Détection des états affectifs lors d'interactions parlées : robustesse des indices non verbaux

Laurence Devillers[*,**] — **Marie Tahon**[*,***] — **Mohamed A. Sehili**[*] — **Agnes Delaborde**[*,**]

[*] *LIMSI-CNRS, 91403 Orsay, France*
[**] *Université Paris-Sorbonne IV, 28 rue Serpente, 75006 Paris, France*
[***] *Conservatoire national des arts et métiers, 292 rue St Martin, 75141 Paris Cedex 3, France*
{laurence.devillers, marie.tahon, mohamed.sehili, agnes.delaborde} @limsi.fr

RÉSUMÉ. *Dans un contexte d'interaction homme-machine, les systèmes de détection des émotions dans la voix doivent être robustes aux variabilités et efficaces en temps de calcul. Cet article présente les performances que nous pouvons obtenir en utilisant uniquement des indices paraverbaux (nonverbaux). Nous proposons une méthodologie pour sélectionner les familles de paramètres robustes, en étudiant trois ensembles de descripteurs testés sur trois corpus de données spontanées collectés dans des contextes d'interaction homme-machine. Le résultat de notre étude met en avant les paramètres perceptifs liés à l'énergie spectrale (énergie par bandes de Bark), en obtenant des performances de détection sur quatre émotions au niveau de l'ensemble des descripteurs de référence du Challenge Interspeech 2009.*

ABSTRACT. *In a Human-Machine Interaction context, automatic in-voice affective state detection systems have to be robust to variabilities and computationally efficient. This paper presents the performance that can be reached using para-verbal (non-verbal) cues. We propose a methodology to select robust parameters families, based on the study of three sets of descriptors, and tested on three different corpora of spontaneous data collected in Human-Machine Interaction contexts. The key finding of our study puts forward perceptive parameters linked to spectral energy, particularly energy on Bark bands, which yield the same performance on a four-emotion detection task as the reference set of descriptors used in the Interspeech 2009 challenge.*

MOTS-CLÉS : *détection des émotions, paramètres non verbaux, énergie par bandes de Bark.*

KEYWORDS: *emotion detection, non verbal parameters, energy on Bark bands.*

1. Introduction

Dès lors que l'on s'intéresse aux échanges communicatifs, il est impossible de faire l'impasse sur la composante affective. Dans de précédentes études, des données ont été collectées dans différents contextes d'interaction pour étudier les canaux verbaux et paraverbaux des émotions dans la communication (Devillers *et al.*, 2005 ; Campbell, 2007 ; Han *et al.*, 2012). Si les activités des locuteurs au niveau verbal s'efforcent tant bien que mal de contrôler les événements émotionnels, les comportements non verbaux ou paraverbaux trahissent souvent un état émotionnel bien présent et quelquefois désynchronisé des paroles émises. La reconnaissance de parole spontanée en interaction est un sujet difficile qui nécessite encore de nombreuses améliorations (Bazillon *et al.*, 2008). Il est donc intéressant, d'un point de vue applicatif, d'évaluer également la robustesse des paramètres paraverbaux. Cette étude porte sur les descripteurs non verbaux sans alignement avec le canal verbal.

Détecter le comportement affectif uniquement dans la voix est un sujet complexe. En effet, la voix, lorsqu'elle véhicule de la parole, est porteuse également de nombreux indices sur l'individu, sur son état physique et mental et sur son état affectif. La voix est très variable d'un individu à un autre et pour un même individu d'une situation à une autre. De nombreux travaux sont menés sur ce sujet depuis plusieurs décennies en psychologie (Scherer, 1986), phonétique et sciences du langage avec un développement important ces trois dernières décennies des approches interactionnistes (Plantin *et al.*, 2000) et depuis une quinzaine d'années en modélisation informatique (Fernandez et Picard, 2003).

La reconnaissance automatique des émotions peut être utilisée pour piloter les stratégies d'un agent conversationnel ou d'un robot. Afin de répondre à cela, il est primordial de collecter et d'analyser des comportements émotionnels dans des contextes réalistes d'interaction avec des utilisateurs potentiels de ces technologies (Devillers et Vidrascu, 2007 ; Devillers *et al.*, 2010 ; Campbell, 2007 ; Han *et al.*, 2012). Les émotions étudiées sont souvent annotées en petit nombre à partir d'étiquettes verbales, comme par exemple la joie, la colère, la tristesse, ou annotées à partir de dimensions continues telle que l'activation (état passif ou actif) et la valence (état positif ou négatif). Ces émotions et ces dimensions affectives détectées automatiquement par un robot lors d'interactions vont lui permettre de modéliser dynamiquement les profils des utilisateurs afin de modifier les stratégies expressives de réponses et les actions en fonction du comportement de l'utilisateur (Delaborde et Devillers, 2010 ; Tahon *et al.*, 2011 ; Buendia et Devillers, 2014).

Il existe un nombre de plus en plus grand d'études dans ce domaine, pourtant pour l'instant aucun standard n'est vraiment appliqué ni en annotation de données, ni en calcul de paramètres. Les seuls efforts portent sur l'évaluation grâce aux défis menés sur des corpus mis à la disposition des chercheurs afin de comparer les performances de leurs systèmes lors des conférences Interspeech de 2009 à 2013 (Schuller *et al.*, 2009a ; Schuller *et al.*, 2010a ; Schuller *et al.*, 2011b ; Schuller *et al.*, 2012). Grâce à ces défis, une communauté scientifique partage un savoir-faire, des outils et

quelques standards d'évaluation. Il faut également mentionner que peu de ces corpus sont collectés en situations réalistes car il est difficile de les rendre libres de droits. Il est pourtant nécessaire de tester les systèmes de détection des émotions sur des corpus spontanés avec de nombreux locuteurs de tous âges. Malgré la mise en place de challenges, on peut constater un manque d'analyse des résultats obtenus en termes de robustesse des paramètres à utiliser. Mais en l'absence de standard pour calculer ces paramètres, il est difficile également de comparer les approches dans la communauté et d'en tirer des analyses pertinentes sur les paramètres non verbaux. La stratégie pour l'instant des chercheurs est souvent de rajouter des paramètres pour améliorer les performances de détection des émotions. Cette approche n'est pas réaliste pour construire des systèmes de détection en temps réel.

Notre but de recherche est de trouver les paramètres non verbaux les plus robustes pour représenter les émotions pour différents types de voix et de contextes mais surtout de minimiser le nombre de familles retenues par souci de temps de calcul dans un système de détection en temps réel. Nous définissons une famille de paramètres acoustiques, prosodiques ou microprosodiques par l'ensemble des paramètres calculés à partir d'une fonction temporelle de bas niveau (énergie, fréquence fondamentale, formants, etc.). Cette fonction temporelle peut être calculée avec différents algorithmes suivant les librairies utilisées. Pour sélectionner les familles de paramètres, il est essentiel de mener des études sur des données spontanées, sur plusieurs corpus, avec plusieurs ensembles d'indices pour améliorer nos connaissances des paramètres non verbaux les plus à même de représenter les inflexions émotionnelles que nous sommes capables de percevoir dans la voix.

Nous présentons dans cet article une étude de plusieurs ensembles et familles de paramètres pour reconnaître quatre états émotionnels : joie, tristesse, colère et un état neutre. Nos travaux s'appuient sur trois collectes de données spontanées menées dans le cadre de projets nationaux : le projet ANR affective avatar (2004-07) (Schuller et Devillers, 2010 ; Brendel *et al.*, 2010), le projet FUI ROMEO (2007-12) (Tahon et Devillers, 2010 ; Tahon *et al.*, 2012b ; Delaborde et Devillers, 2010 ; Delaborde *et al.*, 2009 ; Delaborde et Devillers, 2012) et enfin le projet ANR Tecsan Armen (2009-12) (Chastagnol et Devillers, 2012). L'ensemble de ces données couvre plus de cent cinquante voix de personnes enregistrées dans des conditions d'interaction réalistes. L'âge des personnes varie entre 16 et plus de 90 ans. Nous utilisons également plusieurs librairies *open source* de la communauté : openSMILE[1], Yaafe[2], Aubio[3] , Praat (Boersma, 1993) pour l'analyse des paramètres et Weka (Hall *et al.*, 2009) pour l'apprentissage des modèles afin que les expériences puissent être répliquées.

Les conclusions des tests menés dans cet article montrent l'intérêt de l'approche *pluri-* et *cross-corpus* avec plusieurs librairies d'analyse. Les paramètres liés à l'énergie semblent les plus efficaces et notamment l'énergie par bandes de Bark qui a été

1. http ://openSMILE.sourceforge.net
2. http ://yaafe.sourceforge.net
3. http ://aubio.org

robuste face aux différentes évaluations permettant des performances au niveau de celles obtenues avec l'ensemble utilisé comme référence dans le Challenge Interspeech 2009.

La section 2 présente un état de l'art, la section 3 est dédiée aux ensembles d'indices non verbaux, enfin la section 4 décrit les corpus, le protocole de test et les résultats des expériences en *cross-validation* et en *cross-corpus*. Les conclusions et perspectives sont regroupées dans la section 5.

