Chapitre 1

Reconnaissance acoustique des émotions

1.1. 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 s'intéresser au développement de systèmes de classification automatique des émotions. Cette évolution est née de l'émergence des sciences affectives - « affective computing » [PIC 97] - et de la prise de conscience des applications industrielles potentielles de ce domaine.

Un grand nombre d'applications de la reconnaissance des émotions dans la voix concerne le domaine de l'interaction homme-machine. Les systèmes de dialogue en sont un premier exemple : la détermination de l'état émotionnel de l'utilisateur permet d'adapter la stratégie dialogique. Par exemple, pour les centres d'appel, si l'utilisateur manifeste des signes d'irritation ou de frustration face au répondeur automatique, une stratégie peut être de le diriger vers un opérateur humain [LEE 02], [DEV 05]. Mais les applications de la reconnaissance d'émotions ne se limitent pas aux systèmes de dialogues. Dans le domaine de la santé, certains travaux de recherche académique [IST 03] s'intéressent à la reconnaissance des émotions pour l'aide aux personnes âgées ou aux personnes hospitalisées. Enfin dans le domaine de la sécurité, deux applications sont apparues récemment : la gestion de crise et l'audiosurveillance. En ce qui concerne la gestion de crise, des travaux sont menés pour la prise en charge des émotions des victimes et des sauveteurs par les robots collaboratifs (Search and Rescue) [LOO 07]. Le déploiement de tels robots est en effet crucial dans la gestion de crise et notamment dans des environnements jugés trop dangereux ou inaccessibles

Chapitre rédigé par Chloé CLAVEL et Gaël RICHARD.

pour les sauveteurs. En ce qui concerne la sécurité, 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 [NGH 07]. Certains travaux commencent à s'intéresser à l'intégration de l'information liée aux manifestations émotionnelles contenue dans le flux audio comme complément d'information à la vidéo [CLA 08] [HEN 07].

Ce chapitre décrit les différentes étapes nécessaires à la mise en oeuvre d'un système de reconnaissance des émotions reposant sur des indices acoustiques[1]. La première étape consiste en la collecte et l'annotation de données, cette étape est détaillée dans le chapitre 3. Nous nous focaliserons dans ce chapitre sur la seconde étape dédiée au développement d'un « système » de reconnaissance d'émotions sur ces données, en décrivant d'une part les descripteurs acoustiques destinés à caractériser et discriminer entre elles les différentes manifestations émotionnelles, et d'autre part en répertoriant les techniques de classification (apprentissage et décision) utilisées. Enfin, nous présentons les différents critères à considérer pour l'évaluation d'un système et proposons un état de l'art du domaine en terme de performances.

1.2. Principe d'un système de reconnaissance automatique des émotions

Un système de reconnaissance automatique des émotions repose classiquement sur quatre phases principales (voir Figure 1.1) :

1) *L'extraction de descripteurs acoustiques* qui consiste en un module d'analyse transformant le signal de parole en une séquence de vecteurs acoustiques contenant les valeurs des différents descripteurs (ou paramètres) retenus. L'objectif de cette étape de paramétrisation est d'obtenir une représentation compacte des principales caractéristiques acoustiques du signal de parole qui soit pertinente pour discriminer entre elles les différentes manifestations émotionnelles. L'intensité du signal de parole ou la fréquence centrale des formants sont deux exemples de tels descripteurs acoustiques.

2) Durant la phase *d'apprentissage*, plusieurs vecteurs acoustiques correspondant aux sons d'une même classe sont utilisés pour créer un *représentant* ou un modèle caractéristique de cette classe. Le représentant peut être par exemple obtenu comme le barycentre (ou centroide) des vecteurs acoustiques caractéristiques de la classe considérée. Un modèle permettra souvent de mieux caractériser les distributions de valeurs des vecteurs acoustiques pour chaque classe (ici chaque classe correspond par exemple

1. Certains systèmes reposent sur une fusion des contenus linguistiques et acoustiques [SCH 04]. Nous présentons dans ce chapitre uniquement la reconnaissance acoustique des émotions

à une classe émotionnelle). Ce représentant ou ce modèle est traditionnellement obtenu à partir d'une base de données (dite d'apprentissage) qui aura été préalablement annotée (v. chapitre 3).

3) C'est au cours de la phase de *classification* que les vecteurs acoustiques du signal vocal à analyser sont comparés aux représentants ou modèles de chaque classe. A l'issue de cette phase, une probabilité d'appartenance à chaque classe peut être obtenue pour chaque vecteur acoustique.

4) Enfin, la phase de *décision* associe une classe à un segment de parole en exploitant les probabilités d'appartenance successives des vecteurs acoustiques.

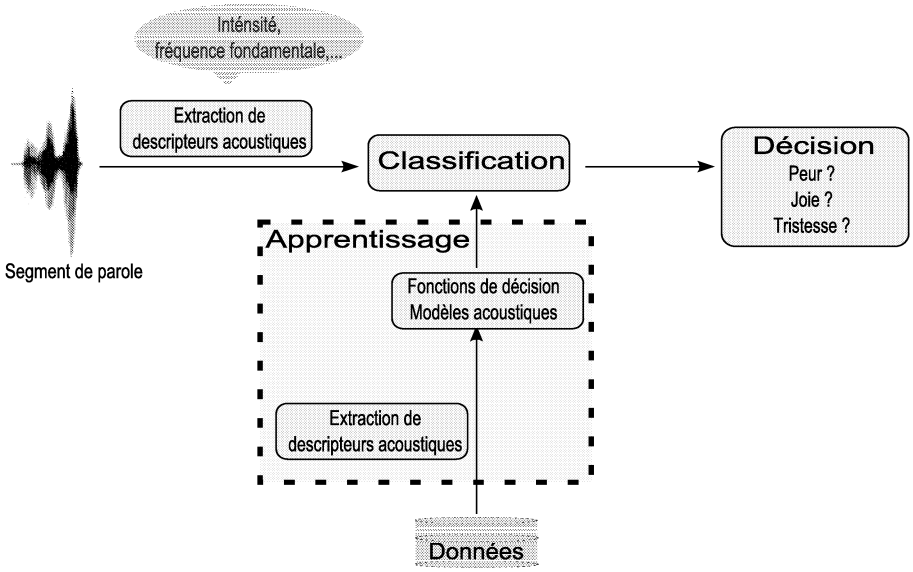

Figure 1.1. *Principe d'un système de reconnaissance d'émotions*

1.3. Descripteurs acoustiques

Le phénomène physique de production de la parole est classiquement représenté par un modèle source-filtre [FAN 60]. Ce modèle simple reste une référence pour décrire les principales caractéristiques acoustiques du signal vocal et c'est dans le cadre de ce modèle de production que seront présentés l'ensemble des paramètres (ou descripteurs) acoustiques utilisés en reconnaissance automatique des émotions.

