Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales

Chloé Clavel

TELECOM PARIS
école nationale supérieure des télécommunications

École Doctorale d'Informatique, Télécommunications et Électronique de Paris

Thèse

présentée pour obtenir le grade de Docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Signal et Images**

Chloé CLAVEL

Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales

Soutenue le 15 mars 2007 devant le jury composé de :

Jean-Paul Haton	Président
Catherine Pelachaud	Rapporteurs
Jean-François Bonastre	
Laurence Devillers	Examinateurs
Ioana Vasilescu	
Gaël Richard, Célestin Sedogbo	Directeurs de thèse

Cette thèse s'est déroulée dans le cadre d'une bourse CIFRE entre :

– Thales Recherche et Technologie (TRT), au sein de l'équipe Human Interaction Team,

– le laboratoire de Traiment du Signal et des Images (TSI) de Telecom-Paris et,

– le LIMSI-CNRS (Paris XI), groupe Traitement du Langage Parlé.

Remerciements

Voici venu le moment de rédiger la page la plus délicate de ce document : les remerciements. Finalement rédiger une thèse est un jeu d'enfants ... Je vous demande donc toute votre indulgence en cas d'oubli ou de maladresse.

Pendant ces trois années, j'ai eu la chance de profiter d'un encadrement exceptionnellement riche et pluridisciplinaire. Je tiens donc à remercier en premier lieu mes cinq encadrants ...

... Célestin Sedogbo (directeur de thèse industriel TRT) pour avoir eu l'initiative de ce sujet de thèse qui m'a tenu en haleine pendant trois ans ;

... Gaël Richard (directeur de thèse académique TSI) pour m'avoir tout d'abord aiguillé vers ce sujet puis pour ses conseils scientifiques et son enthousiasme ;

... Laurence Devillers (encadrante LIMSI) et Ioana Vasilescu (encadrante TSI et LIMSI) pour avoir accepté de participer à la définition du sujet et pour m'avoir fait bénéficier de leur expertise sur un sujet à l'époque précurseur. Ioana m'a fait de plus profiter de son savoir indispensable dans le domaine des sciences humaines et sociales.

... Thibaut Ehrette (encadrant TRT) pour s'être investi dans l'encadrement de cette thèse dans la continuité de la sienne.

Je les remercie tous les cinq pour m'avoir écoutée, conseillée, guidée, soutenue tout au long de cette thèse, tout en me laissant libre dans mes choix de recherche. Avec une reconnaissance spéciale à Ioana et Laurence pour leur grande disponibilité...

Viennent ensuite mes deux rapporteurs et mon président de Jury : Mme Catherine Pelachaud, M. Jean-François Bonastre et M. Jean-Paul Haton. Je les remercie tous les trois pour avoir accepté de faire partie de mon jury, et de m'ouvrir les yeux, grâce à la diversité et la richesse de leurs remarques, sur de nouvelles perspectives de recherche.

Pendant les nombreuses passées à Telecom-Paris, j'ai vu avec admiration se soutenir les thèses de ceux qui m'ont accueillie alors que je n'étais que stagiaire : dans l'ordre chronologique, j'ai nommé, Julie, Thomas, Eduardo, Cléo, Slim, Grégoire. Je salue humblement, ces nouveaux docteurs, sans oublier Raphaël, à l'époque post-doctorant, pour leur encouragement dans ma recherche de thèse puis pour m'avoir fait profiter de leur fraîche expérience de doctorant pendant ces trois années. Je salue également la génération d'après et ceux que je n'ai connus que plus tard : Valentin, Dora, Maria, Tyze, Lionel, Nancy, Christophe, Miguel et son guacamole. Je tiens aussi à remercier Slim et Bertrand pour leurs conseils scientifiques. Sans oublier Laurence, Fabrice et Sophie-Charlotte, les anges gardiens des doctorants.

Un peu plus bas sur le RER B, j'ai pu trouver le soutien quotidien de l'ensemble de mes collègues de Thales. Je les salue tous avec reconnaissance et remercie plus particulièrement (dans l'ordre alphabétique) : Antoine (le parfait stagiaire), Aude, Bénédicte, Cathy, Claire, Frédéric L., Frédéric P., Jérôme, Joëlle, Laurent, Miniar, Nicolas, Thibaut. Je voudrais notamment remercier mon ancien chef d'équipe Olivier Grisvard et ma nouvelle chef d'équipe Juliette Mattioli pour m'avoir permis de terminer ma thèse dans des conditions idéales.

Je remercie également les thésards et chercheurs du groupe Traitement du Language Parlé du LIMSI pour leurs conseils scientifiques.

Mais surtout, parce qu'ils sont la preuve vivante qu'il y avait une vie en dehors de la

thèse, je tiens à remercier mes amis avec qui j'ai pu passer (et j'espère continuer de passer) de grands moments : Marie & Martin, Marlotte, Claire, Aurélien, Marie (Macha), Eléonore, Ngoc-Hue, Lorcan, Kévin, Henri, Benj, Silvia, les Telecommiens nommés ci-dessus, Cédric, Maëva, Marie-Céline. Merci aux cloportes, aux k-têtes de Corbeville, à la fanfare, à Babacar, Fanny et Alejandra.

Si le futur docteur s'engage dans une thèse de son propre chef, prêt à en assumer les conséquences, ce n'est pas le cas de sa famille et de ses proches. Merci donc à toute ma petite famille cristolienne, parisienne, marseillaise, toulousaine, et maurecourtoise. Je ne remercierais jamais assez mes parents et mon frère Grégory pour leur patience, leur écoute et leur enthousiasme.

Bravo à Rémi pour avoir tenu en première ligne, amorti mes doutes, relancé les baisses de régime, bref pour son indispensable présence.

Table des matières

Chapitre 1

Introduction générale

Doter la machine des capacités de compréhension des comportements humains : tel est le défi scientifique autour duquel se rassemblent différentes communautés scientifiques (traitement du signal, traitement automatique du langage, intelligence artificielle, robotique, interaction homme-machine, etc.).

Les informations disponibles sont les signaux acquis par le système via des capteurs (image, son, capteurs physiologiques). Les données manipulées sont donc de très bas niveau : les échantillons sonores ou encore les pixels des images. Entre ces données bas niveau et l'interprétation qu'en font les humains, le fossé est énorme.

L'un des signaux fréquemment utilisé est le signal de parole. La parole est en effet l'une des modalités fondamentales que l'homme utilise pour communiquer. Les systèmes de reconnaissance automatique de la parole donnent à la machine les capacités de transformer le signal sonore en une suite de mots. Le domaine du traitement automatique du langage permet d'accéder au sens de cette suite de mots. Partant de ces outils (relativement efficaces), il est nécessaire d'aller plus loin : la question n'est plus uniquement de savoir ce qui est dit mais aussi de connaître le contexte de prononciation de la phrase. C'est à ce niveau qu'intervient la dimension émotionnelle. Si on ne prend pas en compte l'intonation de la phrase, il est difficile de faire la différence entre une question et une affirmation. De la même façon, selon l'émotion, l'attitude, mais aussi selon la personnalité du locuteur, une même phrase peut avoir un sens différent. Une illustration de ce dernier constat se trouve dans l'utilisation croissante de smileys (ou emoticônes) pour accompagner ses phrases dans le courrier électronique et dans les messages des utilisateurs des groupes de discussion. Les emoticônes sont la représentation graphique d'un visage humain exprimant le sourire, l'étonnement, la colère, etc. et permettent de nuancer le sens d'un texte écrit souvent rapidement, dans un langage proche du langage parlé.

1.1 Le phénomène émotionnel

En quoi consiste le phénomène émotionnel ? La réponse à cette question est un sujet de controverse sur lequel se sont penchés de nombreux psychologues. L'article [Kleinginna et Kleinginna, 1981] retrace les 100 années d'études faites par la communauté des psychologues et conclut qu'il est difficile d'aboutir à un consensus sur la définition du phénomène émotionnel.

Modèles théoriques

Selon Scherer, « les émotions sont les interfaces de l'organisme avec le monde extérieur » et le processus émotionnel se décompose en trois principaux aspects [Scherer, 1984] :

1. L'évaluation de la signification des stimuli par l'organisme (aspect cognitif) ;
2. La préparation aux niveaux physiologique et psychologique d'actions adaptées (aspect physiologique) ;
3. La communication par l'organisme des états et des intentions de l'individu à son environnement social (aspect expressif).

Ces trois aspects, cognitif, physiologique et expressif sont généralement acceptés comme constituants du phénomène émotionnel.

L'une des divisions majeures des théories développées a trait à une différence de point de vue sur les rôles respectifs des processus cognitifs et des modifications physiologiques qui interviennent dans l'expérience émotionnelle.

Pour William James [James, 1884], par exemple, c'est la perception de réactions physiologiques, plus ou moins intenses ou de nature différente, qui conditionnent le vécu d'une émotion. Ainsi, par exemple, les larmes seraient à l'origine de la sensation de tristesse.

A l'inverse, pour les psychologues cognitivistes, le processus cognitif est une étape antérieure et indispensable dans la production de l'émotion.

Une revue des différents modèles théoriques des émotions est donnée dans [Devillers, 2006].

Définition du terme émotion

Chacune des théories qui s'attaquent à la description du phénomène émotionnel implique des définitions du terme émotion plus ou moins restrictives en fonction des aspects pris en compte.

Les distinctions les plus couramment utilisées dans la littérature pour circonscrire le sens du terme émotion sont : *émotions primaires vs. émotions secondaires* et *émotions basiques vs. émotions dérivées*.

Damasio distingue les émotions primaires telles que la peur ou la colère, qui sont des émotions innées, des émotions secondaires ou sociales (honte, fierté, amour, etc.) qui sont acquises au cours de l'existence de l'individu [Damasio, 1995]. La distinction émotions basiques/émotions dérivées repose sur les définitions suivantes :

- les émotions basiques ont des manifestations associées particulières (expressions faciales, tendances comportementales, motifs physiologiques) et sont signifiantes si elles représentent des tendances à l'action et si elles sont reconnaissables par des indices physiologiques ;
- les émotions dérivées sont des combinaisons de plusieurs émotions basiques et sont propres à l'homme.

Le nombre d'émotions basiques varie d'une étude à l'autre, comme l'indique le tableau 1.1. La plupart du temps les émotions basiques incluent la peur, la colère, la joie et la tristesse.

Scherer distingue les émotions dites « full-blown » (abouties : qui impliquent les trois aspects cognitif, physiologique et expressif), des autres *états affectifs* tels que les humeurs, les attitudes. Cette distinction repose sur une analyse des états affectifs en fonction notamment de leur intensité, de leur durée et de leur focus. L' émotion « full-blown » correspond à l'état affectif de durée la plus courte, d'intensité la plus forte et qui est dépendant d'un événement précis (focus).

Cowie utilise le terme *états émotionnels* [Cowie *et al.*, 2001] auquel il donne un sens large en incluant les *états reliés à des émotions* tels que les humeurs. Il distingue ensuite les différents

Theorist	Basic emotions	Basis for inclusion
Arnold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Relation to action tendencies
Ekman, Friesen, and Ellsworth	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Fridja	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray	Rage and terror, anxiety, joy	Hardwired
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James	Fear, grief, love, rage	Bodily involvement
McDougall	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mower	Pain, pleasure	Unlearned emotional states
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness	Do not require prepositional content
Panksepp	Expectancy, fear, rage, panic	Hardwired
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological process
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson	Fear, love rage	Hardwired
Weiner and Graham	Happiness, sadness	Attribution independent

FIG. 1.1 – *Ensembles d'émotions basiques considérés en fonction des théoriciens. Ce tableau est extrait de [Tato, 1999] et est le résultat de l'étude réalisée dans [Orthony et Turner, 1990].*

états émotionnels en fonction de leur description temporelle, de leur focus et du contrôle de la personne.

1.2 Les émotions dans les applications

L'intégration des émotions dans des applications industrielles effectives commence à voir le jour. Selon le domaine applicatif ou la communauté scientifique impliquée, l'angle sous lequel sont abordées les émotions diffère. Le terme émotion présente des variantes telles que les termes affect, états émotionnels, états reliés à des émotions, attitude, humeur, comportement émotionnel, expressivité, etc. qui ont chacune leur propre définition [Devillers, 2006].

Les systèmes de dialogues, les applications et les émotions.

Les émotions interviennent de deux manières opposées dans les systèmes de dialogue, en tant que phénomène susceptible d'altérer la reconnaissance des mots prononcés par l'utilisateur d'une part, et en tant que phénomène permettant de mieux comprendre son comportement, d'autre part.

Dans le premier cas, il s'agit d'améliorer les systèmes de reconnaissance de la parole dans les systèmes de dialogue, en prenant en compte la variabilité émotionnelle du locuteur [Varadarajan *et al.*, 2006] [ten Bosch, 2003]. Les systèmes de reconnaissance automatique de la parole atteignent maintenant des performances dignes de leur industrialisation mais souffrent en effet d'un manque de robustesse aux variations intra-locuteur, qui correspondent à des changements attitudinaux ou émotionnels de l'état du locuteur pendant qu'il s'exprime [Scherer *et al.*, 1998] [Hansen, 1996].

Dans le second cas, la détermination de l'état émotionnel de l'utilisateur permet d'adapter la stratégie dialogique afin de fournir une réponse plus adaptée à sa requête [Lee *et al.*, 2002], [Devillers *et al.*, 2005]. Classiquement, si l'utilisateur présente des signes d'irritation ou de frustration face au répondeur automatique, une stratégie pourrait être de le rediriger vers un opérateur humain.

Les états émotionnels impliqués dans ces deux objectifs diffèrent en fonction du contexte applicatif. Les systèmes de dialogues développés pour des banques ou des services téléphoniques commerciaux, s'intéressent en priorité à des réactions de frustration ou d'irritation, les centres d'appel d'urgence à l'inquiétude ou à l'anxiété [Vidrascu et Devillers, 2005], et les systèmes de dialogue développés pour des applications militaires vont se focaliser sur les émotions telles que le stress cognitif ou physique. C'est le cas notamment des applications comme la reconnaissance de la parole chez des pilotes d'avions ou les pilotes de voitures [Varadarajan *et al.*, 2006] [Fernandez et Picard, 2003].

Les agents artificiels, et les modèles d'interaction émotionnelle.

Les agents artificiels cherchent à analyser et reproduire les comportements humains pour interagir socialement avec l'homme. Ainsi, le robot AIBO de Sony ou le robot Kismet du MIT [Breazeal et Aryananda, 2002] intègre les émotions dans leur modèle d'interaction. Aibo est capable d'exprimer six différents types d'émotion : la joie, la tristesse, la colère, la surprise, la peur et le mécontentement, ainsi que des combinaisons de ces émotions et ses humeurs par un simple jeu de couleur et de son. Kimset, lui, est capable d'exprimer, à travers une voix synthétique et des expressions faciales différentes, des émotions appelées ici, *qualités émotionnelles* : calme, colère, dégoût, peur, joie, tristesse et intérêt.

Une autre forme d'agent artificiel sont les Agents Conversationnels Animés (ACA) qui interviennent dans le but d'améliorer les interfaces homme-machine graphiques traditionnelles [Pelachaud, 2005] ou en tant que personnage intelligent d'un jeu vidéo. Il s'agit à la fois d'analyser les comportements émotionnels des utilisateurs ou des joueurs usuels face à une machine et, en fonction de cette analyse, de générer chez l'agent virtuel des *comportements émotionnels*.

L'indexation des films, le rire et les sons d'horreurs.

L'analyse du contenu affectif des films peut fournir des informations par exemple sur le genre du film. Ainsi, dans [Xu *et al.*, 2005], il s'agit de détecter des événements émotionnels tels que les rires où les manifestations d'horreur à travers la bande audio, dans des comédies et des films d'horreur dans le but de localiser les événements émotionnels forts dans le film.

1.3 Objectifs de recherche

Parce que le contenu émotionnel contribue au décodage sémantique des comportements humains, la reconnaissance automatique des émotions est un sujet qui suscite un intérêt croissant dans le domaine du traitement de la parole. Ce n'est cependant pas une entreprise aisée. Les émotions jouent en effet un rôle implicite dans le processus de communication, en comparaison du message explicite véhiculé par le niveau lexical et le phénomène à reconnaître est un phénomène complexe et subtil, présentant des manifestations très diversifiées et dépendantes de nombreux facteurs (contexte social, culturel, personnalité du locuteur, etc.).

Mes travaux de thèse s'engagent sur ce thème de recherche : la reconnaissance des émotions dans la parole. En considérant que les trois aspects indispensables du phénomène émotionnel sont les aspects cognitif, physiologique, et expressif, nous nous intéressons donc ici à la composante expressive de l'émotion, et plus précisément aux manifestations vocales de l'émotion.

Choix de l'émotion cible : les émotions de type peur

Nous avons choisi de centrer notre étude sur un type de manifestations émotionnelles jusqu'alors peu étudié dans le domaine : *les émotions de type peur*. Ce terme regroupe ici l'ensemble des *états émotionnels* liés à la peur, allant de l'inquiétude à la panique. Nous nous intéressons à la fois à la peur comme émotion primaire ou basique, ainsi qu'à ses dérivées avec des manifestations plus complexes ou plus modérées.

Les manifestations vocales de la peur peuvent être caractérisées par :

- des indices linguistiques (ex : « à l'aide ! ») ;
- des indices paralinguistiques qui incluent : des marqueurs non verbaux (ex : cris, sanglots, halètement, souffle saccadé, etc.) et des indices acoustiques (intensité de la voix, hauteur de la voix).

Nous nous focalisons sur l'étude des *phénomènes acoustiques* qui interviennent lors des manifestations émotionnelles. L'étude de ces phénomènes sera abordée indépendamment du contenu linguistique.

Le contexte étudié : les situations de menace pour la vie humaine

Les *situations anormales* visées sont définies de la manière suivante :

Définition 1. *On appelle* ***situations anormales*** *des événements imprévus, constituant une menace pour la vie humaine, qu'ils soient la conséquence d'une catastrophe naturelle (incendies,*

inondations) ou d'une action humaine (agressions physiques ou psychologiques contre un être humain, prises d'otage, attentats terroristes, etc.).

Cette définition de situations anormales s'est imposée dans le contexte industriel [Andrade *et al.*, 2005] qui concerne la protection civile et la prévention des états critiques.

L'application visée : la surveillance dans les lieux publics

Cette étude est motivée par une application nouvelle dans le domaine de la reconnaissance d'émotions : l'application de surveillance. Il s'agit de permettre à la machine de diagnostiquer les situations anormales, afin d'assister l'homme dans sa tâche de surveillance. A l'heure actuelle, la majorité des systèmes automatiques de surveillance existants s'appuient essentiellement sur la modalité vidéo pour détecter et analyser les situations anormales [Remagnino *et al.*, 2002]. Notre défi est d'intégrer dans ces systèmes l'information liée aux manifestations émotionnelles contenue dans le flux audio comme complément d'information à la vidéo.

1.4 Organisation du document

Ce document est organisé en trois parties. La première partie correspond aux deux étapes préalables, mais non moins fondamentales, nécessaires au développement d'un système automatique de reconnaissance d'émotions : l'acquisition et l'annotation d'un matériel d'étude des émotions de type peur en situations anormales. La deuxième partie est consacrée au développement du système de reconnaissance mis en oeuvre sur ce matériel. La troisième partie propose des éléments pour l'intégration de ce système dans une plateforme effective de détection de situations anormales.

Partie I : Émotions, corpus et annotation

Le premier défi à relever pour la recherche dans le domaine des émotions est la collecte de données émotionnelles. Le phénomène émotionnel est complexe, souvent imprévisible. L'acquisition d'enregistrements illustrant un tel phénomène soulève de nombreuses questions et il est parfois difficile d'obtenir un matériel d'étude qui soit en adéquation avec l'objectif de recherche. L'objet du **chapitre 2** est de s'attaquer à ce problème sous un angle original : illustrer une grande diversité de contextes (situations, locuteurs, etc.) pour un même type d'émotion, les émotions de type peur. L'acquisition de telles données est d'autant plus difficile à mettre en oeuvre, que les émotions de type peur en situations anormales sont des émotions qui surviennent rarement dans des contextes réels.

Le second défi, présenté dans le **chapitre 3**, concerne la définition d'une stratégie d'annotation du contenu émotionnel qui permette l'exploitation des données par un système de reconnaissance d'émotions. Quel aspect du phénomène émotionnel le système devra-t-il apprendre à interpréter ? Comment décrire ce phénomène en un langage qui soit à la fois naturel pour les annotateurs et assimilable par la machine ? De nombreuses études se sont penchées sur le problème de l'annotation du phénomène émotionnel en développant des modèles de représentations complexes qui s'appuient sur les théories des psychologues mais peu se sont consacré à l'exploitation effective de ces annotations par un système. L'originalité de la stratégie proposée est de considérer les manifestations émotionnelles dans leur contexte d'émergence en intégrant le niveau contextuel dans le schéma d'annotation.

Les questions qui se posent une fois le matériel annoté sont les suivantes : quel est le contenu émotionnel et contextuel du matériel d'étude ? Comment le système va-t-il pouvoir exploiter

ce matériel ? Comment va-t-il traiter les divergences des différents annotateurs ? La stratégie d'annotation adoptée fournit-elle des annotations suffisamment fiables pour être exploitables par le système ? C'est à ces questions que se consacre le **chapitre 4**.

Partie II : Analyser et reconnaître les manifestations émotionnelles

Les émotions de type peur sont souvent accompagnées de fortes modifications corporelles telles que la crispation, les tremblements, l'augmentation du rythme cardiaque. Par ailleurs les situations anormales sont propices à des actions du type fuite, poursuite, bagarre, etc.. L'activité physique qui accompagne alors la production du message oral entraîne l'émergence de manifestations non verbales telles que les respirations, les cris. Ces manifestations vont se répercuter sur le contenu acoustique du signal de parole. Le **chapitre 5** est dédié à la caractérisation de ces manifestations émotionnelles par des descripteurs acoustiques pertinents qui seront utilisés par le système de reconnaissance d'émotions. L'originalité de l'étude proposée réside dans la nécessité de caractériser l'ensemble de l'information émotionnelle contenue dans le signal de parole, i.e. à la fois ses portions voisées (ex : voyelles et consonnes voisées) et non voisées (ex : consonnes non voisées, respiration, chuchotement).

La spécificité majeure du système de reconnaissance des émotions qui est développé dans le **chapitre 6** a trait à la fusion des informations acoustiques extraites des portions voisées et des portions non voisées au sein du système. L'enjeu réside également dans l'appréhension de la diversité des contextes d'émergence au sein de la classe peur que l'on cherche à reconnaître.

Partie III : Vers une plateforme de surveillance effective

Le système de reconnaissance des émotions de type peur est destiné à être intégré dans le module audio d'une plateforme de surveillance. Ce module pourra prendre en compte d'autres types d'information pour caractériser une situation anormale. Nous avons divisé les événements audio pertinents en deux catégories :

- les événements de nature vocale (verbale et non-verbale) avec des manifestations émotionnelles typiques telles que l'agressivité ou même la folie chez l'agresseur ou la peur montante, la folie, le désespoir, l'inhibition chez la victime,
- les événements de nature non vocale avec des bruits tels que les explosions ou les coups de feu.

Le **chapitre 7** présente le couplage du système de reconnaissance des émotions de type peur avec un système de détection d'un événement non vocal typique des situations anormales, les coups de feu. L'intérêt de cette approche est la considération des événements anormaux qui correspondent au contexte d'émergence des manifestations émotionnelles comme des éléments permettant de caractériser la situation.

Première partie

Émotions, corpus et annotation

Résumé

Les émotions recherchées sont des émotions de type peur dans des situations dynamiques de menace (du danger potentiel à la menace passée) que nous appelons ici situations anormales. Ces situations sont rares et imprévisibles et par conséquent difficiles à collecter dans la vie réelle. Notre choix s'est porté sur un matériel d'étude original, les films de fiction. Ce matériel fournit une première solution pour l'étude des émotions de type peur en situations anormales et nous permet de poser les hypothèses concernant les types d'émotions émergeant dans ces contextes et leurs manifestations acoustiques. Le corpus développé sur ce matériel, le corpus SAFE illustre des manifestations d'états émotionnels très diversifiées dans les deux contextes d'émergence : situation anormale et situation normale. Sept heures d'enregistrement organisées en 400 séquences audiovisuelles ont été collectées à partir de 30 films récents dans leur langue originale, en anglais.

La stratégie d'annotation est définie en vue d'être transposée à des données réelles de surveillance. Elle fournit notamment une annotation des émotions en contextes dynamiques. Chaque séquence fournissant un contexte particulier est segmentée en une unité d'annotation de base que l'on appelle *segment*. La description de l'émotion au niveau du segment se fait par 4 catégories globales qui ont été définies en isolant l'émotion ciblée par l'application : la peur.
Les contextes social, personnel et situationnel sont intégrés à la description. Leur description nous fournit le matériel nécessaire au contrôle de la diversité émergeant du corpus SAFE. L'originalité de notre approche est de fournir une annotation détaillée de la situation de menace, i.e. des événements déclencheurs des manifestations émotionnelles en décrivant les différents aspects susceptibles d'influencer ces manifestations, tels que la gravité et le degré d'imminence de la menace.

Les tests perceptifs effectués sur un échantillon représentatif du corpus et la comparaison des trois annotations réalisées sur l'ensemble du corpus nous fournissent des éléments pour l'évaluation de la pertinence de la stratégie d'annotation en catégories globales et de la fiabilité des annotations. Les tests perceptifs permettent également de montrer le rôle de la modalité vidéo dans l'annotation.
L'analyse du contenu du corpus SAFE à partir des annotations met en évidence la présence de l'émotion cible : les émotions de type peur. Les segments annotés peur représentent environ 30% du corpus. La corrélation des contextes d'émergence (menace latente, immédiate) et des différentes manifestations de la peur (inquiétude, panique) souligne l'influence du contexte sur les types de manifestations émotionnelles.
Un sous-corpus d'étude est sélectionné. Il est composé des segments annotés peur ou neutre par deux des annotateurs. Une annotation en « condition système » (i.e. sans le support de la vidéo et du contexte temporel fourni par la séquence) est effectuée dans l'objectif d'être comparée avec les résultats obtenus par le système automatique de reconnaissance d'émotions.

Chapitre 2

Les émotions en situations anormales : stratégie d'acquisition

Sommaire

Introduction

Le domaine de l'analyse des émotions dans la parole requiert des données, i.e. des enregistrements de manifestations émotionnelles, qui viennent étayer les différents modèles théoriques du phénomène émotionnel et qui sont nécessaires à la mise en place de modèles computationnels. De telles études sont tributaires de la qualité des données utilisées. L'acquisition de corpus émotionnels qui soient en adéquation avec les objectifs de recherche constitue un sujet de recherche en soi sur lequel collaborent de nombreux laboratoires notamment au travers du réseau d'excellence HUMAINE[1].

Le paragraphe 2.1 répertorie les différents critères auxquels doivent idéalement répondre les corpus émotionnels pour être utilisés dans un objectif de détection automatique d'émotions.

Les conclusions tirées des études sur les bases de données existantes peuvent-elles s'étendre à des contextes pratiques différents ? Notre objectif est de développer un système capable de reconnaître les émotions de type peur. Les données recherchées ont pour vocation la construction d'un modèle (ici, acoustique) de l'émotion cible qui pourra servir de base à la détection. L'utilisation de corpus existants dans cet objectif de recherche est spécifiée dans le paragraphe 2.2.

Nous décrivons ensuite le matériel d'étude original sélectionné ici, *la fiction* et le corpus développé à partir de ce matériel : *le corpus SAFE (Situation Analysis in a Fictional and Emotional corpus)* (paragraphe 2.3).

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2006c]

2.1 Contexte et difficultés

2.1.1 Les critères de qualité

La qualité des données émotionnelles est conditionnée par l'objectif final de recherche [Douglas-Cowie *et al.*, 2003]. Dans un objectif de détection d'émotions, les critères de qualité concernent la manière dont les émotions et leur contexte d'émergence sont représentés dans les bases de données.

Authenticité des émotions illustrées

Les bases de données contiennent des émotions vécues ou des émotions simulées. Le niveau d'authenticité requis dépend de l'application visée. Par exemple, dans [Beller *et al.*, 2006], ce sont expressément des émotions actées, i.e. simulées par des acteurs, qui sont recherchées. En effet, l'objectif de ce travail est de permettre l'utilisation de voix de synthèse pour des applications artistiques, telles que la transformation ou la synthèse de voix d'acteurs pour le doublage de films d'animation.

En revanche, pour une application de détection dans des conditions réelles, la base de données doit idéalement illustrer des émotions vécues dans des contextes réels (i.e. des émotions appelées *real-life*). Or, les émotions vécues en contextes réels sont imprévisibles et il est difficile de contrôler ce qui est collecté.

Une solution adoptée par de nombreuses études pour éviter ce problème est d'utiliser des émotions simulées, en demandant par exemple à des acteurs de simuler les différentes émotions recherchées. Cependant, les manifestations des émotions simulées et des émotions vécues présentent des différences et les modèles appris sur de telles bases de données risquent d'être inutilisables sur

[1] http ://www.emotion-research.net

des données réelles [Batliner *et al.*, 2000]. Malgré ce constat, les bases de données illustrant des émotions simulées sont encore très largement utilisées. La représentativité de ce type de corpus a été évaluée dans l'article [Juslin et Laukka, 2003] : sur les 104 études répertoriées, 87% utilisent des enregistrements d'émotions actées.

Les types d'émotions et leur contexte d'émergence

Le type et la diversité des émotions représentées varient d'un corpus à l'autre. De plus les manifestations d'une émotion vécue sont conditionnées par le contexte dans lequel elle survient [Devillers, 2006]. Il importe donc de circonscrire les manifestations émotionnelles et les contextes d'émergence que l'on cherche à étudier et de vérifier l'adéquation avec le matériel d'étude.

Le contexte d'émergence de l'émotion rassemble :

- le contexte situationnel (lieu, situations et événements déclencheurs de l'émotion) ;
- le contexte d'interaction (ex : homme-homme, homme-machine, téléphonique) ;
- le contexte social (ex : agent/client pour les centres d'appel) ;
- le contexte personnel (ex : sexe, âge du locuteur) ;
- le contexte culturel ;
- le contexte linguistique (langue, dialecte) ;
- le contexte sonore (ex : environnement bruyant) ;
- le contexte intermodal (geste, parole) ;
- le contexte temporel (durée de chaque enregistrement).

Les contextes d'émergence illustrés doivent être aussi diversifiés que le requiert le système. Par exemple, pour des applications de type détection d'émotions dans des centres d'appel, le matériel d'étude doit présenter un nombre significatif de locuteurs différents afin que le système dispose de suffisamment de données pour pouvoir apprendre des modèles suffisamment génériques pour pouvoir reconnaître l'émotion chez un client inconnu. Cette diversité en termes de locuteur n'est pas nécessaire si l'objectif est de développer un système de détection d'émotions qui puisse être adapté à un locuteur spécifique pour des applications telles que la détection du stress chez un pilote d'avion ou un contrôleur aérien.

Les données doivent retranscrire le contexte intermodal (ex : apport du visuel) et temporel du message émotionnel tel qu'il a été produit. Si le canal utilisé par le locuteur pour transmettre son message est multimodal (ex : geste et voix), le support des données devra être idéalement audiovisuel. En revanche, dans le cas des corpus de centre d'appel, le message du locuteur est produit de manière à transmettre toute l'information à travers le canal linguistique et paralinguistique. Seul le support audio est alors nécessaire.

2.1.2 Les émotions recherchées

Nous cherchons à illustrer pour notre étude des émotions de type peur émergeant dans un contexte spécifique, un contexte de menace pour la vie humaine i.e. en situations anormales[2]). Ce besoin est motivé par l'application visée par notre étude : l'application de surveillance. Il est de plus nécessaire que la base de données illustre l'évolution de la menace, du danger potentiel à la menace immédiate, dans laquelle les manifestations émotionnelles émergent. Nous recherchons donc des données illustrant des émotions « prises sur le vif », c'est-à-dire au coeur de l'action, et qui évoluent en fonction de la menace.

De plus les critères requis pour l'application de surveillance sont spécifiques. D'une part, l'application de surveillance nécessite une grande diversité illustrant les variables contextuelles

[2]telles que définies dans le chapitre introductif et dans le glossaire

pertinentes pour l'application (type de locuteurs, contexte social, relation entre les locuteurs, types de lieu). D'autre part, la perspective d'une analyse multimodale impose de disposer d'un corpus audiovisuel. En effet, l'objectif final est d'intégrer le système de détection de la peur dans une plateforme multimodale prenant en compte également des indices visuels.

2.2 Les bases de données émotionnelles et la peur

Les bases de données existantes dans le domaine de la parole émotionnelle[3] peuvent être classées en trois types : actées, élicitées ou *real-life*. A partir de l'état de l'art fourni dans [Douglas-Cowie *et al.*, 2003], nous présentons pour chaque type de corpus un bilan évaluant dans quelle mesure ces corpus présentent des manifestations émotionnelles liées à la peur, ainsi qu'une évaluation de la qualité de ces corpus en fonction de notre objectif de recherche.

2.2.1 Les bases de données actées

La majorité des bases de données mises au point pour l'étude de l'émotion dans la parole font intervenir des émotions simulées. Le réalisme des émotions actées dans ces bases de données est variable et dépend :

- des sujets chargés de simuler les émotions (acteurs professionnels ou non) ;
- du contexte fourni aux sujets pour simuler les émotions (contenu sémantique des phrases prononcées en adéquation avec l'émotion simulée, scénario d'interaction).

Dans certaines bases de données actées, une phrase, un mot, ayant un contenu sémantique neutre voire même une suite delexicalisée sont lus hors contexte dialogique interactionnel par différents sujets non professionnels devant simuler des émotions données. Le réalisme des émotions actées illustrées dans ce type de corpus et leur variabilité en terme de locuteurs et de contexte sont par conséquent très réduits mais permettent par là même un meilleur contrôle de leur contenu. Les études sur ce type de corpus permettent en effet une analyse des variations acoustiques liées à l'émotion qui ne soient pas bruitées par les variations acoustiques liées au contenu linguistique. De ce fait, ces corpus sont abondamment représentés et ce sont, à l'heure actuelle, ceux où des émotions extrêmes, telles que la peur, trouvent le plus d'illustrations. Parmi eux, on trouve essentiellement des corpus audio [Mozziconacci, 1998], [Kienast et Sendlmeier, 2000], [van Bezooijen, 1984], [Abelin et Allwood, 2000], [Yacoub *et al.*, 2003].

D'autres bases de données actées offrent des illustrations plus crédibles : les phrases utilisées présentent un contenu verbal émotionnel cohérent et le contexte introduit autour de la phrase est de plus en plus large.

Les études les plus réalistes sont celles faisant intervenir des émotions jouées par des acteurs professionnels dans des scénarios d'interaction entre les locuteurs. À ce niveau de réalisme, on trouve seulement deux bases de données qui traitent de la peur, fournissant uniquement le support audio ([McGilloway, 1997], [Dellaert *et al.*, 1996]), mais également des corpus audiovisuels [Banse et Scherer, 1996], [Bänziger *et al.*, 2006].

Bien que les corpus actés existants présentent des illustrations d'émotions de type peur, ils restent difficilement exploitables dans le cadre d'applications réelles car ils illustrent pour la plupart des états émotionnels stéréotypés voir caricaturaux, loin de la complexité et de la diversité émotionnelle émergeant dans des interactions réelles.

[3] http ://emotion-research.net/deliverables

2.2.2 Les bases de données élicitées

Il existe différents types de scénario d'élicitation pour recueillir des manifestations émotionnelles naturelles. Le premier type de scénario très répandu dans le domaine de l'interaction homme-machine repose sur le paradigme du Magicien d'Oz : eWIZ [Audibert *et al.*, 2004a], SAL [Douglas-Cowie *et al.*, 2003], SMARTKOM[4]. Alors que le sujet croit communiquer avec son ordinateur, le comportement apparent de l'application est perturbé à distance par un complice humain, le « magicien », afin d'induire certains états émotionnels chez les sujets. Cette technique présente l'avantage d'offrir un contrôle des contenus linguistique et phonétique grâce à l'usage d'un langage de commandes qui contraint les expressions vocales des sujets.

Le deuxième type de scénario plus proche des contextes recherchés repose sur l'intervention d'un complice humain dans une situation réelle, avec par exemple l'étude par Williams et Stevens des interactions à l'intérieur d'un cockpit [Williams et Stevens, 1972].

Ce type de base de données nécessite un scénario d'élicitation des émotions qui peut être difficile à mettre en place selon l'émotion visée. Ainsi, l'induction d'émotions extrêmes de type peur peut s'avérer médicalement dangereuse et peu éthique.

2.2.3 Les bases de données *real-life*

La plupart des corpus *real-life* existants sont des corpus présentant des émotions de la vie de tous les jours [Campbell et Mokhtari, 2003], plus modérées et plus contenues que les émotions recherchées pour notre application. Ceci est dû au fait que les émotions intenses spontanées sont plus rares et plus imprévisibles que les émotions de la vie de tous les jours. Certains corpus spontanés, cependant, fournissent un contenu émotionnel plus intense avec des émotions de type peur. Nous présentons ici les différents corpus illustrant ce type d'émotion afin d'évaluer leur potentiel usage pour une application de surveillance en fonction des contextes illustrés.

Les corpus *real-life* et les émotions « prises sur le vif »

Une solution adoptée afin de répondre au problème du caractère imprévisible de données authentiques pour des émotions intenses, consiste en l'utilisation d'un matériel basé sur les interviews, le locuteur étant invité par l'interviewer à raconter et, par là même, à revivre des expériences émotionnellement intenses : Reeding-Leeds [Roach *et al.*, 1998], le corpus de Belfast [Douglas-Cowie *et al.*, 2003], et un corpus d'interviews après le constat de perte de bagages [Scherer et Ceschi, 2000]. Le corpus EmoTV [Abrilian *et al.*, 2005] contient également des interviews extraits de programmes télévisés et notamment des interviews diffusés lors des informations de 20h, tels que des témoignages après des affaires juridiques, des interviews sur les grèves, sur des problèmes de société, etc. Les interviews permettent d'obtenir des émotions fortes naturelles, mais qui restent cependant des émotions hors situation, car éprouvées dans le cadre d'un témoignage d'événements passés. Ce sont des émotions a posteriori qui auront des manifestations différentes et majoritairement moins extrêmes que celles « prises sur le vif ».

Les corpus *real-life* et la diversité des contextes illustrés

Dans les centres d'appels, des émotions d'intensités plus ou moins fortes sont également illustrées dans les corpus audio. Pour des applications financières les interactions sont très policées et les émotions assez mesurées [Devillers et Vasilescu, 2003], par contre des manifestations beaucoup plus intenses sont présentes lors d'appels d'urgence médicale [Vidrascu et Devillers,

[4]http ://www.phonetik.uni-muenchen.de/Bas/BasMultiModaleng.html#SmartKom

2005]. Différentes illustrations de la peur sont présentes dans ce corpus. Cependant le contexte d'émergence de la peur dans ce corpus est un contexte applicatif spécifique. D'une part, c'est un contexte téléphonique et le message est destiné à être transmis par le canal linguistique et paralinguistique uniquement, d'autre part les appelants tentent de contrôler leurs émotions afin de pouvoir expliquer la situation et obtenir de l'aide.

Un autre type de corpus consiste en des sessions de thérapie et des conversations téléphoniques, des sessions d'évaluation post-thérapie dans le cadre de dépressions et d'états suicidaires [France *et al.*, 2003]. Le contexte illustré dans ce corpus est très spécifique et difficilement exploitable pour une autre application.

Les corpus real-life qui contiennent les émotions les plus intenses sont des corpus illustrant des types de situations éloignées des contextes situationnels, d'interaction et sociaux visés par notre application.

2.3 Le corpus SAFE et le cinéma de fiction

2.3.1 Le cinéma de fiction pour l'illustration d'émotions de type peur

Nous venons de voir à travers un parcours des différentes bases de données existantes que les émotions de type peur sont très peu représentées avec la diversité contextuelle recherchée. Nous avons également souligné que l'authenticité des émotions illustrées est un des critères de qualité requis pour une application destinée à fonctionner sur des données réelles. Malheureusement de telles données sont difficiles à acquérir pour les émotions et les contextes recherchés qui sont plus rares et imprévisibles. Il est par exemple difficile de recueillir en nombre suffisant des enregistrements de catastrophes telles que celles du 11 septembre 2001, compte tenu de leur caractère confidentiel. Les informations télévisées fournissent également des exemples forts des manifestations de ces émotions mais généralement par l'intermédiaire de courts extraits, souvent commentés (voix-off) ou par le récit a posteriori de personnes témoins de l'événement. Compte tenu de ces difficultés, notre choix s'est porté sur un matériel d'étude original, *la fiction*, plus précisément le cinéma de fiction. Ce matériel fournit une première solution pour l'étude des émotions de type peur en situations anormales. Dans les paragraphes suivants, nous présentons les avantages et les inconvénients d'un tel corpus.

Fiction et diversité

Avantages : la fiction fournit un matériel audiovisuel émotionnel de qualité, par la diversité des locuteurs (sexe, milieu social) et des contextes d'émergence des émotions (lieu, contexte d'interaction, contexte social, conditions d'enregistrement etc.). Le cinéma de fiction offre en effet une grande diversité de scénarios. Une telle diversité est rare dans les corpus existants et difficile à obtenir dans des conditions réelles.

Fiction et réalisme des manifestations émotionnelles

Inconvénients vis-à-vis des données réelles : les données de fiction sont des données actées susceptibles d'illustrer des états émotionnels stéréotypés différents de ceux émergeants dans des interactions réelles.

Avantages vis-à-vis des donnnées actées : les émotions illustrées dans le corpus SAFE sont interprétées par des acteurs professionnels dans une situation d'interaction entre personnes et dans le contexte défini par le scénario du film. Dans le cinéma de fiction l'acteur dispose

des objectifs et du contexte dans lequel se trouve son personnage, pour illustrer les différentes émotions correspondant à l'action qu'il est en train de réaliser. Cette mise en situation de l'acteur tend à améliorer le réalisme des émotions jouées [Enos et Hirshberg, 2006] [Bänziger *et al.*, 2006]. Un autre facteur d'amélioration est la durée du tournage qui s'étale sur plusieurs mois et favorise l'identification de l'acteur au personnage.

Fiction et réalisme des conditions d'enregistrement

Inconvénients : le contexte sonore de la parole dans les bandes son de fiction ne répond pas toujours à un souci de fidélité au réel. Le cinéma distingue deux types de bandes son, les bandes son qui utilisent le son direct (son enregistré à la prise de vue) et les bandes son qui utilisent la post-synchronisation [Chion, 1990]. Pour certains films, la *prise de voix* est directe et les sons d'ambiance sont refabriqués par le bruitage ou récupérés dans des sonothèques pour les ajouter plus tard au montage ou au mixage. Par exemple, dans les films d'action, le son accompagnant un coup de poing est souvent exacerbé et peut correspondre à un son de bruitage d'un événement acoustique très différent du coup de poing. Certains réalisateurs sont des opposants fervents à une *prise de voix* post-synchronisée[5]. Ces choix correspondent à des esthétiques individuelles de réalisateurs et il est difficile de faire émerger des tendances claires sur ce sujet. Aux États-Unis, on estime que 60 à 70% du dialogue d'un film est du son direct [Thomas, 2005]. Un autre inconvénient majeur de la fiction est la musique souvent utilisée au cinéma pour rythmer l'action et accentuer les passages émotionnels forts.

Avantages : dans les bandes son qui conservent le son direct, la prise de son tend à refléter la réalité. Les mouvements du locuteur impliquant des variations naturelles dans la reproduction du niveau sonore de sa voix sont la plupart du temps respectées même si le locuteur principal de la scène est souvent audible ce qui ne sera pas forcément le cas en contexte réel, où il sera plus ou moins éloigné du micro fixe utilisé.

Fiction et support contextuel

Avantages : le support contextuel fourni par la fiction est particulièrement riche, à la fois temporel (durée du film), visuel et sonore. Un tel support contextuel est très rare dans les bases de données existantes.

2.3.2 Méthode de sélection des séquences audiovisuelles

Le corpus de fiction développé ici, le corpus SAFE, est basé sur des séquences extraites d'une collection de films récents dans leur langue originale, en anglais[6]. Le corpus de fiction illustre des manifestations d'états émotionnels dans les deux contextes d'émergence : situation anormale (71%) et situation normale (29%) et dans des situations individuelles, de groupe ou de foule.

Les critères de sélection reposent sur la crédibilité du jeu de l'acteur simulant la peur, la qualité de la bande sonore de la séquence (ex : prédominance de la voix sur les bruits et la musique), sur la diversité des types de manifestations émotionnelles en situations anormales et de leur stimuli, sur le degré de réalisme des situations illustrées. En ce qui concerne ce dernier critère, les films policiers (« thrillers »), les films d'horreur ou d'action, les drames psychologiques, les films

[5] cf Straub et Huillet, Cahier du cinéma n°260-261 : « Le film doublé trompe. ... les lèvres qui remuent sur l'écran ne sont pas les lèvres qui prononcent les paroles que l'on entend, ... »

[6] anglais américain pour la plupart, voir chapitre 4

reconstituant certaines périodes récentes de l'histoire, les films sur des catastrophes climatiques ou des faits divers constituent de très bons candidats pour notre corpus.

Définition 2. *On appelle* ***séquence****, une portion de film référant à un même contexte situationnel (ex : prise d'otage, inondation du métro, etc.). La durée des séquences dépend de la façon dont ce même contexte situationnel est illustré dans le film. Un contexte situationnel récurrent dans le film mais illustré à différents moments donnera lieu à plusieurs séquences.*

400 séquences audiovisuelles en anglais de 8 secondes à 5 minutes, soit un total de 7 heures d'enregistrement, sont ainsi sélectionnées. La durée des séquences dépend de la manière dont la situation est illustrée et segmentée dans le film. Ces séquences ont été extraites de 30 films différents, dont la liste est fournie en annexe A. Les manifestations émotionnelles représentées sont variées. En dehors des émotions ciblées par notre application, c'est à dire les émotions de type peur survenant lors de situations anormales, le corpus illustre d'autres états émotionnels et interactions orales survenant lors de situations normales.

2.4 Conclusion

Nous avons identifié ici nos besoins en termes de corpus émotionnel. Nous recherchons des enregistrements d'émotions de type peur dans des contextes de menace. Cette première étape préalable au développement d'un système de reconnaissance des émotions dans la parole est fondamentale. La qualité des modèles computationnels utilisés par la suite dépendra fortement de la qualité des données collectées et de leur adéquation avec l'application visée : la surveillance.

Le matériel d'étude utilisé est original : le cinéma de fiction. Ce matériel offre une première solution pour l'étude des émotions de type peur en situations anormales. Les contextes d'émergence illustrés dans le cinéma de fiction sont très variés et cette diversité aurait été difficile à obtenir avec des données réelles. Par rapport aux corpus actés classiques, le réalisme de l'émotion jouée est accrue par la qualité du jeu des acteurs professionnels, jeu qui est enregistré dans une *situation d'interaction* entre personnes et dans le contexte défini par le scénario du film.

Sept heures d'enregistrements illustrant les deux contextes d'émergence, situations normales vs. situations anormales, ont été collectées. Le corpus SAFE est le seul corpus audiovisuel qui illustre une telle diversité de contextes (locuteur, lieu, situation).

L'étude réalisée par la suite est effectuée sur ce corpus, i.e. sur des données actées qui peuvent s'avérer stéréotypées et différentes des émotions vécues dans des conditions réelles. Les études démontrant le fossé existant entre émotions actées et émotions vécues utilisent des données actées enregistrées en laboratoire avec des scénarios très restreints ou quasi inexistants fournis aux acteurs pour l'interprétation des émotions [Bänziger *et al.*, 2006]. Il n'existe à notre connaissance aucune étude comparant des émotions interprétées dans des fictions avec des émotions réelles. Cette étude pourrait être une perspective particulièrement intéressante de ce travail.

Chapitre 3

Les émotions en situations anormales : stratégie d'annotation

Sommaire

Introduction

Une fois le matériel collecté, la seconde étape nécessaire au développement d'un système de reconnaissance d'émotion est l'annotation du contenu émotionnel de la base de données. Cette étape d'annotation permettra en effet l'exploitation des données pour la modélisation des différents états émotionnels à reconnaître.

L'enjeu de cette étape réside dans l'élaboration d'une stratégie d'annotation destinée à guider les annotateurs dans leur tâche et qui permette de limiter les divergences liées à la subjectivité de l'annotateur. Les contextes socioculturel et psychologique du récepteur (l'auditeur ou l'annotateur) entraînent une sensibilité différente dans la perception des émotions.

La tâche d'annotation des émotions est également rendue difficile par la complexité du message oral communiqué par l'émetteur (le locuteur). Selon Scherer [Scherer *et al.*, 1980], la parole émotionnelle est conditionnée par deux effets pouvant donner lieu à des manifestations contradictoires : une excitation physiologique accrue « pousse » les vocalisations dans une certaine direction (effet *push*), alors que les tentatives conscientes de contrôle les « tirent » dans une autre direction et consistent en l'adoption de styles de langage culturellement acceptés (effets *pull*).

Les différentes théories de la communication [Chung, 2000] témoignent également de cette complexité à l'émission et la réception.

Le paragraphe 3.1 dresse un bilan des deux approches catégorielle et dimensionnelle couramment utilisées dans la littérature pour décrire les émotions. Ces deux approches sont combinées au sein de notre stratégie d'annotation qui présente la particularité d'intégrer une description du contexte d'émergence des émotions. La stratégie adoptée pour la description du niveau contextuel est expliquée dans le paragraphe 3.3. La stratégie de description du niveau émotionnel est ensuite détaillée dans le paragraphe 3.2.

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2004], [Clavel *et al.*, 2006c]

3.1 Les descripteurs émotionnels - Bilan

3.1.1 Descripteurs catégoriels

L'approche catégorielle consiste en la dénomination des émotions par des items lexicaux adaptés et prédéfinis. C'est la manière la plus intuitive pour décrire des émotions spécifiques, en utilisant des catégories issues du langage courant.

Catégories émotionnelles et type de corpus

La définition de catégories émotionnelles consiste à tracer des frontières absolues dans l'espace perceptif. Chaque catégorie émotionnelle correspond alors à un prototype [Kleiber, 1990] auquel peut se rattacher d'autres manifestations émotionnelles similaires.

L'établissement de telles frontières va fortement dépendre du matériel émotionnel utilisé. Dans le cas des corpus entièrement simulés : on part de prototype prédéfini que l'on cherche à illustrer. Toutes les manifestations émotionnelles contenues dans les corpus convergent donc fortement vers ce prototype. Pour les corpus spontanés le processus est inversé il s'agit de regrouper des manifestations émotionnelles non contrôlées autour de prototype abstrait. La complexité de ce regroupement est accrue par la diversité des contextes d'émergence des émotions en parole spontanée.

Les « Big Six »

Un grand nombre d'études consacrées à l'analyse des émotions dans la parole fait référence à un nombre minimal d'émotions dites basiques qui varie d'un modèle théorique à l'autre[7]. Les « Big Six » d'Ekman [Ekman, 1999] (colère, peur, tristesse, joie, dégoût et surprise) constituent les catégories d'émotions les plus utilisées, surtout dans le cadre d'études réalisées sur des corpus actés. Cependant ce type de description s'avère insuffisant pour la description des états émotionnels plus subtils tels que ceux présents dans des corpus réels, où les émotions peuvent s'avérer mélangées entre elles ou avec des manifestations correspondant plutôt à des attitudes [Devillers *et al.*, 2005].

Les listes de termes émotionnels

Des listes plus exhaustives ont également été établies, qu'elles se basent sur des études théoriques ou des études empiriques. Nous citerons par exemple la liste de termes en anglais établie dans [Whissel, 1989] ou le cône des émotions de Plutchik (figure 3.1) (étude théorique) et celle en français établie dans l'article [Galati et Sini, 1995] (étude empirique) dont l'objectif est de répertorier les termes utilisés dans le langage courant pour décrire directement une émotion ou y faire référence. L'un des objectifs du réseau d'excellence HUMAINE est de proposer des listes consensuelles de catégories.

3.1.2 Descripteurs dimensionnels

Né de la psychologie, ce mode de description représente les états émotionnels sur des axes abstraits. Plusieurs dimensions ont été proposées, chacune reposant sur des théories différentes. La dimension de *valence* est la plus utilisée. Elle consiste en la description des états émotionnels sur un axe positif/négatif. Une émotion est considérée comme positive ou négative, lorsque le locuteur semble avoir une évaluation positive, respectivement négative, des événements, des choses ou des personnes. Par exemple, les termes suivants sont placés de gauche à droite sur l'axe négatif-positif : dégoût, colère, peur, surprise, excitation, joie.

Une variante de la *valence* est la dimension du *plaisir/déplaisir* qui est relative au degré de plaisir que procure le sentiment ressenti. Les émotions agréables ou positives, comme par exemple la joie, accompagnent la survenue ou l'anticipation d'événements gratifiants. Les émotions négatives, comme la peur, sont associées à l'expérience ou l'anticipation d'événements désagréables (punition, danger...).

L'axe *activation* découle de la théorie de Darwin [Darwin, 1872] selon laquelle les états émotionnels sont caractérisés par des dispositions à agir d'une manière plutôt qu'une autre. Osgood [Osgood *et al.*, 1975] définit l'activation comme un état d'excitation, de calme à fortement excité, lié à l'intensité. L'activation représente ainsi le niveau d'excitation corporelle, qui s'exprime par des réactions physiologiques, comme l'accélération du coeur, la transpiration.

Cette dernière dimension est couramment couplée avec la *valence* pour former l'*espace activation-évaluation*. Les termes émotionnels peuvent être décrits par des coordonnées de points dans cet espace [Whissel, 1989] et forment un motif approximativement circulaire. Les deux dimensions activation et évaluation capturent une large proportion des variations émotionnelles. Ainsi, le logiciel Feeltrace [Cowie *et al.*, 2000] permet l'annotation selon ces deux dimensions de manière continue (voir figure 3.2). Cependant les termes référant à des émotions au sens strict ne

[7] voir chapitre 1

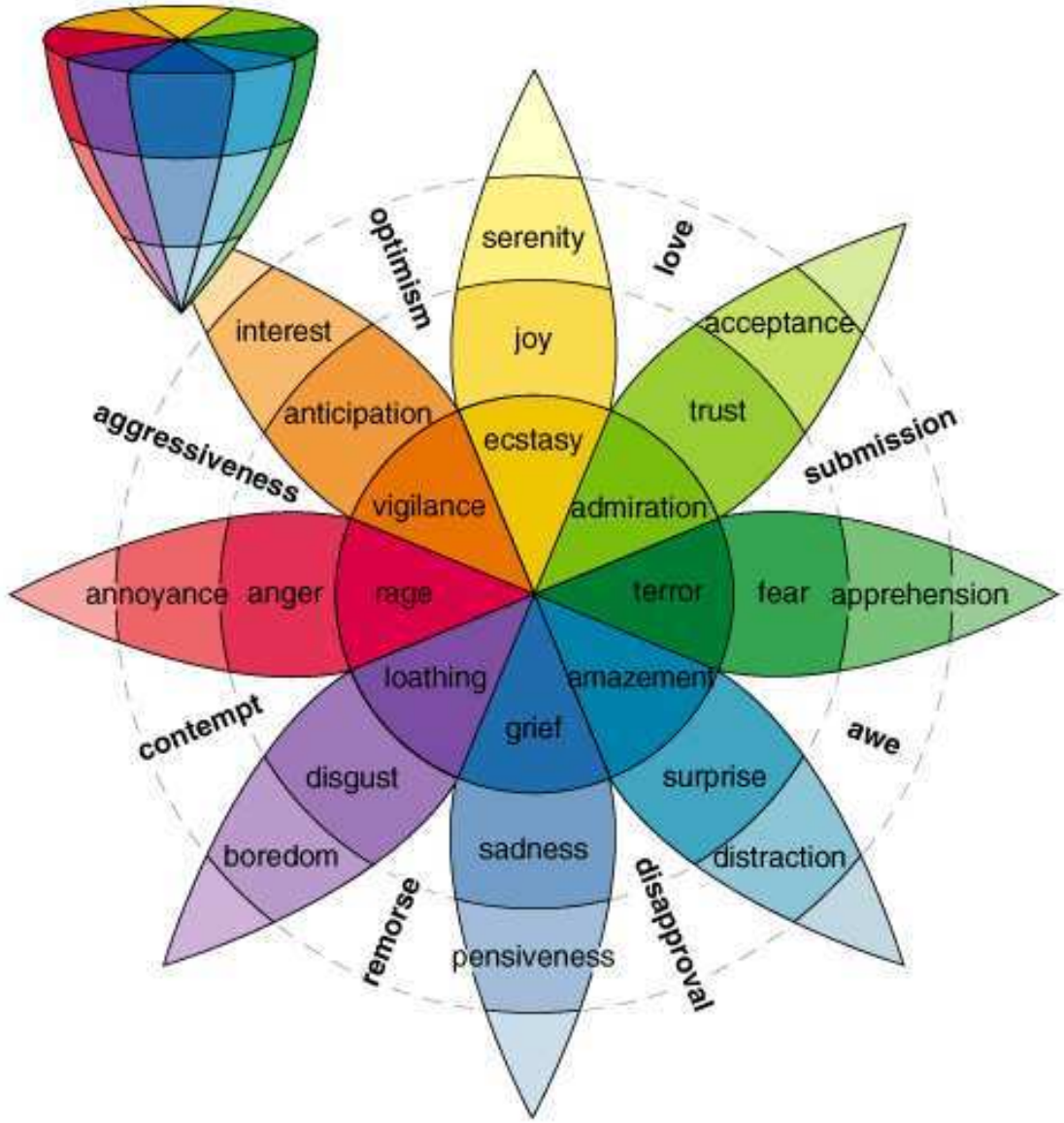

FIG. 3.1 – *Cône des émotions de Plutchik tiré de [Plutchik, 1984]. Les émotions élémentaires sont placées dans une roue. Les émotions secondaires correspondent à des mélanges d'émotions primaires (ex : la soumission résulte d'un mélange entre la peur et l'acceptation). La roue peut être transformée en cône afin de représenter les différents degrés d'intensité des émotions primaires et secondaires.*

sont pas également distribués dans cet espace. Ainsi la peur et la colère sont quasi-superposées à la fois sur l'axe de valence et l'axe activation.

