, le score est de 46.0%, à comparer à 35.0% et 8.6% pour ARMEN_1 et ARMEN_2 respectivement. Quelques remarques peuvent être faites : la Colère est beaucoup mieux reconnue pour les voix jeunes que pour les voix âgées, mais c'est le contraire pour la Joie. Concernant les voix normales, la Tristesse est moins bien reconnue que pour les voix dégradées. Une expérience cross-corpus a également été menée avec la stratégie Deux-à-deux pour les critères d'âge et de qualité vocale, dont les résultats sont disponibles dans le tableau 6.XII. En entraînant le classifieur sur les voix normales et en testant sur les voix dégradées, le rappel pour la classe Tristesse grimpe à 59.7% alors qu'il n'est que de 20.7% lorsque l'on fait le contraire. Cela suggère que les voix dégradées dans ce corpus expriment la classe Tristesse d'une manière différente des voix normales, mais qu'une partie est commune. Il faudrait cependant vérifier d'éventuels effets de différence de taille de données d'apprentissage pour pouvoir conclure.

6.3 Conclusion

Les expériences menées autour des approches cross-corpus sont donc très encourageantes pour surmonter le problème de la rareté des données qui handicape le domaine de l'*affective computing* depuis longtemps. Les résultats obtenus suggèrent aussi qu'une robustesse accrue passera par une diminution substantielle des performances reportées

Ensemble considéré	Score moyen	Colère	Joie	Neutre	Tristesse
ARMEN_1	45.5%	56.5%	45.5%	45.0%	35.0%
ARMEN_2	35.5%	26.1%	68.2%	39.0%	8.6%
ARMEN_1+2	47.0%	49.8%	46.1%	46.0%	46.0%
Moins de 60 ans	44.8%	52.9%	39.3%	46.1%	40.9%
Plus de 60 ans	43.7%	38.9%	47.2%	45.4%	43.3%
Voix "normales"	44.1%	48.8%	50.3%	44.1%	33.3%
Voix dégradées	44.1%	38.8%	42.2%	47.4%	48.2%

Tableau 6.XI – Résultats UAR pour les conditions Intra et Pooling (niveau de chance : 25%).

Configuration (Apprentissage/Test)	Score moyen	Colère	Joie	Neutre	Tristesse
Voix normales / dégradées	40.7%	40.2%	29.0%	33.7%	59.7%
Voix dégradées / normales	37.9%	45.5%	49.1%	36.2%	20.7%
Moins / plus de 60 ans	39.4%	34.8%	44.8%	35.2%	42.9%
Plus / moins de 60 ans	42.4%	59.1%	37.0%	41.6%	31.8%

Tableau 6.XII – Résultats UAR (niveau de chance : 25%) pour la condition Deux-à-deux avec les critères d'âge et de qualité vocale.

dans la littérature, qui ne reflètent pas la réalité du comportement des systèmes sur le terrain. Un point intéressant est le couplage des approches cross-corpus avec les techniques de sélection de paramètres, que ce soit pour fiabiliser les résultats de sélection, très dépendants du corpus utilisé, ou pour tenter de calculer une similarité entre corpus.

CHAPITRE 7

SÉLECTION AUTOMATIQUE DE PARAMÈTRES

La tendance actuelle dans le domaine de la reconnaissance automatique des émotions est d'adopter une approche du type "force brute", c'est-à-dire de générer des centaines voire des milliers de paramètres pour représenter les données, dans l'espoir que le classifieur utilisé ensuite sera capable de déterminer quels paramètres seront pertinents pour la tâche de classification. Cette explosion du nombre de paramètres dans les systèmes est illustrée par la figure 7.1 (les données sont disponibles dans le tableau 7.I).

Outre les problèmes théoriques posés par cette approche (notamment la "malédiction de la dimensionnalité"), elle éloigne la perspective de pouvoir un jour mieux comprendre quels paramètres décrivent le mieux un état émotionnel donné (colère, tristesse) car elle rend l'interprétation des modèles encore difficile. Nous nous sommes donc intéressés aux techniques permettant de sélectionner les paramètres pertinents pour une tâche de classification donnée, dans le but de construire des modèles plus simples, plus compacts et plus faciles à comprendre et expliquer.

Référence	Année	Nombre de paramètres
Challenge Interspeech 2013 [233]	2013	6373
Challenge Interspeech 2012 [231]	2012	6125
Challenge Interspeech 2011 [232]	2011	4368
Challenge Interspeech 2010 [230]	2010	1582
Challenge Interspeech 2009 [229]	2009	384
Devillers et Vidrascu [79]	2006	~100
Vidrascu et Devillers [262]	2005	~50
Ververidis et Kotropoulos [261]	2004	119
Schuller et al. [227]	2003	20
McGilloway et al. [174]	2000	32

Tableau 7.I – Évolution du nombre de paramètres dans les systèmes de reconnaissance automatique des émotions dans la voix.

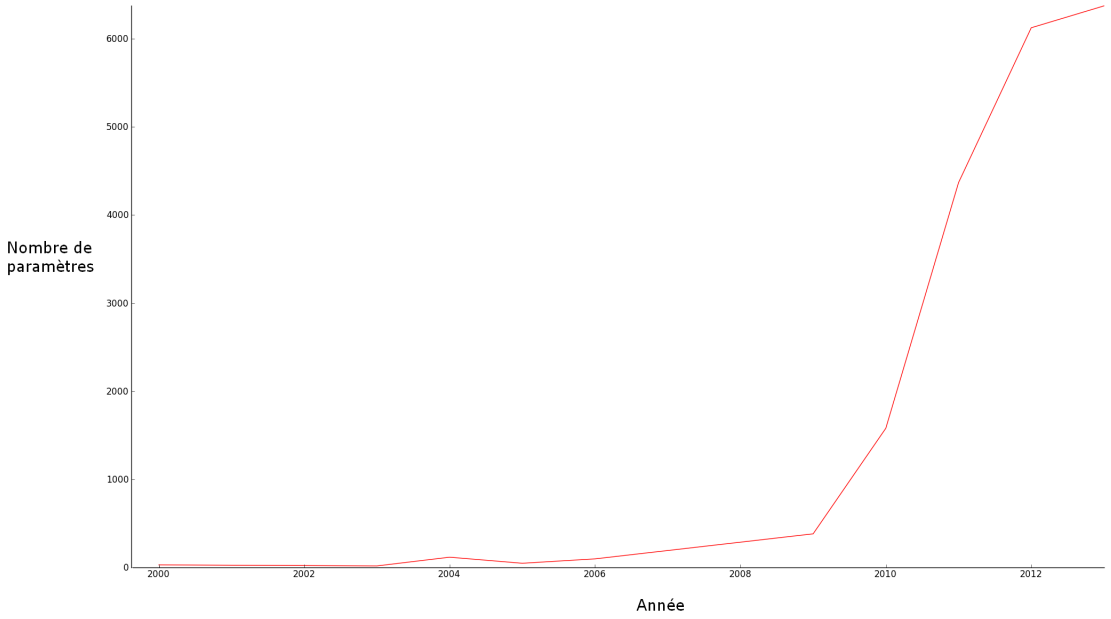

Figure 7.1 – Évolution du nombre de paramètres dans les systèmes de reconnaissance automatique des émotions dans la voix.

7.1 État de l'art

La sélection automatique de paramètres (SAP) désigne un ensemble de techniques destinées à simplifier la représentation des données en minimisant le nombre de paramètres utilisés dans les modèles d'apprentissage automatique, en trouvant les paramètres pertinents et en éliminant les paramètres redondants ou inadéquats. Elle peut satisfaire différents objectifs [117] :

- Optimiser les modèles en termes de mémoire ou de vitesse d'exécution ;
- Rendre les modèles plus robustes et plus performants (en luttant contre le problème de la "malédiction de la dimensionnalité" [18]) ;
- Minimiser les risques de sur-apprentissage (équivalent à augmenter le pouvoir de généralisation), d'autant plus présents que les modèles sont construits à partir de techniques d'apprentissage sophistiquées et en haute dimension ;

- Tenter de trouver une base de paramètres pertinents et explicatifs d'un phénomène. En effet l'approche utilisée en général (générer des paramètres en force brute puis appliquer un classifieur puissant) fait qu'il est très difficile d'interpréter les modèles appris ; cela serait plus aisé en diminuant fortement le nombre de paramètres utilisés.

On peut également citer d'autres objectifs [165] :

- Simplifier la visualisation des données pour de la sélection de modèles ;
- Réduire le bruit.

On peut insister sur la "malédiction de la dimensionnalité" : si en théorie et lorsque l'on travaille avec la population complète ou que l'on connaît la distribution, le fait d'ajouter des dimensions n'augmente pas la probabilité de mauvaise classification, en pratique on ne dispose que d'un ensemble fini d'échantillons de la population considérée [256]. Le fait d'ajouter des paramètres à la représentation des données est donc loin d'être anodin. En adoptant un autre point de vue, on peut montrer qu'en haute dimension, les données n'occupent qu'une fraction de plus en plus réduite de l'espace ; de plus la notion de similarité ou de distance perd sa signification ; enfin le risque de surapprentissage avec des classifieurs sophistiqués (arbres de décision, SVM...) augmente [84]. Il est donc nécessaire de travailler à réduire la dimension des données.

La SAP fait partie du champ plus large des techniques de réduction de dimensionnalité, avec lequel elle partage les mêmes objectifs. Il existe de nombreuses techniques de réduction de dimensionnalité. On peut établir une catégorisation sur la base de deux critères : supervisé/non-supervisé et transformation/sélection des paramètres [63]. Nous nous intéressons aux techniques de sélection supervisées. Parmi elles, on peut encore opérer une distinction entre les techniques du type "filtre", qui utilisent le calcul d'un critère simple (entropie, gain d'information...) pour ordonner les paramètres de manière individuelle, les techniques du type "wrapper", qui utilisent pour critère la performance d'un classifieur utilisé comme une boîte noire, et les techniques "embarquées" qui utilisent la régularisation au sein d'un algorithme d'apprentissage spécifique [146][117].

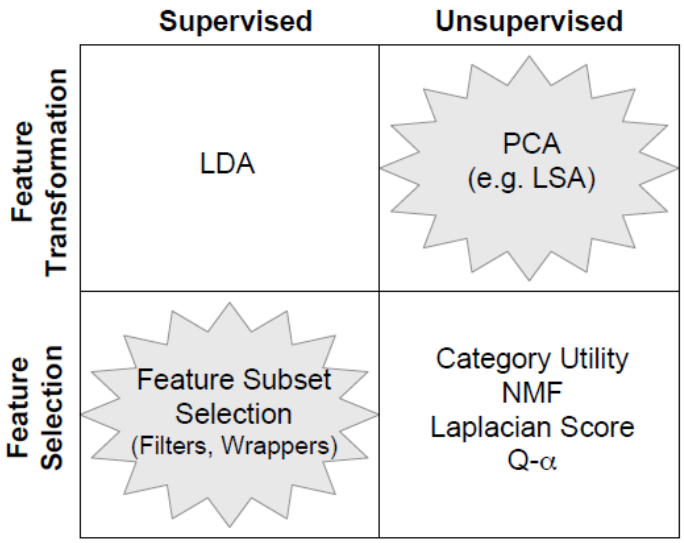

Figure 7.2 – Catégorisation des approches de SAP (d'après [63])

Nous nous intéressons spécifiquement aux méthodes de type "wrapper". Dans cette approche, la sélection de paramètres est vue comme un problème d'optimisation général, où l'on cherche, parmi l'espace des sous-ensembles de paramètres possibles (au nombre de 2N pour N paramètres), le sous-ensemble optimal au vu d'un critère donné (la performance de classification dans la plupart des cas) [210]. L'évaluation exhaustive de tous les ensembles n'étant pas possible, le problème est réputé NP-difficile [145]. Il existe de nombreuses heuristiques pour trouver un sous-ensemble satisfaisant en un temps raisonnable, fondées sur des approches stochastiques (algorithmes génétiques, recuit simulé...) ou gloutonnes (best-first, branch-and-bound, sélection avant et élimination arrière...) ; de nombreuses heuristiques sont applicables puisque la formulation reste générale [277].

7.1.1 Algorithmes de sélection séquentielle

Les algorithmes de sélection séquentielle ont pour principe d'ajouter ou de retirer un paramètre à un ensemble de paramètres existant. A chaque étape, le paramètre sélectionné est celui qui optimise le critère choisi.

7.1.1.1 Sélection avant

Dans le cas de la sélection séquentielle avant (SFS - *Sequential Forward Selection*) [268], l'algorithme sélectionne le paramètre qui apporte le meilleur gain de performance de classification par rapport à l'ensemble de paramètres courant. En notant $S = \{x_i\}$ l'ensemble des N paramètres du problèmes, S_k l'ensemble de paramètres courant, contenant k paramètres ; $J(S)$ le critère à optimiser, c'est-à-dire la performance de classification en utilisant les paramètres contenus dans S, on peut formaliser l'algorithme SFS de la manière suivante :

Algorithm 1 Sélection avant

$S_0 \leftarrow \{\}$
for $k = 0$ to $N-1$ **do**
  $\hat{x} \leftarrow \arg\max_{x \in S \setminus S_k} J(S_k \cup \{x\}) - J(S_k)$
  $S_{k+1} \leftarrow S_k \cup \{\hat{x}\}$
end for

La suite des (S_k) donne les meilleurs ensembles à k paramètres.

7.1.1.2 Sélection arrière

Dans le cas de la sélection séquentielle arrière (SBS - *Sequential Backward Selection*) [171], l'algorithme cherche à retirer un paramètre d'un ensemble existant. Le paramètre choisi est celui qui apporte la plus petite perte de performance de classification (donc moins pertinent par rapport à la tâche de classification). En reprenant les notations précédentes, on peut formaliser l'algorithme SBS de la manière suivante :

Algorithm 2 Sélection arrière

$S_N \leftarrow S$
for $k = N$ to 1 **do**
  $\hat{x} \leftarrow \arg\min_{x \in S_k} J(S_k) - J(S_k \setminus \{x\})$
  $S_{k-1} \leftarrow S_k \setminus \{\hat{x}\}$
end for

7.1.1.3 Performances et critiques

Les algorithmes SFS et SBS sont réputés mieux fonctionner que les approches de type filtre qui ne tiennent pas compte des possibles "effets cocktail" entre paramètres, car elles utilisent généralement des métriques univariées pour ordonner les paramètres. Par exemple, si l'on considère le problème "XOR" (voir figure X), une méthode de type filtre utilisant par exemple le coefficient de corrélation supprimera les paramètres qui auraient été reconnus utiles dans un classifieur et donc sélectionnés par une approche wrapper. De la même manière, des paramètres redondants seront conservés s'ils une bonne corrélation à la classe, alors qu'une approche wrapper pourra n'en sélectionner qu'un seul.

Cependant, SFS et SBS sont décriés pour leurs lourds besoins computationnels. En effet, en prenant l'exemple de SFS, à l'étape k, $N-k$ ensembles doivent être évalués pour trouver le meilleur paramètre x. Au total, ce sont donc $(N2+N)/2$ évaluations qui sont nécessaires pour SFS (idem pour SBS), là où les approches de type filtre n'en nécessitent que N. De plus, une "évaluation" dans le premier cas désigne un apprentissage de classifieur, tâche beaucoup plus lourde que dans le deuxième cas, où c'est généralement un calcul statistique relativement simple qui est mené.

Au-delà de la charge intensive en calculs, les approches séquentielles sont soumises au problème plus grave du sur-apprentissage, car ce sont les mêmes données qui sont utilisées pour chaque évaluation.

SFS et SBS sont de plus soumis au nesting effect, terme qui désigne le fait qu'un paramètre sélectionné pour l'ajout (ou le retrait), ne peut plus ensuite être retiré (ou ajouté). Ces algorithmes sont sous-optimaux et courent le risque de tomber dans des minimums locaux [277]. L'approche "plus-l-moins-r" a été proposée pour résoudre ce problème en alternant l étapes de SFS et r étapes de SBS [245] ; mais le prix à payer est le réglage des hyper-paramètres l et r.

7.1.2 Sélection flottante

Les méthodes de sélection flottantes de type SFFS (Sequential Floating Forward Selection) ou SFBS (Sequential Floating Backward Selection) [210] corrigent ce défaut en permettant une sélection avant ou arrière dynamique : dans le cas de SFFS, une ou plusieurs étapes de SBS sont effectuées après chaque étape de SFS si elles permettent d'améliorer la performance du critère (fonctionnement inverse pour SFBS). Les algorithmes SFFS et SFBS sont formalisés ci-dessous.

7.1.2.1 Algorithme SFFS

SFFS fonctionne en essayant de retirer un ou plusieurs paramètres après chaque paramètre ajouté, en comparant les performances aux meilleurs ensembles précédemment trouvés. On peut formaliser ce comportement en trois étapes :

- Une étape d'inclusion, correspondant à l'algorithme SFS, où l'algorithme cherche à trouver le paramètre $\hat{x}$ qui permettra le meilleur gain de performance à partir de l'ensemble S_k, parmi les $N-k$ paramètres possibles ; l'ensemble S_{k+1} est alors formé par l'union de S_k et de $\{\hat{x}\}$.

- Une étape d'exclusion conditionnelle, correspondant à l'algorithme SBS, où l'on cherche à éliminer le paramètre $\hat{x}$ le moins significatif dans l'ensemble S_{k+1}. Si la performance de l'ensemble $S_{k+1} \setminus \{\hat{x}\}$ est inférieure à celle de S_k, on retourne à l'étape d'inclusion en incrémentant k. Sinon, l'ensemble $S_{k+1} \setminus \{\hat{x}\}$ devient le nouvel ensemble S_k et on passe à l'étape suivante.

- Ici l'étape d'exclusion conditionnelle est répétée tant qu'elle rapporte des performances. On reprend alors l'étape d'inclusion avec le dernier ensemble S'_k.

Le pseudo-code pour l'algorithme SFFS est proposé ci-dessous.

7.1.2.2 Algorithme SFBS

SFBS est l'algorithme symétrique de SFFS, comme SBS l'était pour SFS : au lieu de réaliser plusieurs étapes d'exclusion après chaque inclusion, plusieurs étapes d'inclusion

Algorithm 3 SFFS

```
S_0 ← {}
k ← 0
while k < N − 1 do
    // SFS step
    x̂ ← arg max_{x∈S\S_k} J(S_k ∪ {x}) − J(S_k)
    S_{k+1} ← S_k ∪ {x̂}

    // SBS steps
    while k > 1 do
        x̂ ← arg min_{x∈S_{k+1}} J(S_{k+1}) − J(S_{k+1} \ {x})
        if J(S_{k+1} \ {x̂}) > J(S_k) then
            S_k ← S_{k+1} \ {x̂}
            k ← k − 1
        else
            BREAK
        end if
    end while
    k ← k + 1
end while
```

sont réalisées après chaque étape d'exclusion. L'algorithme démarre donc avec tous les paramètres. Le pseudo-code pour l'algorithme SFBS est proposé ci-dessous.

Algorithm 4 SFBS

```
S_N ← S
k ← N
while k > 1 do
  // SBS step
  x̂ ← arg min_{x∈S_k} J(S_k) − J(S_k \ {x})
  S_{k−1} ← S_k \ {x̂}

  // SFS steps
  while k < N − 1 do
    x̂ ← arg max_{x∈S\S_{k−1}} J(S_{k−1} ∪ {x}) − J(S_{k−1})
    if J(S_{k−1} ∪ {x̂}) > J(S_k) then
      S_k ← S_{k−1} ∪ {x̂}
      k ← k + 1
    else
      BREAK
    end if
  end while
  k ← k − 1
end while
```

7.1.2.3 Performances et critiques

SFFS et SFBS ont été pendant de nombreuses années (et sont toujours dans une certaine mesure) considérées comme des techniques de pointe [277]. Plusieurs études montrent leur efficacité par rapport à des techniques plus simples comme SFS ou les approches filtre [210][150][129]. Cependant on peut leur adresser un certain nombre de critiques :

- l'algorithme est assez compliqué à mettre en oeuvre ;
- le nombre d'évaluations est élevé ;

- l'approche séquentielle fait qu'il est long d'obtenir des sous-ensembles de cardinal non-trivial ;
- l'aspect déterministe de l'algorithme est une faiblesse car des sous-ensembles potentiellement prometteurs ne seront jamais évalués car inatteignables par un "chemin SFFS" dans le parcours des sous-ensembles possibles ;
- l'initialisation de l'algorithme (évaluation de tous les paramètres individuels) devient prohibitive en grande dimension.

Certaines de ces critiques ont été prises en compte dans le développement d'améliorations de SFFS. Elles sont présentées ci-dessous.

7.2 SFFS-SSH

L'approche SFFS-SSH (SFFS *with Set Similarity Heuristic*) [31] modifie l'heuristique de recherche en intégrant une mesure de similarité et une approche gloutonne pour accélerer SFFS. J'ai participé à son élaboration, son implémentation et à sa mise en oeuvre. Alors que tous les paramètres possibles sont considérés pour l'ajout à chaque étape de SFFS (recherche exhaustive), l'idée est ici d'ordonner les paramètres par leur gain de performance estimé décroissant et de s'arrêter dans la recherche dès qu'un meilleur résultat a été trouvé (approche de type Best-First).

Le gain de performance est estimé de la manière suivant : pour le paramètre-candidat courant x, on cherche dans l'historique l'ensemble le plus similaire à l'ensemble courant S_k parmi ceux à qui on a précédemment ajouté le paramètre x. On note cet ensemble $\hat{S}_k$ et le gain estimé pour x est alors $J(\hat{S}_k \cup \{x\}) - J(\hat{S}_k)$. La similarité entre ensembles est calculée par l'indice de Jaccard (voir équation 7.1).

$$I_{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7.1}$$

Malgré la perte d'exploration entraînée par ces changements, les résultats ont montré une accélération d'environ 80% par rapport à SFFS, et ce sans perte de qualité significative [31].

Cet algorithme a été implémenté et mis en oeuvre sur de nouveaux corpus notamment dans le cadre de la compétition InterSpeech 2012 [50]. Parmi les modifications apportées, on peut noter une étape supplémentaire parrallélisée de réglage d'hyper-paramètres du classifieur et l'adaptation de l'algorithme à la sélection de familles complètes de paramètres. Cette démarche avait pour but de surmonter le problème du trop grand nombre de paramètres (plus de 6 000) au regard du nombre d'exemples dans le corpus d'apprentissage (environ 250).

Cependant, certaines critiques faites à SFFS, notamment son aspect déterministe, sont toujours valables pour SFFS-SSH. C'est pour cette raison qu'un nouvel algorithme de sélection a été développé, s'inspirant de SFFS et de ses variantes.

7.3 H-SFFS

Cet algorithme (*Hop*-SFFS), que j'ai conçu et implémenté en m'inspirant de SFFS-SSH, tente de résoudre les problèmes de SFFS et SFFS-SSH, particulièrement l'impossibilité de gérer un nombre de dimensions trop élevé (initialisation trop coûteuse) et l'aspect déterministe du choix des candidats.

Les idées principales sont les suivantes :

- ne plus travailler avec un seul paramètre seulement pour l'ajout et le retrait, mais avec des sous-ensembles de taille variable, de manière à obtenir une exploration plus importante et à relâcher l'aspect déterministe du parcours SFFS ;

- organiser une exploration à plusieurs "échelles" pour disposer de points de repère dans l'arbre des sous-ensembles possibles ; on explorera d'abord l'arbre à "grande échelle" en ajoutant ou retirant un nombre M de paramètres, puis on utilisera une échelle plus petite (avec un diviseur de M) et ainsi de suite jusqu'à une exploration avec un seul paramètre ;

- l'heuristique de SFFS-SSH est généralisée pour travailler avec l'ajout/retrait de plusieurs paramètres ; elle est de plus modifiée pour que des sous-ensembles n'exis-

tant pas dans l'historique soient évalués après ceux existant, dans un ordre aléatoire ;

- seul un nombre limité de sous-ensembles candidats est considéré à chaque étape ; avec un sous-ensemble courant de cardinal k, ces candidats sont générés aléatoirement parmi l'ensemble des (n-k, M) candidats possibles (moins ceux déjà présents dans l'historique).

On remarque que de cette manière, l'algorithme n'a pas de phase d'initialisation : en partant de l'ensemble vide, des sous-ensembles de taille M sont considérés directement pour l'ajout. De plus on perd le côté déterministe car les candidats sont générés aléatoirement, donc l'algorithme est optimal asymptotiquement.

7.3.1 Fonctionnement de l'algorithme H-SFFS : pseudo-code

7.3.2 Résultats

L'algorithme H-SFFS a été évalué en comparaison avec plusieurs algorithmes de sélection de type wrapper :

- un algorithme de sélection trivial, qui évalue aléatoirement un sous-ensemble parmi les 2^N possibles à chaque étape (*random*) ;
- l'algorithme SFFS classique [210] (*SFFS*) ;
- l'algorithme SFFS-SSH [31] (*SFFS_SSH*) ;
- les scores obtenus en entraînant un modèle avec tous les paramètres sont également utilisés (*all_features*).

Quatre corpus ont été utilisés pour l'évaluation : Armen1, Armen2, JEMO et un mélange des corpus Armen et JEMO. Les algorithmes ont été laissés à calculer pendant une période de temps assez longue pour chaque corpus (environ deux jours). Étant données la charge du cluster de calcul et d'éventuelles erreurs, le nombre d'ensembles de paramètres évalués varie entre chaque algorithme et corpus.