Une grande partie des descripteurs acoustiques utilisés pour caractériser les différents états émotionnels est destinée à modéliser les altérations du signal acoustique

liées à des modifications physiologiques à la base de la glotte. C'est le cas des paramètres prosodiques et des paramètres de qualité de voix (voix soufflée, voix grinçante, voix dure, voix tendue). Nous présentons dans ce paragraphe un bilan de l'utilisation de ces descripteurs dans les études sur la parole émotionnelle. Ce paragraphe présentera également des paramètres (formants, coefficients cepstraux, ...) représentant les modifications du signal acoustique liées aux modifications du conduit vocal (e.g. la partie filtre du modèle source/filtre).

Les modifications corporelles/physiologiques qui accompagnent certains états émotionnels, vont fortement influer sur le mode de production du message oral du locuteur. Par exemple, dans le cas de la peur, les modifications physiologiques typiques sont l'augmentation du pouls, de la pression du sang et de la sécheresse de la bouche, et se manifestent par une voix plus forte, 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 [PIC 97].

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 dans les diverses évaluations d'un stimulus externe (théorie de l'évaluation [SCH 03a], *« appraisal theory »*). Ce sont ces diverses évaluations qui déterminent 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 »*).

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 [OUD 03] ou dans [CLA 08], les descripteurs utilisés sont souvent plus nombreux, incluant des descripteurs de plus *bas niveau* pour lesquels il est difficile d'identifier un corrélat perceptif (voir chapitre 8). 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.

Contenu voisé vs. non voisé

Les études dédiées aux 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 fortes qui sont souvent accompagnées de fortes modifications corporelles telles que la crispation, les tremblements, l'augmentation du rythme cardiaque. L'activité physiologique et/ou 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. C'est pour cette raison que dans [CLA 08], l'impact des manifestations émotionnelles sur le contenu acoustique est étudié à la fois sur le contenu voisé et non voisé.

Unité temporelle d'analyse de l'émotion

Au delà du choix de descripteurs acoustiques pertinents pour caractériser le contenu émotionnel, il se pose une question essentielle : 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. Il existe deux approches. La première – comme par exemple dans [VID 05]– utilise des descripteurs de type statistique proposant une modélisation globale de chaque descripteur (par exemple la moyenne, le minimum, etc.) sur différentes durées temporelles, comme la syllabe, le mot, ou la phrase.

La seconde approche repose sur l'extraction des descripteurs acoustiques sur chaque fenêtre d'analyse [AMI 96] ou sur la combinaison d'une modélisation au niveau de la fenêtre avec une modélisation globale des descripteurs [VLA 07] (fenêtres de 25ms, tour de parole) [CLA 08] (fenêtres de 40ms, trajectoires voisées et non voisées[2], segments homogènes[3]).

Les modélisations au niveau de la fenêtre d'analyse et de la trajectoire 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. En revanche, les modélisations globales reposent sur une segmentation fine de la parole qui peut par exemple être obtenue à l'aide d'un système de reconnaissance automatique de la parole. Ces analyses présentent l'inconvénient d'être peu robustes aux phénomènes

2. Une trajectoire voisée (respectivement non voisée) est une succession de fenêtres adjacentes voisées (respectivement non voisées)

3. Le segment est défini dans ce travail comme une portion de tour de parole avec un contenu émotionnel homogène

tels que la coarticulation, des voix non modales, des manifestations non verbales telles que les cris, etc. qui surviennent lors de manifestations émotionnelles fortes spontanées, ce qui peut éventuellement poser problème selon le type de données traitées.

1.3.1. *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 mais aussi plus 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 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 de la voix [DEL 96],[AMI 96] qui correspondent en outre à une modélisation des changements supra-segmentaux[4] du signal acoustique produits par les modifications physiologiques à la base de la glotte [SCH 98].

La fréquence fondamentale ou pitch

En parole, la fréquence fondamentale correspond à la fréquence de vibration des cordes vocales lors de la production des sons dits voisés. Elle est directement 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 $\mathcal{I}$ 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. [KLA 08] [HES 84].

L'une des méthodes les plus populaires – celle utilisée notamment par le logiciel Praat [BOE 05] – est une approche temporelle 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 :

4. en fonction de l'accentuation, de l'intonation ; appelé également niveau prosodique

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

La période T est estimée en recherchant la valeur de m pour laquelle $r_s(m)$ est maximale dans un intervalle choisi a priori $m \in [T_{min} : T_{max}]$. 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.

La figure 1.2, tirée de [CLA 07b], illustre le comportement de la fréquence fondamentale sur deux exemples contenant différents degrés d'intensité émotionnelle[5].

Ces deux *segments* correspondent au même mot « Josh » prononcé par une même locutrice lors d'une situation qui évolue. Pour le premier exemple, 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 cet exemple sont difficilement perceptibles et on note la montée de la fréquence fondamentale caractéristique d'une question. Pour le second exemple, 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 (ou énergie) permet de fournir une mesure de la force sonore de la voix (faible ou forte). L'intensité en décibel (dB) est en général – notamment sous le logiciel Praat [BOE 05] – calculée sur une portion de signal de longueur N de la façon suivante :

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

où w est une fenêtre d'analyse.

La figure 1.3 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.

5. La dimension intensité dérive de la dimension *activation* définie par [OSG 75] comme un état d'excitation, de calme à fortement excité, voir chapitre 3.

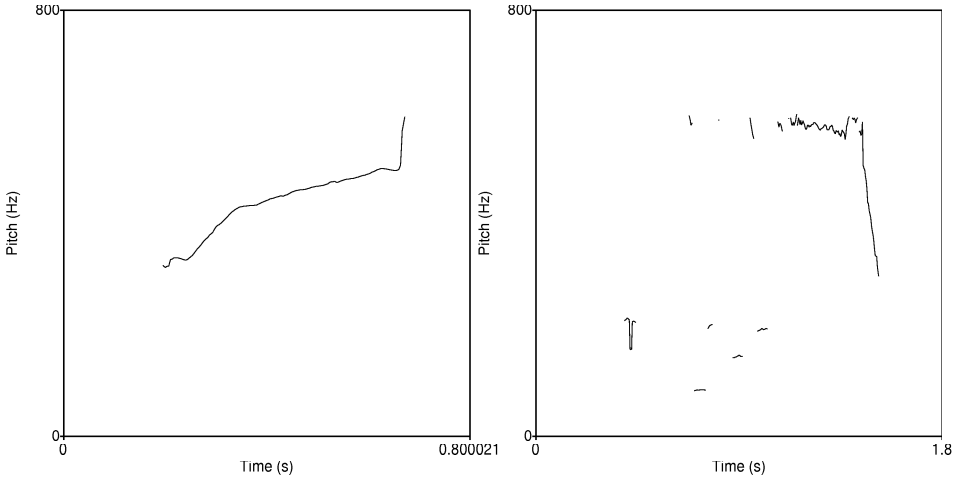

Figure 1.2. *Exemple de contour de la fréquence fondamentale estimée par la méthode de l'autocorrélation implémentée dans le logiciel Praat sur : à gauche, l'exemple « Josh ? » correspondant à de la peur avec une faible intensité (inquiétude) et à droite l'exemple « Joooosh ! » correspondant à de la peur avec une forte intensité (panique).*