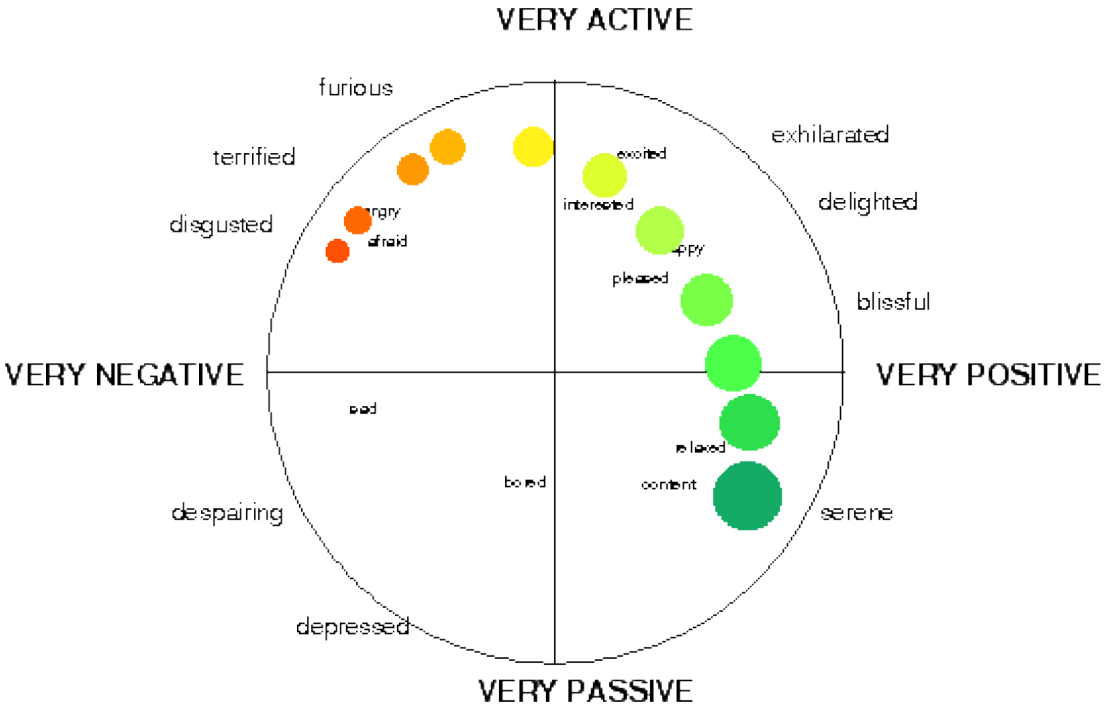

FIG. 3.2 – *Exemple d'annotation des variations émotionnelles au cours du temps sous Feeltrace. Des étiquettes verbales correspondant aux différents états émotionnels sont placées sur et dans le cercle pour permettre à l'utilisateur de se repérer sur ces deux dimensions.*

Une troisième dimension, nommée *contrôle*, *dominance* [Russell, 1997] ou *power* [Pereira., 2000], est ajoutée à cet espace. Elle correspond à l'effort du locuteur pour contrôler son émotion et permet de distinguer les émotions provoquées par le sujet lui-même ou par l'environnement, par exemple le mépris de la peur.

3.1.3 Le point de vue système

La description en catégories permet de fournir une stratégie de description qui soit plus facilement assimilable par l'annotateur, car il fait appel à des mots issus du langage courant. Dans l'objectif de développer un système de reconnaissance d'émotions, l'utilisation de catégories en sortie du système permet également de fournir à l'utilisateur des descriptions qui lui soient compréhensibles. De plus, le choix des catégories peut être guidé par l'application en isolant les catégories d'émotion que l'on veut que le système reconnaisse et choisir le niveau de finesse dans la discrimination des émotions auquel le système doit ou peut accéder.

3.2 Stratégie d'annotation du contenu émotionnel

Nous avons choisi de combiner les deux types d'approches décrites précédemment pour la description du contenu émotionnel du corpus. Cette démarche rejoint celle adoptée par le lan-

guage d'annotation EARL[8] défini récemment dans le cadre du réseau d'excellence HUMAINE. L'objectif premier de notre stratégie d'annotation est de proposer une annotation des états émotionnels présents dans le corpus en des catégories pertinentes du point de vue de l'application. L'enjeu est de s'affranchir des spécificités du corpus d'étude, en proposant une stratégie d'annotation, transposable à d'autres corpus dédiés à la même application. Ce niveau de généricité vis-à-vis du corpus peut être augmenté par une généricité vis-à-vis de l'application : les descripteurs correspondant à ce niveau de généricité sont des descripteurs transposables à d'autres applications.

3.2.1 Le *segment* : unité d'annotation

La stratégie d'annotation adoptée se focalise sur la description des manifestations orales de l'émotion. Dans cet objectif, les séquences audiovisuelles du corpus SAFE sont considérées comme des interactions orales entre plusieurs locuteurs, qu'il va falloir segmenter au préalable en tours de locuteurs. L'unité d'annotation de l'émotion, i.e. l'intervalle temporel sur lequel l'émotion est décrite ici, constitue une subdivision de ces tours de locuteurs et est appelé le *segment*.

Définition 3. *On appelle* ***segment****, un tour de parole ou une partie du tour de parole avec un contenu émotionnel homogène. Ainsi dans le cas d'un long monologue présentant des ruptures dans les manifestations émotionnelles, le tour de parole sera segmenté en plusieurs parties.*

La rupture est définie comme un changement des manifestations émotionnelles en fonction des descripteurs catégoriels et dimensionnels décrits ci-dessous.

3.2.2 Des descripteurs catégoriels intégrant différents niveaux de généricité vis-à-vis du corpus

La description de l'émotion au niveau du *segment* se fait par des catégories globales qui ont été définies en isolant l'émotion ciblée par l'application : la peur. Cette catégorie regroupe l'ensemble des états émotionnels qui se rattachent à la peur, incluant non seulement les manifestations primaires de la peur, qui interviennent comme une réaction de survie à la menace, mais aussi ses manifestations plus modérées et plus complexes en fonction du degré d'imminence de la menace.

Les émotions sont définies ici à partir de classes dérivées de la distinction positif/négatif (voir paragraphe 3.1.2). Les quatre classes proposées sont les suivantes :

- peur : cette classe regroupe tous les états émotionnels qui se rattachent à la peur de l'inquiétude à la terreur.
- autres émotions négatives : les émotions regroupées dans cette classe correspondent à l'ensemble des émotions négatives autres que la peur. Celles-ci peuvent survenir aussi bien dans des situations normales que des situations anormales ;
- neutre : cette classe regroupe les segments où peu ou aucune émotion ne semble être présente ;
- émotions positives : cette classe regroupe les segments où des émotions positives sont présentes, ces émotions surviennent principalement dans des situations normales.

Le concept d'état neutre étant problématique, nous l'utiliserons comme dans [Devillers, 2006] avec la définition suivante :

Définition 4. ***L'état neutre*** *est défini comme un état de valence non négatif et non positif d'activation faible et de contrôle élevé.*

[8] Emotion Annotation Representation Language

C'est le niveau d'activation qui va donc déterminer la catégorisation d'un segment en l'état neutre. Le seuil d'activation considéré va dépendre du référentiel utilisé et des émotions étudiées pour l'application. Ainsi dans notre application, les émotions recherchées sont globalement assez intenses.

Les manifestations émotionnelles présentes dans le corpus sont parfois mélangées entre elles. C'est le cas notamment de la peur et de la colère qui sont fréquemment présentes en même temps dans un segment. La présence d'émotions mélangées a été également constatée dans des données réelles [Devillers *et al.*, 2005]. Leur présence dans notre corpus montre la complexité des émotions illustrées dans le cinéma de fiction, une complexité proche de la complexité émergente en contextes réels.

Afin de pouvoir décrire cette complexité, l'annotateur a ainsi la possibilité de choisir deux catégories, la première catégorie étant considérée comme prédominante par rapport à l'autre.

De plus l'annotation de chacune de ces catégories globales est suivie d'une annotation en sous-catégories permettant à l'annotateur de préciser la ou les catégories globales sélectionnées. Ces sous-catégories (cf. schémas 3.3, 3.4 et 3.5) vont permettre d'accompagner l'annotation pour une analyse plus fine des catégories globales. La démarche utilisée pour le choix de ces sous-catégories sera présentée dans le chapitre suivant.

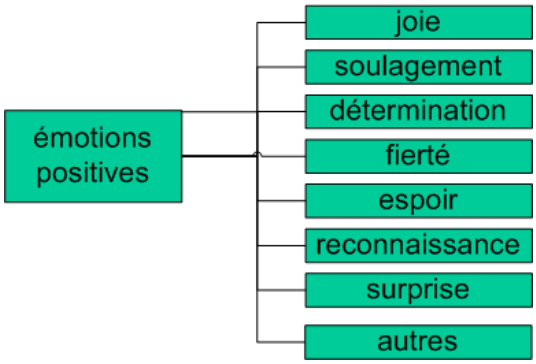

FIG. 3.3 – *Sous-catégories de la classe émotions positives*

D'un point de vue système, cette structure hiérarchique d'annotation permet de fournir différents niveaux de discrimination : une première distinction des différents états émotionnels avec les catégories globales (par exemple une première distinction peur/autres), puis une distinction plus fine des différents états émotionnels survenant dans les situations anormales avec l'annotation des sous-catégories. Par ailleurs, les systèmes de détection basés sur l'apprentissage des différentes classes nécessitent de disposer de suffisamment de données pour représenter chacune des classes, ce qui est peu faisable dans le cas de classes trop nombreuses.

3.2.3 Des descripteurs dimensionnels intégrant différents niveaux de généricité vis à vis de l'application

La description dimensionnelle des différents états émotionnels présents dans le segment est réalisée sur une échelle discrète. Une échelle continue comme celle utilisée dans le logiciel feeltrace permettrait une description plus fine des émotions mais plus complexe à traiter par un système de reconnaissance des émotions. De plus le logiciel ANVIL n'est pas adapté pour une annotation continue.

Nous avons choisi de considérer les trois dimensions abstraites suivantes :

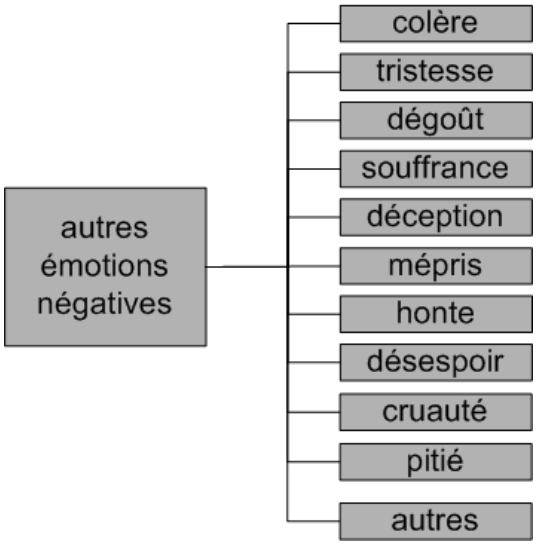

FIG. 3.4 – *Sous-catégories de la classe autres émotions négatives*

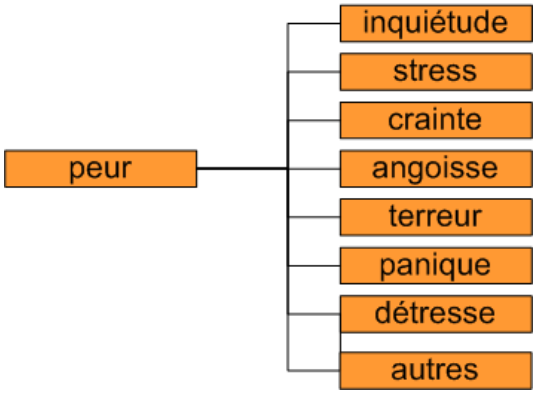

FIG. 3.5 – *Sous-catégories de la classe peur*

Intensité

Cette dimension consiste en l'évaluation du degré de l'intensité de l'émotion (fort ou faible) sur une échelle allant de 0 (neutre) à +3. Afin de fournir un référentiel adapté aux émotions présentes dans le corpus, des exemples typiques de peur sont fournis pour chaque niveau d'intensité (protoInt1, protoInt2 et protoInt3). Nous avons préféré utiliser la dimension *intensité* plutôt que la dimension *activation* car elle est plus axée sur les manifestations orales de l'émotion. Elle a été utilisée dans cet objectif notamment dans [Craggs et Wood, 2004].

Évaluation

Cette dimension consiste en l'évaluation du degré de valence (cf paragraphe 3.1.2) de l'émotion sur un axe négatif-positif. L'échelle proposée ici va de -3 du côté négatif de l'axe à 3 du côté positif de l'axe.

Réactivité

Cette dimension a été créée pour les besoins de l'application. Elle permet d'annoter les différents degrés de réactivité du locuteur face à la menace, c'est-à-dire son comportement face au stimulus engendrant ces manifestations de l'émotion. Cette troisième dimension permet la distinction des émotions en situations anormales, et notamment les différentes manifestations de la peur. Elle s'appuie sur la théorie d'Ekman [Ekman, 2003], selon laquelle les manifestations de la peur ne sont pas les mêmes si la victime fait face à la menace ou non. Elle correspond à une adaptation de la troisième dimension *contrôle* présentée précédemment en fonction de notre objectif de recherche. Un champ *none* (vide) est proposé pour le cas où le locuteur ne se trouve pas en présence de menace. Pour chaque degré de réactivité, voici à titre indicatif un exemple de comportement :

- degré 0 : le locuteur est en état de choc et est soit inhibé (pas de manifestations) soit il réagit de façon incohérente (folie).
- degré 1 : le locuteur réagit vocalement (cris, paroles), mais n'entreprend pas ou ne peut pas entreprendre de réaction physique/gestuelle pour contrôler la situation.
- degré 2 : le locuteur réagit vocalement (cris, paroles etc.) et physiquement sans pour autant contrôler la situation.
- degré 3 : le locuteur agit par la parole et par le geste et ceci a un effet sur le stimulus, il contrôle la situation.

L'annotation selon ces trois dimensions présente un niveau plus élevé d'abstraction qui assure une généricité vis-à-vis du corpus. Si la troisième dimension réactivité est spécifique à l'application, les deux premières dimensions intensité et évaluation, sont des dimensions couramment utilisées et transposables à tout type d'émotions [Cowie et Cornelius, 2003], [Pereira., 2000].

Chaque dimension fournit une perspective d'analyse différente permettant à l'annotateur de décomposer ses classes émotionnelles. Ce sont des dimensions indépendantes des catégories : ce n'est pas l'intensité de la peur qui est annotée ici mais l'intensité de l'émotion sachant qu'elle est plus intense en moyenne dans le cas de la peur.

3.3 Stratégie d'annotation du contexte d'émergence des émotions

L'objet de ce paragraphe est d'une part de décrire comment la stratégie d'annotation intègre la description des manifestations émotionnelles dans leur contexte multimodal et temporel,

et d'autre part de présenter la stratégie d'annotation du contexte situationnel, personnel et social.

3.3.1 Description des manifestations émotionnelles dans leur contexte multimodal et temporel

La description locale du contenu émotionnel de chaque segment est réalisée au sein de la description globale de la séquence dans le niveau global du segment. Ce qui signifie que l'annotateur dispose du contexte temporel associé à la séquence pour évaluer les émotions.

De plus, le canal vidéo, couplé avec le canal audio, est utilisé en tant que support à l'annotation du contenu audio de la séquence. C'est ANVIL (Annotation for Video and Language) [Kipp, 2001] en tant qu'outil d'annotation pour le dialogue multimodal, qui a été choisi comme logiciel d'annotation. ANVIL permet à l'annotateur de disposer du support vidéo de la séquence lors de son annotation. L'aspect multimodal des émotions est en effet essentiel dans le processus de communication. La transmission du message oral passe par une corrélation entre parole et geste [McNeill, 2005].

L'annotation d'une séquence du corpus selon la stratégie d'annotation développée sous ANVIL est présentée sur la figure 3.6. Le choix de ce logiciel est également motivé par la perspective d'intégrer notre système de reconnaissance des émotions dans une plateforme multimodale de surveillance. ANVIL offre en effet la possibilité d'annoter les manifestations visuelles de la situation anormale telles que les mouvements de foule, la présence d'objets anormaux, les mouvements du locuteur, ses gestes ou même ses expressions faciales [Abrilian *et al.*, 2005].

3.3.2 Description du contexte situationnel

La situation est décrite dans son déroulement temporel à l'intérieur de la séquence en fonction de son caractère normal ou anormal. Si la situation est considérée comme anormale, i.e. si une menace est présente, son intensité (ou gravité) et son degré d'imminence (latente, potentielle, immédiate, passée) sont annotées comme représenté sur la figure 3.7.

De plus, une catégorisation des types de menace est proposée en répondant aux questions suivantes :

- Si menace : est-elle connue de la ou les victimes ?
- Si la menace est connue des victimes : la cause de la menace est-elle humaine ou naturelle ?
- Si la menace est humaine et connue des victimes : l'agresseur est-il présent dans la séquence ?
- Si la menace est humaine et connue des victimes : la victime a-t-elle reconnu l'agresseur comme un familier ? (i.e. la victime connaissait-elle l'agresseur avant l'agression ?)

Pour résumer, ces questions sont récapitulées dans le schéma 3.8.

En parallèle, pour chaque menace, le type de situation correspondant est précisé (ex : prise d'otage, séquestration, braquage, inondation etc.). Le but est de permettre d'identifier les séquences qui illustrent la même menace dans un même film d'une part et de fournir une typologie des types de menaces d'autre part.

L'unité d'annotation correspondant à l'annotation du champ menace correspond à un regroupement de segments à l'intérieur des séquences en fonction de l'évolution de la situation.

3.3.3 Description du contexte personnel et social

Le sexe du locuteur et sa position dans l'interaction (victime, agresseur, ou autres) sont également pris en compte par le schéma d'annotation en fonction des annotations fournissant des

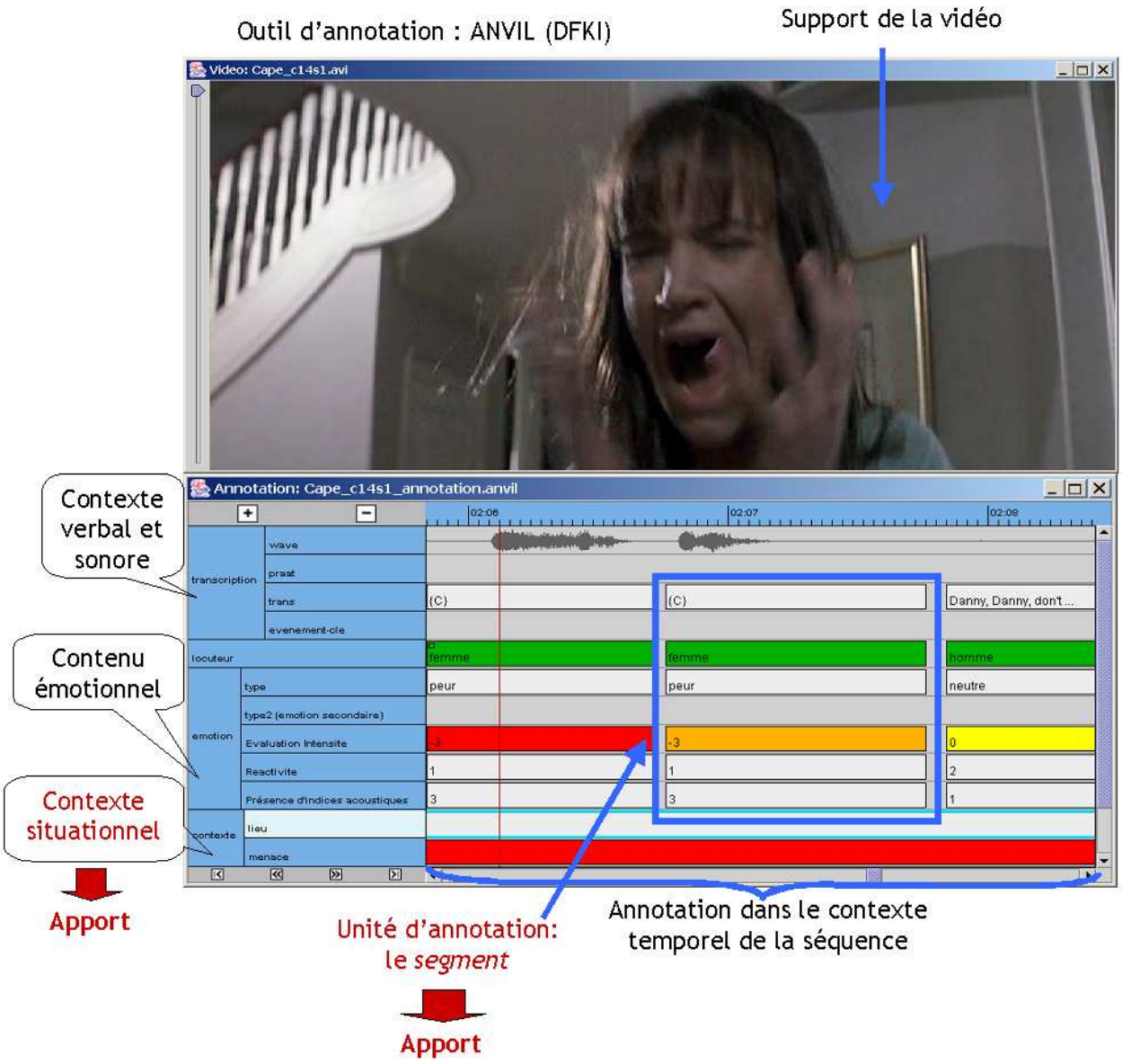

FIG. 3.6 – *Schéma d'annotation sous ANVIL*

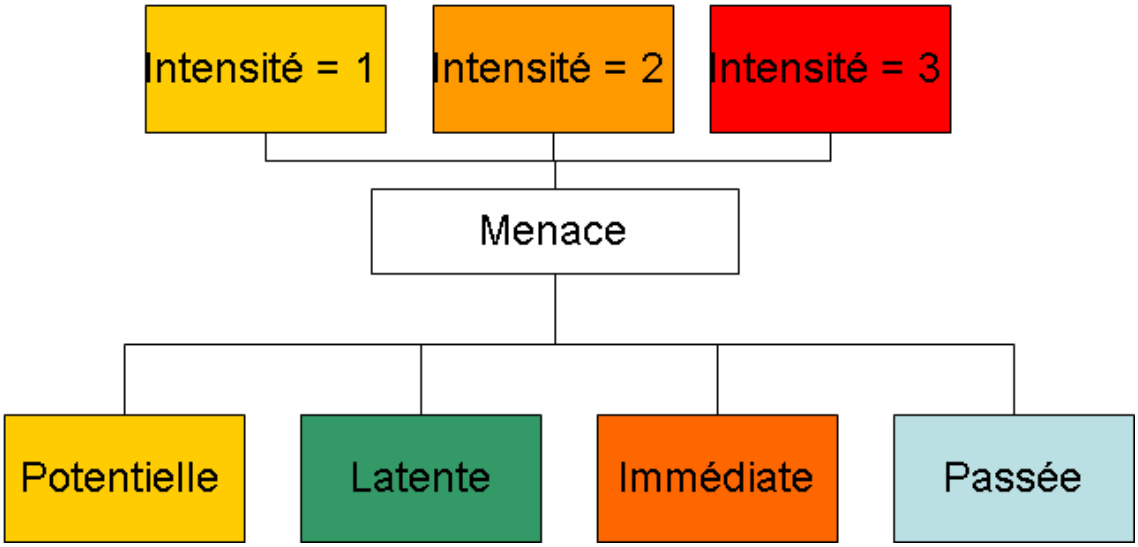

FIG. 3.7 – *Description du degré d'imminence et de la gravité de la menace*

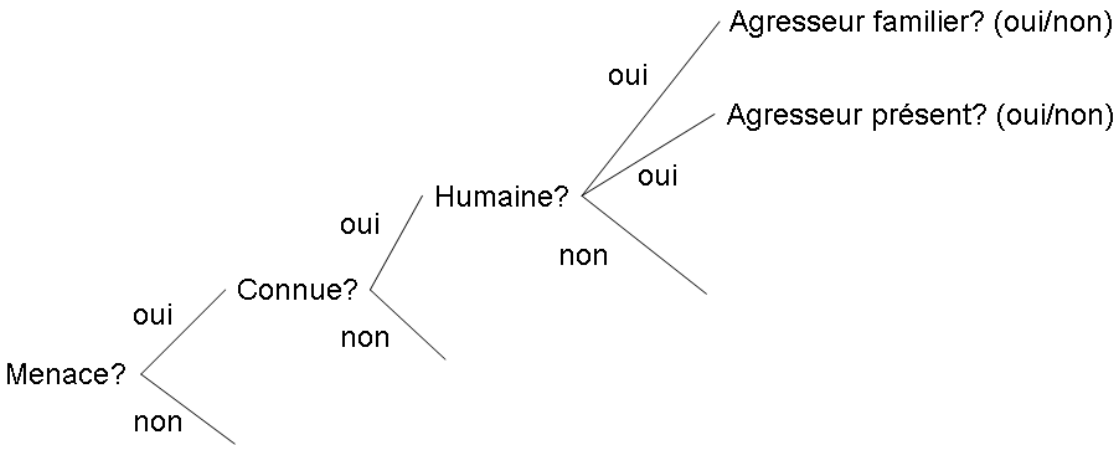

FIG. 3.8 – *Catégorisation des différents types de menace*

informations sur le profil des personnes présentes lors de la situation. La présence de superposition vocale de plusieurs locuteurs (*overlap*) est également signalée.

Le niveau dialogique est intégré de la manière suivante : dans chaque séquence, un nom est attribué à chaque locuteur (locuteur 1, 2, 3, ...), et pour chaque segment, est indiqué soit le locuteur correspondant, soit s'il s'agit d'un groupe ou d'une foule. Cette annotation permet ainsi de déterminer le type de l'interaction (monologue, dialogue, groupe, foule) et de décrire les interactions entre les personnages à l'intérieur de la scène. De plus les locuteurs par film seront annotés de manière à pouvoir être identifiés à travers les séquences, ce qui offre la possiblité de construire des profils de locuteurs.

3.3.4 Description du contexte verbal et sonore

Le contenu verbal est transcrit à l'aide des sous-titres fournit par les DVD selon des règles de transcription inspirées de la norme LDC (Linguistic Data Consortium[9]) et décrite dans l'annexe B. Le contenu vocal, non verbal, tel que les respirations, les cris, est également transcrit ainsi que le contenu non vocal en rapport avec les situations anormales tels que les coups de feu, ou les explosions.

Des meta-descripteurs permettent également de stocker :

- une évaluation de la qualité de la parole dans les segments par une note allant de Q0 (parole peu intelligible due à une mauvaise qualité d'enregistrement ou un environnement sonore trop bruité) à Q3 (parole parfaitement intelligible) ;
- une description de l'environnement sonore intervenant en même temps que la parole par les labels suivants : « propre » (pas de bruit de fond ni de musique), « bruit seul », « musique seule », « bruit et musique ».

3.4 Conclusion

Deux approches sont couramment utilisées dans la littérature pour décrire les émotions : *l'approche dimensionnelle* et *l'approche catégorielle*.

La stratégie d'annotation adoptée repose sur ces deux approches pour proposer une *description à deux niveaux (contextuel et émotionnel)* qui soit adaptée à l'application de surveillance :

→ La stratégie d'annotation adoptée intègre la description du contenu audio de la séquence dans son contexte multimodal : le canal vidéo, couplé avec le canal audio, est utilisé en tant que support à l'annotation du contenu audio de la séquence, la vidéo fournissant ainsi le support contextuel nécessaire à l'interprétation des émotions.

→ La stratégie d'annotation intègre la description des émotions dans leur contexte d'émergence. Plusieurs facteurs permettent en effet de décrire le contexte d'émergence de l'émotion au niveau de la séquence. Ces facteurs concernent à la fois la situation (intensité, imminence, nature de la menace) et la relation entre les locuteurs (victime ou agresseur). Chaque séquence fournissant un contexte particulier est ensuite segmentée en une unité d'annotation de base que l'on appelle *segment*. Cette segmentation permet de capturer l'évolution des manifestations émotionnelles en fonction du déroulement de la situation.

→ La description de l'émotion au niveau du segment se fait par des catégories globales qui ont été définies en isolant l'émotion ciblée par l'application : la peur. Les descripteurs dimensionnels intègrent en plus des dimensions classiques (intensité et évaluation), une dimension adaptée à l'application : *la réactivité*.

[9] http ://www.ldc.upenn.edu/

Chapitre 4

Le corpus SAFE : fiabilité de la stratégie d'annotation et contenu

Sommaire

Introduction

Nous avons vu dans le chapitre précédent que l'annotation des corpus en émotion est une tâche complexe, sujette à la subjectivité de la catégorisation humaine. Dans le cas du corpus exploité ici, nous partons de l'hypothèse que cette subjectivité est accrue par la diversité du matériel émotionnel utilisé, issu de films différents. Le schéma d'annotation adopté propose donc des catégories émotionnelles destinées à permettre à l'annotateur d'organiser les différentes manifestations émotionnelles présentes dans le corpus. Par la suite, il s'agit de s'assurer que des annotations fournies par des annotateurs différents offrent une convergence suffisante pour être exploitables par un système de classification automatique d'émotion, i.e. d'évaluer la fiabilité des annotations.

Cette évaluation nécessite soit de disposer d'une annotation du corpus par un large panel d'annotateurs [Cowie et Cornelius, 2003] soit de réaliser des tests perceptifs ou des validations croisées sur un sous-ensemble du corpus [Devillers et Vasilescu, 2003]. Pour notre étude, la quantité de données à annoter s'élève à 7 heures d'enregistrement. Compte tenu du coût de l'annotation d'un tel corpus, la stratégie adoptée a été de faire appel à un nombre réduit d'annotateurs (trois annotateurs) et de proposer en parallèle une validation perceptive du schéma d'annotation sur un sous-ensemble du corpus par un échantillon représentatif de personnes. Les résultats de ces tests perceptifs sont présentés dans le premier paragraphe de ce chapitre. Afin d'évaluer la fiabilité, nous avons par ailleurs examiné la convergence des trois annotations dans le deuxième paragraphe.

Le schéma d'annotation propose des descripteurs choisis en fonction de l'objectif final de cette étude qui est la détection d'émotions de type peur pour la détection de situations anormales. Ces descripteurs vont permettre de faire émerger les spécificités du corpus, et d'évaluer l'adéquation entre les émotions illustrées et l'application visée. Dans cet objectif, le paragraphe 3 de ce chapitre est dédié à la description du corpus. Le paragraphe 4 propose un bilan du corpus et de ses annotations sous l'angle de leur exploitation par le système de reconnaissance des émotions de type peur.

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2004], [Clavel *et al.*, 2006b], [Clavel *et al.*, 2006a]

4.1 Validation du schéma par des tests perceptifs

Une première étape de validation de la stratégie d'annotation est effectuée au moyen de tests perceptifs [Clavel *et al.*, 2004]. Ceux-ci sont réalisés sur un lot de 20 séquences tests extraites de 6 films différents choisis de manière à illustrer un échantillon représentatif de la base de données.

Les objectifs de ces tests perceptifs sont :

- évaluer la présence de l'émotion cible, i.e. d'émotions intenses de type peur dans les données de fiction sélectionnées,
- évaluer le rôle de la modalité vidéo dans l'annotation, en comparant la catégorisation d'une émotion en fonction de la présence ou non de ce support,
- évaluer la pertinence du segment[10] comme unité d'annotation en fournissant aux sujets le segment isolé de la séquence pour l'annotation du contenu émotionnel. Il s'agit

[10] le segment est l'unité d'annotation définie dans le chapitre précédent comme un tour de locuteur ou une portion de tour de locuteur avec un contenu émotionnel homogène

de s'assurer que le segment contient suffisamment d'information pour caractériser une émotion.

4.1.1 Protocole de test

Base de test

La base de test utilisée est constituée de 40 segments sélectionnés parmi les 20 séquences. Cette sélection est effectuée de telle sorte que chaque classe (peur, autres émotions négatives, neutre, émotions positives) soit illustrée par 10 stimuli répartis en 5 stimuli avec des locuteurs masculins et 5 stimuli avec des locuteurs féminins. Les segments sont extraits de séquences illustrant soit des situations anormales (pour 22 segments), soit des situations normales (pour 18 segments). Les situations considérées comme anormales sont des situations de menace pour la vie humaine.

Conditions expérimentales

Les stimuli sont présentés dans un ordre aléatoire aux sujets. Deux conditions expérimentales sont considérées :

- +vidéo : les segments sont présentés aux sujets avec le support audio et vidéo
- -vidéo : les segments sont présentés aux sujets avec uniquement le support audio

Une phase de familiarisation constituée de 5 stimuli précède le test, afin d'entraîner les sujets et de leur permettre de mettre en place leur échelle de référence. Les réponses à ces 5 stimuli ne seront pas exploitées comme résultats du test.

Sujets

22 sujets ont participé aux tests perceptifs, 11 pour la condition +vidéo et 11 pour la condition -vidéo. L'ensemble des sujets est de langue maternelle française avec une maîtrise de l'anglais suffisante pour les tests. Par ailleurs la transcription des stimuli leur est fournie comme support, afin de pallier les inégalités de compréhension du contenu linguistique. L'influence du facteur « maîtrise de la langue anglaise » sur la perception n'a pas été étudiée ici mais serait cependant un sujet d'étude intéressant à approfondir.

Consignes

Chaque sujet doit effectuer les tâches suivantes pour chaque stimulus : nommer l'émotion perçue et l'évaluer selon les 3 axes intensité/évaluation/réactivité, définies dans le chapitre précédent. La dénomination de l'émotion par les sujets est réalisée sans restriction particulière. Les sujets peuvent écouter/voir les stimuli autant de fois qu'ils le désirent.

4.1.2 Résultats

Dénomination du contenu émotionnel

Les termes utilisés par les sujets pour nommer le contenu émotionnel des segments sont très variés. Il en ressort les trois points suivants :

- Les six émotions de base, les big-six (ex : « peur », « colère », « tristesse », « joie », « surprise », « dégoût ») ont spontanément été utilisées par les sujets ;

– Ces six émotions et les termes utilisés par les sujets s'y référant peuvent être regroupés dans les quatre catégories globales définies dans la stratégie d'annotation (peur, autres émotions négatives, neutre, émotions positives). Un grand nombre de termes correspond à des nuances des émotions cible, i.e les émotions de type peur (ex : « angoisse », « anxiété », « terreur », « panique », etc.) ;
– Les sujets ont annoté des segments par des termes référant à de la peur mêlée à d'autres émotions (ex : « peur-colère », « peur-tristesse »), ce qui confirme la grande diversité des illustrations de peur présentes dans le corpus.

Par ailleurs, certains termes font plutôt référence à des comportements du locuteur (ex : « commandement », « dangereux », « directif », « contrôle de soi »), à des attitudes (ex : « alerte », « amicale »). Les sujets utilisent majoritairement des substantifs (ex : « amusement », « confusion »), des adjectifs (ex : « amusée », « confus ») ou même des verbes (ex : « se braquer »).

La notion d'émotion semble difficile à circonscrire par un sujet naïf. Elle est associée, au delà de la classe émotionnelle même, à des attitudes, comportements ou descriptifs de la situation. On retrouve donc ici le problème de la catégorisation et des différents aspects pris en compte pour définir l'état affectif présenté dans le chapitre 1.

Évaluation de l'émotion selon les trois dimensions.

Les figures ci-dessous présentent le pourcentage d'attributions des différents niveaux pour chaque dimension, intensité (figure 4.1), évaluation (figure 4.2), et réactivité (figure 4.3) en fonction des conditions de test (+/- vidéo) et des contextes d'émergences du stimulus (situations anormales[11] vs. situations normales).

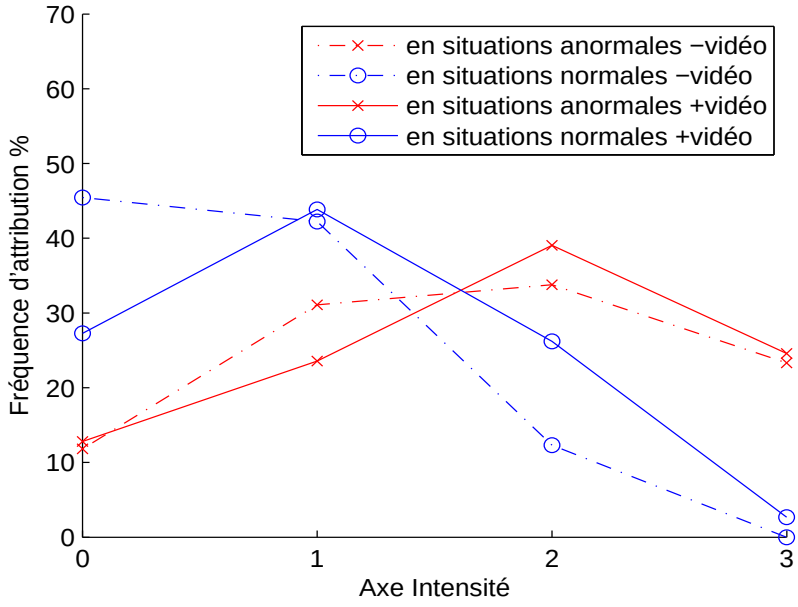

FIG. 4.1 – *Pourcentage d'attribution des degrés d'intensité en situations anormales (×) vs. en situations normales (o) sur les 40 segments (22 segments en situations anormales et 18 segments en situations normales) et pour les deux conditions de tests (11 sujets pour + vidéo et 11 sujets pour - vidéo). Les intervalles de confiance à 95% sont de rayon ≤ 6%*

[11] telles que définies dans le chapitre introductif et dans le glossaire

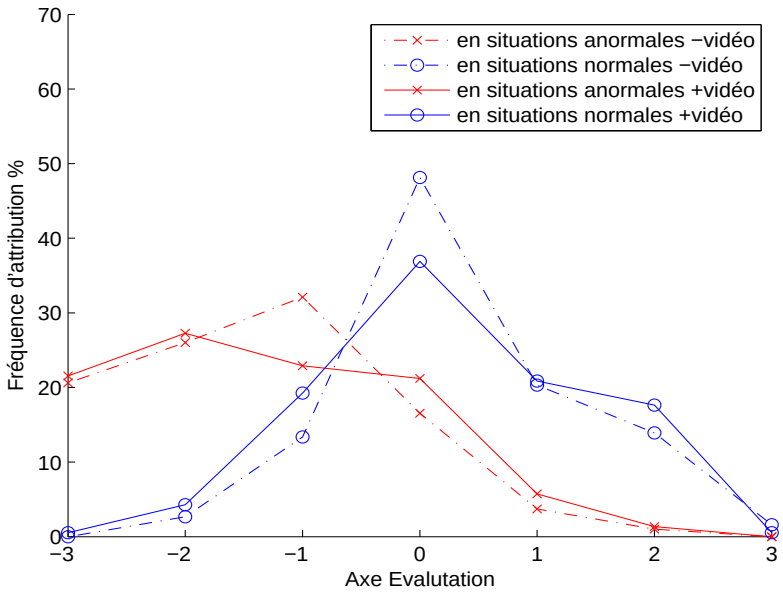

FIG. 4.2 – *Pourcentage d'attribution des niveaux sur l'axe évaluation en situations anormales (×) vs. en situations normales (o) sur les 40 segments (22 segments en situations anormales et 18 segments en situations normales) et pour les deux conditions de tests (11 sujets pour + vidéo et 11 sujets pour - vidéo). Les intervalles de confiance à 95% sont de rayon ≤7%.*

Le pourcentage d'attribution du degré deg_i de chaque dimension s'exprime de la manière suivante :

$$\%attr_{deg_i} = \frac{\sum_{k=1}^{N_{suj}} n_{k,i} * 100}{N_{suj} * N_{seg}}$$

où $n_{k,i}$ est le nombre de segments annotés de degré i par le k^e sujet, N_{suj} le nombre total de sujets et N_{seg} le nombre de segments.

Pour chaque pourcentage, un intervalle de confiance à 95% peut être calculé [Jolion, 2006]. Il correspond à l'intervalle dans lequel la valeur du pourcentage est attendue avec une probabilité de 95%. Les intervalles de confiance permettent d'obtenir les valeurs pour lesquelles le résultat est plausible. La formule utilisée pour leur calcul intègre la valeur calculée du pourcentage, le nombre de sujets des tests perceptifs et le nombre de segments testé. Le rayon de l'intervalle de confiance s'exprime de la manière suivante :

$$r = 1,96\sqrt{\frac{\%attr_{deg_i}(100 - \%attr_{deg_i})}{N_{suj} * N_{seg}}}$$

L'intervalle de confiance du pourcentage est alors : $I = [\%attr_{deg_i} - r; \%attr_{deg_i} + r]$

Les stimuli correspondant à des situations anormales sont évalués comme plus intenses que ceux correspondant à des situations normales (cf. figure 4.1). Ce résultat est dû aux types de films sélectionnés pour la constitution du corpus qui illustrent peu d'émotions intenses en situations normales, comme par exemple la joie.

Concernant l'axe évaluation (figure 4.2) les émotions en situations normales sont majoritairement perçues avec un niveau 0, i.e. ni négatives ni positives, ou faiblement négatives (niveau -1) ou positives (niveau 1). En situations anormales, elles sont évaluées sans surprise majoritairement comme négatives.

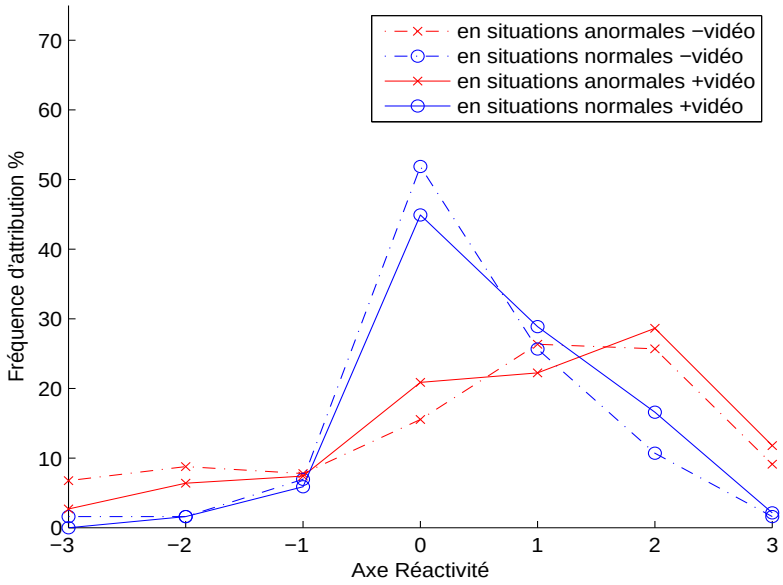

FIG. 4.3 – *Pourcentage d'attribution des niveaux de réactivité en situations anormales (×) vs. en situations normales (o) sur les 40 segments (22 segments en situations anormales et 18 segments en situations normales) et pour les deux conditions de tests (11 sujets pour + vidéo et 11 sujets pour - vidéo). Les intervalles de confiance à 95% sont de rayon ≤7%.*

Pour l'axe réactivité[12], le même phénomène que pour l'axe évaluation est observé pour les émotions en situations normales (cf. figure 4.3). En revanche la réactivité pour les émotions en situations anormales est majoritairement évaluée comme active.

$\Longrightarrow$ *Les émotions émergeant en situations anormales sont perçues comme plus intenses, plus négatives qu'en situations normales. Le comportement du locuteur face à la menace est majoritairement considéré comme actif sur l'axe actif/passif.*

Les segments en situations normales ont été annotés neutre (intensité=0) dans une proportion signicativement plus grande par les sujets dans la condition -vidéo (pourcentage d'attributions de 45%± 6% (intervalle de confiance à 95%)) que par les sujets dans la condition +vidéo (pourcentage d'attributions de 28%± 6%). Les segments en situations normales sont annotés par un degré d'intensité plus élevé dans la condition + vidéo que dans la condition -vidéo. En effet, le pourcentage d'attribution du degré 2 est de 11%± 4% dans la condition -vidéo et de 28%± 6% dans la condition +vidéo (différence significative).

$\Longrightarrow$ *La présence d'émotion semble être difficile à évaluer avec le support audio seul en situations normales et les émotions en situations normales sont perçues comme plus intenses avec le support vidéo.*

[12]Lors de la première annotation et des tests perceptifs l'échelle utilisée pour l'axe réactivité se divisait en deux sur l'axe actif/passif avec un champ *none* (vide) pour les situations normales. Cette échelle a été réajustée à l'issue des tests perceptifs (voir ajustements)

En situations anormales, le pourcentage d'attributions du degré 0 de la dimension intensité (neutre) par les sujets est de 11%± 4% dans les deux conditions. La présence d'émotion en situations anormales semble plus facile à évaluer avec le support audio seul. De manière générale, pour les trois dimensions, les courbes associées à chacune des deux conditions +/- vidéo présentent une plus grande proximité pour les segments en situations anormales que pour les émotions en situations normales.

$\Longrightarrow$ *En situations anormales, l'audio semble être un canal privilégié des manifestations émotionnelles*

4.1.3 Conclusion – Validation des objectifs

Ces tests perceptifs consistent en l'annotation par 22 sujets de 40 segments sélectionnés pour constituer un échantillon représentatif du corpus. Ces tests perceptifs nous ont permis de valider sur ces 40 segments les points suivants :

- La présence de l'émotion cible : par l'utilisation de termes référant à des émotions de type peur lors de l'étape de dénomination, d'une part et par la forte proportion de segments annotés d'intensité 2 ou 3 en situations anormales, d'autre part.
- Le rôle de la modalité vidéo pour l'annotation : les émotions sont perçues comme plus intenses avec l'aide du support vidéo et pour les situations normales, la vidéo représente un réel support à l'annotation. Cependant pour les situations anormales les deux courbes (+/- vidéo) sont proches. Cette proximité met ainsi en évidence le poids de l'audio dans l'évaluation des émotions ciblées par l'application selon les 3 dimensions.
- La pertinence du segment comme unité d'annotation : la proximité des courbes obtenues pour les deux groupes de sujets (+/- vidéo) confirme également que le support contextuel fourni par le segment est porteur d'une quantité d'informations suffisante pour une évaluation consistante des émotions selon les trois dimensions.

4.1.4 Conclusion – Ajustements

Ajustement de l'axe réactivité

A l'issue de ces tests perceptifs, compte tenu de la faible proportion d'attribution du côté « passif » de l'axe, l'échelle utilisée initialement pour la réactivité a été remplacée pour les deuxième et troisième annotations par une mesure du degré de réactivité du locuteur face à la situation avec une échelle allant de 0 à 3.

Sous-catégories émotionnelles et doubles étiquettes

L'utilisation par les sujets de termes référant aux différentes nuances de peur ou à des émotions mêlées, nous a ammenés, pour le schéma d'annotation final, à proposer des sous-catégories émotionnelles en plus des catégories globales et à laisser à l'annotateur la possibilité de choisir deux étiquettes[13].

La liste des sous-catégories émotionnelles proposée par la stratégie d'annotation finale a donc été établie en se basant sur :

- la liste d'émotions proposées par le réseau d'excellence HUMAINE[14] ;
- la liste des termes les plus utilisés par les sujets des tests perceptifs ;

[13] voir chapitre précédent

[14] http ://emotion-research.net/ws/summerschool1/emotion%20words/

inquiétude	état pénible déterminé par l'attente d'un évènement, d'une souffrance que l'on craint, par l'incertitude où on est
anxiété	état d'angoisse considéré surtout dans son aspect psychique
stress	tension nerveuse, contrainte de l'organisme face à un choc (événement soudain, traumatisme, sensation forte, bruit, surmenage)
crainte	sentiment d'inquiétude déterminé par l'idée d'un danger existant ou potentiel
angoisse	malaise psychique et physique né du sentiment de l'imminence d'un danger
terreur	peur extrême qui paralyse
horreur	impression violente causée par la vue et ou la pensée d'une chose qui fait peur ou qui répugne
frayeur	peur très vive, généralement passagère et peu justifiée
effroi	grande frayeur souvent mêlée d'horreur
panique	peur qui trouble subitement et violemment l'esprit
détresse	peur accompagnée d'un sentiment d'abandon et d'impuissance

TAB. 4.1 – *Les nuances de peur et leur définition*

Termes émotionnels	(1)	(2)
colère	oui (anger)	oui
tristesse	oui (sadness)	oui
dégoût	oui(disgust)	oui
souffrance	non	oui
déception	oui(disappointment)	oui
mépris	oui(contempt)	non
honte	oui (shame)	non
désespoir	oui (despair)	oui
cruauté	oui (cruelty)	non
pitié	non	oui
joie	oui (joy)	oui
soulagement	oui (relief)	oui
détermination	oui (determination)	oui
fierté	oui (satisfaction)	oui
espoir	oui (hopeful)	oui
reconnaissance	oui (greed)	oui
surprise	oui (surprise)	oui

TAB. 4.2 – *Termes émotionnels sélectionnés pour la liste des sous-catégories*

– la liste des nuances de la peur présentée dans le tableau 4.1 qui a été établie à l'aide du dictionnaire le Robert.

Nous avons sélectionné parmi les deux premières listes les termes référant à des manifestations que nous avons jugées pertinentes compte tenu de l'application, c'est à dire susceptibles de survenir en situations anormales. Le tableau 4.2 indique pour chacun des termes sélectionnés, si celui-ci est présent dans la liste proposée par HUMAINE (colonne 1), ou s'il a été utilisé par les sujets des tests perceptifs (colonne 2).

4.2 Validation du schéma par la confrontation des annotations

L'annotation du corpus SAFE a été réalisée par trois annotateurs. L'annotateur 1 (Ann1) est bilingue anglais/français de langue maternelle anglaise. Sa tâche consiste en le découpage de la séquence en segments et en la description du contenu de la séquence selon le schéma d'annotation décrit dans le chapitre précédent. Il est notamment chargé d'effectuer la transcription du contenu verbal et non verbal[15] de chaque segment et de fournir une description de son contenu émotionnel. Deux autres annotations du contenu émotionnel des séquences ont été réalisées par deux annotateurs français (Ann 2 et Ann3) parlant anglais couramment. Ils se sont basés, pour cette annotation, sur les segments prédéfinis par la segmentation de la séquence fournie par Ann1. Les profils des trois annotateurs sont détaillés à titre indicatif dans le tableau 4.3.

	sexe	âge	profession
Ann1	masculin	35 ans	traducteur
Ann2	masculin	26 ans	ingénieur du son
Ann3	masculin	30 ans	chercheur en acoustique

TAB. 4.3 – *Profil des trois annotateurs*

Compte tenu du coût de l'annotation, nous n'avons pas proposé de protocole de validation de l'étape de segmentation réalisée par l'annotateur 1. Il serait également intéressant à terme de pouvoir disposer d'un plus grand nombre d'annotateurs afin de renforcer la fiabilité des annotations.

En résumé, chacun des trois annotateurs a sélectionné pour chaque segment :

– une catégorie globale parmi les quatre catégories proposées (peur, autres émotions négatives, neutre, émotions positives),
– un niveau d'intensité parmi les quatre niveaux possibles 0, 1, 2 et 3,
– un niveau sur l'axe évaluation parmi les sept niveaux possibles -3, -2, -1, 0, 1, 2 et 3,
– un niveau sur l'axe réactivité[16] parmi les quatre niveaux possibles 0, 1, 2 et 3.

Dans cette partie, nous nous proposons de comparer les annotations obtenues à l'issue de chacune des trois annotations. L'objectif est d'une part d'évaluer le degré de fiabilité du schéma d'annotation et d'autre part d'évaluer la possiblité d'extraire de ces différentes annotations une information convergeante pour le système de classification automatique développé par la suite. A cet effet nous proposons une évaluation du degré d'accord entre les différents annotateurs et une analyse des recouvrements entre les différentes catégories émotionnelles et entre les différents niveaux des descripteurs dimensionnels.

[15] le contenu non verbal correspond ici au non verbal acoustique (cris, respiration)

[16] L'annotateur 1 a utilisé pour l'axe réactivité l'échelle actif/passif utilisée également par les sujets des tests perceptifs et ajustée à cette occasion. Seuls Ann2 et Ann3 ont donc utilisé l'échelle définitive sur les 4 niveaux 0, 1, 2 et 3.

4.2.1 Comment mesurer un degré de fiabilité ?

Le degré d'accord entre les différents annotateurs peut être mesuré par le taux d'accord τ qui s'exprime par la formule suivante :

$$\tau = \frac{\sum_{k=1}^{K} n_k}{N_{seg}}$$

où N_{seg} correspond au nombre total de segments et n_k au nombre de segments annotés comme appartenant à la catégorie k par la totalité des annotateurs. Cependant le taux d'accord ne prend en compte ni le nombre de jugements possibles ni le nombre d'annotateurs. Il ne permet pas non plus d'intégrer une éventuelle proximité entre les jugements différents.