Algorithm 5 H-SFFS

```
Require: max_power, max_iterations, max_candidates
  offset ← 0
  power ← max_power
  H ← {}
  for scale^max_power times do
    for power = max_power to 0 do
      S ← arg max_{V∈H, |V|=offset} J(V)
      hop_size ← scale^power
      iterations ← 0
      while iterations < max_iterations do
        // Forward Search
        Generate at most max_candidates candidate sets : randomly pick hop_size features from Z \ S, max_candidates times

        Sort the candidate sets according to the set similarity metric
        S* ← {}
        nr_eval ← 0
        for each candidate set S' do
          Evaluate J(S')
          H ← H ∪ {S'}
          nr_eval ← nr_eval + 1
          if J(S') > J(S*) then
            S* ← S'
          end if
          if J(S') > J(S) then
            S* ← S'
            BREAK
          end if
        end for
        S ← S*
        iterations ← iterations + nr_eval
        Do a Backward Search

        // Backward Search
        Generate at most max_candidates candidate sets : randomly pick hop_size features from S, max_candidates times
        Sort the candidate sets according to the set similarity metric
        for each candidate set S' do
          Evaluate J(S')
          H ← H ∪ {S'}
          nr_eval ← nr_eval + 1
          if J(S') > J(S*) then
            S* ← S'
          end if
          if J(S') > J(S) then
            S* ← S'
            BREAK
          end if
        end for
        iterations ← iterations + nr_eval
        if J(S*) > max_{V∈H, |V|=|S*|} J(V) then
          S ← S*
          Do a Backward Search
        else
          Do a Forward Search
        end if
      end while
      offset ← offset + 1
    end for
  end for
```

Le corpus JEMO a été collecté au LIMSI dans le cadre du projet ANR Affective Avatar[1]. Il a été enregistré en laboratoire, dans le cadre d'un jeu affectif, utilisant un modèle simple de détection de cinq émotions (colère, peur, tristesse, joie et neutre) et d'une échelle binaire d'activation (faible/forte) ; une boucle de retour affichait le résultat détecté par le système au joueur, qui devait tenter de faire deviner au système une émotion qu'on lui demandait d'exprimer, sans orientation lexicale particulière [30]. Après segmentation, annotation par deux annotateurs, retrait des segments non-consensuels et sous-échantillonnage de la classe Neutre, le corpus comprend 62 locuteurs (31 hommes et 31 femmes) âgés de 18 à 60 ans, pour une durée totale de 47 minutes, soit 1276 segments. Ce corpus est intéressant en complémentarité des corpus ARMEN car le mode d'élicitation utilisé fait que les données sont plus actés, plus prototypiques, où la prototypicalité peut être définie comme "le fait d'être reconnu de manière fiable par un ensemble d'annotateurs donné" [187].

7.3.2.1 Résultats pour le corpus ARMEN_1

Le nombre d'ensembles de paramètres évalués pour chaque algorithme est donné dans le tableau 7.II ci-dessous. Il est aussi représenté graphiquement sous forme d'histogramme, en fonction de la taille des ensembles (ci figure 7.3). Le fonctionnement séquentiel des algorithmes SFFS, SFFS-SSH et H-SFFS apparaît clairement à la différence de l'algorithme Random qui teste des ensembles de toute taille.

Une première comparaison entre les algorithmes séquentiels (SFFS, SFFS-SSH et H-SFFS) montre que si SFFS atteint de meilleurs scores sur le corpus d'apprentissage, les meilleures performances sur le corpus de développement sont atteintes par H-SFFS. Cela est bien mis en évidence sur la figure 7.4 (seuls les meilleurs résultats pour chaque taille d'ensemble sont représentés). On peut également voir que l'algorithme H-SFFS a une exploration beaucoup plus importante en termes de profondeur de l'arbre des ensembles possibles.

Ajoutons maintenant l'algorithme *Random* à la comparaison. Tout d'abord, on voit

[1] `http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-07-TLOG-0001`

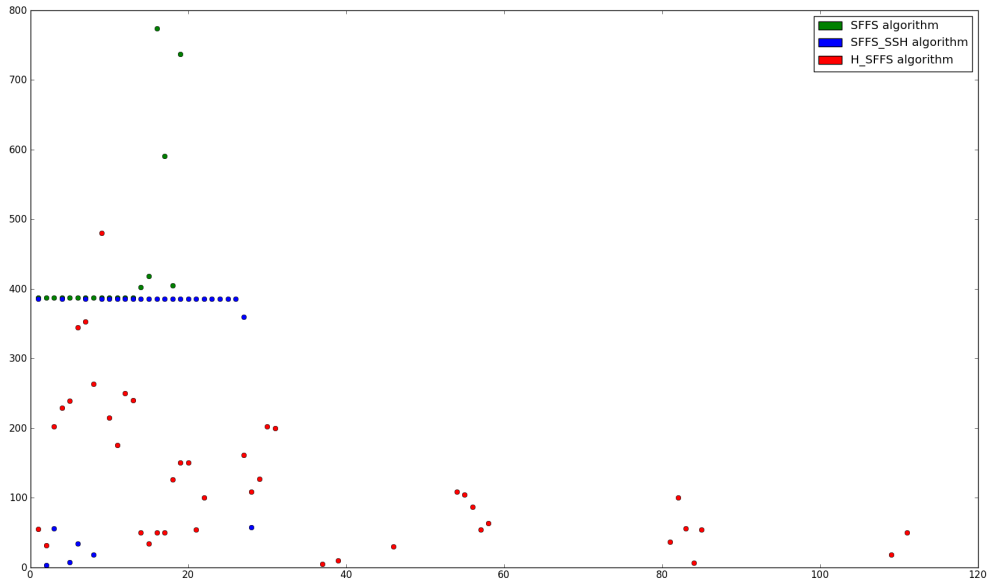

Figure 7.3 – Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_1.

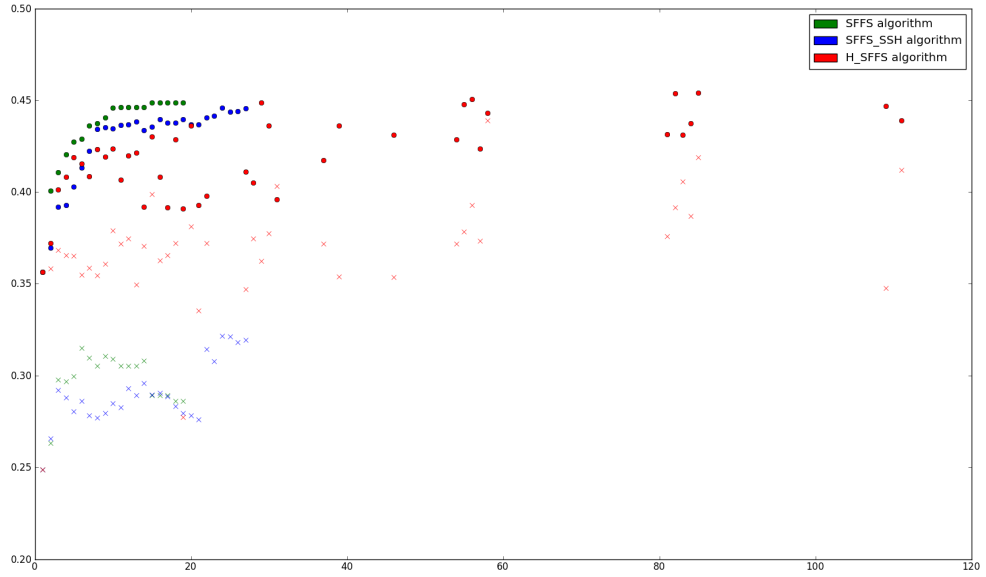

Figure 7.4 – Comparaison des performances sur les algorithmes séquentiels pour le corpus ARMEN_1 (les points correspondent aux scores sur le corpus d'apprentissage et les croix aux scores sur le corpus de développement).

alors sur la figure 7.5 que, par construction, *Random* a un pouvoir d'exploration beaucoup plus important. Ensuite, s'il paraît beaucoup moins efficace que SFFS ou SFFS-SSH sur les scores d'apprentissage, il l'est beaucoup plus en développement : le problème de sur-apprentissage de SFFS et SFFS-SSH est ici encore plus flagrant. Si l'on vérifie la répartition des évaluations en fonction de la taille d'ensemble (cf figure 7.6), on voit bien que la répartition de l'algorithme *Random* est quasi-uniforme, donc beaucoup moins d'ensembles sont testés pour chaque taille, réduisant d'autant le risque de sur-apprentissage de deuxième espèce (voir section 7.5).

En ce qui concerne les scores en test, on assiste globalement à une chute des scores. Les ensembles de développement et de test étant construits de manière à être indépendants en termes de locuteurs et vu leurs proportions respectives (80%, 10% et 10%), il est possible que le corpus de test contienne des segments d'un locuteur sensiblement différent des autres, faisant qu'une sélection de paramètres pertinente pour le corpus d'apprentissage ou de développement ne l'est pas pour le corpus de test.

La différence moyenne entre les scores entre apprentissage et développement d'une part, et apprentissage et test d'autre part est donnée en détail dans le tableau[2] 7.II, confirmant l'observation précédente. Si l'on observe le la différence apprentissage-développement, le sur-apprentissage est moins important pour H-SFFS que pour SFFS ou SFFS-SSH. Cela est vraisemblablement dû au fait que moins d'ensembles sont évalués pour une taille d'ensemble donnée.

[2]La moyenne est calculée sur les meilleurs ensembles par taille, donc sur plus d'ensembles pour H-SFFS que pour SFFS et SFFS-SSH.

Algorithme	Nombre d'ensembles évalués	Différence absolue dév. (points)	Différence relative dév.	Différence absolue test (points)	Différence relative test
Random	4086	-1.04	-2.48%	-10.7	-27.1%
SFFS	8357	-13.75	-31.68%	-13.6	-31.6%
SFFS-SSH	8619	-13.68	-31.98%	-14.2	-33.0%
H-SFFS	5419	-5.11	-12.18%	-15.8	-37.6%

Tableau 7.II – Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement sur les meilleurs ensembles pour le corpus ARMEN_1.

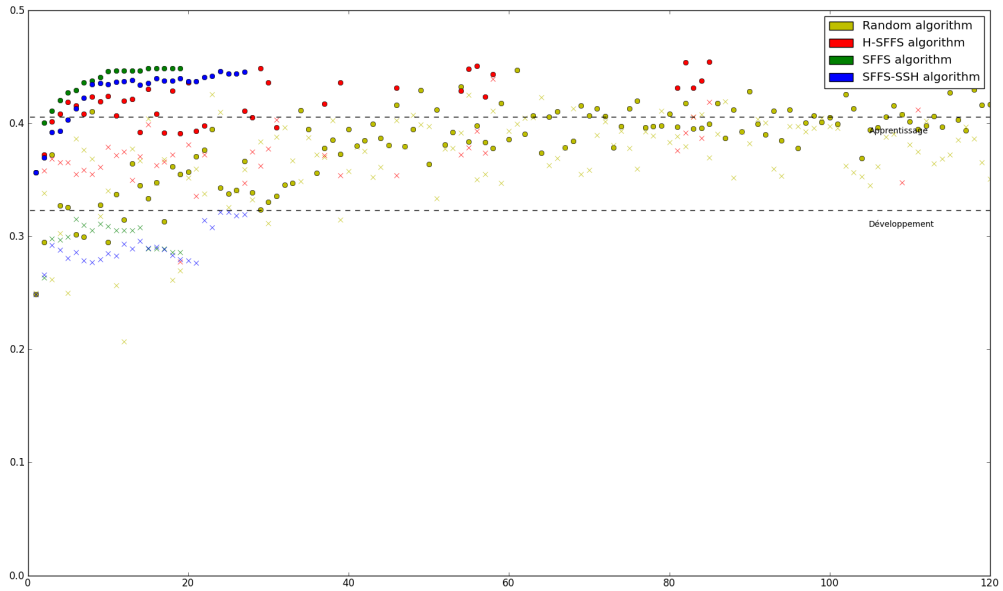

Figure 7.5 – Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN_1 (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement et les triangles aux scores sur le corpus de test).

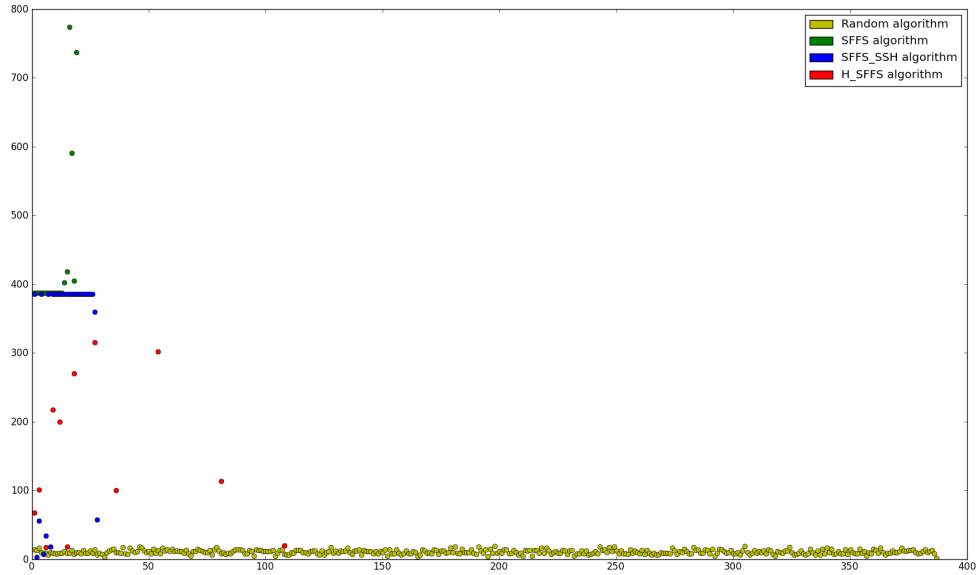

Figure 7.6 – Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_1 .

Pour mieux comparer les performances entre les algorithmes deux-à-deux, nous avons calculé combien de fois en pourcentage un algorithme donnait une meilleure performance qu'un autre. Ce calcul est valable seulement pour les tailles d'ensembles pour lesquelles il existe une performance pour les deux algorithmes calculés. Les résultats sont présentés dans le tableau 7.III. Dans la comparaison avec SFFS et SFFS-SSH, il faut garder à l'esprit que le pourcentage est calculé sur relativement peu d'exemples à cause du faible pouvoir d'exploration de ces algorithmes. On peut ainsi voir que H-SFFS surpasse les trois autres algorithmes en développement.

Une illustration des cinq meilleurs ensembles par algorithme est donnée sur la figure 7.7

7.3.2.2 Résultats pour le corpus ARMEN_2

Le nombre d'ensembles de paramètres évalués pour chaque algorithme est représenté ci-dessous sous forme d'histogramme, en fonction de la taille des ensembles (ci figure 7.8). Il est aussi donné dans le tableau 7.IV avec la différence moyenne absolue et relative entre les scores en apprentissage et en développement d'une part, et en apprentissage et en test d'autre part. Les observations faites précédemment sur l'évitement du sur-apprentissage supérieur pour H-SFFS par rapport à SFFS et SFFS-SSH sont ici moins valables car la perte de pouvoir de généralisation est importante, caractérisée par une chute d'environ 40% des scores pour tous les algorithmes. Cela est probablement dû à la faible taille des données.

Comparaison	Apprentissage	Développement	Test
H-SFFS >Random	97.6%	61.9%	61.9%
H-SFFS >SFFS	0%	89.5%	15.8%
H-SFFS >SFFS-SSH	21.7%	91.3%	39.1%
SFFS-SSH >Random	100%	22.2%	63.0%
SFFS-SSH >SFFS	0%	15.8%	63.2%
SFFS >Random	100%	31.8%	84.2%

Tableau 7.III – Comparaison des performances entre les algorithmes deux-à-deux pour le corpus ARMEN_1.

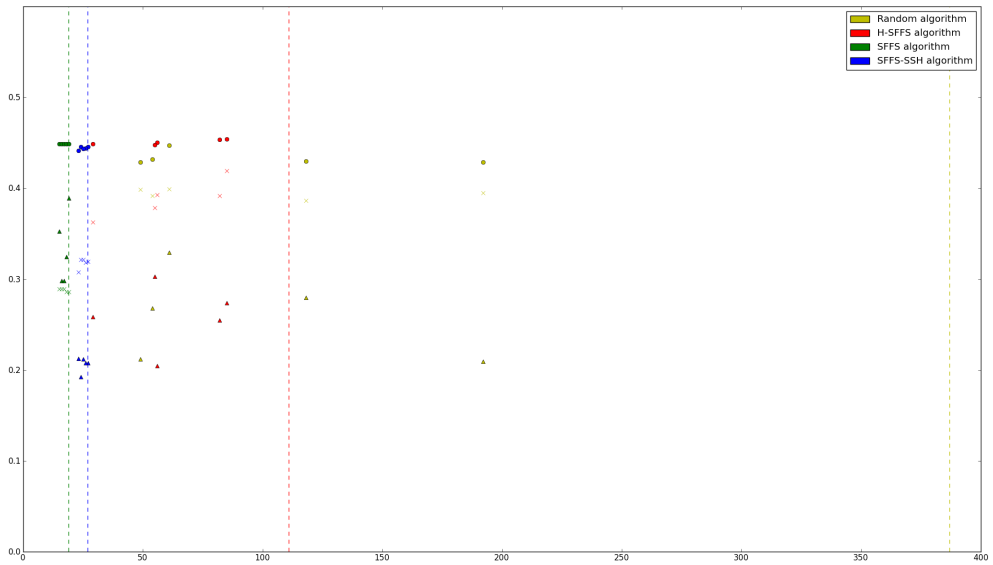

Figure 7.7 – Représentation des 5 meilleurs sets pour le corpus ARMEN_1 en fonction de leur taille et de leur performance en apprentissage (points), développement (croix), et test (triangles). Les lignes verticales correspondent à la taille maximale d'ensemble atteinte pour chaque algorithme.

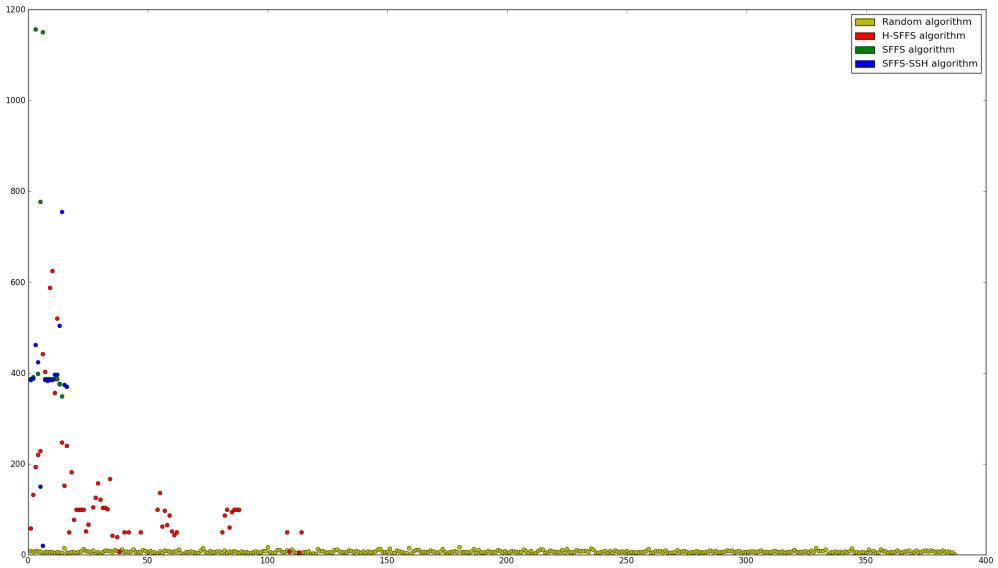

Figure 7.8 – Répartition des ensembles de paramètres évalués en fonction de leur taille pour le corpus ARMEN_2.

Algorithme	Nombre d'ensembles évalués	Différence absolue dév. (points)	Différence relative dév.	Différence absolue test (points)	Différence relative test
Random	2784	-14.9	-41.1%	-11.8	-32.4%
SFFS	7307	-17.3	-42.6%	-19.5	-48.6%
SFFS_SSH	6166	-15.5	-41.7%	-13.5	-37.3%
H_SFFS	8716	-17.0	-38.8%	-18.4	-42.8%

Tableau 7.IV – Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement sur les meilleurs ensembles pour le corpus ARMEN_2.

Néanmoins, comme on peut l'observer sur la figure 7.9 et le tableau 7.V, H-SFFS surpasse en performances de développement SFFS et SFFS-SSH dans la majorité des cas, et Random avec une courte avance. SFFS et SFFS-SSH sont toujours moins performants que Random en développement et en test à cause de leur problème de sur-apprentissage. En test, ces résultats sont à nuancer même si H-SFFS reste équivalent ou supérieur aux autres algorithmes.

7.3.2.3 Résultats pour le corpus JEMO

Les performances comparées des algorithmes sur le corpus JEMO sont représentées sur la figure 7.10. On peut par ailleurs voir dans le tableau 7.VI que H-SFFS a des performances supérieurs à SFFS-SSH et Random en test, mais pas à SFFS (cependant il n'y a que six points de comparaison possible dans ce cas, c'est donc à nuancer). On peut par contre remarquer que la différence entre apprentissage et développement ou test est très marquée globalement, illustrant la tendance des algorithmes de sélection au

Comparaison	Apprentissage	Développement	Test
H_SFFS >Random	98.3%	53.3%	65.0%
H_SFFS >SFFS	30.8%	76.9%	53.8%
H_SFFS >SFFS_SSH	87.5%	93.8%	50.0%
SFFS_SSH >Random	93.8%	18.8%	0.5%
SFFS_SSH >SFFS	7.7%	46.2%	61.5%
SFFS >Random	100%	38.5%	46.2%

Tableau 7.V – Comparaison des performances entre les algorithmes deux-à-deux pour le corpus ARMEN_2.

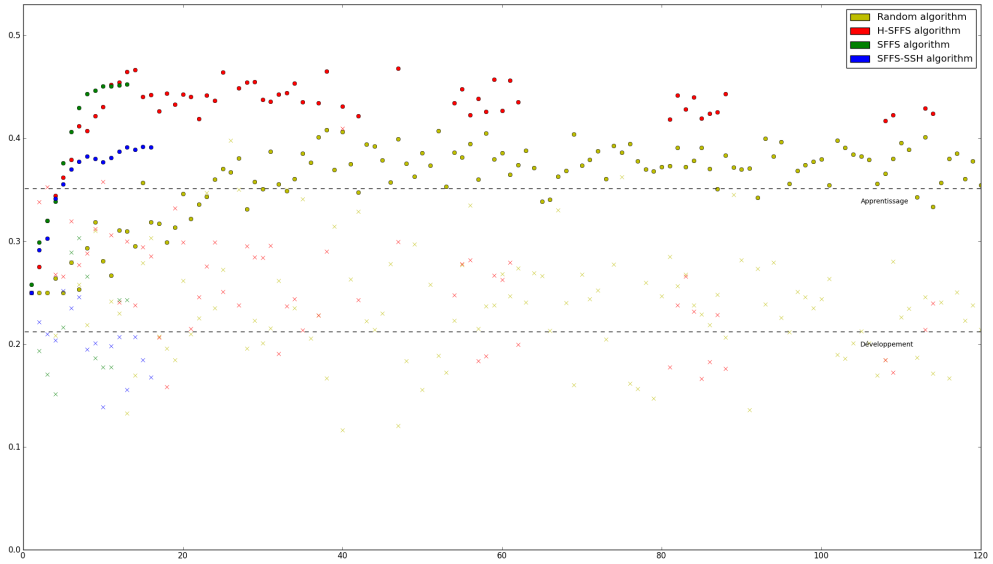

Figure 7.9 – Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN_2 (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement, et les triangles aux scores sur le corpus de test).

sur-apprentissage.

7.3.2.4 Résultats pour le corpus ARMEN + JEMO

Les résultats pour le corpus ARMEN + JEMO sont représentés dans la figure 7.11 ci-dessous. Il s'agit du corpus comprenant le plus gros volume de données, les algorithmes sont donc moins sujets au sur-apprentissage comme on peut le voir en détail dans le tableau 7.VII

Cela se reflète aussi dans le fait que les meilleurs scores sont assez proches, même si H-SFFS obtient le meilleur score (cf figure 7.12).

7.3.2.5 Résumé des résultats

Le tableau 7.VIII ci-dessous présente un récapitulatif des meilleurs scores atteints en test sur les quatre corpus pour les quatre algorithmes de sélection, ainsi que la taille du meilleur ensemble de paramètres (entre parenthèse). L'algorithme H-SFFS obtient la

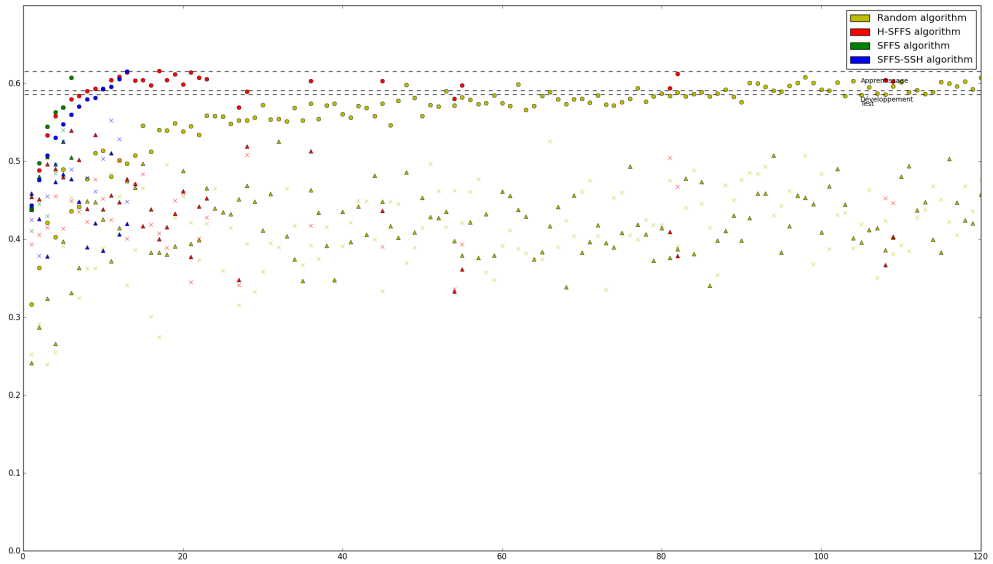

Figure 7.10 – Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus JEMO (les points correspondent aux scores sur le corpus d'apprentissage, les croix aux scores sur le corpus de développement, et les triangles aux scores sur le corpus de test).