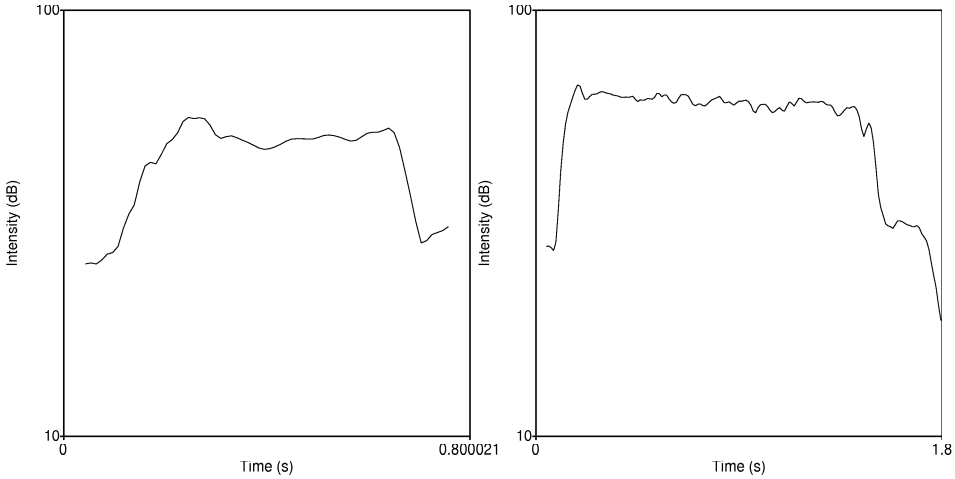

Figure 1.3. *Exemple de contour d'intensité sous Praat calculé sur : à gauche, l'exemple « Josh ? » correspondant à de la peur avec une faible intensité (inquiétude) et à droite l'exemple « Joooosh ! » correspondant à de la peur avec une forte intensité (panique).*

Descripteurs du rythme

Le troisième descripteur classique de la prosodie est un descripteur du rythme lié au débit d'élocution [KOZ 65] 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 calculée que si l'on dispose d'une segmentation du signal de parole en syllabes ou phonèmes (ce qui suppose dans ce cas l'aide d'un système de reconnaissance de la parole). Cette segmentation peut s'avérer complexe selon les conditions d'enregistrement. Une alternative [CLA 08] peut être de calculer la durée de la trajectoire voisée (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.

1.3.2. *Les descripteurs de qualité de voix*

Les variations de qualité de voix, bien que souvent associées à des voix pathologiques, sont également la conséquence des modifications physiologiques provoquées par les changements émotionnels du locuteur. Les descripteurs de qualité de voix sont de plus en plus utilisés pour caractériser les variations émotionnelles.

Le NAQ (Normalized Amplitude Quotient)

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[6]. Ainsi, dans [CAM 03], 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 ce chapitre, le rôle des descripteurs modélisant la source glottale pour la reconnaissance automatique des émotions est clairement mis en évidence sur des données correspondant à des enregistrements d'une locutrice japonaise dans ses interactions de la vie de tous les jours. Sur ces enregistrements, le comportement du NAQ est analysé en fonction du type d'interlocuteur et du degré de politesse et d'«effort» fourni par la locutrice pour communiquer. Le NAQ est également utilisé dans [AUD 04] pour la caractérisation des émotions.

En modélisant le flux glottal 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*), qui correspond au temps de fermeture de la glotte, peut

6. Ces travaux nécessitent cependant des conditions d'enregistrement idéales, comme une proximité du microphone à la bouche.

s'exprimer 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 coefficient AQ dépend de la fréquence fondamentale. Le NAQ correspond au quotient d'amplitude décorrélé de la fréquence fondamentale (F_0) :

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

De plus amples détails sur le calcul de ce coefficient sont donnés dans [ALK 02].

La modulation fréquentielle ou jitter

Le jitter est décrit comme une déviation de la fréquence fondamentale (voir section 1.3.1), et correspond à la mesure d'un bruit sur la fréquence (on parle aussi parfois d'oscillations autour de la fréquence fondamentale). En musique il sert, avec le shimmer ou modulation d'amplitude, à modéliser l'attaque, le soutien et le relâchement des notes [VER 03]. Il est également utilisé pour l'étude des voix pathologiques (voix grinçante) et a été utilisé notamment par [FRA 03] pour la caractérisation d'état dépressif et par [CLA 07a] pour la caractérisation d'émotions de type peur. La figure 1.2 met notamment en évidence la présence de fortes modulations de la fréquence fondamentale sur le cri, « Jooosh ! ».

Une approche classique pour l'estimation du jitter consiste à calculer le rapport suivant :

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

où T_n est le n^e intervalle d'une période de signal et N le nombre d'intervalles.

La modulation en amplitude ou shimmer

Souvent utilisé de pair avec le jitter, le shimmer modélise, lui, la modulation d'amplitude du signal vocal s. Comme pour le jitter, la figure 1.3 met en évidence la présence de fortes modulations de l'intensité sur le cri, « Jooosh ! ».

Une estimation du shimmer, similairement au jitter, peut être obtenue en calculant, sur une portion de signal de longueur N, le rapport suivant :

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

où A_n est l'amplitude moyenne calculée sur l'intervalle T_n, et N le nombre d'intervalles.

Taux de fenêtres non voisées

Le taux de fenêtres non voisées correspond à la proportion de fenêtres considérées comme non voisées sur une portion du signal de parole. En pratique, ce paramètre peut être calculé de la manière suivante : une fenêtre est considérée comme non voisée si la « force du voisement » (score de la fonction d'autocorrélation, voir section 1.3.1) est inférieure à un certain seuil (typiquement fixé dans l'intervalle [0.45 - 0.6]).

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). Par exemple, l'algorithme présenté dans [YEG 98] et repris dans [CLA 07a] pour caractériser les émotions de type peur 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.

Cet algorithme repose sur une décomposition du signal vocal en deux composantes distinctes : l'une périodique résultant de la vibration quasi-périodique des cordes vocales et l'autre apériodique regroupant l'ensemble des bruits de parole (bruit rose fricatif, sifflant ou plosif).

Le descripteur 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 :

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

1.3.3. *Les descripteurs spectraux et cepstraux*

Une grande partie des descripteurs bas niveau reposent sur une analyse du spectre du signal et incluent 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

L'analyse acoustique du conduit vocal met en évidence la présence de résonances. Ces résonances, variables en fonction de la position des différents articulateurs (langue, lèvres, ...), sont appelées les formants. Ils sont notamment caractérisés par une fréquence centrale (ou fréquence du formant) et une largeur de bande mesurée à $-6dB$

du sommet. Sur le contenu voisé, la position des deux premiers formants est indépendante de la hauteur (fréquence fondamentale) et est caractéristique d'une voyelle particulière. 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. En pratique, les premiers formants peuvent être estimés à l'aide d'une analyse par prédiction linéaire (ou LPC pour *Linear Prediction Coding*).

Les paramètres formantiques sont très largement utilisés pour décrire les émotions, comme par exemple dans [FRA 03], [KIE 00] ou [CLA 08] pour la caractérisation acoustique des émotions de type peur. La pertinence de ces paramètres peut être expliquée, d'une part par leur aspect de descripteurs de la qualité vocale, et d'autre part par la relation qui existe entre le premier formant et le degré d'ouverture des voyelles que l'on peut prévoir plus élevé dans le cas d'émotions extrêmes.