2. État de l'art

2.1. *Corpus*

L'enjeu d'un système de détection des émotions est d'intégrer des émotions dans les systèmes computationnels. Les corpus émotionnels sont d'une importance capitale pour la mise au point de ces nouvelles technologies d'interaction homme-machine. Rappelons qu'un corpus est un ensemble de données recueillies et annotées pour étudier une question de recherche spécifique. On peut différencier quatre types de corpus utilisés dans la communauté de recherche sur les émotions contenant :

1) des émotions actées (jouées par des acteurs professionnels ou non) hors contexte et actées en contexte de fiction (film, théâtre) (Clavel *et al.*, 2007) : les corpus disponibles dans la communauté sont souvent de petite taille avec peu de locuteurs, quelques-uns même avec un contenu linguistique fixe. Parmi les corpus les plus cités, on peut faire références aux corpus actés en allemand et danois le *Berlin Emotional Speech Database* (Burkhardt *et al.*, 2005), *the Danish Emotional Speech Database* (Engberg *et al.*, 1997) comprenant respectivement quatre et dix locuteurs. En français, on peut se référer au corpus de données actées GEMEP (Bänziger *et al.*, 2012) ;

2) des émotions spontanées induites : les corpus issus d'interfaces artificielles avec, par exemple, un robot permettant d'induire des émotions spontanées en mode magicien d'Oz (*Wizard of Oz* (WoZ) : interaction avec un système simulé par un humain) (Delaborde *et al.*, 2009 ; Delaborde *et al.*, 2010 ; Delaborde et Devillers, 2012 ; Tahon *et al.*, 2011 ; Grimm *et al.*, 2008 ; Soury et Devillers, 2012 ; Chastagnol *et al.*, 2013) ou les systèmes d'interaction avec des machines (sans WoZ) comme les corpus SEMAINE (McKeown *et al.*, 2012), Herme (Han *et al.*, 2012) ;

3) des émotions spontanées : les corpus *real life* d'interaction entre humains (Devillers *et al.*, 2005 ; Devillers *et al.*, 2010).

Les bases de données émotionnelles disponibles dans la communauté sont référencées dans plusieurs articles qui synthétisent les travaux réalisés sur la reconnaissance des émotions (Ververidis et Kotropoulos, 2003 ; Cowie *et al.*, 2005 ; Schuller *et al.*, 2010b ; Batliner *et al.*, 2011). La plupart des expériences sur les émotions ont été effectuées sur des données artificielles, enregistrées par des acteurs avec, souvent, peu de classes émotionnelles parmi les émotions primaires définies par (Ekman, 1999) : colère, joie, tristesse, peur, dégoût, surprise. Suivant l'étude de Justin et Laukka Jus-

tin et Laukka (2003), sur cent quatre expériences d'études sur les émotions, 87 % sont des expressions produites par des acteurs. Les résultats obtenus dans la communauté scientifique se rapportent donc majoritairement à des données actées, en témoignent les workshops *corpora for Research on Emotion and Affect* qui encouragent les chercheurs à utiliser des données réalistes (Devillers *et al.*, 2006 ; Devillers et Martin, 2008 ; Devillers *et al.*, 2010).

2.2. *Détection des émotions*

Un système de détection des signaux affectifs est composé de trois modules : un premier module de traitement du signal, un module de détection à partir d'un modèle obtenu par apprentissage et enfin un module d'évaluation. Le premier module est composé d'un segmenteur du signal, d'un extracteur de séries temporelles bas niveau : fréquence fondamentale F_0, *Mel Frequency cepstral coefficients* (MFCCs), etc., et d'un module de calcul des fonctionnelles (variations mélodiques maximales, plage de variation de l'intensité, moyenne des MFCCs...) sur les séries temporelles. Le deuxième module consiste à prédire un score de reconnaissance des formes en utilisant un ou plusieurs modèles obtenus par apprentissage à l'aide de méthodes statistiques : *Support Vector Machine* (SVM), *Neural Network* (NN), *Gaussian Mixtures Models* (GMM), *Hidden Markov Models* (HMMs). Enfin, le dernier module est un module d'évaluation. Il s'agit de comparer l'émotion prédite par le système avec l'émotion annotée par un humain et de calculer un score de bonne reconnaissance. Plusieurs mesures comme le taux de classification WA *Weighted Accuracy* qui représente le nombre d'instances bien reconnues par rapport au nombre total d'instances ou le taux de classification UA *Unweighted Accuracy* qui est le même calcul effectué sur des ensembles équilibrés (avec le même nombre d'échantillons dans chaque classe), sont régulièrement utilisées pour évaluer les performances des systèmes. La F-mesure est également largement utilisée, elle combine la précision et le rappel et permet de pondérer le taux de reconnaissance par le taux de faux positif (cf. équation 1). Cette mesure très largement utilisée dans la communauté permet de prendre en compte à la fois le rappel et la précision. Nous utiliserons cette mesure pour les expériences décrites dans cette article tout en donnant les taux de reconnaissance (*Accuracy*).

$$F = \frac{2 \cdot (\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})} \quad [1]$$

Depuis 1997, de nombreux travaux ont été menés sur la détection des émotions avec des boîtes à outils *open source* permettant de calculer jusqu'à quelques milliers de paramètres acoustiques, citons par exemple le système openSMILE (Schuller *et al.*, 2009b) permettant d'entraîner des modèles par apprentissage, ou encore la plateforme Weka (Hall *et al.*, 2009). Peu de ces travaux ont été effectués pour une application particulière avec un système de détection en temps réel utilisant des modèles entraînés sur des corpus de données spontanées ou réalistes. Les scores de détection obtenus dans la communauté sont très variables et sont totalement liés aux données du

corpus. Sur des corpus de données actées, ils peuvent atteindre plus de 90 % de bonne détection mais de tels systèmes seront très mauvais pour prédire des émotions dans des contextes naturels. Avec des données réalistes, les taux sont plus faibles. Notre étude porte sur l'utilisation de plusieurs corpus collectés en situations réalistes pour tester la robustesse des indices acoustiques avec une évaluation en *cross-validation* et en *cross-corpus* sous Weka.

2.3. *Études* pluri- *et* cross-corpus

Pour améliorer les performances des systèmes et les adapter à des applications réelles, il existe de nombreux défis de recherche liés à la généricité et à la robustesse des signaux comme par exemple la prise en compte de signaux multimodaux et la prise en compte du contexte (types de locuteurs, types de tâches, etc.). L'amélioration des performances est également liée à l'utilisation de corpus de données de grande taille et de corpus réalistes. Pour pallier le manque de données, des expériences *cross-corpus* sont actuellement menées.

Effectivement, les expériences portant sur un seul corpus ne permettent pas d'estimer l'influence des sources de variabilités sur la reconnaissance des émotions puisque ces variabilités sont maintenues constantes (par exemple la salle, le type de locuteur, le scénario, le contenu linguistique et paralinguistique, etc.). Plusieurs études *pluri-* et *cross-corpus* ont été réalisées, généralement en utilisant un ou plusieurs corpus pour l'apprentissage et un corpus de test. Ce protocole entraîne une baisse des taux de reconnaissance (Schuller *et al.*, 2010c) mais ces taux sont plus réalistes. Plusieurs expériences ont été menées soit en mélangeant plusieurs corpus pour l'apprentissage (Schuller *et al.*, 2010b ; Eyben *et al.*, 2010 ; Marchi *et al.*, 2012 ; Devillers *et al.*, 2010) soit en adaptant un corpus de référence de manière non supervisée (Zhang *et al.*, 2011), soit encore en faisant un vote majoritaire des résultats obtenus avec un corpus en apprentissage et plusieurs corpus de test (Schuller *et al.*, 2011c ; Tahon *et al.*, 2011 ; Sun et Moore, 2013). Notre étude porte sur une expérience *pluri-corpus* avec une méthodologie de *cross-validation* et de *cross-corpus* pour l'évaluation.

3. Analyse non verbale

Les descripteurs paraverbaux utilisés en détection des émotions (Scherer, 1986) sont empruntés au domaine de la prosodie, de la phonétique, de la reconnaissance de la parole, du discours ou encore de la musique et ont servi, par exemple, à mesurer différents aspects de la phonation et de l'articulation. L'ensemble de ces descripteurs regroupe des paramètres dans le domaine fréquentiel (par exemple, la fréquence fondamentale F_0), dans le domaine de l'amplitude (par exemple, l'énergie), dans le domaine temporel (par exemple, le rythme) et dans le domaine spectral (par exemple, l'enveloppe spectrale ou l'énergie par bandes spectrales). De nombreuses études (Schuller *et al.*, 2010c ; Scherer, 1986) montrent l'intérêt de combiner ces paramètres.

Nous présentons dans cette partie trois ensembles de descripteurs obtenus à partir de différentes librairies qui regroupent des paramètres de ces différents domaines. Ils seront utilisés pour mener des tests sur trois corpus dans la section « Expériences » de ce document. Les descripteurs sont calculés en grande majorité grâce aux librairies *open source* : openSMILE, Yaafe, Aubio et Praat. Les paramètres connus dans la communauté pour être important pour la reconnaissance des émotions sont utilisés dans ces ensembles. Ces trois ensembles se recoupent pour certains paramètres, cependant ils ont chacun leurs spécificités. Les trois tableaux présentés plus bas détaillent le contenu des descripteurs classés par famille. Nous avons sélectionné huit grandes familles de descripteurs alliant prosodie (fréquence fondamentale, énergie et qualité vocale), spectre (descripteurs d'enveloppe spectrale et énergie par bandes spectrales), cepstre (MFCCs surtout utilisé en reconnaissance de la parole), nombre de passage par zéros (ZCR, *zero crossing rate*) et formants.