Il existe différents mesures permettant de pallier ces éventuels biais. Nous avons choisi ici d'utiliser pour mesurer le degré d'accord entre nos trois annotateurs, le coefficient kappa [Carletta, 1996] [Bakeman et Gottman, 1997] pour les catégories globales et le coefficient alpha de Cronbach [Cronbach, 1951] pour les trois dimensions. Le coefficient kappa, noté ici κ correspond au taux d'accord corrigé de ce qu'il serait sous le simple effet du hasard :

$$\kappa = \frac{\bar{p_o} - \bar{p_e}}{1 - \bar{p_e}}$$

où $\bar{p_o}$ est la proportion d'accord observée et $\bar{p_e}$ est la proportion d'accord aléatoire, c'est à dire la probabilité que les annotateurs s'accordent par chance. Ces deux proportions se calculent de la manière suivante :

$$\bar{p_0} = \frac{1}{N_{seg}} \sum_{i=1}^{N_{seg}} p_{seg_i}$$

avec

$$p_{seg_i} = \frac{1}{N_{ann}(N_{ann} - 1)} \sum_{k=1}^{K} n_{ik}(n_{ik} - 1)$$

et,

$$\bar{p_e} = \sum_{k=1}^{K} p_{cl_k}^2$$

avec

$$p_{cl_k} = \frac{1}{N_{seg} N_{ann}} \sum_{i=1}^{N_{seg}} n_{ik}$$

où N_{ann} est le nombre d'annotateurs et n_{ik} le nombre d'annotateurs ayant annoté le i^e segment par la catégorie k.

p_{cl_k} correspond à la proportion globale des segments attribués à la classe k et la proportion p_{seg_i} correspond à la mesure de concordance entre les N_{ann} annotateurs pour chaque segment i.

Dans le cas d'un accord parfait entre les annotateurs le kappa atteint sa valeur maximale qui est de 1. Le kappa est égal à 0 lorsque l'accord est identique à celui obtenu par chance. Un kappa négatif correspond à un accord plus mauvais que celui obtenu par chance. Un classement de l'accord en fonction de la valeur de Kappa est proposé dans [Landis et Koch, 1977] et récapitulé dans le tableau 4.4. Le kappa est une mesure qui a été utilisée par de nombreuses études sur l'annotation des corpus en émotion [Cowie et Cornelius, 2003] [Craggs, 2004] [Shafran *et al.*, 2003] [Devillers *et al.*, 2005].

Accord	Kappa
Excellent	$\geq 0,81$
Bon	0,80-0,61
Modéré	0,60-0,41
Médiocre	0,40-0,21
Mauvais	0,20-0
Très mauvais	< 0

TAB. 4.4 – *Degré d'accord en fonction des valeurs de Kappa*

Pour les échelles numériques utilisées pour les trois dimensions (intensité, évaluation et réactivité), le coefficient alpha de Cronbach permet d'évaluer la fiabilité ou cohérence interne de l'échelle utilisée. La cohérence interne correspond à l'homogénéité existante entre les annotations des différents annotateurs, i.e. une échelle a une cohérence interne tant que les différents annotateurs fournissent des annotations suffisamment corrélées entre elles. Le coefficient alpha de Cronbah, noté ici α, se calcule de la manière suivante :

$$\alpha = \frac{N_{ann}.\bar{r}}{1 + (N_{ann} - 1).\bar{r}}$$

où $\bar{r}$ est l'inter corrélation moyenne entre les annotateurs.

4.2.2 Catégories : de la difficulté d'une catégorisation neutre/émotion

La figure 4.4 représente la manière dont les segments du corpus ont été répartis selon les quatre catégories émotionnelles par chacun des trois annotateurs.

Si les distributions des catégories émotionnelles à travers le corpus sont assez similaires pour le premier annotateur (Ann1) et le troisième annotateur (Ann3), le second annotateur semble avoir adopté une stratégie d'annotation différente en termes de catégories : le second annotateur (Ann2) a tendance à attribuer une plus forte proportion de segments (42%) à la classe neutre que les deux autres annotateurs (Ann1 : 24% et Ann3 : 22%).

Le calcul des kappa (cf tableau 4.5) pour évaluer le degré d'accord entre chaque paire d'annotateurs confirme cette différence de stratégie de la part du second annotateur. La paire d'annotateur Ann1/Ann3 obtient en effet un kappa supérieur à ceux obtenus par les deux autres paires d'annotateurs.

	Kappa
Ann1 vs. Ann2	0,47
Ann1 vs. Ann3	0,54
Ann2 vs. Ann3	0,48
Ann1 vs. Ann2 vs. Ann3	0,49

TAB. 4.5 – *Calcul du kappa entre les 4 catégories émotionnelles pour chaque paire d'annotateurs et entre les trois annotateurs sur l'ensemble des 4275 segments du corpus SAFE*

Le kappa obtenu pour les trois annotateurs 0,49 correspond à un accord modéré d'après le tableau 4.4 donné dans l'article [Landis et Koch, 1977]. Cette valeur est acceptable étant donné que l'annotation des émotions est une tâche subjective pour laquelle il est difficile d'obtenir des degrés d'accord élevés. Il n'est cependant pas évident de pouvoir comparer les kappas obtenus pour des tâches d'annotation différentes ou sur des données différentes. Notons également que

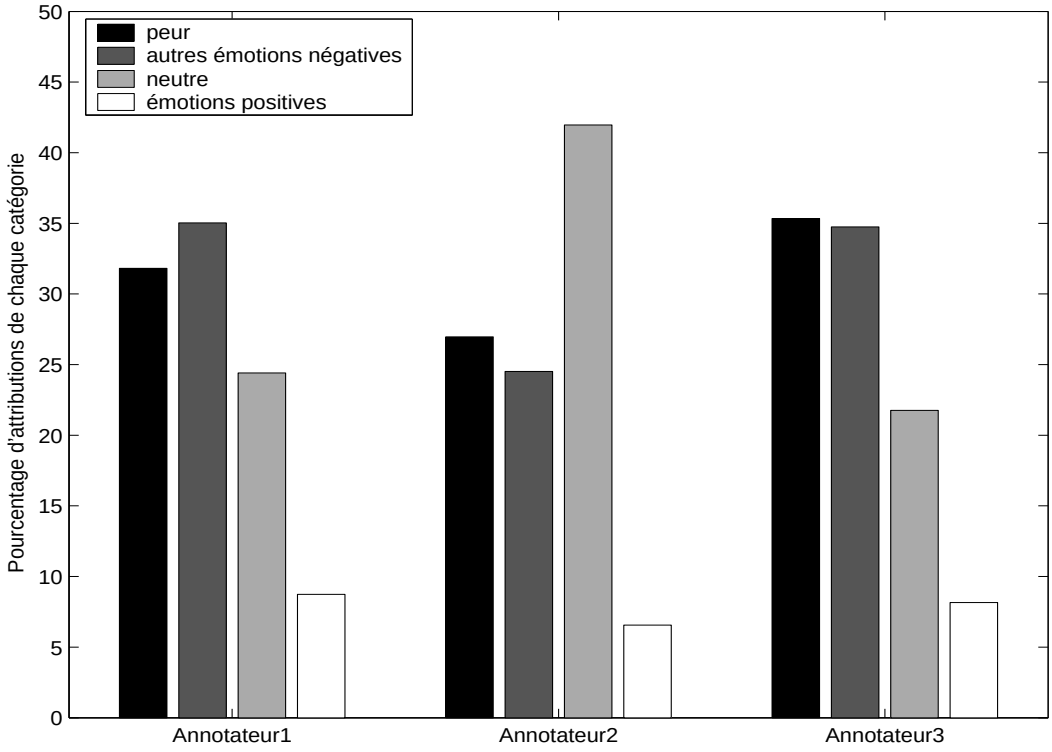

FIG. 4.4 – *Pourcentage d'attributions des différentes catégories émotionnelles en fonction des trois annotateurs sur l'ensemble des 4275 segments du corpus SAFE*

les kappas calculés entre 2 ou 3 annotateurs sont dépendants de la paire ou du triplet d'annotateurs considérés. L'idéal serait évidemment de disposer d'un kappa calculé sur un grand nombre d'annotateurs.

Citons à titre d'exemple quelques valeurs de kappa obtenues sur différentes bases de données. Le kappa obtenu dans [Douglas-Cowie *et al.*, 2003] sur la base de données de Belfast annotée par trois annotateurs oscille entre 0,42 et 0,6 en fonction de la paire d'annotateurs considérée pour la catégorisation en 4 catégories globales, qui consistent en le regroupement de catégories plus fines. Les kappas obtenus sur un corpus de dialogue homme-machine annoté par deux annotateurs dans [Shafran *et al.*, 2003] vont de 0,32 à 0,42 en fonction des catégories considérées. Dans [Vidrascu et Devillers, 2006], les kappas sont évalués sur un corpus de dialogue homme-homme (20 heures, 2 annotateurs) et le kappa obtenu est de 0,57 pour les appelants et de 0,35 pour les agents.

La figure 4.5 va nous permettre d'identifier les confusions entre Ann2 et Ann1, la paire d'annotateurs qui obtient le kappa le plus faible. A chaque catégorie émotionnelle est associé un histogramme (histogramme de confusion). L'histogramme associé à la catégorie peur représente la distribution des segments classés peur par Ann2 en fonction des classes assignées par Ann1.

La confusion majeure se situe au niveau de la catégorisation du neutre. 53% des segments annotés neutre par Ann2, ont été annotés comme porteurs d'une émotion (peur, autres émotions négatives, ou émotions positives) par Ann1. La seconde cause de désaccords entre les deux annotateurs est liée à une confusion entre peur et autres émotions négatives. 22% des segments annotés autres émotions négatives par Ann2 ont été annotés peur par Ann1 et 16% des segments annotés peur ont été annotés autres émotions négatives.

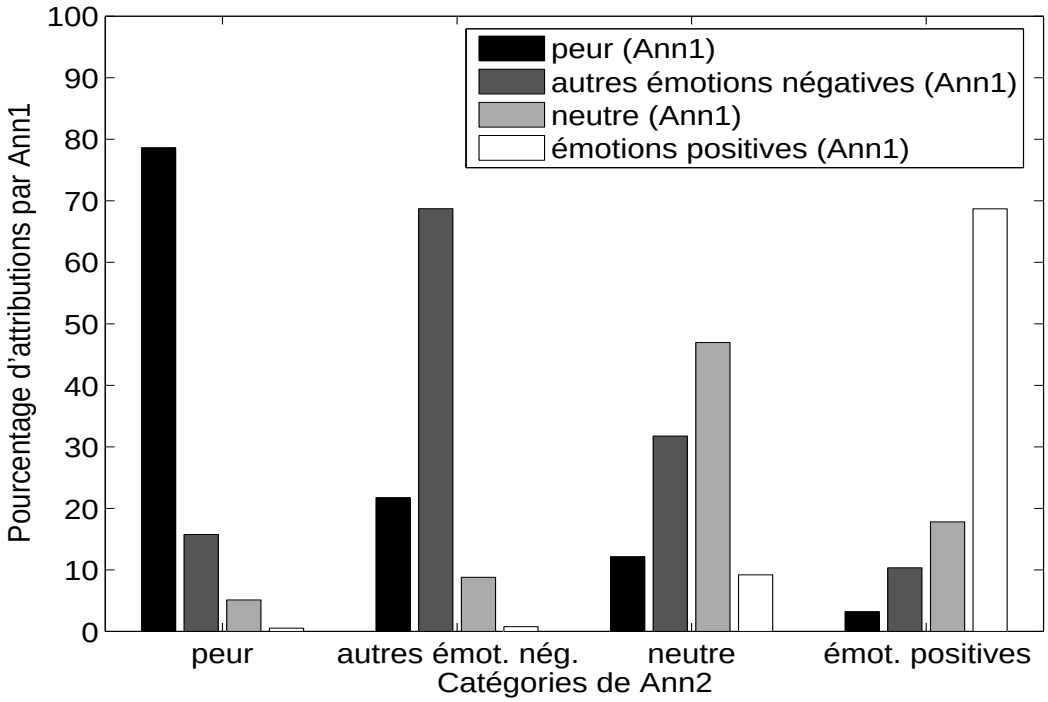

FIG. 4.5 – *Histogramme de confusion entre les deux premiers annotateurs pour les quatre catégories émotionnelles sur l'ensemble des 4275 segments du corpus SAFE*

4.2.3 Dimensions : de la difficulté d'établir un référentiel commun

Intensité

La figure 4.6 représente la distribution dans le corpus des différents niveaux de l'axe intensité en fonction de l'annotateur. Les problèmes constatés dans le paragraphe précédent concernant la catégorisation émotion/neutre se retrouvent sur cette figure, la catégorie neutre correspondant à un niveau 0 sur l'échelle d'intensité. Le nombre de segments évalués d'intensité 0 par Ann2 est nettement plus élevé que pour les deux autres annotateurs. Si Ann1 et Ann3, comme nous l'avons vu précédemment fournissent la même proportion de segments d'intensité 0, la distribution des autres niveaux d'intensité est assez différente pour ces deux annotateurs. Ann1 attribue notamment beaucoup plus fréquemment le degré 3 (35% des segments ont été annotés d'intensité 3 par Ann1 contre 9% par Ann2 et 14% par Ann3).

Pour chaque paire d'annotateurs, nous avons calculé le coefficient kappa et le coefficient alpha de Cronbach (cf tableau 4.6), afin d'évaluer le degré d'accord entre les différents annotateurs sur l'axe intensité. Le premier considère de manière égale les désaccords entre les quatre niveaux d'intensité, le second prend en compte les proximités entre les différents niveaux d'intensité, en pénalisant par exemple plus un désaccord entre le niveau 0 et le niveau 3 qu'entre le niveau 1 et le niveau 2 cf paragraphe 4.2.1.

Si les trois annotateurs semblent difficilement s'accorder sur les quatre niveaux d'intensité (kappas peu élevés), leurs annotations semblent cependant assez corrélées (mesures de Cronbach élevées). Les désaccords sont donc plutôt liés à l'utilisation d'un référentiel différent pour l'axe intensité, i.e. un segment évalué d'intensité 3 par Ann1 serait évalué assez systématiquement

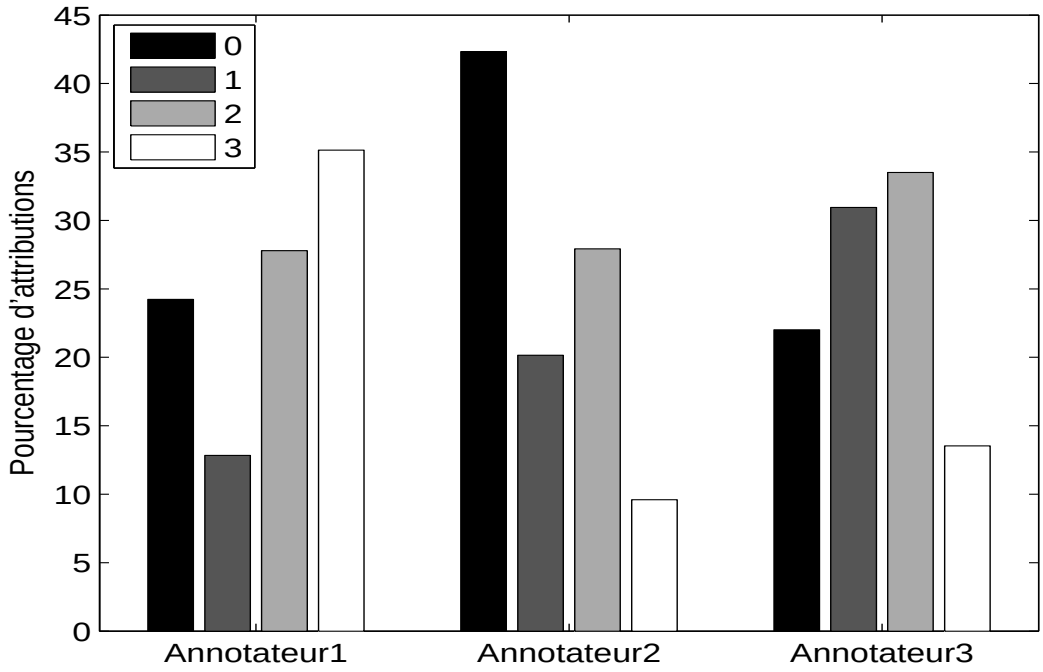

FIG. 4.6 – *Pourcentage d'attributions des 4 niveaux d'intensité en fonction de l'annotateur sur l'ensemble des 4275 segments du corpus SAFE*

	Kappa	Cronbach
Ann1 vs. Ann2	0,20	0,62
Ann1 vs. Ann3	0,24	0,66
Ann2 vs. Ann3	0,32	0,78
Ann1 vs. Ann2 vs. Ann3	0,26	0,77

TAB. 4.6 – *Axe intensité : calcul du kappa et du coefficient alpha de Cronbach pour chaque paire d'annotateurs sur l'ensemble des 4275 segments du corpus SAFE*

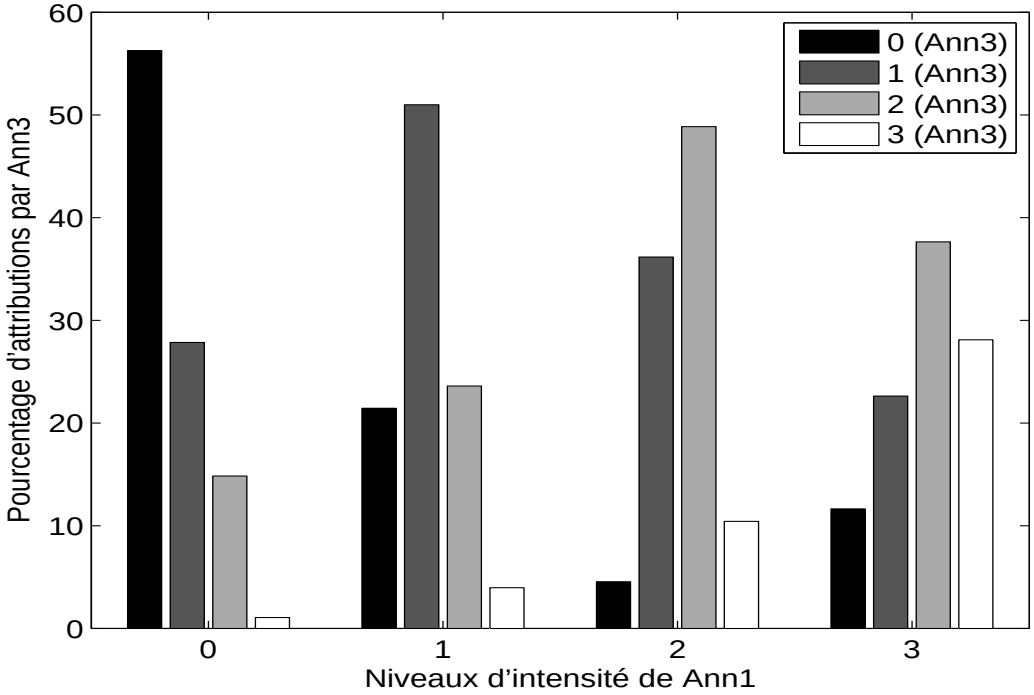

FIG. 4.7 – *Histogramme de confusion entre Ann1 et Ann3 pour les quatre niveaux d'intensité sur l'ensemble des 4275 segments du corpus SAFE*

comme d'intensité 2 par Ann3.

Afin de confirmer ce résultat, l'histogramme de confusion entre Ann1 et Ann3 est représenté sur la figure 4.7. On retrouve sur l'histogramme le plus à droite la distribution des segments évalués d'intensité 3 par Ann1 en fonction des annotations de Ann3. Ces segments ont été majoritairement (37%) évalués d'intensité 2 par Ann3. De manière générale, Ann3 a évalué les segments comme moins intenses, qu'Ann1. A noter également, la frontière entre le degré 0 correspondant à la classe neutre et les segments émotionnels d'intensité 1 semble fragile : 28% des segments évalués d'intensité 0 par Ann1 ont été évalués d'intensité 1 par Ann3 et 21% de ceux évalués d'intensité 1 par Ann1 ont été évalués d'intensité 0 par Ann3.

Évaluation

Comme pour la dimension intensité, nous présentons ici les différents résultats permettant d'évaluer et d'analyser les trois annotations, dans un premier temps avec la distribution des segments du corpus en fonction des sept degrés de l'axe évaluation pour chaque annotateur (cf figure 4.8), ensuite avec le calcul du kappa et du coefficient alpha de Cronbach (cf tableau 4.7) pour chaque paire d'annotateur, et enfin avec l'histogramme de confusion (cf figure 4.9) entre Ann1 et Ann3 qui ont obtenus le kappa le plus faible.

Quelque soit l'annotateur considéré, peu de segments ont été évalués comme positifs de niveau 3. Outre le pic de segments neutres du second annotateur, les distributions ont un aspect similaire, ce qui semble être confirmé par des valeurs des coefficients kappa et alpha de Cronbach un peu plus élevées que pour l'intensité.

Les confusions de Ann1 et Ann3 sont majoritairement situées entre les niveaux -3 et -2.

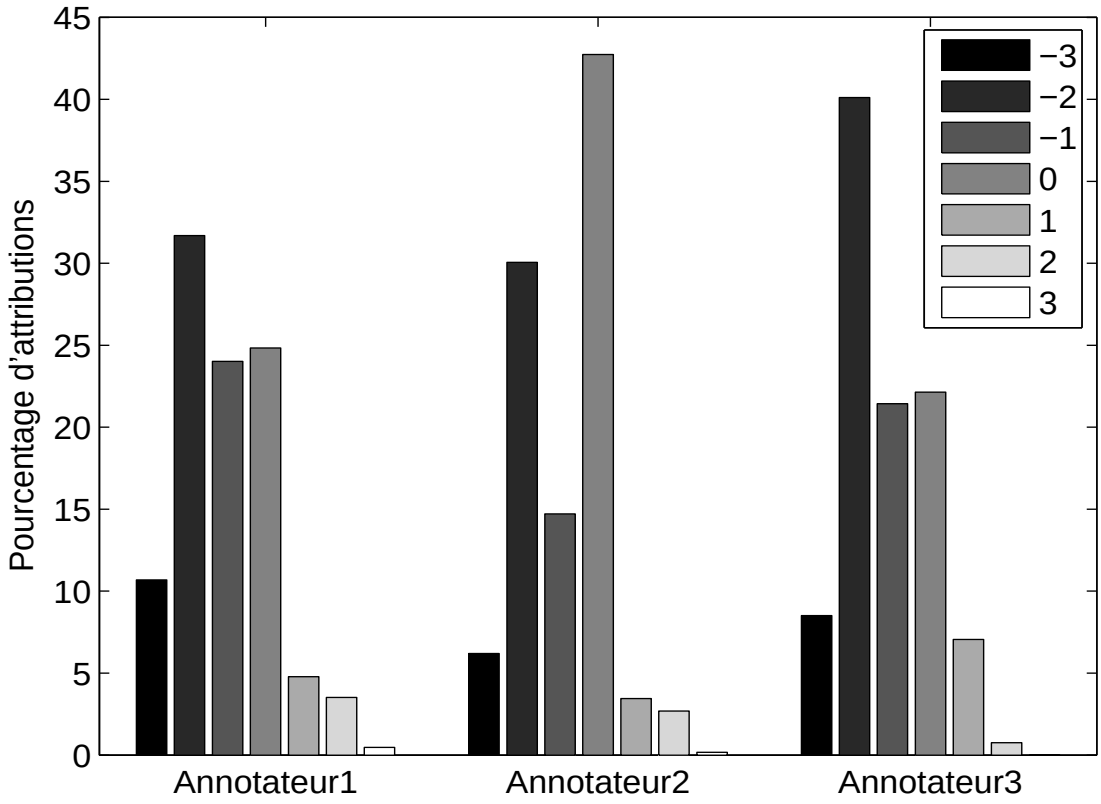

FIG. 4.8 – *Pourcentage d'attributions des différents niveaux sur l'axe évaluation sur l'ensemble des 4275 segments du corpus SAFE*

	Kappa	Cronbach
Ann1 vs. Ann2	0,33	0,80
Ann1 vs. Ann3	0,30	0,82
Ann2 vs. Ann3	0,33	0,79
Ann1 vs. Ann2 vs. Ann3	0,32	0,86

TAB. 4.7 – *Axe évaluation : calcul du kappa et du coefficient alpha de Cronbach pour chaque paire d'annotateurs sur l'ensemble des 4275 segments du corpus SAFE*

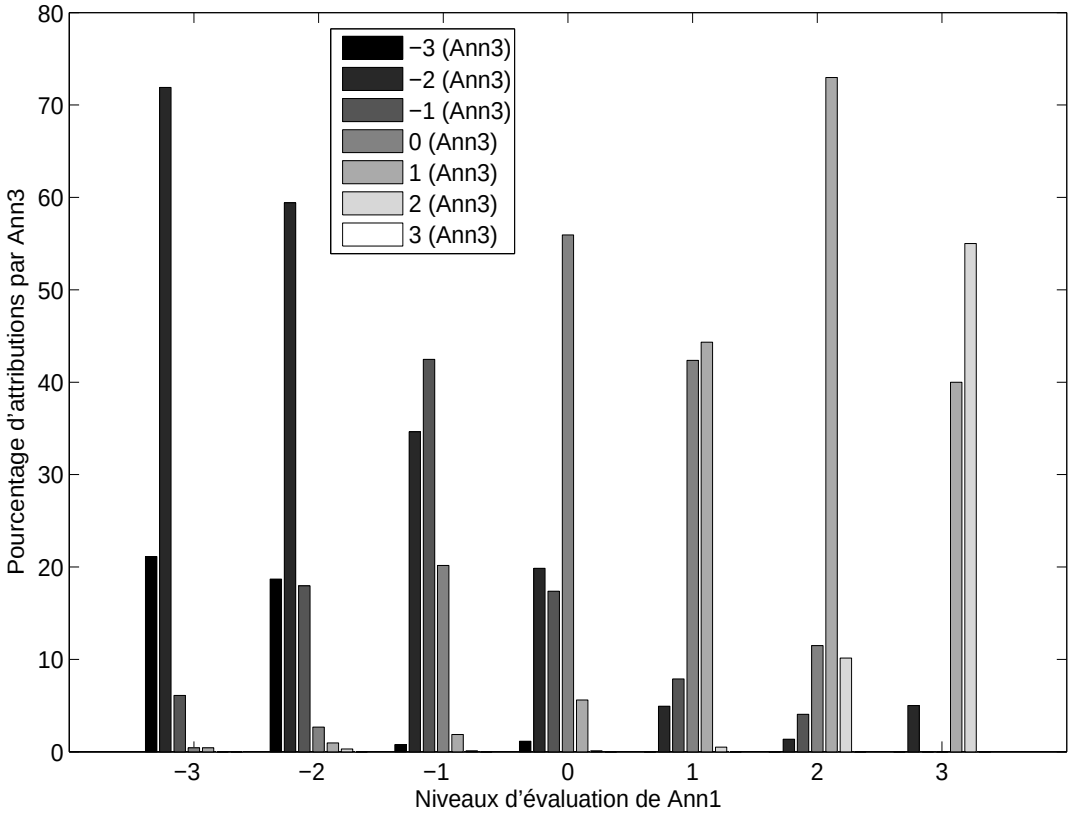

FIG. 4.9 – *Histogramme de confusion entre Ann1 et Ann3 pour les sept niveaux de l'axe évaluation sur l'ensemble des 4275 segments du corpus SAFE*

Les segments évalués de niveau -3 par Ann1 sont annotés de niveau -2 par Ann3. Le troisième annotateur semble préférer de manière générale le niveau intermédiaire -2 pour les émotions négatives. Une grande partie des segments évalués de niveau -1 par Ann1 sont évalués de niveau -2 par Ann3. Les confusions entre ces deux annotateurs même si elles sont nombreuses sont majoritairement des confusions entre deux niveaux de l'axe évaluation très proches. Ceci peut expliquer le fait que la mesure de Cronbach obtenue pour ces deux annotateurs est la plus élevée alors que le coefficient kappa est le plus faible. D'ailleurs pour les quatre catégories globales dérivées de l'axe positif-négatif, ce sont également ces deux annotateurs qui ont obtenu le kappa le plus élevé.

Réactivité

Les mêmes figures et tableaux sont présentés pour l'axe réactivité pour Ann2 et Ann3 uniquement. Comme nous l'avons précisé précédemment, l'échelle de réactivité avait été corrigée à l'issue de la première annotation.

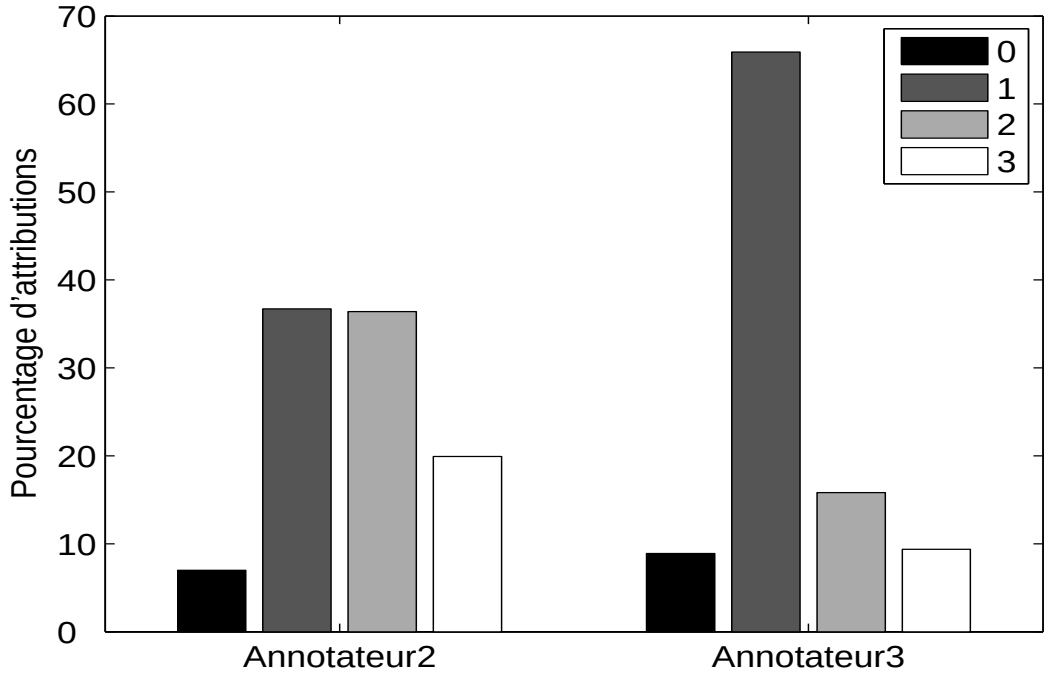

FIG. 4.10 – *Pourcentage d'attributions des différents niveaux sur l'axe réactivité sur l'ensemble des 4275 segments du corpus SAFE*

	Kappa	Cronbach
Ann2 vs. Ann3	0,14	0,55

TAB. 4.8 – *Axe réactivité : calcul du kappa et du coefficient alpha de Cronbach pour Ann2 et Ann3 sur l'ensemble des 4275 segments du corpus SAFE*

Le coefficient kappa et le coefficient alpha de Cronbach sont moins élevés pour cette dimension. Il apparait donc ici que cette dimension est plus sujette à la subjectivité. Ceci peut

être dû au fait que la réactivité a la particularité d'intégrer le comportement du locuteur face au stimulus, contrairement aux deux dimensions précédentes qui se focalisent sur l'émotion. On peut voir sur la figure 4.10 que la majorité des segments sont annotés avec un degré de réactivité de 1 ou 2 par les deux annotateurs avec une forte proportion de segments (presque 70%) annotés avec un degré de réactivité de 2 par le second annotateur. L'annotateur 2 a une distribution des segments plus étalée sur l'axe réactivité.

L'histogramme de confusion 4.11 montre en effet que Ann2 a tendance à annoter les segments avec un degré de réactivité supérieur à Ann3. La majorité des segments annotés avec un degré de réactivité 0 par Ann3 sont annotés avec un degré de réactivité de 1 par Ann2 et 37% des segments annotés avec un degré de réactivité 1 par Ann3 sont annotés 2 par Ann2. Le degré 3 rarement attribué par Ann3 (9%) est plus fréquent chez Ann2 (20%). L'annotateur 3 semble avoir des difficultés à discriminer les différents degrés de réactivité dans le comportement des locuteurs présents dans le corpus. Il serait donc intéressant pour cette dimension de pouvoir disposer d'une annotation supplémentaire.

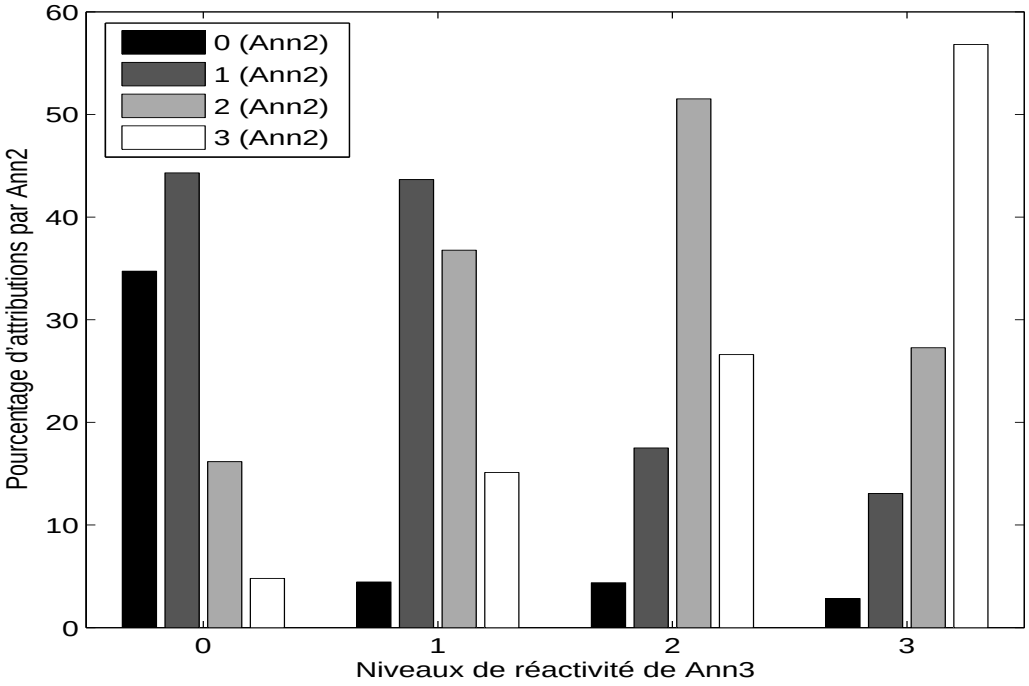

FIG. 4.11 – *Histogramme de confusion entre Ann2 et Ann3 pour les quatre niveaux de réactivité sur l'ensemble des 4275 segments du corpus SAFE*

4.2.4 Bilan

L'annotation selon les quatre catégories globales proposées par notre schéma permet d'obtenir une convergence satisfaisante entre les différents annotateurs. Ce choix de catégories globales a d'ailleurs été adopté par d'autres études [Douglas-Cowie *et al.*, 2003], [Shafran *et al.*, 2003], [Devillers *et al.*, 2005] pour favoriser la convergence des annotations en vue du développement d'un système automatique de classification.

En revanche, les annotations selon les descripteurs dimensionnels posent un certain nombre de problèmes. Les kappas obtenus pour les trois dimensions sont en effet beaucoup plus faibles (entre 0,14 pour la réactivité et 0,32 pour l'évaluation). Les trois annotations selon les trois dimensions s'avèrent cependant corrélées. Cette corrélation est mesurée par le coefficient alpha de Cronbach qui oscille entre 0,55 et 0,86. Chaque annotateur semble utiliser un référentiel qui lui est propre sur chacun des trois axes. De plus ce référentiel a tendance à évoluer au fur et à mesure de l'annotation.

Il est donc difficile d'intégrer directement cette annotation dimensionnelle dans le système de classification automatique qui sera développé. Cependant elle fournit une base pour l'analyse des divergences entre les différents annotateurs pour la catégorisation globale. L'annotation dimensionnelle peut également être utilisée pour l'analyse des erreurs de notre système.

4.3 Contenu du corpus SAFE

L'objectif de cette partie est de mettre en lumière, à travers la description du contenu du corpus, l'adéquation entre les émotions illustrées et l'application visée. A cet effet nous présenterons le contenu global du corpus en termes de locuteurs, de situations et de types d'environnements sonores illustrés ainsi que le contenu émotionnel du corpus.

4.3.1 Contenu global

Les 400 séquences du corpus comptent 5275 segments d'une durée pouvant aller de 40 ms à 80 s avec une moyenne de 4 secondes par segment. La durée d'un segment, l'unité d'annotation définie dans le chapitre précédent, est dépendante des interactions dialogiques entre locuteurs et des variations émotionnelles à l'intérieur d'un même tour de locuteur, ce qui explique sa forte variabilité. L'ensemble des segments représente 85% de la durée totale du corpus, soit 6 heures de parole. Les 15% restant correspondent à des portions de séquences pendant lesquelles aucun des locuteurs ne parle, donc soit à des silences ou du bruit. Les séquences ont été sélectionnées de manière à illustrer une situation particulière normale ou anormale. Le nombre de segments par séquence varie de 1 à 53 avec une moyenne de 13 segments par séquence.

La diversité des locuteurs du corpus SAFE

Le corpus SAFE contient plus de 400 locuteurs différents, dont la répartition en terme de segments est illustrée par le diagramme 4.12. Les 20% dits d'*overlap* correspondent à des recouvrements entre locuteurs incluant des manifestations au niveau du groupe (2%) et de la foule (2%).

L'accent anglais majoritairement illustré dans le corpus est l'anglais américain (70% des données). Les 30% restant correspondent à d'autres accents anglais (britanique, irlandais, canadien) ou à des accents étrangers (français, scandinave, allemand).

La diversité des situations du corpus SAFE

71% des segments du corpus correspondent à des situations anormales. Ces situations anormales sont de natures variées, avec à la fois des catastrophes naturelles de type incendie, inondation, éruption volcanique et des menaces physiques et/ou psychologiques causées par des agressions sur des personnes (prise d'otage, séquestration, agression physique, etc.). Ces situa-

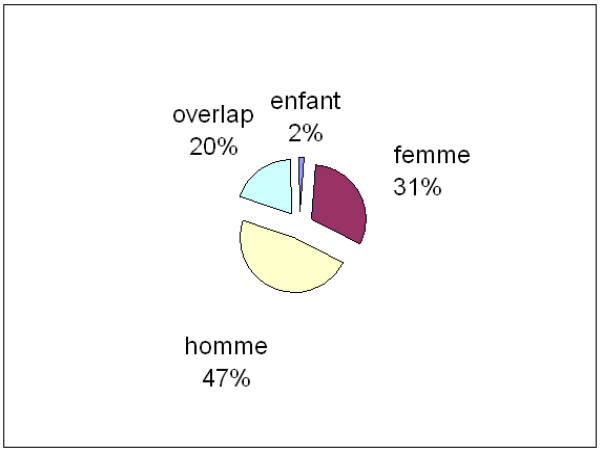

FIG. 4.12 – *Répartition des locuteurs dans le corpus SAFE*

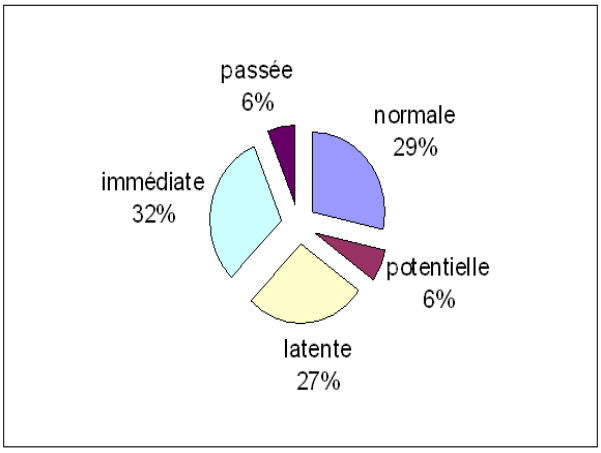

FIG. 4.13 – *Répartition des menaces dans le corpus SAFE en fonction de leur degré d'imminence*

tions sont également représentées avec différents degrés d'imminence (menace potentielle, latente, immédiate, passée) dans les proportions représentées sur le diagramme 4.13.

Les menaces latentes et immédiates sont les menaces les plus représentées dans notre corpus, car elles correspondent à des degrés d'imminence au centre des situations anormales illustrées dans la plupart des séquences. Lorsque les menaces sont encore à l'état potentiel, elles sont souvent illustrées brièvement en début de séquence, même si dans quelques cas isolés, la menace reste à l'état potentiel tout le long de la séquence sans aboutir à une réelle menace. De la même façon, les menaces passées interviennent surtout brièvement en fin de séquence et sont donc moins représentées dans notre corpus.

La diversité des contextes sonores du corpus SAFE

Les environnements sonores sont très variables d'un film à l'autre, d'une séquence à l'autre et voire même à l'intérieur d'une même séquence, comme en témoignent les diagrammes suivants. Les environnements sonores sont en effet fortement dépendants, entre autres, des situations illustrées et de l' évolution de ces situations au cours d'une même séquence. Le premier diagramme (4.14) montre la répartition des segments du corpus en fonction des types d'environnements sonores dans lesquels ils ont été enregistrés (B= bruit seul, M = musique seule, B & M = bruit et musique, propre = sans bruit ni musique) et le second (4.15) en fonction de la qualité de la parole (Q0=mauvaise qualité à Q3=bonne qualité).

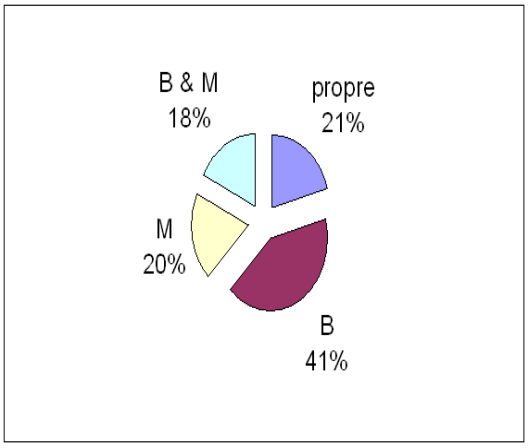

FIG. 4.14 – *Répartition des segments en fonction de l'environnement sonore.*

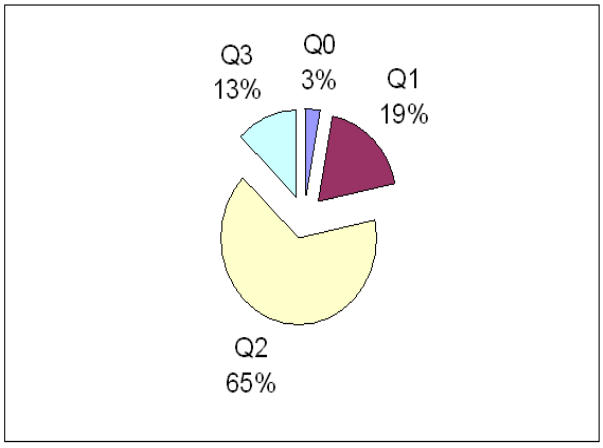

FIG. 4.15 – *Répartition des segments en fonction de la qualité de parole.*

4.3.2 Le contenu émotionnel

Dans la partie précédente, les figures 4.4, 4.6, 4.8 et 4.10 illustrent la répartition des segments du corpus selon l'annotateur considéré en termes de catégories émotionnelles et de dimensions. Si la peur, les émotions négatives et l'état neutre sont équitablement représentés dans le corpus, les émotions positives sont, elles, peu fréquentes et évalués comme peu intenses quel que soit l'annotateur considéré. Ceci est dû aux types de films sélectionnés qui illustrent des situations peu propices à l'émergence d'émotions positives.

Présence de peur extrême

La corrélation entre descripteurs catégoriels et dimensionnels permet une meilleure visibilité des types de manifestations engendrés par chacune des classes. La figure 4.16 représente la distribution des segments de chaque catégorie en fonction de la première dimension, l'intensité. Cette distribution est obtenue après une moyenne des trois annotations.

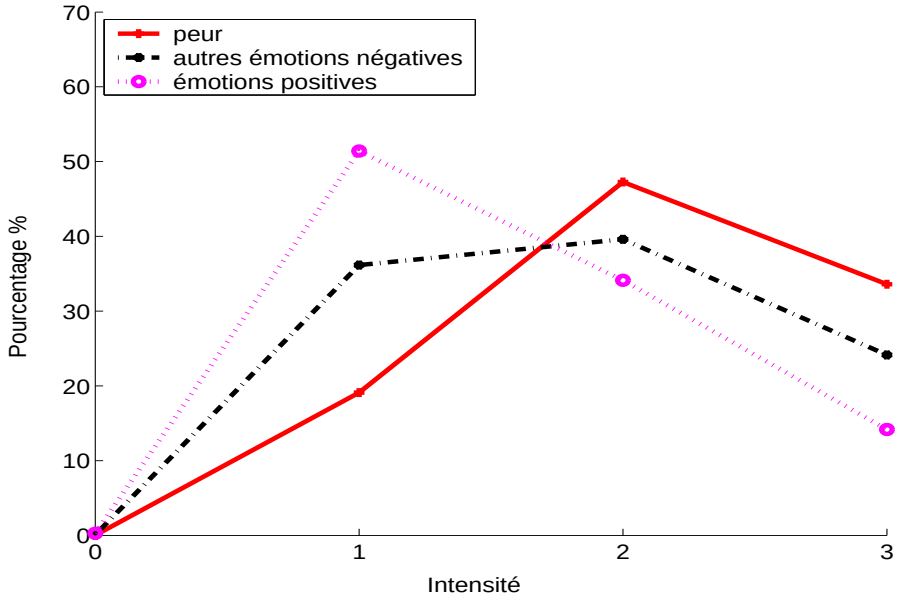

FIG. 4.16 – *Distribution des segments de chaque catégorie en fonction de l'intensité (moyenne des trois annotations)*

Les émotions de type peur sont perçues comme plus intenses que les autres émotions. 80% des segments de peur sont annotés d'intensité 2 ou 3 alors que la majorité des autres émotions sont annotées d'intensité 1 ou 2. De plus la présence de cris pour les émotions de type peur (139 segments contiennent des cris) confirme la présence de peur extrême.

À noter que les émotions positives présentes dans le corpus sont en majorité peu intenses.

Des émotions en situations dynamiques

Les manifestations émotionnelles illustrées dans le corpus varient au fur et à mesure de l'évolution de la situation. La corrélation des annotations des catégories émotionnelles avec les annotations de la situation fournit un matériel riche pour l'analyse des différentes réactions émotionnelles à un type de situation. Nous avons choisi d'étudier ici les différentes manifestations émotionnelles en fonction du degré d'imminence de la menace. L'histogramme 4.17 répertorie pour chaque degré d'imminence de la menace la répartition des segments en fonction de chaque catégorie et le tableau 4.9 en fonction de chaque sous-catégorie de peur.

La peur est l'émotion majoritaire dans le cas de menaces latente (Lat.), immédiate (Imm.) et passée (Pass.) et est en revanche peu illustrée en situations normales (Norm.). Les situations normales présentent majoritairement des segments neutre avec également une forte représentation d'émotions négatives et positives. Les menaces latente et passée sont propices à l'émergence d'émotions négatives autres que la peur. Ce résultat montre la diversité des comportements face à la menace.

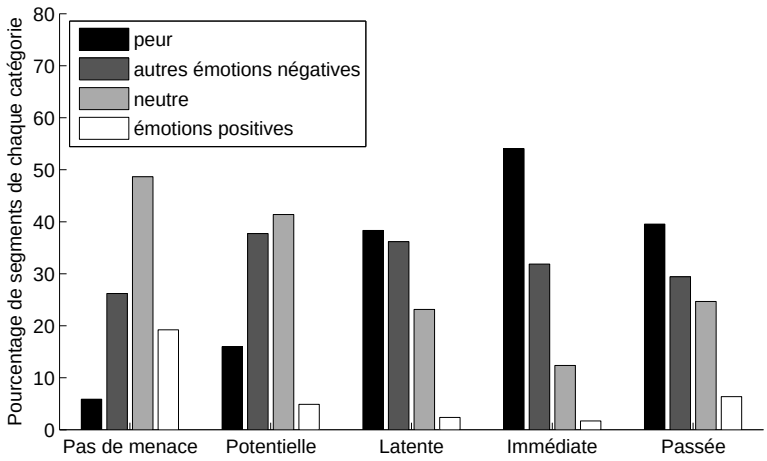

FIG. 4.17 – *Distribution des segments pour chaque degré d'imminence en fonction des catégories*

Les sous catégories les plus fortement représentées sont l'inquiétude, la panique et le stress. Lors de menaces immédiates, la sous-catégorie majoritaire devient la panique avec 31.8% de segments annotés panique.

En situations normales ou de menace potentielle, la sous-catégorie majoritaire est l'inquiétude. On ne trouve pas de détresse ni de peur-souffrance, et très peu de terreur et de stress en situations normales. Ces sous-catégories interviennent soit exclusivement lors de menaces immédiates (peur-souffrance) soit essentiellement lors de menaces qu'elles soient immédiates ou latentes (terreur, détresse et stress). À noter la présence de peur mêlée à d'autres émotions, comme la colère (majoritairement), la surprise, la tristesse et la souffrance.

En situations de menace passée, les sous-catégories majoritaires sont l'inquiétude et la panique. Les menaces passées illustrées dans le corpus sont des menaces qui surviennent plutôt en fin de séquence après une menace immédiate et correspondent donc à un passé proche de la menace immédiate. Ceci explique la présence de manifestations similaires à celles obtenues en menace immédiate, avec cependant une plus forte proportion d'inquiétude. En situation de menace passée on trouve en effet des manifestations émotionnelles liés à l'inquiétude des conséquences ou des dégâts de la menace.

émot. / men.	inqu.	stress	ang.	détr.	pan.	terr.	peur-colère	peur-souffrance	peur-surprise
Norm.	**64,1**	4,3	5,4	0,0	12,0	7,6	3,3	0,0	2,2
Pot.	**61,2**	4,1	6,1	0,0	10,2	6,1	10,2	0,0	2,0
Lat.	**50,9**	9,7	8,3	5,5	17,3	4,2	3,1	0,0	0,0
Imm.	25,2	13,7	6,7	6,0	**31,8**	8,1	3,5	2,7	0,8
Pass.	**42,0**	4,0	4,0	8,0	**22,0**	6,0	6,0	0,0	0,0

TAB. 4.9 – *Répartition des segments en % en fonction de chaque catégorie pour chaque degré d'imminence (Pot. = potentielle, Lat. = latente, Imm. = immédiate, Pass = passée, Norm. = situations normales, inqu. = inquiétude, ang. = angoisse, det. = détresse, pan. = panique, terr. = terreur)*

4.3.3 Le poids des indices acoustiques dans les segments du corpus

L'annotation des émotions est influencée par les informations contenues à la fois dans la modalité audio à travers le contenu acoustique, dans la modalité visuelle et par le contexte sémantique de la séquence. Les tests perceptifs présentés dans le paragraphe 4.1 dans la seconde partie de ce chapitre ont permis d'évaluer le rôle de la modalité vidéo dans l'annotation du contenu émotionnel du corpus. Si les émotions sont perçues comme plus intenses avec l'aide du support vidéo, l'impact de la modalité audio reste dans la majorité des cas suffisamment important pour permettre une évaluation des émotions ciblées par l'application selon les 3 dimensions.

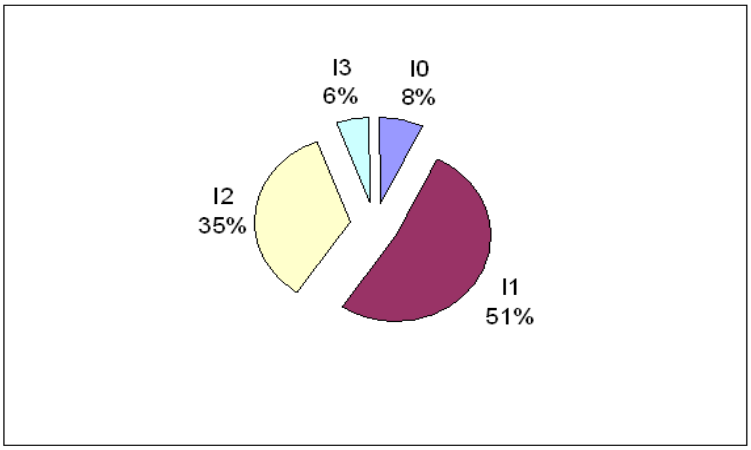

FIG. 4.18 – *Poids des indices acoustiques dans les segments du corpus (de I0 = aucun indice acoustique présent dans ce segment à I3 = indices acoustiques fortement présents)*

Afin d'évaluer plus précisément l'impact des indices acoustiques dans l'annotation du contenu émotionnel, nous avons demandé à l'un des annotateurs, Ann3, d'apprécier sur une échelle allant de 0 à 3 la présence d'indices acoustiques[17] dans chaque segment catégorisé comme non neutre. La répartition de ces segments en fonction des degrés de l'échelle permettant d'évaluer la présence d'indices acoustiques est représenté sur le diagramme 4.18. Seuls 8% des segments ont été évalués comme porteurs d'une émotion sans cependant contenir d'indices acoustiques indiquant la présence d'une émotion. Le reste des segments contiennent des indices plus ou moins présents avec une majorité (52%) de segments ayant des indices acoustiques présents avec une note de 1/3. Cette note est cependant corrélée avec l'intensité de l'émotion et est donc à mettre en parallèle avec la distribution des segments en fonction de l'intensité pour Ann3 dans la figure 4.6. En effet plus l'intensité du contenu émotionnel du segment est élévée, plus les indices acoustiques sont susceptibles d'être présents dans ce segment.

4.4 Corpus et annotations – le point de vue du système

4.4.1 Choix des classes d'émotions traitées

La classe d'émotion visée par notre système est la classe des émotions de type peur. Les quatre classes d'émotion (peur, *autres émotions négatives*, neutre, émotions positives) sont chacune illustrées avec une forte variabilité (locuteurs, situations, contextes sonores) inter- et intra-

[17] les indices linguistiques ne sont pas masqués ici. L'annotateur a uniquement pour consigne de les ignorer pour l'évaluation de la note

classe. Ces variables sont suceptibles d'influencer les caractéristiques acoustiques des segments d'une même classe, pouvant ainsi rendre difficile la distinction inter-classe.

La modélisation acoustique des émotions de type peur peut se faire de deux manières : par rapport à l'état neutre ou par rapport aux autres émotions et l'état neutre. Ce dernier cas est plus subtil car il requiert une distinction entre peur/émotions négatives et peur/émotions positives. La distinction peur/émotions négatives est particulièrement complexe en raison de la variabilité des contextes qui exclut l'émergence d'une peur bien caractérisée. Au contraire, la peur est parfois mélangée à d'autres émotions négatives ou attitudes (peur-colère, peur-panique, peur-tristesse, etc.). Une des causes majeures de désaccord entre les annotateurs s'est d'ailleurs avérée dans le paragraphe 4.2.2 être liée à une confusion entre ces deux classes d'émotion. La distinction peur/émotions positives peut sembler plus aisée. Cependant les manifestations acoustiques de ces deux classes d'émotion peuvent être similaires : il est parfois difficile de distinguer un cri de joie d'un cri de peur avec uniquement des indices acoustiques.

Le développement d'un système de reconnaissance des émotions de type peur sur un tel corpus est une entreprise ambitieuse compte tenu des avancées dans ce domaine. Les systèmes de reconnaissance d'émotions existants, comme nous le verrons dans l'état de l'art du chapitre 6 commencent à travailler sur des données avec un grand nombre de locuteurs. Cependant les contextes sonores et les situations illustrées dans le corpus SAFE présentent un degré d'hétérogénéité qui ne trouve, à notre connaissance, pas d'équivalent dans les corpus utilisés par la communauté.

Une des motivations de notre étude est le contrôle de cette diversité lors du développement du système de reconnaissance. Le cas de la discrimination des émotions de type peur de l'état neutre est le cas plus stable que nous avons choisi de traiter afin de mieux pouvoir contrôler les problèmes susceptibles d'émerger de la forte variabilité interne de chacune des deux classes.

Les différentes études présentées par la suite seront donc effectuées sur un sous-ensemble de ce corpus contenant uniquement les segments annotés peur ou neutre.

4.4.2 Choix des annotations considérées

Le système de reconnaissance d'émotion qui est développé ici doit apprendre à discriminer automatiquement les deux classes peur et neutre. Il doit utiliser pour cela l'information contenue dans le signal sonore. Cet apprentissage est réalisé à partir de l'ensemble des segments du corpus annotés peur ou neutre. Or la composition de cet ensemble varie selon l'annotateur qui a effectué les annotations.