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 [SHA 03] [KWO 03].

Les MFCC appartiennent à la famille des descripteurs cepstraux qui se basent sur une représentation cepstrale du signal. Le cepstre présente l'avantage, dans une certaine mesure[7], de permettre une séparation des contributions respectives de la source et du conduit vocal.

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 et prend notamment en compte la plus grande sélectivité en fréquence de l'oreille dans les basses fréquences.

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 bande Mel m. A cet effet, des filtres triangulaires de largeur de bande constante et régulièrement espacés sur l'échelle Mel sont couramment utilisés. Enfin, les coefficients MFCC sont obtenus en effectuant une transformée en cosinus discrète du logarithme des coefficients obtenus précédemment :

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

7. La séparation est moins précise dans les aigus que dans les graves

où M_b est le nombre de filtres triangulaires.

La contribution du conduit vocal est principalement présente dans les premiers coefficients cepstraux ce qui explique qu'en pratique seuls les premiers coefficients sont conservés (typiquement entre 12 et 15 coefficients).

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 [ZWI 57]. Il existe plusieurs formules donnant la fréquence en Bark en fonction de la fréquence en Hertz dont celle proposée dans [SER 84] :

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

En pratique, il peut être calculé en découpant le signal en différentes bandes de fréquences de largeur égale à un ou deux Bark [EHR 04]. Pour chaque bande i, l'énergie s'exprime de la manière suivante :

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

où m 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 différents types de voix (ex : criarde, agressive, sensuelle) dans [EHR 04] ou pour l'étude des manifestations acoustiques des émotions dans [KIE 00] (colère, joie, peur, ennui et tristesse) et [CAI 03] (rires et acclamations).

Centroïde spectral

Le centroïde spectral ou balance spectrale peut être utilisé, comme dans [EHR 04], en tant que descripteur du « timbre » de la voix. Dans [KIE 00], 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 exploité, entre autres, pour la caractérisation et la reconnaissance automatique des instruments de musique [MCA 99] [ESS 05]. 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.

1.4. Classification automatique des émotions

Les systèmes de classification automatique des émotions reposent sur des méthodes dites d'apprentissage, en raison de leur capacité d'apprendre (ou de caractériser) –à partir d'une quantité de données suffisante – 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 [DUD 73]. Nous présentons ici quelques algorithmes de classification en soulevant les différents problèmes inhérents aux émotions.

1.4.1. *Sélection de descripteurs*

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 hétérogénéïté dans les valeurs peut avoir des conséquences sur le comportement des algorithmes de réduction de l'espace des descripteurs – par exemple une sélection des descripteurs les plus discriminants – et d'apprentissage : 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 envisagées pour éviter ce biais. Certaines ont d'ailleurs déjà été utilisées avec succès en reconnaissance des émotions :

– la technique de normalisation dite normalisation min-max [CLA 08],

– la normalisation sigma-mu,

– la normalisation par genre/par locuteur/par phonème [DEV 05].

Réduction de l'espace de représentation des données

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 [DUD 73]. 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 ainsi indispensable si un grand nombre de paramètres est choisi (on pourra par exemple consulter [GUY 03]).

Pour réduire l'espace des descripteurs, deux options se présentent :

– la projection de l'espace de représentation des données sur un espace de dimension plus petit (ex : analyse en composantes principales, analyse discriminante).

– la sélection du sous-ensemble des descripteurs le plus discriminant (ex : algorithme de sélection de Fisher, algorithme génétique) ; cette option présente 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.

Algorithme de sélection de Fisher : cette méthode dont la simplicité et l'efficacité ont été démontrées à maintes reprises (voir [ESS 05] pour son utilisation dans un système de reconnaissance des instruments de musique) est dérivée de l'analyse discriminante de Fisher dont on peut trouver une description dans [DUD 73]. Elle consiste à maximiser le rapport FDR (*Fisher Discriminant Ratio*) de la dispersion inter-classe et la dispersion intra-classe pour chaque descripteur séparément :

$$FDR_{d_i} = \frac{(\mu_{i,classe1} - \mu_{i,classe2})^2}{\sigma^2_{i,classe1} + \sigma^2_{i,classe2}}$$

où $\mu_{i,classe1}$ et $\mu_{i,classe2}$ sont les moyennes des valeurs correspondant aux descripteurs d_i pour chacune des classes et $\sigma^2_{i,classe1}$ et $\sigma^2_{i,classe2}$ les variances correspondantes.

Une sélection multi-classes par l'algorithme de Fisher avec une décomposition en deux étapes est également possible[8] :

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 :

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

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

8. http ://www.kyb.mpg.de/bs/people/spider/. C'est notamment l'option choisie par la boite à outils Spider

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 les 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 peut être utilisé en deux étapes comme dans [CLA 07a] :

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.

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* [PEE 03]) ou des algorithmes génétiques qui proposent une procédure systématique de parcours des sous-ensembles de descripteurs possibles.

1.4.2. *Algorithmes d'apprentissage*

L'utilisation de méthodes de classification dans le domaine des émotions est un sujet de recherche émergeant. Nous répertorions ici les différentes méthodes de classification utilisées dans ce domaine et présentons plus en détails quelques unes de ces méthodes.

Les Séparateurs à Vastes Marges - SVM

Les SVM ont été introduits par Vapnik en 1995 [VAP 95]. Cette méthode est donc une alternative récente pour la classification. Initialement prévue pour résoudre des problèmes de classification à deux classes, il existe aujourd'hui des généralisations multi-classes [HSU 02].

Le principe des SVM peut se schématiser 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 1.4). 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 ».

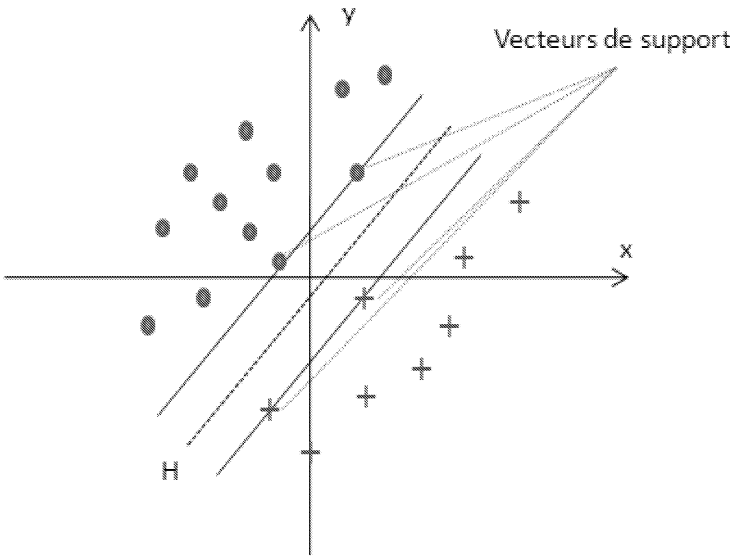

Figure 1.4. *Hyperplan séparateur et vecteurs de support dans un espace à deux dimensions*

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. Hormis le noyau linéaire $k(x, y) = x.y$, les noyaux les plus couramment utilisés sont les noyaux polynomiaux $K(x, y) = (1 + x.y)^d$ (polynôme de degré d) et gaussiens $k(x, y) = \exp(\frac{-\|x-y\|^2}{2\sigma^2})$.