Comparaison	Apprentissage	Développement	Test
H_SFFS >Random	100%	82.4%	64.7%
H_SFFS >SFFS	16.7%	0%	33.3%
H_SFFS >SFFS_SSH	76.9%	15.4%	76.9%
SFFS_SSH >Random	100%	100%	61.5%
SFFS_SSH >SFFS	16.7%	50.0%	16.7%
SFFS >Random	100%	100%	100%

Tableau 7.VI – Comparaison des performances entre les algorithmes deux-à-deux pour le corpus JEMO.

Algorithme	Nombre d'ensembles évalués	Différence absolue dév. (points)	Différence relative dév.	Différence absolue test (points)	Différence relative test
Random	1547	-5.3	-12.2%	-6.2	-14.3%
SFFS	20404	-10.7	-22.1%	-6.3	-12.8%
SFFS_SSH	14019	-8.0	-16.8%	-7.0	-14.6%
H_SFFS	20334	-7.8	-16.7%	-6.5	-13.9%

Tableau 7.VII – Nombre d'ensembles évalués et différence moyenne entre les scores en apprentissage et développement d'une part et apprentissage et test d'autre part sur les meilleurs ensembles pour le corpus ARMEN + JEMO.

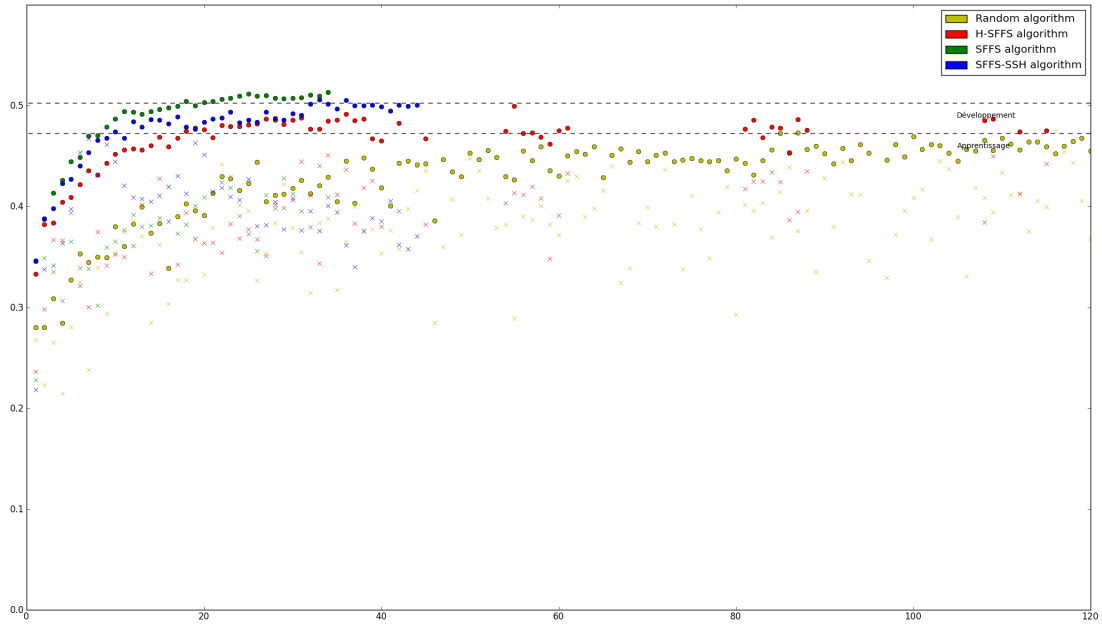

Figure 7.11 – Comparaison des performances sur les algorithmes séquentiels et l'algorithme *Random* pour le corpus ARMEN + JEMO (zoom).

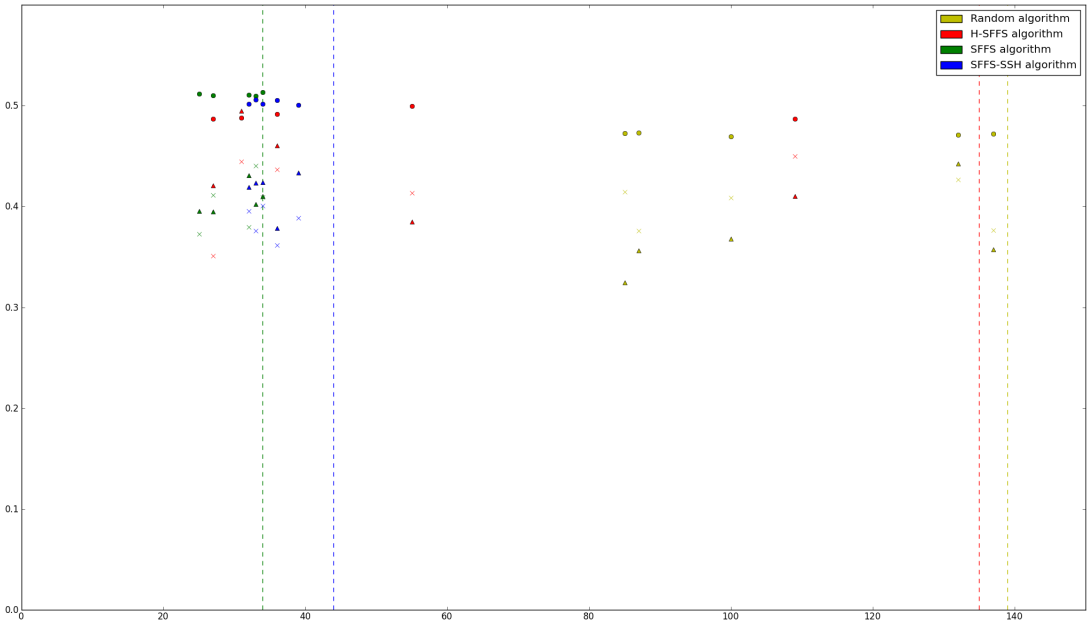

Figure 7.12 – Représentation des 5 meilleurs sets pour le corpus ARMEN + JEMO en fonction de leur taille et de leur performance en apprentissage (points), développement (croix) et test (triangles). Les lignes verticales correspondent à la taille maximale d'ensemble atteinte pour chaque algorithme.

meilleure performance sur le corpus comprenant le plus de données, ARMEN + JEMO ; sur les trois autres corpus, il obtient des scores supérieurs ou équivalents aux deux autres algorithmes séquentiels. L'algorithme simple *Random* obtient de très bonnes performances, en particulier sur les corpus comprenant le moins de données (ARMEN_1 et ARMEN_2). Comme la dégradation moyenne de sa performance en développement ou test est globalement inférieure à celle des autres algorithmes, il peut donc être comme une première étape de sélection intéressante. C'est ce qui sous-tend l'idée de l'exploration accrue et moins guidée dans un premier temps de l'algorithme H-SFFS.

7.3.2.6 Visualisation du fonctionnement des algorithmes

Le fonctionnement des algorithmes séquentiels est illustré dans les figures ci-dessous. Dans la figure 7.13, le processus de sélection pour l'algorithme SFFS (calculé sur le corpus ARMEN_1) est détaillé sous forme de graphe. L'axe des abscisses représente la performance de classification et l'axe des ordonnées la taille de l'ensemble considéré. Chaque noeud (départ à partir du noeud en bas à gauche) correspond au meilleur ensemble trouvé lors d'une étape de sélection avant (vers le haut) ou d'élimination (vers le bas). La couleur rouge indique que l'étape *Backward* a renvoyé un ensemble qui n'améliorait pas les scores suffisamment. On voit que le processus est bien séquentiel et assez simple.

Sur la figure 7.14, on a représenté le processus de sélection de l'algorithme SFFS-SSH. L'algorithme progresse par plus de petits sauts par rapport à SFFS, ce qui est dû à son heuristique gloutonne.

Sur la figure 7.15, le processus de sélection de l'algorithme H-SFFS est représenté en

Algorithme	ARMEN_1	ARMEN_2	JEMO	ARMEN + JEMO
Random	49.5% (7)	37.3% (4)	55.5% (247)	46.9% (114)
SFFS	38.9% (19)	25.4% (12)	52.5% (5)	48.7% (13)
SFFS_SSH	37.7% (5)	35.8% (5)	51.1% (11)	44.7% (9)
H_SFFS	38.7% (21)	34.9% (42)	54.0% (6)	49.5% (31)

Tableau 7.VIII – Résumé des meilleurs résultats (et tailles d'ensembles correspondantes) atteints en test.

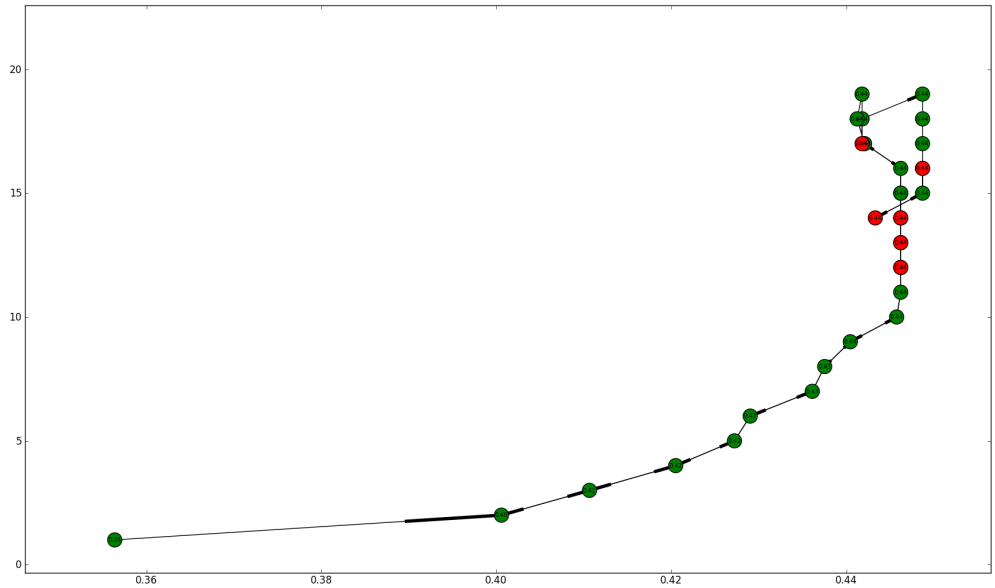

Figure 7.13 – Fonctionnement de l'algorithme SFFS.

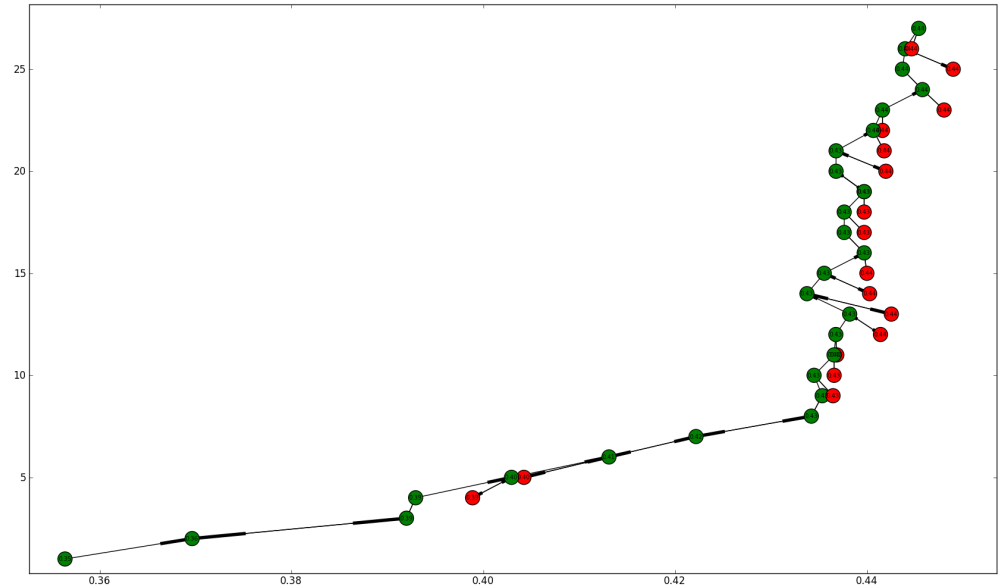

Figure 7.14 – Fonctionnement de l'algorithme SFFS-SSH.

fonction du temps (nombre d'étapes en abscisse et nombre de paramètres en ordonnée). On voit clairement les sauts de taille progressivement décroissante s'enchaîner. Sur la figure 7.16, le fonctionnement du même algorithme est représenté, cette fois-ci avec la performance en abscisse (axe transformé pour être plus visible). On peut constater que le processus est beaucoup plus compliqué et voir que globalement l'algorithme atteint des scores de plus en plus élevés, tout en explorant plus profondément.

7.4 Analyse des paramètres sélectionnés

Nous avons utilisé l'algorithme H-SFFS sur les corpus ARMEN_1, ARMEN_2 et JEMO pour comparer les résultats de la sélection de paramètres. La répartition des paramètres dans les différentes familles de paramètres (MFCC, ZCR, etc, et les dérivées) est représentée sur la figure 7.17 pour le meilleur ensemble de paramètres par corpus. La répartition normale des paramètres est également indiquée pour comparaison. On constate que pour le corpus ARMEN_2 et à la différence d'ARMEN_1 et JEMO, aucun paramètre du type F0 et dérivée de l'énergie n'est sélectionné. Ce constat se répète sur la figure 7.18 où sont représentées les répartitions des paramètres pour le meilleur ensemble à 20 paramètres pour chaque corpus. Cela corrobore les impressions acquises lors d'expériences précédentes, comme quoi le contenu d'ARMEN_2 est moins expressif, avec une forte proportion de neutre, ce qui pourrait expliquer que la dérivée de l'énergie ne soit pas sélectionnée.

Nous avons appliqué le calcul de similarité entre corpus, fondé sur la similarité des paramètres les plus pertinents par corpus, présenté au chapitre 6 aux corpus ARMEN_1, ARMEN_2 et JEMO. Selon cette mesure, ARMEN_2 semble moins similaire à JEMO qu'ARMEN_1. Cela confirme également le fait que le contenu de JEMO est plus prototypique et donc plus éloigné d'ARMEN_2.

La similarité entre corpus basée sur les paramètres les plus pertinents paraît donc intéressante car elle confirme des observations quantitatives et qualitatives.

Un autre type de visualisation permet de comprendre mieux les différences entre ces trois corpus du point de vue des paramètres pertinents. Sur les figures 7.19, 7.20 et 7.21,

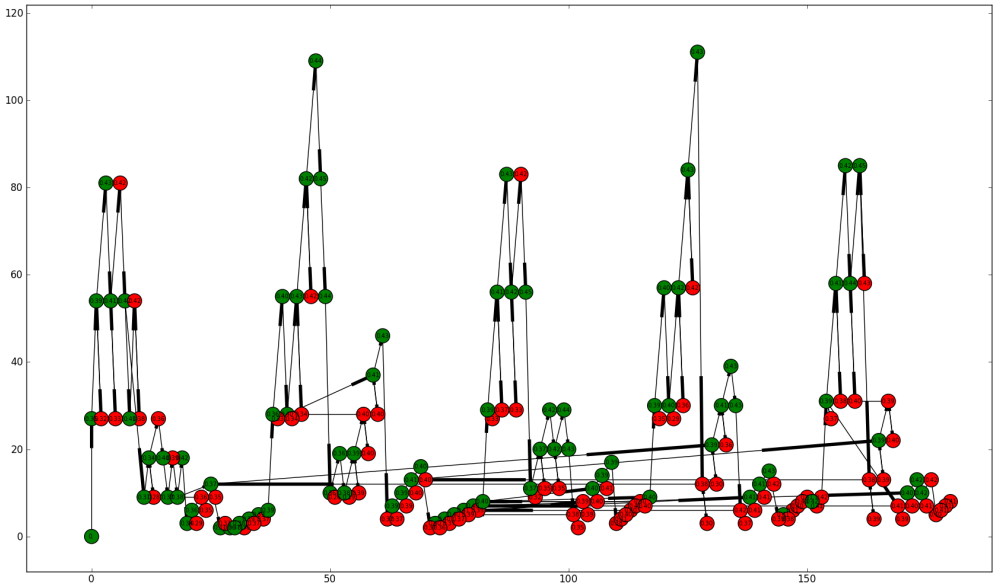

Figure 7.15 – Fonctionnement des étapes de l'algorithme H-SFFS.

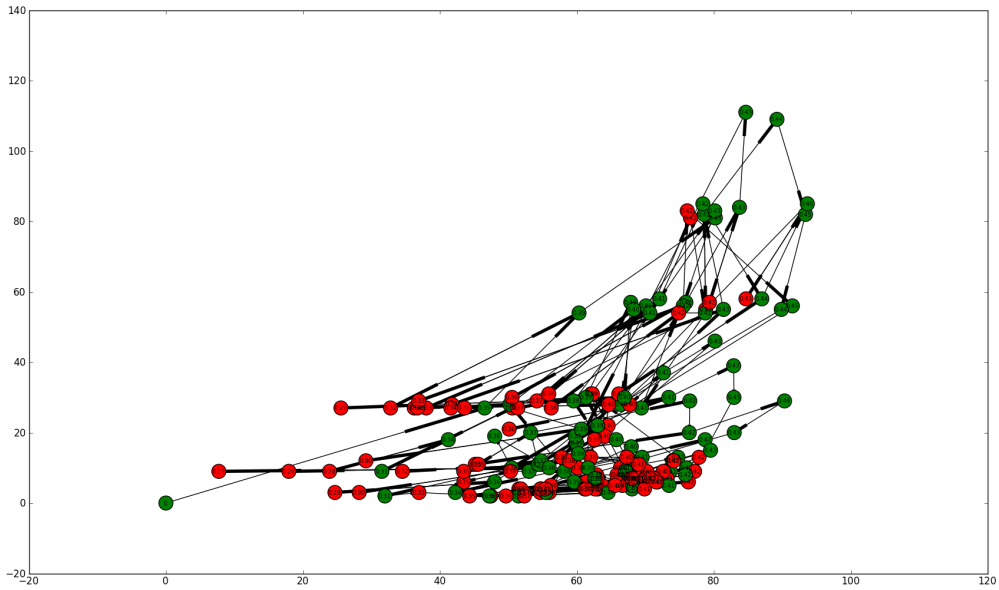

Figure 7.16 – Fonctionnement de l'algorithme H-SFFS.

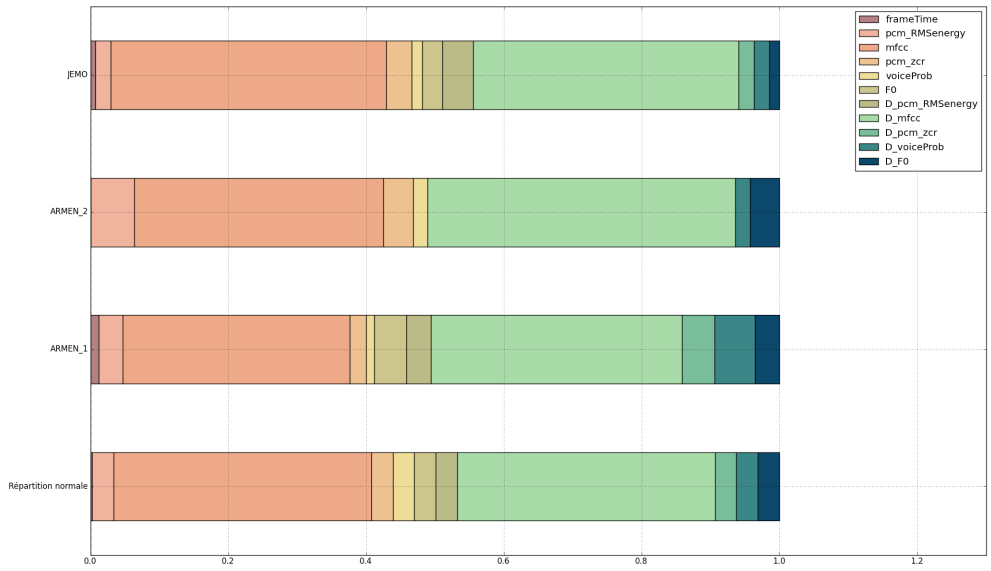

Figure 7.17 – Comparaison des proportions entre types de paramètres pour différents résultats de sélection (meilleur résultat choisi).

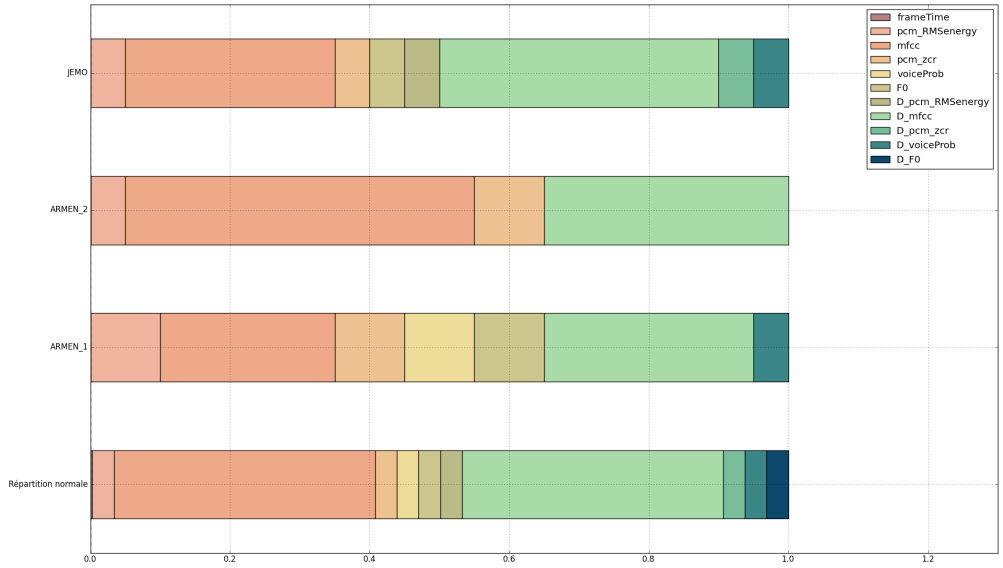

Figure 7.18 – Comparaison des proportions entre types de paramètres pour différents résultats de sélection (meilleur résultat pour 20 paramètres).

	ARMEN_1 vs ARMEN_2	ARMEN_1 vs JEMO	ARMEN_2 vs JEMO
Similarité de Jaccard	0.053	0.053	0.026

Tableau 7.IX – Scores de similarité fondés le meilleur ensemble à 20 paramètres pour chaque corpus.

on a représenté les meilleurs ensembles de paramètres, de la taille 1 à 20. On voit alors de manière frappante la différence entre les corpus ARMEN, pour lesquels l'algorithme H-SFFS a sélectionné des paramètres de tous les types de manière très répartie, et le corpus JEMO, pour lequel la sélection est beaucoup plus propre, avec souvent un seul paramètre par famille comme c'est le cas pour l'énergie et sa dérivée temporelle, le ZCR et sa dérivée, la F0 et la dérivée de la présence de voix. Cela peut être interprété par le fait que le contenu est suffisamment simple pour être correctement représenté par un seul paramètre de haut niveau, alors qu'il en faut plusieurs pour les corpus ARMEN. On peut également remarquer que la sélection est très stable pour JEMO alors qu'elle est plus chaotique pour ARMEN_1 et ARMEN_2. Si le nombre de paramètres à choisir est contraint (c'est-à-dire qu'il faudrait plus de paramètres pour représenter l'information de manière satisfaisante) et que pour une famille donnée, chaque paramètre apporte un peu d'information de manière assez équitablement répartie, on peut comprendre qu'avec les variations dues à la validation croisée, dont les plis sont choisis aléatoirement à chaque évaluation, même si elles sont faibles, on puisse retrouver ce genre de résultat.

Des différences plus subtiles apparaissent entre ARMEN_1 et ARMEN_2 : le ZCR et sa dérivée sont sélectionnés beaucoup plus souvent pour le premier que pour le second, idem pour l'énergie et sa dérivée. Les dérivées des MFCC sont très utilisées pour ARMEN_2 alors que c'est moins le cas pour ARMEN_1.

On voit donc que les paramètres pertinents d'un corpus peuvent être une source très intéressante d'informations lorsqu'on cherche à le comparer à d'autres. Les différences observées entre les corpus ARMEN, au contenu spontané, et le corpus JEMO, au contenu plus acté, sont des tendances qui peuvent suggérer une nouvelle manière de faire la distinction entre ces deux types de données. Il serait intéressant de pouvoir répliquer ces résultats sur d'autres corpus.