3.1. *Prosodie*

La fréquence fondamentale F_0 (ou *pitch*) peut être extraite avec les outils Praat, Aubio ou openSMILE suivant différents algorithmes avec différents protocoles de gestion des erreurs d'extraction. Afin de s'approcher au mieux de la perception humaine, la fréquence fondamentale peut être donnée en semi-tons (échelle logarithmique) : $F_0(st) = \frac{12}{\ln 2} \ln F_0(Hz)$ (référence à 1 Hz). La fréquence fondamentale correspond à la fréquence de vibration des cordes vocales, plusieurs descripteurs peuvent en découler et apporter des informations complémentaires. Plusieurs études montrent que la famille liée à l'intonation est importante pour la reconnaissance des émotions mais un certain nombre d'auteurs soutiennent que des modèles d'intonation spécifiques reflètent des émotions spécifiques, tandis que d'autres ne partagent pas cette idée et pensent que la F_0 est continuellement, plutôt que catégoriquement, affectée par les émotions.

L'énergie est très importante dans la reconnaissance des émotions, elle est incluse dans tous les outils d'extraction de paramètres. La forme la plus courante est l'énergie ou l'intensité RMS, cependant elle ne correspond pas à ce que perçoit l'oreille humaine contrairement à l'énergie perçue ou *loudness*. Le paramètre *loudness* global est calculé à partir des *loudness* spécifiques par bandes et correspond à l'énergie du signal à laquelle on applique un filtre de l'oreille humaine (ici un filtre par bandes de Bark). Ce descripteur est généralement donné en décibels.

La qualité vocale est une problématique complexe qui ne fait pas consensus. Quatre paramètres sont souvent utilisés et peuvent être calculés de différentes façons. Le *jitter* (resp. le *shimmer*) mesure les microvariations de la fréquence fondamentale (de l'énergie) suivant les périodes de fermeture de la glotte. Ces paramètres sont plus couramment utilisés sur des voix pathologiques. Le rapport harmonique sur bruit (HNR) permet d'avoir des informations sur le voisement du signal, cependant ce paramètre dépend très fortement du bruit de fond. Le rapport entre la durée des parties

voisées et la durée des parties non voisées (*punvoiced*) peut être également considéré comme un paramètre de rythme. Plus il est élevé, plus les phonèmes voisés seront longs par rapport aux phonèmes non voisés. Les paramètres de qualité vocale sont assez nombreux dans la librairie de Praat, et également dans les dernières versions d'openSMILE et de Yaafe. La qualité vocale a montré son intérêt pour la reconnaissance de la valence notamment (Tahon *et al.*, 2012a ; Gendrot, 2004).

3.2. *Spectre*

Les descripteurs d'enveloppe spectrale cherchent à représenter le timbre d'un signal. Il existe de nombreux descripteurs sous les librairies Yaafe et openSMILE. Beaucoup de ces descripteurs ont été initialement développés pour l'analyse de signaux musicaux (Peeters, 2004). Les descripteurs spectraux peuvent être extraits directement sur le résultat d'une transformée de Fourier rapide (*Fast Fourier Transform* (FFT)), ou bien sur une enveloppe spectrale. Il existe différentes manière de déterminer l'enveloppe spectrale : l'estimation de la *true-envelope* (Villavicencio *et al.*, 2009) ou une estimation à partir des coefficients LPC (*Linear Predictive Coding*). Cette dernière opération permet d'obtenir des informations relatives au conduit vocal (Ruiz *et al.*, 2008). Les fréquences *roll-off* (5 %, 25 %, 50 %, 75 % et 95 %) ainsi que les pentes spectrales (sur tout le spectre, sur 0-500 Hz et sur 500-1 500 Hz) et le barycentre spectral permettent de décrire au mieux la forme du spectre. Ces descripteurs ont été initialement conçus pour l'étude de timbres musicaux.

L'énergie par bandes spectrales peut prendre des formes très diverses suivant les librairies. Les bandes peuvent être linéaires en Hertz (Batliner *et al.*, 2006), centrées sur des fréquences fixes ou bien dépendantes de la fréquence fondamentale comme dans les bandes harmoniques (Xiao, 2008). Elles peuvent être logarithmiques ou sur des échelles perceptives Mel ou Bark. Pour aller jusqu'à la fréquence de 8 000 Hz, vingt et une bandes de Bark sont suffisantes. L'énergie est généralement donnée en décibels à partir du carré de l'amplitude spectrale du signal, mais elle peut être aussi définie par des modèles de perception plus complexes (Moore *et al.*, 1997). La plupart des librairies contiennent plusieurs paramètres d'énergie par bandes spectrales. Yaafe n'utilise que les *loudness* spécifiques, alors que openSMILE calcule les bandes de Bark.

3.3. *Cepstre*

Les descripteurs cepstraux, et particulièrement les MFCCs, sont très largement utilisés en traitement automatique de la parole et plus spécifiquement en reconnaissance de la parole. Ils sont connus pour leur robustesse au bruit de fond et leur forte dépendance au contenu lexical. Alors que la plupart des recherches sur les émotions dans la voix utilisent les coefficients cepstraux (Dumouchel *et al.*, 2009 ; Schuller *et al.*, 2011a), ils ne sont pas forcément utiles pour toutes les applications (Tahon

et al., 2011). Les coefficients MFCCs représentatifs du cepstre peuvent être extraits grâce aux différentes librairies Praat, Yaafe ou openSMILE.

3.4. *Autres*

Les formants sont principalement utilisés pour la reconnaissance automatique des phonèmes. Les valeurs des fréquences de résonance du conduit vocal ne semblent pas être pertinentes pour la reconnaissance des émotions, cependant les différences entre les formants (en semi-tons) peuvent l'être (Batliner *et al.*, 2006).

Le ZCR est généralement inclus dans les librairies d'extraction de paramètres audio. Il a l'avantage d'être de très bas niveau et donc facilement implémentable. Ce paramètre permet d'avoir des informations sur le voisement, la quantité de bruit d'un signal, etc.

3.5. *Ensemble 1 (E1-384-IS09)*

Cet ensemble de 384 paramètres (cf. tableau 1) est largement utilisé dans la communauté internationale et permet de situer les performances. Ces 384 paramètres sont extraits avec l'outil openSMILE et ont été utilisés pour le Challenge Interspeech 2009 sur la détection des émotions (Schuller *et al.*, 2009a). Il contient plusieurs paramètres de bas niveau : valeur de la fréquence fondamentale, probabilité de voisement, ZCR, énergie RMS (énergie brute non pondérée par un filtre perceptif). Cet ensemble a la particularité d'inclure les dérivées (Δ) de l'ensemble des descripteurs de bas niveau. À tous les paramètres de bas niveau, douze fonctionnelles sont appliquées : maximum, minimum, étendue (différence entre le maximum et le minimum), positions des maxima et minima, moyenne, coefficients 1 et 2 d'une régression linéaire, ainsi que le coefficient de corrélation Q, écart-type (std), *kurtosis* (estimation de l'étalement d'une distribution autour de la valeur moyenne), *skewness* (estimation de l'asymétrie d'une distribution autour de sa valeur moyenne).

3.6. *Ensemble 2 (E2-334-LIMSI)*

Cet ensemble (Sehili, 2013) de 334 paramètres (cf. tableau 2) utilise deux autres librairies (Yaafe et Aubio). Il est principalement axé sur le descripteur d'énergie perçue (ou *loudness*) (240 paramètres sur 334) extrait sur vingt-quatre bandes de Bark. Il contient également plusieurs descripteurs de qualité vocale issus des calculs de *jitter* et *shimmer*. Outre les statistiques classiques, un ensemble de dix fonctionnelles est appliqué aux descripteurs de bas niveau : maximum, minimum, moyenne, médiane, écart-type (std), *kurtosis* (estimation de l'étalement d'une distribution autour de la valeur moyenne), *skewness* (estimation de l'asymétrie d'une distribution autour de sa valeur moyenne), pente (pente de la distribution), barycentre (barycentre de la dis-

LLD	Fonctionnelles	Voisée
Qualité vocale		**0**
Probabilité de voisement	12 fonc.	12
ΔProbabilité de voisement	12 fonc.	12
F_0 (Hz)	12 fonc.	12
ΔF_0 (Hz)	12 fonc.	12
Fréquence fondamentale F_0		**48**
Énergie RMS	12 fonc.	12
ΔÉnergie RMS	12 fonc.	12
Énergie		**24**
Spectre		**0**
Énergie par bandes spectrales		**0**
MFCC 1-12	12 fonc.	144
ΔMFCC 1-12	12 fonc.	144
Cepstre		**288**
Formants		**0**
ZCR	12 fonc.	12
ΔZCR	12 fonc.	12
ZCR	.	**24**

Tableau 1. *Ensemble 1 des descripteurs paraverbaux (E1-384-IS09). Douze fonc. : min., max. ,étendue, position max., position min., moy., régression linéaire coefficients 1 et 2, corrélation Q, std,* kurtosis*,* skewness

tribution), étendue (estimation de l'étalement de la distribution pondéré autour de sa valeur moyenne).