Il est donc crucial de s'assurer que les segments utilisés pour l'apprentissage répondent aux critères suivants :

1. Cet ensemble doit contenir un nombre suffisant de segments ayant des *indices audio perceptibles*. En effet certains segments ont pu être annotés uniquement à partir d'indices visuels. Or, si trop de segments ne possèdent aucun indice audio perceptible, l'apprentissage réalisé sur ces données risque d'être biaisé ;

2. Cet ensemble doit être composé de segments dont les annotations par les deux classes peur ou neutre sont suffisamment *fiables* ;

3. Cet ensemble doit contenir une *quantité* suffisante de données afin que les modèles appris soient suffisamment génériques pour être utilisés sur de nouvelles données.

Compromis entre fiabilité et quantité

Concernant le choix des segments pour constituer l'ensemble d'apprentissage, nous disposons de trois annotations. Le nombre d'annotateurs étant inférieur au nombre de classes, une stratégie de sélection par vote majoritaire n'est pas adaptée ici, car elle ne permet pas d'établir un consensus entre les trois annotations. Un même segment peut en effet être catégorisé en trois classes différentes.

Pour notre étude, il est en revanche possible de considérer les quatre cas de figure suivants :

1. prendre les segments annotés peur ou neutre par les trois annotateurs ;
2. prendre les segments annotés peur ou neutre par deux annotateurs choisis au préalable (Ann1 et Ann2, Ann2 et Ann3, ou encore Ann1 et Ann3) ;
3. prendre tous les segments annotés peur ou neutre par au moins deux annotateurs ;
4. prendre tous les segments annotés peur ou neutre par au moins un des annotateurs.

Le premier cas de figure fournit le sous-ensemble du corpus présentant les annotations les plus fiables. Le sous-ensemble obtenu est en effet composé de segments pour lesquels les trois annotateurs s'accordent sur les deux classes peur et neutre. Le dernier cas de figure permet de disposer d'un plus grand nombre de segments.

C'est le deuxième cas de figure que nous avons choisi de considérer ici, car il constitue un bon compromis entre la quantité de segments considérés et le degré de fiabilité de leur annotation. Le sous-ensemble étudié est composé des segments sur lesquels les deux annotations (Ann1 et Ann2) les plus divergentes[18] s'accordent. Les autres cas de figure n'ont pas été étudiés dans le cadre de cette thèse. Cependant, il serait intéressant d'analyser l'influence du cas de figure considéré sur les performances du système dans le cadre d'une étude ultérieure.

Annotation en « condition système » : audio sans contexte

Demander à Ann3 d'attribuer une note relative à la présence d'indices acoustiques sur chaque segment nous a permis d'analyser l'impact de la modalité audio dans les manifestations émotionnelles présentes dans le corpus. Cependant, il est difficile de s'abstraire des indices fournis par les autres modalités et d'évaluer la saillance d'une modalité indépendamment des autres. Les tests perceptifs nous fournissent également une évaluation de l'impact de la modalité audio dans les manifestations de la peur.

Nous voulons ici plus spécifiquement évaluer le poids des indices audio présents dans les segments peur et neutre qui seront utilisés par le système et vérifier que ces indices sont suffisamment pertinents pour permettre une catégorisation en deux classes peur et neutre.

Une annotation supplémentaire du sous ensemble du corpus constitué des segments annotés peur ou neutre par au moins l'un des deux annotateurs Ann1 et Ann2 a été réalisée ici par un annotateur « expert » dans le domaine du traitement de la parole, appelé par la suite AnnSyst (annotateur « système »). Celui-ci est mis dans les mêmes conditions que celles de notre système automatique de classification peur/neutre. Chaque segment isolé de sa séquence lui est présenté avec comme seul support l'audio. Celui-ci doit alors classer le segment dans la classe peur ou neutre qui lui semble la plus appropriée.

Plus l'annotation fournie par AnnSys est proche de celle fournie par les deux annotateurs, plus le nombre de segments présentant des indices audio permettant à eux seuls une catégorisation peur/neutre est élevé. Le poids des indices audio est donc évalué par le calcul du kappa entre les

[18] les kappas entre chaque paire d'annotateurs ont été calculés et le kappa le plus faible (0,47) a été obtenu par la paire d'annotateurs Ann1/Ann2.

annotations de AnnSys et celles fournies par la paire d'annotateurs Ann1 et Ann2. Les kappa entre les annotations de AnnSys et les annotations de Ann1 puis de Ann2 sont également calculés afin de comparer les deux stratégies d'annotation. Les trois valeurs de kappa obtenues sont présentées dans le tableau 4.10.

	Kappa
Ann1 vs. AnnSys	0,24
Ann2 vs. AnnSys	0,36
Ann1 ∩ Ann2 vs. AnnSys	0,57

TAB. 4.10 – *Calcul du kappa avec l'«annotateur système» sur les segments annotés peur ou neutre selon les annotations de Ann1 (1679 segments pour la classe peur et 1287 pour la classe neutre), de Ann2 (1422 segments pour la classe peur et 2213 segments pour la classe neutre) puis de Ann1 ∩ Ann2 (1118 segments pour la classe peur et 1039 segments pour la classe neutre)*

Ce tableau nous conforte dans le choix de considérer l'ensemble correspondant à l'intersection des deux annotations les plus divergentes. Le kappa obtenu avec l'« annotateur système » sur cet ensemble est de 0,57 et est en effet supérieur aux kappas obtenus sur les ensembles constitués en ne considérant qu'un seul annotateur. Cette valeur de kappa assure ainsi une divergence minimale entre les annotations utilisées par l'ensemble d'apprentissage et les annotations basées uniquement sur l'audio.

En outre l'« annotation système » semble être plus proche de l'annotation fournie par Ann2 ($\kappa = 0,36$) que de celle fournie par Ann1 ($\kappa = 0,24$), ce qui laisse supposer que les annotations fournies par Ann2 reposent plus sur les indices audio que celles d'Ann1 qui semblent au contraire plus influencées par le contexte.

Plus précisément, la comparaison des deux annotations initiales et de l'« annotation système » montre que 54% des segments annotés peur par Ann1 et que 43% de ceux annotés peur par Ann2 sont annotés neutre par AnnSys. En revanche presque tous les segments annotés neutre par Ann1 ou Ann2 sont également annotés neutre par AnnSys. La proportion de segments annotés neutre en considérant l'« annotation système » augmente donc considérablement, ce qui explique les valeurs faibles de kappa obtenus par Ann1 et Ann2.

Dans le paragraphe 4.3.3, nous avons également émis l'hypothèse suivante : plus l'émotion portée par le segment est intense, plus les indices acoustiques sont susceptibles d'être présents dans ce segment. Afin de vérifier cette hypothèse l'annotation de la catégorie peur est ici corrélée avec l'annotation en intensité fournie par Ann1 et Ann2, de manière à considérer les trois catégories suivantes : *peur1*, *peur2* et *peur3* pour les peurs annotées avec des niveaux d'intensité respectivement de 1, 2 et 3.

Les figures 4.19 et 4.20 illustrent les histogrammes de confusion respectivement entre AnnSys et Ann1 et entre AnnSys et Ann2. Il ressort de ces histogrammes que :

- les *peur3* de Ann1 et Ann2 sont presque toutes (plus de 85% d'entre elles) annotées peur par AnnSys ;
- les *peur2* de Ann2 sont majoritairement (près de 60% d'entre elles) annotées peur par AnnSys, contrairement à celles de Ann1 : 65% d'entre elles ont été annotées neutre par AnnSys ;
- les *peur1* de Ann1 et Ann2 sont presque toutes (plus de 80% d'entre elles) annotées neutre par AnnSys ;

Les peurs annotées d'intensité 1 et 2 semblent donc contenir des indices audio plus difficilement perceptibles que les peurs annotées d'intensité 3. De plus, l'annotation fournie par Ann2

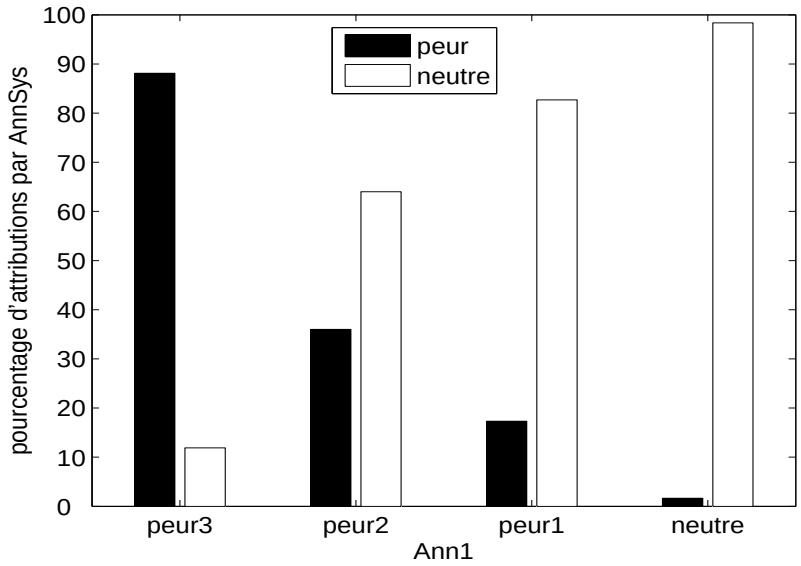

FIG. 4.19 – *Histogramme de confusion entre Ann1 et AnnSys sur les segments annotés peur ou neutre selon les annotations de Ann1 (1679 segments pour la classe peur et 1287 pour la classe neutre)*

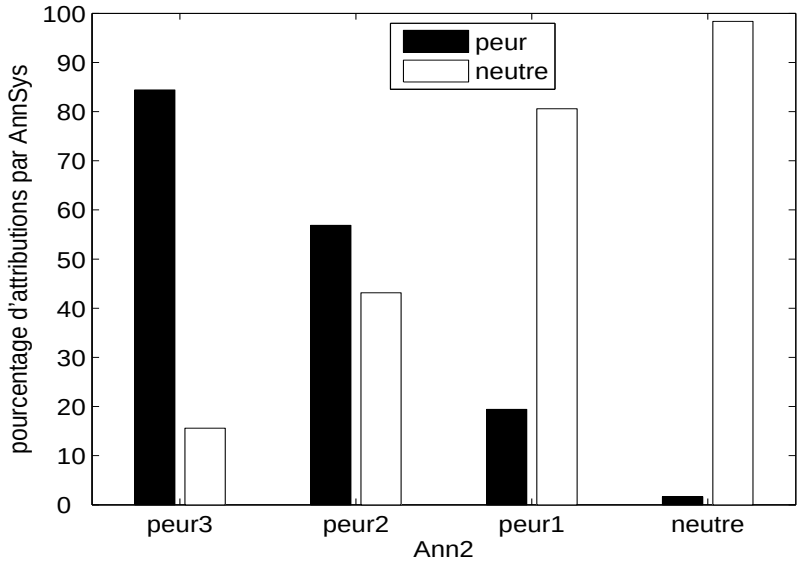

FIG. 4.20 – *Histogramme de confusion entre Ann2 et AnnSys sur les segments annotés peur ou neutre selon les annotations de Ann2 (1422 segments pour la classe peur et 2213 segments pour la classe neutre)*

semble plus reposer sur les indices audio que celles d'Ann1 : les segments annotés *peur2* par Ann2 ont presque tous des indices audio perceptibles.

4.5 Conclusion

Les tests perceptifs, la comparaison des annotations, et l'analyse du contenu du corpus nous permettent de valider la stratégie de sélection des séquences adoptée pour constituer le corpus SAFE : l'émotion cible de notre application, la classe des émotions de type peur, est bien illustrée dans le corpus. Les segments annotés peur représentent environ 30% du corpus.

Les tests perceptifs nous ont également permis de valider la stratégie d'annotation adoptée et notamment d'évaluer le rôle de la vidéo pour l'annotation en comparant les résultats issus de l'annotation des segments dans deux conditions : + vidéo (i.e. avec le support audiovisuel) et - vidéo (avec le support audio uniquement). Les émotions sont perçues globalement comme plus intenses avec le support de la vidéo et les sujets ne disposant que de l'audio ont annoté plus de segments comme neutres. Le contexte temporel fourni aux sujets lors des tests est restreint au contexte fourni par le segment. La proximité des annotations de ces segments obtenue par les deux groupes de sujets (+/- vidéo) témoigne de la quantité d'information disponible dans un segment.

Le kappa obtenu entre les trois annotateurs sur l'annotation en catégories globales est de 0,49. Cette valeur de kappa correspond à un accord modéré dans l'échelle des kappas qu'il est difficile de comparer avec d'autres kappas obtenus pour des tâches d'annotation différentes ou sur des données différentes. Il ressort cependant des différentes études existantes que l'annotation des émotions est une tâche subjective pour laquelle il est difficile d'obtenir des degrés d'accord élevés. La conclusion que nous pouvons tirer de la comparaison des trois annotations est la suivante : l'annotation selon les quatre catégories globales proposées par notre schéma offre une convergence satisfaisante entre les trois annotateurs, plus facilement exploitable par notre système que l'annotation dimensionnelle.

Les différentes manifestations de peur collectées par l'intermédiaire de notre corpus de fiction sont très variées. La corrélation de la description dimensionnelle, de la description en catégories et des informations sur la situation en présence données par le champ menace, nous fournit une visibilité des variables susceptibles d'intervenir lors de la modélisation de cette classe. La classe peur est illustrée pour différents niveaux d'intensité, de valence et de réactivité, pour différents types de menace, pour différents types de locuteur et dans différents environnements sonores.

Enfin, un sous-corpus d'étude est sélectionné. Il est composé des segments annotés peur ou neutre par deux des annotateurs. Une annotation en « condition système » (i.e. sans le support de la vidéo et du contexte temporel fourni par la séquence) a été effectuée. Le kappa obtenu par cette annotation sur ce sous-corpus est de 0,57. Cette valeur nous fournit une référence concernant les performances humaines de catégorisation pour notre système sur le corpus SAFE.

Le système de reconnaissance des émotions développé par la suite sur ce corpus devra surmonter les problèmes suivants :

- des conditions d'enregistrements variables dans un même film et a fortiori entre les films,
- un grand nombre de locuteurs,
- une forte variabilité en termes de contextes illustrés.

Ces problèmes liés aux données sont également susceptibles d'intervenir lors du développement d'un système de surveillance effectif. Par ailleurs la stratégie d'annotation est validée ici sur le corpus SAFE dont la diversité accroît la difficulté de la tâche d'annotation. Cependant, la

validation de la stratégie d'annotation sur des données réelles peut entraîner l'apparition de nouveaux problèmes liés à l'annotation. Cette validation pourrait dans le cadre d'un travail ultérieur faire l'objet d'une étude préalable nécessaire pour l'adaptation du système de reconnaissance des émotions à des données réelles.

Deuxième partie

Analyser et reconnaître les manifestations émotionnelles

Les différentes étapes intervenant dans la conception d'un système de classification d'émotions sont récapitulées sur le schéma de la figure 4.21. Les chapitres de cette partie associés à chaque étape sont également indiqués.

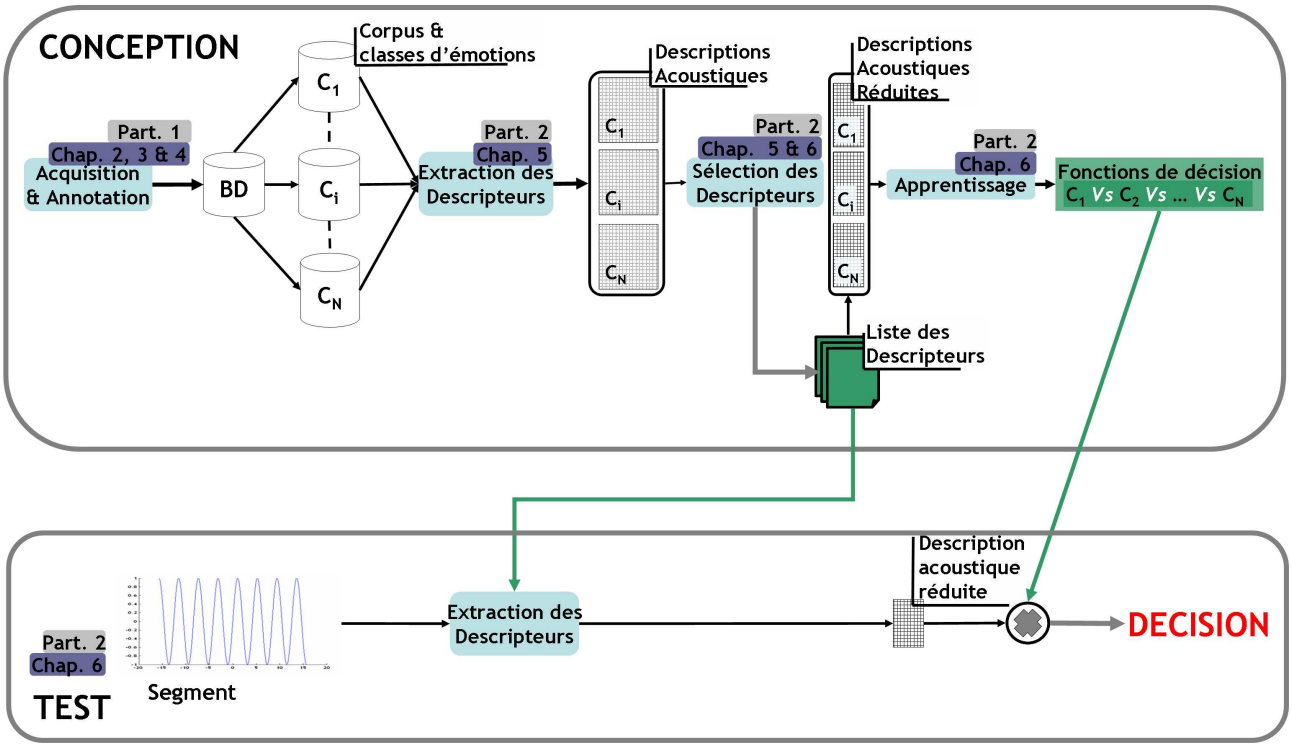

FIG. 4.21 – *Les différentes étapes nécessaires à la mise en place d'un système de reconnaissance automatique d'émotion. Les chapitres de cette partie associés à chaque étape sont indiqués en bleu.*

Résumé

La proportion de parole non voisée est particulièrement importante lors des manifestations des émotions ciblées par notre étude : les émotions de type peur. Notre apport réside dans la considération des descripteurs acoustiques liés aux deux contenus : le contenu voisé et le contenu non voisé. Les descripteurs que nous avons sélectionnés incluent des descripteurs prosodiques, des descripteurs modélisant l'effort vocal et des descripteurs spectraux et cepstraux.
Nous avons choisi de modéliser le contenu émotionnel du corpus avec différents niveaux de temporalité. Les descripteurs acoustiques sont tout d'abord modélisés au niveau de la fenêtre d'analyse, puis les dérivées et les statistiques (ex : écart type, moyenne, etc.) sont extraites au niveau de la trajectoire (voisée ou non voisée) afin de modéliser l'évolution temporelle de chacun des descripteurs. L'approche proposée repose sur l'extraction d'une liste conséquente de descripteurs à différents niveaux temporels afin de sélectionner les descripteurs acoustiques qui s'avèrent les plus efficaces.

Le système de reconnaissance des émotions, développé sur notre corpus, traite un grand nombre de locuteurs inconnus, dans des environnements sonores et contextes variés. Il a pour cible la reconnaissance de la classe émotionnelle qui rassemble les émotions de type peur et repose sur une classification peur/neutre des segments prédéfinis lors de l'annotation. L'évaluation des performances est réalisée grâce au protocole *Leave One Film Out*, qui assure une indépendance complète entre ensemble d'apprentissage et ensemble de test en termes de source contextuelle (profils de locuteur, les situations illustrées et les types de prise de son).
Notre système consiste en la fusion de deux classifieurs voisé et non voisé. Le poids attribué à chacun des classifieurs est fixé pour chaque segment en fonction de son taux de voisement. Cette approche permet de faire baisser le taux d'égale erreur de 32% à 29%. Ce taux d'erreur correspond à un kappa (taux de reconnaissance prenant en compte le hasard) de 0,53, proche du kappa obtenu par l'annotateur dans les « conditions système » dans le chapitre 4. Les résultats obtenus sont très encourageants compte tenu de la diversité des données étudiées.
Nous menons une analyse des comportements du système afin de mieux comprendre l'influence de la diversité inhérente au corpus SAFE sur les performances du système. Nous étudions notamment les performances obtenues en fonction des contextes situationnels liés à la menace.

Chapitre 5

Analyse acoustique des émotions de type peur

Sommaire

Introduction

La première question qui se pose lorsque l'on cherche à accéder à une interprétation haut niveau du signal de parole est : comment transformer les échantillons sonores du signal de parole en une représentation qui soit pertinente vis-à-vis de l'information à reconnaître ?

Cette transformation constitue en effet une première étape pour des objectifs tels que reconnaître la parole, le locuteur, les émotions, la langue. Le domaine du traitement de la parole s'appuie généralement sur les modèles physiques de production de la parole ou sur les modèles de perception et propose une grande quantité de descripteurs acoustiques.

Nous présentons ici les descripteurs acoustiques que nous avons choisis afin de caractériser les émotions cibles de notre application, i.e. les émotions de type peur. La démarche adoptée est la suivante :

1. sélectionner une liste de descripteurs pertinents pour la caractérisation du contenu émotionnel (paragraphe 5.2) ;
2. évaluer leur efficacité pour la caractérisation des émotions de type peur sur les données du corpus (paragraphe 5.4).

L'ensemble des descripteurs acoustiques évalués comme étant les plus efficaces pourra être ainsi utilisé par le système de reconnaissance des émotions présenté dans le chapitre suivant.

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2006d]

5.1 Le signal de parole et les émotions

Nous proposons ici un bilan des descripteurs acoustiques couramment utilisés pour décrire le signal de parole et notamment pour caractériser les manifestations émotionnelles dans la parole. Les études existantes dans ce domaine ont deux objectifs principaux : la synthèse ou la reconnaissance des émotions dans la parole. Il s'agit de déterminer les paramètres[19] ou descripteurs acoustiques caractéristiques d'une émotion donnée pour permettre soit de les manipuler dans une voix de synthèse, soit de les utiliser pour apprendre des modèles acoustiques d'émotion qui seront utilisés pour sa reconnaissance.

5.1.1 Le signal de parole et ses modes de production

Le signal de parole est un signal réel, continu, d'énergie finie, et quasi-stationnaire (i.e. ses caractéristiques statistiques varient lentement sur des durées allant de 5 à 100 ms) au cours du temps. Cependant les sons produits peuvent être très différents impliquant des changements de structure du signal. Celui-ci est en effet tantôt périodique (plus exactement pseudo-périodique) pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt impulsionnel dans les phases explosives des sons occlusifs [Calliope, 1997].

Les variations structurelles du signal de parole que l'on vient de décrire sont dues à des modes de production différents des sons de parole. Afin de mieux comprendre le rôle des différents descripteurs acoustiques, nous proposons ici une brêve description des processus impliqués dans la production de la parole. Ceux-ci peuvent se diviser en trois étapes :

1. génération d'une énergie ventilatoire,

[19] Dans le domaine de la synthèse et parfois dans le domaine de la reconnaissance, les descripteurs sont appelés « paramètres ». Nous préférons ici utilisés le terme descripteur, le terme paramètre interviendra dans le chapitre suivant pour référer aux paramètres du système.

2. vibration des cordes vocales et/ou apparition de bruits d'explosion ou de friction grâce à l'énergie ventilatoire générée par l'expiration phonatoire,
3. réalisation d'une gestuelle articulatoire au niveau des cavités supraglottiques (conduit vocal et fosses nasales).

Le modèle de production le plus utilisé en traitement de la parole est le modèle source/filtre qui assimile les deux premières étapes à la source, i. e. à une étape de génération et la troisième étape à une étape de filtrage.

La seconde étape donne lieu à différents modes de phonation en fonction de la manière dont les cordes vocales sont impliquées :

- le voisement (les cordes vocales sont en vibration) ;
- l'absence de voisement : les cordes vocales sont en position écartée et ne vibrent pas ;
- l'aspiration : courte période non voisée qui se produit pendant et immédiatement après un relâchement articulatoire dans les cavités supraglottiques ;
- le murmure : les cordes vocales vibrent écartées ;
- la laryngalisation : seule une petite partie des cordes vocales est en vibration ;
- l'occlusion glottale : les cordes vocales sont maintenues l'une contre l'autre en position d'adduction ;
- le chuchotement : les cordes vocales sont en contact ou assez rapprochées.

En fonction de ces différents modes de phonation, les sons émis peuvent être organisés en distinguant sons voisés et sons non voisés. Les contenus voisés et non voisés du signal sont définis de la manières suivante :

Définition 5. *On appelle **contenu voisé**, les portions du signal de parole qui ont été produites avec vibration des cordes vocales. Ces portions contiennent les voyelles et les consonnes voisées.*

Définition 6. *On appelle **contenu non voisé**, les portions du signal de parole qui ont été produites sans vibration des cordes vocales. Ces portions contiennent les consonnes non voisées, les aspirations, les respirations, mais également des voyelles chuchotées ou prononcées sans voisement.*

5.1.2 Descripteurs acoustiques et émotions

Une grande partie des descripteurs acoustiques utilisés pour caractériser les différents états émotionnels est destinée à modéliser les modifications du signal acoustique liées à des modifications physiologiques à la base de la glotte. C'est le cas des descripteurs de la qualité de voix (voix soufflée, voix grinçante, voix dure, voix tendue) et des descripteurs prosodiques. Nous présentons dans ce paragraphe un bref bilan de l'utilisation de ces descripteurs dans les études sur la parole émotionnelle. Les descripteurs prosodiques et de qualité de voix utilisés pour notre étude seront détaillés dans le paragraphe 5.2. Ce paragraphe présentera également des descripteurs (formants, coefficients cepstraux, ...) modélisant les modifications du signal acoustique liées à la partie filtre.

Modifications physiologiques et expression vocale de l'émotion

Les modifications physiologiques se produisent à deux niveaux : sur les organes intervenant au niveau de la source (ex : poumon, trachée, muscles de la repiration, glotte, larynx, cordes vocales) et sur les organes intervenant au niveau du filtre (ex : conduit vocal, fosses nasales, langue, muscles des mandibules, lèvres).

Les modifications corporelles/physiologiques qui accompagnent certains états émotionnels, vont fortement influer sur le mode de production du message oral du locuteur [Picard, 1997]. Par

exemple, dans le cas de la peur, les modifications physiologiques typiques sont l'augmentation du pouls et de la pression du sang et la sécheresse de la bouche, et se manifestent par une voix plus forte, et plus aigüe et un débit plus rapide, au contraire de l'ennui et de la tristesse qui sont corrélés avec un abaissement du rythme cardiaque et se manifestent par une voix plus grave, moins intense et un débit plus lent.

Scherer décrit également les modifications de descripteurs acoustiques (fréquence fondamentale, énergie) de l'expression vocale de l'émotion comme la conséquence des changements physiologiques de l'organisme (voir figure 5.1) dans les diverses évaluations d'un stimulus externe (théorie de l'évaluation [Scherer, 2003], *« appraisal theory »*). Selon cette théorie, l'émotion résulte des diverses évaluations d'un stimulus externe qui détermine l'organisation des comportements émotionnels :

- la situation est-elle nouvelle ? (*« novelty check »*) ;
- la situation est-elle plaisante ? (*« intrinsec pleasantness check »*) ;
- favorise-t-elle l'atteinte des buts de l'individu ? (*« goal/need significance check »*) ;
- l'individu dispose-t-il des ressources nécessaires pour y faire face ? (*« coping potential check »*) ;
- la situation est-elle compatible avec ses normes personnelles et socio-culturelles ? (*« norm/self compatibility check »*).

Qualité de voix et émotions

Ainsi, dans [Campbell et Mokhtari, 2003], les auteurs ont développé un algorithme pour la mesure de descripteurs de qualité de voix prenant en compte des modifications physiologiques dans la parole émotionnelle. La qualité de voix est mesurée par le NAQ (Normalized Amplitude Quotient) qui est décrit ici comme un indicateur du niveau de souffle dans la voix : plus le NAQ est élevé plus la voix est soufflée. Dans cet article, ils mettent en évidence le rôle des descripteurs modélisant la source glottale pour reconnaître les émotions. Les données utilisées pour cette étude sont des enregistrements d'une locutrice japonaise dans ses interactions de la vie de tous les jours. Le NAQ est également utilisé dans [Audibert *et al.*, 2004b] pour la caractérisation des émotions.

Les détails du calcul sont présentés dans [Alku *et al.*, 2002]. Le flux glottal est modélisé par une pulsation triangulaire pendant la phase d'ouverture et par un signal nul pendant la phase de fermeture. Le quotient d'amplitude (AQ – *Amplitude Quotient*) correspond au temps de fermeture de la glotte et s'exprime de la manière suivante :

$$AQ = \frac{fac}{dpeak}$$

où fac correspond à l'amplitude maximum du flux glottal et $dpeak$ à la pente de décroissance du flux glottal.

Le NAQ correspond au quotient d'amplitude décorrélé de la fréquence fondamentale (F_0) :

$$NAQ = log(AQ) + log(F_0)$$

Prosodie et émotions

Si les descripteurs de la qualité de voix commencent à être utilisés pour la caractérisation des émotions, les premières études sur la parole émotionnelle se sont basées sur une analyse de la prosodie, avec des descripteurs comme la hauteur (ou fréquence fondamentale) ou l'intensité

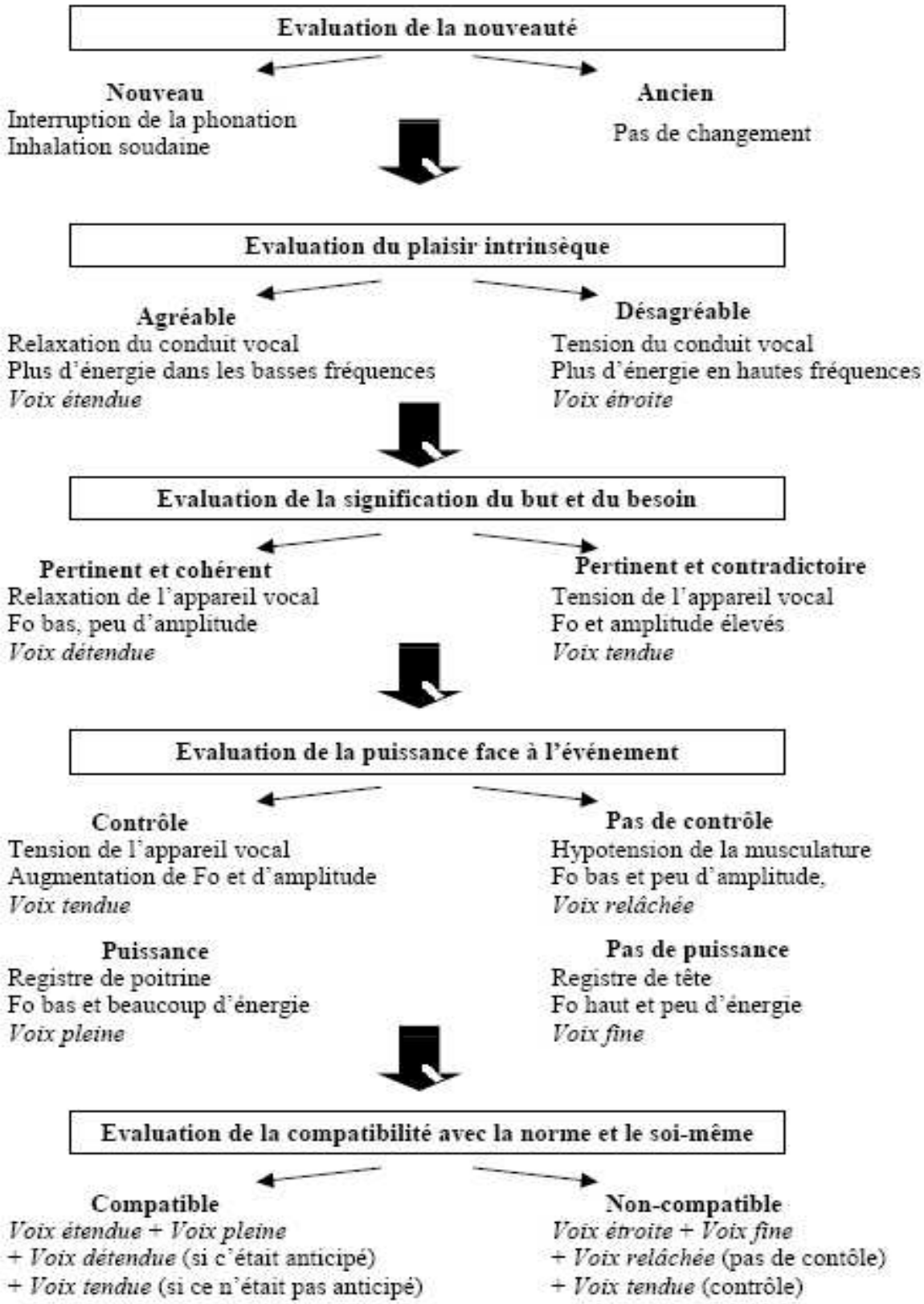

FIG. 5.1 – *Prédiction des changements vocaux dans les différentes étapes d'évaluation tiré de [Chung, 2000]*

de la voix [Dellaert *et al.*, 1996], [Amir et Ron, 1996], qui correspondent en outre à une modélisation des changements supra-segmentaux[20] du signal acoustique produits par les modifications physiologiques à la base de la glotte [Scherer *et al.*, 1998]. Ces descripteurs que nous utilisons pour notre étude sont présentés en détail dans le paragraphe 5.2.

Descripteurs bas niveau vs. haut niveau

L'intensité et la hauteur de la voix ou les descripteurs de la qualité de voix sont des descripteurs dits de *haut niveau* des manifestations émotionnelles, dans le sens où ils fournissent un niveau d'interprétation élevé. Ces descripteurs *haut niveau* sont les plus utilisés dans le domaine de la synthèse ou de l'analyse de parole émotionnelle.

Cependant, dans un objectif de reconnaissance des émotions par des algorithmes d'apprentissage comme dans [Oudeyer, 2003], les descripteurs utilisés sont souvent plus nombreux, incluant des descripteurs plus *bas niveau* pour lesquels il est difficile d'identifier un corrélat perceptif. De nombreux descripteurs *bas niveau* sont extraits et utilisés par le système de reconnaissance d'émotions que nous avons développé. Si l'information acoustique véhiculée par de tels descripteurs est moins explicite que celle correspondant aux descripteurs prosodiques par exemple, cette information s'avère cependant utile et est intégrée dans les modèles acoustiques construits par les algorithmes d'apprentissage.

Les descripteurs bas niveau constituent une grande partie des descripteurs spectraux présentés dans le paragraphe 5.2.

Contenu voisé vs. non voisé

La plupart des études existantes dans le domaine de la recherche des descripteurs acoustiques des émotions dans la parole, centrent leurs travaux sur le contenu voisé de la parole qui se révèle, notamment avec les descripteurs prosodiques, être porteur d'une grande partie des informations caractéristiques des émotions du locuteur. Cependant, selon les émotions étudiées, le contenu non voisé peut également véhiculer des informations pertinentes sur les manifestations émotionnelles. C'est le cas des émotions de type peur qui sont souvent accompagnées de fortes modifications corporelles telles que la crispation, les tremblements, l'augmentation du rythme cardiaque. Par ailleurs, les situations anormales sont propices à des actions du type fuite, poursuite, bagarre, etc. L'activité physique qui accompagne alors la production du message oral entraîne l'émergence de manifestations non verbales telles que des respirations plus fortes, des cris. Ces manifestations vont se répercuter sur les deux types de contenu : voisé et non voisé.

Notre approche sur cet aspect est décrite dans le paragraphe 5.3.3 et consiste à considérer les descripteurs acoustiques liés aux deux contenus (voisé et non voisé) afin de caractériser les émotions cibles.

5.1.3 Unité temporelle d'analyse de l'émotion

Au delà du choix de descripteurs acoustiques pertinents pour caractériser le contenu émotionnel, la question essentielle qui se pose est la suivante : sur quelle durée temporelle doit-on considérer ces descripteurs ?

Le comportement de chaque descripteur est en effet dépendant de la fenêtre temporelle sur laquelle il est considéré. La plupart des travaux utilise des descripteurs de type statistique proposant une modélisation globale de chaque descripteur (par exemple la moyenne, le minimum,

[20] en fonction de l'accentuation, de l'intonation ; appelé également niveau prosodique

etc.) sur différentes durées temporelles, comme la syllabe, le mot, ou la phrase. Ces analyses reposent donc sur une segmentation fine de la parole.

Le type de corpus utilisé ici et ceux sur lesquels notre système est destiné à être utilisé pour une application de surveillance sont des corpus qui présentent des manifestations extrêmes d'émotions telles que les émotions de type peur. Les manifestations liées à l'émotion cible rendent difficile la segmentation du signal de parole en des unités phonétiques distinctes avec notamment des phénomènes tels que la coarticulation, des voix non modales, des manifestations non verbales telles que les cris, etc. C'est pourquoi nous avons choisi de ne pas procéder à une segmentation de la parole comme prétraitement au calcul des descripteurs acoustiques.

D'autres analyses sont basées sur des descripteurs acoustiques extraits sur chaque fenêtre d'analyse [Amir et Ron, 1996]. Elles présentent l'avantage de ne pas faire de présupposé sur la structure de la parole et ne nécessitent pas de connaissance sur le contenu linguistique associé au signal de parole.

L'approche choisie ici est d'utiliser des descripteurs proposant une modélisation à différents niveaux temporels, comme présenté dans le paragraphe 5.3.3.

5.2 Choix de descripteurs acoustiques pour la caractérisation des émotions de type peur

Nous avons choisi pour caractériser le contenu acoustique des émotions de type peur des descripteurs acoustiques parmi les trois familles suivantes : les descripteurs prosodiques, les descripteurs de qualité de voix, les descripteurs spectraux. Nous utilisons les logiciels Praat [Boersma et Weenink, 2005] et Matlab pour l'extraction des descripteurs. Le détail et les paramètres du calcul de chacun des descripteurs est présenté ci-dessous.

5.2.1 Les descripteurs prosodiques

Initialement utilisés dans le cadre de la reconnaissance vocale ou de la synthèse vocale, les descripteurs prosodiques classiques (hauteur et intensité de la voix, durée des syllabes) ont pour vocation la structuration du flux de parole. Les descripteurs prosodiques sont souvent utilisés pour l'identification d'éléments segmentaux déterminés et peuvent fournir également des indices sur la structure syntaxique de la phrase. En synthèse ils permettent de rendre le signal synthétique plus naturel et notamment intelligible en indiquant les grandes articulations de la phrase.

Les descripteurs prosodiques permettent de modéliser les accents, le rythme, l'intonation, la mélodie de la phrase et sont ainsi très pertinents pour la modélisation de l'état émotionnel du locuteur. Les descripteurs prosodiques, tels que le pitch ou l'intensité sont fréquemment utilisés dans le domaine de la parole émotionnelle [Kwon *et al.*, 2003] [McGilloway, 1997] [Schuller *et al.*, 2004] et ont fait leur preuve dans le cadre de la modélisation de la peur [Scherer, 2003], [Devillers et Vasilescu, 2003], [Batliner *et al.*, 2003].

La fréquence fondamentale ou pitch

En parole, la fréquence fondamentale ou pitch caractérise les parties voisées du signal de parole et est liée à la sensation de hauteur de la voix (aigüe ou grave). Les parties voisées ont une structure pseudo-périodique et sur ces portions, le signal est généralement modélisé comme la somme d'un signal périodique T et d'un bruit blanc. La fréquence fondamentale est l'inverse de la période T, $F_0 = \frac{1}{T}$.

Il existe plusieurs méthodes pour l'estimation de la fréquence fondamentale : les méthodes temporelles (autocorrélation, fonctions de différences moyennées (ASDF – Average Square Difference Function)), les méthodes d'estimation par maximum de vraisemblance, les méthodes reposant sur une analyse du cepstre, etc. [Calliope, 1997].

La fréquence fondamentale est ici extraite à l'aide du logiciel Praat [Boersma et Weenink, 2005], qui utilise une méthode temporelle d'estimation du pitch consistant à rechercher des ressemblances entre des versions décalées du signal observé s. Ces ressemblances sont évaluées par la fonction d'autocorrélation, définie de la manière suivante :

$$r_s(m) = \begin{cases} \frac{\sum_{n=0}^{N-1-m} s(n)s(n+m)}{\sqrt{\sum_{n=0}^{N-1-m} s(n)^2}\sqrt{\sum_{n=0}^{N-1-m} s(n+m)^2}} & si\ m \geq 0 \\ r_s(-m) & sinon \end{cases}$$

La période T est estimée en recherchant la plus petite valeur de m pour laquelle $r_s(m)$ est maximale. La fonction d'autocorrélation donne également une estimation de « la force de voisement » du signal de parole : plus $r_s(T)$ est proche de 1 (la valeur maximum possible), plus le signal se rapproche d'un signal de période T.

Praat utilise également une méthode de résolution du saut d'octave robuste permettant d'éliminer les doublements de fréquence.

La figure 5.2 illustre le comportement de la fréquence fondamentale sur deux *segments* du corpus tous deux annotés par la catégorie globale *peur*, le premier ayant été annoté par la sous-catégorie *inquiétude* et le second par la sous-catégorie *panique*.

Ces deux *segments* correspondent au même mot « Josh » prononcé par une même locutrice au cours d'une même séquence. Pour le premier segment, la locutrice vient de se rendre compte de la disparition de son ami Josh et l'appelle une première fois. Les indices acoustiques de la peur présents dans ce segment sont difficilement perceptibles et on note la montée de la fréquence fondamentale caractéristique d'une question. Pour le second segment, la locutrice, après de nombreux appels sans réponse, hurle le nom de son ami. Dans le second cas, le contour de la fréquence fondamentale est beaucoup moins lisse avec de nombreux sauts fréquentiels, les erreurs d'estimation mises à part, et une fréquence fondamentale en moyenne plus élevée.

L'intensité

L'intensité correspond à la variation de l'amplitude de signal de parole causée par une énergie plus ou moins forte provenant du diaphragme et provoquant une variation de la pression de l'air sous la glotte. Ce descripteur permet de fournir une mesure de la force sonore de la voix (faible ou forte). L'intensité en décibel (dB) est ici calculée sur une portion de signal de longueur N à l'aide du logiciel Praat de la manière suivante :

$$I = 10\log(\sum_{n=1}^{N} s^2(n)w(n))$$

où w est une fenêtre d'analyse gaussienne.

La figure 5.3 illustre le comportement de l'intensité sur les deux mêmes exemples que dans le paragraphe précédent. Dans le cas de la panique, l'intensité moyenne est plus élevée avec plus de modulations.

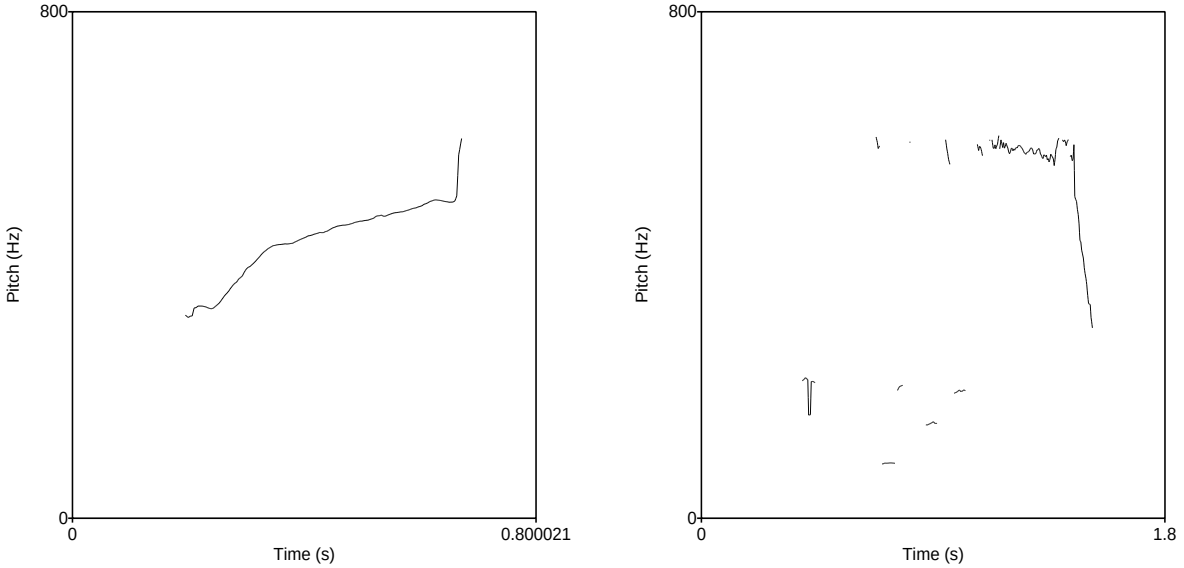

FIG. 5.2 – *Exemple de contour de la fréquence fondamentale sous Praat calculé sur : (gauche) le segment « Josh ? » annoté peur avec une intensité de 1 (sous-catégorie : inquiétude) et (droite) le segment « Joooosh ! » annoté peur avec une intensité de 3 (sous-catégorie : panique).*

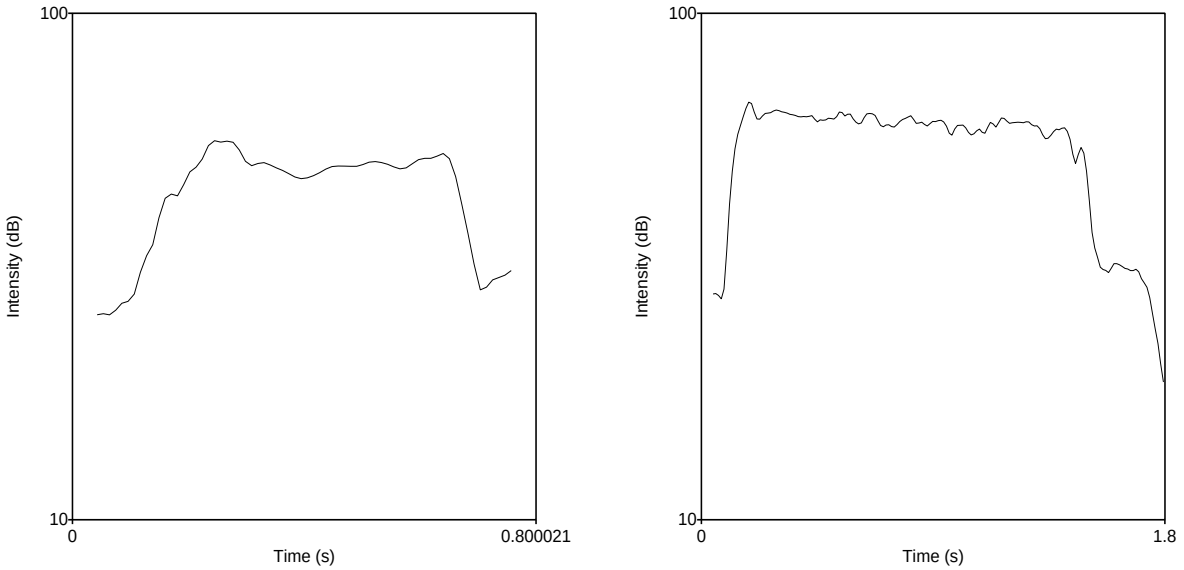

FIG. 5.3 – *Exemple de contour d'intensité sous Praat calculé sur : (gauche) le segment « Josh ? » annoté peur avec une intensité de 1 (sous-catégorie : inquiétude) et (droite) le segment « Joooosh ! » annoté peur avec une intensité de 3 (sous-catégorie : panique).*

Durée de la trajectoire voisée

Le troisième descripteur classique de la prosodie est un descripteur du rythme, i.e. de la cadence de la phrase et se mesure par le nombre d'unités vocales par unités de temps, par exemple le nombre de syllabes ou de phonèmes par minute. Cependant cette mesure du rythme ne peut être calculé que si l'on dispose d'une segmentation du signal de parole en syllabes. Pour des raisons indiquées dans le paragraphe 5.3.3, nous avons fait le choix de proposer une analyse acoustique qui ne s'appuie pas sur cette segmentation.

La durée calculée ici correspond à la durée de chaque trajectoire voisée. La trajectoire voisée correspond aux fenêtres voisées consécutives. Ce descripteur permet de caractériser les différences de débit dans le flux de parole, une forte proportion de longues trajectoires voisées étant le signe d'un débit plus lent.

5.2.2 Les descripteurs de qualité de voix

Les variations de qualité de voix sont entre autre la conséquence des modifications physiologiques provoquées par les changements émotionnels du locuteur. La qualité de voix, comme nous l'avons vu précédemment, commence à être utilisée pour caractériser les variations émotionnelles. Certains travaux sur la parole émotionnelle visent à étudier les interactions fortes qu'il pourrait y avoir entre la configuration glottale et l'expression de l'émotion. Cependant la plupart de ces travaux nécessitent des conditions d'enregistrement idéales, comme une proximité du micro à la bouche, qui nous ont semblé difficile à obtenir dans un contexte réel de surveillance. Compte tenu de ces contraintes, nous avons choisi de modéliser la qualité de voix par les descripteurs suivants :

La modulation fréquentielle ou jitter

Le jitter est décrit comme une déviation de la fréquence fondamentale, et correspond à la mesure d'un bruit sur la fréquence. En musique il sert, avec le shimmer ou modulation d'amplitude, à modéliser l' attaque, le soutien et le relâchement des notes [Verfaille, 2003]. Il est également utilisé pour l'étude des voix pathologiques (voix grinçante) et a été utilisé notamment par [France *et al.*, 2003] pour la caractérisation d'émotions. La figure 5.2 met notamment en évidence la présence de fortes modulations de la fréquence fondamentale sur le cri, « Jooosh ! ».

Le jitter permet de modéliser ces oscillations autour de la fréquence fondamentale de la voix. Il est calculé ici à l'aide de Praat. La formule utilisée pour son calcul sur une portion du signal de longueur N est :

$$Jitter = \frac{\sum_{n=2}^{N-1} |2T_n - T_{n-1} - T_{n+1}|}{\sum_{i=1}^{N} T_n}$$

où T_n est la période estimée à l'instant n.

La modulation en amplitude ou *shimmer*

Souvent utilisé de pair avec le jitter, le shimmer modélise, lui, la modulation d'amplitude. Comme pour le jitter, la figure 5.3 met en évidence la présence de fortes modulations de l'intensité sur le cri, « Jooosh ! ». Ces modulations sont caractérisées ici à l'aide de Praat qui permet le calcul du shimmer sur une portion de signal de longueur N :

$$Shimmer = \frac{\sum_{n=2}^{N-1} |2A_n - A_{n-1} - A_{n+1}|}{\sum_{i=1}^{N} A_n}$$

où A_n est l'amplitude de la période T_n.

Taux de fenêtres non voisées

Le taux de fenêtres non voisées est calculé sous Praat et correspond à la proportion de fenêtres considérées comme non voisées sur une portion du signal de parole. Une fenêtre est considérée comme non voisée si la « force du voisement » (score de la fonction d'autocorrélation) est inférieure à un certain seuil fixé sous Praat à 0,45.

Le rapport harmonique sur bruit

L'idée de ce descripteur est de trouver un indicateur du niveau de souffle dans la voix en mesurant le rapport harmonique sur bruit du signal de parole par le descripteur connu sous le nom de PAP (Périodique APériodique) ou HNR (Harmonic to Noise Ratio). L'algorithme est celui présenté dans [Yegnanarayana *et al.*, 1998] et repris dans [Ehrette, 2004] qui repose sur l'estimation du degré de remplacement des harmoniques par du bruit, c'est à dire le rapport entre l'énergie acoustique des composantes harmoniques du signal sur l'énergie acoustique des composantes du bruit.

L'algorithme comprend une partie d'analyse du signal et une partie de resynthèse du signal en ses deux composantes : le bruit s_{bruit} est défini comme le complémentaire de la partie harmonique $s_{harmonique}$ du signal s_{vocal} avant passage dans le conduit, suivant la modélisation suivante :

$$s_{vocal} = s_{harmonique} + s_{bruit}$$

Le bruit dans la voix correspond à la fois au bruit rose fricatif, sifflant ou plosif et à la composante non périodique. La modélisation du signal vocal est alors réajustée de la manière suivante :

$$s_{vocal} = s_{periodique} + s_{aperiodique}$$

Le PAP est adapté à l'estimation de la contribution du bruit, en considérant à la fois le bruit dû aux irrégularités des oscillations des cordes vocales (non harmonicité de l'onde glottique) et au bruit additif. Il s'exprime par la formule suivante :

$$PAP = 10\log(\frac{Energie_{periodique}}{Energie_{aperiodique}})$$

Praat propose une autre méthode pour le calcul du HNR. Les descripteurs extraits ici intègrent les deux méthodes : le PAP de Yegnanarayana implémenté sous le logiciel Matlab et le HNR calculé sous Praat.

5.2.3 Les descripteurs spectraux et cepstraux

Nous avons également choisi d'utiliser des descripteurs de plus bas niveau reposant sur une analyse du spectre du signal : les descripteurs spectraux (descripteurs formantiques, énergie en bande de Bark, centroïde spectral) et les descripteurs cepstraux (MFCC – *Mel Frequency Cepstral Coefficients*).

Les formants et leur largeur de bande

Un formant est un pic d'amplitude dans le spectre d'un son composé de fréquences harmoniques, inharmoniques et/ou de bruit. Le conduit vocal présente des fréquences de résonance, ce qui se manifeste dans le spectre par l'apparition de pics formantiques. La largeur de bande du formant est définie comme la largeur de la bande du spectre entre les points à -6 dB par rapport à la crête du formant.

Sur le contenu voisé, la position des deux premiers formants est indépendante de la hauteur (fréquence fondamentale) et est caractéristique d'une voyelle particulière. Une voyelle est en effet produite en imposant une position particulière aux différents articulateurs (lèvres, langue, ...).

Les formants et leurs largeurs de bande sont également intéressants sur les fenêtres non voisées car ils fournissent une modélisation du conduit vocal et sont par là même des descripteurs de la qualité de voix.

Les deux premiers formants F_1 et F_2 et leurs largeurs de bande Bw_1 et Bw_2 sont calculés ici à l'aide du logiciel Praat par une analyse LPC (Linear Prediction Coding - codage prédictif linéaire).

L'algorithme utilisé sous Praat pour calculer les coefficients de prédiction linéaire est l'algorithme de Burg [Calliope, 1997] qui permet d'obtenir des largeurs de bande de chacun des formants.

Très largement utilisés comme descripteurs des émotions, comme par exemple dans [France *et al.*, 2003], [Kienast et Sendlmeier, 2000], les formants nous ont semblé très pertinents dans le cadre de la caractérisation acoustique des émotions de type peur, d'une part pour leurs aspects de descripteurs de la qualité vocale, d'autre part car le premier formant peut permettre également de modéliser le degré d'ouverture des voyelles que l'on peut prévoir plus élevé dans le cas des cris. Par ailleurs, certains états émotionnels provoquent des changements dans l'articulation produisant des modifications plutôt à court terme et segmentales du signal acoustique mesurées par la fréquence et la largeur de bande des formants [Scherer *et al.*, 1998].

Mel-Frequency Cepstral Coefficients (MFCC)

La paramétrisation MFCC est une paramétrisation très répandue dans le domaine du traitement de la parole, que ce soit en reconnaissance automatique de la parole, en reconnaissance du locuteur ou en reconnaissance des langues. Les MFCC ont également été utilisés dans le domaine de la reconnaissance des émotions [Shafran *et al.*, 2003] [Kwon *et al.*, 2003].