Dans ce dernier cas, la variance σ du noyau permet de régler la précision de l'hyperplan séparateur. Un σ petit signifie que l'hyperplan séparateur est très proche des vecteurs supports et risque le surapprentissage, un σ très grand correspond à une situation où l'hyperplan séparateur devient linéaire, et peut entraîner un sous apprentissage.

La classification d'un nouvel exemple de test est donnée par sa position dans l'espace de re-description 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.

Les modèles de mélange de gaussiennes – GMM

Les GMM sont couramment utilisés 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 ([SAN 05], [FRE 01], [BAR 03]). Il a été également utilisé avec succès pour la reconnaissance des émotions. 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 considérées 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 classiquement 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 ensuite réalisée à l'aide de l'algorithme « Expectation Maximisation » (E-M) [DEM 77].

La classification peut être réalisée à partir d'une règle de décision basée sur le maximum a posteriori. Pour chaque classificateur, un score *a posteriori* (SAP) est associé à chaque classe d'émotion à reconnaître pour chaque segment de parole. Ce score correspond à la moyenne des log-probabilités *a posteriori*, calculée en multipliant les probabilités obtenues sur chaque fenêtre d'analyse. Ainsi, le score obtenu

pour le classificateur est :

$$SAP(C_q) = \frac{\sum_{n=1}^{N} \log p(C_q/x_n)}{N}$$

où x_n est le vecteur d'observation correspondant à la fenêtre d'analyse n 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 :

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

Si on considère que les classes C_q sont équiprobables[9], nous avons $p(C_q) = \frac{1}{q}$ et le problème se ramène à maximiser le score *a posteriori* modifié de la manière suivante (ce qui se ramène à un score au maximum de vraisemblance) :

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

1.5. Performances et évaluation

À l'heure actuelle, il est difficile de comparer les performances obtenues par les systèmes déployés dans 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 et poursuivie en 2007 dans le cadre du réseau d'excellence HUMAINE[10] [BAT 06]. 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. 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 cependant 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. En effet, les performances des systèmes de classification développés par les différents laboratoires dépendent de nombreux facteurs que nous répertorions ci-après.

---

9. La probabilité de chaque classe peut être réglée différemment si on dispose d'*a priori* sur la fréquence des classes considérées.

10. http ://www5.informatik.uni-erlangen.de/Forschung/Projekte/HUMAINE/ ?language=en

Premier Facteur : les données et les classes d'émotions

Les performances des systèmes sont fortement dépendantes 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 ? La première difficulté pour comparer les différentes performances obtenues vient donc de la diffusion restreinte, voire quasi inexistantes, des bases de données émotionnelles. Certaines études récentes s'attaquent à ce problème et commencent à proposer des solutions [CUL 08], [BAT 08].

Par ailleurs, il n'existe à ce jour que très peu d'études menées avec des données issues de corpus enregistrés dans des contextes réels et les scores élevés de classification affichés jusque maintenant sont exclusivement obtenus sur des données actées. Citons par exemple, dans [SCH 04], le score de 93% atteint pour une classification en 7 classes. Les données naturelles illustrent une plus grande diversité de manifestations émotionnelles que les données actées enregistrées en laboratoire, ce qui laisse présager une chute de performance lors du passage à des données réelles. Cette baisse a été constatée dans [BAT 00]. Dans [SHA 07], les auteurs comparent également les performances obtenues sur deux bases de données Kismet et BabyEars qui illustrent sensiblement les même classes émotionnelles. Les performances obtenues sur la base de données actées (Kismet) tournent autour des 90% pour une discrimination entre trois classes (approbation, attention, interdiction) alors que sur la base de données présentant des interactions naturelles entre parents et enfants le score descend aux alentours de 65%.

Cependant, les scores élevés ne sont pas forcément à mettre en relation avec le caractère acté des données. Le caractère caricatural et stéréotypé des données actées facilite, certes, la classification. Mais la variabilité des émotions illustrées est également un facteur de complexité pour la modélisation acoustique. Ainsi, les scores élevés sont souvent liés à une illustration restreinte des manifestations émotionnelles (locuteurs, situations) présentes dans les données enregistrées en laboratoire au sein d'une même classe émotionnelle. Sur un corpus acté (corpus SAFE constitués de séquences extraites du cinéma de fiction) illustrant une grande diversité de contextes en termes de locuteurs, d'environnements sonores, de situations et par conséquent de types de manifestations émotionnelles (40 films, 400 locuteurs) – cette richesse du contexte participe d'ailleurs à un jeu plus réaliste –, les performances obtenues sont autour de 70% pour une discrimination en deux classes [CLA 08]

Les performances obtenues dépendent également du nombre de classes émotionnelles considérées, une classification en 10 classes étant plus ambitieuse qu'une classification en 2 classes. Elles dépendent également du choix des classes émotionnelles considérées, une discrimination joie/satisfaction étant en général plus difficile qu'une discrimination joie/neutre. Les performances dans [DEV 07] sont de 55% avec des indices paralinguistiques pour 5 classes (Peur, Colère, Tristesse, Soulagement, Neutre)

et atteignent 80% pour deux classes de valence (négative/neutre) avec des indices paralinguistiques. Parmi les rares études qui traitent plus de 4-5 états émotionnels, citons par exemple [BAT 03] qui en considèrent 7 pour détecter les émotions dans des données actées en contexte obtenues dans le cadre d'un système de dialogue homme-machine avec un kiosque multimodal. Les performances actuelles des systèmes de reconnaissance des émotions sont encore trop basses pour envisager une discrimination sur un nombre de classes supérieur à 7 sur des données réelles ou présentant une forte variabilité intra-classe.

Second facteur : le problème des « vérités terrains »

L'évaluation des performances d'un système de classification ou de reconnaissance consiste en l'analyse des confusions entre les annotations (« vérités terrains ») et les décisions du système. En reconnaissance du locuteur, par exemple, les annotations qui servent de référence pour l'évaluation des performances sont fixes et fiables : le locuteur associé à l'enregistrement sonore est connu dans la base de données de test. Il s'agit donc de comparer la sortie du système de reconnaissance du locuteur avec l'annotation. En reconnaissance de la parole, les annotations de référence sont soumises à l'erreur humaine qui est estimée à 2% (voir chapitre 8). . Dans le domaine des émotions, nous ne pouvons pas parler d'erreur humaine proprement dite, le problème est beaucoup plus complexe : l'étape d'annotation du contenu émotionnel des enregistrements sonores est soumise à la subjectivité dans la perception des émotions. Il est par conséquent difficile de faire converger les différentes annotations entre elles. Les performances obtenues vont donc être dépendantes des annotations servant de référence. La fiabilité des annotations obtenues dépend, entre autres, de l'adéquation du schéma d'annotation avec les données en fonction de la tâche visée (voir chapitre 3). La convergence entre les différentes annotations pour les différentes classes d'émotions considérées se mesure couramment à l'aide du kappa[11] [COW 03], [CLA 08], [CRA 04], [DEV 05]. Dans [CLA 07b], le système de classification des émotions de type peur versus la classe neutre a été évalué en confrontant les réponses des trois annotateurs à la réponse du système. Les performances obtenues avec le taux d'égale erreur[12] oscillent entre 30% et 35%, soit une différence d'environ 5%. Pour pallier ce problème de subjectivité, [STE 05] propose l'utilisation d'une mesure d'entropie pour l'évaluation des performances.