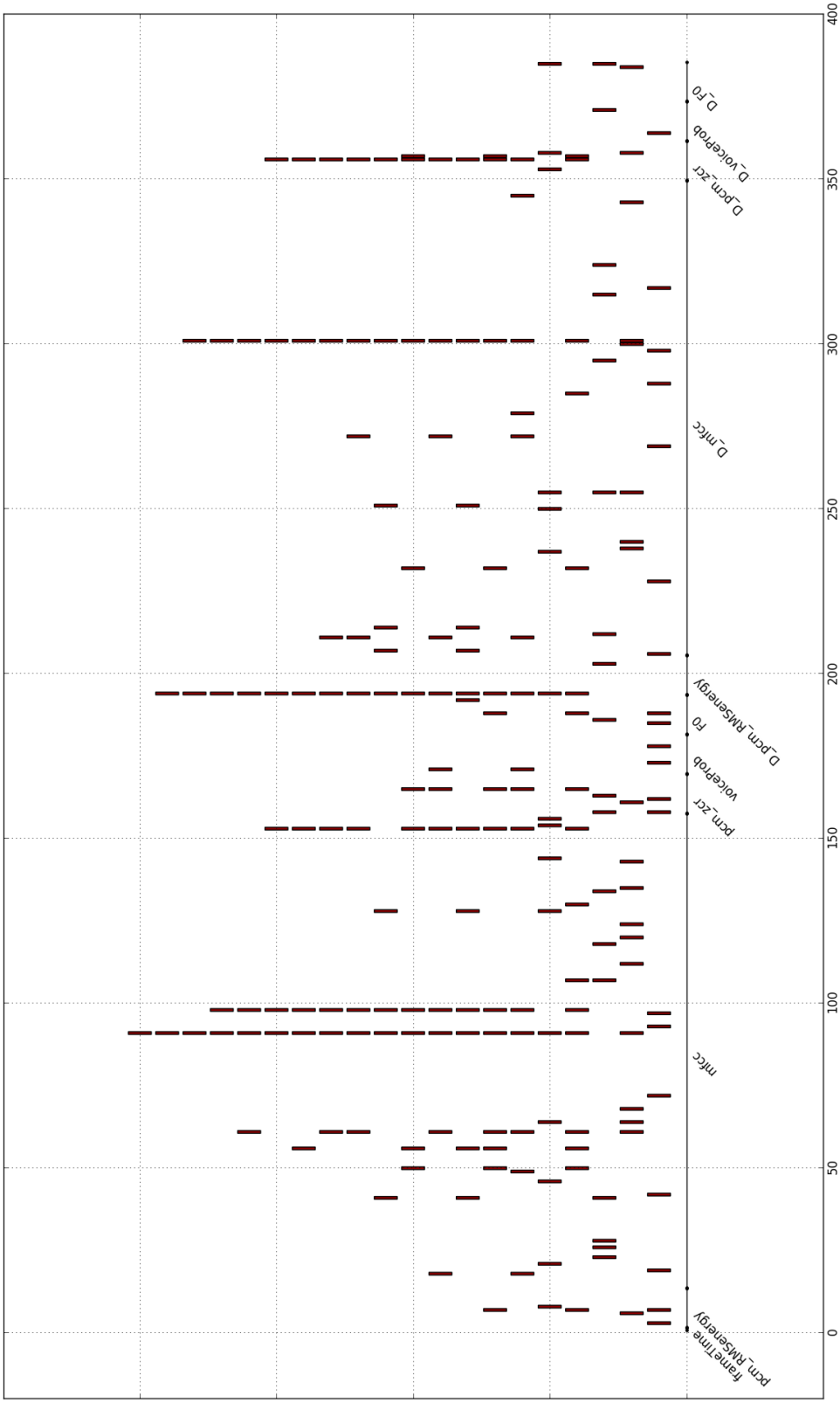

Figure 7.19 – Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus ARMEN_1.

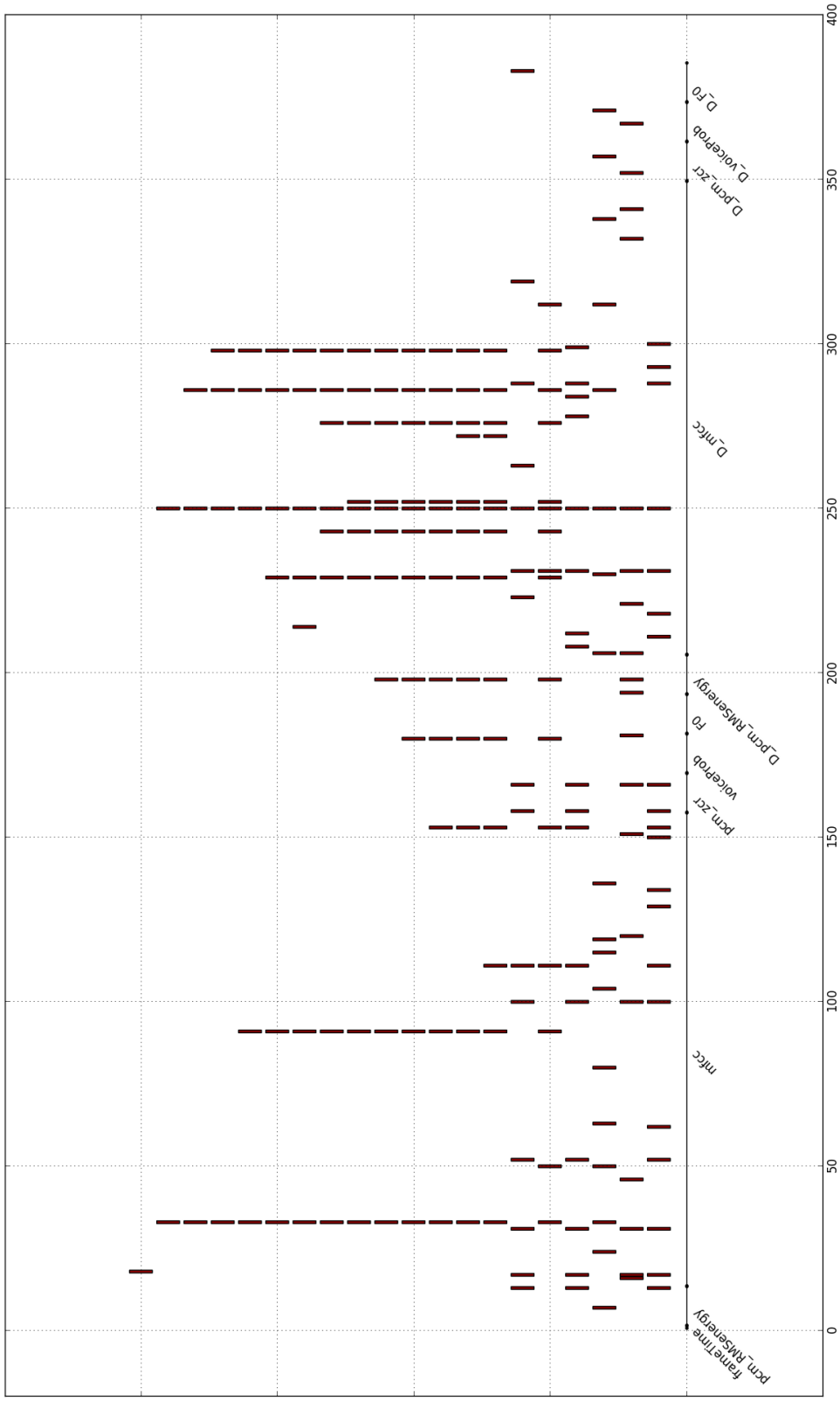

Figure 7.20 – Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus ARMEN_2.

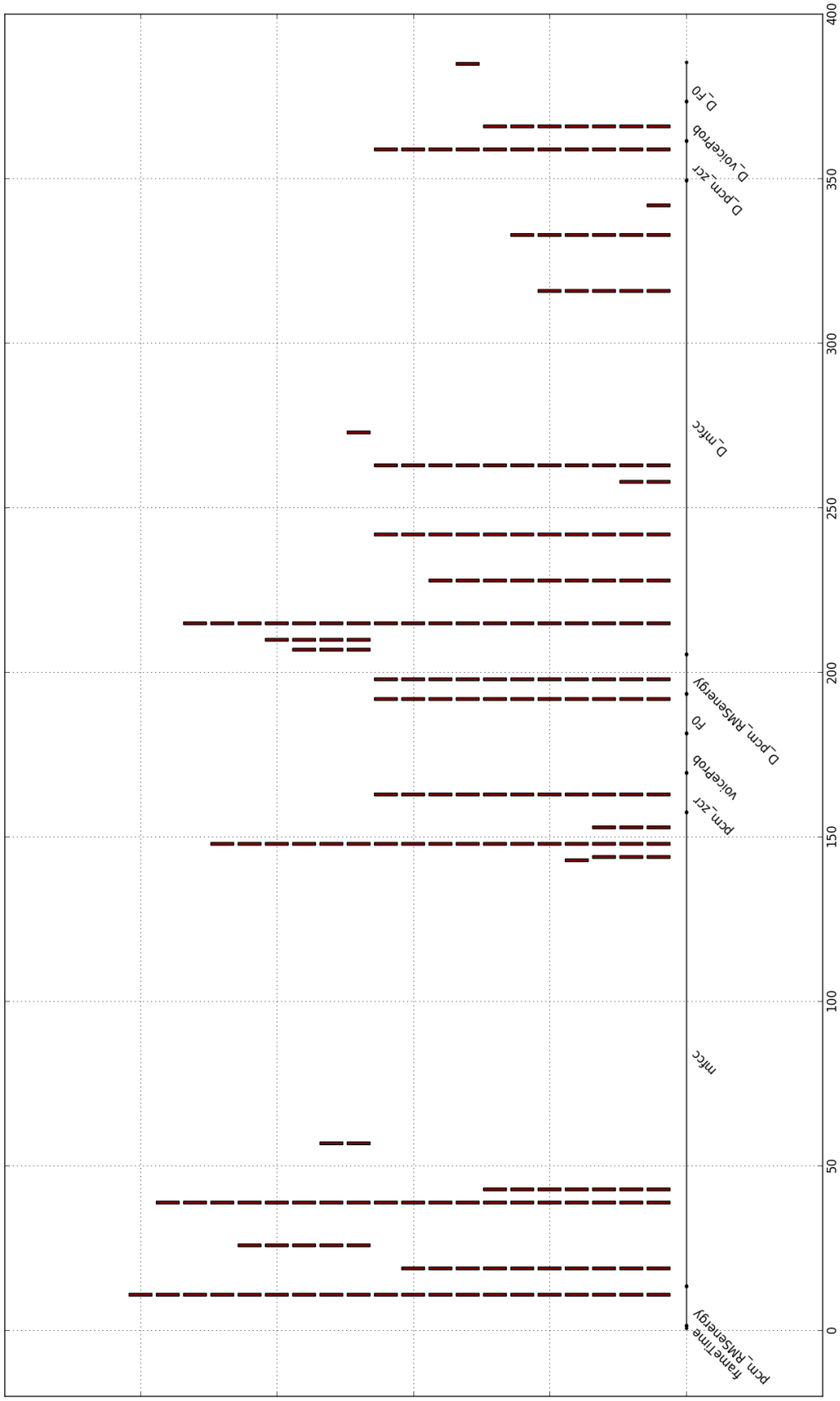

Figure 7.21 – Représentation des meilleurs ensembles de paramètres de taille 1 à 20 (de haut en bas) sélectionnés par l'algorithme H-SFFS pour le corpus JEMO.

7.5 Méthodologie : combattre le sur-apprentissage

Au vu des problèmes de sur-apprentissage rencontrés de manière générale dans l'application des algorithmes de sélection de paramètres, nous revenons dans cette section sur des notions théoriques.

Le sur-apprentissage est un concept fondamental dans le domaine des statistiques et de l'apprentissage automatique, théorisé notamment à l'aide de la dimension de Vapnik-Chervonenkis et des inégalités de Hoeffding [2]. Il intervient lorsque le modèle choisi est trop puissant par rapport au motif à apprendre ou lorsque trop peu de données sont disponibles pour représenter le phénomène à apprendre. Le risque est alors d'apprendre le bruit ou de mémoriser les particularités des données d'apprentissage ; la conséquence est la perte de pouvoir de généralisation, c'est-à-dire l'incapacité à classifier correctement des données non précédemment vues. Ce phénomène est illustré sur les figures 7.22 et 7.23.

Sur la figure 7.22, la cible d'apprentissage, un polynôme du second degré (en vert), est échantillonnée avec erreur (bruit gaussien). Deux hypothèses sont construites à partir de ces échantillons : un polynôme du second degré et un de degré 10. On voit bien que si l'erreur intra-corpus (sur les échantillons) est meilleure avec la deuxième hypothèse, plus complexe, l'erreur extra-corpus explose. Sur la figure 7.23, deux hypothèses (polynômes de degré trois) sont construites pour approximer la même cible que précédemment. Dans un cas, seuls trois échantillons bruités sont disponibles (en marron) et la performance extra-corpus est très mauvaise. Dans l'autre cas, dix fois plus d'échantillons sont disponibles et l'hypothèse est très proche de la cible malgré le bruit.

Les solutions pour éviter le sur-apprentissage sont de disposer de plus de données d'apprentissage, ce qui n'est pas forcément possible, de mettre en place des procédures du type validation croisée pour utiliser la totalité des données, ou de pénaliser la complexité des hypothèses (régularisation).

Un autre problème appelé "sur-apprentissage de seconde espèce" [214] apparaît dans l'utilisation des techniques wrapper mais également plus largement dans la sélection de modèles et également dans les statistiques (*multiple testing* [53]). Ce problème ne se

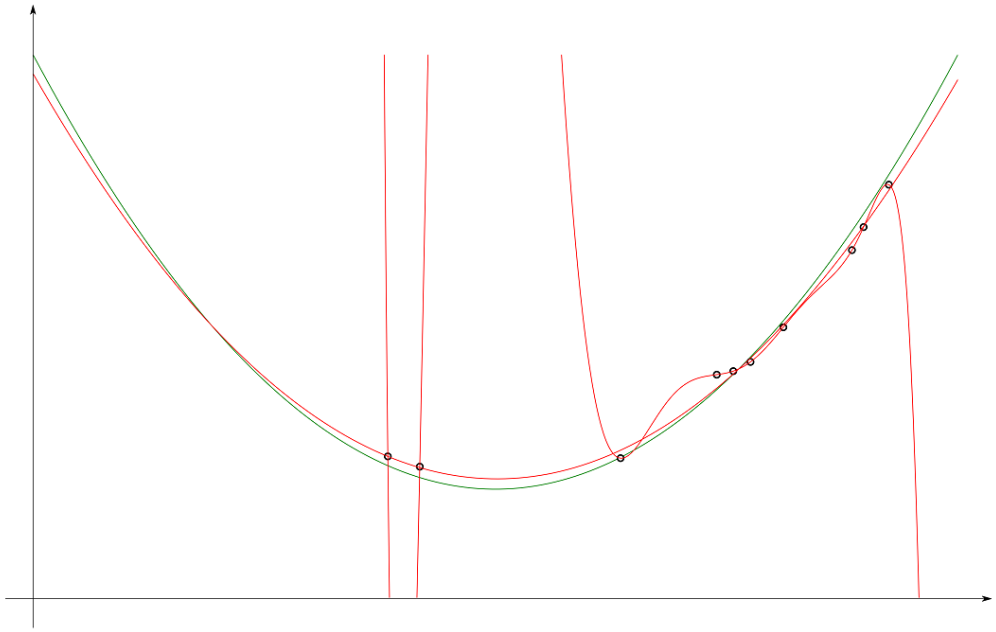

Figure 7.22 – Illustration du phénomène de sur-apprentissage avec un modèle trop complexe et un modèle adapté à la complexité de la cible.

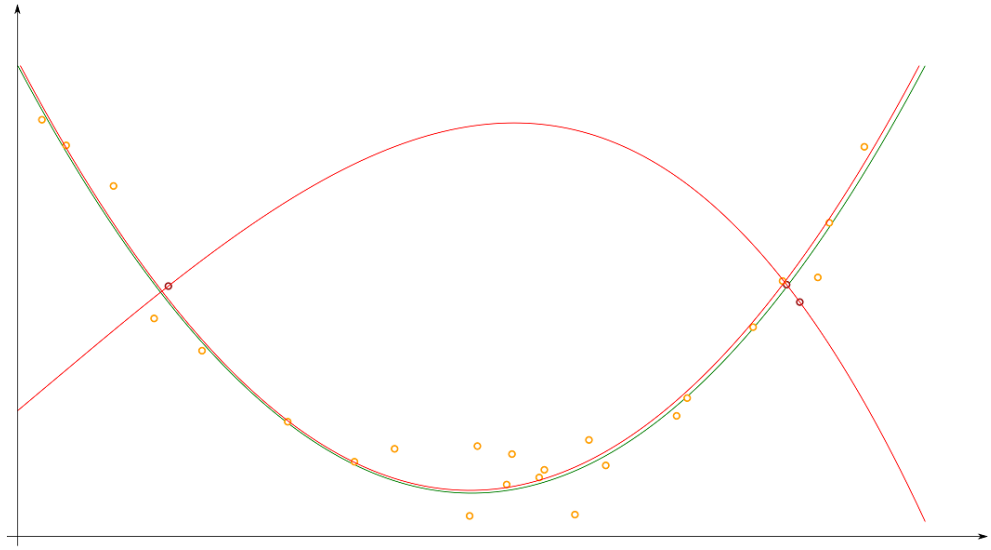

Figure 7.23 – Illustration du phénomène de sur-apprentissage avec trop peu de données d'apprentissage.

présente pas lorsque que l'on utilise un modèle trop complexe, mais lorsqu'on teste trop de modèles sur les mêmes données avant de choisir le meilleur [191]. Même la validation croisée n'est pas à l'abri du phénomène, le constat a été fait en termes frappants :

> A naive intensive use of cross-validation, perhaps over many thousands of models, may produce a deceptively good lowest-error model, in a manner similar to overfitting of data.
>
> *D'après Moore et Lee (1994) [183]*

En effet il est possible, mais peu probable, qu'un modèle donne un bon score en validation croisée sur un ensemble de données, mais un mauvais score sur d'autres données similaires mais non précédemment vue (la validation croisée est un estimateur non-biaisé de la performance sur la population complète ; sa variance décroit avec la taille des données de validation croisée [2]). Il est donc probable de tomber sur un tel modèle si l'on en essaie beaucoup et si la taille est des données de validation croisée est faible [145]. Une illustration simple du problème permet de mieux ancrer l'idée : un trader qui "bat le marché" dix années d'affilée peut être considéré comme un génie, mais si l'on considère qu'il y a 1 000 traders en activité et que chacun a 50% de chances d'obtenir un bon résultat sur une année donnée, il est très probable que l'un d'entre eux se révèlera aussi chanceux que notre "génie" par pure chance [84]. En fait, c'est le contraire qui serait étonnant, c'est-à-dire qu'il n'y ait jamais de série gagnante...

Dans les techniques wrapper, c'est donc bien l'utilisation à outrance des scores de validation croisée sur les mêmes données (souvent des milliers de fois) qui pose problème, car les scores annoncés sont beaucoup trop optimistes et généralisent mal.

Les solutions proposées pour lutter contre ce phénomène sont des méthodologies simples mais efficaces : il est par exemple possible d'effectuer une évaluation finale sur un autre set de données jamais vues durant la phase de sélection [214]. Une comparaison schématique de cette méthodologie par rapport à la méthodologie classique est donnée ci-dessous (figure X [242]). Une autre possibilité est d'utiliser une procédure du type percentile-cv où l'on choisit le modèle correspondant au k-percentile des scores (avec k à fixer) plutôt que celui donnant le meilleur score [191].

Malgré la description récurrente et précise du phénomène dans la littérature [147], de nombreux utilisateurs des techniques de sélection utilisent encore une mauvaise méthodologie d'évaluation, conduisant à des scores trop optimistes [102].

7.6 Conclusion

Un nouvel algorithme de sélection séquentielle a donc été conçu et évalué par rapport à des algorithmes existants. Il possède plusieurs avantages, comme le fait de s'affranchir d'une étape d'initialisation parfois trop longue à calculer, un pouvoir d'exploration accru. Une exploration de la similarité entre corpus basée sur cet algorithme a été réalisée et a permis de mettre à jour des différences dans les paramètres pertinents entre corpus, dans le nombre de paramètres nécessaires pour représenter correctement l'information portée par un type de paramètre et dans la stabilité de la sélection. Ces résultats suggèrent de nouvelles pistes pour la distinction automatique de contenu acté ou spontané et de manière plus générale pour la caractérisation et la comparaison de corpus.

Troisième partie

Système de dialogue émotionnel avec un personnage virtuel

CHAPITRE 8

IMPLÉMENTATION

Ce chapitre présente des aspects d'implémentation du prototype pour le système complet, c'est-à-dire le système de dialogue incluant la reconnaissance des émotions et l'AVE. Les travaux présentés ci-dessous sont le fruit de la collaboration avec les équipes des thèmes *Dimensions Affectives et Sociales dans les Interactions Parlées* et *Agents Virtuels et Émotions* du LIMSI et la PME Voxler[1].

8.1 Fonctionnalités

Le prototype développé est prévu pour posséder les fonctionnalités suivantes :

- Fonctionnement avec un flux audio continu.
- Fonctionnement en temps-réel.

8.2 Architecture et détail des modules

L'architecture logicielle du prototype a été conçue en collaboration avec la PME Voxler. Elle se décompose en plusieurs modules, dont l'agencement est représenté sur la figure 8.1 :

Un module de traitement de l'audio, qui gère le flux continu d'audio, ainsi que la reconnaissance de la parole. Les fonctions de segmentation et de classification des émotions, sur lesquelles je suis plus particulièrement intervenu et qui sont détaillées plus bas, font partie de ce module.

Un module de gestion du dialogue, comprenant la mémoire sous forme d'arbres de dialogue, dont le format a été mis au point conjointement avec Voxler.

[1]Je tiens à remercier chaleureusement Damien Henry, Pedro Cardoso, Olivier Veneri, Aymeric Zils et Nicolas Delorme de Voxler, ainsi que Céline Clavel et Matthieu Courgeon du LIMSI pour leur aide et le partage de leur expertise dans ce travail d'intégration.

Un module pour l'AVE, commandé par le dialogue.

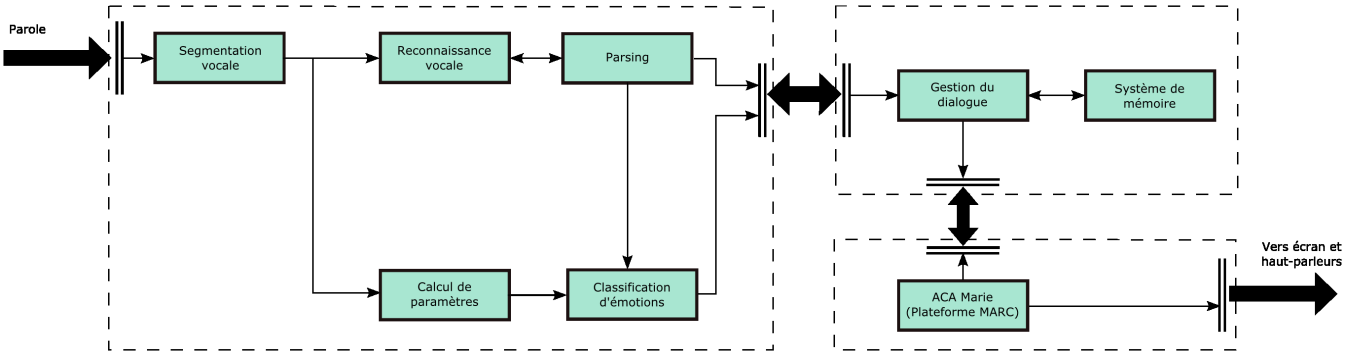

Figure 8.1 – Architecture du prototype.

Il existe de nombreux travaux en architecture de systèmes de dialogue intégrant la reconnaissance de la parole, la détection des émotions dans la parole [203] ou un agent virtuel expressif [224]. Ces architectures existantes intègrent cependant rarement toutes ces composantes pour permettre une interaction affective en temps-réel avec un agent virtuel. C'est le prototype d'un tel système qui a été réalisé ici.

Les modules audio et de mémoire devant communiquer intensément, l'API du système complet a été définie avec Voxler.

Le système étant encore à l'état de prototype, des choix de simplicité ont été faits. Par exemple, la reconnaissance de la parole n'est pas en domaine ouvert, le lexique de reconnaissance est également très limité (entre 10 et 20 mots reconnus par noeud), on parle plutôt de *word-spotting*. La reconnaissance est basée sur des grammaires prédéfinies encodées dans les arbres de dialogue. Ce choix respecte également des contraintes de robustesse.

L'ensemble a été implémenté en plusieurs langages (C++ majoritairement) et fait appel à des librairies externes.

8.2.1 Gestion du flux audio

Le flux audio entrant est stocké dans un *buffer* avant analyse et segmentation. Après la segmentation, le reste de la chaîne audio travaille avec des segments dont la durée est

de l'ordre de la seconde.

8.2.2 Segmentation de la voix

Le but de ce module est de détecter la présence de parole et d'extraire les segments audios correspondants du flux audio. Il a été implémenté conjointement avec la PME Voxler. L'algorithme implémenté est basé sur une machine à états observant le dépassement de seuils sur plusieurs mesures calculées, notamment l'énergie et le ZCR ; son fonctionnement est décrit sur la figure 8.2. Cette approche est couramment décrite dans la littérature [211, 213, 272].

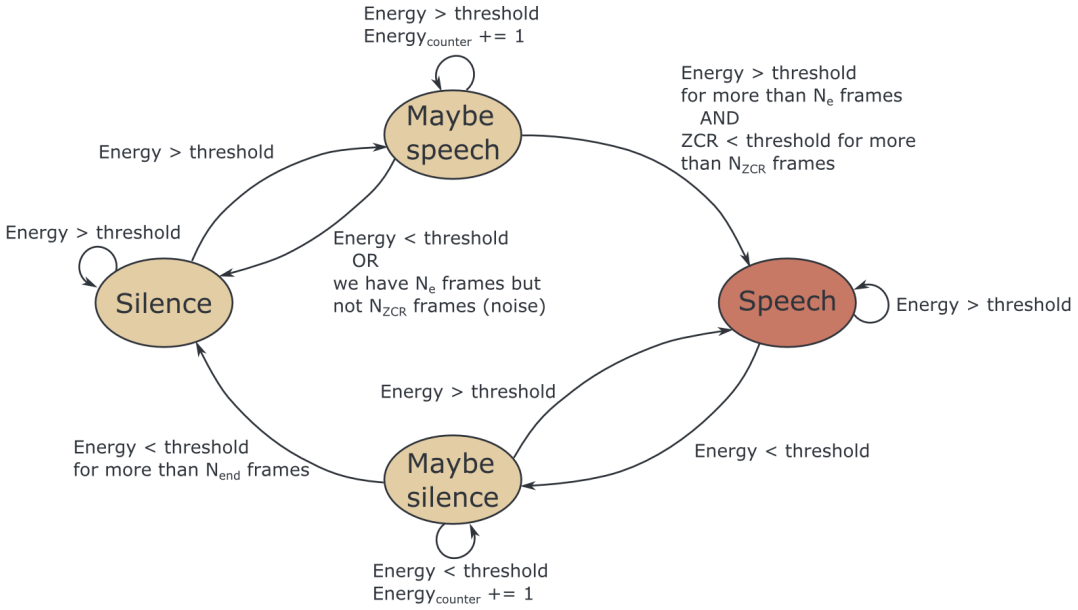

Figure 8.2 – Schéma de fonctionnement de l'algorithme de segmentation.