Les MFCC appartiennent à la famille des descripteurs cepstraux qui se basent sur une représentation cepstrale du signal. Le cepstre présente l'avantage de permettre une séparation des contributions respectives de la source et du conduit vocal (voir paragraphe 5.1.1). En effet, si le signal de parole $s(t)$ est représenté sous la forme de la convolution du signal source $g(t)$ par la réponse impulsionnelle du filtre représentant le conduit vocal $h(t)$:

$$x(t) = g * h(t)$$

le cepstre, défini comme la transformée de Fourier inverse (TF^{-1}) du logarithme du spectre d'amplitude $|S(\omega)|$, s'exprime comme la somme de deux termes, l'un caractéristique de la source et l'autre caractéristique de l'enveloppe spectrale :

$$c(\tau) = TF^{-1}|S(f)| = TF^{-1}|G(f)| + TF^{-1}|H(f)|$$

Le paramètre τ est homogène à un temps et est appelé quéfrence. Il est montré dans [d'Alessandro, 2002] que les coefficients cepstraux correspondant aux basses quéfrences représentent ainsi principalement la contribution du filtre.

Les MFCC s'obtiennent en utilisant, pour le calcul du cepstre, une échelle fréquentielle non linéaire tenant compte des particularités de l'oreille humaine, l'échelle des fréquences Mel. L'échelle Mel correspond à une approximation de la sensation psychologique de hauteur d'un son qui prend notamment en compte la caractéristique suivante : la sélectivité en fréquence est plus grande dans les graves que dans les aigus. L'échelle des fréquences Mel s'obtient par l'expression suivante :

$$m(f) = 2595 \log(1 + \frac{f}{700})$$

où f est la fréquence en Hertz.

Pour chaque trame du signal sonore, le spectre d'amplitude $S(k)$ est intégré par bandes Mel pour obtenir un spectre d'amplitude modifié $\tilde{a}_m$ $m = 1..M_b$, représentant l'amplitude de la m^e bande Mel. A cet effet, Praat utilise des filtres triangulaires de largeur de bande constante et régulièrement espacés sur l'échelle Mel. Enfin, les coefficients MFCC sont obtenus en effectuant une transformée en cosinus discrète du logarithme des coefficents obtenus précédemment :

$$\tilde{c}(\tau) = \sum_{m=1}^{M_b} \log(\tilde{a}(m) \cos[\tau(m - \frac{1}{2})\frac{\pi}{M_b}])$$

où M_b est le nombre de filtres triangulaires. Nous avons choisi de garder les 12 premiers coefficients $\tilde{c}(\tau)$.

L'énergie en bandes de Bark

L'énergie en bandes de Bark repose sur une autre échelle perceptive couramment utilisée, l'échelle de la tonie dont l'unité est le Bark. Le spectre du signal est pour ce descripteur découpé en différentes bandes de fréquence déterminées par cette échelle. L'échelle Bark est basée sur les bandes critiques telles qu'elles sont perçues par l'oreille et la formule donnant la tonie en Bark en fonction de la fréquence en Hertz utilisée ici repose sur l'approximation suivante présentée dans [Serkey et Hanson, 1984] :

$$f_{Bark} = 6 \arcsin(\frac{f_{Hz}}{600})$$

Nous utilisons ce descripteur tel qu'il a été calculé par [Ehrette, 2004] en découpant le signal en différentes bandes de fréquences de largeur égale à un ou deux Barks. Pour chaque bande i, l'énergie s'exprime de la manière suivante :

$$EBB(i) = \sum_{f_{inf}(i)}^{f_{sup}(i)} |S(f)|^2$$

où i est le numéro de la bande considérée.

L'énergie dans des bandes de fréquences particulières est un descripteur qui a été utilisé pour expliquer les différentes catégories perceptives de voix dans [Ehrette, 2004] ou dans [Kienast et Sendlmeier, 2000] et [Cai *et al.*, 2003] pour l'étude des manifestations acoustiques des émotions.

Centroïde spectral

Le centroïde spectral ou balance spectrale est utilisé ici, comme dans [Ehrette, 2004], en tant que descripteur du « timbre » de la voix. Dans [Kienast et Sendlmeier, 2000], ce descripteur est calculé sur les fricatives non voisées et permet de mesurer le degré de constriction. Plus

généralement utilisé pour décrire une sensation de brillance auditive, il a été également utilisé, entre autres, pour la reconnaissance d'instruments de musique [Essid, 2005]. Il correspond au moment spectral d'ordre 1 et est calculé de la manière suivante :

$$C_s = \frac{\sum_k kEBB(k)}{\sum_k EBB(k)}$$

où k est le numéro de la bande de fréquence et a_k l'énergie de cette bande.

Le calcul des MFCC, des énergies en bandes de Bark et du centroïde spectral a été implémenté sous Matlab.

5.3 Paramètres d'extraction des descripteurs

Les descripteurs présentés ci-dessus ont été utilisés en tant que descripteurs acoustiques du contenu émotionnel du corpus SAFE. Nous présentons dans ce paragraphe les paramètres du signal audio traité et les paramètres de calcul des descripteurs.

5.3.1 Paramètres d'échantillonnage du signal

Les séquences audio du corpus SAFE ont une fréquence d'échantillonnage de 48kHz et possèdent ainsi une bande passante de 24 kHz. Chaque échantillon est codé sur 16 bits. Cette qualité d'enregistrement est largement suffisante pour traiter des signaux de parole sachant qu'une oreille humaine performante perçoit des fréquences comprises entre 20Hz et 20kHz.

5.3.2 Normalisation du signal

Les conditions d'enregistrement peuvent être très variables d'un film à l'autre. Afin de limiter l'effet de cette variabilité sur les performances de classification, une normalisation de la forme d'onde du signal associée à chaque séquence est effectuée. Le signal normalisé $\hat{s}$ est obtenu à partir du signal $s(n)$ d'origine en réalisant les deux opérations suivantes :

1. $\tilde{s}(n) = s(n) - \bar{s}(n)$ où $\bar{s}(n) = \frac{1}{L}\sum_{l=0}^{L-1} s(l)$ est la moyenne du signal sur l'ensemble de la séquence,
2. $\hat{s}(n) = \frac{\tilde{s}(n)}{\max_n |\tilde{s}(n)|}$.

5.3.3 Choix de l'unité d'analyse : description sur des durées temporelles variables

L'originalité de notre description du contenu émotionnel repose sur les deux aspects suivants. D'une part, nous considérons à la fois les portions voisées et *non voisées* du signal de parole (cf définitions 5 et 6). D'autre part, les descripteurs sont calculés sur des durées temporelles variables : fenêtres, trajectoires, ou segment[21].

1. Le segment est découpé en une succession de *fenêtres d'analyse* de 40 ms, se chevauchant entre elles sur 30ms (i.e. avec un recouvrement de 75%). La « force de voisement » de chaque fenêtre est calculée sous Praat à l'aide de la fonction d'autocorrélation et comparée à un seuil afin de déterminer si celle-ci est voisée ou non. Certains descripteurs comme la fréquence fondamentale ne sont pertinents que pour les fenêtres d'analyse déterminées

[21] unité d'annotation définie dans le chapitre 2, le segment est un tour de parole ou une portion de tour de parole avec un contenu émotionnel homogène.

comme voisées. Le tableau 5.1 répertorie les descripteurs calculés sur les fenêtres voisées et/ou non voisées.

2. Les fenêtres d'analyse adjacentes voisées ou non voisées sont regroupées en *trajectoire* respectivement voisée ou non voisée. Les dérivées et statistiques d'ordre 1 (moyenne, minimum, maximum, plage), 2 (écart-type), 3 (aplatissement), et 4 (asymétrie) de chaque descripteur sont calculées pour chaque trajectoire.
3. Le jitter, shimmer et taux de fenêtres non voisées sont calculés sur l'ensemble du *segment*.

Le tableau 5.1 liste l'ensemble des 534 descripteurs calculés et précise l'unité d'analyse utilisée pour chacun ainsi que la condition de voisement nécessaire à leur calcul.

Descripteurs	fen. V.	fen. NV.	traj. V.	traj. NV	segment
F_0, dF_0	x				
$statF_0$, $statdF_0$			x		
Int, $dInt$	x	x			
$statInt$, $statdInt$			x	x	
HNR,$dHNR$	x	x			
$statHNR$, $statdHNR$			x	x	
PAP,$dPAP$	x	x			
$statPAP$, $statdPAP$			x	x	
$Jitter$					x
$Shimmer$					x
Taux de fenêtres non voisées					x
$MFCCk$, $dMFCCk$ ($k \in [1:12]$)	x	x			
$statMFCCk$, $statdMFCCk$			x	x	
$EBBk$, $dEBBk$ ($k \in [1:12]$)	x	x			
$statEBBk$, $statdEBBk$			x	x	
C_s, dC_s	x	x			
$statC_s$, $statdC_s$			x	x	
F_1, dF_1	x	x			
$statF_1$, $statdF_1$			x	x	
Bw_1, dBw_1	x	x			
$statBw_1$, $statdBw_1$			x	x	
F_1, dF_1	x	x			
$statF_1$, $statdF_1$			x	x	
Bw_2, dBw_2	x	x			
$statBw_2$, $statdBw_2$			x	x	

TAB. 5.1 – *Liste des 534 descripteurs utilisés en fonction de la condition de voisement et unité d'analyse associée (d = dérivée, stat = moyenne, minimum, maximum, écart-type, plage, aplatissement, asymétrie, V= voisées, NV = non voisées)*

5.3.4 Choix de normalisation des descripteurs

La plage de valeurs que peut prendre un descripteur varie fortement d'un descripteur à un autre. Par exemple, si le taux de fenêtres non voisées varie entre 0 et 1, la fréquence fondamentale prend des valeurs de l'ordre de plusieurs centaines. Cette hétérogénité dans les valeurs peut avoir

des conséquences sur le comportement des algorithmes de réduction et d'apprentissage qui seront utilisés dans le chapitre suivant : les descripteurs ayant des valeurs élevées risquent d'avoir plus de poids que ceux atteignant des valeurs plus faibles.

Des techniques de normalisation peuvent ainsi être utilisées pour éviter ce biais. Nous utilisons ici la technique de normalisation dite normalisation min-max. Les valeurs de chaque descripteur sont ramenées dans l'intervalle $[-1; 1]$ en divisant chacune de ces valeurs par le maximum en valeur absolue atteint sur les données pour ce descripteur.

5.4 Évaluation de la pertinence des descripteurs acoustiques pour la modélisation des émotions de type peur

Dans le chapitre 4, nous avons expliqué que nous préférons nous focaliser pour le système de reconnaissance sur le cas plus stable d'une discrimination des émotions de type peur de l'état neutre. L'étude présentée ici intervient comme une étape préalable au développement du système de reconnaissance. La pertinence des descripteurs acoustiques est par conséquent évaluée pour la modélisation des émotions de type peur par rapport à l'état neutre.

Sous-corpus d'étude

Cette étude est effectuée sur un sous-ensemble du corpus SAFE contenant uniquement les segments annotés *peur* ou *neutre*[22].

La qualité de parole des *segments* du corpus SAFE étant très variable, nous avons également fait le choix de restreindre notre étude aux *segments* du corpus ayant une qualité d'enregistrement de la parole acceptable (i.e. annotés Q2 ou Q3[23]). Une qualité insuffisante dans l'enregistrement de la parole risque de biaiser l'estimation des descripteurs acoustiques du signal de parole, rendant l'analyse acoustique des émotions trop délicate pour pouvoir en tirer des conclusions intéressantes. En revanche, sont illustrés dans le sous-corpus, des environnements sonores réalisés dans des lieux très variés, en extérieur ou en intérieur.

Les recouvrements entre les locuteurs et les manifestations au niveau de la foule, bien qu'importants car susceptibles d'être fréquents pour une application de surveillance, restent cependant des cas particulièrement complexes à traiter. Il paraît en effet difficile de détecter une émotion dans un flot de parole pouvant présenter des manifestations émotionnelles divergentes superposées. Dans le cas de manifestations émotionnelles superposées convergentes, les descripteurs acoustiques qui pourraient être efficaces sur de telles données seront très spécifiques. Une telle étude nécessiterait un matériel plus adapté avec une plus grande proportion de *segments* illustrant des manifestations émotionnelles au niveau de la foule. Nous avons donc opté pour une modélisation des manifestations émotionnelles au niveau de l'individu. Les *segments* correspondant à des recouvrements entre locuteurs (« overlap ») ont été retirés du sous-corpus d'étude.

Le tableau 5.2 présente la quantité de données associée à chaque classe dans le sous-corpus d'étude en termes de segments, trajectoires et fenêtres. Nous appellerons par la suite ce sous-corpus d'étude *SAFE_1*. Le contenu du sous-corpus est divisé en deux parties, la première contenant les fenêtres voisées et la seconde les fenêtres non voisées. Les descripteurs acoustiques sont évalués de manière séparés sur chacun des deux contenus voisé et non voisé.

Les données des deux classes sont réparties de la manière suivante en terme de voisement :

– 33% des fenêtres associées à la classe *peur* sont voisées,

[22] en considérant l'intersection des annotations de Ann1 et Ann2 (cf chapitre 4)

[23] Rappel : la qualité de chaque segment a été annotée par une note allant de Q0 à Q3

Classe	nombre de segments	nombre de trajectoires	nombre de fenêtres	durée
Peur	381	2891	113 385	19 min
Neutre	613	5417	181 615	30 min
Total	994	8308	295 000	49 min

TAB. 5.2 – *Sous-corpus d'étude SAFE_ 1*

– 47% des fenêtres associées à la classe *neutre* sont voisées.

Fisher Discriminant Ratio

Nous utilisons pour cette étude le FDR (Fisher Discriminant Ratio). Il permet de comparer les capacités de chaque paramètre à distinguer les deux classes (*peur* et *neutre*) en mesurant le chevauchement de leurs fonctions de densité de probabilité. Le FDR est calculé ici pour chaque descripteur d_i sur les deux contenus voisé et non voisé séparément :

$$FDR_{d_i} = \frac{(\mu_{i,neutre} - \mu_{i,peur})^2}{\sigma^2_{i,neutre} + \sigma^2_{i,peur}}$$

où $\mu_{i,neutre}$ et $\mu_{i,peur}$ sont les moyennes des valeurs correspondant aux descripteurs d_i pour chacune des classes et $\sigma^2_{i,neutre}$ et $\sigma^2_{i,peur}$ les variances correspondantes.

5.4.1 Contenu voisé

Nous considérons ici uniquement le contenu voisé des segments. Pour chaque famille de descripteurs, nous présentons le descripteur qui est ressorti comme le plus pertinent et le FDR associé dans le tableau 5.3. Le classement complet des descripteurs pour chaque famille sera présenté dans le paragraphe 6.3.1.

Famille de descripteurs	descripteur le plus pertinent	FDR
Prosodiques	fréquence fondamentale (moyenne sur une trajectoire)	0,42
Qualité de voix	jitter sur un segment	0,11
Spectraux	centroïde spectral (moyenne sur une trajectoire)	0,10

TAB. 5.3 – *Descripteurs évalués comme les plus pertinents pour chaque famille sur le contenu voisé de SAFE_ 1 (tableau 5.2) (FDR calculé sur environ 123 000 fenêtres).*

Pour les familles prosodique et spectrale, les descripteurs sélectionnés comme les plus efficaces sont ceux correspondant à des statistiques calculées sur des trajectoires (la moyenne de la fréquence fondamentale et la moyenne du centroïde spectral). Pour les descripteurs de qualité de voix, c'est le jitter calculé au niveau du segment qui est le plus efficace. Il ressort de ce tableau que le descripteur associé à la fréquence fondamentale a un pouvoir de discrimination mesuré par le FDR nettement supérieur à celui du jitter et du descripteur associé au barycentre spectral sur le sous-corpus d'étude.

La figure 5.4 compare la répartition de chacun des trois descripteurs ci-dessous pour les deux classes *peur* et *neutre* grâce à des diagrammes dits de « boîte à moustaches » [Tukey, 1977]. Le diagramme de boîte consiste en un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. Les « moustaches » correspondent aux lignes pointillées noires et ont

une longueur qui vaut 1,5 fois l'écart interquartile[24]. Les croix rouges correspondent aux valeurs en dehors des « moustaches ».

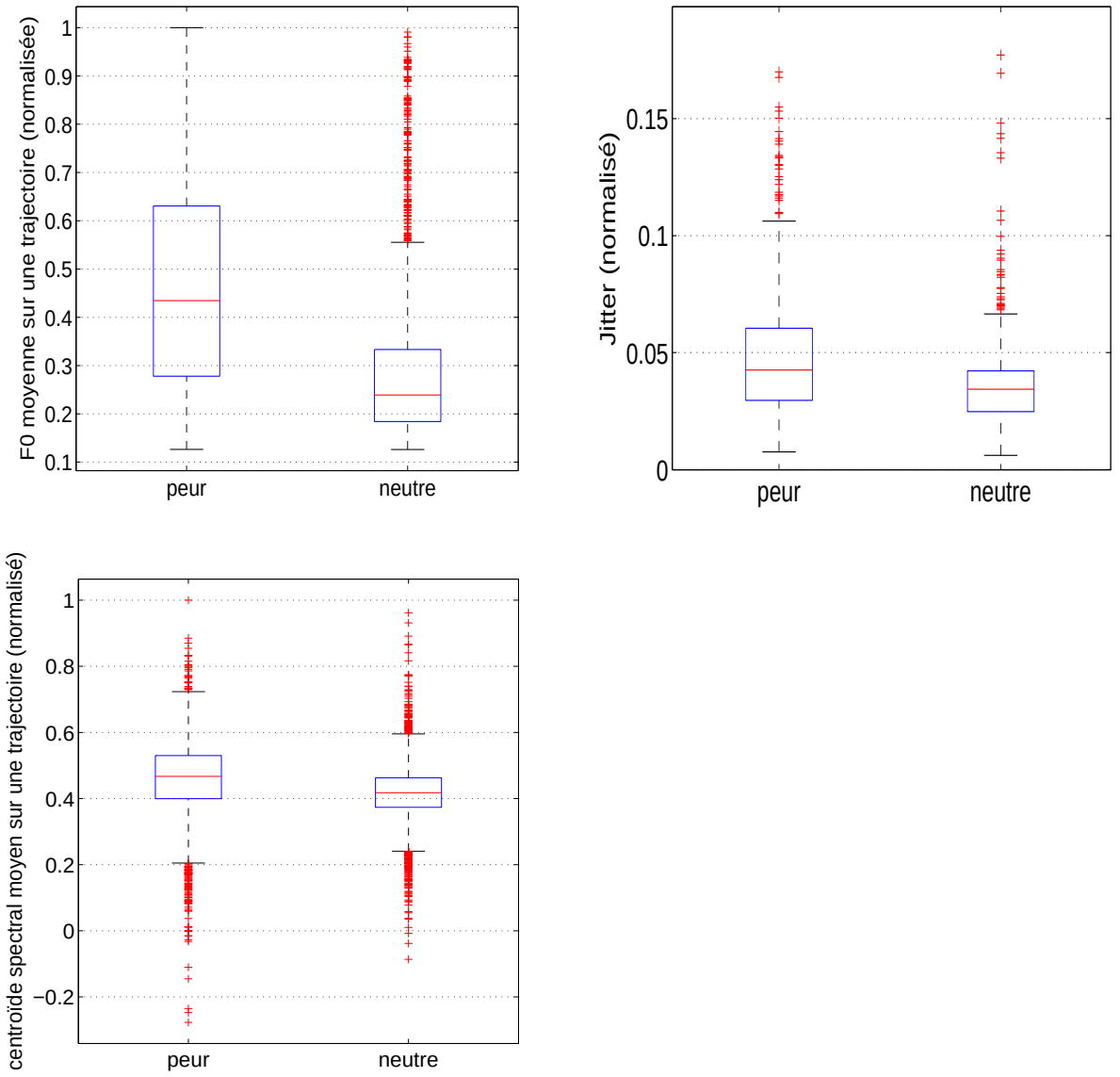

FIG. 5.4 – *Diagrammes de boîte à moustaches pour les trois descripteurs les plus efficaces sur le contenu voisé de SAFE_1 (tableau 5.2).*

La fréquence fondamentale a tendance à être plus élevée pour la classe *peur* que pour le *neutre*. Ce résultat rejoint les résultats obtenus par d'autres études sur les manifestations acoustiques de la peur [Williams et Stevens, 1972], [Devillers et Vasilescu, 2003], [Scherer *et al.*, 1998] et peut s'expliquer par les modifications physiologiques intervenant dans le cas de la *peur* (augmentation du rythme cardiaque etc.). De plus la plage de valeurs atteintes par la fréquence

[24]Ces diagrammes sont utilisés dans des secteurs où les données peuvent le plus souvent être modélisées en utilisant une loi normale ; dans ce cas, la théorie montre que les extrémités des « moustaches » sont voisines du premier et 99e centile : ces diagrammes sont surtout utilisés pour détecter la présence de données exceptionnelles.

fondamentale pour la classe *peur* est beaucoup plus grande que pour la classe *neutre* (cf longueur des « moustaches »).

Il a été également constaté que la peur se manifeste également par plus d'énergie dans les hautes fréquences [Scherer *et al.*, 1998], ce qui explique la pertinence du barycentre spectral qui atteint en effet des valeurs plus élevées dans le cas de la *peur*.

Le jitter atteint des valeurs plus extrêmes dans le cas de la *peur*. On peut s'attendre en effet à ce que dans le cas de *segments* contenant des cris, le jitter soit particulièrement élevé, la modulation fréquentielle étant très forte dans ce cas, comme on a pu le voir sur la figure 5.2.

5.4.2 Contenu non voisé

Nous considérons ici le contenu non voisé des segments. Pour chaque famille de descripteurs, nous présentons le descripteur qui est ressorti comme le plus pertinent et le FDR associé dans le tableau 5.4. Comme pour le contenu voisé, le classement complet des descripteurs pour chaque famille sera présenté dans le paragraphe 6.3.1.

Famille de descripteurs	descripteur le plus pertinent	FDR
Prosodiques	Int (plage des valeurs sur une trajectoire)	0,05
Qualité de voix	taux de fenêtres non voisées sur un segment	0,13
Spectraux	EBB6 (plage des valeurs sur une trajectoire)	0,13

TAB. 5.4 – *Descripteurs évalués comme les plus efficaces pour chaque famille sur le contenu non voisé de SAFE_1 (tableau 5.2) (FDR calculé sur environ 172 000 fenêtres).*

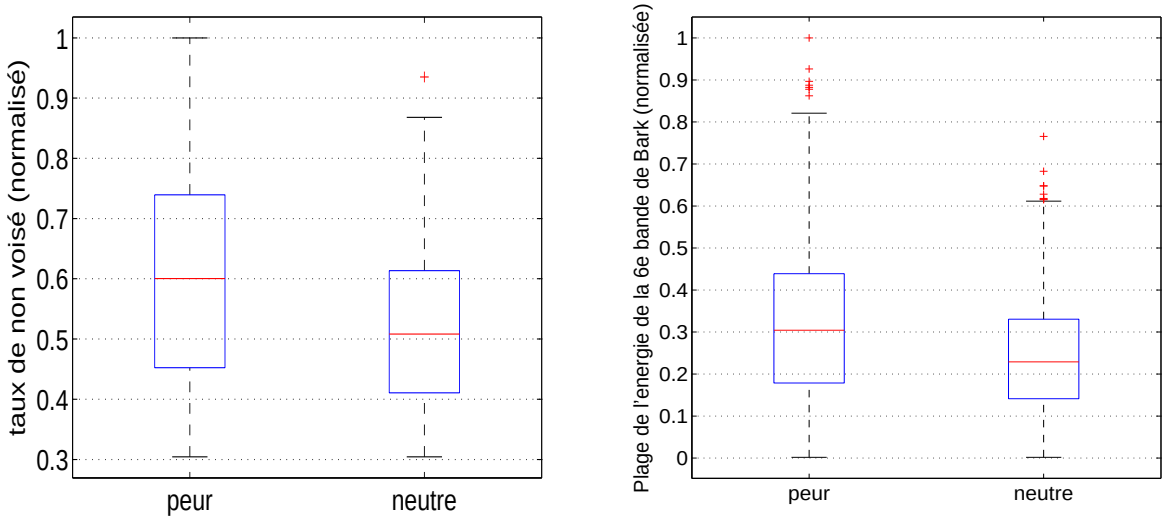

FIG. 5.5 – *Diagrammes de boîte à moustaches pour les deux descripteurs évalués comme les plus efficaces sur le contenu non voisé de SAFE_1 (tableau 5.2).*

Sur le contenu non voisé, les descripteurs prosodiques qui se réduisent à l'intensité et ses statistiques au niveau de la trajectoire apparaissent comme peu discriminants. L'intensité est par ailleurs un descripteur qui pose problème car ses valeurs sont altérées par de nombreux autres facteurs comme le bruit environnant, la distance du microphone à la bouche, malgré la

normalisation du signal effectuée et décrite dans le paragraphe précédent. Ce descripteur s'est également avéré peu discriminant pour le contenu voisé.

La figure 5.4.2 présente les diagrammes de « boîte à moustaches » pour le taux de fenêtres non voisées et pour la plage de valeurs sur les trajectoires de l'énergie de la sixième bande de Bark (930-1080 Hz).

Le taux de fenêtres non voisées est plus élevé pour la *peur*. On peut interprêter ce résultat par le fait que les *segments* de *peur* contiennent plus de respirations. Dans [Cahn, 1989], l'auteur énumère les corrélats physiologiques de la peur et cite parmi ceux-ci une augmentation du taux de respirations.

5.5 La fréquence fondamentale et les formants : la sensibilité au locuteur et au contenu linguistique

La plupart des descripteurs utilisés couramment pour caractériser les émotions sont dépendants d'autres variables telles que le bruit environnant, le locuteur, le contenu linguistique, etc. Aussi, la fréquence fondamentale est statistiquement plus élevée pour les femmes et les enfants que pour les hommes. Le volume des poumons, la position du larynx, par exemple, varient fortement avec le sexe, l'âge et l'individu. Notamment le larynx est plus élevé chez la femme ce qui aura des conséquences lors de la production vocalique (décalages formantiques).

Comme nous l'avons vu dans le chapitre 3, le corpus présente une forte variabilité en terme de locuteurs, d'environnements sonores. Il s'agit dans ce paragraphe de décrire le comportement des deux descripteurs acoustiques qui nous ont semblé les plus sensibles aux « variables-bruit », i.e. avec une forte dépendance à d'autres variables que l'émotion : les formants et leur dépendance au contenu linguistique et la fréquence fondamentale et la dépendance au locuteur.

5.5.1 Les formants et la sensibilité au contenu linguistique

Les premiers et deuxièmes formants varient en fonction de la voyelle prononcée et sur une même voyelle en fonction du locuteur. Nous avons pourtant considéré ces descripteurs pour l'analyse acoustique des manifestations émotionnelles sans effectuer de normalisation par phonème.

L'étude du comportement de ces descripteurs en fonction de l'émotion sur une même voyelle permet d'extraire les paramètres discriminants de cette émotion indépendamment des variations entre les voyelles. Cette étude s'avère plus délicate lorsqu'elle est menée en aveugle, i.e. sans connaissance a priori de la voyelle prononcée.

Dans un objectif d'apprentissage du comportement de ces descripteurs, si le nombre de données sur lesquelles l'apprentissage est réalisé est suffisamment grand, ces descripteurs ont cependant un sens.

Nous vérifions donc que le contenu phonétique est approximativement le même pour chaque classe, compte tenu de la quantité de données associée à chaque classe. Dans cet objectif, la transcription du contenu verbal fournie lors de l'annotation est soumise à un outil de conversion de graphèmes en phonèmes[25]. La répartition ainsi obtenue des principales voyelles de l'anglais (voir tableau 5.5) pour chaque classe est représentée sur la figure 5.6. La différence de comportement des formants entre les deux classes n'est donc pas due à une différence entre les contenus phonétiques des deux classes mais peut être due à des modifications dans l'articulation qui varient en fonction de la classe d'émotion considérée.

[25]http ://www.cs.cmu.edu/ lenzo/t2p

Phonème	Exemple
/aa/	odd
/iy/	eat
/uw/	two
/ae/	at
/er/	hurt
/ih/	it
/eh/	Ed
/ao/	ought
/ah/	hut
/uh/	hood

TAB. 5.5 – *La transcription phonétique associée aux principales voyelles de l'anglais*

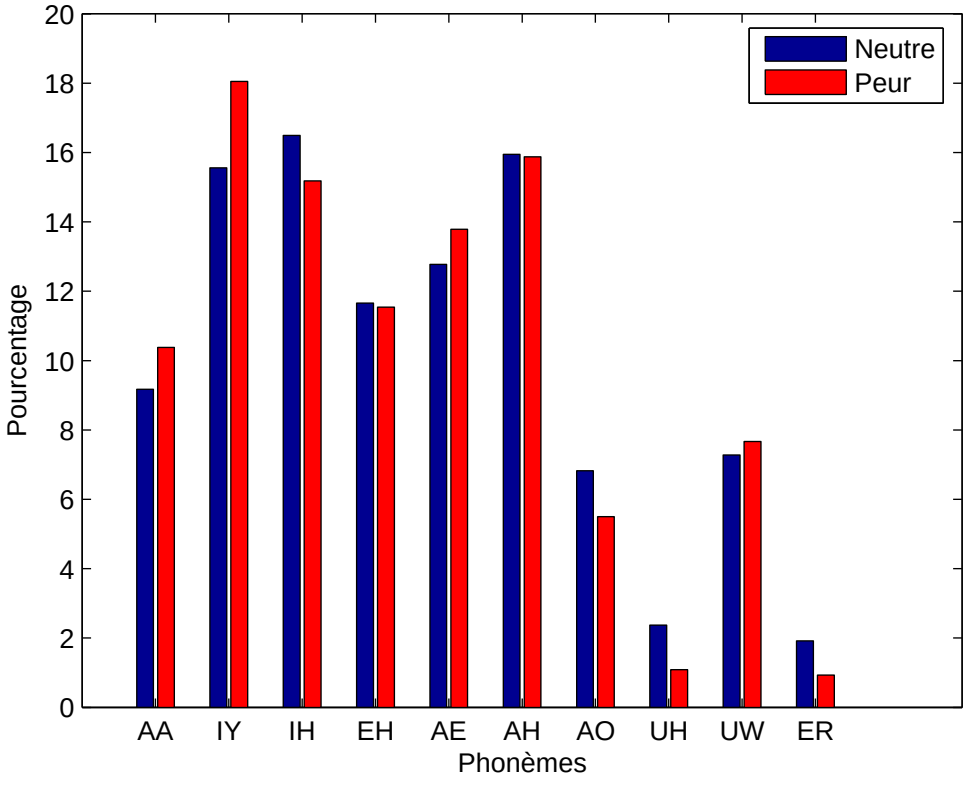

FIG. 5.6 – *Répartition des voyelles pour chaque classe émotionnelle sur SAFE_1 (tableau 5.2)*

5.5.2 La fréquence fondamentale et la sensibilité au locuteur

La fréquence fondamentale moyenne varie d'un locuteur à un autre et les différences peuvent être particulièrement importantes entre un homme et une femme. L'étude des variations de ce descripteur en fonction de l'émotion peut être fortement biaisée si notre étude porte sur plusieurs locuteurs, sans connaissance a priori du locuteur intervenant. De nombreuses études utilisent pour pallier ce biais une normalisation de ce descripteur par locuteur [Devillers et Vidrascu, 2006], [Wred et Shriberg, 2003]. Cette normalisation présuppose cependant une connaissance a priori du locuteur considéré. Or, dans le cas d'une application de surveillance l'identité du locuteur n'est pas connue a priori et le nombre de locuteurs à traiter peut être particulièrement élevé. C'est pourquoi nous n'avons pas effectué de normalisation par locuteur.

La répartition de chaque classe en fonction du sexe du locuteur est présentée dans le tableau 5.6 pour le contenu voisé[26]. Le sous-corpus étudié dans ce paragraphe et appelé *SAFE_ 2* correspond au sous-corpus *SAFE_ 1* auquel nous avons retiré les 47 segments correspondant à des locuteurs enfants. Le nombre de données correspondant à une femme pour la classe *peur* est deux fois plus élevé que celui correspondant à un homme. Pour la classe *neutre*, la tendance s'inverse. La pertinence évaluée dans le paragraphe précédent du descripteur $meanF_0$ (moyenne de la fréquence fondamentale sur une trajectoire) pourrait être due à ce déséquilibre.

	Femme		Homme	
	nbr. de *segments*	nbr. de fenêtres	nbr. de *segments*	nbr. de fenêtres
Neutre	182	26 972	418	55 969
Peur	221	24 882	126	11 004
Total	403	51 854	544	66 973

TAB. 5.6 – *Répartition des données de chaque classe du corpus SAFE_ 2 en fonction du sexe des locuteurs (contenu voisé uniquement)*

Nous avons donc recalculé le FDR obtenu par le descripteur $meanF_0$ pour chaque sexe séparément. Les résultats sont présentés dans le tableau 5.7.

	Femme	Homme
FDR de $meanF_0$	0,62	0,16

TAB. 5.7 – *FDR du descripteur $meanF_0$ pour chaque classe en fonction du sexe des locuteurs (contenu voisé uniquement) sur SAFE_ 2 (tableau 5.6)*

Le descripteur $meanF_0$ est évalué comme très efficace pour les femmes, son efficacité sur la totalité du sous-corpus d'étude n'est pas due uniquement à un déséquilibre dans la répartition homme/femme des deux classes. En revanche chez les hommes, ce descripteur s'avère moins discriminant avec un FDR moins élevé mais reste quand même classé deuxième parmi tous les descripteurs. Les manifestations de la *peur* chez les hommes sont en effet moins extrêmes que chez les femmes dans notre corpus.

5.6 Conclusion

Le cadre applicatif qu'est la surveillance dans les lieux publics fait émerger de nouveaux problèmes. Premièrement, les émotions cible, i.e. les émotions de type peur présentent des ma-

[26] La fréquence fondamentale est calculée sur le contenu voisé uniquement

nifestations extrêmes qui entraînent des modifications aussi bien sur les portions voisées et les portions non voisées du signal de parole. Ces manifestations font également émerger des phénomènes de coarticulations et des phénomènes non verbaux tels que les cris et les fortes respirations qui nous ont ammenés à ne pas faire reposer notre analyse acoustique sur une segmentation préalable du signal de parole (choix de l'unité de l'analyse, technique de normalisation indépendante du contenu phonétique).

Deuxièmement le nombre de locuteurs susceptibles d'intervenir pour une telle application est difficilement contrôlable. Les techniques de normalisation utilisées se doivent donc d'être indépendantes du locuteur. Or certains descripteurs tels que la fréquence fondamentale sont très sensibles au changement de locuteur. Cette contrainte augmente donc la difficulté de la tâche de description acoustique des manifestations émotionnelles.

Troisièmement les conditions d'enregistrement et le contexte sonore dans lequel émerge la parole (bruit, musique) dans le corpus SAFE sont des conditions qui sont loin d'être idéales pour une analyse acoustique. Le choix des descripteurs considérés est conditionné par cette particularité du corpus. Certains descripteurs comme le NAQ nécessitent en effet de très bonnes conditions avec une bonne proximité du micro avec la bouche. Les enregistrements susceptibles d'intervenir pour notre application présenteront également des caractéristiques différentes de celles obtenues lors d'enregistrement en laboratoire qu'il sera important de prendre en compte.

Sur les portions voisées, le descripteur qui émerge nettement comme le plus efficace pour un problème de discrimination peur/neutre est la moyenne de la fréquence fondamentale sur une trajectoire voisée. La trajectoire semble donc fournir une unité d'analyse pertinente pour la modélisation des manifestations émotionnelles qui a de plus l'avantage de ne pas reposer sur une segmentation a priori du signal de parole en unités phonétiques ou en segments d'annotation.

Sur les portions non voisées, le descripteur le plus efficace est le taux de fenêtres non voisées sur un segment.

Les descripteurs acoustiques qui sont utilisés pour décrire le contenu émotionnel du corpus SAFE ont été sélectionnés dans l'objectif de développer un système de reconnaissance des émotions de type peur. C'est en fonction de ces contraintes que nous avons choisi les descripteurs acoustiques présentés dans ce chapitre et que nous avons fixé les paramètres de calcul de ces descripteurs (choix de l'unité d'analyse, normalisation, etc.). Nous verrons dans le chapitre suivant comment les plus efficaces de ces descripteurs seront utilisés par notre système de reconnaissance des émotions.

Chapitre 6

Reconnaissance des émotions pour l'analyse et la détection de situations anormales

Sommaire

Introduction

Depuis quelques années, les études sur la parole émotionnelle vont au delà d'une analyse des manifestations vocales des différents états émotionnels et commencent à développer des systèmes de classification automatique des émotions. Cette évolution est née de la prise de conscience des applications industrielles potentielles du domaine des sciences affectives avec l'apparition d'un nouveau champ de recherche, le domaine de l'« affective computing » [Picard, 1997].

Les systèmes de classification sont basés sur des méthodes dites d'apprentissage, dont nous présentons un bref tour d'horizon dans le paragraphe 6.1. L'objectif de ce premier paragraphe est également de faire un bilan des systèmes de classification automatique existants.

Nous décrivons ensuite le système de reconnaissance automatique des états émotionnels liés à la peur qui a été développé sur un sous-corpus d'étude (paragraphe 6.2). Les données utilisées par ce système correspondent à une représentation du signal de parole par les descripteurs acoustiques décrits dans le chapitre précédent. Les paramètres du système sont fixés dans le paragraphe 6.3. Cette étape a pour vocation de valider le choix des descripteurs considérés en optimisant les performances du système sur le corpus de développement.

Dans le paragraphe 6.4, nous proposons une analyse du comportement du système sur le corpus SAFE. Cette analyse est cruciale : elle fournit les éléments permettant d'appréhender l'adaptation du système à d'autres données. Le paragraphe 6.5 propose et évalue de nouvelles stratégies de classification.

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2006c], [Clavel *et al.*, 2006d], [Clavel *et al.*, 2007]

6.1 Etat de l'art en reconnaissance des émotions dans la parole

Les systèmes de classification automatique des émotions reposent sur les méthodes dites d'apprentissage, car elles sont basées sur une procédure d'apprentissage capable à partir d'une quantité de données suffisante, de caractériser les propriétés acoustiques de chaque classe d'émotion.

On distingue deux types de classification : supervisée et non supervisée. Lors d'une classification supervisée la classe de chaque objet (représentée par son étiquette) est fournie au programme d'apprentissage en même temps que les données. Lors d'une classification non supervisée, les classes sont déterminées automatiquement en fonction de la structure des données. Les systèmes de classification automatique d'émotions utilisent essentiellement des méthodes supervisées où les classes considérées sont des classes d'émotions souvent déterminées en fonction de l'application visée. Il existe de nombreuses techniques de classification qui sont détaillées dans [Duda et Hart, 1973].

L'utilisation de méthodes de classification dans le domaine des émotions est un sujet de recherche émergeant. À l'heure actuelle, il est difficile de comparer les performances obtenues par les systèmes déployés par les différents laboratoires de recherche. Aucune campagne d'évaluation comme celles menées dans les domaines de la reconnaissance de la parole, de l'identification du locuteur et de l'identification de langues n'a été menée jusqu'à présent.

Un premier pas cependant a été franchi avec une initiative de coopération appelée CEICES (Combining Efforts for Improving automatic Classification of Emotional user States) lancée en 2005 par le FAU Erlangen en Allemagne qui rassemble les partenaires du réseau d'excellence

HUMAINE[27].

Cette initiative repose sur le principe suivant : une base de données annotée (fichiers audio, dictionnaire phonétique, segmentations manuelles en mots et labels émotionnels) est fournie aux différents partenaires. Les résultats sont présentés dans l'article [Batliner *et al.*, 2006], dont l'objectif est avant tout d'augmenter les taux de reconnaissance en rassemblant les descripteurs acoustiques des différents partenaires et en combinant les classificateurs utilisés. Les auteurs soulignent qu'il est difficile de comparer les méthodes de classification et l'efficacité des ensembles de descripteurs utilisés séparément, compte tenu de la diversité des procédures de normalisation et de transformation des descripteurs.

Par ailleurs, les performances des différents systèmes ne dépendent pas seulement des méthodes de classification et des descripteurs acoustiques utilisés, mais aussi :

- de la base de données sur laquelle les algorithmes ont été testés : est ce que ce sont des bases de données illustrant des émotions actées ou spontanées ? Quelle est la qualité d'enregistrement ? Quel est le degré de diversité en terme de locuteurs et de contextes illustrés ? ;
- du nombre de classes émotionnelles considérées, une classification en 10 classes étant plus ambitieuse qu'une classification en deux classes émotionnelles ;
- des classes émotionnelles considérées, une discrimination peur/colère étant en général plus subtile qu'une discrimination peur/neutre (voir paragraphe 4.4.1) ;
- des conditions d'apprentissage : le locuteur de l'échantillon testé est-il présent dans la base de données utilisées pour l'apprentissage ? ;
- des techniques d'extraction des descripteurs : l'extraction des différents descripteurs reposent-elles sur des a priori sur le locuteur ou sur le contenu linguistique[28] (technique de normalisation, unité d'analyse) ;

Il est souvent difficile de rassembler de telles informations dans la description des différents systèmes. Cependant, certains articles mettent l'accent sur l'étude d'un des facteurs de dépendance présentés ci-dessus.

6.1.1 Conditions d'apprentissage et performances

L'article [Schuller *et al.*, 2004] compare les deux conditions d'apprentissage (dépendant vs. indépendant du locuteur) sur une même base de données avec les mêmes descripteurs et la même méthode de classification. Les performances chutent de 89% ou 93% à 76% ou 75% selon la méthode utilisée lorsque le système devient indépendant du locuteur (voir tableau 6.1).

Dans [Lee *et al.*, 2002], l'ensemble d'apprentissage est tiré aléatoirement, le locuteur de l'échantillon testé peut se trouver dans l'ensemble d'apprentissage. Cependant, la base de données étudiée contient un grand nombre de locuteurs (1200), limitant ainsi la dépendance du système au locuteur.

6.1.2 Algorithme d'apprentissage et performances

Dans l'article [Schuller *et al.*, 2004], on trouve également une confrontation des deux méthodes d'apprentissage, l'une appartenant à la classe des algorithmes génératifs, les GMM (*Gaussian Mixture Models*), l'autre reposant sur une approche discriminative les SVM (*Support Vector Machine*). Les deux méthodes obtiennent des performances similaires, 75% pour les GMM et 76% pour les SVM.

[27] http ://www5.informatik.uni-erlangen.de/Forschung/Projekte/HUMAINE/ ?language=en
[28] Voir chapitre 4

Article	Méthode	loc.	BD	Classes	Dep. loc.	Performances
[Schuller *et al.*, 2004]	SVM	13	actée	7	non	76%
[Schuller *et al.*, 2004]	SVM	13	actée	7	oui	93%
[Schuller *et al.*, 2004]	GMM	13	actée	7	non	75%
[Schuller *et al.*, 2004]	GMM	13	actée	7	oui	89%
[Lee *et al.*, 2002]	SVM	1200	réelle	2	oui	74%
[Lee *et al.*, 2002]	AD	1200	réelle	2	oui	75%
[Lee *et al.*, 2002]	kppv	1200	réelle	2	oui	74%

TAB. 6.1 – *Performances selon la méthode de classification utilisée, selon le type de base de données (avec des émotions actées vs. vécues), le nombre de locuteurs différents considérés (loc.), le nombre de classes, des conditions d'apprentissage (dépendance au locuteur ou non).*

Sur la même base de données, en 2003, les auteurs dans [Schuller *et al.*, 2003] avaient comparé la classification par GMM avec la méthode des HMM continus. La première méthode utilise des statistiques globales pour les caractéristiques dérivées de l'échelle des fréquences et le contour d'énergie du signal de parole. La deuxième méthode introduit une complexité temporelle en appliquant les modèles de Markov continu, en considérant des caractéristiques instantanées de plus haut niveau au lieu de statistiques globales. Dans ce cadre, les performances avaient été obtenues sans évaluation inter-locuteur et se sont avérées être finalement meilleures avec les GMM.

Dans [Lee *et al.*, 2002], trois méthodes différentes ont été testées : l'analyse discriminante, les machines à vecteur support et les k plus proches voisins. Les performances obtenues par les trois méthodes de classification sont relativement équivalentes.

Une autre étude [Dellaert *et al.*, 1996] au contraire obtient des performances plus contrastées en fonction de la méthode de classification utilisée avec une confrontation des k plus proches voisins (kppv) et de l'analyse discriminante linéaire (AD). Les meilleures performances sont obtenues cette fois avec les k plus proches voisins. Cette méthode a cependant été évaluée dans [Petrushin, 2003] qui compare réseau de neurones et k plus proches voisins et obtient des performances nettement meilleures avec les réseaux de neurones (70% au lieu de 55%). Le système de classification utilise en fait plusieurs réseaux de neurones sur différents sous-ensembles d'apprentissage (bootstrap).

À noter que d'autres techniques de classification non présentées dans le tableau telles que les arbres de décision [Yacoub *et al.*, 2003] [Maeireizo et Litman, 2004] [Oudeyer, 2003] ont également été développées et ont donné des résultats acceptables. Dans [Vidrascu et Devillers, 2005], des expériences ont été réalisées sur des données réelles (centres d'appel) avec différents algorithmes, pricipalement des SVM et des arbres de décision. Les résultats n'ont pas montré de différences significatives entre les performances obtenues par les algorithmes.

Il est donc difficile de dégager de ces résultats la supériorité d'une méthode de classification par rapport à une autre, les résultats semblant fortement dépendre de la base de données utilisée. Cependant, nous retiendrons pour la suite que les GMM, qui ont été choisis pour notre système de classification, obtiennent des résultats acceptables et souvent comparables à d'autres méthodes de classification [Batliner *et al.*, 2004].

6.1.3 Émotions simulées vs. vécues, nombre de classes et performances

Les émotions traitées dans les deux articles [Schuller *et al.*, 2004] et [Schuller *et al.*, 2003] sont au nombre de 7 : colère, peur, joie, dégoût, surprise, tristesse, neutre. Ce sont les « Big-six »

auxquelles s'ajoute la classe neutre. Les performances sont très bonnes pour un problème de classification à sept classes. A noter cependant que ces résultats sont obtenus sur des données actées enregistrées en laboratoire.

Dans [Lee *et al.*, 2002], les données étudiées sont des données réelles et correspondent à un corpus de dialogues homme-machine pour une application commerciale. On peut voir que les taux de reconnaissance sont inférieurs à ceux obtenus dans [Schuller *et al.*, 2004], sachant que ce système considère un nombre de classes émotionnelles très inférieur (deux classes d'émotions : émotions négatives vs. autres émotions contre 7 dans le système précédent). La chute des performances due au passage à des données réelles a été également constaté dans [Batliner *et al.*, 2000].

6.1.4 Techniques de normalisation et performances

La dépendance des performances au type de normalisation utilisée n'a pas été évaluée à notre connaissance. Nous avons vu dans le chapitre précédent qu'une normalisation par locuteur permettait de réduire les différences de comportement des descripteurs acoustiques entre locuteurs mais suppose cependant une connaissance a priori du locuteur de l'échantillon testé. Ce qui peut être le cas pour certaines applications [Oudeyer, 2003], mais pas pour notre application de surveillance dans les lieux publics.

6.2 Système de classification – synopsis

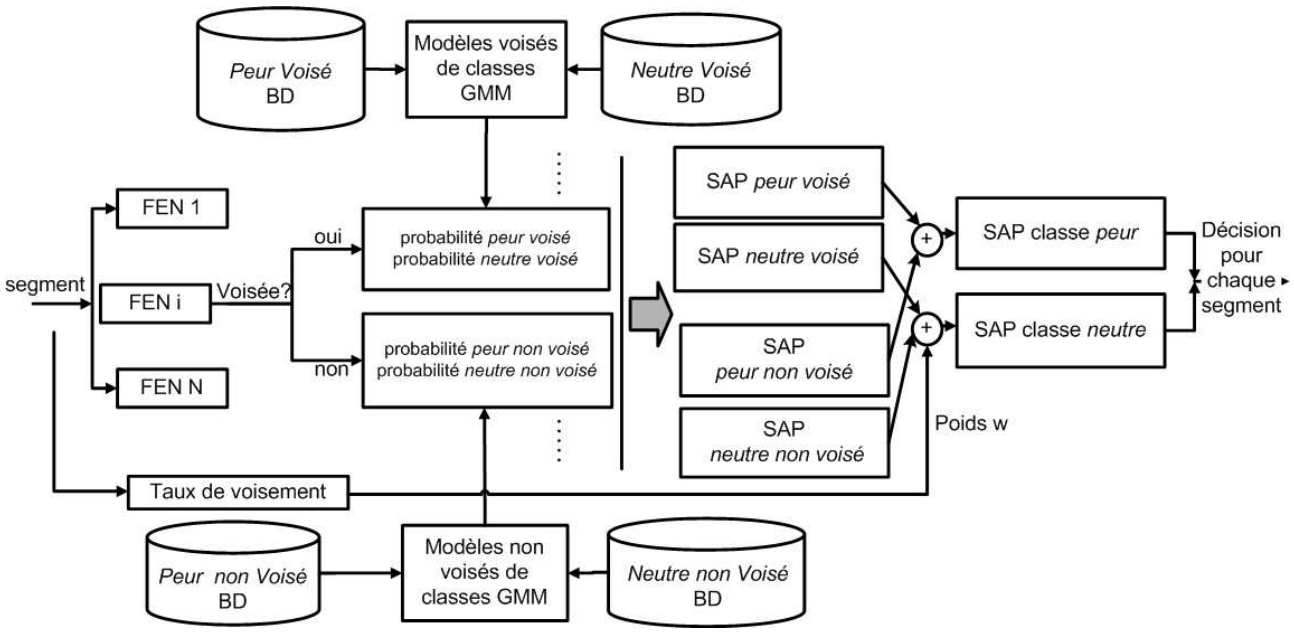

FIG. 6.1 – *Schéma de fusion des deux classificateurs voisé et non voisé lors de la décision*

Comme expliqué dans le chapitre précédent, nous avons choisi de nous focaliser dans un premier temps sur le problème de la discrimination peur/neutre. Le système de reconnaissance de la peur implémenté se base sur la classification peur/neutre des segments prédéfinis lors de l'annotation. Nous rappelons que le segment est l'unité d'annotation choisie et définie dans le

second chapitre. Un segment correspond à un tour de locuteur ou une portion de tour de locuteur avec un contenu émotionnel homogène.

Le schéma décrit dans le résumé de cette partie (figure 4.21) présente le fonctionnement du système qui peut être divisé en trois étapes : la réduction de l'espace des descripteurs, la modélisation de chacune des classes et la classification des segments qui repose sur les modèles précédemment construits. Chacune des étapes est détaillée dans les trois paragraphes suivants.

La méthode implémentée, les GMM, repose sur une modélisation des classes par un mélange de gaussiennes. Le système à base de GMM consiste en la fusion de deux classificateurs voisé et non voisé (cf figure 6.1), exploitant les contenus respectivement voisés et non voisés[29] du segment.

6.2.1 Réduction de l'espace de représentation des données

Le nombre des descripteurs utilisés et présentés dans le chapitre précédent est relativement élevé (534 descripteurs).

En théorie, augmenter le nombre de descripteurs pourrait permettre d'améliorer les performances du système. Cependant, en pratique, l'utilisation d'un trop grand nombre de descripteurs, au delà du problème de complexité engendré par une dimension élevée de l'espace de représentation des données, peut en fait aboutir à une baisse des performances [Duda et Hart, 1973]. Ce phénomène est appelé la *malédiction de la dimensionalité*. Cette étape de réduction de la dimension de l'espace de représentation des données préalable aux étapes d'apprentissage et de décision du système de classification est indispensable.

Bilan des méthodes existantes

Pour réduire l'espace des descripteurs, deux options se présentent :
- la sélection du sous-ensemble des descripteurs le plus discriminant (ex : algorithme de sélection de Fisher, algorithme génétique) ;
- la projection de l'espace de représentation des données par projection sur un espace de dimension plus petit (ex : analyse en composantes principales, analyse discriminante).

Afin d'avoir une meilleure visibilité des données traitées par le système de reconnaissance des émotions, nous avons choisi d'utiliser la première option, i.e. une méthode de sélection des descripteurs les plus pertinents. Cette option présente également l'avantage de pouvoir directement extraire les descripteurs les plus pertinents à l'étape de test alors que les méthodes de projection nécessitent le calcul préalable de l'ensemble des descripteurs de l'échantillon testé.

Nous avons choisi l'algorithme de sélection de Fisher, pour sa simplicité et son efficacité dans le domaine du traitement audio [Essid, 2005].

Algorithme de sélection de Fisher

Cette méthode est dérivée de l'analyse discriminante de Fisher dont on peut trouver une description dans [Duda et Hart, 1973]. Cependant, la maximisation du rapport de la dispersion inter-classe et la dispersion intra-classe se fait pour chaque descripteur séparément avec le FDR (Fisher Discriminant Ratio) présenté dans le paragraphe 5.4.

Nous utilisons donc ici l'algorithme de sélection fourni par la toolbox Spider[30] et qui se décompose pour une sélection multi-classe en deux étapes :

[29] cf chapitre 5 pour la définition des deux contenus.
[30] http ://www.kyb.mpg.de/bs/people/spider/

1. Pour chaque descripteur i, et pour chaque classe k, un score intermédiaire est estimé à partir des données de chacune des classes selon la formule suivante :

$$s_i^k = \sum_{l=1}^{K} FDR_{l,k}(i)$$

où K est le nombre de classes et $FDR_{l,k}$ est le critère de Fisher calculé pour les deux classes l et k :

$$FDR_{l,k}(i) = \frac{(\mu_{i,l} - \mu_{i,k})^2}{\sigma_{i,l}^2 + \sigma_{i,k}^2}$$

où $\mu_{i,l}$ et $\mu_{i,k}$ sont les moyennes du descripteur i sur les données associées aux classes l et k et $\sigma_{i,l}^2$ and $\sigma_{i,k}^2$ les variances.

2. Les scores $s_{i\ 1\leq i\leq I;1\leq k\leq K}^k$ sont ensuite triés par ordre décroissant et les N premiers descripteurs distincts associés aux scores les plus élevés sont ainsi sélectionnés.

Cet algorithme permet de sélectionner les descripteurs les plus discriminants sans cependant prendre en compte d'éventuelles relations qui les lient. Afin d'éviter que l'ensemble final des descripteurs sélectionnés ne présente de trop fortes redondances, l'algorithme de Fisher est utilisé en deux étapes :

1. Une première sélection est effectuée pour chaque famille de descripteurs (prosodiques, qualité de voix et spectraux) séparément. $1/5^e$ des descripteurs est ainsi sélectionné pour chaque famille, formant un premier ensemble composé d'une centaine de descripteurs.
2. La seconde sélection est effectuée en appliquant une nouvelle fois l'algorithme de Fisher sur l'ensemble des descripteurs sélectionnés à l'étape précédente. 40 descripteurs issus des trois familles sont ainsi sélectionnés.

Il existe cependant des méthodes plus sophistiquées de sélection de descripteurs qui permettent d'évaluer un sous-ensemble de descripteurs par rapport à un autre en éliminant les descripteurs redondants. C'est le cas notamment de l'algorithme IRMFSP (*Inertia Ratio Maximization using Feature Space Projection*) ou des algorithmes génétiques qui proposent une procédure systématique de parcours des sous-ensembles de descripteurs possibles.

6.2.2 Modélisation par Mélange de Gaussiennes (GMM-Gaussian Mixture Models)

Le modèle de mélange de gaussiennes est couramment utilisé dans le domaine de la reconnaissance de la parole. Depuis une dizaine d'années, ce modèle est aussi devenu l'approche dominante pour les systèmes de vérification du locuteur ([Sanchez-Soto, 2005], [Fredouille *et al.*, 2001], [Barras et Gauvain, 2003]). Il a été également utilisé avec succès pour la reconnaissance des émotions comme nous l'avons vu dans l'état de l'art au début de ce chapitre.