3.7. *Ensemble 3 (E3-293-LIMSI)*

Nous proposons un ensemble hybride (perceptif et cepstral) (Tahon, 2012) de 293 descripteurs (cf tableau 3). Cet ensemble regroupe un certain nombre de descripteurs issus de la communauté scientifique, que ce soit dans le domaine de la parole et des émotions, ou dans le domaine musical. Certains descripteurs sont redondants, ce qui permet de les comparer afin d'étudier ceux qui sont les plus robustes. Les descripteurs sont généralement une combinaison entre des fonctions de bas niveau (LLD) et des fonctionnelles (fonctions statistiques standard) appliquées sur les LLD. Étant donné que les descripteurs ont des comportements différents suivant le type de signal, nous proposons de les extraire sur trois types de signaux simples à identifier : les signaux voisés, non voisés, ou bien l'ensemble du signal sonore. Les descripteurs sont extraits pour chaque sous-partie du signal et ensuite moyennés sur l'ensemble des

LLD	Fonctionnelles	Voisée	Tout
Jitter absolu	moy./std	2	
Jitter relatif	moy./std	2	
Shimmer	moy./std	2	
Shimmer (dB)	moy./std	2	
Punvoiced			1
Qualité vocale		**8**	**1**
F_0	moy./std × (moy., range, max., min., médiane, pente)	12	
VoicedFrames			1
VoiceBreaks			1
VoiceDegreeOfBreaks			1
Fréquence fondamentale F_0		**12**	**3**
Énergie		**0**	**0**
Roll Off 95%	10 fonc.		10
Décroissance spectrale	10 fonc.		10
Variation spectrale	10 fonc.		10
Étalement spectral	10 fonc.		10
Acuité perceptive (barycentre sur *loudness*)	10 fonc.		10
Étendue perceptive (sur *loudness*)	10 fonc.		10
Spectre		**0**	**60**
Loudness spécifique 0-24 (dB)	10 fonc.		240
Énergie par bandes spectrales		**0**	**240**
Cepstre		**0**	**0**
Formants		**0**	**0**
ZCR	10 fonc.		**10**

Tableau 2. *Ensemble 2 des descripteurs paraverbaux (E2-334-LIMSI). Dix fonc. : min., max., moy., médiane, std,* kurtosis*,* skewness*, pente, barycentre, étendue*

sous-parties d'un même type. Par exemple, on fera la moyenne sur les parties voisées du minimum de la fréquence fondamentale obtenu sur chaque partie voisée. Afin de réduire les erreurs d'estimation sur la F_0, seules les parties voisées de plus de 4 ms sont considérées. Pour chaque partie voisée, la moyenne, l'écart-type (std), le minimum et le maximum de la fréquence fondamentale sont calculés. À ces indices classiques, nous avons ajouté le glissando $G = \frac{F_{max} - F_{min}}{T_{max} - T_{min}}$ (pour une fréquence en semi-tons) qui correspond à la phase descendante de la courbe d'intonation (t Hart, 1981), qui est la plus perceptible et les variations au sein d'une même partie voisée (interF0) et entre

deux parties voisées (intraF0). Les descripteurs de qualité vocale sont extraits grâce au logiciel Praat (Boersma, 1993).

LLD	Fonctionnelles	Voisée	Non voisée	Tout
JitterLocalPraat				1
ShimmerLocalPraat				1
HNRPraat				1
PunvoicedPraat				1
Qualité vocale				**4**
F_0 (st)	4 fonc.	4		
Glissando (st/s)	moy.	1		
Intra/interF0 (st)		2		
Fréquence fondamentale F_0		**7**		
Total *loudness* (Bark, dB)	moy./std	2	2	2
Énergie		**2**	**2**	**2**
Roll Off 5, 25, 50, 75, 95 % (st)	moy./std	10	10	10
Pente totale, [0-500], [500-1 500] (st/Hz)	moy./std	6	6	6
Barycentre	moy/std	2	2	2
Spectre		**18**	**18**	**18**
Bandes de Bark 0-21 (dB)	moy./std	42	42	42
Bandes harmoniques 1-5 (dB)	moy./std	10		
Énergie par bandes spectrales		**52**	**42**	**42**
MFCC 0-12	moy/std	26	26	26
Cepstre		**26**	**26**	**26**
F2-F1 (st)	4 fonc.	4		
F3-F2 (st)	4 fonc.	4		
Formants		**8**		
ZCR		**0**	**0**	**0**

Tableau 3. *Ensemble 3 des descripteurs paraverbaux (E3-293-LIMSI). Quatre fonc. : min., max., moyenne, std*

3.8. *Robustesse des paramètres*

Le but de ce travail est de sélectionner une ou des familles de descripteurs parmi les plus robustes dans les ensembles étudiés. La répartition des descripteurs en famille pour les trois ensembles est résumée dans le tableau 4. Nous définissons la robustesse d'une famille de paramètres par le fait qu'elle soit sélectionnée avec un haut rang d'explication pour plusieurs corpus. Pour cette étude sur le pouvoir d'explication des

Ensemble/familles **#total paramètres**	E1 + E2 + E3 1 011	E1(ISO9) 384	E2 334	E3 293
EBBark	376	0	240	136
Cepstre	366	288	0	78
Spectre	114	0	60	54
F_0	70	48	15	7
ZCR	34	24	10	0
Énergie	30	24	0	6
QV	13	0	9	4
Formants	8	0	0	8

Tableau 4. *Répartition des paramètres dans les huit familles (EBBark : énergie par bandes spectrales, QV : qualité vocale)*

paramètres, nous utilisons l'algorithme GainRatioAttributeEval disponible sous Weka (Hall *et al.*, 2009) qui évalue le gain de chaque descripteur par rapport à la tâche pour l'ensemble regroupant tous les descripteurs (E1 + E2 + E3), soit 1 011 descripteurs. Pour mesurer le rang d'explication d'une famille, nous utilisons trois indices : le pourcentage de paramètres explicatifs retenus (dont le score est supérieur à 0), la valeur moyenne du score, le nombre de paramètres retenus.

Il existe beaucoup de redondances entre les paramètres au sein de ces ensembles mais aussi de spécificités car les paramètres ne sont pas toujours calculés de la même manière. L'objectif de ce travail n'est pas un travail exhaustif sur la comparaison de tous les descripteurs. Notre but est de façon pragmatique de sélectionner une ou quelques familles de paramètres parmi les plus robustes et qui auront une performance au niveau de l'ensemble de référence (Challenge Interspeech 2009) et pourront être calculées en peu de temps. Nous avons privilégié un nombre réduit de familles avec beaucoup de paramètres, à un nombre réduit de paramètres issus de beaucoup de familles différentes pour ces raisons de gain en temps de calcul.

4. Expériences

Nous avons mené des expériences sur trois corpus collectés en situation réaliste dans le cadre de plusieurs projets nationaux. Plus de cent cinquante voix ont été enregistrées avec de grandes variabilités acoustiques dues notamment à l'âge. L'âge des personnes enregistrées varie entre 16 ans (dans JEMO) et plus de 90 ans (dans ARMEN). À l'heure actuelle, il existe une grande diversité de classifieurs disponibles : modèles de mélanges de gaussiennes, réseaux de neurones, arbres de décision ou encore machines à vecteurs supports (SVM) (Asa et Weston, 2010). Dans l'étude présentée, les auteurs ont choisi de réaliser leurs expériences de reconnaissance avec des SVM standard avec un noyau polynomial. Les expériences de reconnaissance ont

toutes été menées grâce à la plate-forme Weka (Hall *et al.*, 2009) avec l'algorithme SVM (SMO) en utilisant le protocole de *cross-validation* sur dix sous-ensembles ou en *cross-corpus*. La configuration de l'algorithme est la configuration standard, les données sont normalisées entre –1 et 1.

4.1. *Corpus*

Les trois corpus utilisés ont été segmentés et annotés en double annotation suivant le même protocole et schéma d'annotation (Devillers *et al.*, 2010). L'annotation émotionnelle réalisée porte uniquement sur les informations transmises sur le canal paralinguistique. Le κ (kappa) de Cohen (ou alpha de Cronbach sur des données continues) est généralement utilisé pour mesurer le degré de concordance entre les annotations attribuées par deux juges. Seules les données consensuelles sont ensuite utilisées pour entraîner les modèles d'apprentissage. La qualité et la fiabilité des annotations sont fondamentales : elles sont vérifiées grâce à des tests de concordance entre annotateurs mais également grâce à des tests perceptifs menés auprès d'autres observateurs généralement sur une partie du corpus tirée au hasard. Cette partie du travail sur l'annotation est souvent moins valorisée, alors qu'elle est pourtant essentielle. Il est fréquent de voir publiés des scores de kappa assez faibles ($< 0{,}5$) pour l'annotation d'émotions sur des données réalistes (non jouées par des acteurs). De faibles degrés de concordance entre les annotateurs sur des données réalistes montrent alors la difficulté de la tâche d'annotation et la diversité des émotions présentes. Dans la vie de tous les jours, les émotions sont rarement discrètes et primaires, elles sont souvent mélangées, subtiles et suivent des processus dynamiques. Plus de 40 % des segments des trois corpus collectés n'ont pas été pris en compte dans cette étude : nous n'avons gardé que les données consensuelles et les classes ont été constituées avec un nombre d'échantillons comparables. Le tableau 7 résume les informations relatives aux trois corpus utilisés.

Toutes les classes émotionnelles utilisées lors de l'annotation des trois corpus sont regroupées en quatre macroclasses : colère, joie, tristesse et un état neutre non expressif pour des raisons de taille de corpus. Dans la classe colère, on trouve aussi bien de l'énervement, de l'impatience, de l'agacement, de la colère froide que de la colère chaude. La classe joie regroupe des annotations positives d'amusement, de soulagement, de satisfaction et de joie franche avec des rires. La tristesse est une classe qui regroupe également différents types de tristesse exprimée par des voix lentes, hésitantes avec peu d'énergie, voire dépressive.