11. Le kappa correspond au taux d'accord entre les différents annotateurs corrigé de telle façon qu'un kappa nul correspond à un taux d'accord qui serait obtenu uniquement par chance [CAR 96].

12. Pour le système à base de GMM, la balance entre le taux de fausses détections et le taux de faux rejets va dépendre du seuil de décision choisi. Il existe un seuil pour lequel la valeur prise par le taux de fausses détections est égal au taux de faux rejets. Cette valeur correspond au taux d'égale erreur (TEE).

Troisième facteur : prétraitements manuels

Les différentes études menées sur la classification d'émotions peuvent reposer sur des prétraitements manuels dont l'automatisation est susceptible de faire chuter les performances. Tout d'abord, pour l'extraction des descripteurs, différentes stratégies sont adoptées dont la plupart repose sur une connaissance a priori du contenu linguistique. Le signal sonore, une fois aligné avec la transcription du contenu linguistique, peut être segmenté en phrases, en mots voire même en syllabes. Cette segmentation peut ensuite servir de base à la modélisation acoustique du signal. Pour le développement d'un système de classification effectif, cette approche fait l'hypothèse qu'une segmentation fine du signal de parole puisse être réalisée automatiquement sans chute de performance, ce qui n'est pas le cas pour tous les types de corpus.

Enfin, pour l'étape de normalisation des descripteurs, certains descripteurs comme la fréquence fondamentale varient d'un locuteur à un autre et les différences peuvent être particulièrement importantes entre un homme et une femme. L'étude des variations de ces descripteurs en fonction de l'émotion peut être fortement biaisée si l'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 comme dans [DEV 06], [WRE 03]. Cette normalisation présuppose cependant une connaissance préalable du locuteur considéré. Ce qui peut être le cas pour certaines applications où le logiciel de détection d'émotions pourrait être adapté à un locuteur spécifique (détection du stress chez un pilote [VAR 06], certains systèmes de dialogues). Cependant, dans de nombreux cas applicatifs (surveillance, centre d'appel) l'identité du locuteur n'est pas connue a priori et le nombre de locuteurs à traiter peut être particulièrement élevé.

Quatrième facteur : algorithme d'apprentissage

Différentes études ont été menées afin de comparer différents algorithmes d'apprentissage pour un même problème de classification. Ainsi, dans l'article [SCH 04], deux méthodes d'apprentissage sont confrontées sur des données actées en laboratoire, l'une appartenant à la classe des algorithmes génératifs, les GMM, l'autre reposant sur une approche discriminative les SVM. Les deux méthodes obtiennent des performances similaires, 75% pour les GMM et 76% pour les SVM. Ce résultat se retrouve également dans [CLA 07b] qui compare ces deux méthodes pour une classification peur/neutre sur un corpus de fiction présentant des données très diversifiées.

Dans [SCH 03b], la classification par GMM est comparée à la méthode des HMM (*Hidden Markov Models*) 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 continus, en considérant des caractéristiques instantanées de plus haut niveau au lieu de statistiques globales. Dans ce cadre, les performances – obtenues sans évaluation inter-locuteur– se sont avérées être finalement meilleures avec les GMM.

De manière générale, les méthodes d'apprentissage utilisées fournissent des résultats équivalents et il est donc difficile de dégager de ces résultats la supériorité d'une méthode de classification. C'est le cas dans [LEE 02], où trois méthodes différentes ont été testées : l'analyse discriminante, les machines à vecteur support et les k plus proches voisins. En revanche, [PET 03] compare réseau de neurones et k plus proches voisins (kppv) et obtient des performances nettement meilleures avec les réseaux de neurones (70% au lieu de 55%). Dans [VID 05], ce sont les SVM et les arbres de décision qui sont comparés sur des données réelles (centres d'appel) : les résultats ne présentent pas de différences significatives. Le système de classification utilise en fait plusieurs réseaux de neurones sur différents sous-ensembles d'apprentissage (bootstrap). L'article de [SHA 07] fournit une bonne comparaison des performances en testant différents algorithmes (kppv, SVM, arbres de décision Ada-boost) sur plusieurs bases de données. Les algorithmes qui permettent d'obtenir les meilleures performances varient en fonction de la base de données testée. Cependant cette variation est négligeable au regard de l'intervalle de confiance élevé dû à la faible taille des bases de données.

Cinquième facteur : conditions d'apprentissage

L'évaluation des performances des systèmes de classification est réalisée sur des bases de données pouvant avoir des caractéristiques très variables comme présenté précédemment. La répartition de la base de données en ensemble d'apprentissage et ensemble de test doit être réalisée en cohérence avec l'objectif applicatif. Par exemple, le système doit-il être dépendant du locuteur ? Si tel n'était pas le cas, il serait alors primordial de s'assurer que le locuteur de l'échantillon testé est présent dans la base de données utilisée pour l'apprentissage.

L'article [SCH 04] 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.

1.6. Conclusion

Les attentes dans le domaine de la reconnaissance automatique des émotions dans la voix sont fortes, et le domaine émergeant. Les travaux existants commencent à

prendre en compte les différents problèmes scientifiques soulevés en vue de l'utilisation effective de tels systèmes dans des applications. Un bout du chemin a été parcouru, mais il reste de multiples enjeux à traiter. Avant de présenter ces enjeux, nous récapitulons ci-dessous les étapes 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) (premier facteur du paragraphe 1.5) ;

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) (second facteur du paragraphe 1.5) ;

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) (second facteur du paragraphe 1.5) ;

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) (troisième facteur du paragraphe 1.5) ;

5) L'*apprentissage par le système* d'un problème de classification par le choix et le paramétrage de méthodes de classification (quatrième facteur du paragraphe 1.5) ;

6) L'évaluation des performances du système selon des *protocoles* définis compte tenu des contraintes applicatives (conditions de dépendance ou d'indépendance au locuteur, au contexte, etc.) (cinquième facteur du paragraphe 1.5) ;

Selon les applications, il ne s'agira pas de faire de la *reconnaissance* d'émotions (systèmes de dialogues) mais plutôt de la *détection* d'émotions (surveillance, gestion de crise, applications médicales, robotique, etc.). Les tours de parole ne sont également pas toujours clairement séparés en fonction de l'application. Un des défis qu'il reste à relever est le traitement des différents types de recouvrements entre locuteurs. Il s'agit de les prendre en compte au sein d'une stratégie d'annotation et de permettre au système de reconnaissance de les traiter.