Les GMM consistent en la modélisation, pour chaque classe C_q, des données x_d sous la forme d'une somme pondérée par les coefficients $w_{m,q}$ de fonctions de densité de probabilité gaussiennes $p_{m,q}(x)$.

$$p(x/C_q) = \sum_{m=1}^{M} w_{m,q} p_{m,q}(x)$$

avec $\sum_{m=1}^{M} w_{m,q} = 1$ pour chacune des classes q considérées et où M est le nombre de composantes de densité considéré pour le modèle. Chaque composante s'exprime en fonction de sa moyenne $\mu_{m,q}$ et de sa matrice de covariance $\Sigma_{m,q}$:

$$p_{m,q} = \frac{1}{(2\pi)^{1/2} \left|\Sigma_{m,q}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_{m,q})^T(\Sigma_{m,q})^{-1}(x-\mu_{m,q})\right]$$

La matrice de covariance utilisée est diagonale, i.e. les modèles sont appris en considérant les observations associées à chacun des descripteurs de manière indépendante.

Pour chaque classe, chacune des composantes du mélange modélise une région différente de l'espace des données appelée aussi *cluster*.

L'apprentissage consiste en l'estimation à partir des observations d'une même classe des paramètres des gaussiennes qui composent le modèle de cette classe. Pour chaque classe C_q, les paramètres à estimer sont :

- les poids $(w_{m,q})_{m=1,..M}$ associés à chacune des M composantes du mélange,
- les moyennes et matrices de covariance de chacune des composantes du mélange :

$$(\mu_{m,q}, \Sigma_{m,q})_{m=1,..M}$$

Les paramètres sont initialisés par l'algorithme des k-moyennes permettant d'obtenir des valeurs approximatives des paramètres des gaussiennes de la classe. L'estimation des paramètres est réalisée à l'aide de l'algorithme appelé « Expectation Maximisation » (E-M) [Dempster *et al.*, 1977]. Le nombre d'itérations de l'algorithme E-M effectué pour l'apprentissage est de 10.

Ici, pour chaque classe et chaque condition de voisement, un modèle par mélange de gaussiennes est appris (cf figure 6.1) : PeurV, PeurNV, NeutreV, NeutreNV (où V correspond à « Voisé » et NV à « Non Voisé »).

6.2.3 Décision

La classification est réalisée à partir d'une règle de décision basée sur le maximum a posteriori. Pour chaque classificateur (voisé ou non voisé), un score *a posteriori* (SAP) est associé à chaque classe peur ou neutre pour chaque segment.

Calcul du score *a posteriori* pour chaque classificateur

Ce score correspond à la moyenne des log-probabilité *a posteriori*, calculée en multipliant les probabilités obtenues sur chaque fenêtre d'analyse. Ainsi, le score obtenu pour le classificateur voisé est :

$$SAP_V(C_q) = \frac{\sum_{n=1}^{N_{f_V}} \log p(C_q/x_n)}{N_{f_V}}$$

où x_n est le vecteur d'observation correspondant à la n^e fenêtre d'analyse voisée du segment et $p(C_q/x_n)$ la probabilité *a posteriori* correspondant à cette fenêtre. Celle-ci s'exprime d'après la formule de Bayes sous la forme suivante :

$$p(C_q/x) = \frac{p(C_q)p(x/C_q)}{p(x)}$$

Nous avons considéré les classes C_q équiprobables, donc $p(C_q) = \frac{1}{q}$. Cependant la probalilité de chaque classe peut être règlée si on dispose d'*a piori* sur les classes considérées.

Le score *a posteriori* est destiné à être maximisé, celui-ci est donc modifié de la manière suivante :

$$S\tilde{A}P_V(C_q) = \frac{\sum_{n=1}^{N_{f_V}} \log(p(x_n/C_q))}{N_{f_V}}$$

Un score *a posteriori* est calculé de la même manière pour le classificateur non voisé sur les fenêtres non voisées du segment.

Fusion des scores des deux classificateurs

En fonction de la proportion r de fenêtres voisées contenue dans le segment ($r \in [0;1]$), un poids ($w = 1 - r^\alpha$) est attribué à chacun des deux scores de manière à obtenir le score *a posteriori* final correspondant au segment :

$$S\tilde{A}P_{final}(C_q) = (1-w) * S\tilde{A}P_V(C_q) + w * S\tilde{A}P_{NV}(C_q)$$

Le paramètre α permet de régler la vitesse de décroissance du poids lorsque le taux de voisement augmente (cf figure 6.2). Le segment est ainsi attribué à la classe C_{q_0} pour laquelle le score *a posteriori* $SAP(C_q)$ est maximal :

$$q_0 = \arg \max_{1 \in [1:q]} S\tilde{A}P_{final}(C_q)$$

Il est également possible d'introduire un seuil de décision. La différence des scores de probabilité

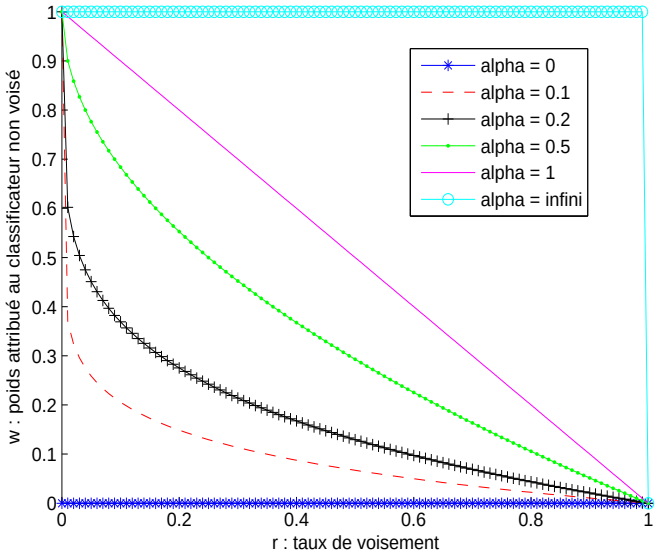

FIG. 6.2 – *Décroissance du poids $w = 1 - r^\alpha$ attribué au classificateur non voisé en fonction du taux de voisement r pour différentes valeurs du paramètre α.*

pour les deux classes peur et neutre est comparée à un seuil de décision, afin de déterminer la classe à laquelle le segment appartient.

Le principe de fusion du classificateur voisé et du classificateur non voisé est résumé dans la figure 6.1.

6.2.4 Protocole d'évaluation

Sous-corpus d'étude

Le sous-corpus utilisé ici est le même que celui sur lequel a été effectué l'analyse acoustique du chapitre précédent[31]. Une description détaillée du contenu de ce sous-corpus en termes de quantité de données a été présentée dans ce même chapitre.

Protocole de test : *Leave One Film Out*

Lors de l'état de l'art nous avons vu que le protocole de test avait une forte influence sur les performances, en fonction par exemple de la présence ou non du locuteur testé dans la base utilisée pour la modélisation des données par mélange de gaussiennes. Nous avons choisi ici le protocole de test *Leave One Film Out* qui assure que, pour chaque segment testé, l'ensemble d'apprentissage utilisé ne contient aucun segment du film duquel il est extrait. Autrement dit, le segment testé est issu d'une source contextuelle nouvelle (i.e d'un film nouveau), non représentée dans les données ayant servies à apprendre les modèles des différentes classes. Pour chaque film, un contexte spécifique est en effet défini en terme de profil de locuteur, de situations illustrées et de prise de sons. Pour cela, le sous-corpus d'étude est divisé en 30 sous-ensembles, correspondant aux 30 tirages possibles de 29 films parmi les 30 films du sous-corpus. 30 apprentissages sont ainsi réalisés à partir de chacun de ces sous-ensembles. Ce protocole est utilisé pour la modélisation par mélange de gaussiennes des données représentées sous la forme des descripteurs sélectionnés comme les plus pertinents (voir paragraphe 6.3.1).

Évaluation des résultats

Les performances du système obtenues pour les deux classificateurs seront présentés par la suite comme dans le tableau 6.2 qui représente la matrice de confusion entre les annotations des annotateurs (« manuel ») et les décisions du système (« automatique »). L'antidiagonale de la matrice de confusion correspond aux taux de faux rejets (FR) et de fausses détections (FD) de la classe peur (tableau 6.2).

automatique / manuel	Neutre	Peur
Neutre	taux de reconnaissance de la classe neutre	taux de fausses détections (FD)
Peur	taux de faux rejet (FR)	taux de reconnaissance de la classe peur

TAB. 6.2 – *Matrice de confusion pour l'évaluation de la classification peur/neutre*

La représentativité de chaque classe dans le sous-corpus étudié n'est pas équilibrée. Le taux d'erreur qui correspond au ratio entre le nombre de segments mal classés et le nombre total de segments est une mesure globale qui ne prend pas en compte ce déséquilibre. Cette mesure n'est donc pas utilisée ici. Nous calculons plutôt à partir de la matrice de confusion un taux d'erreur moyen par classe (TEM) qui s'exprime de la manière suivante :

$$TEM = \frac{FR + FD}{2}$$

[31] segments annotés peur ou neutre par Ann1 et Ann2, ayant été évalués comme ayant une qualité de parole de 2 ou 3 sur 3, sans recouvrement entre locuteurs

Pour le système à base de GMM, la balance entre le taux de fausses détections et le taux de faux rejets (i.e. le taux de fausse détection de la classe neutre) va dépendre du seuil de décision choisi. Il existe un seuil pour lequel la valeur prise par le taux de fausses détections est égale au taux de faux rejets (les taux de faux rejets et de fausses détections présentés dans la matrice de confusion sont obtenus avec une valeur de ce seuil de 0). Cette valeur correspond au taux d'égale erreur (TEE).

Nous proposons également une mesure qui prenne en compte le hasard. Le kappa présenté dans le chapitre 4 est calculé ici et correspond au taux d'accord entre les décisions du système et les classes attribuées par les annotateurs, corrigé de ce qu'il serait sous le simple effet du hasard.

Pour chaque taux d'erreur err, le rayon r de l'intervalle de confiance [Bengio et Mariéthoz., 2004] à 95%[32] $I = [err - r; err + r]$ est calculé selon la formule suivante :

$$r = 1,96\sqrt{\frac{err(1-err)}{N_{seg}}}$$

où N_{seg} est le nombre de segments sur lequel le taux d'erreur est calculé.

6.3 Réglage des paramètres du système et résultats

Le système de classification peur/neutre a été évalué selon le protocole d'évaluation décrit dans le paragraphe précédent. Afin de valider les résultats obtenus avec les GMM, une autre méthode, les SVM a été testée et est présentée dans l'annexe C.

6.3.1 Les descripteurs sélectionnés

Les tableaux 6.3 et 6.4 listent pour chaque famille de descripteurs l'ensemble des 40 descripteurs sélectionnés par l'algorithme de Fisher décrit dans la partie précédente pour la discrimination peur/neutre pour les contenus voisés et les contenus non voisés. Cette sélection est réalisée sur l'ensemble des données du sous-corpus.

Pour le contenu voisé, nous avions vu en comparant le critère de Fisher (FDR) dans le chapitre précédent que les descripteurs prosodiques étaient de loin les plus efficaces. Ceci se confirme ici avec la totalité des descripteurs prosodiques obtenus à l'issue de la seconde sélection. Ceux-ci sont essentiellement liés à la fréquence fondamentale, les descripteurs liés à l'intensité ne sont pas sélectionnés.

En ce qui concerne les descripteurs de qualité de voix, le jitter et le shimmer ont tous deux été sélectionnés. En revanche, le descripteur du niveau de souffle dans la voix (HNR ou PAP) n'a pas été sélectionné.

Les descripteurs plus bas niveau, les descripteurs spectraux ont été préférés aux descripteurs HNR et PAP avec en tête le centroïde spectral. Les descripteurs spectraux, initialement beaucoup plus nombreux, se retrouvent donc très représentés dans l'ensemble final des descripteurs sélectionnés. Les descripteurs cepstraux (MFCC) ressortent comme plus pertinents ici que les descripteurs décrivant directement l'energie spectrale (EBB). Les formants sont également largement représentés dans l'ensemble final des descripteurs qui sera soumis au système de classification.

Pour le contenu non voisé, l'energie en Bande de Bark (EBB) est cette fois plus représentée que les descripteurs cepstraux. A noter également la sélection du descripteur mesurant le rapport periodique-aperiodique (PAP).

[32] voir paragraphe 4.1.2

Famille	Nini2/Nini1	Descripteurs sélectionnés	Nfinal/Nini2
Prosodiques	7/33	$meanF_0$, $minF_0$, F_0, $maxF_0$, $stdevdF_0$, $rangedF_0$, $rangeF_0$	7/7
Qualité de voix	8/37	$Jitter$, $Shimmer$	2/8
Spectraux	93/464	$meanC_s$, $minMFCC1$, $meanMFCC4$, $maxF_1$, $minMFCC4$, $mindF_1$, $mindF_2$, $meanMFCC1$, $rangedF_1$, $rangedF_2$, $rangeF_1$, $rangeF_2$, $MFCC4$, $MFCC1$, $stdevF_2$, $maxdF_1$, $maxdF_2$, $maxMFCC4$, $maxC_s$, $minMFCC3$, $skewEBB3$, $meanEBB3$, $maxF_2$, $stdevdMFCC11$, $stdevdF_2$, $kurtdF_1$, $minF_2$, $kurtF_1$, $rangeMFCC1$, $stdevdMFCC6$, $minMFCC6$	31/93

TAB. 6.3 – *Liste des 40 descripteurs sélectionnés pour le contenu voisé de SAFE_1 (tableau 5.2). Nini1 = nombre descripteurs extraits, Nini2 = nombre de descripteurs soumis à la 2e sélection* ($Nini2 = \lceil \frac{Nini1}{5} \rceil$)*, Nfinal = nombre de descripteurs sélectionnés à la suite des deux sélections, range = plage, stdev = écart type, kurt = aplatissement, skew = disymétrie*

Famille	Nini2/Nini1	Descripteurs sélectionnés	Nfinal/Nini2
Prosodiques	4/16	$rangeInt$	1/4
Qualité de voix	8/36	$tauxNonVoise$, $kurtdPAP$	2/8
Spectraux	93/464	$rangeEBB6$, $rangeEBB7$, $rangeEBB10$, $stdevEBB6$, $stdevEBB7$, $rangeEBB8$, $rangeEBB5$, $stdevEBB10$, $rangeEBB11$, $rangeEBB9$, $stdevEBB8$, $rangeEBB12$, $maxMFCC3$, $stdevEBB5$, $stdevEBB9$, $rangeEBB4$, $rangeMFCC10$, $rangeMFCC12$, $stdevEBB11$, $minEBB10$, $rangeMFCC8$, $minEBB7$, $minEBB6$, $rangeEBB3$, $rangeMFCC6$, $minEBB8$, $rangedEBB6$, $minMFCC11$, $stdevEBB12$, $stdevEBB4$, $minEBB9$, $meanMFCC3$, $rangeMFCC11$, $rangedEBB5$, $maxBw_1$, $rangedEBB4$, $maxBw_2$	37/93

TAB. 6.4 – *Liste des 40 descripteurs sélectionnés pour le contenu non voisé de SAFE_1 (tableau 5.2).*

Ce sont presque essentiellement des descripteurs statistiques sur la trajectoire qui ont été sélectionnés comme les plus pertinents. L'unité d'analyse qu'est la trajectoire semble donc être adaptée à la modélisation du contenu émotionnel du corpus SAFE.

6.3.2 Paramétrage des GMM

La mise en place du système de classification peur/neutre par les GMM impose le réglage de deux paramètres, le poids attribué à chacun des classifieurs (voisé et non voisé) lors de la décision et le nombre de gaussiennes du modèle.

Poids des classifieurs voisé vs. non voisé

La méthode utilisée lors de la décision dans le cas des GMM est décrite dans le paragraphe 6.2.3.

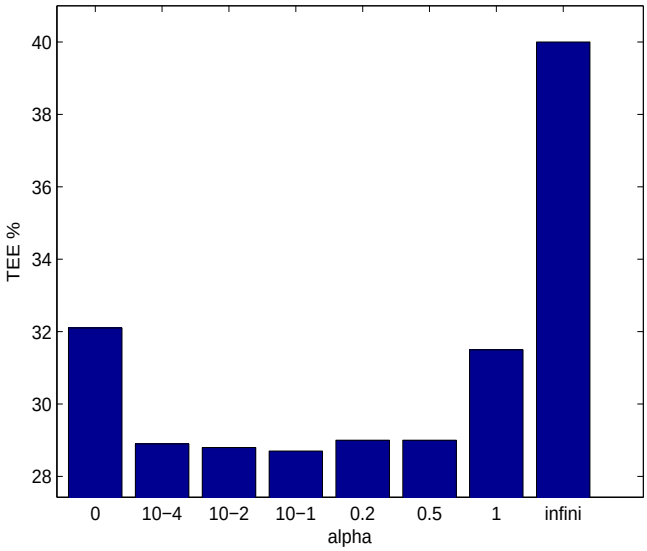

FIG. 6.3 – *Taux d'égale erreur en fonction de α ($w = 1 - r^{\alpha}$) qui règle le poids attribué au classifieur non voisé par rapport au classifieur voisé sur SAFE_1 (tableau 5.2) (intervalles de confiance à 95% près : rayon $\leq$ 3%)*

La figure 6.3 présente les taux d'égale erreur de la classification peur/neutre obtenus pour différents valeurs du paramètre α duquel dépend le poids ($w = 1 - r^{\alpha}$) attribué à chaque classifieur. Le nombre de gaussiennes est fixé à 8 pour cette expérience. L'influence du nombre de gaussiennes sur les résultats est étudiée dans le paragraphe suivant.

Le classifieur voisé est plus efficace que le classifieur non voisé. Le cas où seul le classifieur non voisé est considéré ($\alpha = \infty$) est celui qui donne les performances les moins bonnes avec un taux d'égale erreur atteignant 40%. L'utilisation exclusive du classifieur non voisé revient en effet à ignorer l'information importante portée par le contenu voisé.

A l'inverse, si le classifieur voisé est utilisé en priorité (le classifieur non voisé est utilisé uniquement lorsque les segments ne contiennent pas de fenêtres voisées, $\alpha = 0$), le taux d'égale erreur est de 32%.

Les meilleurs résultats ($TEE = 29\%$) sont obtenus lorsque le classifieur non voisé est utilisé avec un poids qui décroît rapidement lorsque le taux de voisement augmente, i.e. pour $\alpha = 0.1$ (cf figure 6.2).

Nombre de gaussiennes

Le tableau 6.5 présente les résultats obtenus en fonction du nombre de gaussiennes utilisées pour modéliser les deux classes. Le paramètre α permettant de régler le poids attribué au classifieur non voisé est ici réglé à 0.1. Ce tableau montre l'influence du nombre de gaussiennes sur les

Nombre de gaussiennes	4	8	16
TEE	31%	29%	30%

TAB. 6.5 – *Taux d'égale erreur (TEE) pour le système de classification peur vs. neutre en fonction du nombre de gaussiennes considérées avec les GMM sur SAFE_1 (tableau 5.2) (intervalles de confiance à 95% près : rayon ≤ 3%).*

performances. Celles-ci restent similaires aux intervalles de confiance près lorsque le nombre de gaussiennes utilisé pour modéliser les différentes classes augmente. Sur le sous-corpus d'étude, les meilleures performances ne sont pas obtenues pour le nombre de gaussiennes testé le plus élevé et sont obtenues pour un nombre de gaussiennes égal à 8. Ceci peut être dû au fait que les données apprises sont des données bruitées. Si le nombre de gaussiennes s'avère trop élevé par rapport au nombre de données, le risque est en effet d'aboutir à de la sur-généralisation.

Résultats optimisés

Les meilleurs résultats ont donc été obtenus avec un nombre de gaussiennes fixé à 8 et un α pour la fusion des deux classifieurs de $0,1$. Ces résultats sont présentés dans le tableau 6.6 et serviront de référence pour la suite de ce chapitre.

automatique / manuel	Neutre	Peur
Neutre	71%	29%
Peur	30%	70%
TEM	29%	
TEE	29%	
κ	0,53	

TAB. 6.6 – *Matrice de confusion, taux d'erreur normalisé, taux d'égale erreur pour le système de classification peur vs. neutre sur SAFE_1 (tableau 5.2) (intervalles de confiance à 95% près : rayon ≤ 3%)*

Le kappa correspondant est de 0,53. Ce kappa peut être comparé au kappa obtenu par l'« annotateur système » dans le chapitre 4 qui était de 0,57. Les résultats obtenus par le système sont très encourageants dans la mesure où ils sont presque aussi bons que ceux obtenus par un annotateur placé dans les mêmes conditions.

6.4 Analyse des comportements du système

Les résultats de la partie précédente sont des résultats globaux fournissant les taux de reconnaissance de chaque classe. La stratégie d'annotation développée dans le chapitre 2 fournit cependant d'autres champs d'annotations que les annotations en classes d'émotions. Ces annotations peuvent ainsi être corrélées avec les annotations en deux classes peur/neutre afin d'obtenir des sous classes regroupant les données de chaque classe en des sous ensembles. L'objectif de cette partie est d'identifier les comportements spécifiques du système sur ces sous ensembles.

6.4.1 Comportements du système en fonction du degré d'imminence de la menace

La description du contenu du corpus fournie dans le chapitre 4 a mis en évidence la diversité des manifestations émotionnelles présentes dans notre corpus, notamment en fonction des types de menaces illustrés. Le tableau 6.7 illustre les comportements du système de classification en fonction du degré d'imminence de la menace sur la reconnaissance de la classe peur. La classe peur a un taux de reconnaissance globale de 70%. Si seuls sont considérés les segments de peur survenant lors de menaces latentes, le taux de reconnaissance chute à 62% avec un intervalle de confiance à 95% de ±8%. En revanche si seuls sont considérés les segments survenant lors de menaces immédiates, les performances sont largement meilleures et le taux de reconnaissance de la peur atteint 78% avec un intervalle de confiance à 95% de ±5%. Les performances obtenues sur

		%de segments testés	taux de reconnaissance
peur	pas de menace	7%	61%±18%
	menace potentielle	4%	64%±24%
	menace latente	33%	60%±8%
	menace immédiate	50%	78%±5%
	menace passée	5%	71%±18%

TAB. 6.7 – *Répartition et taux de reconnaissance avec leur intervalle de confiance à 95% des segments de la classe peur en fonction du degré d'imminence de la menace sur SAFE_1 (tableau 5.2)*

les menaces passées s'avèrent meilleures. Ces segments correspondent plutôt à des manifestations émotionnelles assez intenses et correspondent à un passé proche de la situation de menace (voir tableau 4.9). Cependant, le nombre de segments en menace passée est insuffisant pour assurer la reproductibilité du résultat sur d'autres données[33].

Les manifestations émotionnelles survenant en situations normales (pas de menace), ou lors de menace potentielle ou latente, correspondent, comme nous l'avons vu dans le chapitre 3, plutôt à des manifestations modérées de la peur de type inquiétude ou anxiété. Dans ces segments les manifestations acoustiques de la peur sont moins fortes, ce qui explique la différence de performances entre les différents contextes d'émergence de la peur.

[33]les intervalles de confiance sont fournis à titre indicatif pour les sous-ensembles *pas de menace*, *menace potentielle* et *menace passée* malgré un nombre de segments insuffisant pour l'utilisation de la formule donnée dans le paragraphe 6.2.4

6.4.2 Comportements du système en fonction des annotations de référence

Le chapitre 4 a soulevé un problème propre à l'étude des émotions dans la parole : le problème de la subjectivité des annotations. Ce problème resurgit lors du développement d'un système de classification d'émotions pour les deux raisons suivantes :

1. Les modèles acoustiques associés à chaque classe ou les fonctions de décision sont appris à partir d'une base de données d'apprentissage. Or cette base de données d'apprentissage est conditionnée par les annotations effectuées par le ou les annotateurs considérés. Par conséquent les modèles acoustiques ou fonctions de décision qui vont en dériver peuvent être très différents en fonction des annotations considérées ;
2. L'évaluation des performances du système consiste en l'analyse des confusions entre les annotations et les décisions du système, comme présenté dans le paragraphe 6.2.4. Les performances obtenues vont donc être dépendantes des annotations servant de référence. Ce problème est également abordé dans [Steidl *et al.*, 2005].

Le premier problème a déjà été abordé dans le chapitre 4. Nous nous focalisons ici sur le second problème.

L'évaluation des performances du système est effectuée par la matrice des confusions entre le système et chaque annotateur et est présentée dans les tableaux 6.8, 6.9, 6.10. Les performances obtenues avec le TEE oscillent entre 30% (en considérant les annotations de Ann3 comme référence) et 35% (en considérant les annotations de Ann1 comme référence), soit une différence d'environ 5%. Ces performances restent cependant similaires aux intervalles de confiance près.

Ann1 \ Système	Neutre	Peur
Neutre	70%	30%
Peur	39%	61%
TEM	34%	
TEE	35%	
κ	0,48	

TAB. 6.8 – *Matrice de confusion pour le système de classification peur vs. neutre en utilisant comme référence les annotations de Ann1 (704 segments pour la classe neutre et 631 segments pour la classe peur, intervalles de confiance à 95% : rayon $\leq$ 4%)*

Ann2 \ Système	Neutre	Peur
Neutre	69%	31%
Peur	32%	68%
TEM	32%	
TEE	32%	
κ	0,45	

TAB. 6.9 – *Matrice de confusion pour le système de classification peur vs. neutre en utilisant comme référence les annotations de Ann2 (1322 segments pour la classe neutre et 518 segments pour la classe peur, intervalles de confiance à 95% : rayon $\leq$ 4%)*

La distribution entre les classes change en fonction de l'annotation de référence et donc aussi l'accord lié au hasard. Le kappa correspondant à chaque annotateur de référence est éga-

Ann3 \ Système	Neutre	Peur
Neutre	72%	29%
Peur	31%	69%
TEM	30%	
TEE	30%	
κ	0,53	

TAB. 6.10 – *Matrice de confusion pour le système de classification peur vs. neutre en utilisant comme référence les annotations de Ann3 (352 segments pour la classe neutre et 309 segments pour la classe peur, intervalles de confiance à 95% : rayon $\leq$ 5%)*

lement précisé. Les kappas obtenus entre l'« annotation système » et les annotations de Ann1 ($\kappa = 0,25$) et Ann2 ($\kappa = 0,36$) étaient beaucoup plus faibles que ceux obtenus ici. Ce résultat est assez surprenant et signifie que le système a un comportement plus proche de celui des annotateurs qui disposaient du contexte que l'annotateur système. Ceci pourrait être dû au fait que la reconnaissance de la peur par le système ne repose pas que sur des descripteurs haut niveau du signal audio mais aussi sur des descripteurs plus bas niveau qui modélisent une information acoustique difficilement perceptible par l'homme, même expert.

Il pourrait être intéressant de valider cette hypothèse en proposant d'autres annotations en « conditions système ».

6.5 Analyse de l'imminence de la menace par la reconnaissance de la peur

6.5.1 Objectif

Le système de classification proposé précédemment suit un schéma de classification binaire qui permet de discriminer les émotions de type peur de l'état neutre. Dans le paragraphe 6.4.1, nous avons mis en évidence les différences des performances sur les sous ensembles de segments de peur en fonction de l'imminence de la menace.

Le système obtient de très bonnes performances lorsque la menace est immédiate. En revanche, lorsque la menace est latente les performances baissent sensiblement pour la reconnaissance de la classe peur. Les manifestations de la peur illustrées dans le corpus SAFE sont très variées. En particulier, dans le chapitre 4, nous avons pu voir comment les manifestations émotionnelles évoluaient en fonction de la menace, les menaces potentielle et latente étant plus favorables à l'émergence de peur de type inquiétude ou anxiété et les menaces immédiates à l'émergence de peur de type panique ou terreur.

La classe peur rassemble donc des segments de peur ayant des signatures acoustiques très différentes en fonction du type de menace durant lequel ils surviennent. L'objectif de cette partie est de construire des modèles acoustiques spécifiques pour chacun des sous ensembles de la classe peur déterminés en fonction de l'imminence de la menace. Etant donnée la quantité insuffisante de données dont nous disposons pour les sous-ensembles de segments de la classe peur survenant en menace potentielle et en menace passée, nous avons considéré les deux sous-classes suivantes[34] :

[34]Nous ne créons pas ici de modèle spécifique pour les peurs survenant en situations normales, celles-ci n'étant pas visées par le système de détection de situations anormales

– sous-classe PeurLatPot : sous-classe contenant les segments survenant en menace potentielle ou latente,
– sous-classe PeurImmPass : sous-classe contenant les segments survenant en menace immédiate ou passée (passé proche),

Ce regroupement se justifie également par la proximité situationnelle entre les menaces latentes et potentielles, d'une part, et entre les menaces immédiate et passée, d'autre part.

Nous voulons maintenant essayer d'extraire des informations sur l'imminence de la menace. Jusqu'à maintenant, le système de reconnaissance des émotions de type peur était destiné à signaler la présence d'une situation anormale. Il nous a semblé intéressant à ce stade du travail d'évaluer la possibilité d'extraire une information supplémentaire sur l'imminence de la menace en reconnaissant entre elles les différentes sous-classes associées aux différentes manifestations émotionnelles de la peur.

Cette information sur la menace pourrait en effet venir enrichir le système de détection de situations anormales par un système d'analyse de la situation permettant de diagnostiquer l'imminence de la menace afin d'aider l'homme à prendre la décision appropriée pour limiter les dommages. Les mesures à prendre sont conditionnées par le degré d'imminence de la menace.

6.5.2 Principe

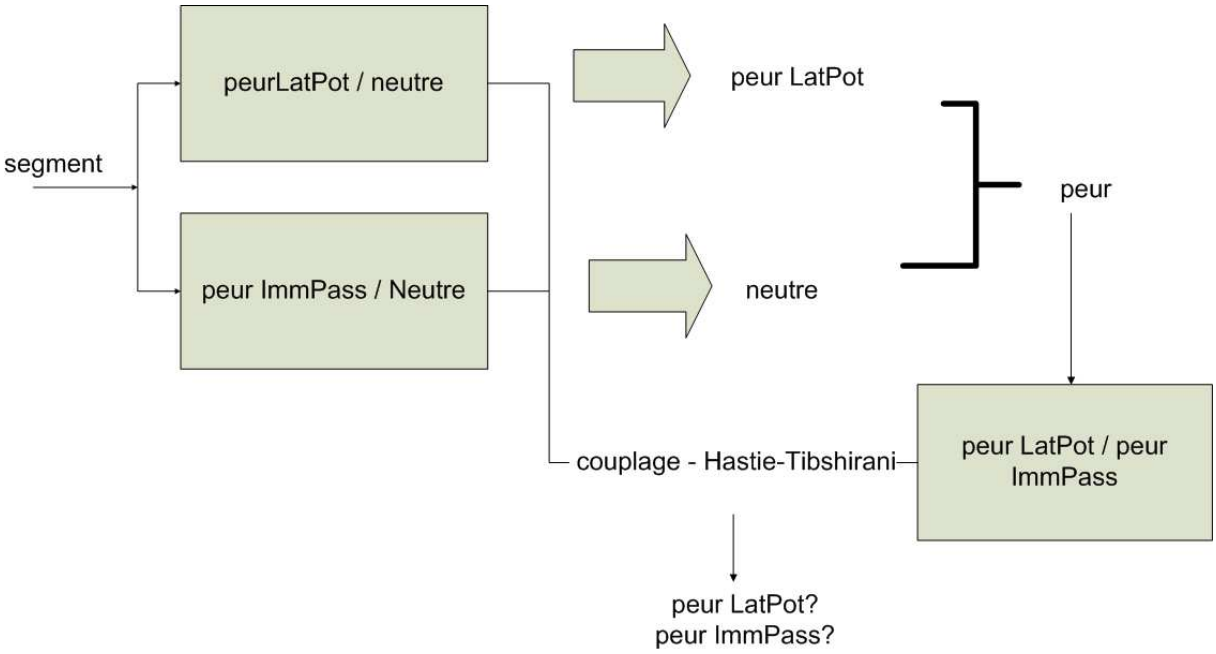

FIG. 6.4 – *Exemple de fonctionnement du classifieur*

Un classifieur est associé à chacune des sous-classes de la classe peur. La classification peur/neutre va maintenant reposer sur la fusion de deux classifieurs binaires : peurImmPass vs. neutre et peurLatPot vs. neutre.

Pour chaque segment le score *a posteriori* correspondant aux deux classes *peurLatPot*, *peurImmPass* est calculé et comparé au score *a posteriori* de la classe neutre. La classification peur/neutre est effectué en utilisant la règle de décision suivante : le segment est classé peur s'il a été classé peur, i.e. *peurLatPot* ou *peurImmPass* par l'un des deux classifieurs.

Les segments qui ont été reconnus comme peur sont soumis à un classifieur supplémentaire *peurLatPot* vs. *peurImmPass* comme illustré sur la figure 6.4. La classification *peurLatPot* vs. *peurImmpass* est effectuée en utilisant l'approche proposée par Hastie et Tibshirani [Hastie, 1996] afin de réaliser un couplage optimal des trois classifieurs. Cette approche a été implémentée par Slim Essid dans [Essid *et al.*, 2006] pour la reconnaissance d'instruments de musique par GMM.

Pour chaque observation x_t, i.e. pour chaque fenêtre, les probabilités de chaque classe *peurLatPot*, $p(C_1|x_t)$ et *peurImmPass*, $p(C_2|x_t)$ sont estimées selon le modèle suivant :

$$p_{i,j}(C_i|x_t) = \frac{p(C_i|x_t)}{p(C_i|x_t) + p(C_j|x_t)}$$

où $p_{i,j}(C_i|x_t)_{1\leq i,j\leq 3}$ correspond à la probabilité que x_t appartienne à C_i en considérant le classifieur binaire $\{C_i, C_j\}$.

Les scores *a posteriori* de chaque classe associés à chaque segment sont ensuite comparés pour réaliser la classification finale.

6.5.3 Résultats

Les résultats de la classification sont présentés dans le tableau 6.11.

man. \ autom.	neutre	peurLatPot	peurImmPass
neutre	61%	21%	18%
peurLatPot	34%	**34%**	32%
peurImmPass	13%	27%	**60%**

TAB. 6.11 – *Matrice de confusion pour la classification peurLatPot vs. peurImmPass sur SAFE_1 (tableau 5.2). Les intervalles de confiance pour les classe peurImmPass et peurLatPot sont de rayon* $\leq$ *7%.*

60% des segments annotés *peurImmPass* sont correctement reconnus par le système comme des manifestations de peur correspondant à des menaces immédiates ou passées et 34% comme des menaces latentes ou potentielles.

Ces premiers résultats nous ammènent à la reflexion suivante. Extraire des informations sur le contexte telles que le degré d'imminence de la menace à travers l'analyse des manifestations émotionnelles, est une tâche délicate. Nous nous sommes focalisé sur les manifestations acoustiques de l'émotion. Utilisée seule, l'information acoustique semble insuffisante pour extraire une information contextuelle de si haut niveau. Il serait donc intéressant pour améliorer ces résultats de corréler l'information acoustique avec d'autres informations telles que les informations linguistiques, visuelles ou contextuelles.

6.6 Conclusion

Les performances des systèmes de reconnaissance des émotions existants dépendent fortement :

- du corpus sur lequel ils ont été développés (nombre de locuteurs, diversité des contextes illustrés, degré d'authenticité des émotions)
- du degré de complexité requis par l'application : indépendance au locuteur, utilisation d'outil de segmentation, alignement avec le texte, etc.

Si les systèmes basés sur des analyses dépendantes du locuteur, reposant sur un alignement avec la transcription et menées sur des corpus actés, obtenus en laboratoire, peuvent dépasser les 90% de réussite, les rares systèmes, qui s'attaquent à des données plus complexes avec de réelles contraintes applicatives, voient leurs performances chuter.

Les performances obtenues par notre système de reconnaissance des émotions de type peur sont donc très encourageantes compte tenu de la diversité des contextes d'émergence de la peur illustrés dans le corpus SAFE. Les résultats obtenus ont pu être validés en utilisant deux méthodes de classification les GMM et les SVM. L'originalité de notre système réside dans la considération des deux types de contenu du signal de parole, le contenu voisé et le contenu non voisé. Cette approche a permis de faire baisser le taux d'égale erreur de 32% à 29%.

Nous avons mené une analyse des comportements du système afin de mieux comprendre l'influence de la diversité inhérente au corpus SAFE sur les performances du système. Nous avons notamment étudié les performances obtenues en fonction des contextes situationnels liés à la menace. Le taux de reconnaissance des émotions de type peur est de 78% sur les segments de peur survenant en menace immédiate. Le système a cependant plus de difficultés à reconnaître les émotions de type peur lorsqu'elles surviennent en menace latente avec un taux de reconnaissance de 60%. Celles-ci sont en effet moins bien exprimées au niveau acoustique,

Nous avons également évalué la dépendance des performances aux annotations servant de référence et comparé les performances humaines de catégorisation avec les performances du système de classification. Reconnaître des émotions automatiquement est un problème complexe, car les annotations sur lesquelles repose le système sont subjectives. Nous soulignons ici l'importance de la prise en compte de cette subjectivité dans l'évaluation des différents systèmes.

Enfin, nous avons exploré la possibilité d'accéder au niveau contextuel par les manifestations émotionnelles. Si les peurs en menace latente s'expriment différemment des peurs en menace immédiate, est-il possible de déduire de ces différences une information sur le degré d'imminence de la menace ? Nous avons tenté de répondre à cette question en utilisant uniquement l'information acoustique locale (contenue dans le segment). Cette information reste cependant insuffisante pour accéder à un niveau d'interprétation aussi élevé que le niveau contextuel. Ceci fait émerger l'intérêt de considérer non seulement la dimension dynamique des manifestations émotionnelles en étudiant la progression des manifestations émotionnelles au sein d'une même séquence mais aussi des informations linguistiques ou visuelles afin de modéliser la situation.

Les résultats du système sont obtenus sur le corpus SAFE, qui correspond à un corpus de développement nous ayant permis d'ajuster les paramètres du système afin d'obtenir les meilleures performances. La validation de ces résultats sur un corpus de test constitué de données réelles de surveillance permettrait de tester la portabilité des modèles acoustiques construits à partir du corpus SAFE. Le passage à des données réelles pourrait entraîner une chute des performances. En contrepartie, l'adaptation à un environnement sonore spécifique celui du lieu à surveiller pourrait améliorer les performances. Ces deux hypothèses relatives à l'évolution des résultats sur des données réelles de surveillance pourront être vérifiées dans un travail ultérieur dès l'acquisition de telles données.

Troisième partie

Vers une plateforme de surveillance effective

Résumé

L'objectif applicatif des différentes études présentées dans cette thèse est l'intégration du système de reconnaissance de la peur dans une plateforme de surveillance dédiée à la détection et à l'analyse des situations anormales. Les informations extraites de la modalité audio sont destinées à être corrélées aux informations issues de la modalité vidéo afin de pouvoir être analysées par un système de fusion de l'information pour la détection et l'analyse de la situation.

Dans cette partie, nous illustrons le fonctionnement du module audio d'une telle plateforme sur le corpus SAFE à travers un démonstrateur. Le module audio visé repose sur les manifestations à la fois vocales et non vocales, incluant en parallèle un système de détection des émotions et un système de détection d'événements anormaux.

Nous avons abordé le problème de la détection d'un événement typique de la situation anormale : le coup de feu. Ceci nous a permis de définir une méthodologie pour la détection d'événements en environnement bruité dans un flux audio. Les performances du système ont été évaluées pour un type de lieu, le marché.

Un démonstrateur combinant les deux systèmes de détection a été développé. Il utilise la plateforme multi-agent OAA (Open Agent Architecture). Cette plateforme permet la gestion des différents agents (acquisition de la séquence à traiter, système de détection de coups de feu et système de détection de la peur).

Chapitre 7

Système de détection et d'analyse des situations anormales pour la surveillance dans les lieux publics

Sommaire

Introduction

L'automatisation des systèmes de surveillance est un sujet de recherche qui rassemble de nombreux laboratoires et industriels [ADVISOR, 2000], [SERKET, 2005]. La tâche de surveillance est réalisée jusqu'à maintenant essentiellement par l'homme qui est souvent chargé de la surveillance de plusieurs lieux en parallèle. Celui-ci est par conséquent soumis à une charge cognitive préjudiciable à sa vigilance. Le développement de systèmes automatiques de surveillance est donc motivé par la nécessité d'assister l'homme dans sa tâche de surveillance. Ces systèmes répondent également aux besoins de la police d'accéder rapidement aux données enregistrées pour établir la traçabilité d'un évènement donné.

Les systèmes de surveillance se focalisent cependant essentiellement sur des indices visuels avec par exemple la détection d'objets anormaux, d'intrusions, la détection de mouvements de foule ou de comportements anormaux par le suivi d'individu. Cependant pour certaines situations, l'image ne suffit pas. C'est le cas notamment des manifestations de situations anormales qui ont lieu hors du champ des caméras, i.e. dans les angles morts, ou encore lorsque les évènements anormaux ont des manifestations visuelles peu saillantes. Le coup de feu est une illustration pertinente de ce dernier cas.

En se focalisant sur l'image pour détecter les situations anormales, la plupart des systèmes ignore alors une grande partie de l'information qui pourrait être disponible et utile pour la détection de situations anormales. Le couplage audio/vidéo permet d'avoir une analyse exhaustive de la situation.

Le système de reconnaissance de la classe *peur* décrit dans le chapitre précédent est destiné à être intégré à une plateforme de surveillance multimodale que nous décrivons ici dans le paragraphe 7.1. Nous nous sommes focalisé sur le module audio de cette plateforme. Il nous a paru crucial dans un objectif de surveillance de considérer en plus de l'information contenue dans le signal de parole, l'information non vocale à travers la détection d'événements audio caractéristiques de comportements anormaux (paragraphe 7.2). Le module audio ainsi constitué a donné lieu à un démonstrateur qui est présenté dans le paragraphe 7.3.

Publication(s) associée(s) à ce chapitre: [Clavel *et al.*, 2005]

7.1 Plateforme multimodale de surveillance – Synopsis

La plateforme de surveillance dans laquelle notre module audio est destiné à être inséré a un double rôle : donner l'alarme dans un premier temps, puis, dans un second temps fournir des éléments pour l'analyse de la situation. Le schéma d'une telle plateforme est illustré dans la figure 7.1. Les alarmes vont permettre de signaler la présence d'une situation anormale et d'orienter ainsi l'attention de l'utilisateur sur le lieu où le problème a été localisé. Fournir à l'utilisateur des éléments d'analyse de la situation peut l'aider ensuite à prendre rapidement une décision appropriée pour limiter les dommages. Dans le chapitre 6, nous avons vu par exemple que le diagnostique de l'incidence de la menace peut être utile pour la gestion des situations anormales par l'homme.

Les informations extraites des modalités audio et vidéo sont destinées à être analysées simultanément par un système de fusion de l'information. La stratégie adoptée ici repose sur l'extraction d'un maximum d'événements des deux flux audio et vidéo qui une fois extraits peuvent être fusionnés. Cette stratégie est appelée « late fusion » par opposition à la stratégie de fusion dite de « early fusion » qui consiste en une fusion des deux modalités effectuée dès l'étape

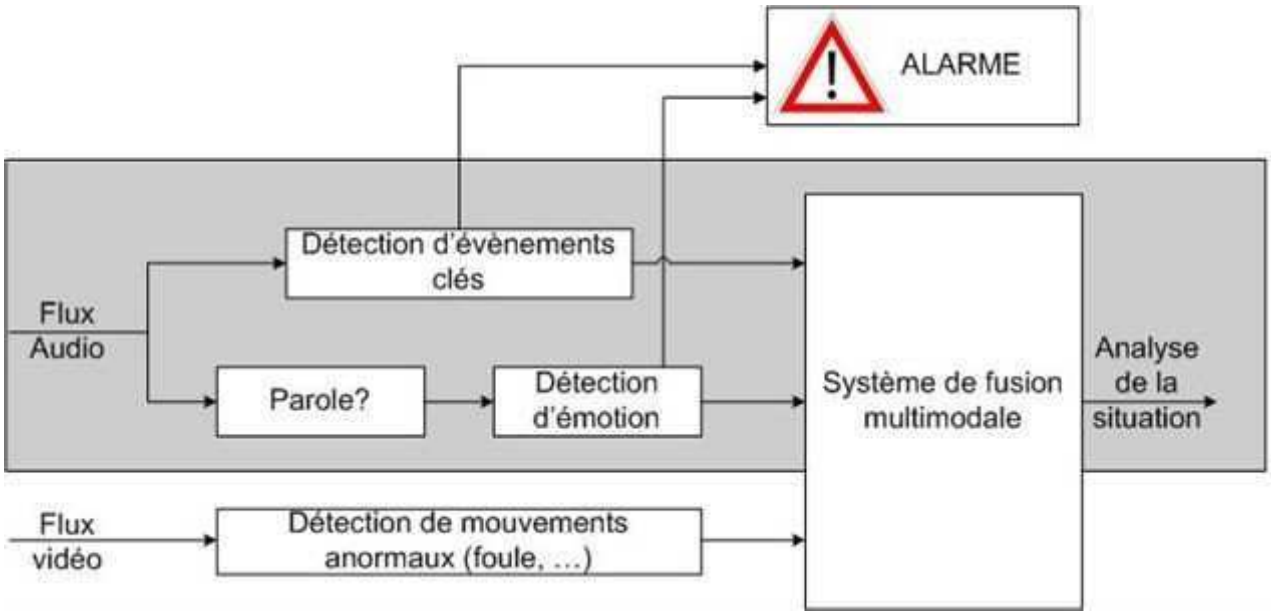

FIG. 7.1 – *Système de détection et d'analyse des situations anormales*

de représentation du signal audiovisuel [Snoek *et al.*, 2005] par des descripteurs.

Nous nous sommes concentré sur l'extraction d'informations du flux audio et seul le module audio a été développé ici. L'extraction d'informations du flux vidéo et le couplage audio vidéo n'a pas été abordé dans le cadre de cette étude.

Le module audio se base sur les manifestations à la fois vocales et non vocales des situations anormales en considérant le contenu émotionnel de la parole et des événements acoustiques clé. Le comportement des locuteurs présents en situations anormales entraîne en effet l'émergence de bruits plus ou moins violents, comme les coups de feu, coups de poing, etc. Ces bruits accompagnent les manifestations émotionnelles en situations anormales et peuvent ainsi fournir des indices pour la détection et la caractérisation de ces situations. Nous avons choisi de considérer ces bruits comme des événements-clé permettant de déclencher l'alarme du système de détection de situation anormale.

7.2 Détection d'événements anormaux

Nous avons choisi pour la détection d'événements anormaux, le cas d'un événement acoustique typique : le coup de feu. L'une des difficultés majeures recontrée lors du développement d'un tel système est liée au fait que le bruit environnant est souvent non stationnaire et peut être fort par rapport à l'événement à détecter. Le cas du coup de feu nous a permis de définir une méthodologie pour la détection d'événements en environnement bruité *dans un flux audio*. C'est cette méthodologie qui est présentée ici après un bref bilan des méthodes utilisées en détection d'événements audio.

7.2.1 La détection/classification audio – Bilan

La détection ou la classification d'événements audio à des fins d'indexation est un sujet d'étude très répandu dans la communauté scientifique, que ce soit pour l'indexation de documents

audiovisuels [Cai *et al.*, 2003], ou pour l'indexation de documents audio [Essid *et al.*, 2006]. La détection de ruptures dans le signal sonore à partir de la bande audio d'un film peut par exemple fournir des indices concernant la structure temporelle du film [Nam et Tewfik, 2002].

Par ailleurs la détection d'événements audio commence à apparaitre dans des applications spécifiques de surveillance telles que la surveillance médicale [Vacher *et al.*, 2004].

Les méthodes utilisées pour ces différentes applications peuvent être regroupées en trois classes :

- l'utilisation de fonction de détection ou de seuillage,
- l'utilisation de méthode de classifications d'événements audio,
- l'utilisation de méthode basée sur la détection de nouveauté.

La première classe de méthode consiste à rechercher dans le signal sonore les variations brusques en niveau d'énergie. Ce niveau ne nécessite pas d'apprentissage. Un exemple de l'utilisation de ce type de méthode est donné dans [Nam et Tewfik, 2002] pour la détection des actions violentes dans les films. La fonction de détection utilisée se base sur l'utilisation du critère d'entropie de l'énergie.

La seconde classe de méthodes, les méthodes de classification de sons, consiste en l'apprentissage des différentes classes de sons à reconnaître. Ce type de méthode est par exemple utilisé dans [Guo et Li, 2003] pour classifier différents sons tels que des sons d'instruments de musique, des bruits de foules, etc.. L'article [Vacher *et al.*, 2004] utilise une méthode de classification après une étape préalable de détection. Son système de détection d'événements se décompose en deux phases, une première phase de détection permettant de localiser les variations brusques du signal audio et une phase de classification dont le rôle est d'identifier le signal détecté.

Enfin la troisième classe de méthodes, la détection de nouveauté [Markou et Singh, 2003] est un mélange entre les deux méthodes précédentes : un modèle pour l'état normal du signal est appris et la « nouveauté » est détectée lorsque la distance à ce modèle dépasse un seuil fixé au préalable.

La méthode utilisée et décrite ci-dessous s'inspire des deux dernières classes de méthodes : un modèle acoustique de l'événement audio à détecter (événement anormal) est construit et comparé au modèle acoustique de l'état normal correspondant ici à l'environnement sonore en situation normale.

7.2.2 Le système de détection de coup de feu

Le flux audio est segmenté en segments de 0,5 s. avec un recouvrement de 50% entre les segments successifs, comme illustré sur la figure 7.2. Chaque segment est ensuite soumis à un système de classification qui considère les deux classes *coup de feu*/*normal*. La classe *normal* permet de modéliser l'environnement sonore en situations normales. L'architecture du système de classification est présentée sur le schéma 7.3. Le système comprend :

- un module d'extraction des descripteurs,
- un module d'apprentissage des modèles associés à chaque classe à l'aide des GMM (Gaussian Mixture Models),
- un module de classification qui, à partir des modèles appris, classe les segments audio successifs.

Extraction des descripteurs

Les descripteurs sont extraits sur des fenêtres d'analyse d'une durée de 20 ms avec un recouvrement de 50% entre les fenêtres. Les descripteurs utilisés ici ont été choisis parmi ceux

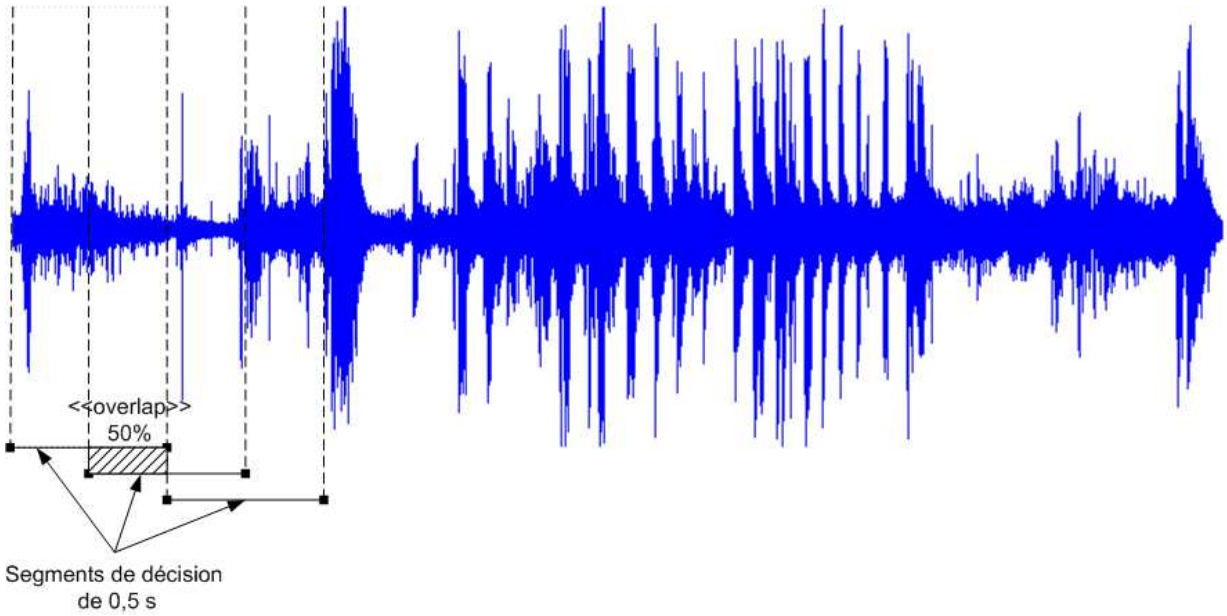

FIG. 7.2 – *Segmentation du flux audio en segments de 0,5 s avec un recouvrement de 50% entre les fenêtres*

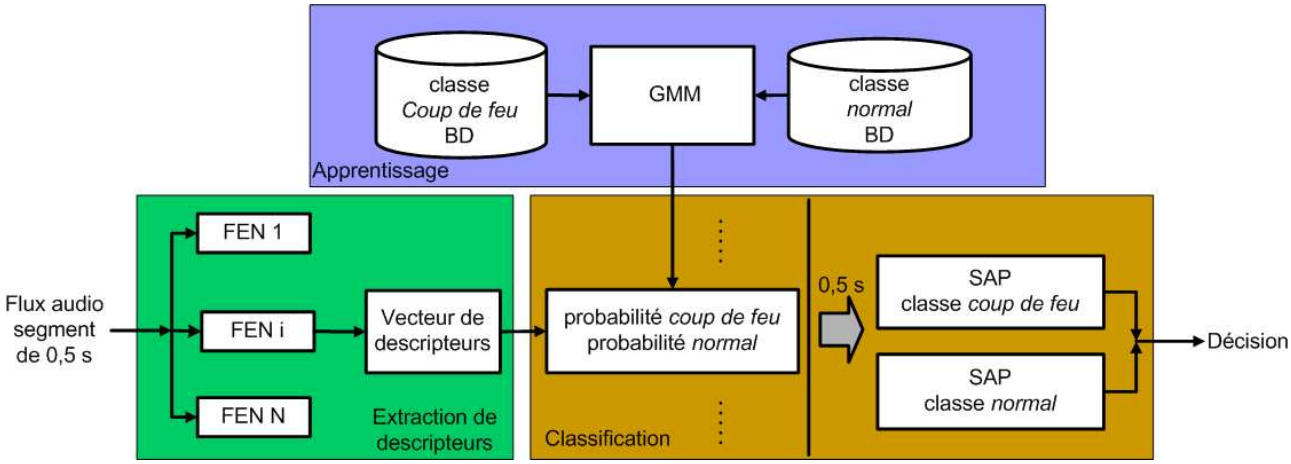

FIG. 7.3 – *Système de détection de coups de feu*

les plus couramment utilisés en traitement de l'audio et qui nous ont semblé les plus pertinents pour la détection de coups de feu :

- l'énergie à court-terme décrit l'énergie du signal sur une fenêtre d'analyse ;
- les huit premiers coefficients MFCC (Mel Frequency Cepstral Coefficient) dont une description est fournie dans le chapitre 5 ;
- les deux premiers moments statistiques du spectre du signal. Le premier moment est également appelé centroïde spectral et est décrit dans le chapitre 5, le second modélise l'étendue spectrale ;
- les dérivées premières et secondes de chacun des descripteurs ci-dessus.

La dimension de l'espace des descripteurs est ici réduite via une analyse en composantes principales. Seules les 13 premieres composantes sont gardées ici comme significatives. Chaque fenêtre d'analyse est ainsi représentée par une vecteur de dimension de 13.

Apprentissage

L'apprentissage est réalisé ici par une modélisation sous la forme d'une somme de gaussiennes (GMM). Un modèle est construit pour chaque classe. La méthode est la même que celle utilisée par notre système de reconnaissance des émotions et est décrite dans le chapitre 6. Le nombre de gaussiennes utilisées ici pour chaque classe est de 16. Les fenêtres de silence sont éliminées de la base de données d'apprentissage. La détection des fenêtres de silence est effectuée en éliminant toutes les fenêtres présentant une amplitude maximale faible (1000 fois plus petite) par rapport à l'amplitude maximale globale.

Classification

La décision finale concernant la classe d'appartenance coup de feu ou normal de chaque segment de 0,5 s est prise en comparant les scores a posteriori[35] (SAP) obtenus pour chaque modèle de classes. Les fenêtres détectées comme étant des fenêtres de silence ne sont pas considérées pour le test.