La calibration de cet algorithme n'est pas triviale car cinq paramètres doivent être réglés. Une procédure de calibration basée sur une recherche aléatoire dans l'espace des paramètres, suivant les récentes avancées en optimisation de paramètres [21], a été conçue. Elle utilise une mesure simple (pourcentage de recouvrement) pour déterminer la validité d'une segmentation correspondant à un jeu de paramètres, en comparai-

son avec une segmentation manuelle faisant office de *gold standard*. Cette procédure a l'avantage d'être rapide, mais la mesure utilisée ne pénalise pas des erreurs comme le fait de découper un segment de la référence en plusieurs petits segments. Les différentes erreurs de segmentation possibles sont représentées sur la figure 8.3.

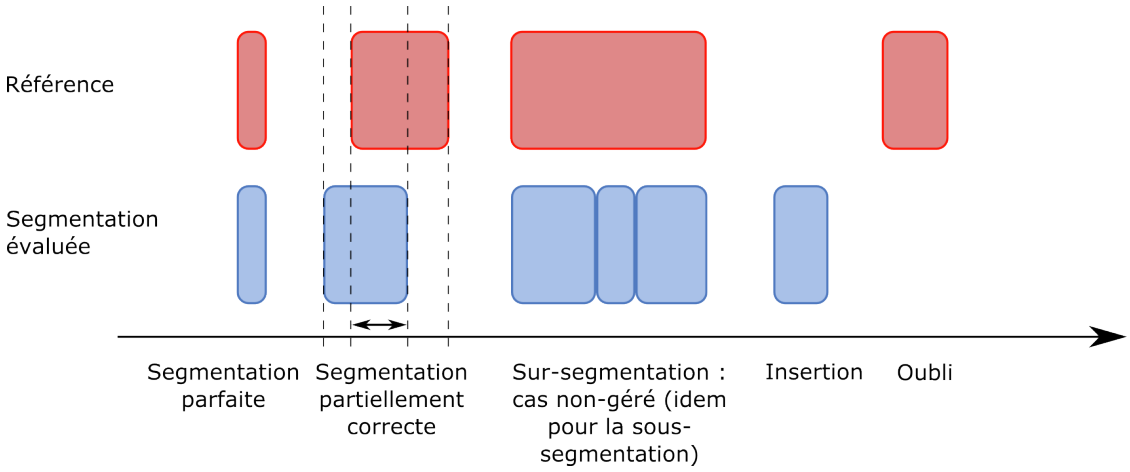

Figure 8.3 – Les différentes erreurs de segmentation possibles.

8.2.3 *Word-spotting*, grammaires et arbres de dialogue

Le module de *word-spotting* permet de comprendre les réponses de l'utilisateur. Un petit nombre de termes pertinents sont recherchés dans les réponses ; ces termes permettent ensuite de choisir la réponse correspondant le mieux dans l'arbre de dialogue, à l'étape donnée.

La première étape dans la construction de ce module est de rassembler le vocabulaire pertinent à partir des scénarios et des réponses des sujets pendant les collectes de données. Ce vocabulaire a ensuite servi à créer les grammaires, c'est-à-dire de représenter à la fois la forme et le contenu des réponses possibles pour pouvoir les comprendre. Ces grammaires sont encodées dans un langage du type expressions régulières sous un format XML puis compilées sous forme de fichiers binaires avec un outil de Voxler. Un exemple de règle pour une phrase de l'utilisateur se présentant est donné dans la figure 8.4 ci-dessous.

```
<Grammar ID="A015">
    <Rule Out="PRENOM">(Je (m'appelle | suis | me présente) | mon (nom | prénom) est)[$GARBAGE]</Rule>
    <Out>Rule.Out</Out>
</Grammar>
```

Figure 8.4 – Exemple de règle pour la présentation de l'utilisateur. Les alternatives au sein d'un groupe délimité par des parenthèses sont séparées par le caractère "barre verticale". Le placement de gauche à droite donne l'ordre séquentiel et donc la forme de la phrase. Le jeton "[$GARBAGE]" est utilisé pour n'importe quel contenu que l'on ne cherche pas à reconnaître, un peu comme le *wildcard* "*" en informatique. Cela signifie ici que le système ne comprend pas le prénom de l'utilisateur.

Les scénarios sont représentés en mémoire par un arbre de dialogue, encodé sous un format XML conçu spécialement avec Voxler. Une représentation de l'arbre de dialogue pour le scénario d'alerte est donnée sur la figure 8.5. Chaque noeud de l'arbre, correspondant à une phrase du système, est représenté par une boîte bleue. Le parcours de l'arbre, c'est-à-dire le choix de la prochaine phrase du système, dépend des réponses de l'utilisateur ; ce sont les transitions entre les noeuds, représentées par un trait horizontal sur la flèche reliant deux noeuds. À chaque phrase de transition correspond une règle de la grammaire.

En interaction, chaque transition d'un noeud se voit assigner un score de confiance en fonction du segment audio de l'utilisateur analysé ; celle avec le plus haut score est sélectionnée. Cependant, si le score de toutes les transitions est inférieur à un seuil particulier, le système dispose d'un mécanisme pour demander la répétition. Le comportement du système, incluant ce processus, est représenté sur la figure 8.6 ci-dessous.

8.2.4 Détection des émotions

L'organisation du module de détection d'émotions est décrite sur la figure 8.7. Une première version du module a utilisé des outils open-source : openEAR pour l'extraction de paramètres [98] et libSVM pour la classification [44]. Un modèle statistique de reconnaissance des émotions a été entraîné en utilisant les données des collectes du projet ARMEN et en les combinant avec le corpus JEMO.

Les paramètres acoustiques sont calculés pour chaque segment audio entrant ; ils

sont ensuite normalisés par rapport au modèle. Cette représentation numérique est comparée au modèle par le classifieur (SVM) qui prend une décision et assigne une étiquette émotionnelle au segment. Le processus complet est très rapide (inférieur à une demi-seconde).

8.2.5 AVE : contrôle et expressions

Nous avons utilisé la plate-forme MARC et plus particulière l'agent Mary (représenté sur la figure 8.8) pour l'AVE. La plate-forme interprète des instructions au format BML pour animer l'agent dans ses mouvements, expressions et pour déclencher les phrases prononcées par l'agent. Le logiciel MARC est exécuté en parallèle de notre système, éventuellement sur une machine distante. Un serveur UDP sert donc à la transmission des commandes BML.

La conception des expressions faciales a été une étape importante. Après plusieurs itérations, les expressions ont été validées et encodées sous forme de scripts BML. Quelques unes sont représentées sur la figure 8.8, exprimées par l'agent Mary.

Concernant les phrases prononcées par l'agent, il s'agit de phrases pré-enregistrées, générées par un programme de synthèse vocale. La synchronisation des lèvres (*lip-sync*) est faite en temps réel en analysant l'audio ; il s'agit d'une fonctionnalité de MARC.

8.3 Conclusion

Un prototype du système complet, incluant reconnaissance de mots-clés, arbre de dialogue, agent virtuel expressif et reconnaissance des émotions a donc été implémenté en collaboration avec les partenaires du projet ARMEN. L'intérêt de l'intégration de tous ces modules est de disposer d'un système permettant de collecter des données de manière automatique. Il pourra également être utilisé dans des études utilisateurs pour examiner plus précisément l'impact de l'agent virtuel dans son apparence ou de la stratégie dialogique globale du système.

Une prochaine étape, prévue dans l'architecture, permettra d'intégrer le système de dialogue à la plate-forme robotique, permettant ainsi une remontée d'informations sur la

localisation du robot ou l'état de la tâche entreprise. Des évaluations techniques du robot d'un point de vue navigation automatique et manipulation d'objets ont été effectuées dans un contexte clinique ; une suite possible du projet est l'évaluation clinique du robot avec le système d'interface complet.

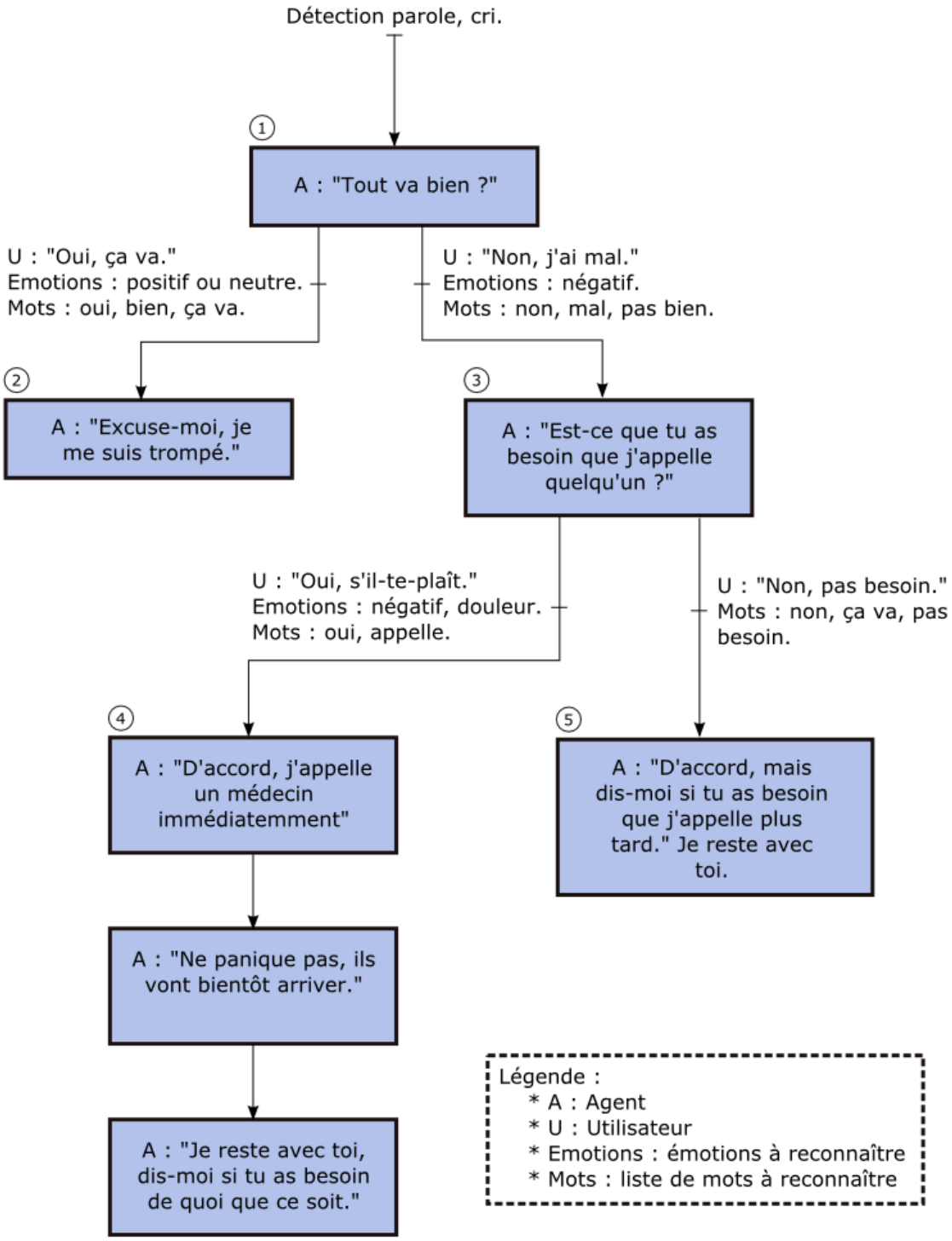

Figure 8.5 – Représentation graphique du scénario d'alerte.

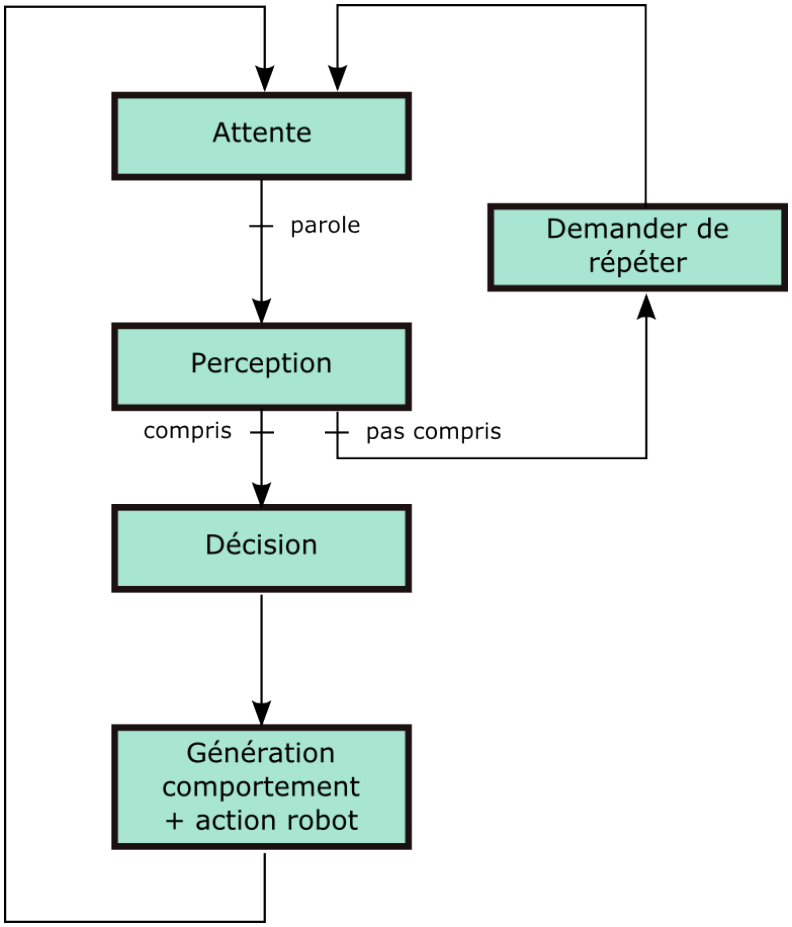

Figure 8.6 – Comportement du système en interaction.

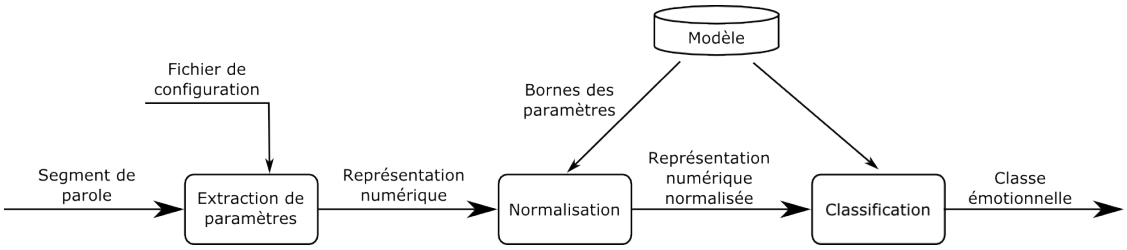

Figure 8.7 – Architecture du module de détection d'émotions.

Figure 8.8 – Capture d'écran de l'agent Mary, utilisant la plate-forme MARC, arborant plusieurs expressions (au centre, expression neutre ; à partir du haut et dans le sens des aiguilles d'une montre : angoisse, doute, curiosité, intérêt, expression positive, agacement).

CHAPITRE 9

VERS UNE MESURE OBJECTIVE DE L'ENGAGEMENT

Nous nous intéressons principalement à des interactions informelles avec les utilisateurs du système, sans véritable tâche à accomplir. À l'heure actuelle, il n'existe pas de méthode standard et automatique d'évaluation de telles interactions, particulièrement en ce qui concerne la perception de l'interaction, la satisfaction de l'utilisateur et son engagement avec le système.

Il existe plusieurs définitions d'"engagement" dans une interaction ; par exemple cela peut être compris comme une sensation de présence et l'impression d'être "captivé" par l'expérience [168]. Une définition plus complète envisage l'engagement comme le "processus par lequel deux participants ou plus établissent, maintiennent et terminent leur connexion perçue ; ce processus inclut le contact initial, la négociation d'une collaboration, la vérification que l'autre prend toujours part à l'interaction, la décision de rester impliqué ou non et de quand terminer la connexion" [240]. Dans les interactions en face-à-face, ce processus se manifeste à la fois dans la communication verbale et non-verbale [241]. Dans notre cas, nous sommes plutôt intéressés par l'évaluation de l'implication des interlocuteurs dans l'interaction et de leur envie de la maintenir.

Il existe un réel besoin en matière de mesure automatique de l'engagement. Nous décrivons ci-dessous une approche développée pour l'estimer automatiquement dans l'interaction à partir des signaux vocaux uniquement.

9.1 Description de l'approche

L'idée est de partir d'une évaluation perceptive de l'interaction par l'utilisateur et de tenter un "reverse-engineering" à partir de mesures objectives de bas-niveau effectuées sur le signal audio. Nous ne sommes pas vraiment intéressés par une évaluation en temps-réel dans un premier temps, mais plutôt par une évaluation globale, car l'engagement se construit tout au long de l'interaction. Cependant, il est possible d'étendre

cette approche au temps-réel pour disposer d'une métrique permettant de mieux gérer et planifier le déroulement de l'interaction.

Le moyen le plus direct d'obtenir l'avis des utilisateurs est simplement de leur demander. Cela a été fait sous forme de questionnaire pendant la collecte du corpus Armen2 : tous les sujets ont noté l'interaction et le personnage virtuel sur plusieurs critères, des adjectifs positifs ou négatifs, à l'aide d'une échelle de Likert à 5 niveaux. Tous les détails concernant le questionnaire et ses résultats sont disponibles dans la section 5.2.2.3.

Ensuite, des mesures ont été extraites du signal et de sa segmentation. On peut distinguer deux types de mesures : des mesures relatives au dialogue (temps de réponse, nombre de tours, etc) et des annotations émotionnelles. L'ensemble des mesures ainsi que les items des questionnaire utilisés par la suite est décrit dans le tableau 9.I.

Des caractéristiques des locuteurs (âge et sexe) ont également été utilisées, pour étudier les éventuelles différences. Les résultats de cette étude ont été présentés lors d'un workshop et publiés en tant que chapitre d'un livre[46, 47].

Nous avons trouvé des descriptions d'approches similaires dans la littérature, par exemple pour l'évaluation de systèmes basée sur l'expérience de l'utilisateur, dans le cas des systèmes de communication vocale du type VoIP [22], ou pour un système d'*e-learning* [83].

9.2 Résultats

Avec ces données disponibles, nous avons ensuite calculé leur matrice de corrélation croisée. Elle est représentée sur la figure 9.1 et les mesures précises sont rassemblées dans le tableau 9.II [1]. Nous avons sélectionné les corrélations significatives, c'est-à-dire ayant une valeur absolue supérieure à 0.4, la valeur critique pour un test bilatéral de signification statistique à $\alpha = 0.05$ étant de 0.396 dans notre cas. Ces valeurs sont

[1] Des identifiants numériques sont utilisés pour des raisons de compacité ; les correspondances se trouvent dans le tableau 9.I. Il faut également noter que seules les 14 premières colonnes de la matrice sont données car c'est principalement la corrélation entre les mesures objectives (indices 1 à 14) et celles provenant du questionnaire (indices 15 à 25) qui nous intéresse. De plus la matrice est symétrique, donc toutes les données peuvent être retrouvées.

Type	Mesure	ID
Dialogue	Nombre de tours	1
	Durée totale de parole	2
	Nombre de réponses avec un temps de réponse court (<1s)	3
	Temps de réponse moyen	4
	Temps total de superposition	5
Annotations émotionnelles	Valeur d'activation moyenne	6
	Valeur d'activation maximale	7
	Valeur d'activation minimale	8
	Proportion des segments annotés "colère"	9
	Proportion des segments annotés "joie"	10
	Proportion des segments annotés "neutre"	11
	Nombre de tours marqués "superposition"	12
Caractéristiques du locuteur	Âge	13
	Sexe (1 : masculin, 2 : féminin)	14
Questionnaire	Personnage virtuel perçu comme communicatif	15
	Personnage virtuel perçu comme bavard	16
	Personnage virtuel perçu comme sympathique	17
	Personnage virtuel perçu comme bizarre	18
	Personnage virtuel perçu comme amusant	19
	Personnage virtuel perçu comme intéressant	20
	Personnage virtuel perçu comme émotif	21
	Interaction perçue comme captivante	22
	Interaction perçue comme distrayante	23
	Interaction perçue comme répétitive	24
	Interaction perçue comme lente	25

Tableau 9.I – Mesures utilisées

marquées en gras dans le tableau 9.II.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.000	**0.693**	**0.575**	-0.157	**0.599**	**0.436**	**0.701**	-0.224	0.279	0.194	-0.184	**0.639**	0.273	0.161
2	**0.693**	1.000	0.356	-0.152	**0.572**	0.363	**0.507**	-0.105	0.056	0.239	-0.044	**0.589**	0.193	0.300
3	**0.575**	0.356	1.000	**-0.709**	0.204	**0.421**	**0.406**	-0.080	0.375	**0.460**	**-0.583**	0.206	0.116	0.155
4	-0.157	-0.152	**-0.709**	1.000	-0.140	-0.207	-0.012	-0.091	-0.217	-0.331	**0.506**	-0.129	0.008	-0.077
5	**0.599**	**0.572**	0.204	-0.140	1.000	**0.427**	0.376	-0.089	0.099	0.165	-0.259	**0.956**	**0.414**	0.289
6	**0.436**	0.363	**0.421**	-0.207	**0.427**	1.000	**0.644**	0.291	**0.514**	0.241	-0.361	**0.535**	0.062	-0.130
7	**0.701**	**0.507**	**0.406**	-0.012	0.376	**0.644**	1.000	-0.332	**0.461**	0.204	-0.286	**0.471**	0.310	-0.017
8	-0.224	-0.105	-0.080	-0.091	-0.089	0.291	-0.332	1.000	0.035	-0.152	-0.053	-0.120	**-0.478**	-0.145
9	0.279	0.056	0.375	-0.217	0.099	**0.514**	**0.461**	0.035	1.000	0.076	-0.328	0.149	-0.005	-0.266
10	0.194	0.239	**0.460**	-0.331	0.165	0.241	0.204	-0.152	0.076	1.000	**-0.400**	0.192	-0.172	0.198
11	-0.184	-0.044	**-0.583**	**0.506**	-0.259	-0.361	-0.286	-0.053	-0.328	**-0.400**	1.000	-0.210	-0.073	-0.115
12	**0.639**	**0.589**	0.206	-0.129	**0.956**	**0.535**	**0.471**	-0.120	0.149	0.192	-0.210	1.000	**0.408**	0.212
13	0.273	0.193	0.116	0.008	**0.414**	0.062	0.310	**-0.478**	-0.005	-0.172	-0.073	**0.408**	1.000	0.072
14	0.161	0.300	0.155	-0.077	0.289	-0.130	-0.017	-0.145	-0.266	0.198	-0.115	0.212	0.072	1.000
15	0.188	0.392	0.137	-0.080	0.163	**0.437**	0.157	0.163	0.254	**0.413**	0.068	0.317	-0.010	0.039
16	0.144	0.152	0.117	-0.183	0.201	**0.507**	0.190	0.131	**0.445**	0.301	-0.122	0.357	-0.133	-0.389
17	0.258	0.065	0.334	**-0.471**	0.298	0.275	0.009	0.340	0.168	0.256	-0.367	0.311	-0.202	-0.105
18	-0.201	-0.059	-0.245	**0.552**	-0.380	-0.085	0.092	-0.190	-0.149	-0.007	0.274	-0.379	-0.119	-0.074
19	-0.153	-0.176	0.180	-0.268	-0.091	0.182	-0.166	0.131	-0.028	**0.426**	-0.208	-0.047	**-0.464**	-0.042
20	-0.139	-0.321	-0.121	-0.028	0.181	-0.096	-0.202	0.179	**-0.410**	0.269	-0.262	0.109	-0.059	-0.025
21	-0.033	-0.186	0.080	0.072	-0.275	0.106	-0.016	0.222	0.019	0.229	0.060	-0.271	**-0.494**	**-0.420**
22	0.378	0.241	0.333	-0.229	0.285	0.369	0.090	0.306	0.274	0.211	-0.263	0.247	**-0.400**	-0.181
23	-0.076	-0.178	0.073	0.050	-0.159	0.297	0.068	0.094	0.288	0.336	-0.115	-0.052	**-0.433**	**-0.514**
24	-0.223	0.023	-0.203	0.326	-0.328	-0.341	-0.113	-0.153	-0.111	0.139	0.214	-0.354	**-0.507**	0.161
25	-0.076	0.098	0.098	0.063	0.073	0.243	0.060	0.287	**0.441**	-0.033	-0.155	-0.052	-0.156	0.122

Tableau 9.II – Matrice de corrélation croisée entre les différentes mesures relevées.