Par ailleurs, les systèmes de reconnaissance acoustique des émotions peuvent être utilisés en parallèle d'un système de reconnaissance basé sur des indices vidéo (voir chapitre 4). Les seuls indices acoustiques ne sont souvent pas suffisants pour la détection d'émotion [CLA 04]. Il est par exemple difficile – pour l'homme et a fortiori pour le système – de distinguer un cri de joie d'un cri de peur. La vidéo et plus largement le contexte (la situation, le contenu linguistique, i.e. ce qui est dit) fournissent

une information supplémentaire voire même nécessaire à la reconnaissance automatique d'émotions. La corrélation de ces différentes sources d'information est un enjeu important pour la détection d'émotion [ZEN 07].

1.7. Bibliographie

[ALK 02] ALKU P., BCKSTRM T., VILKMAN E., « Normalized amplitude quotient for parametrization of the glottal flow », *Journal of Acoustical Society of America*, vol. 112, n°2, p. 701-710, 2002.

[AMI 96] AMIR N., RON S., « Towards an automatic classification of emotions in speech », *Proc. of ICSLP*, Philadelphie, 1996.

[AUD 04] AUDIBERT N., ROSSATO S., AUBERGÉ V., « Paramétrisation de la qualité de voix : EEG vs. filtrage inverse », *Actes des Journées d'Étude sur la Parole*, Fs, p. 53-56, 2004.

[BAR 03] BARRAS C., GAUVAIN J.-L., « Feature and score normalization for speaker verification of cellular data », *Proc. of ICASSP*, Hong-Kong, 2003.

[BAT 00] BATLINER A., FISHER K., HUBER R., SPILKER J., NTH E., « Desperately seeking emotions or : Actors, wizards and human beings », *Proc. of ISCA Workshop on Speech and Emotion*, Belfast, p. 195-200, 2000.

[BAT 03] BATLINER A., FISCHER K., HUBER R., SPILKER J., NÖTH E., « How to find trouble in communication », *Speech Communication*, vol. 40, n°1-2, p. 117-143, 2003.

[BAT 06] BATLINER A., STEIDL S., SCHULLER B., SEPPI D., LASKOWSKI K., VOGT T., DEVILLERS L., VIDRASCU L., AMIR N., KESSOUS L., AHARONSON V., « Combining Efforts for Improving Automatic Classification of Emotional User States », *Proc. of IS-LTC*, Ljubljana, 2006.

[BAT 08] BATLINER A., STEIDL S., NTH E., « Releasing a thoroughly annotated and processed spontaneous emotional database : the FAU Aibo Emotion Corpus. », *Proc. of Workshop on Corpora for Research on Emotion and Affect LREC*, Marrakech, p. 28-31, 2008.

[BOE 05] BOERSMA P., WEENINK D., Praat : doing phonetics by computer [Computer program], from http ://www.praat.org/, Rapport, 2005.

[CAI 03] CAI R., LU L., ZHANG H.-J., CAI L.-H., « Highlight Sound Effects Detection in Audio Stream », *Proc. of ICME*, Baltimore, 2003.

[CAM 03] CAMPBELL N., MOKHTARI P., « Voice quality : the 4h Prosodic Dimension », *Proc. of International Congress on Phonetic Sciences*, Barcelone, 2003.

[CAR 96] CARLETTA J., « Assessing agreement on classification tasks : the kappa statistic », *Computational Linguistics*, vol. 22, n°2, p. 249-254, 1996.

[CLA 04] CLAVEL C., VASILESCU I., DEVILLERS L., EHRETTE T., « Fiction Database for Emotion Detection in Abnormal Situations », *Proc. of ICSLP*, Jeju, p. 2277-2280, 2004.

[CLA 07a] CLAVEL C., DEVILLERS L., RICHARD G., VASILESCU I., EHRETTE T., « Abnormal situations detection and analysis through fear-type acoustic manifestations », *Proc. of ICASSP*, Honolulu, 2007.

[CLA 07b] CLAVEL C., Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales, PhD thesis, Doctorat Signal et Images, ENST - TSI Traitement du Signal et des Images, 2007.

[CLA 08] CLAVEL C., VASILESCU I., DEVILLERS L., RICHARD G.AND EHRETTE T., « Fear-type emotions recognition for future audio-based surveillance systems », *Speech Communication*, vol. 50, p. 487-503, 2008.

[COW 03] COWIE R., CORNELIUS R., « Describing the emotional states that are expressed in speech », *Speech Communication*, vol. 40, n°1-2, p. 5-32, 2003.

[CRA 04] CRAGGS R., « Annotating Emotion in dialogue - Issues and Approaches », *Proc. of CLUK Research Colloquium*, 2004.

[CUL 08] CULLEN C., VAUGHAN B., KOUSIDIS S., MCAULEY J., « Emotional Speech Corpus Construction, Annotation and Distribution », *Proc. of Workshop on Corpora for Research on Emotion and Affect LREC*, Marrakech, p. 32-37, 2008.

[DEL 96] DELLAERT F., POLZIN T., WAIBEL A., « Recognizing Emotion in Speech », *Proc. of ICSLP*, Philadelphie, 1996.

[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., « Maximum likelihood from incomplete fata via the EM algorithm », *Journal of the Royal Statistical Society*, vol. 39, n°1, p. 1-38, 1977.

[DEV 05] DEVILLERS L., VIDRASCU L., LAMEL L., « Challenges in real-life emotion annotation and machine learning based detection. », *Journal of Neural Networks*, vol. 18, n°4, p. 407-422, 2005.

[DEV 06] DEVILLERS L., VIDRASCU L., « 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, special issue « Interaction Emotionnelle »*, vol. 20, n°4-5, p. 447-476, 2006.

[DEV 07] DEVILLERS L., VIDRASCU L., « *Speaker characterization* », Chapitre Emotion recognition, Springer-Verlag, 2007.

[DUD 73] DUDA R., HART P. E., *Pattern Classification and Scence Analysis*, Wiley-Interscience, 1973.

[EHR 04] EHRETTE T., Les voix des services telecoms, de la perception la modélisation, PhD thesis, Universit Paris XI, 2004.

[ESS 05] ESSID S., Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique, PhD thesis, Telecom-Paris, 2005.

[FAN 60] FANT G., *Acoustic theory of speech production*, Paris, La Hague, 1960.

[FRA 03] FRANCE D., SHIAVI R., SILVERMAN S., SILVERMAN M., WILKES D., « Acoustical properties of speech as indicators of depression and suicidal risks », *IEEE Transactions on Biomedical Engeneering*, vol. 47, n°7, p. 829-837, 2003.

[FRE 01] FREDOUILLE C., BONASTRE J., MERLIN T., « Bayesian approach based decision in speaker verification », *Proc. of a speaker odyssey, The speaker recognition workshop*, Grce, p. 77-81, 2001.

[GUY 03] GUYON I., ELISSEEFF A., « An introduction to feature and variable selection », *Journal of Machine Learning Research*, vol. 3, 2003.