4.1.1. *JEMO*

Le corpus JEMO a été enregistré en laboratoire pour obtenir des émotions réalistes en contexte de jeu dans le cadre du projet ANR Affective Avatar. Le jeu consistait à faire reconnaître à la machine une émotion (colère, joie, tristesse, peur ou un état neutre) sans qu'aucun contexte ne soit indiqué (Brendel *et al.*, 2010). Les émotions collectées sont alors prototypiques, c'est-à-dire que le taux d'accord inter-annotateur

est haut sur ce corpus malgré un support lexical totalement libre. Ce corpus de 29 minutes a été enregistré en décembre 2010 au LIMSI. Cinquante-neuf locuteurs entre 16 et 48 ans ont participé à l'enregistrement (trente hommes et vingt-neuf femmes).

4.1.2. *ARMEN*

Pour disposer de données émotionnelles spontanées au plus proche de la réalité, un système de collecte simulant une interaction naturelle et mettant en œuvre un agent virtuel expressif embarqué sur un robot a été développé. Il a été mis en œuvre pour recueillir deux corpus émotionnels, avec la participation de près de quatre-vingts patients de centres médicaux de la région de Montpellier en 2010 et 2011, dans le cadre du projet ANR Tecsan ARMEN (Chastagnol *et al.*, 2013). Soixante-dix-sept personnes, quarante-huit hommes et vingt-neuf femmes, entre 18 ans et plus de 90 ans ont participé à cette collecte. La moitié des données ont été collectées auprès de personnes de plus de 60 ans. Ces données ont été utilisées dans l'exploration d'approches pour la résolution du problème de la généralisation des performances des systèmes de détection des émotions à d'autres données (Chastagnol, 2013).

4.1.3. *IDV-HR*

Le corpus IDV-HR (Tahon *et al.*, 2011) a été enregistré dans l'appartement témoin de l'Institut de la Vision (Paris) dans le cadre du projet FUI ROMEO. Vingt-deux locuteurs (onze hommes et onze femmes de 28 à 80 ans) ont participé à cet enregistrement. Lors d'une séance d'enregistrement le participant est assis face au robot NAO. Un Magicien d'Oz commande le robot depuis une salle cachée et sélectionne la réponse la plus adaptée parmi plusieurs choix possibles. Chaque réponse est encodée linguistiquement suivant le comportement du robot. Le robot propose au participant une série de scénarios proches du réveil du matin, au cours desquels le participant simule différents états de santé (en forme, en mauvaise santé, en situation d'urgence, de déprime, de joie pour un événement à venir). Chaque série est jouée plusieurs fois, NAO ayant des comportements à chaque fois différents (directif, dubitatif, encourageant, aimable, neutre ou encore empathique), auxquels le sujet réagit spontanément. Deux extraits sont transcrits dans les tableaux 5 et 6, présentant les annotations réalisées par deux personnes sur les émotions perçues dans la voix (et non dans le lexique). Comme on peut le constater dans le tableau 5, les annotations sont parfois non consensuelles. Nous avons indiqué les deux étiquettes annotées sur chaque tour de parole.

4.2. *Protocole expérimental*

Nous avons effectué la comparaison des trois ensembles de paramètres sur les trois corpus, puis combiné les ensembles de descripteurs deux à deux, puis regroupé les trois ensembles. Les performances des combinaisons d'ensembles de paramètres montrent les ensembles les plus complémentaires. L'ensemble global des descripteurs

Locuteur	Transcription	Annotations
Nao	Comment vas-tu ce matin ?	
Sujet	*Ben écoute ce matin, ça va.*	Peur, Neutre
	Parce que j'avais un projet, puis euh j'ai un rendez-vous, tout ça, avec une équipe sympa. Ça va. Ça va. C'est bien. Je suis content.	Tristesse, Neutre
	Je vais me lever.	Joie, Joie
Nao	Ça me fait très plaisir que tu sois content. Qu'est-ce qui te rend si content ?	
Sujet	*Ben je... écoute, t'écoutes pas mon pauvre, je t'ai dit là...*	Tristesse, Neutre
	J'ai une raison de me lever, un rendez-vous avec une équipe formidable, et voilà.	Joie, Joie

Tableau 5. *Extrait d'IDV-HR : Nao (comportement empathique) discute avec le sujet (un homme de 66 ans) pendant le scénario « joie pour un événement à venir ».*

Locuteur	Transcription	Annotations
Nao	C'est catastrophique, tu as un problème. Mon pauvre, est-ce que tu as très mal ?	
Sujet	*Ben euh pour l'instant oui, et puis surtout je me sens con, de m'être fait mal.*	Colère, Colère
Nao	C'est terrible pour toi. Où est-ce que tu as mal précisément ?	
Sujet	*Ben à l'arcade sourcilière tu vois pas là non ?*	Colère, Colère
Nao	Tu dois vraiment souffrir. Qu'est-ce je peux faire pour t'aider ?	
Sujet	*Ben écoute appelle-moi une infirmière ou un médecin, je sais pas quoi, mais c'est à toi de savoir hein.*	Colère, Colère

Tableau 6. *Extrait d'IDV-HR : Nao (comportement empathique) discute avec le sujet (un homme de 66 ans) pendant le scénario « urgence médicale ».*

a été ensuite utilisé sur les trois corpus pour sélectionner la ou les familles les plus robustes. Des tests en *cross-corpus* ont également été menés. Les corpus JEMO et ARMEN étant utilisés en apprentissage et le corpus IDV-HR en test.

4.2.1. *Comparaison des trois ensembles de paramètres sur les trois corpus*

Les résultats obtenus avec la mesure d'Accuracy sont très similaires à ceux obtenus avec la F-mesure. Les résultats en *cross-corpus* (tableau 9) sont moins bons mais

Corpus	#locuteurs (H/F) âge	Durée (s) min/max/moy	#seg.	Col.	Joi.	Tri.	Neu.
JEMO	59 (30H/29F) de 16 à 48	41 min 0,10/9,82/1,96	1 249	291	310	307	341
ARMEN	77 (48H/29F) de 18 à 90 et plus	71 min 0,25/7,04/2,35	1 807	403	506	318	580
IDV-HR	22 (11H/11F) de 28 à 80	82 min 0,24/5,73/2,37	2 063	512	508	495	551

Tableau 7. *Répartition des segments consensuels dans les trois corpus. Durée totale, durée minimale, maximale et moyenne des segments.*

Corpus JEMO (59 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384-ISO9	60,8	60,6	66,7	54,4	67,0	54,1
E2-334-LIMSI	62,4	62,1	71,6	53,5	68,5	54,5
E3-293-LIMSI	60,5	60,1	70,3	48,8	56,3	65,1
Corpus ARMEN (77 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384-ISO9	49,1	48,9	49,3	48,7	54,4	38,9
E2-334-LIMSI	50,9	50,4	56,0	49,2	54,6	37,9
E3-293-LIMSI	52,6	52,2	55,6	51,3	57,4	39,9
Corpus IDV-HR (22 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384-ISO9	41,8	41,8	39,4	50,5	40,9	36,3
E2-334-LIMSI	41,3	41,4	38,3	48,9	41,6	36,4
E3-293-LIMSI	40,6	40,6	36,6	36,7	46,0	42,9

Tableau 8. *Détection des émotions en* cross-validation *(en dix tests) sur les trois ensembles de descripteurs*

restent au-dessus du hasard. On peut remarquer des différences importantes suivant les émotions.

Les résultats (tableau 8) obtenus en validation croisée avec dix tests sont du même ordre pour les trois ensembles de paramètres (E1, E2 et E3) qui contiennent de 293 à 384 descripteurs pourtant différents (cf tableau 4) quel que soit le corpus utilisé, avec des différences de quelques points dans la répartition par émotion. La meilleure performance obtenue (par F-mesure) est d'environ 60 % pour la détection de quatre classes obtenue par *cross-validation* sur dix tests pour le corpus JEMO qui est le plus « prototypique ». Les deux autres corpus ARMEN et IDV-HR contiennent tous les deux des voix de sujets âgés enregistrés en interaction avec un robot dans leur cadre de vie quotidienne (EPHAD, IDV, centre de rééducation Propara) et obtiennent des scores de détection moins élevés respectivement d'environ 50 % pour ARMEN et

Corpus JEMO (59 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384-ISO9	31,0	30,7	26,5	36,9	29,1	30,4
E2-334-LIMSI	33,8	32,9	41,5	26,0	40,9	22,0
E3-293-LIMSI	31,8	31,6	29,0	24,8	36,1	35,8
Corpus ARMEN (77 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384-ISO9	33,5	32,2	33,4	41,8	31,8	21,2
E2-334-LIMSI	35,6	34,2	38,9	41,2	37,4	18,2
E3-293-LIMSI	36,0	34,0	37,5	42,8	38,1	16,2

Tableau 9. *Détection des émotions en* cross-corpus *(test sur IDV-HR) sur les trois ensembles de descripteurs*

40 % pour IDV-HR. Les émotions sont exprimées de façons plus subtiles et mélangées. La tristesse est l'émotion la moins facilement reconnue avec les paramètres utilisés.

4.2.2. *Combinaison des ensembles de descripteurs acoustiques*

Afin d'étudier la complémentarité des ensembles de descripteurs, nous les avons groupés deux à deux créant trois nouveaux ensembles de 627 à 718 descripteurs, et également un ensemble contenant tous les 1 011 descripteurs (tableau 10). Nous avons effectué des tests en *cross-validation* et en *cross-corpus*. Le corpus IDV-HR a été utilisé pour le test en *cross-corpus*.