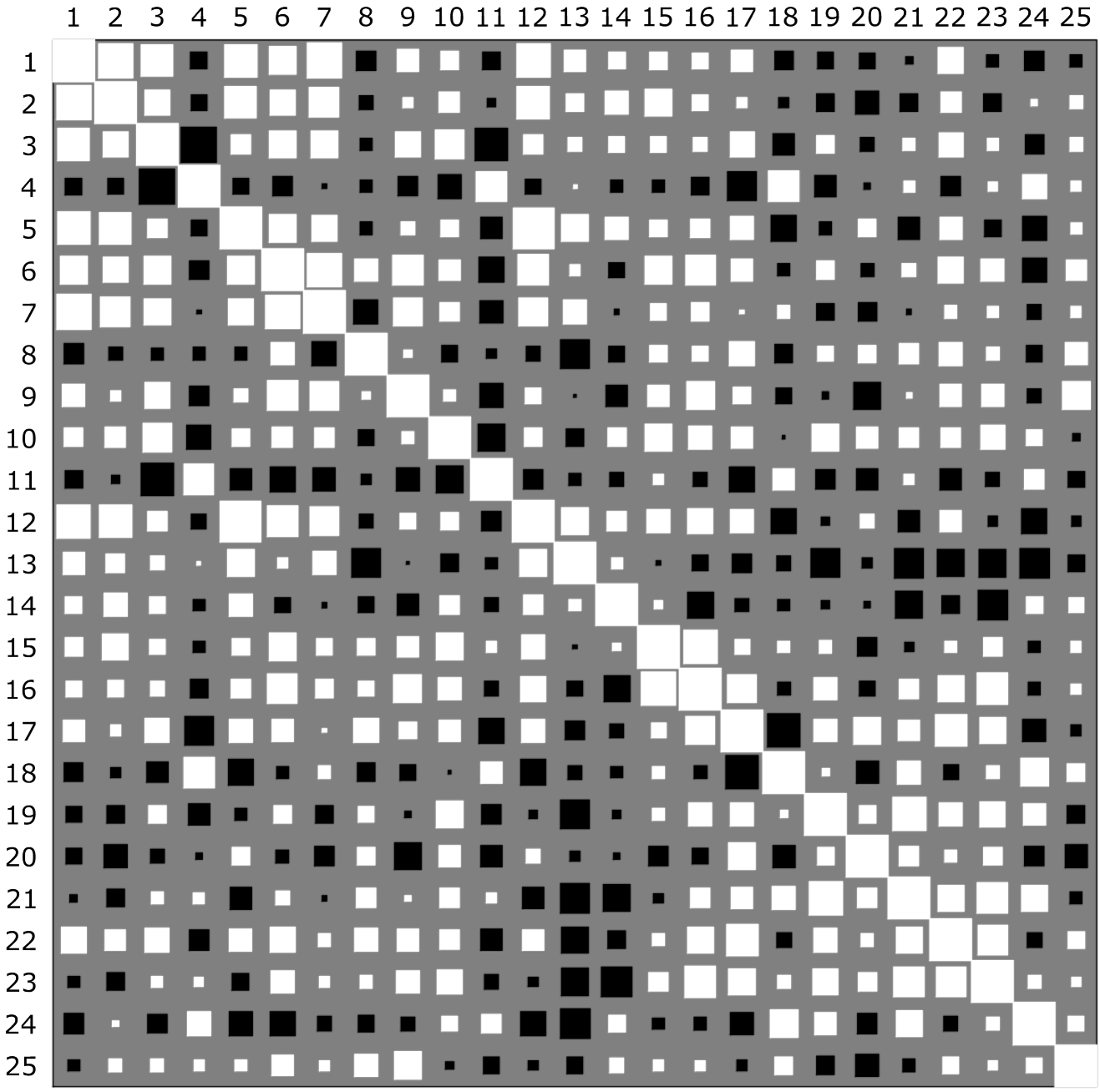

Figure 9.1 – Diagramme de Hinton pour la matrice de corrélation. L'aire de chaque carré est proportionnelle à la valeur absolue du coefficient de corrélation correpondant dans la matrice et la couleur correspond au signe (blanc pour positif, noir pour négatif).

On peut décrire certaines de ces corrélations car elles sont intéressantes. Tout d'abord des phénomènes connus sont confirmés par nos données. Par exemple, le nombre de tours de parole pour un locuteur donné est corrélé positivement avec le nombre de tours ayant un temps de réponse rapide (corrélation de 0.58). Par ailleurs, les locuteurs qui parlent plus et plus souvent génèrent un plus grand nombre de segments en superposition, c'est-à-dire qu'ils interrompent plus souvent le personnage virtuel (corrélation autour de 0.60). L'âge des locuteurs semble lié au nombre et à la durée de segments en superposition.

D'autres résultats révèlent des liens entre les annotations émotionnelles et les mesures dialogiques : les locuteurs qui expriment un niveau moyen d'activation plus élevé ont aussi un nombre de tours de parole et une durée totale de parole plus importants ; ils génèrent aussi un nombre plus élevé de segments en superposition et répondent plus fréquemment avec un temps de réponse court. Si l'on compare maintenant les valeurs d'activation annotées avec les réponses du questionnaire, on voit que les locuteurs avec ces caractéristiques trouvent également le personnage virtuel plus communicatif, bavard et moins ennuyant. L'âge ne paraît pas avoir beaucoup d'influence sur ce point (seule la valeur minimale d'activation est négativement corrélée à l'âge avec une valeur de 0.48) et le sexe pas du tout.

Par ailleurs, les locuteurs avec une proportion plus importante de segments annotés "colère" montrent des valeurs maximale et moyenne d'activation plus importantes ; ils perçoivent également le personnage virtuel comme plus communicatif mais moins intéressant, et l'interaction comme plus lente. Cela pourrait être une expression de frustration par rapport au système qu'ils jugeraient trop lent à répondre malgré son intérêt. Les locuteurs avec une proportion plus importante de segments annotés "joie" ont un temps de réponse plus rapide et trouvent le personnage virtuel plus communicatif et plus amusant. Au contraire, les locuteurs avec une proportion plus importante de segments annotés "neutre" ont un temps de réponse plus importante, traduisant peut-être leur manque d'enthousiasme pour l'intégration. De la même manière, en renversant la perspective, les locuteurs avec un temps de réponse plus long trouvent le personnage virtuel moins sympathique et plus bizarre.

Un point intéressant : l'âge des locuteurs est négativement corrélé avec la percep-

tion du personnage virtuel comme amusant et émotif, et l'interaction comme captivante et distrayante. Mais paradoxalement, l'âge est également négativement corrélé avec la perception de l'interaction comme répétitive. Il se pourrait que les personnes âgées sont plus sensibles à la nouveauté du système, tout en l'appréciant moins que les personnes plus jeunes qui ont a priori plus l'habitude de systèmes automatiques. Ce fait est clairement repérable sur la figure 9.2, où l'on a représenté l'évaluation de l'interaction d'après le questionnaire, en ventilant les résultats sur les deux centres médicaux impliqués dans l'expérience et pour lesquels les populations sont très différentes en termes d'âge (cf figure 9.3). Par contre, il existe très peu de différences par rapport au sexe des participants.

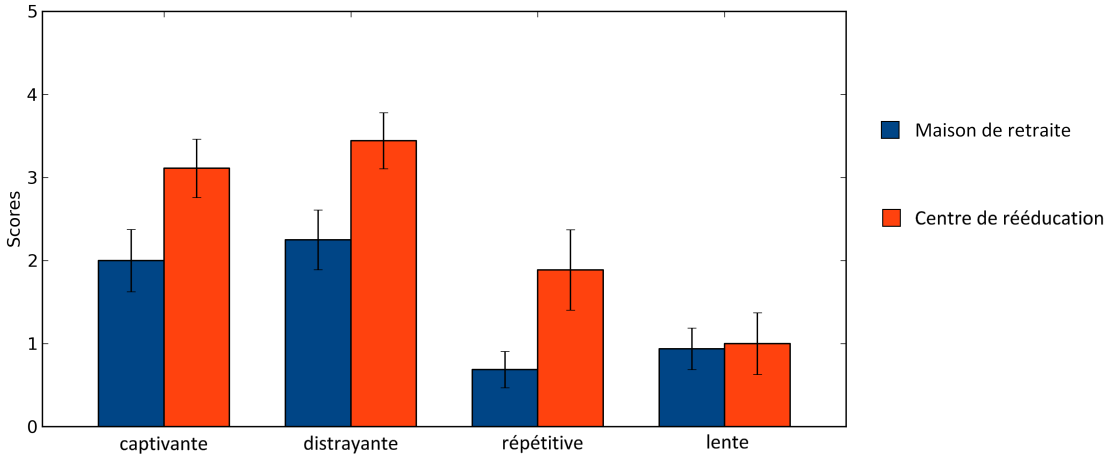

Figure 9.2 – Évaluation de la qualité de l'interaction par les participants à l'expérience.

Des tests de régression ont été effectuées pour tenter de prédire les réponses au questionnaire, mais il y a trop peu de données (25 patients) pour pouvoir donner des scores fiables.

9.3 Discussion

Cette étude a donc permis de montrer que les réponses au questionnaire sont assez corrélées à toutes les mesures qui ont été extraites : à la fois au niveau dialogique (temps de réponse, nombre de tours de paroles, superposition des voix), au niveau de l'émotion exprimée (activation et étiquettes émotionnelles). On a de plus observé qu'il existait des

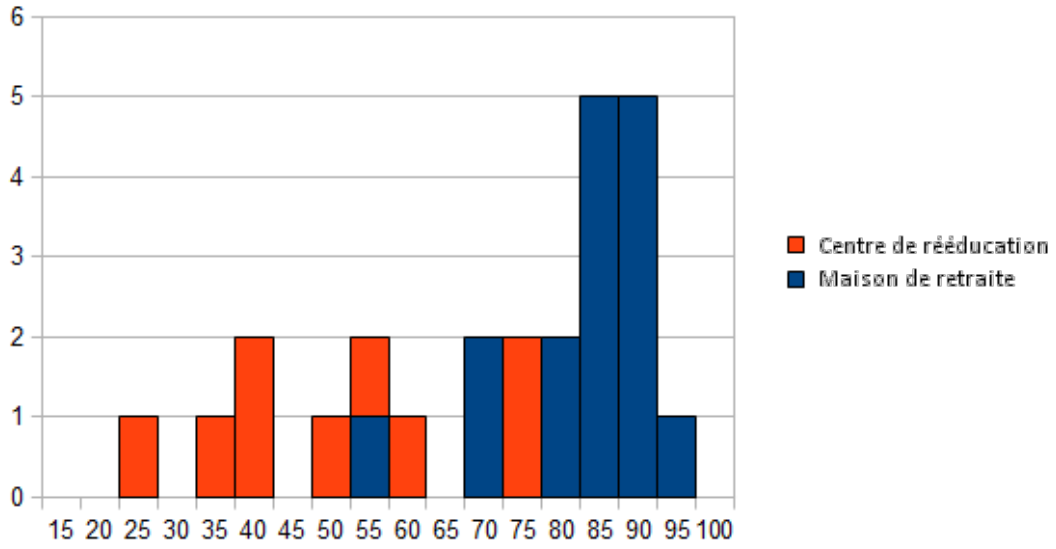

Figure 9.3 – Répartition des participants à l'expérience selon leur âge.

différences significatives selon l'âge des utilisateurs du système. Il faudrait donc prévoir pour ce genre de système des stratégies adaptées aux utilisateurs.

Même si cette étude est une preuve de concept, nous atteignons assez rapidement les limites de ce qui est réalisable avec le peu de données dont nous disposons. Il serait intéressant de continuer à explorer cette approche sur des corpus de données avec beaucoup plus de locuteurs.

CHAPITRE 10

CONCLUSION

10.1 Contexte et rappel des objectifs de recherche

Dans le cadre de cette thèse, je me suis intéressé à plusieurs besoins et challenges du domaine de l'*affective computing*, qui cherche notamment à dépasser les difficultés liées à la rareté des données émotionnelles, leur qualité, et tente d'avancer sur la compréhension des paramètres acoustiques permettant de décrire au mieux l'expression vocale émotionnelle. Des approches ont été imaginées et commencent à être explorées pour répondre à ces challenges, comme le recours à des techniques d'élicitation ou des montages de type "Magicien d'Oz" dans les protocoles de collecte de données ; la mise en commun des bases de données émotionnelles pour l'entraînement de systèmes de reconnaissance automatique des émotions (approches "cross-corpus") ; la sélection automatique de paramètres pertinents, technique issue du domaine du *machine learning* développée depuis plusieurs années. Par ailleurs, la détection des émotions s'inscrit plus largement dans le cadre d'interactions naturelles et affectives avec les machines, dont l'implémentation réussie intégrera à la fois des stratégies dialogiques, émotionnelles, empathiques et des interfaces expressives et engageantes sous la forme de robots ou agents virtuels.

Toutes ces problématiques se retrouvent dans les besoins du projet collaboratif ARMEN de l'ANR, qui fait office à la fois d'inspiration et de contexte applicatif pour les travaux développés au cours de cette thèse. Ce projet a pour but de développer un prototype de robot-assistant pour des personnes âgées et handicapées, destiné à les aider dans la vie quotidienne pour des tâches physiques qu'ils ne peuvent accomplir seuls. La nécessité d'une interaction simple et naturelle avec des utilisateurs non-experts rend alors évident le besoin d'une interface langagière, notamment pour expliquer les actions du robot. Des tests fonctionnels pour ARMEN et pour un projet ANR précédent, AVISO, ont d'ailleurs montré que les utilisateurs insistaient sur l'importance d'une interaction

plus conviviale. Un besoin de simplicité émerge donc. De plus, par la diversité des utilisateurs potentiels en termes d'âge, de qualité vocale et d'expression émotionnelle, ce projet s'inscrit dans la recherche de modèles statistiques optimisés et robustes, notamment en termes de paramètres utilisés.

10.2 Résumé des contributions

Pour répondre à ces challenges, j'ai développé mes contributions autour de grands axes.

Axe données émotionnelles : pour répondre à la spécificité du projet ARMEN en termes de population cible, deux corpus de données ont été collectés dans des centres médicaux avec des usagers potentiels du système. Des protocoles utilisant une approche "Magicien d'Oz" ont été conçus, organisés autour de scénarios écrits en collaboration avec l'association APPROCHE ; des logiciels de collecte de données ont été implémentés et j'ai participé à la collecte des données proprement dite. J'ai ensuite établi et supervisé les étapes de segmentation et d'annotation des données. Au total, 77 patients ont été interviewés, soit près de 27 heures d'enregistrements réparties sur 8 jours entre juin 2010 et juin 2011.

Axe optimisation de système de reconnaissance automatique des émotions : des approches "cross-corpus" ont été menées pour maximiser l'utilité des données collectées et de celles déjà en possession de l'équipe du thème *Dimensions Affectives et Sociales dans les Interactions Parlées* du LIMSI. Les études réalisées ont montré l'intérêt réel de ces approches, qui permettent d'atteindre une masse critique de données et partant, une simplification des modèles de détection ainsi qu'une amélioration du pouvoir de généralisation et de la robustesse. Parallèlement, des techniques de sélection de paramètres ont été explorées pour mieux comprendre quels paramètres sont plus pertinents pour un corpus ou un groupe d'utilisateurs donnés ; ces travaux ont donné lieu à une participation au Challenge InterSpeech 2012, avec une amélioration significative des scores par rapport au système basique proposé, et ce en utilisant une fraction des paramètres acoustiques. Un nouvel algorithme

de sélection a été développé, en s'inspirant de travaux précédemment menés sur la sélection flottante de paramètres notamment. Des tests ont montré l'équivalence ou la supériorité de cet algorithme par rapport aux algorithmes séquentiels existants (SFFS, SFFS-SSH) en termes de performance, de rapidité et d'évitement du sur-apprentissage. Il a ensuite été appliqué à la caractérisation de corpus différents.

Axe système de dialogue : une mesure d'engagement dans l'interaction à partir d'indices dialogiques et émotionnels de bas niveau a été explorée sur les données collectées. Un prototype du système de dialogue émotionnel a été développé en collaboration avec les équipes des thèmes *Dimensions Affectives et Sociales dans les Interactions Parlées* et *Agents Virtuels et Émotions* du LIMSI et la PME Voxler.

Ces contributions ont résulté en l'écriture de plusieurs articles acceptés en ateliers, conférences nationales et internationales, ainsi qu'un article de revue et un chapitre de livre, dont la liste est détaillée en Annexe I. J'ai participé à plusieurs conférences et ateliers pour exposer ces travaux sous forme de posters et présentations orales (cf Annexe II).

10.3 Perspectives - Discussion

Les approches menées sont prometteuses et elles nécessitent encore des efforts supplémentaires pour être validées. Les stratégies cross-corpus ont besoin d'être évaluées sur un nombre plus important de corpus. Bien qu'étant identifiées comme l'une des récentes tendances dans le domaine [236], elles restent encore assez limitées, l'état-de-l'art présenté sur le sujet dans le chapitre 6 en étant, à notre connaissance, une image quasi-exhaustive. Actuellement, le caractère privé des données lorsqu'elles sont spontanées ou collectées dans un contexte particulier est souvent un frein pour leur partage au sein de la communauté. Une prochaine étape pourrait passer par un partage des données, anonymisées et représentées sous forme de paramètres acoustiques supra-segmentaux, garantissant l'impossibilité de reconstruire les données audio de manière intelligible, devrait être possible ; l'établissement d'un standard pour l'ensemble de ces paramètres formant la représentation numérique ne devrait pas être trop compliquée, la plupart des

acteurs du domaine utilisant une base commune lors de la participation aux Challenges InterSpeech, fondée sur le schéma de codage du CEICES [16]. L'utilisation de données de synthèse, proposée comme alternative [226], me semble un recours moins intéressant quand des bases de données de qualité existent. Une banque de corpus de qualité disponibles pour la communauté, ainsi qu'un benchmark standardisé constitueraient un progrès indéniable.

Concernant l'annotation des données, la difficulté du processus a été illustrée (cf chapitre 5). La mise à contribution d'un grand nombre d'annotateurs non-experts (*crowdsourcing*) paraît être une piste encourageante [243, 244, 253] et son application a été étudiée récemment d'un point de vue éthique, économique et légal [3]. Couplée à des techniques d'apprentissage semi-supervisé, cela pourrait permettre de construire une base fiable pour des ressources immenses pour être entièrement annotées et, par effet de levier, de disposer d'une très grande quantité de données, potentiellement en perpétuelle expansion et rendant ainsi la collecte traditionnelle de données émotionnelles obsolète sauf pour des cas très spécifiques. L'exploration de techniques d'apprentissage en ligne, permettant aux modèles de continuer de se raffiner en mettant à profit les données rencontrées après leur déploiement, est également un sujet crucial [126].

Avec l'augmentation du nombre de corpus, des mesures de similarité deviendront nécessaires, pour juger par exemple de la complémentarité de ressources pour construire un système de reconnaissance dédié à une tâche particulière. Des premiers travaux ont exploré cette voie [32] et ont été appliqués dans cette thèse. Ils suggèrent que le problème est soluble, mais il demande encore à être validé sur plus de données.

Les méthodes de sélection de paramètres, après avoir eu bonne presse à la fin des années 1990 et au début des années 2000, ont paradoxalement souffert lors du *Feature Selection Challenge* de la conférence NIPS en 2003 [118]. En effet le gagnant du challenge avait utilisé une très forte proportion (environ 80%) des paramètres proposés et généralement les systèmes similaires ont obtenus de bons scores, ce qui a remis en question l'intérêt de la sélection. Cependant, de l'aveu même des organisateurs, la mesure utilisée pour classer les participations, incluant plusieurs critères comme le taux d'erreur et le pourcentage de paramètres utilisés, était trop biaisée en faveur de la perfor-

mance brute. Or les participants utilisant des algorithmes éliminant plus de paramètres obtiennent également des performances très intéressantes et ce avec une très faible proportion de paramètres (moins de 5%), ne sélectionnant que les paramètres pertinents et rejetant plus de 97% des faux paramètres délibérément introduits (*probes*). La performance des participants en développement et en test (en rouge et vert respectivement) en termes de taux d'erreur est illustrée sur la figure 10.1 ci-dessous, en fonction du pourcentage de paramètres sélectionnés. La recherche de la technique de sélection de paramètres parfaite est toujours ouverte ; on entrevoit cependant une résolution possible du problème de sur-adaptation des algorithmes de sélection aux données utilisées avec des approches cross-corpus et des méthodologies plus rigoureuses que celles parfois appliquées dans la littérature [214]. Concernant l'algorithme développé dans cette thèse, des tests plus avancés sont nécessaires pour son évaluation, en utilisant plus de données et des sondes (paramètres artificiels introduits délibérément).

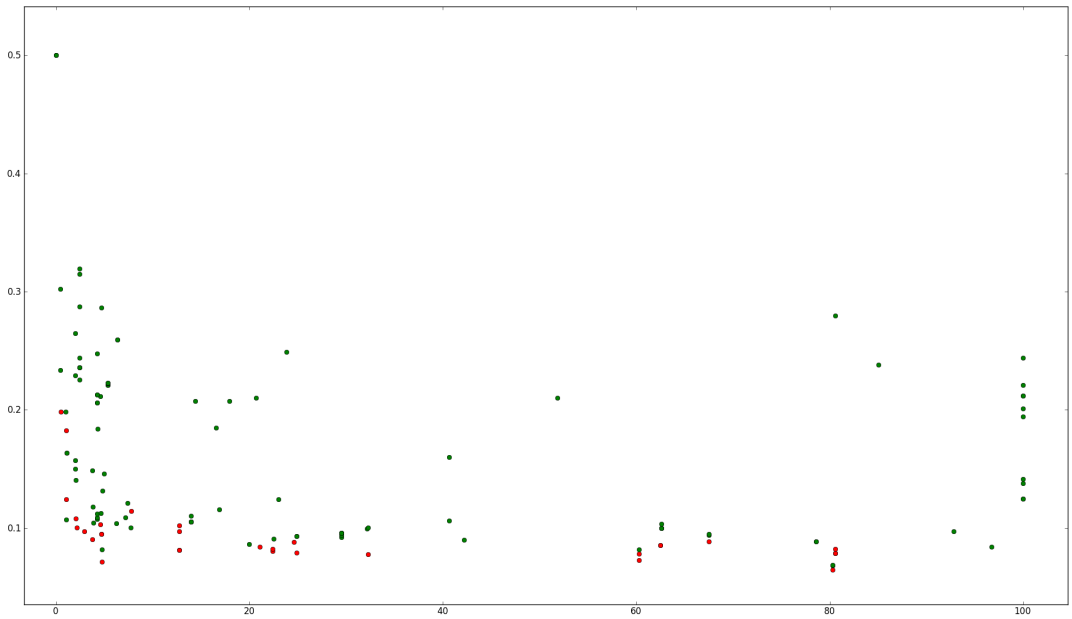

Figure 10.1 – Résumé des performances des participants du challenge de sélection de paramètres de la conférence NIPS 2003.

L'utilisation actuelle de l'agent virtuel MARC a été bien perçue par les utilisateurs.

Il reste cependant encore de nombreuses perspectives en termes de recherche sur les interactions affectives avec les agents virtuels concernant par exemple, la génération proprement dite d'expressions multimodales d'émotions (expressions faciales, parole, regard), ainsi que l'adaptation dynamique et personnalisée de ces expressions aux comportements des utilisateurs et à leurs émotions.

10.4 Conclusion générale

Les travaux développés dans cette thèse ont montré qu'il existait encore de nombreuses voies d'amélioration pour la conception de systèmes de reconnaissance d'émotions dans la voix performants et des systèmes de dialogue affectif. Si nous nous sommes limités à l'expression vocale des émotions, il existe de nombreux autres champs de recherche dans le domaine de l'*affective computing*, ne serait-ce que pour la détection d'émotions : à partir du vocabulaire utilisé, des expressions faciales, de la postures et des gestes, ou encore de variables physiologiques, toutes ces modalités étant utilisées à différents niveaux par les humains dans la communication interpersonnelle. Pour être jugées aussi intelligentes qu'un humain dans une interaction naturelle, les machines devront donc savoir les exploiter et les coordonner efficacement. Des versions émotionnelles du test de Turing ont d'ailleurs été imaginées [212]. Quoiqu'il en soit, la conception de ces machines ne pourra se faire qu'en étant centrée sur des cas d'utilisation et surtout des utilisateurs réels, la maxime de Protagoras, philosophe précurseur de Socrate, *"L'homme est la mesure de tout chose"*, s'appliquant ici particulièrement bien.

Pour finir, il me paraît indispensable de prendre du recul. Les machines sont en voie d'acquérir des capacités de plus en plus importantes (reconnaissance des personnes, du langage, des émotions, établissement de profils individuels, etc) et de se démocratiser, touchant une population de plus en plus large. Ces capacités, comme toute technologie pervasive et puissante, font naître des interrogations d'ordre éthique. Malheureusement, comme cela a été souligné dans la littérature [167], ce problème de l'éthique est souvent abordé par le biais des *lois d'Asimov*, issues de ses romans célèbres, ou par spéculation des usages futurs car la réalité de la robotique, avec les aspirateurs automatiques

Roomba, est souvent loin de l'imaginaire développé dans la littérature de science-fiction et le cinéma à grand spectacle. Or certains domaines, comme celui des robots-assistants, encouragés par les financements publics et privés, évoluent très rapidement. Le caractère personnel de telles machines, acquis en rentrant dans la sphère privée, donne naissance à des problèmes très concrets qui peuvent être étudiés dès maintenant : en mémorisant des informations à caractère privé voire intime, ce qui sera très probablement le cas vu le comportement des sujets interviewés lors des collectes de données menées lors de cette thèse, faisant parfois de véritables confessions à un système pourtant très simple, des problèmes d'anonymisation, de sécurisation et de protection des données émergent. Par ailleurs, la robotique d'assistance, poussée pour des raisons de coûts, de démographie ou d'indisponibilité de personnel qualifié, pose très fortement la question de la disparition de la relation humaine avec les membres les plus fragiles de la société, ainsi que des interrogations sur les droits de l'homme, le potentiel usage de la contrainte physique par les robots, la responsabilité au yeux de la loi et plus généralement le bien-être physique et psychologique des utilisateurs. Ces questions deviennent encore plus prégnantes pour des applications comme la garde d'enfants ou la prise en charge d'une partie de l'éducation, comme c'est déjà le cas en Corée par exemple [119]. Des structures spéciales de réflexion sur l'éthique et les usages de la technologie, en particulier la robotique personnelle, existent, notamment en France avec la création du CERNA (Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene[1]) et de comités dédiés aux nouvelles technologies de l'information au CNRS et à l'INRIA [86, 170]. Cependant il paraît nécessaire de poser le problème sous la forme d'un débat de société et de travailler à la vulgarisation de ces enjeux.

[1] `https://www.allistene.fr/cerna-2/`

Bibliographie

BIBLIOGRAPHIE

[1] Sarkis Abrilian, Laurence Devillers et Jean-Claude Martin. Emotv1 : Annotation of emotions in real-life video interviews : Variability between coders. Dans *5th Int. Conf. on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, 2006.

[2] Yaser S. Abu-Mostafa, Malik Magdon-Ismail et Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN 1600490069, 9781600490064.

[3] Gilles Adda, Joseph J Mariani, Laurent Besacier et Hadrien Gelas. Economic and ethical background of crowdsourcing for speech. *Crowdsourcing for Speech Processing : Applications to Data Collection, Transcription and Assessment*, pages 303–334, 2013.

[4] Vamshi Ambati, Stephan Vogel et Jaime G Carbonell. Active learning and crowdsourcing for machine translation. Dans *LREC*, volume 11, pages 2169–2174, 2010.