7.2.3 Base de données et protocole

Nous avons vu dans le chapitre 2 que l'acquisition de données audio réelles illustrant des situations anormales est difficile compte tenu du caratère rare et imprédictible de telles situations. Le corpus SAFE décrit dans le chapitre 4 illustre des situations anormales avec des coups de feu. Cependant le son associé aux coups de feu dans les films de fiction est souvent peu réaliste et correspond la plupart du temps à un son de bruitage, i.e. à un autre événement acoustique que celui produit par un réel coup de feu dans l'environnement de la scène filmée.

Le problème du matériel utilisé pour le système de détection des événements anormaux se pose donc également ici.

Matériel utilisé

Etant données ces difficultés nous avons choisi de générer des données artificielles qui soient les plus proches possibles des situations réelles. Pour cela nous avons utilisé des enregistrements sonores de lieux publics et de coups de feu tirés du CD [Mercier, 1989]. Ce CD réalisé pour Radio France propose une librairie de sons, incluant des enregistrements réels de coups de feu

[35] Voir chapitre 6

correspondant à différentes armes et différentes prises de sons dans un grand nombre de lieux publics (station de métro, hall d'aéroport, stade, marché).

Le contenu des données de coups de feu ainsi collectées est décrit dans le tableau 7.1. Un total de 134 coups de feu, soit une durée totale de 296 secondes est ainsi rassemblé. Ce tableau donne la quantité de données en nombre de coups de feu et en durée pour chaque type d'arme.

arme	Pistolet	Fusil	Mitraillette	Grenade	Coup de canon
nombre de coups de feu	5	15	79	8	27
durée totale	5s	24s	134s	28s	105s

TAB. 7.1 – *La base de données de coups de feu*

Les enregistrements du type de lieu le plus illustré dans le CD, le marché, sont sélectionnés pour cette étude. Ils sont organisés en 4 plages d'enregistrement. La durée totale de ces enregistrements est de 797 secondes, soit environ 13 minutes et 17 secondes.

Base de données d'apprentissage et base de données de test

A partir de ces différents enregistrements nous avons constitué une base de données d'apprentissage et une base de données de test. Le protocole d'acquisition de ces bases de données est représenté sur la figure 7.4.

La base de données utilisée pour l'apprentissage de la classe *normal* est constituée des 75 dernières secondes de chacun des 4 enregistrements. La durée restante est segmentée en séquences de 30 secondes qui sont utilisées pour la base de données de test.

La base de données de test est issue d'un mixage entre les coups de feu et les séquences provenant des enregistrements du marché. Chacun des 134 coups de feu est inséré dans une séquence choisie aléatoirement parmi les séquences de la base de test. L'insertion du coup de feu est effectuée à un instant de la séquence choisi aléatoirement. 134 séquences sont ainsi générées. Dans chacune de ces séquences le coup de feu est inséré pour des rapports signal sur bruit (RSB) de 5, 10, 15 et 20dB. Le RSB correspond au rapport entre la puissance du signal de coup de feu et la puissance du bruit environnant calculée sur la portion de signal dans laquelle le coup de feu a été inséré.

Les séquences de test ainsi obtenues fournissent une simulation de situations anormales (dans le cas de l'apparition d'un coup de feu) dans des lieux publics aussi proches que possible de la réalité. Leur caractère artificiel facilite l'évaluation du système dans la mesure où le RSB peut être contrôlé et où les vérités terrains, i.e. la localisation du coup de feu dans la séquence de test, sont obtenues sans nécessiter une annotation préalable. Ces vérités terrain vont servir de référence pour l'évaluation du système automatique de détection.

Protocole *Leave One Shot Out*

L'ensemble d'apprentissage et l'ensemble de test utilisé pour la classe normal sont distincts. Pour la classe coup de feu, le protocole d'évaluation appelé *Leave One Out* est ici appliqué : le coup de feu inséré dans la séquence testée est supprimé de la base de données d'apprentissage utilisée pour l'apprentissage de la classe coup de feu. 134 apprentissages différents sont ainsi effectués.

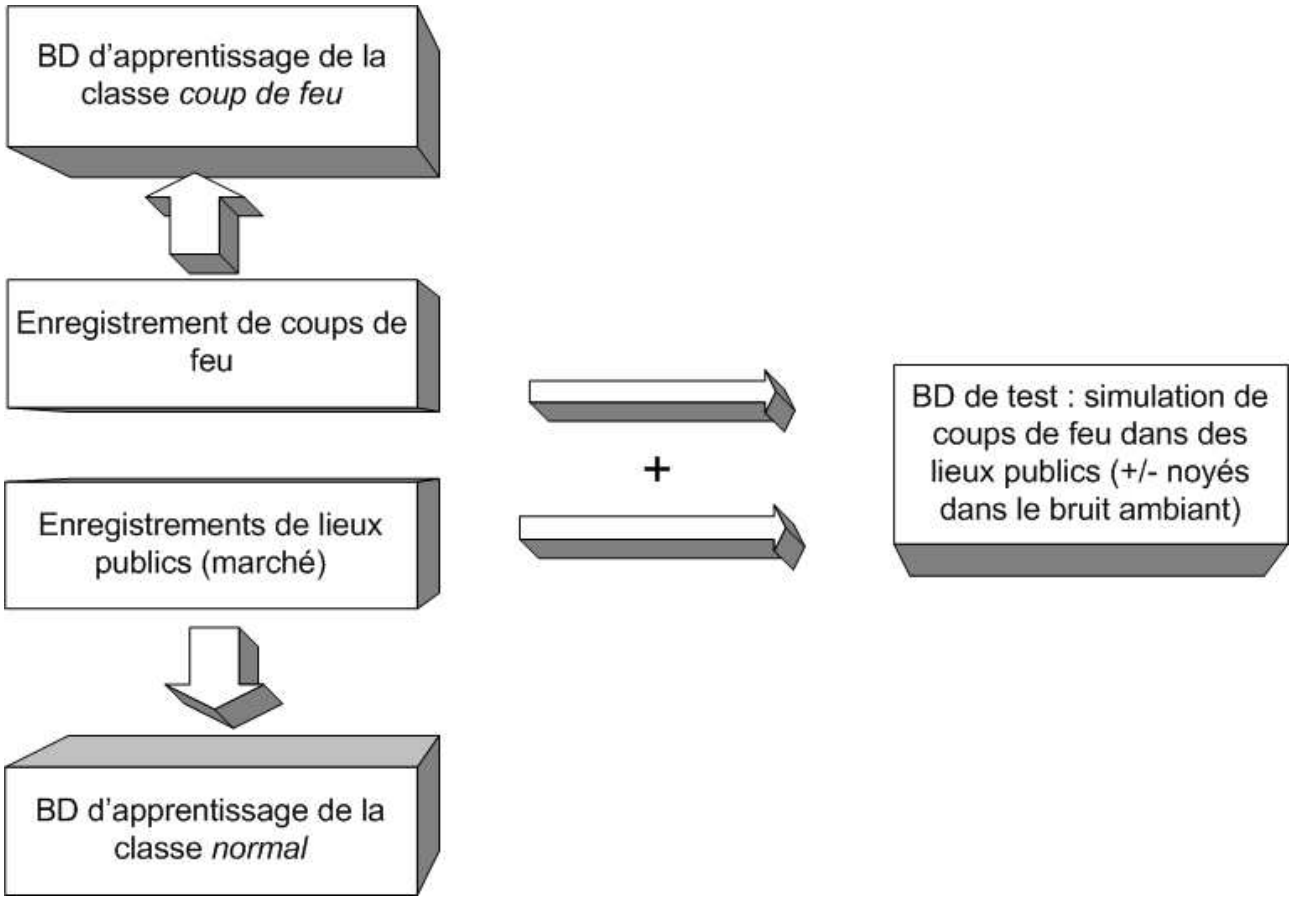

FIG. 7.4 – *Mise en place de la base de données d'apprentissage et de la base de données de test*

Évaluation des résultats

Les sorties du système de détection sont comparées aux vérités terrains obtenues. Les résultats sont donnés par le calcul du taux de faux rejets (FR) et du taux de fausses détections (FD) qui sont définis par les formules suivantes.

$$FR = \frac{nombre\ de\ faux\ rejets}{nombres\ d'événements\ à\ détecter}$$

$$FD = \frac{nombre\ de\ fausses\ détection}{nombre\ de\ segments\ de\ décision}$$

7.2.4 Expérimentations et résultats

Les performances du système ont d'abord été évaluées en utilisant pour la classe coup de feu les données d'apprentissage décrites dans le paragraphe précédent appelées ici base de données « propre ». Puis dans une perspective d'amélioration de ces performances nous proposons également d'apprendre les modèles de coups de feu dans leur environnement sonore. Quatre nouvelles bases de données de coups de feu mixés dans le bruit environnant sont générées pour quatre RSB différents. Les coups de feu sont en effet insérés dans des portions d'enregistrements de marché pour des rapports signal sur bruit allant de 5dB à 20dB avec un pas de 5dB. La

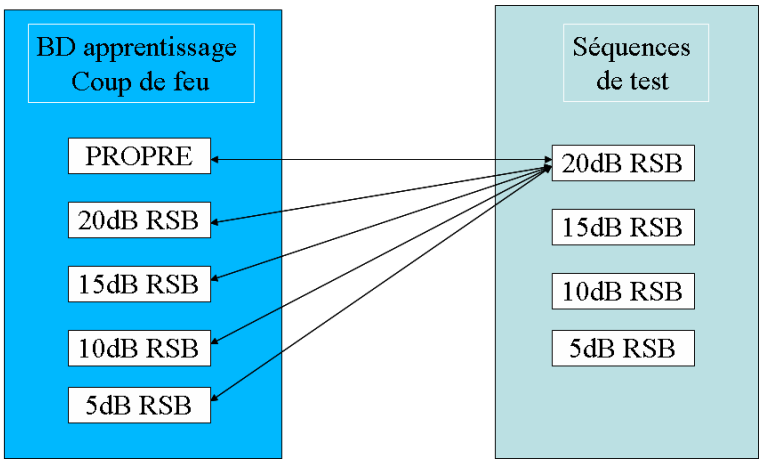

FIG. 7.5 – *Protocole de test*

figure 7.5 résume les différentes expérimentations effectuées. Les séquences de tests sont divisées en 4 groupes, chaque groupe contenant 134 séquences correspondant aux 134 coups de feu insérés pour un RSB donné. Chaque groupe de séquences est testé pour chaque condition d'apprentissage de la classe coup de feu.

Les résultats de ces différentes expérimentations sont présentés sur la figure 7.6. Plus l'événement à détecter est noyé dans le bruit environnant, i.e plus le RSB des séquences de test est bas, plus les performances du système se dégradent. La courbe noire correspondant aux performances obtenues sur les séquences de test ayant un RSB de 5dB est en effet au dessus de la courbe rouge associée aux séquences dans lesquelles le coup de feu est inséré avec un RSB de 20dB. En particulier la base de données d'apprentissage dite « propre » offre des résultats insuffisants sur les séquences les plus bruitées en termes de faux rejets. À l'inverse l'utilisation de modèles acoustiques construits sur des bases de données trop bruitées pour la classe coup de feu

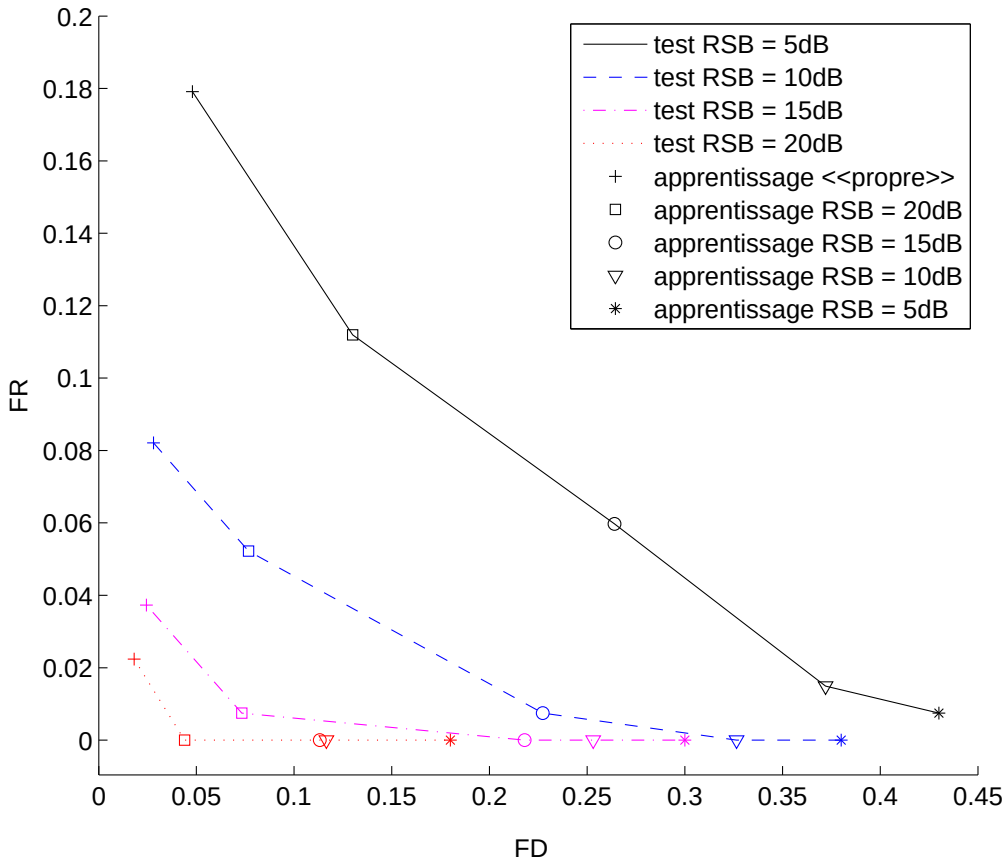

FIG. 7.6 – *Taux de faux rejets en fonction du taux de fausses détections pour chacune des expérimentations. Chaque courbe correspond aux performances obtenues pour une même condition de test, i.e. pour un RSB donné*

entraîne une augmentation considérable du taux de fausses détections qui atteint 43% pour les séquences les plus bruitées (RSB = 5dB) testées avec les modèles de coup de feu les plus bruités.

Cette expérimentation illustre le compromis nécessaire entre les faux rejets et les fausses détections en utilisant des modèles de coups de feu appris à partir d'une base de données ayant le RSB le plus approprié. Pour une application de surveillance il est particulièrement important d'obtenir un taux de faux rejets aussi faible que possible, les fausses alarmes étant moins « graves » que les oublis.

Il émerge de ces différentes expérimentations que le meilleur compromis est obtenu pour toutes les conditions de test en utilisant la base de données d'apprentissage fournissant un RSB de 20dB. Le taux de fausses détections, avec cette base de données d'apprentissage ne dépasse pas les 11% et le taux de fausses détections reste également en dessous des 15%.

Ces résultats obtenus sur des données expérimentales pourront être validés sur des données réelles. Dans ce contexte, il pourra être également nécessaire d'adapter la conception du système afin de prendre en compte les problèmes engendrés par la rareté des événements à détecter.

7.3 Démonstrateur

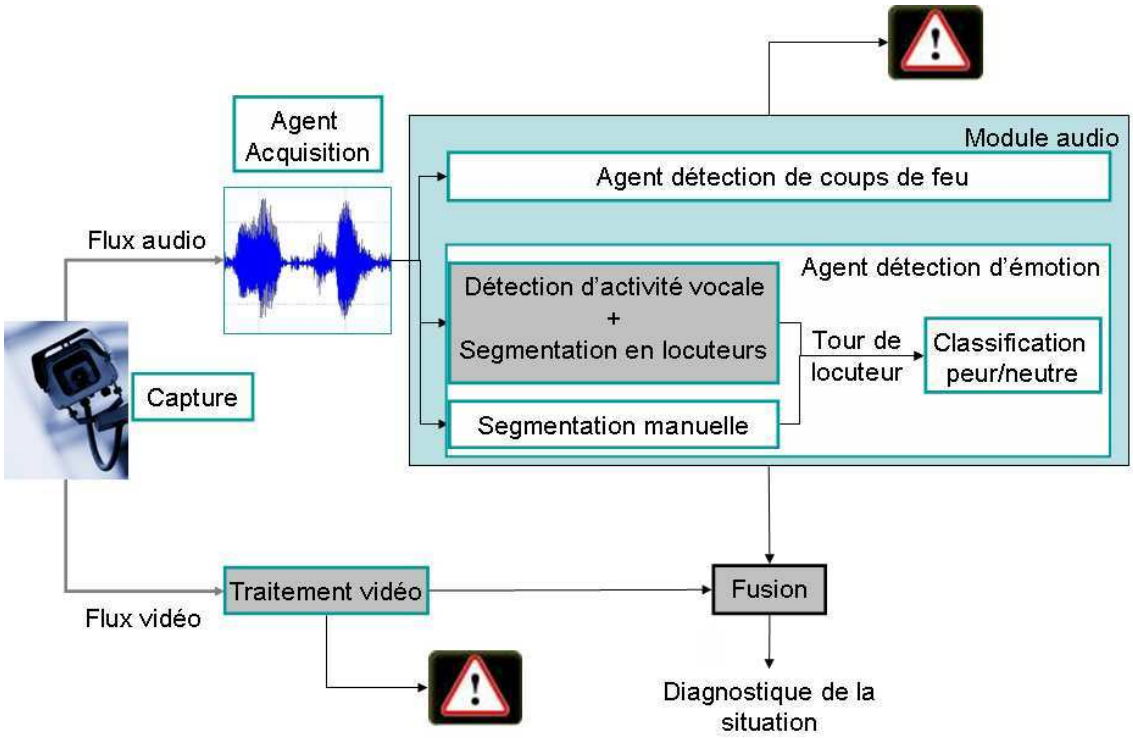

FIG. 7.7 – *Fonctionnement détaillé du démonstrateur*

Un démonstrateur du module audio de détection de situations anormales combinant le système de détection de coup de feu et le système de reconnaissance des émotions de type peur a été développé. Il utilise la plateforme multi-agent OAA [Cheyer et Martin, 2001] (Open Agent Architecture). Cette plateforme est celle préconisée pour les démonstrateurs développés au sein de l'équipe HIT[36] de Thales Recherche et Technologies. Elle présente l'avantage d'offrir les fonc-

[36]Human Interaction Team

tionnalités telles que le fonctionnement distribué permettant de répartir la charge des traitements sur des serveurs spécialisés en laissant l'affichage au poste client, ou encore le fonctionnement par agents autonomes communiquant par requête via un superviseur. Dans le cadre de ce démonstrateur, la plateforme permet la gestion des différents agents (acquisition de la séquence à traiter, système de détection de coups de feu et système de détection de la peur, affichage des résultats grâce à une interface java) de manière indépendante.

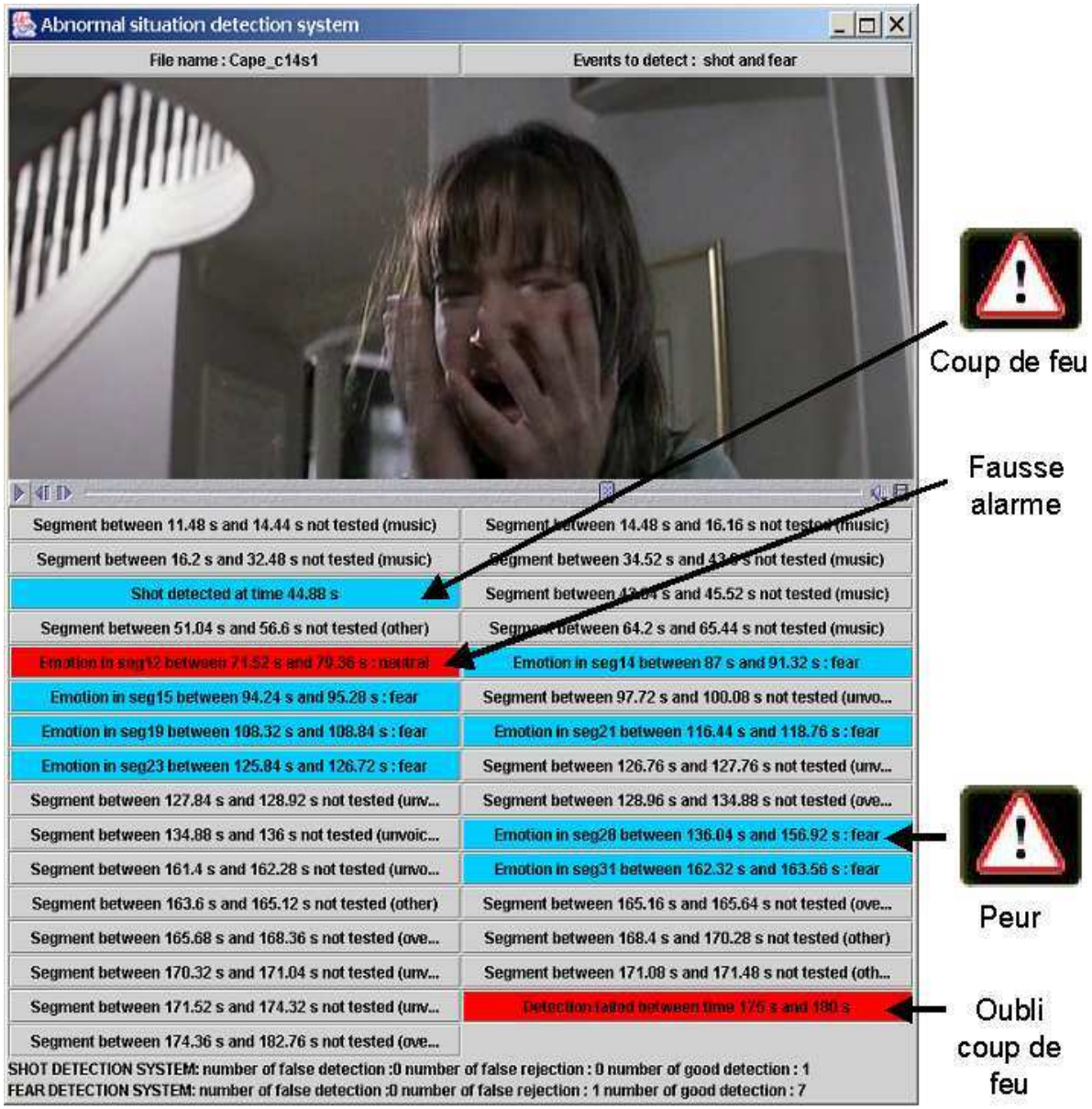

FIG. 7.8 – *Interface du démonstrateur*

Le schéma 7.7 présente le fonctionnement détaillé du système de détection de situations anormales. Le flux audio est traité en parallèle par le système de détection de coups de feu et le système de détection d'émotions. Les boîtes non grisées correspondent aux différents modules développés ici. Comme nous l'avons vu dans le paragraphe 7.2.2, le système de détection de coup de feu fournit une décision sur des intervalles de 0,5 s. du segment. En revanche, le système

de détection d'émotions s'appuie lui sur une segmentation du flux audio en tours de parole qui est réalisée ici manuellement. Le remplacement de cette segmentation manuelle par un outil permettant automatiquement de détecter l'activité vocale et les changements de locuteurs est à l'étude actuellement dans le cadre d'un stage réalisé au sein de l'équipe HIT de Thales Recherche et Technologies.

Le fonctionnement du module audio est illustré sur des séquences du corpus SAFE. Au fur et à mesure du déroulement de la séquence, l'interface (figure 7.8) affiche les différents événements audio censés intervenir dans la séquence et obtenus à partir des annotations. L'événement apparaît en bleu si le système automatique de détection de situations anormales l'a correctement détecté et en rouge dans le cas contraire et dans le cas de fausses alarmes.

Nous avons vu dans le chapitre 6 que le système de reconnaissance des émotions de type peur développé s'appuie sur une classification peur/neutre. Seuls les segments annotés peur ou neutre par l'un des annotateurs sont donc traités par le démonstrateur. Les segments correspondant à des recouvrements entre locuteurs (*overlap*) ou ayant été annotés comme présentant une mauvaise qualité d'enregistrement de la parole[37] ne sont pas encore traités. Ces segments apparaissent en gris sur l'interface.

7.4 Conclusion

Nous avons montré dans ce chapitre comment l'information véhiculée par la modalité audio ignorée jusqu'alors par les systèmes de surveillance peut être intégrée par ceux-ci pour le diagnostique de situations anormales. Le démonstrateur développé ici intègre notamment une analyse des manifestations émotionnelles dans la parole et des événements sonores caractéristiques de comportements anormaux. Ces premiers développements constituent une étape à l'intégration de cette étude à un système de surveillance effectif. Ils permettent de soulever des problèmes nouveaux tels que le passage d'un système de reconnaissance d'émotion à un système de détection d'émotion, ou encore l'adaptation du système de détection de situations anormales à un type de lieu, i.e. à un type d'environnement sonore particulier (parking, marché, métro).

Les événements anormaux du type coup de feu et les émotions de type peur dans des données réelles de surveillance sont a priori plutôt rares. Par exemple, l'évaluation des performances sur de telles données pourrait soulever de nouveaux problèmes quant au seuil permettant d'équilibrer fausses détections et faux rejets. Il serait donc intéressant en guise de perspectives de réévaluer les performances obtenues pour les deux systèmes (reconnaissance des émotions de type peur et détection de coup de feu) sur des données réelles correspondant directement à l'application. Par ailleurs, il pourra être également nécessaire d'adapter la conception du système afin de prendre en compte les problèmes engendrés par la rareté des événements à détecter.

[37] i.e. annotés Q0 ou Q1 voir chapitre 4

Quatrième partie

Conclusion et perspectives

Chapitre 8

Conclusion et perspectives

Sommaire

8.1 Apports de la méthodologie

Le premier apport de cette thèse réside en la définition d'une *méthodologie* pour la conception d'un système automatique de classification d'émotions.

Notre méthodologie s'appuie sur des savoir-faire issus de disciplines connexes telles que l'informatique, le traitement du signal, la psychologie, la linguistique. L'enjeu de cette thèse a été de circonscrire ces savoir-faire et d'en extraire les informations utiles pour notre objectif de recherche.

Nous récapitulons ci-dessous les étapes importantes à considérer lors de la conception d'un système automatique de classification d'émotions :

1. Le choix et l'acquisition d'un *matériel d'étude* ainsi que l'évaluation de sa qualité en termes d'adéquation avec l'objectif de recherche (types d'émotions et de contextes recherchés) ;
2. La définition d'une *stratégie d'annotation* qui permettent l'exploitation des données par le système (décrire le contenu émotionnel du corpus en un langage interprétable par la machine, adapté à la tâche visée, offrant des annotations fiables et pertinentes pour la compréhension des comportements du système) ;
3. La gestion des problèmes de *subjectivité des annotations* (les annotations considérées pour former l'ensemble d'apprentissage, les annotations servant de référence pour le test, création d'un référentiel en termes de performances humaines pour la catégorisation d'émotions) ;
4. La représentation des données sous la forme de *descripteurs acoustiques efficaces* pour le problème de classification considéré et extraits en adéquation avec les contraintes applicatives (techniques de normalisation, choix de l'unité temporelle d'analyse) ;
5. L'*apprentissage par le système* d'un problème de classification par le choix et le paramétrage de méthodes de classification ;
6. L'évaluation des performances du système selon des *protocoles* définis selon les contraintes applicatives (conditions de dépendance ou d'indépendance au locuteur, au contexte, etc.) ;
7. L'*analyse des comportements du système* en fonction du type de données testées en vue de son adaptation à d'autres données. Identification de ses faiblesses et de ses atouts.

L'efficacité de cette méthodologie a été validée sur un exemple applicatif : l'application de surveillance. Cette application innovante dans le domaine de la reconnaissance d'émotions présente l'avantage de motiver de nouveaux besoins en corpus émotionnels et et nous a permis de relever de nouveaux défis en termes de système de reconnaissance des émotions. Par ailleurs, le système de reconnaissance a été intégré au sein d'un démonstrateur destiné à illustrer le module audio d'une plateforme de surveillance. Ce démonstrateur inclut également un système de détection des événements anormaux caractéristiques des situations anormales.

Corpus émotionnel et annotation des émotions

L'originalité de nos travaux de recherche réside dans la nature des émotions et des contextes recherchés : les émotions de type peur en situations de menace. L'étude de ce type d'émotions se heurte à un problème majeur : leur caractère rare et imprévisible complexifie la tâche de collecte de données émotionnelles authentiques. De plus les données de surveillance existantes présentent à l'heure actuelle peu d'enregistrements audio et les contextes de menace recherchés sont rares dans de telles données. Nous avons choisi les films de fiction comme matériel de base pour la

conception de notre système. Le corpus SAFE développé par nos soins constitue une première en termes de bases de données émotionnelles par la diversité des contextes illustrés pour une même émotion cible. Si les bases de données émotionnelles commencent à présenter une grande diversité de locuteurs, les environnements sonores et les situations illustrés restent restreints. Le corpus SAFE intervient sur cet aspect en fournissant des situations anormales de nature variée telles que des prises d'otage et des incendies pour différents degrés d'imminence allant de la menace potentielle à la menace passée. Dans ces situations émergent des manifestations très variées de l'émotion cible : inquiétude, panique, terreur, peur-colère, etc.

$\Longrightarrow$ *Notre corpus, le corpus SAFE présente une grande diversité de contextes d'émergence pour une même émotion cible.*

Les 7 heures de séquences audiovisuelles du corpus SAFE fournissent des contextes, correspondant aussi bien à des situations normales qu'anormales. Cette variété de manifestations émotionnelles offre la possibilité d'étendre l'utilisation de ce corpus à d'autres fins de recherche.

La richesse du contexte d'émergence des émotions est intégrée dans notre stratégie d'annotation. L'évolution de la situation et celle des manifestations émotionnelles sont annotées en parallèle permettant l'étude des corrélations entre manifestations émotionnelles et contexte situationnel. Cette stratégie d'annotation se focalise sur la description des manifestations orales de l'émotion par le découpage des séquences audiovisuelles en *segments*, définis en fonction des changements émotionnels dans chaque tour de parole. La description est aidée par la vidéo qui permet un décodage plus précis des émotions en fournissant un complément contextuel. L'enjeu d'une telle stratégie est de s'affranchir des spécificités du matériel d'étude en proposant non seulement des descripteurs adaptés à l'application et transposables à des données réelles de surveillance (génériques vis-à-vis du corpus d'étude) mais également des descripteurs transposables à tout type d'application (génériques vis-à-vis de l'application).

$\Longrightarrow$ *Notre stratégie d'annotation intègre la description des manifestations émotionnelles dans leur contexte d'émergence.*

Système de reconnaissance des émotions

Une telle diversité est requise pour une application de surveillance dont l'objectif est de pouvoir détecter des situations anormales très variées. La conception d'un système de reconnaissance des émotions de type peur sur des données présentant une telle hétérogénéité constitue l'une des contributions majeure de cette thèse. Les contraintes applicatives nous ont poussés à affranchir notre analyse acoustique d'un alignement avec la transcription et de techniques de normalisation basées sur une connaissance a priori du locuteur. Les performances ont été évaluées dans des conditions d'indépendance au locuteur. Malgré ces contraintes, les performances obtenues sont très encourageantes avec un taux d'égale erreur de 29% pour une discrimination peur/neutre.

$\Longrightarrow$ *Notre système de reconnaissance des émotions de type peur est indépendant du locuteur et du contexte et obtient malgré ces contraintes des performances qui dépassent les 70% de bonne reconnaissance de la peur par rapport au neutre.*

L'originalité de la méthode utilisée repose sur la stratégie adoptée pour la représentation du signal de parole sous la forme de descripteurs acoustiques. Premièrement nous avons extrait

des descripteurs acoustiques non seulement sur les portions voisées du signal, comme le font la plupart des études, mais aussi sur ses portions non voisées. La considération de ces deux types de contenu au sein du système de classification permet de faire baisser le taux d'égale erreur de 32% à 29%.

Deuxièmement, nous avons choisi de considérer des descripteurs acoustiques modélisant des contenus très variés : les descripteurs prosodiques, les descripteurs de la qualité de voix et les descripteurs spectraux et cepstraux. Les descripteurs les plus efficaces sont sélectionnés pour chaque contenu afin de fournir une représentation du signal de parole pertinente pour la tâche de reconnaissance des émotions de type peur. Cette méthode qui repose sur l'extraction d'un grand nombre de descripteurs puis sur leur sélection en fonction du problème de classification posé présente l'avantage d'être transposable à tout autre problème de classification d'émotions.

$\Longrightarrow$ *La méthode de classification utilisée offre la possibilité de considérer l'intégralité de l'information (voisée et non voisée) contenue dans le signal de parole.*

8.2 Perspectives de recherche

Les attentes dans le domaine du traitement automatique des émotions par la machine sont énormes, et le domaine émergeant. Parce que nos travaux interviennent en vue d'une application effective, ils ont pour vocation première de mettre en lumière le chemin qu'il reste à parcourir. Un bout du chemin a été parcouru dans le cadre de cette thèse, et nos travaux laissent la place à de multiples perspectives.

8.2.1 Les perspectives à court-terme

Extension à un système multi-classes

Nous avons mis en place un système de reconnaissance de la peur par rapport à l'état neutre. L'extension de ce système à un problème multi-classes, incluant la colère, permettrait d'une part de discriminer la peur des autres émotions et d'autre part d'affiner l'analyse de la situation, pour une application de surveillance. La mise en place d'un système multi-classes plus subtil sur des données très hétérogènes constitue un sujet de recherche qui, malgré sa complexité, mérite d'être traité à court-terme.

Le traitement des recouvrements entre locuteurs

Un des défis qu'il reste à relever pour le fonctionnement du système dans toutes les situations est une analyse des différents types de recouvrements entre locuteurs et leur prise en compte par la stratégie d'annotation et par le système de reconnaissance. Nous avons isolé ces recouvrements et nous avons distingué les recouvrements entre un groupe de locuteurs des recouvrements correspondant à des manifestations au niveau de la foule. A court terme la perspective est de voir si les modèles acoustiques construits sur des émotions au niveau de l'individu sont portables sur les données présentant des recouvrements (notamment dans le cas d'un locuteur prédominant). A plus long terme, il s'agira d'améliorer la reconnaissance de la peur sur de telles données en construisant des modèles acoustiques spécifiques, notamment dans le cas de la foule. Pour cela il sera nécessaire d'acquérir des données illustrant des manifestations émotionnelles de foule en quantité suffisante.

Adaptation à des données réelles

Les données sur lesquelles repose notre étude sont des données tirées des films de fiction et qui ont été choisies pour le réalisme des émotions jouées par les différents acteurs et pour la diversité des contextes illustrés. La méthodologie d'annotation et le système de reconnaissance des émotions ont été évalués avec succès sur ce corpus de développement. Le système de reconnaissance des émotions de type peur conçu ici devra cependant fonctionner sur des données réelles de surveillance. Or, les émotions du corpus SAFE sont des émotions actées qui peuvent s'avérer différentes des émotions vécues dans des conditions réelles. Une des perspectives à court-terme de nos travaux qui nous semblent la plus importante est l'étude de la portabilité des modèles acoustiques construits à partir de notre matériel d'étude sur des données illustrant des émotions vécues dans des contextes réels. Ces données pourront être idéalement des données de surveillance ou des extraits télévisés de situations anormales, tels que présentés dans les « no comment » d'Euronews.

Système de classification

Il existe de nombreuses méthodes de classification. Nous en avons testé deux indépendamment, les GMM et les SVM. La fusion des résultats obtenus par des classifieurs différents pourrait être une piste pour l'amélioration des performances. Les techniques de classification utilisées par notre système pourraient être améliorées en s'appuyant sur l'expertise acquise dans les domaines de la vérification du locuteur ou de l'identification de langue (techniques de normalisation des descripteurs, adaptation des modèles, algorithmes de réduction de l'espace de réprésentation des descripteurs, etc.).

Passage à un système de détection

Comme souligné dans ce manuscrit, le système développé est pour l'instant un système de *reconnaissance* des émotions. Pour passer à un système de *détection*, nous avons pensé à l'utilisation d'outils de segmentation en locuteurs afin d'isoler les portions de parole associées à un locuteur. Chacune de ces portions pourra être segmentée en des intervalles de décision qui seront soumis en entrée du système de reconnaissance. Notre choix qui a été de proposer des descripteurs acoustiques du signal dont l'unité d'analyse (fenêtre de 40ms, trajectoire voisée, trajectoire non voisée) ne repose pas sur une segmentation manuelle facilitera le passage à un système de détection.

8.2.2 Les perspectives à long-terme

La dimension dynamique

Les manifestations acoustiques de la peur ont été étudiées en fonction de l'imminence de la menace en séparant dans tout le corpus les peurs en menace latente, immédiate ou passée. Nous n'avons pas encore pris en compte la dimension dynamique des manifestations émotionnelles. Le corpus SAFE présente des séquences qui illustrent l'évolution d'une situation donnée. Il serait donc particulièrement intéressant de tirer profit de cet atout en étudiant la progression des manifestations émotionnelles au sein d'une même séquence, voire même au sein de plusieurs séquences d'un même film qui sont associées à une même situation de menace.

Les annotations du comportement

Le schéma d'annotation proposé permet une description des manifestations émotionnelles et du stimulus (la situation de menace) associé. La nouvelle dimension introduite ici, la réactivité permet une annotation du comportement du locuteur face à la menace et donc de modéliser les réactions du locuteur face à la situation. Le choix d'une annotation dimensionnelle rend cependant cet aspect difficile à exploiter. L'une des améliorations susceptibles d'être apportées au schéma d'annotation est l'intégration d'une description plus fine du comportement du locuteur par la définition de catégories de comportement qui peuvent s'avérer plus maniables qu'une dimension pour l'annotation.

Les indices linguistiques de la peur

Nous nous sommes focalisé sur le contenu acoustique pour la reconnaissance de la peur. Cependant certains de nos segments contiennent des marques lexicales spécifiques qui peuvent être attribuées à la peur. C'est le cas typiquement du mot « Help ! ». La détection de ces indices lexicaux pourrait ainsi être utilisée en complément des indices acoustiques pour la reconnaissance des émotions de type peur.

Étude des corrélations entre audio et vidéo

L'intégration du module audio dans des systèmes de surveillance automatique existants et basés sur les informations vidéos pourra également être envisagée. Cette intégration pourra être l'occasion d'explorer les possibles corrélations entre les deux modalités vidéo et audio pour l'analyse de la situation.

Cinquième partie

Annexes

Annexe A

Corpus et Outils

Outils d'extraction et de sélection

Le rip et l'encodage des chapitres du DVD sélectionnés au format Xvid se font à l'aide du logiciel NeodivX. Les passages des chapitres qui nous intéressent sont sélectionnés et extraits grâce au logiciel VirtualDub, qui permet également l'extraction de la séquence sonore associée à la vidéo au format wav.

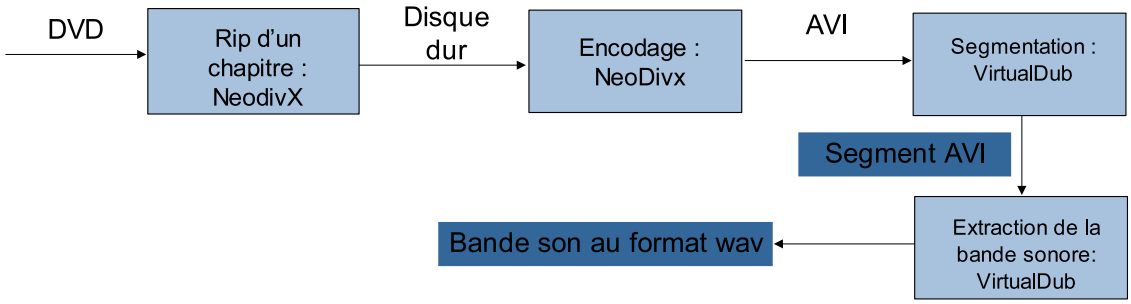

FIG. A.1 – *Protocole d'extraction des séquences à partir des DVD*

Outils d'annotation

Notre choix en ce qui concerne le logiciel d'annotation s'est arrêté sur un outil d'annotation pour le dialogue multimodal, ANVIL (Annotation for Video and Language)[38] du dfki. ANVIL permet de créer notre propre schéma d'annotation à l'aide du format xml (Extensive Markup Language) qui permet l'adaptation de l'interface d'annotation. L'annotateur dispose ainsi d'une interface utilisateur intuitive pour annoter chaque séquence du corpus. Une fois la séquence annotée, les descripteurs sélectionnés sont stockés dans un fichier d'interprétation, également sous la forme d'un fichier xml.

Le logiciel Anvil fournit également des facilités de gestion de la base de données constituées des séquences annotées, permettant d'accéder aux séquences répondant à une requête donnée (ex : toutes les séquences présentant de la peur intense éprouvée par un enfant)

[38] http ://www.dfki.de/ kipp/anvil/

Liste des films et time codes des séquences

58 minutes pour vivre (Die hard 2)	
réalisateur :	Renny Harlin
année de production :	1990
séquences extraites :	3
identifiant :	58min
Sixième sens (The Sixth Sense)	
réalisateur :	M. Night Shyalmalan
année de production :	1999
séquences extraites :	5
identifiant :	6sens
Le Projet Blair Witch (The Blair witch project)	
réalisateurs :	Daniel Myrick, Eduardo Sanchez
année de production :	1999
séquences extraites :	8
identifiant :	blairwitch
Les Nerfs à vif (Cape Fear)	
réalisateur :	Martin Scorsese
année de production :	1991
séquences extraites :	11
identifiant :	cape
Cube	
réalisateur :	Vicenzo Natali
année de production :	1997
séquences extraites :	5
identifiant :	cube
Mulholand Drive	
réalisateur :	David Lynch
année de production :	2001
séquences extraites :	22
identifiant :	mulholand

Phone Game (Phone booth)	
réalisateur :	Joël Schumacher
année de production :	2002
séquences extraites :	8
identifiant :	phone
La liste de Schindler (Schindler's List)	
réalisateur :	Steven Spielberg
année de production :	1993
séquences extraites :	12
identifiant :	schindler
Volcano	
réalisateur :	Mick Jackson
année de production :	1997
séquences extraites :	21
identifiant :	volcano
Les affranchis (Goodfellas)	
réalisateur :	Martin Scorsese
année de production :	1990
séquences extraites :	9
identifiant :	affranchis
Dancer in the Dark	
réalisateur :	Lars Von Trier
année de production :	2000
séquences extraites :	24
identifiant :	dancer
Les dents de la mer (Jaws)	
réalisateur :	Steven Spielberg
année de production :	1975
séquences extraites :	18
identifiant :	dents

Dogville	
réalisateur :	Lars Von Trier
année de production :	2002
séquences extraites :	11
identifiant :	dogville
Dommage collatéral (Collateral Damage)	
réalisateur :	Andrew Davis
année de production :	2002
séquences extraites :	10
identifiant :	dommage
Le fugitif (The Fugitive)	
réalisateur :	Andrew Davis
année de production :	1993
séquences extraites :	3
identifiant :	fugitif
Gangs of New York	
réalisateur :	Martin Scorsese
année de production :	2002
séquences extraites :	19
identifiant :	gangs
Hannibal	
réalisateur :	Ridley Scott
année de production :	2000
séquences extraites :	20
identifiant :	hannibal
Dites-leur que je suis un homme (A Lesson Before Dying)	
réalisateur :	Joseph Sargent
année de production :	1999
séquences extraites :	6
identifiant :	homme

Identity	
réalisateur :	James Manglod
année de production :	2002
séquences extraites :	19
identifiant :	identity
Marathon man	
réalisateur :	John Schlesinger
année de production :	1976
séquences extraites :	8
identifiant :	marathon
Un frisson dans la nuit (Play Misty for me)	
réalisateur :	Clint Eastwood
année de production :	1972
séquences extraites :	12
identifiant :	misty
Narc	
réalisateur :	Joe Carnahan
année de production :	2001
séquences extraites :	26
identifiant :	narc
Le pianiste (The Pianist)	
réalisateur :	Roman Polanski
année de production :	2001
séquences extraites :	12
identifiant :	pianist
L'aventure du Poseidon (The Poseidon adventure)	
réalisateur :	Ronald Neame
année de production :	1972
séquences extraites :	16
identifiant :	poseidon

Rose Red	
réalisateur :	Craig R.Baxley
année de production :	2002
séquences extraites :	21
identifiant :	rose
Rosemary's baby	
réalisateur :	Roman Polanski
année de production :	1968
séquences extraites :	18
identifiant :	rose
Seven	
réalisateur :	David Fincher
année de production :	1995
séquences extraites :	11
identifiant :	seven
Shining	
réalisateur :	Stanley Kubrick
année de production :	1980
séquences extraites :	13
identifiant :	shining
Le silence des agneaux (The silence of the Lambs)	
réalisateur :	Jonathan Demme
année de production :	1990
séquences extraites :	5
identifiant :	silence
Speed	
réalisateur :	Jan De Bont
année de production :	1994
séquences extraites :	23
identifiant :	speed

FIG. A.2 – *Liste et références des films du SAFE Corpus*

Film	Séquence	N° Chapitre	Start	Stop	Longueur
A Lesson Before Dying	homme_c2s1	2	00:10,80	01:21,94	01:11,14
A Lesson Before Dying	homme_c3s1	3	04:19,85	06:18,89	01:59,04
A Lesson Before Dying	homme_c4s1	4	01:50,78	03:46,54	01:55,76
A Lesson Before Dying	homme_c7s1	7	04:03,35	06:44,09	02:40,74
A Lesson Before Dying	homme_c9s1	9	03:42,01	04:15,55	00:33,54
A Lesson Before Dying	homme_c11s1	11	01:27,68	03:45,22	02:17,54
Cape Fear	Cape_c2s1	2	02:06,36	02:43,48	00:37,12
Cape Fear	Cape_c3s1	3	00:01,80	01:33,00	01:31,20
Cape Fear	Cape_c4s1	4	01:14,48	02:16,60	01:02,12
Cape Fear	Cape_c6s1	6	00:20,20	01:35,08	01:14,88
Cape Fear	Cape_c7s1	7	03:39,08	04:03,44	00:24,36
Cape Fear	Cape_c8s1	8	02:38,76	03:11,40	00:32,64
Cape Fear	Cape_c9s1	9	00:51,84	01:52,40	01:00,56
Cape Fear	Cape_c11s1	11	09:12,33	10:25,49	01:13,16
Cape Fear	Cape_c14s1	14	10:17,32	13:24,43	03:07,11
Cape Fear	Cape_c16s1	16	00:28,48	02:08,24	01:39,76
Cape Fear	Cape_c17s1	17	00:00,00	00:37,08	00:37,08
Collateral Damage	dommage_c1s1	1	01:10,84	01:39,44	00:28,60
Collateral Damage	dommage_c1s2	1	01:45,88	03:12,56	01:26,68
Collateral Damage	dommage_c1s3	1	03:21,32	04:08,84	00:47,52
Collateral Damage	dommage_c1s4	1	04:08,84	04:37,88	00:29,04
Collateral Damage	dommage_c1s5	1	04:43,00	04:59,92	00:16,92
Collateral Damage	dommage_c3s1	3	06:46,68	07:48,48	01:01,80
Collateral Damage	dommage_c4s1	4	03:46,56	04:50,60	01:04,04
Collateral Damage	dommage_c9s1	9	06:55,24	07:14,64	00:19,40
Collateral Damage	dommage_c9s2	9	05:28,04	08:20,68	02:52,64
Collateral Damage	dommage_c12s1	12	07:16,64	07:53,00	00:36,36
Cube	Cube_c2s1	2	01:10,23	02:15,15	01:04,92
Cube	Cube_c3s1	3	00:05,92	00:15,12	00:09,20
Cube	Cube_c3s2	3	00:23,52	00:49,16	00:25,64
Cube	Cube_c4s1	4	00:00,00	00:14,84	00:14,84
Cube	Cube_c18s1	18	00:00,00	00:56,40	00:56,40
Dancer in the Dark	dancer_c2s1	2	00:21,96	00:57,64	00:35,68
Dancer in the Dark	dancer_c2s2	2	04:14,72	05:24,36	01:09,64
Dancer in the Dark	dancer_c2s3	2	05:39,44	06:04,40	00:24,96
Dancer in the Dark	dancer_c2s4	2	06:04,92	07:26,08	01:21,16
Dancer in the Dark	dancer_c3s1	3	00:00,00	00:24,44	00:24,44
Dancer in the Dark	dancer_c3s2	3	01:03,08	01:50,24	00:47,16
Dancer in the Dark	dancer_c3s3	3	0:05	0:07	01:47,02

Dancer in the Dark	dancer_c4s1	4	00:00,0	01:40,0	01:40,0
Dancer in the Dark	dancer_c4s2	4	01:33,0	03:46,1	02:13,1
Dancer in the Dark	dancer_c4s3	4	06:45,8	07:35,2	00:49,4
Dancer in the Dark	dancer_c5s1	5	00:22,4	01:08,9	00:46,5
Dancer in the Dark	dancer_c5s2	5	01:13,0	02:24,9	01:11,9
Dancer in the Dark	dancer_c6s1	6	14:40,8	15:51,4	01:10,6
Dancer in the Dark	dancer_c8s1	8	01:39,2	04:35,8	02:56,7
Dancer in the Dark	dancer_c8s2	8	04:35,0	10:09,7	05:34,7
Dancer in the Dark	dancer_c10s1	10	00:00,0	00:51,7	00:51,7
Dancer in the Dark	dancer_c10s2	10	03:35,8	04:57,8	01:22,0
Dancer in the Dark	dancer_c12s1	12	01:34,7	02:27,0	00:52,3
Dancer in the Dark	dancer_c12s2	12	02:25,7	06:11,4	03:45,7
Dancer in the Dark	dancer_c14s1	14	00:11,2	01:45,5	01:34,3
Dancer in the Dark	dancer_c14s2	14	05:53,7	08:29,4	02:35,7
Dancer in the Dark	dancer_c15s1	15	01:11,1	02:17,0	01:05,9
Dancer in the Dark	dancer_c16s1	16	03:40,2	06:46,4	03:06,2
Dancer in the Dark	dancer_c16s2	16	09:19,0	11:09,1	01:50,1
Dancer in the Dark	dancer_c17s1	17	01:49,7	04:45,4	02:55,6
Die hard 2	58min_c3s1	3	00:37,6	01:05,6	00:28,0
Die hard 2	58min_c4s1	4	00:05,2	01:18,0	01:12,8
Die hard 2	58min_c12s1	12	06:37,1	06:50,6	00:13,5
Dogville	dogville_c2s1	2	03:21,0	04:22,1	01:01,1
Dogville	dogville_c4s1	4	04:51,9	06:38,4	01:46,5
Dogville	dogville_c11s1	11	00:59,3	02:08,0	01:08,7
Dogville	dogville_c11s2	11	02:07,2	03:58,9	01:51,7
Dogville	dogville_c13s1	13	01:15,2	03:12,3	01:57,1
Dogville	dogville_c16s1	16	00:08,2	01:24,2	01:16,0
Dogville	dogville_c18s1	18	06:03,6	07:38,1	01:34,5
Dogville	dogville_c18s2	18	07:35,8	09:05,4	01:29,6
Dogville	dogville_c19s1	19	05:03,7	06:00,4	00:56,7
Dogville	dogville_c19s2	19	07:33,1	08:49,1	01:16,0
Dogville	dogville_c19s3	19	09:15,0	10:16,2	01:01,3
Gangs of New York	gangs_c2s1	2	03:53,8	05:54,2	02:00,4
Gangs of New York	gangs_c4s1	4	00:00,0	01:01,5	01:01,5
Gangs of New York	gangs_c4s2	4	01:02,3	01:48,8	00:46,5
Gangs of New York	gangs_c4s3	4	03:49,0	04:30,3	00:41,3
Gangs of New York	gangs_c4s4	4	07:09,2	07:47,2	00:38,0
Gangs of New York	gangs_c5s1	5	00:00,0	00:52,2	00:52,2
Gangs of New York	gangs_c5s2	5	06:19,8	07:04,1	00:44,3
Gangs of New York	gangs_c6s1	6	00:33,0	01:32,9	01:00,0

Gangs of New York	gangs_c6s2	6	04:58,4	05:44,7	00:46,2
Gangs of New York	gangs_c7s1	7	02:50,9	04:27,5	01:36,6
Gangs of New York	gangs_c11s1	11	05:44,7	06:01,7	00:17,0
Gangs of New York	gangs_c11s2	11	06:17,8	08:18,4	02:00,6
Gangs of New York	gangs_c12s1	12	03:22,6	06:22,4	02:59,8
Gangs of New York	gangs_c13s1	13	00:00,0	01:32,6	01:32,6
Gangs of New York	gangs_c14s1	14	00:31,8	01:02,0	00:30,3
Gangs of New York	gangs_c14s2	14	04:58,0	06:59,3	02:01,3
Gangs of New York	gangs_c15s1	15	00:49,6	01:52,2	01:02,6
Gangs of New York	gangs_c15s2	15	02:33,4	03:22,6	00:49,2
Gangs of New York	gangs_c17s1	17	01:20,8	02:49,6	01:28,8
Goodfellas	affranchis_c3s1	3	02:35,9	03:03,0	00:27,2
Goodfellas	affranchis_c4s1	4	01:58,1	02:08,3	00:10,1
Goodfellas	affranchis_c7s1	7	00:54,8	01:23,6	00:28,8
Goodfellas	affranchis_c8s1	8	00:00,0	00:10,8	00:10,8
Goodfellas	affranchis_c13s1	13	00:23,9	01:34,6	01:10,7
Goodfellas	affranchis_c13s2	13	01:50,7	02:25,5	00:34,7
Goodfellas	affranchis_c13s3	13	02:50,0	03:18,7	00:28,8
Goodfellas	affranchis_c16s1	16	00:28,2	01:00,5	00:32,3
Goodfellas	affranchis_c19s1	19	00:00,0	01:43,5	01:43,5
Hannibal	hannibal_c3s1	3	00:28,5	01:08,7	00:40,2
Hannibal	hannibal_c3s2	3	04:20,8	04:47,0	00:26,2
Hannibal	hannibal_c4s1	4	00:00,0	00:16,8	00:16,8
Hannibal	hannibal_c4s2	4	01:50,8	03:03,7	01:12,8
Hannibal	hannibal_c7s1	7	00:11,8	00:41,4	00:29,6
Hannibal	hannibal_c9s1	9	01:31,5	02:17,9	00:46,4
Hannibal	hannibal_c11s1	11	00:07,6	01:04,4	00:56,9
Hannibal	hannibal_c12s1	12	00:37,6	01:57,5	01:20,0
Hannibal	hannibal_c14s1	14	00:00,0	01:07,4	01:07,4
Hannibal	hannibal_c15s1	15	03:08,6	04:11,4	01:02,8
Hannibal	hannibal_c17s1	17	02:14,6	02:57,8	00:43,2
Hannibal	hannibal_c18s1	18	01:32,0	03:13,1	01:41,1
Hannibal	hannibal_c20s1	20	00:27,6	01:25,7	00:58,1
Hannibal	hannibal_c20s2	20	02:23,3	03:45,8	01:22,5
Hannibal	hannibal_c21s1	21	03:12,6	04:07,6	00:55,0
Hannibal	hannibal_c21s2	21	04:16,7	04:50,7	00:34,0
Hannibal	hannibal_c25s1	25	00:53,9	01:43,5	00:49,6
Hannibal	hannibal_c29s1	29	06:07,0	07:39,2	01:32,2
Hannibal	hannibal_c29s2	29	07:43,0	08:26,2	00:43,2
Hannibal	hannibal_c30s1	30	00:17,0	01:19,1	01:02,1