Les performances obtenues pour les trois corpus sont très proches des performances de l'ensemble qui sert de référence, montrant clairement une forte redondance entre les ensembles de descripteurs. L'association des trois ensembles ne permet pas d'obtenir de meilleurs résultats sur aucun corpus. La performance la plus élevée est obtenue par la combinaison des deux ensembles E2 et E3, développés au LIMSI à partir des librairies Aubio, Praat et Yaafe, pour les trois corpus : 64,8 % au lieu de 61,1 % avec E1-ISO9 (JEMO), 52,7 % au lieu de 49,1 % pour E1-ISO9 (ARMEN) et 43,4 % au lieu de 41,8 % pour E1-ISO9 (IDV-HR). Cependant, très peu d'améliorations sont constatées en ajoutant les différents ensembles et il est difficile de savoir quels sont les paramètres les plus robustes.

Nous obtenons de meilleurs résultats en *cross-corpus* si l'apprentissage est fait sur ARMEN avec E2+E3. Plus particulièrement, nous améliorons la F-mesure sur la détection de la classe Tristesse, ce qui n'est pas le cas avec l'ensemble E1.

4.2.3. *Sélection des familles de paramètres*

La sélection des paramètres est effectuée sur la combinaison des trois ensembles de descripteurs (E1+E2+E3). Afin de faire une analyse des paramètres les plus explicatifs et les plus robustes à travers les trois corpus, il est possible d'utiliser plusieurs algorithmes sous Weka comme par exemple GainRatioAttributeEval. Nous avons étudié le score (gainRatio) obtenu par les descripteurs d'une même famille, puis pondéré

Corpus JEMO (59 loc.)	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	60,6	66,7	54,4	67,0	54,1
E1+E2-718(384+334)	60,6	69,6	55,7	66	53,4
E1+E3-677(384+293)	62	70,3	54,2	66,7	56,7
E2+E3-627(334+293)	64,8	77	59,1	68,7	54,8
E1+E2+E3-1011(384+334+293)	62,4	79,3	54,2	66,7	56,7
Corpus ARMEN (77 loc.)	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	49,1	50,6	49,7	54,6	35,9
E1+E2-718(384+334)	48,9	49,3	48,7	54,4	38,9
E1+E3-677(384+293)	50,5	54,3	48,9	54,3	40,3
E2+E3-627(334+293)	52,7	58,3	53	56,5	38,3
E1+E2+E3-1011(384+334+293)	51,8	57,6	52,4	54	39,7
Corpus IDV-HR (22 locuteurs)	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	41,8	39,4	50,5	40,9	36,3
E1+E2-718(384+334)	43,3	43	50,5	42,1	39,2
E1+E3-677(384+293)	42,6	40,3	51,5	40,4	38,1
E2+E3-627(334+293)	43,4	39,6	50	45	38,7
E1+E2+E3-1011(384+334+293)	42,8	40,1	50,5	42,9	37,6

Tableau 10. *Détection des émotions en combinant deux ensembles et trois ensembles en* cross-validation *(sur dix tests)*

Corpus JEMO (59 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	31,0	30,7	26,5	36,9	29,1	30,4
E1+E2-718(384+334)	32,0	31,3	33,5	38,7	28,0	24,9
E1+E3-677(384+293)	31,7	31,6	36,4	31,7	34,5	23,2
E2+E3-627(334+293)	31,5	35,2	33,5	33,2	23,3	31,4
E1+E2+E3-1011(384+334+293)	32,1	31,8	34,0	34,7	34,7	23,2
Corpus ARMEN (77 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	33,5	32,2	33,4	41,8	31,8	21,2
E1+E2-718(384+334)	34,1	32,2	32,9	43,4	32,5	19,5
E1+E3-677(384+293)	35,6	33,8	37,0	43,3	36,6	17,3
E2+E3-627(334+293)	37,3	38,4	44,1	38,4	22,6	36,1
E1+E2+E3-1011(384+334+293)	35,3	33,4	37,7	44,1	32,5	18,6

Tableau 11. *Détection des émotions en combinant deux ensembles et trois ensembles en* cross-corpus *(test sur IDV-HR)*

ce score par le nombre de descripteurs et enfin privilégié dans nos choix pour la suite des expériences la famille qui a potentiellement le plus de paramètres sélectionnés pour les trois expériences parallèles avec les différents corpus. La répartition des para-

mètres explicatifs en pourcentages dans les huit familles de paramètres pour les trois corpus JEMO, ARMEN et IDV-HR est résumée dans le tableau 12.

Familles/Corpus	JEMO	ARMEN	IDV-HR
%PE (1011)	70,4 %	75,3 %	27 %
EBBark(376)	82,5 %(307-0,082)	79,7 %(300-0,052)	27,4 %(102-0,029)
Cepstre(366)	57,1 %(209-0,057)	69 %(253-0,039)	21 %(78-0,021)
Spectre(114)	78 %(68,4-0,052)	71 %(80-0,049)	17 %(19-0,023)
F_0**(70)**	85,7 %(60-0,059)	90 %(63-0,004)	52 %(37-0,028)
ZCR(34)	64,7 %(22-0,055)	76,5 %(26-0,037)	52 %(18-0,024)
Énergie(30)	73,6 %(22-0,133)	80 %(24-0,057)	40 %(12-0,038)
QV(13)	46 %(6-0,06)	100 %(13-0,038)	61,5 %(8-0,026)
Formants(8)	100 %(8-0,009)	25 %(2-0,026)	0 %(0-0,000)

Tableau 12. *Répartition des paramètres explicatifs (PE en pourcentages) dans les huit familles, le nombre de paramètres et la moyenne du score sur les descripteurs de la famille sont donnés entre parenthèses (EBBark : énergie par bandes spectrales, QV : qualité vocale)*

Les paramètres d'énergie (énergie globale et énergie par bandes spectrales) sont dans cette expérience les plus explicatifs et les plus robustes par rapport aux différentes expériences sur les trois corpus. Une famille avec un grand nombre de descripteurs explicatifs est aussi un critère sélectif. La famille sélectionnée allie un bon score moyen (GainRatio moyen) et un grand nombre de paramètres sélectionnés pour les trois corpus. Il s'agit de l'énergie par bandes spectrales (376 paramètres) calculée à partir du *loudness* de l'ensemble E2 qui contient 240 paramètres et des énergies par bandes de Bark de l'ensemble E3 qui contient 136 paramètres calculés sur les parties voisées, non voisées, et sur tout le signal.

Nous n'avons pas mené une analyse plus poussée de l'ensemble des familles de paramètres pour les classer. Nous pouvons cependant faire quelques remarques sur les scores obtenus dans le tableau 12. On peut s'étonner du peu de paramètres sélectionnés pour le corpus IDV-HR par rapport à ARMEN et JEMO, les émotions dans ce corpus sont plus mélangées et moins expressives que dans les autres corpus. Deux familles ont une taille de même ordre : l'énergie par bandes spectrales et le cepstre et obtiennent des résultats très différents. Le score moyen et le nombre de paramètres sélectionnés pour le cepstre pour les trois corpus sont plus faibles. Les scores sont également plus faibles pour les paramètres calculés sur la fréquence fondamentale ou sur l'enveloppe spectrale. D'autres familles ont également donné des scores intéressants mais sont instables d'un corpus à l'autre (par exemple les formants, la qualité vocale), ces familles sont de plus décrites avec peu de paramètres.

La famille sélectionnée (l'énergie par bandes spectrales, 376 paramètres) a ensuite été utilisée seule pour la détection des quatre émotions et a permis d'obtenir des performances au niveau des résultats de référence présentés dans le tableau 13. Nous avons pu constater également expérimentalement que les bandes de Bark (EBBark) les plus importantes sont calculées sur tout le signal et non spécifiquement sur les parties voisées ou non voisées, ce qui permet de réduire encore l'ensemble des paramètres à 282 (240 + 42) paramètres calculés. Il est probable qu'il y ait également des redondances dans les paramètres *loudness*, nous n'avons pas cherché à réduire cet ensemble de paramètres car il est peu coûteux de calculer les fonctions statistiques sur une série temporelle.

Nous avons effectué des tests en *cross-validation* et en *cross-corpus*. Le corpus IDV-HR a été utilisé pour le test en *cross-corpus*.

Corpus JEMO (59 loc.)	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	60,6	66,7	54,4	67,0	54,1
E2+E3-627(334+293)	64,8	77	59,1	68,7	54,8
E2+E3-376(L.+EBBark136)	62,7	75,2	58,5	68,2	48,8
E2+E3-282(L.+EBBark42)	61,1	73,3	55,4	65,8	50
E2-240(Loudness)	56,2	65,4	49,8	63,1	43,3
Corpus ARMEN (77 loc.)	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	48,9	49,3	48,7	54,4	38,9
E2+E3-627(334+293)	52,7	56,5	53	58,3	38,3
E2+E3-376(L.+EBBark136)	53,7	57,7	52,7	58,7	41,3
E2+E3-282(L.+EBBark42)	53,3	59	51,1	57,9	41,1
E2-240(Loudness)	50,1	54,2	48	57,3	35,2
Corpus IDV-HR (22 loc.)	F-mes.	Col.	Joi.	Neu.	Tri.e
E1-384(IS09-référence)	41,8	39,4	50,5	40,9	36,3
E2+E3-627(334+293)	43,4	39,6	50	45	38,7
E2+E3-376(L.+EBBark136)	41,1	41,7	45,7	43,9	32,5
E2+E3-282(L.+EBBark42)	40,9	42,4	44,5	43,6	32,3
E2-240(Loudness)	39,1	38	42,7	42,4	32,7

Tableau 13. *Détection des émotions avec la famille de paramètres sur l'énergie par bandes spectrales (regroupant les* loudness *(L.) spécifiques sur les Barks et l'énergie par bandes de Bark (EBBark) en* cross-validation *(dix tests)*

Les résultats sont meilleurs que ceux obtenus avec la librairie openSMILE en *cross-corpus*. Il semble intéressant de ne garder que l'énergie par bandes spectrales (lignes 4 et 5 pour chaque corpus). En revanche, retirer le set E3-Bark pénalise la reconnaissance (notamment au niveau de la reconnaissance de la tristesse).