[HEN 07] VAN HENGEL P.W.J. AND ANDRINGA T., « Verbal aggression detection in complex social environments », *Proc. of IEEE Conference on Advanced Video and Signal Based Surveillance*, Londres, p. 15 - 20, Septembre 2007.

[HES 84] HESS W., O'SHAUGHNESSY D., « Pitch Determination of Speech Signals : Algorithms and Devices », *The Journal of the Acoustical Society of America*, vol. 74, n°4, p. 1277-1278, October 1984.

[HSU 02] HSU C.-W., LIN C.-J., « A comparison of methods for multi-classe support vector machines », *IEEE Transactions on Neural Networks*, vol. 13, n°2, p. 415-425, mars 2002.

[IST 03] ISTRATE D., Dtection et Reconnaissance des Sons pour la Surveillance Médicale, PhD thesis, Universit de Grenoble, 2003.

[KIE 00] KIENAST M., SENDLMEIER W.-F., « Acoustical analysis of spectral and temporal changes in emotional speech », *Proc. of ISCA ITRW on Speech and Emotion*, Belfast, p. 92-97, 2000.

[KLA 08] KLAPURI A., « Multipitch analysis of polyphonic music and speech signals using an auditory model », *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, n°2, p. 255-266, Fvrier 2008.

[KOZ 65] KOZHEVNIKOV V. A., CHISTOVICH I. A., *Speech Production and Perception*, Joint Publication Research Service, Washington, DC, 1965.

[KWO 03] KWON O.-W., CHAN K., HAO J., LEE T.-W., « Emotion Recognition by Speech Signals », *Proc. of Eurospeech*, Genve, p. 125-128, 2003.

[LEE 02] LEE C., NARAYANAN S., PIERACCINI R., « Classifying emotions in human-machine spoken dialogs », *Proc. of ICME*, vol. 1, Lausanne, p. 737- 740, 2002.

[LOO 07] LOOIJE R., NEERINCS M., KRUIJFF G.-J. M., « Affective Collaborative robots for Safety and Crisis Management in the Field », *Proc. of ISCRAM*, p. 497-505, 2007.

[MCA 99] MCADAMS S., « Perspectives on the Contribution of Timbre to Musical Structure », *Comput. Music J.*, vol. 23, n°3, p. 85–102, MIT Press, 1999.

[NGH 07] NGHIEM A., BREMOND F., THONNAT M., VALENTIN. V., « ETISEO, an evaluation project for video surveillance systems », *In Proc. of IEEE International Conference on Advanced Video and Signal based Surveillance*, Londres, p. 476-481, Septembre 2007.

[OSG 75] OSGOOD C., MAI W. H., MIRON M., *Cross-cultural Universals of Affective Meaning*, University of Illinois Press, Urbana, 1975.

[OUD 03] OUDEYER P.-Y., « The production and recognition of emotions in speech : features and algorithms », *International Journal of Human Computer Interaction, special issue on Affective Computing*, vol. 59, n°1-2, p. 157-183, 2003.

[PEE 03] PEETERS G., « Automatic Classification of Large Musical Instrument Databases Using Hierachical Classifiers with Inertia Ratio Maximization », *In Proc. of AES 115th Convention*, New-York, USA, p. 1-13, Octobre 2003.

[PET 03] PETRUSHIN V., « Emotion in Speech : Recognition and Application to Call Center », *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, p. 7-10, 2003.

[PIC 97] PICARD R., *Affective Computing*, MIT Press, Cambridge, MA., 1997.

[SAN 05] SANCHEZ-SOTO E., Réseaux Bayesiens Dynamiques pour la Vérification du Locuteur, PhD thesis, Tlcom-Paris, 2005.

[SCH 98] SCHERER K. R., JOHNSTONE T., SANGSUE J., « L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole », *Actes des Journes d'Études sur la Parole*, Matigny, p. 249-257, 1998.

[SCH 03a] SCHERER K., « Vocal communication of emotion : a review of research paradigms », *Speech Communication*, vol. 40, n°1-2, p. 227-256, 2003.

[SCH 03b] SCHULLER B., RIGOLL G., LANG M., « Hidden Markov Model-based Speech Emotion Recognition », *Proc. of ICASSP*, Hong Kong, p. 1-4, 2003.

[SCH 04] SCHULLER B., RIGOLL G., LANG M., « Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture », *Proc. of ICASSP*, vol. 1, Montreal, p. I- 577-80, 2004.

[SER 84] SERKEY A., HANSON B., « Improved 1-Bark Auditory Filter », *Journal of Acoustical Society of America*, vol. 75, n°6, p. 1902-1904, 1984.

[SHA 03] SHAFRAN I., RILEY M., MOHRI M., « Voice Signatures », *Proc. of ASRU Workshop*, St Thomas, p. 31- 36, 2003.

[SHA 07] SHAMI M., VERHELST W., « An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech », *Speech Commun.*, vol. 49, n°3, p. 201–212, Elsevier Science Publishers B. V., 2007.

[STE 05] STEIDL S., LEVIT M., BATLINER A., NTH E., H. N., « "Of all things the measure is man", Automatic classification of emotions and inter-labeler consistency », *Proc. of ICASSP*, Philadelphie, p. 317 - 320, 2005.

[VAP 95] VAPNIK V., *The Nature of Statistical Learning Theory*, Springer, 1995.

[VAR 06] VARADARAJAN V., HANSEN J., AYAKO I., « UT-SCOPE - A corpus for Speech under Cognitive/Physical task Stress and Emotion », *Proc. of LREC Workshop en Corpora for Research on Emotion and Affect*, Gnes, p. 72-75, 2006.

[VER 03] VERFAILLE V., Effets audionumriques adaptatifs - théorie, mise en oeuvre et usage en création musicale numérique, PhD thesis, Universit Aix-marseille II, 2003.

[VID 05] VIDRASCU L., DEVILLERS L., « Detection of real-life emotions in call centers », *Proc. of Eurospeech*, Lisbonne, p. 1841-1844, 2005.

[VLA 07] VLASENKO B., SCHULLER B., WENDEMUTH A., RIGOLL G., « Frame vs. Turn-Level : Emotion Recognition from Speech Considering Static and Dynamic Processing », *Affective Computing and Intelligent Interaction*, Lisbonne, Portugal, p. 139-147, Septembre 2007.

[WRE 03] WRED B., SHRIBERG E., « Spotting 'Hot Spots' in Meetings : Human Judgements and Prosodic Cues », *Proc. of Eurospeech*, Gnes, p. 2805-2808, 2003.

[YEG 98] YEGNANARAYANA B., D'ALESSANDRO C., DARSINO V., « An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components », *IEEE Transactions on Speech and Audio Processing*, vol. 6, n°1, p. 1-11, 1998.

[ZEN 07] ZENG Z., PANTIC M., ROISMAN G. I., HUANG T. S., « A survey of affect recognition methods : audio, visual and spontaneous expressions », *Proc. of the 9th international conference on Multimodal interfaces*, New York, NY, USA, ACM, p. 126–133, 2007.

[ZWI 57] ZWICKER E., FLOTTORP G., STEVENS S., « Critical bandwidth in loudness summation », *J. Acoust. Soc. Am.*, vol. 29, p. 548-557, 1957.