[5] Yasmine Arafa et Abe Mamdani. Scripting embodied agents behaviour with cml : character markup language. Dans *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 313–316. ACM, 2003.

[6] Nicolas Audibert, Véronique Aubergé et Albert Rilliard. Ewiz : contrôle d'émotions authentiques. *Actes des Journées d'Étude sur la Parole*, pages 49–52, 2004.

[7] Marie Avril, Mohamed Chetouani et Nicolas Sabouret. Étude d'une collaboration socio-affective entre une personne et le robot jazz. *Revue Interfaces Numériques*, 2(1):57–76, 2013.

[8] Rainer Banse et Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[9] Claude Barras, Edouard Geoffrois, Zhibiao Wu et Mark Liberman. Transcriber : a free tool for segmenting, labeling and transcribing speech. Dans *First international conference on language resources and evaluation (LREC)*, pages 1373–1376, 1998.

[10] Anton Batliner, Jan Buckow, Richard Huber, Volker Warnke, Elmar Nöth et Heinrich Niemann. Prosodic feature evaluation : brute force or well designed. Dans *Proc. 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, 1999.

[11] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker et Elmar Nöth. Desperately seeking emotions or : Actors, wizards, and human beings. Dans *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[12] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker et Elmar Nöth. How to find trouble in communication. *Speech communication*, 40(1):117–143, 2003.

[13] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell et Michael Wong. "you stupid tin box" - children interacting with the aibo robot : A cross-linguistic emotional speech corpus. Dans *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, 2004.

[14] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth et Jürgen Haas. From emotion to interaction : lessons from real human-machine-dialogues. Dans *Affective Dialogue Systems*, pages 1–12. Springer, 2004.

[15] Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson et Noam Amir. The automatic recognition of emotions in speech. Dans *The HUMAINE Handbook (Cognitive Technologies)*, pages 71–99. Springer, 2010.

[16] Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous et al. Whodunnit–searching for the most important feature types signalling

emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28, 2011.

[17] Charles Bell. *The Anatomy and Philosophy of Expression*. Third édition, 1844.

[18] Richard Bellman. *Adaptive control processes : a guided tour*, volume 4. Princeton University Press, Princeton, 1961.

[19] Catherine Belzung. *Biologie des émotions*. De Boeck, Bruxelles, Belgique, 2007.

[20] Jeremy Bentham. *An introduction to the principles of morals and legislation*. Clarendon Press, 1879. Original work published in 1780.

[21] James Bergstra et Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281–305, 2012.

[22] Abhishek Bhattacharya, Wanmin Wu et Zhenyu Yang. Quality of experience evaluation of voice communication : an affect-based approach. *Human-centric Computing and Information Sciences*, 2(1):1–18, 2012.

[23] Timothy Bickmore et Daniel Schulman. A virtual laboratory for studying long-term relationships between humans and virtual agents. Dans *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 297–304. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[24] Timothy Bickmore, Daniel Schulman et Langxuan Yin. Engagement vs. deceit : Virtual humans with human autobiographies. Dans *Intelligent Virtual Agents*, pages 6–19. Springer, 2009.

[25] Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr et Tim Heeren. "it's just like you talk to a friend" : relational agents for older adults. *Interacting with Computers*, 17(6):711–735, 2005.

[26] Frances R. Bilous et Robert M. Krauss. Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*, 8(3):183–194, 1988.

[27] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2002.

[28] Cynthia Breazeal. Role of expressive behaviour for robots that learn from people. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 364 (1535):3527–3538, 2009.

[29] Cynthia L Breazeal. *Designing Sociable Robots*. MIT Press, Cambridge, MA, 2002.

[30] Mátyás Brendel, Riccardo Zaccarelli et Laurence Devillers. Building a system for emotions detection from speech to control an affective avatar. Dans *Proceedings of LREC 2010*, pages 2205–2210, 2010.

[31] Mátyás Brendel, Riccardo Zaccarelli et Laurence Devillers. A quick sequential forward floating feature selection algorithm for emotion detection from speech. Dans *INTERSPEECH*, pages 1157–1160, 2010.

[32] Mátyás Brendel, Riccardo Zaccarelli, Björn Schuller et Laurence Devillers. Towards measuring similarity between emotional corpora. Dans *Proc. 3rd ELR0A Internat. Workshop on EMOTION*, pages 58–64, 2010.

[33] Joost Broekens, Marcel Heerink et Henk Rosendal. Assistive social robots in elderly care : a review. *Gerontechnology*, 8(2):94–103, 2009.

[34] Cyril Brom et J Lukavsky. Towards virtual characters with a full episodic memory ii : The episodic memory strikes back. Dans *Proc. Empathic Agents, AAMAS workshop*, pages 1–9, 2009.

[35] Fabio Brugnara, Daniele Falavigna et Maurizio Omologo. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12(4):357–370, 1993.

[36] Michael Buhrmester, Tracy Kwang et Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data ? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[37] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[38] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann et Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. Dans *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[39] Zoraida Callejas et Ramón López-Cózar. Improving acceptability assessment for the labelling of affective speech corpora. Dans *INTERSPEECH*, pages 2863–2866, 2009.

[40] Joana Campos et Ana Paiva. May : my memories are yours. Dans *Intelligent Virtual Agents*, pages 406–412. Springer, 2010.

[41] Zoraida Callejas Carrión. *On the development of Adaptive and Portable Spoken Dialogue Systems : Emotion Recognition, Language Adaptation and Field Evaluation.* Thèse de doctorat, Universidad de Granada, 2008.

[42] Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000.

[43] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost et Matthew Stone. Animated conversation : rule-based generation of facial expression, gesture & spoken

intonation for multiple conversational agents. Dans *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM, 1994.

[44] Chih-Chung Chang et Chih-Jen Lin. LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27), 2011.

[45] Patrick Charaudeau. *Grammaire du sens et de l'expression*. Hachette, 1992.

[46] Clément Chastagnol, Céline Clavel, Matthieu Courgeon et Laurence Devillers. Designing an emotion detection system for a socially intelligent human-robot interaction. Dans *Proceedings of the International Worshop on Spoken Dialog Systems (IWSDS 2012)*, 2012.

[47] Clément Chastagnol, Céline Clavel, Matthieu Courgeon et Laurence Devillers. Designing an emotion detection system for a socially intelligent human-robot interaction. Dans Joseph Mariani, Sophie Rosset, Martine Garnier-Rizet et Laurence Devillers, éditeurs, *Natural Interaction with Robots, Knowbots and Smartphones*, pages 199–211. Springer New York, 2014.

[48] Clément Chastagnol et Laurence Devillers. Analysis of anger across several agent-customer interactions in french call centers. Dans *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4960–4963. IEEE, 2011.

[49] Clément Chastagnol et Laurence Devillers. Collecting spontaneous emotional data for a social assistive robot. Dans *Proceedings of ES^3 2012 workshop, as part of LREC 2012*, 2012.

[50] Clément Chastagnol et Laurence Devillers. Personality traits detection using a parallelized modified SFFS algorithm. 2012.

[51] Herbert H. Clark. *Using language*, volume 23. Cambridge University Press, 1996.

[52] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[53] Paul R Cohen, David Jensen et al. Overfitting explained. Dans *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 115–122, 1997.

[54] William S. Condon et William D. Ogston. A segmentation of behavior. *Journal of psychiatric research*, 5(3):221–235, 1967.

[55] Matthieu Courgeon, Stéphanie Buisine et Jean-Claude Martin. Impact of expressive wrinkles on perception of a virtual character's facial expressions of emotions. Dans *Intelligent Virtual Agents*, pages 201–214. Springer, 2009.

[56] Matthieu Courgeon et Céline Clavel. Marc : a framework that features emotion models for facial animation during human–computer interaction. *Journal on Multimodal User Interfaces*, pages 1–9, 2013.

[57] Matthieu Courgeon, Céline Clavel et Jean-Claude Martin. Appraising emotional events during a real-time interactive game. Dans *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. ACM, 2009.

[58] Matthieu Courgeon, Jean-Claude Martin et Christian Jacquemin. Marc : a multimodal affective and reactive character. Dans *Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments*, 2008.

[59] Roddy Cowie et Randolph R Cornelius. Describing the emotional states that are expressed in speech. *Speech communication*, 40(1):5–32, 2003.

[60] Roddy Cowie, Ellen Douglas-Cowie, Jean-Claude Martin et Laurence Devillers. The essential role of human databases for learning in and validation of affectively competent agents. Dans K. Scherer, T. Bänziger et E. Roach, éditeurs, *A Blueprint for Affective Computing : a Sourcebook and Manual*, pages 151–165. Oxford University Press, 2010.

[61] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey et Marc Schröder. 'feeltrace' : An instrument for recording perceived emotion in real time. Dans *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[62] Richard Coyne. Computers, metaphors and change. *Architectural Research*, 97, 1995.

[63] Pádraig Cunningham. Dimension reduction. Dans *Machine learning techniques for multimedia*, pages 91–112. Springer, 2008.

[64] Charles Darwin. *The expression of emotions in man and animals*. 1872.

[65] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay et Iain Werry. What is a robot companion-friend, assistant or butler ? Dans *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1192–1197. IEEE, 2005.

[66] Frederique De Vignemont et Tania Singer. The empathic brain : how, when and why ? *Trends in cognitive sciences*, 10(10):435–441, 2006.

[67] Agnes Delaborde et Laurence Devillers. Use of nonverbal speech cues in social interaction between human and robot : emotional and interactional markers. Dans *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pages 75–80. ACM, 2010.

[68] Émilie Delaherche et Mohamed Chetouani. Multimodal coordination : exploring relevant features and measures. Dans *Proceedings of the 2nd international workshop on Social signal processing*, pages 47–52. ACM, 2010.

[69] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux et David Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3):349–365, 2012.

[70] Frank Dellaert, Thomas Polzin et Alex Waibel. Recognizing emotion in speech. Dans *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[71] Eugene d'Eon, David Luebke et Eric Enderton. Efficient rendering of human skin. Dans *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 147–157. Eurographics Association, 2007.

[72] René Descartes. *Les passions de l'âme*. 1649.

[73] L. Devillers, R. Cowie, J.C. Martin, E. Douglas-Cowie, S. Abrilian et M. McRorie. Real life emotions in french and english tv video clips : an integrated annotation protocol combining continuous and discrete approaches. Dans *5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, 2006.

[74] Laurence Devillers. Les émotions dans les interactions homme-machine : perception, détection et génération, 2006. Habilitation à Diriger des Recherches en Informatique de l'Université Paris XI - Orsay, soutenue le 4 décembre 2006 au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, LIMSI-CNRS (UPR3251).

[75] Laurence Devillers et Jean-Claude Martin. Corpus émotionnels : de l'acquisition à la modélisation. Dans Catherine Pelachaud, éditeur, *Émotions*. Hermès, 2010.

[76] Laurence Devillers et Ioana Vasilescu. Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. Dans *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[77] Laurence Devillers, Iona Vasilescu et Laurence Vidrascu. Anger versus fear detection in recorded conversations. Dans *Proceedings of Speech Prosody*, pages 205–208, 2004.

[78] Laurence Devillers, Christophe Vaudable et Clément Chastagnol. Real-life emotion-related states detection in call centers : a cross-corpora study. Dans *INTERSPEECH*, pages 2350–2353, 2010.

[79] Laurence Devillers et Laurence Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. Dans *INTERSPEECH*, pages 801–803, 2006.

[80] Laurence Devillers, Laurence Vidrascu et Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Journal of Neural Networks*, 18(4):407–422, 2005.

[81] Laurence Devillers, Laurence Vidrascu et Omar Layachi. Automatic detection of emotion from vocal expression. pages 232–244. Oxford University Press, 2010.

[82] Thomas Dixon. "emotion" : The history of a keyword in crisis. *Emotion Review*, 4(4):338–344, 2012.

[83] Sidney K D'Mello, Scotty D Craig, Amy Witherspoon, Bethany Mcdaniel et Arthur Graesser. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1-2):45–80, 2008.

[84] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[85] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie et Peter Roach. Emotional speech : Towards a new generation of databases. *Speech communication*, 40(1): 33–60, 2003.

[86] Gilles Dowek, David Guiraud, Claude Kirchner, Daniel Le Métayer et Pierre-Yves Oudeyer. Rapport sur la création d'un comité d'éthique en sciences et technologies du numérique. Rapport technique, Groupe de réflexion sur la création potentielle d'un comité d'éthique à l'INRIA, 2009.

[87] Guillaume-Benjamin Duchenne. *Mécanisme de la physionomie humaine, ou Analyse électro-physiologique de ses différents modes d'expression, par le Dr Duchenne (de Boulogne). Mémoire communiqué aux Académies des sciences et de médecine*. Asselin, 1862.

[88] Starkey Duncan et George Niederehe. On signalling that it's your turn to speak. *Journal of experimental social psychology*, 10(3):234–247, 1974.

[89] P Ekman. Are there basic emotions ? *Psychol Rev*, 99(3):550–553, 1992.

[90] Paul Ekman. *Emotions revealed : Recognizing faces and feelings to improve communication and emotional life.* Times Books/Henry Holt and Co, 2003.

[91] Paul Ekman et Wallace V Friesen. Facial action coding system : A technique for the measurement of facial movement. palo alto. *CA : Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From appraisal to emotion : Differences among unpleasant feelings. Motivation and Emotion*, 12:271–302, 1978.

[92] Wallace V. ; Ellsworth Phoebe Ekman, Paul ; Friesen. *Emotion in the human face : Guidelines for research and an integration of findings*. Pergamon Press, Oxford, England, 1972.

[93] Phoebe C Ellsworth et Klaus R Scherer. Appraisal processes in emotion. Dans R.J. Davidson, K.R. Scherer et H.H. Goldsmith, éditeurs, *Handbook of affective sciences*, pages 572–595. Oxford University Press, New York, 2003.

[94] Inger S Engberg, Anya Varnich Hansen, Ove Andersen et Paul Dalsgaard. Design, recording and verification of a danish emotional speech database. Dans *Eurospeech*, 1997.

[95] Sibylle Enz, Martin Diruf, Caroline Spielhagen, Carsten Zoll et Patricia A Vargas. The social role of robots in the futureâĂŤexplorative measurement of hopes and fears. *International Journal of Social Robotics*, 3(3):263–271, 2011.

[96] Florian Eyben, Anton Batliner, Björn Schuller, Dino Seppi et Stefan Steidl. Cross-corpus classification of realistic emotions some pilot experiments. Dans *Proc. 3rd International Workshop on EMOTION (satellite of LREC) : Corpora for Research on Emotion and Affect, Valetta*, pages 77–82, 2010.

[97] Florian Eyben, Felix Weninger et Björn Schuller. Affect recognition in real-life acoustic conditions - a new perspective on feature selection. Dans *to appear in Proc. INTERSPEECH 2013*, 2013.

[98] Florian Eyben, Martin Wollmer et Bjorn Schuller. Openear - introducing the munich open-source emotion and affect recognition toolkit. Dans *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[99] David Feil-Seifer et Maja J Mataric. Defining socially assistive robotics. Dans *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, pages 465–468. IEEE, 2005.

[100] David Feil-Seifer, Kristine Skinner et Maja J Mataric. Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423–439, 2007.

[101] Ylva Fernaeus, Maria Håkansson, Mattias Jacobsson et Sara Ljungblad. How do you play with a robotic toy animal ? : a long-term study of pleo. Dans *Proceedings of the 9th international Conference on interaction Design and Children*, pages 39–48. ACM, 2010.

[102] Rebecca Fiebrink et Ichiro Fujinaga. Feature selection pitfalls and music classification. Dans *ISMIR*, pages 340–341. Citeseer, 2006.

[103] Susan T Fiske. *Social beings : Core motives in social psychology*. Wiley, 2009.

[104] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971.

[105] Terrence Fong, Illah Nourbakhsh et Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[106] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch et Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12): 1050–1057, 2007.

[107] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman et M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *Biomedical Engineering, IEEE Transactions on*, 47(7):829–837, 2000.

[108] Nico H Frijda. *The emotions*. Cambridge University Press, 1986.

[109] Martine Garnier-Rizet, Gilles Adda, Frederik Cailliau, Jean-Luc Gauvain, Sylvie Guillemin-Lanne, Lori Lamel, Stephan Vanni, Claire Waast-Richard et al. Callsurf : Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. Dans *LREC*, 2008.

[110] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups : A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[111] Maria Gendron et Lisa Feldman Barrett. Reconstructing the past : A century of ideas about emotion in psychology. *Emotion Review*, 1(4):316–339, 2009.

[112] Howard Giles. Accent mobility : A model and some data. *Anthropological linguistics*, 15(2):87–105, 1973.

[113] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz et al. Designing robots for long-term social interaction. Dans *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1338–1343. IEEE, 2005.

[114] Marjorie H Goodwin et Charles Goodwin. Emotion within situated activity. *Communication : An arena of development*, pages 33–53, 2000.

[115] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg et Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. Dans *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages 1033–1036. IEEE, 2006.

[116] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast et Robin Duffy. Creating rapport with virtual agents. Dans *Intelligent Virtual Agents*, pages 125–138. Springer, 2007.

[117] Isabelle Guyon et André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[118] Isabelle Guyon, Steve Gunn, Asa Ben-Hur et Gideon Dror. Result analysis of the nips 2003 feature selection challenge. Dans *Advances in Neural Information Processing Systems*, pages 545–552, 2004.

[119] Jeonghye Han et Dongho Kim. r-learning services for elementary school students with a teaching assistant robot. Dans *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 255–256. IEEE, 2009.

[120] Ursula Hess et Pascal Thibault. Darwin and emotion expression. *American Psychologist*, 64(2):120, 2009.

[121] Geoffrey E Hinton, Simon Osindero et Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[122] Wan Ching Ho et Kerstin Dautenhahn. Towards a narrative mind : The creation of coherent life stories for believable virtual agents. Dans *Intelligent Virtual Agents*, pages 59–72. Springer, 2008.

[123] Wan Ching Ho, Kerstin Dautenhahn, Mei Yii Lim, Patrícia Amâncio Vargas, Ruth Aylett et Sibylle Enz. An initial memory model for virtual and robot companions supporting migration and long-term interaction. Dans *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 277–284. IEEE, 2009.

[124] Mohammed Ehsan Hoque, Matthieu Courgeon, J Martin, Bilge Mutlu et Rosalind W Picard. Mach : My automated conversation coach. Dans *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013)*, 2013.

[125] Chih-Wei Hsu et Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[126] Chengwei Huang, Ruiyu Liang, Qingyun Wang, Ji Xi, Cheng Zha et Li Zhao. Practical speech emotion recognition based on online learning : From acted data to elicited data. *Mathematical Problems in Engineering*, 2013, 2013.

[127] Carroll E Izard. *Human emotions*. New York : Plenum Press, 1977.

[128] Carroll E Izard. More meanings and more questions for the term "emotion". *Emotion review*, 2(4):383–385, 2010.

[129] Anil Jain et Douglas Zongker. Feature selection : Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.

[130] Hazaël Jones et Nicolas Sabouret. An affective model for a virtual recruiter in a job interview context. *Procedia Computer Science*, 15:312–313, 2012.

[131] Patrik N Juslin et Petri Laukka. Communication of emotions in vocal expression and music performance : Different channels, same code ? *Psychological bulletin*, 129(5):770–814, 2003.

[132] Garrett D Kearney et Sati McKenzie. Machine interpretation of emotion : Design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions. *Cognitive Science*, 17(4):589–622, 1993.

[133] Sathiya Keerthi, Vikas Sindhwani et Olivier Chapelle. An efficient method for gradient-based adaptation of hyperparameters in svm models. Dans *Advances in Neural Information Processing Systems*, volume 19, pages 673–680. Schölkopf, B. and Platt, J. and Hoffman, T., Cambridge, MA, 2007.

[134] JF Kelley. An empirical methodology for writing user-friendly natural language computer applications. Dans *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 193–196. ACM, 1983.

[135] Thomas Kemp, Michael Schmidt, Martin Westphal et Alex Waibel. Strategies for automatic segmentation of audio data. Dans *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1423–1426. IEEE, 2000.

[136] Theodore D Kemper. Social constructionist and positivist approaches to the sociology of emotions. *American Journal of Sociology*, pages 336–362, 1981.

[137] Christian Keysers et Valeria Gazzola. Social neuroscience : mirror neurons recorded in humans. *Current Biology*, 20(8):353–354, 2010.

[138] Cory D Kidd et Cynthia Breazeal. A robotic weight loss coach. Dans *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 22, page 1985. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999, 2007.

[139] Cory D Kidd et Cynthia Breazeal. Robots at home : Understanding long-term human-robot interaction. Dans *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3230–3235. IEEE, 2008.

[140] Miriam Kienast et Walter F Sendlmeier. Acoustical analysis of spectral and temporal changes in emotional speech. Dans *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 92–97.

[141] Samuel Kim, Panayiotis G Georgiou, Sungbok Lee et Shrikanth Narayanan. Real-time emotion detection system using speech : Multi-modal fusion of different timescale features. Dans *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 48–51. IEEE, 2007.

[142] Michael Kipp. Anvil - a generic annotation tool for multimodal dialogue. Dans *Proceedings of Eurospeech'2001*, 2001.

[143] Aniket Kittur, Ed H Chi et Bongwon Suh. Crowdsourcing user studies with mechanical turk. Dans *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[144] Paul R Kleinginna Jr et Anne M Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5 (4):345–379, 1981.

[145] Ron Kohavi. *Wrappers for performance enhancement and oblivious decision graphs*. Thèse de doctorat, Stanford University, 1995.

[146] Ron Kohavi et George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[147] Ron Kohavi et Dan Sommerfield. Feature subset selection using the wrapper method : Overfitting and dynamic search space topology. Dans *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, pages 192–197, 1995.

[148] Margarita Kotti et Fabio Paternò. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International journal of speech technology*, 15(2):131–150, 2012.

[149] Michael Kriegel, Ruth Aylett, Pedro Cuba, Marco Vala et Ana Paiva. Robots meet ivas : a mind-body interface for migrating artificial intelligent agents. Dans *Intelligent Virtual Agents*, pages 282–295. Springer, 2011.

[150] Mineichi Kudo et Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern recognition*, 33(1):25–41, 2000.

[151] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao et Te-Won Lee. Emotion recognition by speech signals. Dans *Proceednigs of the Eighth European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003.

[152] Richard S Lazarus, Allen D Kanner, Susan Folkman et al. Emotions : A cognitive-phenomenological analysis. *Theories of emotion*, 1:189–217, 1980.

[153] Charles Le Brun. Conférences sur l'expression des différents caractères des passions. Numbéro vol. 9 dans L'art de connaître les hommes par la physionomie. Depélafol, 1820. Original work published in 1667.

[154] O. Lebec, M.W. Ben Ghezala, V. Leynart, I. Laffont, C. Fattal, L. Devilliers, C. Chastagnol, J.-C. Martin, Y. Mezouar, H. Korrapatti et V. Dupourque. High level functions for the intuitive use of an assistive robot. Dans *Proceedings of the 13th IEEE International Conference on Rehabilitation Robotics (ICORR 2013)*, 2013.

[155] Chul Min Lee, Shrikanth Narayanan et Roberto Pieraccini. Recognition of negative emotions from the speech signal. Dans *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 240–243. IEEE, 2001.

[156] Chul Min Lee et Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.

[157] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis et Sarun Savetsila. Personalization in hri : A longitudinal field experiment. Dans *Human-*

Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on, pages 319–326. IEEE, 2012.

[158] Iulia Lefter, Leon JM Rothkrantz, David A Van Leeuwen et Pascal Wiggers. Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems*, 4(2):148–168, 2011.

[159] Iulia Lefter, L.J.M. Rothkrantz, Pascal Wiggers et David. A. van Leeuwen. Emotion recognition from speech by combining databases and fusion of classifiers. Dans Petr Sojka, Ales Horak, Ivan Kopecek et Karel Pala, éditeurs, *Text and Speech and Dialogue*, volume 6231, page 353–359, Berlin, sep 2010. Springer, Springer. ISBN 3-642-15759-9.

[160] Iolanda Leite, Carlos Martinho et Ana Paiva. Social robots for long-term interaction : A survey. *International Journal of Social Robotics*, pages 1–18.

[161] C. Leroux, O. Lebec, M.W. Ben Ghezala, Y. Mezouar, L. Devillers, C. Chastagnol, J.-C. Martin, V. Leynaert et C. Fattal. Armen : Assistive robotics to maintain elderly people in natural environment. *IRBM*, 34(2), 2013.

[162] Michael Levit, Richard Huber, Anton Batliner et Elmar Nöth. Use of prosodic speech characteristics for automated detection of alcohol intoxication. Dans *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, pages 103–106, 2001.

[163] Mei Yii Lim. *Emotions, behaviour and belief regulation in an intelligent guide with attitude*. Thèse de doctorat, Heriot-Watt University, Edinburgh, UK, 2007.

[164] Mei Yii Lim. Memory models for intelligent social companions. Dans *Human-Computer Interaction : The Agency Perspective*, pages 241–262. Springer, 2012.

[165] Huan Liu et Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Springer, 1998.

[166] Sara Ljungblad, Jirina Kotrbova, Mattias Jacobsson, Henriette Cramer et Karol Niechwiadowicz. Hospital robot at work : something alien or an intelligent colleague ? Dans *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 177–186. ACM, 2012.

[167] Sara Ljungblad, Stina Nylander et Mie Nørgaard. Beyond speculative ethics in hri ? : Ethical considerations and the relation to empirical data. Dans *Proceedings of the 6th international conference on Human-robot interaction*, pages 191–192. ACM, 2011.

[168] Matthew Lombard, Theresa B Ditton, Daliza Crane, Bill Davis, Gisela Gil-Egui, Karl Horvath, Jessica Rossman et S Park. Measuring presence : A literature-based approach to the development of a standardized paper-and-pencil instrument. Dans *Third International Workshop on Presence, Delft, The Netherlands*, 2000.

[169] Karl F MacDorman, Robert D Green, Chin-Chang Ho et Clinton T Koch. Too real for comfort ? uncanny responses to computer generated faces. *Computers in Human Behavior*, 25(3):695–710, 2009.