Identity	identity_c2s1	2	02:17,8	02:38,7	00:20,9
Identity	identity_c2s2	2	02:58,1	03:19,4	00:21,3
Identity	identity_c2s4	2	03:37,8	05:13,9	01:36,1
Identity	identity_c2s5	2	05:37,2	06:47,8	01:10,6
Identity	identity_c3s1	3	00:13,9	00:33,6	00:19,7
Identity	identity_c3s2	3	00:48,6	01:07,0	00:18,4
Identity	identity_c5s1	5	00:03,0	01:18,9	01:15,9
Identity	identity_c6s1	6	00:55,2	01:50,5	00:55,2
Identity	identity_c10s1	10	00:00,0	01:43,0	01:43,0
Identity	identity_c10s2	10	03:49,6	04:03,6	00:13,9
Identity	identity_c11s1	11	01:43,9	02:34,4	00:50,6
Identity	identity_c12s1	12	00:46,8	01:56,3	01:09,5
Identity	identity_c18s1	18	00:21,0	00:46,1	00:25,1
Identity	identity_c22s1	22	00:00,0	01:02,5	01:02,5
Identity	identity_c26s1	26	00:46,9	01:35,4	00:48,4
Identity	identity_c26s2	26	01:56,4	02:40,7	00:44,3
Identity	identity_c27s1	27	00:13,6	00:55,6	00:41,9
Identity	identity_c27s2	27	00:55,6	02:04,5	01:09,0
Jaws	dents_c2s1	2	00:35,4	01:20,5	00:45,1
Jaws	dents_c2s2	2	02:04,0	02:57,8	00:53,8
Jaws	dents_c2s3	2	03:40,3	04:13,3	00:33,0
Jaws	dents_c5s1	5	00:00,0	00:26,2	00:26,2
Jaws	dents_c5s2	5	00:20,7	00:40,4	00:19,8
Jaws	dents_c5s3	5	02:00,0	02:51,1	00:51,2
Jaws	dents_c5s4	5	03:15,4	04:17,5	01:02,1
Jaws	dents_c6s1	6	04:23,7	05:02,9	00:39,2
Jaws	dents_c6s2	6	05:02,9	06:02,7	00:59,8
Jaws	dents_c7s1	7	01:40,7	03:03,2	01:22,4
Jaws	dents_c9s1	9	00:30,7	00:45,2	00:14,5
Jaws	dents_c9s2	9	01:35,2	03:13,1	01:38,0
Jaws	dents_c11s1	11	03:13,3	03:28,1	00:14,8
Jaws	dents_c11s2	11	04:25,6	04:40,6	00:15,0
Jaws	dents_c11s3	11	05:37,2	05:50,3	00:13,1
Jaws	dents_c11s4	11	06:01,9	07:11,6	01:09,6
Jaws	dents_c12s1	12	00:00,0	00:28,7	00:28,7
Jaws	dents_c12s2	12	00:33,1	02:18,8	01:45,7
Marathon man	marathon_c8s1	8	01:31,8	02:17,4	00:45,7
Marathon man	marathon_c12s1	12	00:29,8	01:20,6	00:50,8
Marathon man	marathon_c18s1	18	00:15,4	00:50,0	00:34,6
Marathon man	marathon_c25s1	25	00:30,0	00:55,7	00:25,7

Marathon man	marathon_c25s2	25	00:55,7	01:39,8	00:44,1
Marathon man	marathon_c28s1	28	00:45,6	01:28,8	00:43,2
Marathon man	marathon_c32s1	32	01:30,4	02:36,1	01:05,7
Marathon man	marathon_c33s1	33	01:25,6	04:11,1	02:45,5
Mulholand Drive	Mulho_c3s1	3	02:25,5	03:04,4	00:39,0
Mulholand Drive	Mulho_c7s1	7	00:00,0	00:41,8	00:41,8
Mulholand Drive	Mulho_c7s2	7	00:41,8	02:40,3	01:58,5
Mulholand Drive	Mulho_c12s1	12	00:43,7	01:19,0	00:35,2
Mulholand Drive	Mulho_c13s1	13	00:42,8	01:43,4	01:00,6
Mulholand Drive	Mulho_c15s1	15	00:00,0	01:15,2	01:15,2
Mulholand Drive	Mulho_c16s1	16	00:14,0	01:51,4	01:37,4
Mulholand Drive	Mulho_c18s1	18	01:50,1	02:33,3	00:43,2
Mulholand Drive	Mulho_c21s1	21	01:37,9	03:30,2	01:52,4
Mulholand Drive	Mulho_c22s1	22	01:03,9	02:08,8	01:05,0
Mulholand Drive	Mulho_c23s1	23	00:00,0	01:12,7	01:12,7
Mulholand Drive	Mulho_c26s1	26	02:04,1	02:57,4	00:53,3
Mulholand Drive	Mulho_c28s1	28	00:00,0	01:02,4	01:02,4
Mulholand Drive	Mulho_c29s1	29	00:00,0	00:51,0	00:51,0
Mulholand Drive	Mulho_c30s1	30	00:10,4	01:35,1	01:24,7
Mulholand Drive	Mulho_c31s1	31	00:00,0	02:09,1	02:09,1
Mulholand Drive	Mulho_c35s1	35	01:27,6	01:48,9	00:21,3
Mulholand Drive	Mulho_c37s1	37	00:10,0	00:44,2	00:34,2
Mulholand Drive	Mulho_c37s2	37	03:19,8	04:44,0	01:24,2
Mulholand Drive	Mulho_c37s3	37	08:42,7	09:31,9	00:49,2
Mulholand Drive	Mulho_c45s1	45	00:00,0	00:36,2	00:36,2
Mulholand Drive	Mulho_c45s2	45	01:11,1	01:40,3	00:29,2
Narc	narc_c1s1	1	01:40,6	03:32,9	01:52,3
Narc	narc_c2s1	2	01:54,6	02:33,2	00:38,6
Narc	narc_c2s2	2	05:59,1	06:08,4	00:09,2
Narc	narc_c3s1	3	00:49,6	01:52,1	01:02,5
Narc	narc_c3s2	3	02:00,5	03:18,4	01:17,9
Narc	narc_c3s3	3	03:19,1	03:49,6	00:30,4
Narc	narc_c3s4	3	03:49,3	06:04,4	02:15,1
Narc	narc_c4s1	4	02:17,8	03:23,4	01:05,6
Narc	narc_c5s1	5	04:17,3	06:15,9	01:58,6
Narc	narc_c7s1	7	00:13,3	00:54,5	00:41,2
Narc	narc_c7s2	7	02:24,8	04:45,6	02:20,8
Narc	narc_c7s3	7	05:21,6	06:09,2	00:47,6
Narc	narc_c8s1	8	02:54,9	04:27,3	01:32,4
Narc	narc_c10s1	10	03:19,1	03:27,9	00:08,8

Narc	narc_c10s2	10	03:48,0	05:09,8	01:21,8
Narc	narc_c11s1	11	02:06,4	02:50,5	00:44,1
Narc	narc_c11s2	11	05:58,5	07:16,5	01:18,0
Narc	narc_c12s1	12	00:00,0	00:26,8	00:26,8
Narc	narc_c12s2	12	00:28,6	00:49,2	00:20,5
Narc	narc_c12s3	12	00:52,2	01:40,4	00:48,2
Narc	narc_c12s4	12	04:33,6	05:30,5	00:56,9
Narc	narc_c13s1	13	00:00,0	05:30,5	05:30,5
Narc	narc_c13s2	13	04:27,0	06:11,6	01:44,7
Narc	narc_c14s1	14	00:34,1	02:06,8	01:32,6
Narc	narc_c14s2	14	02:06,2	03:05,3	00:59,1
Narc	narc_c14s3	14	03:31,1	05:07,3	01:36,2
Phone booth	Phone_c3s1	3	00:18,3	01:06,1	00:47,8
Phone booth	Phone_c4s1	4	00:04,7	00:53,9	00:49,2
Phone booth	Phone_c5s1	5	00:00,0	00:39,7	00:39,7
Phone booth	Phone_c10s1	10	02:49,3	03:11,2	00:21,9
Phone booth	Phone_c13s1	13	00:08,0	01:07,5	00:59,6
Phone booth	Phone_c17s1	17	02:50,1	03:41,0	00:51,0
Phone booth	Phone_c21s1	21	01:07,9	01:31,1	00:23,2
Phone booth	Phone_c24s1	24	01:27,6	01:41,6	00:14,0
Play Misty for me	misty_c2s1	2	00:54,2	01:30,4	00:36,2
Play Misty for me	misty_c2s2	2	03:15,7	04:04,9	00:49,2
Play Misty for me	misty_c3s1	3	00:40,3	01:34,7	00:54,4
Play Misty for me	misty_c3s2	3	04:47,1	05:34,0	00:46,9
Play Misty for me	misty_c4s1	4	01:27,8	02:04,7	00:36,9
Play Misty for me	misty_c5s1	5	03:49,5	04:51,7	01:02,2
Play Misty for me	misty_c5s2	5	04:52,0	06:11,0	01:19,0
Play Misty for me	misty_c8s1	8	04:21,7	05:08,4	00:46,7
Play Misty for me	misty_c10s1	10	02:32,2	03:49,3	01:17,1
Play Misty for me	misty_c12s1	12	01:05,6	01:33,5	00:27,8
Play Misty for me	misty_c15s1	15	04:10,2	05:15,2	01:05,0
Play Misty for me	misty_c17s1	17	08:10,3	08:58,0	00:47,7
Rose Red	rose_c1s1	1	03:39,2	04:15,5	00:36,2
Rose Red	rose_c1s2	1	05:28,5	06:33,1	01:04,6
Rose Red	rose_c3s1	3	03:44,1	05:21,0	01:36,9
Rose Red	rose_c7s1	7	03:41,6	05:08,6	01:27,0
Rose Red	rose_c7s2	7	05:30,1	06:02,5	00:32,4
Rose Red	rose_c11s1	11	00:17,7	02:58,0	02:40,3
Rose Red	rose_c11s2	11	06:15,2	07:46,7	01:31,5
Rose Red	rose_c14s1	14	02:09,0	02:48,4	00:39,5

Rose Red	rose_c14s2	14	03:26,1	04:05,2	00:39,1
Rose Red	rose_c14s3	14	05:55,3	06:57,9	01:02,6
Rose Red	rose_c16s1	16	00:00,7	01:04,5	01:03,8
Rose Red	rose_c16s2	16	03:05,7	04:03,8	00:58,0
Rose Red	rose_c19s1	19	01:34,4	02:25,4	00:51,0
Rose Red	rose_c19s2	19	06:09,7	06:59,9	00:50,2
Rose Red	rose_c21s1	21	02:18,2	03:42,5	01:24,3
Rose Red	rose_c21s2	21	04:23,7	05:13,9	00:50,2
Rose Red	rose_c22s1	22	04:32,2	05:56,8	01:24,6
Rose Red	rose_c24s1	24	00:26,5	00:51,6	00:25,1
Rose Red	rose_c24s1	24	01:23,9	01:41,1	00:17,2
Rose Red	rose_c24s3	24	03:02,2	03:18,3	00:16,0
Rose Red	rose_c24s4	24	04:41,8	06:14,1	01:32,3
Rosemary's baby	rosemary_c2s1	2	00:26,0	01:02,6	00:36,7
Rosemary's baby	rosemary_c2s2	2	01:35,2	02:16,0	00:40,8
Rosemary's baby	rosemary_c2s3	2	04:12,4	04:30,2	00:17,9
Rosemary's baby	rosemary_c5s1	5	00:00,0	01:13,7	01:13,7
Rosemary's baby	rosemary_c5s2	5	03:27,1	04:13,8	00:46,6
Rosemary's baby	rosemary_c6s1	6	00:39,2	01:27,1	00:47,9
Rosemary's baby	rosemary_c6s2	6	02:41,2	04:10,1	01:28,8
Rosemary's baby	rosemary_c6s3	6	04:10,1	04:10,1	00:00,0
Rosemary's baby	rosemary_c6s4	6	05:07,6	05:49,4	00:41,8
Rosemary's baby	rosemary_c9s1	9	00:19,2	01:04,5	00:45,2
Rosemary's baby	rosemary_c14s1	14	02:53,3	03:11,8	00:18,6
Rosemary's baby	rosemary_c27s1	27	03:15,8	03:43,2	00:27,4
Rosemary's baby	rosemary_c27s2	27	04:13,8	04:34,6	00:20,8
Rosemary's baby	rosemary_c27s3	27	04:56,8	06:21,8	01:25,0
Rosemary's baby	rosemary_c27s4	27	07:45,0	08:17,7	00:32,7
Rosemary's baby	rosemary_c28s1	28	06:14,3	06:36,6	00:22,3
Rosemary's baby	rosemary_c28s2	28	06:57,7	07:24,8	00:27,2
Rosemary's baby	rosemary_c28s3	28	07:26,3	08:45,8	01:19,5
Schindler's List	Schindler_c3s1	3	00:39,9	01:44,4	01:04,6
Schindler's List	Schindler_c5s1	5	01:44,2	02:15,9	00:31,7
Schindler's List	Schindler_c5s2	5	02:36,3	02:55,0	00:18,8
Schindler's List	Schindler_c9s1	9	00:00,0	01:03,9	01:03,9
Schindler's List	Schindler_c9s2	9	01:15,8	02:40,6	01:24,7
Schindler's List	Schindler_c11s1	11	02:12,0	03:25,6	01:13,7
Schindler's List	Schindler_c14s1	14	01:56,6	03:28,4	01:31,8
Schindler's List	Schindler_c14s2	14	06:12,9	07:13,5	01:00,6
Schindler's List	Schindler_c15s1	15	01:16,8	02:40,6	01:23,7

Schindler's List	Schindler_c16s1	16	05:45,1	06:19,6	00:34,5
Schindler's List	Schindler_c20s1	20	01:05,2	03:22,7	02:17,5
Schindler's List	Schindler_c20s2	20	03:21,6	04:31,1	01:09,4
Seven	seven_c2s1	2	01:30,3	02:58,9	01:28,6
Seven	seven_c13s1	13	03:13,0	04:04,8	00:51,8
Seven	seven_c15s1	15	00:25,2	01:18,7	00:53,5
Seven	seven_c22s1	22	00:40,6	02:45,5	02:04,9
Seven	seven_c23s1	23	00:00,0	02:45,5	02:45,5
Seven	seven_c27s1	27	01:07,6	01:17,1	00:09,6
Seven	seven_c27s2	27	01:41,2	01:51,9	00:10,8
Seven	seven_c27s3	27	02:10,2	02:49,2	00:39,0
Seven	seven_c30s1	30	00:36,0	01:32,0	00:56,0
Seven	seven_c34s1	34	00:52,5	01:34,0	00:41,4
Seven	seven_c35s1	35	00:21,5	03:25,8	03:04,3
Shining	shining_c2s1	2	00:33,0	01:06,3	00:33,3
Shining	shining_c2s2	2	01:13,8	01:59,3	00:45,5
Shining	shining_c6s1	6	00:08,7	03:13,6	03:04,9
Shining	shining_c11s1	11	01:40,4	03:21,7	01:41,3
Shining	shining_c12s1	12	03:22,3	05:10,6	01:48,3
Shining	shining_c18s1	18	01:02,5	04:00,2	02:57,7
Shining	shining_c21s1	21	00:09,1	00:53,4	00:44,3
Shining	shining_c22s1	22	00:03,3	02:16,3	02:13,0
Shining	shining_c29s1	29	00:00,0	01:30,0	01:30,0
Shining	shining_c29s2	29	02:00,7	05:36,3	03:35,6
Shining	shining_c30s1	30	01:25,9	03:55,6	02:29,7
Shining	shining_c33s1	33	01:38,8	03:23,6	01:44,8
Shining	shining_c34s1	34	00:35,1	00:56,0	00:20,9
Speed	speed_c2s1	2	00:00,0	00:36,0	00:36,0
Speed	speed_c2s2	2	01:20,6	01:36,0	00:15,4
Speed	speed_c4s1	4	00:55,8	02:05,2	01:09,4
Speed	speed_c6s1	6	00:00,0	01:46,6	01:46,6
Speed	speed_c8s1	8	01:05,2	02:01,4	00:56,2
Speed	speed_c8s2	8	02:01,9	03:13,6	01:11,6
Speed	speed_c9s1	9	00:17,8	00:47,6	00:29,8
Speed	speed_c9s2	9	00:56,9	01:37,6	00:40,7
Speed	speed_c10s1	10	00:43,5	01:25,6	00:42,0
Speed	speed_c10s2	10	02:49,9	03:07,2	00:17,3
Speed	speed_c11s1	11	05:30,6	07:18,6	01:48,0
Speed	speed_c12s1	12	00:00,0	01:25,3	01:25,3
Speed	speed_c12s2	12	01:46,6	02:56,6	01:10,0

Speed	speed_c13s1	13	02:05,4	02:58,2	00:52,8
Speed	speed_c14s1	14	02:32,5	03:51,3	01:18,8
Speed	speed_c15s1	15	00:36,9	01:56,0	01:19,1
Speed	speed_c16s1	16	00:29,8	00:52,3	00:22,5
Speed	speed_c16s2	16	01:43,0	02:12,4	00:29,4
Speed	speed_c17s1	17	00:05,5	00:37,2	00:31,7
Speed	speed_c17s2	17	01:24,0	03:19,0	01:55,0
Speed	speed_c18s1	18	01:22,4	02:03,2	00:40,8
Speed	speed_c26s1	26	03:26,4	05:05,7	01:39,3
Speed	speed_c31s1	31	00:00,0	00:59,2	00:59,2
The Blair witch project	Blairwitch_c2s1	2	00:12,1	00:32,9	00:20,8
The Blair witch project	Blairwitch_c2s2	2	00:33,1	00:53,9	00:20,8
The Blair witch project	Blairwitch_c2s3	2	06:03,0	06:31,4	00:28,4
The Blair witch project	Blairwitch_c4s1	4	00:00,0	00:21,3	00:21,3
The Blair witch project	Blairwitch_c5s1	5	02:09,9	03:17,6	01:07,8
The Blair witch project	Blairwitch_c5s2	5	05:40,4	06:42,9	01:02,5
The Blair witch project	Blairwitch_c6s1	6	06:16,9	06:52,6	00:35,7
The Blair witch project	Blairwitch_c8s1	8	00:00,0	00:38,2	00:38,2
The Fugitive	fugitif_c5s1	5	00:33,1	01:12,7	00:39,6
The Fugitive	fugitif_c19s1	19	01:17,2	02:25,2	01:08,0
The Fugitive	fugitif_c36s1	36	02:34,2	03:47,0	01:12,8
The Pianist	pianist_c2s1	2	00:16,9	01:22,7	01:05,8
The Pianist	pianist_c3s1	3	00:07,8	01:15,8	01:08,0
The Pianist	pianist_c3s2	3	02:46,4	04:08,7	01:22,2
The Pianist	pianist_c3s3	3	04:02,3	05:32,2	01:29,9
The Pianist	pianist_c4s1	4	00:00,0	01:08,2	01:08,2
The Pianist	pianist_c4s2	4	01:16,0	02:28,5	01:12,5
The Pianist	pianist_c4s3	4	08:24,2	09:34,9	01:10,8
The Pianist	pianist_c4s4	4	11:23,3	12:05,8	00:42,5
The Pianist	pianist_c6s1	6	00:50,4	02:05,3	01:15,0
The Pianist	pianist_c7s1	7	03:21,6	05:14,7	01:53,1
The Pianist	pianist_c9s1	9	00:20,1	01:40,4	01:20,3
The Pianist	pianist_c11s1	11	05:31,2	06:53,9	01:22,7
The Poseidon adventure	poseidon_c4s1	4	00:15,0	01:17,0	01:01,9
The Poseidon adventure	poseidon_c8s1	8	00:50,4	02:24,2	01:33,8
The Poseidon adventure	poseidon_c8s2	8	03:18,8	04:03,0	00:44,2
The Poseidon adventure	poseidon_c8s3	8	04:03,0	04:36,0	00:33,0
The Poseidon adventure	poseidon_c9s1	9	00:00,0	00:51,2	00:51,2
The Poseidon adventure	poseidon_c9s2	9	00:55,2	01:17,9	00:22,7
The Poseidon adventure	poseidon_c10s1	10	01:07,1	01:15,5	00:08,4

The Poseidon adventure	poseidon_c10s2	10	01:29,5	02:07,9	00:38,4
The Poseidon adventure	poseidon_c11s1	11	00:26,8	01:10,4	00:43,6
The Poseidon adventure	poseidon_c11s2	11	01:18,6	01:50,0	00:31,4
The Poseidon adventure	poseidon_c12s1	12	06:04,8	07:32,4	01:27,7
The Poseidon adventure	poseidon_c13s1	13	02:43,5	03:11,1	00:27,6
The Poseidon adventure	poseidon_c15s1	15	01:45,6	02:59,0	01:13,4
The Poseidon adventure	poseidon_c17s1	17	04:25,0	05:44,4	01:19,4
The Poseidon adventure	poseidon_c19s1	19	09:03,8	10:14,9	01:11,1
The Poseidon adventure	poseidon_c23s1	23	00:52,8	02:08,4	01:15,6
The silence of the Lambs	silence_c15s1	15	00:00,0	01:40,5	01:40,5
The silence of the Lambs	silence_c21s1	21	00:01,0	00:47,6	00:46,6
The silence of the Lambs	silence_c29s1	29	05:49,4	07:13,7	01:24,4
The silence of the Lambs	silence_c30s1	30	01:08,7	01:53,3	00:44,6
The silence of the Lambs	silence_c30s2	30	03:53,8	04:21,6	00:27,9
The Sixth Sense	6sens_c2s1	2	01:18,5	01:48,8	00:30,3
The Sixth Sense	6sens_c2s2	2	04:33,3	05:18,9	00:45,6
The Sixth Sense	6sens_c2s3	2	05:19,1	06:19,4	01:00,3
The Sixth Sense	6sens_c11s2	11	03:32,2	04:33,1	01:00,9
The Sixth Sense	6sens_c9s1	9	02:09,3	03:42,4	01:33,0
Volcano	Volcano_c3s1	3	00:09,0	00:43,6	00:34,6
Volcano	Volcano_c3s2	3	01:58,8	02:38,9	00:40,2
Volcano	Volcano_c4s1	4	00:39,0	01:39,6	01:00,6
Volcano	Volcano_c4s2	4	03:05,9	04:01,4	00:55,5
Volcano	Volcano_c5s1	5	00:17,9	01:07,2	00:49,4
Volcano	Volcano_c6s1	6	00:20,0	01:27,4	01:07,3
Volcano	Volcano_c8s1	8	04:25,3	05:00,0	00:34,7
Volcano	Volcano_c9s1	9	00:47,7	00:55,6	00:07,9
Volcano	Volcano_c9s2	9	02:36,2	02:58,6	00:22,4
Volcano	Volcano_c10s1	10	01:21,6	02:18,4	00:56,8
Volcano	Volcano_c10s2	10	04:17,6	05:13,2	00:55,6
Volcano	Volcano_c12s1	12	00:20,9	02:26,2	02:05,3
Volcano	Volcano_c13s1	13	03:42,0	04:49,7	01:07,6
Volcano	Volcano_c13s2	13	06:08,1	06:42,8	00:34,8
Volcano	Volcano_c13s3	13	06:42,2	08:32,4	01:50,2
Volcano	Volcano_c14s1	14	05:08,3	05:33,5	00:25,2
Volcano	Volcano_c14s2	14	05:33,8	06:20,9	00:47,1
Volcano	Volcano_c17s1	17	00:02,3	02:00,4	01:58,1
Volcano	Volcano_c19s1	19	00:32,2	01:27,6	00:55,5
Volcano	Volcano_c19s2	19	01:43,7	02:01,6	00:17,8
Volcano	Volcano_c20s1	20	00:00,6	00:18,0	00:17,4

FIG. A.3 – *Séquences, chapitres et time codes du SAFE Corpus.*

Annexe B

Normes de transcription

Les normes de transcription utilisée pour la transcription de la bande audio sont adaptées de la norme définie par le LDC (Linguistique Data Consortium[39]) et sont récapitulées ci-dessous :

Abréviations

L'annotateur ne doit pas utiliser d'abréviations dans sa transcription. Par exemple, dans la phrase suivante, « I went to the doctor, and all he said was, don't worry, it's natural. », l'annotateur transcrit « doctor » et non « Dr ».

Mots

Les mots mal prononcés par le locuteur doivent être écrits correctement sauf si le locuteur a utilisé un autre mot.

Noms propres

Les noms propres doivent porter un majuscule en début de mot (ex : John)

Adresse internet

Les adresses internet seront transcrites comme dans l'exemple suivant : « www point google point F R »

Nombres

Les nombres ne doivent pas être transcrits en toute lettre mais en chiffre (ex : 60000)

Acronymes

Les acronymes doivent être écrits en majuscules sans espace et précédés du symbole « ». Exemples : FBI.

[39] http ://www.ldc.upenn.edu/

Bruits

Lors de la transcription, les phénomènes sonores tels que les toux, les respirations, la parole inintelligible, etc., sont pris en compte par l'utilisation d'accolades. Ainsi le son produit par le locuteur est transcrit en utilisant la liste suivante de sons sélectionnés ici en fonction de leur pertinence pour notre application.

- L : rires
- Co : toux
- B : respirations
- C : cris
- T : larmes
- S : coups de feu
- NN : autre son produit par le locuteur et non répertorié dans la liste ci-dessus

Si le locuteur produit ces sons alors qu'il est en train de parler, l'annotateur pourra utiliser la liste précédente pour insérer le texte prononcé entre des balises comme dans l'exemple qui suit : « C}Help/C »

Parole inintelligible

L'annotateur utilisera des parenthèses pour signaler les mots ou groupe de mots inintelligibles qu'il aura transcrit intuitivement : « Well, I ((thought)) that it was fine. ». Dans le cas où la parole prononcée de manière inintelligible ne peut pas être devinée, l'annotateur utilisera les parenthèse seule : « (( )) ».

Mots tronqués

Pour les mots tronques qui correspondrait par exemple à des marques de répétitions et d'hésitations, l'annotateur pourra se référer à l'exemple suivant : « I went to the ju- junior league game »

Interjections

Ci-dessous les codes utilisés pour transcrire les différentes interjections de l'anglais :

- mhm
- uh-huh
- uh-oh
- okay
- whoa
- whew
- yeah
- jeeze

Lorsque les interjections sont effectués par une autre personne que le locuteur pendant son intervention, l'annotateur pourra mettre l'interjection entre accolades, ex : {mhm}. À cette liste s'ajoutent les sons utilisés par le locuteur lorsqu'il hésite. Ces sons ne sont pas considérés contrairement aux interjections ci-dessus comme des mots et leur transcription est précédée du symbole « % ».

- %ach

- %ah
- %eee
- %eh
- %ew
- %ha
- %hee
- %huh
- %hm
- %huh
- %um
- %uh
- %oh

Annexe C

Validation des résultats par les SVM

Le système présenté dans le chapitre 6 utilise la méthode de classification des GMM (*Gaussian Mixture Models*) pour la discrimination peur/neutre. Nous avons jugé important de tester une autre méthode de classification, les SVM (*Support Vector Machines*) afin de valider la cohérence des résultats obtenus.

Méthode

Les machines à vecteurs supports ou SVM ont été introduits par Vapnik en 1995 [Vapnik, 1995]. Cette méthode est donc une alternative récente pour la classification. Intialement prévue pour résoudre des problèmes de classification à deux classes, il existe aujourd'hui des généralisations multi-classe [Hsu et Lin, 2002].

Le principe des SVM tels qu'utilisés ici peut se résumer de la manière suivante : pour deux classes d'exemples donnés, il s'agit de séparer les exemples de chaque classe par un hyperplan en maximisant la distance des exemples d'apprentissage à l'hyperplan. Les points les plus proches de l'hyperplan, qui seuls sont utilisés pour la détermination de l'hyperplan sont appelés vecteurs de support (cf figure C.1). La distance que l'on cherche à maximiser est la distance minimale entre l'hyperplan et les exemples d'apprentissage. Cette distance est appelée « marge ».

Parmi les modèles SVM on distingue le cas linéairement séparable et le cas non linéairement séparable. Pour surmonter les inconvénients du cas non linéairement séparable, l'idée des SVM est de transformer l'espace des données, afin de passer d'un problème de séparation non linéaire à un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est effectuée via une fonction, dite fonction noyau.

L'un des paramètres à fixer lors de l'implémentation des SVM va donc être le choix du noyau. Nous avons testé ici deux types de noyaux : le noyau gaussien,

$$k(x,y) = \exp(\frac{-\left\|x-y\right\|^2}{2\sigma^2})$$

et le noyau linéaire

$$k(x,y) = x.y.$$

Dans le cas du noyau gaussien le choix de σ la variance du noyau gaussien permet de régler la précision de l'hyperplan séparateur. Un σ petit signifie que l'hyperplan séparateur est très proche des vecteurs support et risque le surapprentissage, un σ très grand correspond à une situation où l'hyperplan séparateur devient linéaire, et peut entrainer un sous apprentissage.

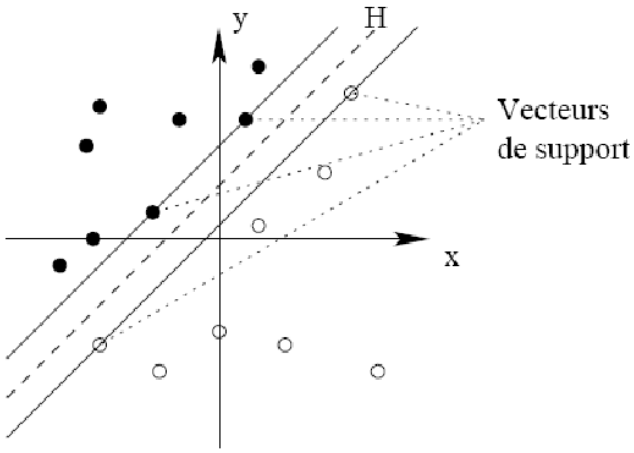

FIG. C.1 – *Hyperplan séparateur et vecteurs de support dans un espace à deux dimensions*

La classification d'un nouvel exemple de test est donnée par sa position dans l'espace de redescription par rapport à l'hyperplan optimal défini lors de l'apprentissage. Les SVM fournissent une distance à l'hyperplan dont le signe détermine la classe de l'exemple testé.

Résultats

Les SVM ont été évalués pour une classification peur/neutre sur le corpus SAFE. Plusieurs paramétrages des SVM ont été testés pour ce problème dans [Antoniou, 2006]. Le tableau C.1 présente les meilleurs résultats qui ont été obtenus avec le noyau gaussien, pour $\sigma = 10$.

manuel \ automatique	Neutre	Peur
Neutre	73%	27%
Peur	30%	70%
TEM	29%	

TAB. C.1 – *Matrice de confusion pour le système SVM de classification peur vs. neutre sur SAFE_1 (tableau 5.2) (intervalles de confiance à 95% près : rayon ≤ 3%)*

Les performances obtenues avec les GMM dans les mêmes conditions que les SVM, i.e. en n'utilisant que le classificateur voisé, sont présentées dans le tableau C.2. Les performances obte-

manuel \ automatique	Neutre	Peur
Neutre	71%	29%
Peur	31%	69%
TEM	30%	

TAB. C.2 – *Matrice de confusion pour le système de classification GMM peur vs. neutre sur SAFE_1 (tableau 5.2) (intervalles de confiance à 95% près : rayon ≤ 3%)*

nues par les deux classificateurs sont similaires (29% contre 30% avec les GMM) aux intervalles de confiance à 95% près. Ceci nous permet de valider la cohérence des résultats obtenus avec les GMM.

Glossaire

Contenu voisé	Portions du signal de parole qui ont été produites avec vibration des cordes vocales. Ces portions contiennent les voyelles et les consonnes voisées.
Contenu non voisé	Portions du signal de parole qui ont été produites sans vibration des cordes vocales. Ces portions contiennent les consonnes non voisées, les aspirations, les respirations, mais également des voyelles chuchotées ou prononcées sans voisement.
Corpus SAFE	Situation Analysis in a Fictional and Emotional Corpus
EM	Expectation Maximization
État neutre	État de valence (non-négatif et non-positif) d'activation faible et de contrôle élevé.
FDR	Fisher Discriminant Ratio
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
NAQ	Normalized Amplitude Quotient
Paralinguistique	Informations de type acoustique, prosodique, incluent également les manifestations non verbales (cris, rire, etc.).
Real-life	Se dit des bases de données illustrant des émotions vécues en contextes réels.
SAP	Score A Posteriori
Segment	Tour de parole ou partie du tour de parole avec un contenu émotionnel homogène. Ainsi dans le cas d'un long monologue présentant des ruptures dans les manifestations émotionnelles, le tour de parole sera segmenté en plusieurs parties.

Séquence	Portion de film référant à un même contexte situationnel (ex : prise d'otage, inondation du métro, etc.).
Situations anormales	Événements imprévus, constituant une menace pour la vie humaine, qu'ils soient la conséquence d'une catastrophe naturelle (incendies, inondations) ou d'une action humaine (agressions physiques ou psychologiques contre un être humain, prises d'otage, attentats terroristes, etc.).
SVM	Support Vector Machine
TEE	Taux d'Égale Erreur

Table des figures

Liste des tableaux

Bibliographie

[Abelin et Allwood, 2000] A. Abelin et J. Allwood. Cross linguistic interpretation of emotional prosody. Dans *Proc. of ISCA ITRW on Speech and Emotion*, pages 110–113, Belfast, 2000.

[Abrilian *et al.*, 2005] S. Abrilian, L. Devillers, S. Buisine, et J.C. Martin. Emotv1 : Annotation of real-life emotions for the specification of multimodal affective interfaces. Dans *Proc. of HCI International*, Las Vegas, 2005.

[ADVISOR, 2000] ADVISOR. Annotated digital video for surveillance and optimised retrieval. http ://www.inria.fr/MULTIMEDIA/Videotheque/0-Fiches-Videos/517-fra.html, 2000.

[Alku *et al.*, 2002] P. Alku, T. Bäckström, et E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *Journal of Acoustical Society of America*, 112(2) :701–710, 2002.

[Amir et Ron, 1996] N. Amir et S. Ron. Towards an automatic classification of emotions in speech. Dans *Proc. of ICSLP*, Philadelphie, 1996.

[Andrade *et al.*, 2005] E. Andrade, S. Blunsden, et R. Fisher. Simulation of crowd problems for computer vision. Dans *Proc. of First International Workshop on Crowd Simulation*, Lausanne, 2005.

[Antoniou, 2006] E. Antoniou. Reconnaissance des émotions de type peur - machines à vecteurs support. Master's thesis, Telecom-Paris, 2006.

[Audibert *et al.*, 2004a] N. Audibert, V. Aubergé, et A. Rilliard. Ewiz : contrôle d'émotions authentiques. Dans *Actes des Journées d'Étude sur la Parole*, pages 49–52, Fès, 2004.

[Audibert *et al.*, 2004b] N. Audibert, S. Rossato, et V. Aubergé. Paramétrisation de la qualité de voix : Eeg vs. filtrage inverse. Dans *Actes des Journées d'Étude sur la Parole*, pages 53–56, Fès, 2004.

[Bakeman et Gottman, 1997] R. Bakeman et J.M. Gottman. *Observing Interaction : an introduction to sequential analysis*. Cambridge University Press, 1997.

[Banse et Scherer, 1996] R. Banse et K. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3) :614–636, 1996.

[Barras et Gauvain, 2003] C. Barras et J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. Dans *Proc. of ICASSP*, Hong-Kong, 2003.

[Batliner *et al.*, 2003] A. Batliner, K. Fischer, R. Huber, J. Spilker, et E. Nöth. How to find trouble in communication. *Speech Communication*, 40(1-2) :117–143, 2003.

[Batliner *et al.*, 2000] A. Batliner, K. Fisher, R. Huber, J. Spilker, et E. Nöth. Desperately seeking emotions or : Actors, wizards and human beings. Dans *Proc. of ISCA Workshop on Speech and Emotion*, pages 195–200, Belfast, 2000.

[Batliner *et al.*, 2004] A. Batliner, S. Steidl, B. Schuller, et D. Seppi. "you stupid ting box" - children interacting with the aiob robot : a cross-linguistic emotional speech corpus. Dans *Proc. of LREC*, pages 171–174, Lisbon, 2004.

[Batliner *et al.*, 2006] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, et V. Aharonson. Combining efforts for improving automatic classification of emotional user states. Dans *Proc. of Proc. IS-LTC*, Ljubljana, 2006.

[Beller *et al.*, 2006] G. Beller, T. Hueber, D. Schwarz, et X. Rodet. Speech rates in french expressive speech. Dans *Proc. of Speech Prosody*, Dresdes, 2006.

[Bengio et Mariéthoz., 2004] S. Bengio et J. Mariéthoz.. A statistical significance test for person authentication. Dans *In Proc. of Odyssey 2004 : The Speaker and Language Recognition Workshop*, Tolede, 2004.

[Bänziger *et al.*, 2006] T. Bänziger, H. Pirker, et Scherer K.R.. Gemep - geneva multimodal emotion portrayals : A corpus for the study of multimodal emotional expressions. Dans *Proc. of LREC Workshop on Corpora for Research on Emotion and Affect*, Gênes, 2006.

[Boersma et Weenink, 2005] P. Boersma et D. Weenink. Praat : doing phonetics by computer [computer program], from http ://www.praat.org/. Rapport Technique, 2005.

[Breazeal et Aryananda, 2002] C. Breazeal et L. Aryananda. Recognizing affective intent in robot directed speech. *Autonomous Robots*, 12(1) :83–104, 2002.

[Cahn, 1989] J. E. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, 1989.

[Cai *et al.*, 2003] R. Cai, L Lu, H.-J. Zhang, et L.-H. Cai. Highlight sound effects detection in audio stream. Dans *Proc. of ICME*, Baltimore, 2003.

[Calliope, 1997] Calliope. *La parole et son traitement automatique*. Dunod, 1997.

[Campbell et Mokhtari, 2003] N. Campbell et P. Mokhtari. Voice quality : the 4h prosodic dimension. Dans *Proc. of International Congress on Phonetic Sciences*, Barcelone, 2003.

[Carletta, 1996] J. Carletta. Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistics*, 22(2) :249–254, 1996.

[Cheyer et Martin, 2001] A. Cheyer et D. Martin. The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1) :143–148, 2001.

[Chion, 1990] M. Chion. *L'audio-vision : son et image au cinéma*. Nathan, 1990.

[Chung, 2000] S. Chung. *L'expression et la perception de l'émotion extraite de la parole spontanée : évidences du coréen et de l'anglais*. PhD thesis, Institut de Linguistique et Phonétique Générales et Appliquées - PARIS III, 2000.

[Clavel *et al.*, 2007] C. Clavel, L. Devillers, G. Richard, I. Vasilescu, et T. Ehrette. Abnormal situations detection and analysis through fear-type acoustic manifestations. Dans *Proc. of ICASSP*, Honolulu, 2007. à paraître.

[Clavel *et al.*, 2005] C. Clavel, T. Ehrette, et G. Richard. Events detection for an audio-based surveillance system. Dans *Proc. of ICME*, Amsterdam, 2005.

[Clavel *et al.*, 2004] C. Clavel, I. Vasilescu, L. Devillers, et T. Ehrette. Fiction database for emotion detection in abnormal situations. Dans *Proc. of ICSLP*, Jeju, 2004.

[Clavel *et al.*, 2006a] C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, et G. Richard. Safe corpus : fear-type emotions detection for surveillance application. Dans *Proc. of LREC*, Gênes, 2006.

[Clavel *et al.*, 2006b] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, et C. Sedogbo. The safe corpus : illustrating extreme emotions in dynamic situations. Dans *Proc. of LREC Workshop on Corpora for Research on Emotion and Affect*, Gênes, 2006.

[Clavel *et al.*, 2006c] C. Clavel, I. Vasilescu, G. Richard, et L. Devillers. Du corpus émotionnel au système de détection : le point de vue applicatif de la surveillance dans les lieux publics. *Revue d'Intelligence Artificielle, numéro spécial Interaction Emotionnelle*, 20(4-5) :529–551, 2006.

[Clavel *et al.*, 2006d] C. Clavel, I. Vasilescu, G. Richard, et L. Devillers. Voiced and unvoiced content of fear-type emotions in the safe corpus. Dans *Proc. of Speech Prosody*, Dresdes, 2006.

[Cowie et Cornelius, 2003] R. Cowie et R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2) :5–32, 2003.

[Cowie *et al.*, 2000] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, et M. Schröder. 'feeltrace' : An instrument for recording perceived emotion in real time. Dans *Proc. of Proceedings of the ISCA Workshop on Speech and Emotion : A Conceptual Framework for Research*, pages 19–24, Belfast, 2000.

[Cowie *et al.*, 2001] R. Cowie, E. Douglas-Cowie, G. Tsapatsoulis, N.and Votsis, S. Kollias, W. Fellenz, et J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1) :32–80, 2001.

[Craggs, 2004] R. Craggs. Annotating emotion in dialogue - issues and approaches. Dans *Proc. of CLUK Research Colloquium*, 2004.

[Craggs et Wood, 2004] R. Craggs et M.M. Wood. A categorical annotation scheme for emotion in the linguistic content. Dans *Proc. of Affective Dialogue Systems*, Kloster Irsee, 2004.

[Cronbach, 1951] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 :297–334, 1951.

[d'Alessandro, 2002] C. d'Alessandro. *Analyse, synthèse et codage de la parole*. Hermes, 2002.

[Damasio, 1995] A.R. Damasio. *L'erreur de Descartes, la raison des émotions*. Odile Jacob, 1995.

[Darwin, 1872] C. Darwin. *The Expression of the Emotions in Man and Animals*. Chicago University Press, 1872.

[Dellaert *et al.*, 1996] F. Dellaert, T. Polzin, et A. Waibel. Recognizing emotion in speech. Dans *Proc. of ICSLP*, Philadelphie, 1996.

[Dempster *et al.*, 1977] A. Dempster, N. Laird, et D. Rubin. Maximum likelihood from incomplete fata via the em algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[Devillers, 2006] L. Devillers. Les émotions dans les interactions homme-machine : perception, détection et génération. Thèse d'Habilitation à Diriger des Recherches, Décembre 2006. Université Paris XI, Orsay.

[Devillers et Vasilescu, 2003] L. Devillers et I. Vasilescu. Prosodic cues for emotion characterization in real-life spoken dialogs. Dans *Proc. of Eurospeech*, Genève, 2003.

[Devillers et Vidrascu, 2006] L. Devillers et L. Vidrascu. Représentation et détection des émotions dans des dialogues enregistrés dans un centre d'appel - des émotions complexes dans des données réelles. *Revue d'intelligence artificielle, numéro spécial « Interaction Emotionnelle »*, 20(4-5) :447–476, 2006.

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

[Douglas-Cowie *et al.*, 2003] E. Douglas-Cowie, N. Campbell, R. Cowie, et P. Roach. Emotional speech : Towards a new generation of databases. *Speech Communication*, 40(1-2) :33–60, 2003.

[Duda et Hart, 1973] R. Duda et P. E Hart. *Pattern Classification and Scence Analysis*. ser. Wiley-Interscience, 1973.

[Ehrette, 2004] T. Ehrette. *Les voix des services telecoms, de la perception à la modélisation*. PhD thesis, Université Paris XI, 2004.

[Ekman, 1999] P. Ekman. *Basic Emotions*. John Wiley, New York, handbook of cognition and emotion édition, 1999.

[Ekman, 2003] P. Ekman. *Emotions revealed : Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York : Times Books (US), 2003.

[Enos et Hirshberg, 2006] F Enos et J. Hirshberg. A framework for eliciting emotional speech : Capitalizing on the actor's process. Dans *Proc. of LREC Workshop on Corpora for Research on Emotion and Affect*, Gênes, Italie, mai 2006.

[Essid, 2005] S. Essid. *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. PhD thesis, Telecom-Paris, 2005.

[Essid *et al.*, 2006] S. Essid, G. Richard, et B. David. Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Speech and Audio Processing.*, 14(4) :1401–1412, 2006.

[Fernandez et Picard, 2003] R. Fernandez et R. W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1-2) :145–159, 2003.

[France *et al.*, 2003] D. France, R. Shiavi, S. Silverman, M. Silverman, et D. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risks. *IEEE Transactions on Biomedical Engeneering*, 47(7) :829–837, 2003.

[Fredouille *et al.*, 2001] C. Fredouille, J.F. Bonastre, et T. Merlin. Bayesian approach based decision in speaker verification. Dans *Proc. of a speaker odyssey, The speaker recognition workshop*, pages 77–81, Grèce, 2001.

[Galati et Sini, 1995] D. Galati et B. Sini. *Les structures sémantiques du lexique français des émotions*, volume Les émotions dans les interactions. A. Plantin, 1995.

[Guo et Li, 2003] G. Guo et S. Z. Li. Content-based audio classification and retrieval by support vector machines. *IEEE transactions on neural networks*, 14 :209–216, janvier 2003.

[Hansen, 1996] John H. L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, 20(1-2) :151–173, 1996.

[Hastie, 1996] and Tibshirani R. Hastie, T. Classification by pairwise coupling. Rapport Technique, Stanford University and university of Toronto, 1996.

[Hsu et Lin, 2002] C.-W. Hsu et C.-J. Lin. A comparison of methods for multi-classe support vector machines. *IEEE Transactions on Neural Networks*, 13(2) :415–425, mars 2002.

[James, 1884] W. James. *What is emotion ?* Mind 19, 1884.

[Jolion, 2006] J.M. Jolion. *Probabilités et Statistique*. Cours de 3e année, Département Génie Industriel, Insa Lyon, Mai 2006.

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

[Kienast et Sendlmeier, 2000] M. Kienast et W.-F. Sendlmeier. Acoustical analysis of spectral and temporal changes in emotional speech. Dans *Proc. of ISCA ITRW on Speech and Emotion*, pages 92–97, Belfast, 2000.

[Kipp, 2001] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. Dans *Proc. of Eurospeech*, Aalborg, 2001.

[Kleiber, 1990] Georges Kleiber. *La sémantique du prototype, Catégories et sens lexical*. PUF, Paris, 1990.

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

[Kwon *et al.*, 2003] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, et Te-Won Lee. Emotion recognition by speech signals. Dans *Proc. of Eurospeech*, Genève, 2003.

[Landis et Koch, 1977] R.J Landis et G.G. Koch. The measurement of an observer agreement for categorical data. *Biometrics*, 33 :159–174, 1977.

[Lee *et al.*, 2002] C. Lee, S. Narayanan, et R. Pieraccini. Classifying emotions in human-machine spoken dialogs. Dans *Proc. of ICME*, Lausanne, 2002.

[Maeireizo et Litman, 2004] B. Maeireizo et D. Litman. Co-training for predicting emotions with spoken dialogue data. Dans *Proc. of ACL*, Barcelone, 2004.

[Markou et Singh, 2003] M. Markou et S. Singh. Novelty detection : a review. *Signal Processing*, 83(12) :2481–2497, 2003.

[McGilloway, 1997] S. McGilloway. *Negative symptoms and speech parameters in schizophrenia*. PhD thesis, Queen's University, Belfast, 1997.

[McNeill, 2005] D. McNeill. *Gesture and Thought*. University of Chicago Press, 2005.

[Mercier, 1989] D. Mercier. Sound library. Audivis Distribution, 1989. CD.

[Mozziconacci, 1998] S. Mozziconacci. *Speech variability and emotion*. PhD thesis, Technical University Eidhoven, 1998.

[Nam et Tewfik, 2002] J. Nam et A.H. Tewfik. Event-driven video abstraction and visualization. *Multimedia Tools and Applications*, 16(1-2) :55–77, 2002.

[Orthony et Turner, 1990] A. Orthony et T.J. Turner. What's basic about basic emotion ? *Psychological Review*, 97 :315–331, 1990.

[Osgood *et al.*, 1975] C. Osgood, W. H. mai, et M.S. Miron. *Cross-cultural Universals of Affective Meaning*. University of Illinois Press, Urbana, 1975.

[Oudeyer, 2003] P-Y. Oudeyer. The production and recognition of emotions in speech : features and algorithms. *International Journal of Human Computer Interaction, special issue on Affective Computing*, 59(1-2) :157–183, 2003.

[Pelachaud, 2005] C. Pelachaud. Multimodal expressive embodied conversational agent. Dans *Proc. of ACM Multimedia, Brave New Topics session*, Singapour, 2005.

[Pereira., 2000] C. Pereira.. Dimensions of emotional meaning in speech. Dans *Proc. of ISCA ITRW on Speech and Emotion*, pages 75–80, Belfast, 2000.

[Petrushin, 2003] V. Petrushin. Emotion in speech : Recognition and application to call center. Dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[Picard, 1997] R. Picard. *Affective Computing*. MIT Press, Cambridge, MA., 1997.

[Plutchik, 1984] R. Plutchik. *A General Psychoevolutionary Theory*, volume Approaches to Emotion. Erlbaum, Hillsdale, NJ, 1984.

[Remagnino *et al.*, 2002] P. Remagnino, G.A. Jones, N. Paragios, et C.S. Regazzoni. *Video-Based Surveillance Systems*. Hardcover, 2002.

[Roach *et al.*, 1998] P. Roach, R. Stibbard, J. Osborne, S. Arnfield, et J. Setter. Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, 28 :83–94, 1998.

[Russell, 1997] J. A. Russell. *How shall an emotion be called ?* American Psychological Association,Washington, DC, 1997.

[Sanchez-Soto, 2005] Eduardo Sanchez-Soto. *Réseaux Bayésiens Dynamiques pour la Vérification du Locuteur*. PhD thesis, Télécom-Paris, 2005.

[Scherer, 2003] K. Scherer. Vocal communication of emotion : a review of research paradigms. *Speech Communication*, 40(1-2) :227–256, 2003.

[Scherer et Ceschi, 2000] K. Scherer et G. Ceschi. Studying affective communication in the airport : the case of lost baggage claims. *Personality and Social Psychological Bulletin*, 26(3) :327–339, 2000.

[Scherer, 1984] K. R. Scherer. *On the nature and function of emotion : A component process approach*. Lawrence Erlbaum Associates, Publishers, Londres, 1984.

[Scherer *et al.*, 1998] K. R. Scherer, T. Johnstone, et J. Sangsue. L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole. Dans *Actes des Journées d'Études sur la Parole*, pages 249–257, Matigny, 1998.

[Scherer *et al.*, 1980] U. Scherer, H. Helfrich, et K. R. Scherer. *Internal push or external pull ? Determinants of paralinguistic behavior*. Oxford - New York : Pergamon, 1980.

[Schuller *et al.*, 2003] B. Schuller, G. Rigoll, et M. Lang. Hidden markov model-based speech emotion recognition. Dans *Proc. of ICASSP*, Hong Kong, 2003.

[Schuller *et al.*, 2004] B. Schuller, G. Rigoll, et M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. Dans *Proc. of ICASSP*, Montreal, 2004.

[SERKET, 2005] SERKET. Security keeps threats away. http ://www.research.thalesgroup.com/software/cognitive_solutions/Serket, 2005.

[Serkey et Hanson, 1984] A. Serkey et B.A. Hanson. Improved 1-bark auditory filter. *Journal of Acoustical Society of America*, 75(6) :1902–1904, 1984.

[Shafran *et al.*, 2003] I. Shafran, M. Riley, et M. Mohri. Voice signatures. Dans *Proc. of ASRU Workshop*, St Thomas, 2003.

[Snoek *et al.*, 2005] Cees G.M. Snoek, Marcel Worring, et Arnold W.M. Smeulders. Early versus late fusion in semantic video analysis. Dans *Proc. of ACM Multimedia*, Singapour, 2005.

[Steidl *et al.*, 2005] S. Steidl, M. Levit, A. Batliner, E. Nöth, et Nieman H.. "of all things the measure is man", automatic classification of emotions and inter-labeler consistency. Dans *Proc. of ICASSP*, Philadelphie, 2005.

[Tato, 1999] R. Tato. *Emotion Recognition in Speech Signal*. PhD thesis, University of Manchester, 1999.

[ten Bosch, 2003] L. ten Bosch. Emotions, speech and the asr framework. *Speech Communication*, 40(1-2) :213–225, 2003.

[Thomas, 2005] F. Thomas. *Actes de l'université d'automne - Musique(s) et cinéma(s)*. Direction générale de l'Enseignement scolaire, ministère de l'Éducation nationale, de l'enseignement supérieur et de la recherche édition, mars 2005.

[Tukey, 1977] J. W. Tukey. *Exploratory Data Analysis*. EDA, Reading, 1977.

[Vacher *et al.*, 2004] M. Vacher, D. Istrate, L. Besacier, J.F.Serignat, et E. Castelli. Sound detection and classification for medical telesurvey. Dans *Proc. of IASTED Biomedical Conference*, pages 395–399, Innsbruck, Autriche, février 2004.

[van Bezooijen, 1984] R. van Bezooijen. *Characteristics and recognizability of vocal expressions of emotion*. Foris Publications, Dordrecht, 1984.

[Vapnik, 1995] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Varadarajan *et al.*, 2006] V. Varadarajan, J. Hansen, et I. Ayako. Ut-scope - a corpus for speech under cognitive/physical task stress and emotion. Dans *Proc. of LREC Workshop en Corpora for Research on Emotion and Affect*, Gênes, 2006.

[Verfaille, 2003] V. Verfaille. *Effets audionumériques adaptatifs - théorie, mise en oeuvre et usage en création musicale numérique*. PhD thesis, Université Aix-marseille II, 2003.

[Vidrascu et Devillers, 2005] L. Vidrascu et L. Devillers. Detection of real-life emotions in call centers. Dans *Proc. of Eurospeech*, Lisbonne, 2005.

[Vidrascu et Devillers, 2006] L. Vidrascu et L. Devillers. Real-life emotions in naturalistic data recorded in a medical call center. Dans *Proc. of LREC Workshop on Corpora for Research on Emotion and Affect*, Gênes, 2006.

[Whissel, 1989] C.M. Whissel. *The dictionary of affect in language. Emotion : Theory, Research, and Experience*. New York : Academic Press, 1989.

[Williams et Stevens, 1972] C. E. Williams et K. N Stevens. Emotions and speech : some acoustical correlates. *Journal of the Acoustical Society of America*, 52(4) :1238–1250, 1972.

[Wred et Shriberg, 2003] B. Wred et E. Shriberg. Spotting 'hot spots' in meetings : Human judgements and prosodic cues. Dans *Proc. of Eurospeech*, Genève, 2003.

[Xu *et al.*, 2005] M. Xu, L.-T. Chia, et J. Jin. Affective content analysis in comedy and horror videos by audio emotional event detection. Dans *Proc. of ICME*, Amsterdam, 2005.

[Yacoub *et al.*, 2003] S. Yacoub, S. Simske, X. Linke, et J. Burns. Recognition of emotions in interactive voice response system. Dans *Proc. of Eurospeech*, Genève, 2003.

[Yegnanarayana *et al.*, 1998] B. Yegnanarayana, C. d'Alessandro, et V. Darsino. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech and Audio Processing*, 6(1) :1–11, 1998.

Publications

Articles de journal

> C. Clavel, I. Vasilescu, G. Richard, et L. Devillers : Du corpus émotionnel au système de détection : le point de vue applicatif de la surveillance dans les lieux publics. Dans *Revue d'Intelligence Artificielle, numéro spécial Interaction Emotionnelle*, 20(4-5) :529-551, 2006.

Conférences internationales

> C. Clavel, L. Devillers, G. Richard, I. Vasilescu, et T. Ehrette : Abnormal situations detection and analysis through fear-type acoustic manifestations. Dans *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, 2007, à paraître.

> C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, et G. Richard : Safe corpus : fear-type emotions detection for surveillance application. Dans *International Conference on Language Resources and Evaluation Conference (LREC)*, Gênes, 2006.

> C. Clavel, I. Vasilescu, G. Richard, et L. Devillers : Voiced and unvoiced content of fear-type emotions in the safe corpus. Dans *International Conference on Speech Prosody*, Dresdes, 2006.

> C. Clavel, T. Ehrette, et G. Richard : Event detection for an audio-based surveillance system. Dans *International Conference on Multimedia Expo (ICME)*, Amsterdam, 2005.

> C. Clavel, I. Vasilescu, L. Devillers, et T. Ehrette : Fiction database for emotion detection in abnormal situations. Dans *International Conference on Spoken Language Processing (ICSLP)*, Jeju, 2004.

Workshops

> C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, et C. Sedogbo : The safe corpus : illustrating extreme emotions in dynamic situations. Dans *LREC Workshop on Corpora for Research on Emotion and Affect*, Gênes, 2006.