Corpus JEMO (59 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	31,0	30,7	26,5	36,9	29,1	30,4
E2+E3-627(334+293)	31,5	35,2	33,5	33,2	23,3	31,4
E2+E3-376(L.+EBBark136)	31,7	31,6	29,7	37,6	33,0	25,6
E2+E3-282(L.+EBBark42)	31,4	31,6	31,1	35,7	33,9	25,2
E2-240(Loudness)	30,1	28,8	26,3	39,3	23,4	26,4
Corpus ARMEN (77 loc.)	Ac.	F-mes.	Col.	Joi.	Neu.	Tri.
E1-384(IS09-référence)	33,5	32,2	33,4	41,8	31,8	21,2
E2+E3-627(334+293)	37,3	38,4	44,1	38,4	22,6	36,1
E2+E3-376(L.+EBBark136)	37,4	36,3	37,6	44,6	37,4	25,2
E2+E3-282(L.+EBBark42)	37,7	36,4	39,7	44,7	37,4	23,1
E2-240(Loudness)	35,5	33,1	34,0	43,4	38,3	15,3

Tableau 14. *Détection des émotions avec la famille de paramètres sur l'énergie par bandes spectrales (regroupant les* loudness *(L.) spécifiques sur les Barks et l'énergie par bande de Bark (EBBark) en* cross-corpus *(test sur IDV-HR)*

4.3. *Analyse des résultats sur la famille sélectionnée*

La famille sur l'énergie par bandes spectrales Bark permet d'obtenir des performances comparables à celles obtenues sur les ensembles testés, notamment sur l'ensemble de référence IS09 mélangeant des paramètres cepstraux, spectraux, des paramètres de la fréquence fondamentale et de qualité vocale. Cette famille de descripteurs psycho-acoustiques sur l'énergie par bandes spectrales révèle des performances intéressantes. Elle regroupe deux façons de les calculer : les coefficients de *loudness* et l'énergie par bandes de Bark (21). Les coefficients de *loudness*, utilisés seuls, ont une performance en dessous de l'ensemble de référence. Cependant, si ces coefficients sont couplés à la moyenne et à la déviation standard de l'énergie par bandes de Bark (21) sur tout le signal, les performances (F-mesure) de détection des quatre classes d'émotions sur les parties voisées et non voisées sont améliorées de 56,2 % à 61,1 % pour JEMO, de 50,1 % à 53,7 % pour ARMEN et de 39,1 % à 41,1 % pour IDV-HR.

L'énergie $E(k)$ de la bande de Bark k est calculée comme étant la somme des amplitudes au carré pour chaque échantillon fréquentiel appartenant à la bande k. Ce descripteur est donc un descripteur perceptif de relativement bas niveau. Le *loudness* spécifique $N'(k)$ de la bande k est défini comme étant l'énergie dans la bande k à la puissance 0,23 (Peeters, 2004). Ce descripteur est donc de plus haut niveau que le précédent. Même si ces descripteurs sont très proches, ils ne correspondent pas tout à fait à la même réalité et semblent être complémentaires. Le descripteur de *loudness* dans une bande spécifique est plus proche de ce que peut percevoir l'oreille humaine, que la simple énergie dans cette même bande. Les descripteurs de *loudness* par fréquence, utilisés comme uniques paramètres (240 paramètres) montrent des performances inférieures de quelques pourcents par rapport à la référence.

5. Conclusion

Dans un contexte d'interaction homme-machine et plus particulièrement homme-robot, les variabilités liées à l'environnement acoustique, au locuteur, et au type de tâche, peuvent fortement influencer les systèmes de reconnaissance automatique des émotions mais également les systèmes de transcription automatique. Ces systèmes sont principalement fondés sur l'extraction de descripteurs acoustiques à partir du signal enregistré par la machine, mais également sur le choix des corpus d'apprentissage. La recherche des descripteurs les plus robustes est donc fondamentale dans ce domaine. De plus, pour pouvoir être embarqués, les systèmes ont besoin de temps de calcul réduit, et donc d'un minimum de paramètres et surtout de familles de paramètres à extraire du signal. Dans ce travail, nous avons cherché à déterminer les familles acoustiques les plus robustes pour la détection des émotions suivant trois tâches différentes comportant plus de cent cinquante locuteurs entre 16 et plus de 90 ans.

Pour ce faire, plusieurs ensembles de descripteurs ont été comparés en *cross-validation* et en *cross-corpus*. Une famille acoustique semble remplir ce critère de robustesse, formée de l'énergie par bandes de Bark, ainsi que du *loudness* spécifique défini également sur l'échelle de Bark. Bien que ces deux types de paramètres soient relativement proches, ils semblent être complémentaires, en cela qu'ils reconnaissent mieux une classe d'émotions ou une autre. L'utilisation de cette seule famille sur deux sets de descripteurs combinés permet d'obtenir les mêmes performances que l'ensemble de référence du Challenge Interspeech 2009 sur quatre classes émotionnelles.

Certes l'énergie a depuis longtemps été considérée comme un paramètre important pour la détection des émotions mais les résultats obtenus sur l'énergie par bandes spectrales ouvrent un grand nombre de perspectives encourageantes : il existe des paramètres suffisamment robustes et surtout indépendants de chaque tâche. Ces paramètres renforcent l'importance des modèles perceptifs pour la reconnaissance des émotions. Étant donné que ces paramètres sont extraits du spectre, nous pouvons en déduire que le timbre, dans toute son étendue spectrale, reste plus robuste que des descripteurs d'enveloppe spectrale, que ce soit dans les tests en *cross-validation* ou en *cross-corpus*.

L'étude de la synchronisation entre les canaux verbaux et non verbaux pour la communication des émotions dans ces corpus est également un de nos objectifs à court terme.

Remerciements

Nous remercions tous les chercheurs, étudiants et partenaires des projets ANR Tecsan ARMEN, ANR Affective Avatar, FUI ROMEO et tout particulièrement Mariette Soury et Clément Chastagnol qui ont participé aux collectes de données

6. Bibliographie

Asa B.-H., Weston J., *A user's guide to support vector machines (ch. 13)*, Humana Press, p. 223-239, 2010.

Bänziger T., Mortillaro M., Scherer K., « Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception », *Emotion*, vol. 12(5), p. 1161-1179, 2012.

Batliner A., Steidl S., , Schuller B., Seppi D., Laskowski K., Vogt T., Devillers L., Vidrascu L., Amir N., loic Kessous, Aharonson V., « CEICES : Combining Efforts for Improving automatic Classification of Emotional user States : a «forced co-operation» initiative », *Language and Technologies Conference*, Slovenia, p. 240-245, 2006.

Batliner A., Steidl S., Schuller B., Seppi D., Vogt T., Wagner J., Devillers L., Vidrascu L., Aharonson V., Kessous L., Amir N., « Whodunnit Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech », *Computer Speech and Language (CSL), Special Issue on Affective Speech in real-life interactions*, vol. 25, Issue 1, p. 4-28, 2011.

Bazillon T., Jousse V., Béchet F., Estève Y., Linarès G., Luzzati D., « La parole spontanée : transcription et traitement », *Traitement Automatique des Langues 49(3)*, 2008.

Boersma P., « Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound », *Institute of Phonetics Sciences, University of Amsterdam*, vol. 17, p. 97-110, 1993.

Brendel M., Zaccarelli R., Devillers L., « Building a system for emotions detection from speech to control an affective avatar », *LREC*, Valetta, Malta, 2010.

Buendia A., Devillers L., « From informative cooperative dialogues to long-term social relation with a robot. », *in* J. Mariani, S. Rosset, M. Garnier-Rizet, L. Devillers (eds), *Natural Interaction with Robots, Knowbots and Smartphones*, Springer New York, p. 135-151, 2014.

Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., WeissI B., « A database of german emotionnal speech », *Interspeech*, Lisbon, Portugal, p. 1517-1520, 2005.

Campbell N., « On the Use of NonVerbal Speech Sounds in Human Communication », *Verbal and Nonverbal Communication Behaviours*, vol. 4775, Springer Berlin Heidelberg, p. pp 117-128, 2007.

Chastagnol C., Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes., PhD thesis, Université Paris-Sud, 2013.

Chastagnol C., Clavel C., Courgeon M., Devillers L., « Designing an emotion detection system for a socially-intelligent human-robot interaction », *Towards a Natural Interaction with Robots, Knowbots and Smartphones, Putting Spoken Dialog Systems into Practice (Springer)*, 2013.

Chastagnol C., Devillers L., « Personality traits detection using a parallelized modified SFFS algorithm », *Interspeech*, Portland, Oregon, USA, 2012.

Clavel C., Devillers L., Richard G., Vidrascu I., Ehrette T., « Abnormal situations detection and analysis through fear-type acoustic manifestations », *ICASSP*, vol. IV, Honolulu, HI, U.S.A., p. 21-24, 2007.