[170] Joseph Mariani, Jean-Michel Besnier, Jacques Bordé, Jean-Michel Cornu, Marie Farge, Jean-Gabriel Ganascia, Jean-Paul Haton et Evelyne Serverin. Pour une éthique de la recherche en sciences et technologies de l'information et de la communication (stic). Rapport technique, Comité d'éthique du CNRS, 2009.

[171] Thomas Marill et D Green. On the effectiveness of receptors in recognition systems. *Information Theory, IEEE Transactions on*, 9(1):11–17, 1963.

[172] Philippe Martin. Winpitch ltl, un logiciel multimédia d'enseignement de la prosodie. *Apprentissage des Langues et Systèmes d'Information et de Communication (Alsic)*, 8(2):95–108, 2005.

[173] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado et Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. Dans *ISMIR*, pages 441–446, 2010.

[174] Sinéad McGilloway, Roddy Cowie, Ellen Douglas-Cowie, Stan Gielen, Machiel Westerdijk et Sybert Stroeve. Approaching automatic recognition of emotion from voice : a rough benchmark. Dans *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 207–212, 2000.

[175] Scott W McQuiggan, Jonathan P Rowe et James C Lester. The effects of empathetic virtual characters on presence in narrative-centered learning environments. Dans *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1511–1520. ACM, 2008.

[176] Michael F. McTear. *Spoken dialogue technology : enabling the conversational user interface*. Springer-Verlag, London, 2004.

[177] Albert Mehrabian. Pleasure-arousal-dominance : A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[178] Albert Mehrabian et Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248–252, 1967.

[179] Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, C Muller, Richard Huber, Bernt Andrassy, Josef G Bauer et al. Comparison of four approaches to age and gender recognition for telephone applications. Dans *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages 1089–1092. IEEE, 2007.

[180] Roman Miletitch, Nicolas Sabouret et Magalie Ochs. Susciter l'émotion dans la narration automatique. *TSI-Technique et Science Informatiques*, 31(4):477, 2012.

[181] Marvin Minsky. *Society of mind*. SimonandSchuster, 1988.

[182] Wade J Mitchell, Kevin A Szerszen Sr, Amy Shirong Lu, Paul W Schermerhorn, Matthias Scheutz et Karl F MacDorman. A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1):10, 2011.

[183] Andrew W Moore et Mary S Lee. Efficient algorithms for minimizing cross validation error. Dans *ICML*, pages 190–198, 1994.

[184] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[185] Robert Morris. Crowdsourcing workshop : the emergence of affective crowdsourcing. Dans *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. ACM, 2011.

[186] Emily Mower, Kyu Jeong Han, Sungbok Lee et Shrikanth S Narayanan. A cluster-profile representation of emotion using agglomerative hierarchical clustering. Dans *INTERSPEECH*, pages 797–800, 2010.

[187] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee et Shrikanth Narayanan. Interpreting ambiguous emotional expressions. Dans *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009.

[188] Iosif Mporas et Todor Ganchev. Estimation of unknown speaker's height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.

[189] Christian A Müller, Frank Wittig et Jörg Baus. Exploiting speech for recognizing elderly users to respond to their special needs. Dans *INTERSPEECH*, 2003.

[190] Daniel Neiberg, Kjell Elenius et Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. Dans *INTERSPEECH*, pages 809–812, 2006.

[191] Andrew Y Ng. Preventing" overfitting" of cross-validation data. Dans *ICML*, volume 97, pages 245–253, 1997.

[192] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini et Catherine Pelachaud. Greta : an interactive expressive eca system. Dans *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[193] Radosław Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont et al. Laugh-aware virtual agent and its impact on user amusement. Dans *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 619–626. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[194] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda et Kensuke Kato. Measurement of negative attitudes toward robots. *Interaction Studies*, 7(3):437–454, 2006.

[195] E. Nöth, A. Batliner, A. Kießling, R. Kompe et H. Niemann. Suprasegmental modelling. Dans *Computational Models of Speech Pattern Processing*, pages 181–198. Springer, 1999.

[196] Edgar Osuna, Robert Freund et Federico Girosit. Training support vector machines : an application to face detection. Dans *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1997)*, pages 130–136. IEEE, 1997.

[197] Igor S Pandzic et Robert Forchheimer. *MPEG-4 facial animation : the standard, implementation and applications*. Wiley, 2003.

[198] Maja Pantic et Leon JM Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[199] V Petrushin. Emotion in speech : Recognition and application to call centers. Dans *Proceedings of Artificial Neural Networks in Engineering*, pages 7–10, 1999.

[200] Roberto Pieraccini et Juan Huerta. Where do we go from here ? research and commercial spoken dialog systems. Dans *6th SIGdial Workshop on Discourse and Dialogue*, 2005.

[201] George W Pigman. Freud and the history of empathy. *The International Journal of Psychoanalysis*, 76:237–252, 1995.

[202] Hannes Pirker et Georg Loderer. I said "two ti-ckets" : How to talk to a deaf wizard. Dans *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*, pages 181–186, 1999.

[203] Johannes Pittermann, Angela Pittermann et Wolfgang Minker. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 13(1):49–60, 2010.

[204] Robert Plutchik. *Emotion : A psychoevolutionary synthesis*. Harper & Row, New York, 1980.

[205] Robert Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[206] Tim Polzehl, Shiva Sundaram, Hamed Ketabdar, Michael Wagner et Florian Metze. Emotion classification in children's speech using fusion of acoustic and linguistic features. Dans *INTERSPEECH*, pages 340–343, 2009.

[207] Ken Prepin et Catherine Pelachaud. Basics of intersubjectivity dynamics : Model of synchrony emergence when dialogue partners understand each other. Dans *Agents and Artificial Intelligence*, pages 302–318. Springer, 2013.

[208] Ken Prepin et Nicolas Sabouret. Sélection d'un vocabulaire commun : Une étude autour de l'énaction dans l'interaction entre agents. Dans *Proc. French Workshop on Artificial Companions, Affects and Interactions (WACAI, 2012)*, 2012.

[209] Stephanie D Preston et Frans De Waal. Empathy : Its ultimate and proximate bases. *Behavioral and brain sciences*, 25(01):1–20, 2002.

[210] Pavel Pudil, Jana Novovičová et Josef Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[211] Javier Ramirez, Juan Manuel Górriz et José Carlos Segura. Voice activity detection. fundamentals and speech recognition system robustness. *Robust Speech Recognition and Understanding*, 6(9):1–22, 2007.

[212] Dirk M Reichardt. A definition approach for an "emotional turing test". Dans *Affective Computing and Intelligent Interaction*, pages 716–717. Springer, 2007.

[213] Elias Rentzeperis, Christos Boukis, Aristodemos Pnevmatikakis et Lazaros C Polymenakos. Combining finite state machines and lda for voice activity detection. Dans *Artificial Intelligence and Innovations 2007 : from Theory to Applications*, pages 323–329. Springer, 2007.

[214] Juha Reunanen. *Overfitting in feature selection : Pitfalls and solutions*. Thèse de doctorat, 2012.

[215] Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese et Leonardo Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3 (2):131–141, 1996.

[216] Peter Roach, Richard Stibbard, Jane Osborne, Simon Arnfield et Jane Setter. Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, 28(1-2):83–94, 1998.

[217] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[218] David Sander, Didier Grandjean et Klaus R Scherer. A systems approach to appraisal mechanisms in emotion. *Neural networks*, 18(4):317–352, 2005.

[219] C Schaeffer et T May. Care-o-bot-a system for assisting elderly or disabled persons in home environments. *Assistive technology on the threshold of the new millenium*, 1999.

[220] Klaus R Scherer. Appraisal theory. *Handbook of Cognition and Emotion*, pages 637–663, 1999.

[221] Klaus R Scherer. Appraisal considered as a process of multi-level sequential checking. Dans K. R. Scherer, A. Schorr et T. Johnstone, éditeurs, *Appraisal processes in emotion : Theory, methods, research*, pages 92–120. Oxford University Press, New York, 2001.

[222] Klaus R Scherer. Vocal communication of emotion : A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.

[223] Klaus R Scherer, Tom Johnstone et Gundrun Klasmeyer. Vocal expression of emotion. Dans R.J. Davidson, K.R. Scherer et H.H. Goldsmith, éditeurs, *Handbook of affective sciences*, pages 433–456. Oxford University Press, New York, 2003.

[224] Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Louis-Philippe Morency et al. Perception markup language : towards a standardized representation of perceived nonverbal behaviors. Dans *Intelligent virtual agents*, pages 455–463. Springer, 2012.

[225] Marc Schröder, Laurence Devillers, Kostas Karpouzis, Jean-Claude Martin, Catherine Pelachaud, Christian Peter, Hannes Pirker, Björn Schuller, Jianhua Tao et Ian Wilson. What should a generic emotion markup language be able to represent ? Dans *Affective Computing and Intelligent Interaction*, pages 440–451. Springer, 2007.

[226] Björn Schuller et Felix Burkhardt. Learning with synthesized speech for automatic emotion recognition. Dans *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5150–5153. IEEE, 2010.

[227] Björn Schuller, Gerhard Rigoll et Manfred Lang. Hidden markov model-based speech emotion recognition. Dans *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 2, pages 1–4. IEEE, 2003.

[228] Björn Schuller, Dino Seppi, Anton Batliner, Andreas Maier et Stefan Steidl. Towards more reality in the recognition of emotional speech. Dans *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages 941–944. IEEE, 2007.

[229] Björn Schuller, Stefan Steidl et Anton Batliner. The interspeech 2009 emotion challenge. Dans *INTERSPEECH*, pages 312–315, 2009.

[230] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller et Shrikanth S Narayanan. The interspeech 2010 paralinguistic challenge. Dans *INTERSPEECH*, pages 2794–2797, 2010.

[231] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet et al. The interspeech 2012 speaker trait challenge. Dans *INTERSPEECH*, 2012.

[232] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel et Jarek Krajewski. The interspeech 2011 speaker state challenge. Dans *INTERSPEECH*, pages 3201–3204, 2011.

[233] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi et al. The interspeech 2013 computational paralinguistics challenge : social signals, conflict, emotion, autism. Dans *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013.

[234] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth et Gerhard Rigoll. Cross-corpus acoustic emotion recognition : Variances and strategies. *Affective Computing, IEEE Transactions on*, 1(2):119–131, 2010.

[235] Björn Schuller et Felix Weninger. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. Dans *Acoustics Speech and Si-*

gnal Processing (ICASSP), 2010 IEEE International Conference on, pages 5054–5057. IEEE, 2010.

[236] Björn Schuller et Felix Weninger. Ten recent trends in computational paralinguistics. Dans *Cognitive Behavioural Systems*, pages 35–49. Springer, 2012.

[237] Björn Schuller, Zixing Zhang, Felix Weninger et Gerhard Rigoll. Selecting training data for cross-corpus speech emotion recognition : Prototypicality vs. generalization. Dans *Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*, 2011.

[238] Björn Schuller, Zixing Zhang, Felix Weninger et Gerhard Rigoll. Using multiple databases for training in emotion recognition : To unite or to vote ? Dans *INTERSPEECH*, pages 1553–1556, 2011.

[239] Mohammad Shami et Werner Verhelst. Automatic classification of expressiveness in speech : a multi-corpus study. Dans *Speaker classification II*, pages 43–56. Springer, 2007.

[240] Candace L Sidner, Cory D Kidd, Christopher Lee et Neal Lesh. Where to look : a study of human-robot engagement. Dans *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 78–84. ACM, 2004.

[241] Candy Sidner. Engagement during human-robot dialogues. Dans *Constraints in Discourse*, pages 147–151, 2006.

[242] Pawel Smialowski, Dmitrij Frishman et Stefan Kramer. Pitfalls of supervised feature selection. *Bioinformatics*, 26(3):440–443, 2010.

[243] John Snel, Alexey Tarasov, Charlie Cullen et Sarah Jane Delany. A crowdsourcing approach to labelling a mood induced speech corpus. 2012.

[244] Mohammad Soleymani et Martha Larson. Crowdsourcing for affective annotation of video : Development of a viewer-reported boredom corpus. Dans *Proceedings*

of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010), pages 4–8, 2010.

[245] Stephen D Stearns. On selecting features for pattern classifiers. Dans *Proceedings of the 3rd International Joint Conference on Pattern Recognition*, pages 71–75, 1976.

[246] Stefan Steidl. *Automatic classification of emotion-related user states in spontaneous children's speech*. Thèse de doctorat, University of Erlangen-Nuremberg, 2009.

[247] Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth et Heinrich Niemann. "of all things the measure is man" : Automatic classification of emotions and inter-labeler consistency. Dans *Proc. ICASSP*, volume 1, pages 317–320, 2005.

[248] Bas R Steunebrink, Nieske L Vergunst, Christian P Mol, Frank Dignum, Mehdi Dastani et John-Jules Ch Meyer. A generic architecture for a companion robot. Dans *ICINCO-RA (2)*, pages 315–321, 2008.

[249] Richard L. Street. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139–169, 1984.

[250] Marie Tahon. *Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot*. Thèse de doctorat, Université Paris-Sud, 2012.

[251] Marie Tahon, Agnes Delaborde et Laurence Devillers. Real-life emotion detection from speech in human-robot interaction : Experiments across diverse corpora with child and adult voices. Dans *INTERSPEECH*, pages 3121–3124, 2011.

[252] Marie Tahon et Laurence Devillers. Acoustic measures characterizing anger across corpora collected in artificial or natural context. Dans *Proceedings of the Fifth International Conference on Speech Prosody (ISCA, Speech Prosody 2010Chicago, USA, 2010)*, 2010.

[253] Alexey Tarasov, Sarah Jane Delany et Charlie Cullen. Using crowdsourcing for labelling emotional speech assets. 2010.

[254] Silvan Tomkins. Affect, imagery, consciousness. vol. 1 : The positive affects. 1962.

[255] Silvan Tomkins. Affect, imagery, consciousness. vol. 2 : The negative affects. 1963.

[256] GV Trunk. A problem of dimensionality : A simple example. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (3):306–307, 1979.

[257] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[258] Christophe Vaudable. *Analyse et reconnaissance des émotions lors de conversations de centres d'appels*. Thèse de doctorat, Université Paris Sud-Paris XI, 2012.

[259] Christophe Vaudable et Laurence Devillers. Negative emotions detection as an indicator of dialogs quality in call centers. Dans *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5109–5112. IEEE, 2012.

[260] Christophe Vaudable, Nicolas Rollet et Laurence Devillers. Annotation of affective interaction in real-life dialogs collected in a call-center. Dans *Proceedings of the Third International Workshop on EMOTION, LREC*, 2010.

[261] Dimitrios Ververidis et Constantine Kotropoulos. Automatic speech classification to five emotional states based on gender information. Dans *Proceedings of 12th European Signal Processing Conference*, pages 341–344, 2004.

[262] Laurence Vidrascu et Laurence Devillers. Detection of real-life emotions in call centers. Dans *INTERSPEECH*, pages 1841–1844, 2005.

[263] Laurence Vidrascu et Laurence Devillers. Représentation et détection des émotions dans des données issues de dialogues enregistrés dans des centres d'appels : des émotions mixtes dans des données réelles. *Numéro spÃl'cial "Interaction Émotionnelle", Revue des Sciences et Technologies de l'Information, série Revue d'Intelligence Artificielle*, 20(4–5):447–476, 2006.

[264] Hannes Vilhjálmsson, Nathan Cantelmo, Justine Cassell, Nicolas E Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud et al. The behavior markup language : Recent developments and challenges. Dans *Intelligent virtual agents*, pages 99–111. Springer, 2007.

[265] Kazuyoshi Wada et Takanori Shibata. Living with seal robotsâĂŤits sociopsychological and physiological influences on the elderly at a care house. *Robotics, IEEE Transactions on*, 23(5):972–980, 2007.

[266] Yingxu Wang. Cognitive robots. *Robotics & Automation Magazine, IEEE*, 17(4): 54–62, 2010.

[267] Richard L West et Lynn H Turner. *Introducing communication theory : analysis and application*. Boston : McGraw-Hill, 4th ed édition, 2010.

[268] A Wayne Whitney. A direct method of nonparametric measurement selection. *Computers, IEEE Transactions on*, 100(9):1100–1103, 1971.

[269] Janneke Wilting, Emiel Krahmer et Marc Swerts. Real vs. acted emotional speech. Dans *INTERSPEECH*, 2006.

[270] Martin Wollmer, Björn Schuller, Florian Eyben et Gerhard Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *Selected Topics in Signal Processing, IEEE Journal of*, 4(5):867–881, 2010.

[271] Martin Wöllmer, Felix Weninger, Stefan Steidl, Anton Batliner et Björn Schuller. Speech-based non-prototypical affect recognition for child-robot interaction in reverberated environments. Dans *INTERSPEECH*, pages 3113–3104, 2011.

[272] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park et Chungyong Lee. Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 36(2):180–181, 2000.

[273] Rachel Wood, Paul Baxter et Tony Belpaeme. A review of long-term memory in natural and synthetic systems. *Adaptive Behavior*, 20(2):81–103, 2012.

[274] Wilhelm Max Wundt. *Grundzüge de physiologischen Psychologie*. W. Engelman, 1874.

[275] Nick Yee, Jeremy N Bailenson et Kathryn Rickertsen. A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. Dans *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1–10. ACM, 2007.

[276] Serdar Yildirim, Shrikanth Narayanan et Alexandros Potamianos. Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, 25(1):29–44, 2011.

[277] Silvia Casado Yusta. Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters*, 30(5):525–534, 2009.

[278] Tong Zhang, Mark Hasegawa-Johnson et Stephen E Levinson. Children's emotion recognition in an intelligent tutoring scenario. Dans *Proc. Eighth European Conf. Speech Comm. and Technology (INTERSPEECH)*, 2004.

[279] Zixing Zhang, Jun Deng et Björn Schuller. Co-training succeeds in computational paralinguistics. Dans *to appear in Proc. 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

Annexes

Annexe I

Publications

En cours de publication

- Chastagnol, C., Clavel, C., Courgeon, M., & Devillers, L. Designing an emotion detection system for a socially-intelligent human-robot interaction, In Mariani, J., Devillers, L., Garnier-Rizet, M., & Rosset, S. (Eds.), Towards a Natural Interaction with Robots, Knowbots and Smartphones, Putting Spoken Dialog Systems into Practice, Springer, 2013.

Revues internationales

- Leroux, C., Lebec, O., Ben Ghezala, M.W., Mezouar, Y., Devillers, L., Chastagnol, C., Martin, J.-C., Leynaert, V., & Fattal, C. ARMEN : Assistive robotics to maintain elderly people in natural environment, IRBM, Volume 34, Issue 2, April 2013.

Conférences internationales

- Lebec, O., Ben Ghezala, M., Leynart, V., Laffont, I., Fattal, C., Leroux, C., Devillers, L., Chastagnol, C., Martin, J-C., Mezouar, Y., Korrapatti, H. & Dupourque, V. High level functions for the intuitive use of an assistive robot, 13th International Conference on Rehabilitation Robotics (ICORR 2013).

- Chastagnol, C., & Devillers, L. Personality traits detection using a parallelized modified SFFS algorithm. In Proceedings of InterSpeech 2012.

- Chastagnol, C., & Devillers, L. Analysis of Anger Across Several Agent-Customer Interactions in French Call Centers. In Proceedings of ICASSP 2011, pp 4960–4963.

- Devillers, L., Vaudable, C., & Chastagnol, C. Real-life emotion-related states detection in call centers : a cross-corpora study. In Proceedings of InterSpeech 2010, pp 2350–2353.

Workshops et conférences nationales

- Chastagnol, C., Clavel, C., Courgeon, M., & Devillers, L. Designing an Emotion Detection System for a Socially-Intelligent Human-Robot Interaction. In Proceedings of IWSDS 2012.

- Chastagnol, C., & Devillers, L. Collecting Spontaneous Emotional Data for a Social Assistive Robot. In Proceedings of ES^3 2012 workshop, as part of LREC 2012.

- Chastagnol, C., & Devillers, L. Détection d'émotions dans la voix de patients en interaction avec un agent conversationnel animé. Actes JEP-TALn-RECITAL 2012

Annexe II

Présentations et posters

- Séminaire du GT ACAI 2013 (Paris, France). Présentation orale : *Analyse des dimensions affectives dans une interaction avec un robot assistant dans le cadre du projet ANR ARMEN*.
- Colloque du Centre Expertise National en Robotique (CENRob), 2013 (Paris, France). Présentation d'un poster : *Emotion Detection System for Human-Robot Interaction*.
- Workshop IWSDS 2012 (France). Présentation d'un poster : *Designing an Emotion Detection System for a Socially-Intelligent Human-Robot Interaction*.
- Conférence internationale InterSpeech 2012 (Portland, Oregon, USA). Présentation orale : *Personality traits detection using a parallelized modified SFFS algorithm*.
- Forum Annuel Digiteo 2012 (Paris, France). Participation à un poster : *Affective and social dimensions in spoken interactions*.
- École d'été AERFAISS 2012 (Vigo, Espagne). Présentation d'un poster : *Collecting spontaneous emotional data for a social assistive robot*.
- Conférence internationale ICASSP 2011 (Prague, République tchèque). Présentation d'un poster : *Analysis of Anger Across Several Agent-Customer Interactions in French Call Centers*.
- Conférence internationale LREC 2012 (Istanbul, Turquie). Présentation orale : *Collecting Spontaneous Emotional Data for a Social Assistive Robot*.
- École d'été ICVSS 2010 (Sicile). Présentation d'un poster : *Emotion Detection on Pathological Subjects*.

Annexe III

Théories des émotions : éléments historiques

Le mot "émotion" n'apparaît que relativement tardivement dans la langue française, au cours du 17ème siècle. Cependant l'oeuvre des philosophes antiques fait déjà état de la notion, comme Aristote dans la *Rhétorique* et l'*Éthique à Nicomaque* en citant des exemples de colère, de pitié, de désir et de peur, et en les décrivant comme suivies de plaisir ou de douleur [19]. Ce sont alors les termes de "passions" ou d'"affections" qui sont utilisés et ils ont en commun d'avoir une étymologie connotant la souffrance, la passivité et la maladie. Les stoïciens les conçoivent comme des "maladies de l'âme" qui devaient être traitées avec calme et raison (concept d'*apatheia*). Plus tardivement, les médecins du Moyen-Âge et de la Renaissance les considèrent également comme mauvaises pour la santé [82]. À l'apparition du mot "émotion", celui-ci désigne plutôt un mouvement physique qui est interprété comme un signe externe et corporel des mouvements internes des passions, comme le remarquent le peintre Le Brun et de manière plus élégante, le philosophe Jeremy Bentham :

> Premièrement la passion est un mouvement de l'âme, [...] lequel se fait pour suivre ce que l'âme pense lui être bon, ou pour fuir ce qu elle pense lui être mauvais, et d'ordinaire tout ce qui cause à l'âme de la passion, fait faire au corps quelque action.
>
> *D'après Le Brun (1667) [153]*

> The emotions of the body are received, and with reason, as probable indications of the temperature of the mind.
>
> *D'après Bentham (1789) [20]*

Même si l'utilisation du mot "émotion" a peu a peu migré du domaine du physique au domaine du mental durant la fin du 18ème siècle, peu d'avancées significatives ont été observées jusqu'à la révolution darwinienne. Charles Darwin est essentiellement connu

pour sa théorie de l'évolution - qu'on pourrait qualifier plus justement de mise en évidence des mécanismes de l'adaptation des espèces - développée dans son oeuvre majeure, *"De l'origine des espèces"*. Cependant il a également apporté une contribution majeure au domaine de la théorie des émotions avec son livre *"L'expression des émotions chez l'homme et les animaux"*, publié en 1872 [64].

Il faut tout d'abord noter que Darwin s'intéresse en premier lieu à l'expression faciale des émotions et non aux émotions pour elles-mêmes - il est d'ailleurs intéressant de remarquer qu'il ne définit pas le terme "émotion" [120]. Son but est d'étayer sa théorie évolutionniste et de réfuter certains de ses contemporains aux visions créationnistes comme son ancien professeur Sir Charles Bell qui soutient que l'homme possède des muscles faciaux spécialement conçus pour exprimer des émotions [17]. Il insiste par exemple sur le fait que certaines expressions comme montrer les dents quand on est en colère, ne peuvent pas être expliquées si on ne considère pas que l'homme a pu exister dans une forme plus animale.

Une des ses idées les plus fortes, dans la lignée de sa théorie évolutionniste, est de présenter la faculté d'afficher des expressions faciales comme une adaptation nécessaire tout d'abord pour des raisons de survie (pour protéger l'organisme et le préparer à l'action en situation de danger) puis de communication. Le caractère adaptatif et évolué de ces expressions explique donc pourquoi elles sont semblables chez tous les humains, indépendamment de la culture, et pourquoi elles sont similaires chez des espèces proches comme certains mammifères. Mais il présente également de nombreuses observations, considérées comme des évidences ou d'amusantes curiosités, qui ont par la suite donné naissance à des champs de recherche complets comme par exemple la tendance des gens à imiter les expressions des autres ou la notion que réprimer une expression peut supprimer ou du moins atténuer l'émotion ressentie sous-jacente, ou encore des observations sur les différences entre les sourires spontanés, dits *"de Duchenne"* [87], et intentionnels.

Il relève aussi des problèmes spécifiques au domaine : difficulté d'étudier les expressions, car elles sont souvent très subtiles et rapides (il n'y avait pas de caméra haute-définition ni de ralenti à l'époque...), difficulté pour un observateur de rester neutre si un sujet exprime une émotion forte à cause du phénomène empathie et risque pour l'obser-

vateur d'imaginer des expressions là où il n'y en a pas (donc pas d'observateur objectif et fiable). Le biais dans l'auto-évaluation d'expressions ou d'émotions l'amène par ailleurs à imaginer des protocoles qu'on pourrait qualifier aujourd'hui d'*annotation perceptive*.

En ayant l'intuition de ces questions bien avant qu'elles soient véritablement investiguées[1], Darwin a donc été très influent sur la communauté (plus de 3000 citations), même si certains chercheurs modèrent son importance [111].

[1] La question de l'universalité des expressions faciales sera traitée longuement à partir des années 1970 dans les travaux d'Ekman [92]