Cowie R., Douglas-Cowie E., Cox C., « Beyond emotion archetypes : databases for emotion modelling using neural networks », *Neural Networks*, vol. 18, p. 371-388, 2005.

Delaborde A., Devillers L., « Use of Nonverbal Speech Cues in Social Interaction between Human and Robot : Emotional and Interactional markers », *International Workshop on Affective Interaction in Natural Environements (AFFINE)*, Firenze, Italy, 2010.

Delaborde A., Devillers L., « Impact of the Social Behaviours of the Robot on the User's Emotions : Importance of the Task and the Subject's Age », *Workshop on Affect, Compagnons Artificiels, Interaction*, Grenoble, France, 2012.

Delaborde A., Tahon M., Barras C., Devillers L., « A Wizard-of-Oz game for collecting emotional audio data in a children-robot interaction », *AFFINE'09*, Boston, MA, U.S.A., 2009.

Delaborde A., Tahon M., Barras C., Devillers L., « Affective Links in a Child-Robot Interaction », *LREC*, Valetta, Malta, 2010.

Devillers L., Cowie R., Martin J.-C., Douglas-Cowie E., Abrilian S., McRorie M., « Real-life emotions in French and English TV video clips : an integrated annotation protocol combining contiuous and discrete approaches », *LREC*, Genoa, Italy, 2006.

Devillers L., Martin J.-C., « Coding Emotional Events in Audiovisual Corpora », *LREC*, Marrakech, Morocco, 2008.

Devillers L., Vidrascu L., « Positive and negative emotional states behind laughs in spontaneous spoken dialogs », *Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, Germany, p. 37-40, 2007.

Devillers L., Vidrascu L., Lamel L., « Challenges in real-life emotion annotation and machine leraning based detection », *Journal of Neural Networks, Special Issue on Emotion and Brain*, vol. 18 (4), p. 407-422, 2005.

Devillers L., Vidrascu L., Layachi O., *Automatic detection of emotion from vocal expression*, Oxford University Press., chapter A blueprint for an affectively competent agent, Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing., 2010.

Dumouchel P., Dehak N., Attabi Y., Dehak R., Boufaden N., « Cepstral and Long-Term Features for Emotion Recognition », *Interspeech*, Brighton, U.K., p. 344-347, 2009.

Ekman P., *Handbook of cognition and emotion*, Wiley, U.K., chapter Basic emotion, 1999.

Engberg I. S., Hansen A. V., Andersen O., Dalsgaard P., « Design, recording and verification of a danish emotional speech database », *Eurospeech*, Rhodes, Greece, p. 1695-1698, 1997.

Eyben F., Batliner A., Schuller B., Seppi D., Steidl S., « Cross-corpus classification of realistic emotions : some pilot experiments », *LREC, Workshop on EMOTION : Corpora for Research on Emotion and Affect*, ELRA, Valetta, Malta, p. 77-82, 2010.

Fernandez R., Picard R. W., « Modeling drivers' speech under stress », *Speech Communication*, vol. 40, p. 145-159, 2003.

Gendrot C., « Rôle de la qualité de la voix dans la simulation des émotions : une étude perceptive et physiologique », 2004.

Grimm M., Kroschel K., Narayanan S., « The Vera am Mittag German audio-visual emotional speech database », *in* IEEE (ed.), *International Conference on Mulitmedia and Expo (ICME)*, Hannover, Germany, p. 865-868, 2008.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA Data Mining Software : An Update », *SIGKDD Explorations*, 2009.

Han J. G., Gilmartin E., Looze C. D., Vaughan B., Campbell N., « Speech & Multimodal Resources : The Herme Database of Spontaneous Multimodal Human-Robot Dialogues », *LREC*, Istanbul, Turkey, 2012.

Justin P., Laukka P., « Communication of emotions in vocal expression and music performance : different channels, same code ? », *Psychological Bulletin*, vol. 129 (5), p. 770-814, 2003.

Marchi E., Batliner A., Schuller B., « Speech, emotion, age, language, task and typicality : trying to disentangle performance and future relevance », *Workshop on Wide Spectrum Social Signal Processing (ASE/IEEE International Conference on Social Computing)*, Amsterdam, Netherlands, 2012.

McKeown G., Valstar M., Cowie R., Pantic M., Schröder M., « The SEMAINE database : annotated multimodal records of emotionally coloured conversations between a person and a limited agent », *IEEE Transactions on Affective Computing*, vol. 3, Issue 1, p. 5-17, 2012.

Moore B., B.Glasberg, Baer T., « A Model for the Prediction of Thresholds Loudness and Partial Loudness », *J. Audio Eng. Soc.*, vol. 45, p. 224-240, 1997.

Peeters G., « A large set of audio features for sound description (similarity and classification) in the CUIDADO project », *Ircam*, 2004.

Plantin C., Doury M., Traverso V., *Collection Ethologie et Psychologie*, Presses Universitaires de Lyon, chapter Les émotions dans les interactions, 2000.

Ruiz R., de Hugues P. P., Legros C., « Analysing cockpit and laboratory recordings to determine fatigue levels in pilots' vocies », *Journal of the Acoustical Society of America*, vol. Vol. 123, Issue 5, p. 3070-3070, 2008.

Scherer K. R., « Vocal affect expressions : A Review and a Model for Future Research », *Psychological Bulletin*, vol. 99 (2), p. 143-165, 1986.

Schuller B., Batliner A., Steidl S., Seppi D., « Recognising realistic emotions and affect in speech : state of the art and lessons learnt from the first challenge », *Speech Communication, Special Issue on "Sensing Emotion and Affect - Facing Realism in Speech Processing"*, vol. 53 (9/10), p. 1062-1087, 2011a.

Schuller B., Devillers L., « Incremental acoustic valence recognition : an inter-corpus perspective on features, matching and performance in a gating paradigm », *Interspeech*, Makuhari, Chiba, Japan, 26 - 30 sept, 2010.

Schuller B., Steidl S., Batliner A., « The INTERSPEECH 2009 Emotion Challenge », *Interspeech*, Brighton, U.K., 2009a.

Schuller B., Steidl S., Batliner A., Burkhardt F., Devillers L., Müller C., Narayanan S., « The INTERSPEECH 2010 Paralinguistic Challenge », *Interspeech*, Makuhari, Chiba, Japan, p. 2830-2833, 26 - 30 sept, 2010a.

Schuller B., Steidl S., Batliner A., Nöth E., Vinciarelli A., Burkhardt F., van Son R., Weninger F., Eyben F., Bocklet T., Mohammadi G., Weiss B., « The INTERSPEECH 2012 Speaker Trait Challenge », *Interspeech*, Portland, Oregon, USA, 2012.

Schuller B., Steidl S., Batliner A., Schiel F., Krajewski J., « The INTERSPEECH 2011 Speaker State Challenge », *Interspeech*, Firenze, Italy, 2011b.

Schuller B., Vlasenko B., Eyben F., Rigoll G., Wendemuth A., « Acoustic emotion recognition : a benchmark comparison of performances », *Automatic Speech Recognition and Understanding Workshop (ASRU/IEEE)*, Merano, p. 552-557, 2009b.

Schuller B., Vlasenko B., Eyben F., Wöllmer M., Stühlsatz A., Wendemuth A., Rigoll G., « Cross-corpus acoustic emotion recognition : variances and strategies », *Transaction on Affective Computing (IEEE)*, vol. 1, Issue 2, p. 119-131, 2010b.

Schuller B., Zaccarelli R., Rollet N., , Devillers L., « CINEMO - A French spoken language resource for complex emotions : facts and baselines », *LREC*, Valetta, Malta, 2010c.

Schuller B., Zhang Z., Weninger F., Rigoll G., « Using multiple databases for training emotion recognition : to unite or to vote ? », *Interspeech*, Florence, Italy, August, 2011c.

Sehili M. E. A., Reconnaissance des sons de l'environnement dans un contexte domotique, PhD thesis, Telecom SudParis, 2013.

Soury M., Devillers L., « Collecte de données pour la détection du stress dans les interactions sociales », *Workshop Affects, Compagnons Artificiels et Interactions (WACAI 2012)*, 2012.

Sun R., Moore E. I., « Using ROVER for Multiple Databases Training at the Decision Level for Binary Emotional Recognition », *ICASSP*, 2013.

t Hart J., « Differential sensitivity to pitch distance, particularly in speech », *Journal of the Acoustical Society of America*, vol. 69 (3), p. 811-821, 1981.

Tahon M., Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot, PhD thesis, Unversité Paris-Sud, 2012.

Tahon M., Degottex G., Devillers L., « Usual voice quality features for emotionnal valence detection », *Speech Prosody*, Shanghai, China, 2012a.

Tahon M., Delaborde A., Devillers L., « Real-life Emotion Detection from Speech in Human-Robot Interaction : Experiments across Diverse Corpora with Child and Adult Voices », *Interspeech*, Firenze, Italia, 2011.

Tahon M., Delaborde A., Devillers L., « Corpus of children voices for mid-level social markers and affect bursts analysis », *LREC*, Istanbul, Turckey, 2012b.

Tahon M., Devillers L., « Acoustic measures characterizing anger across corpora collected in artificial or natural context », *Speech Prosody*, Chicago, USA, 2010.

Ververidis D., Kotropoulos C., « A review of emotionnal speech databases », *Penhellenic Conf. on Informatics (PCI)*, n° 9th, Thessaloniki, Greece, p. 560-574, 2003.

Villavicencio F., Röbel A., Rodet X., « Applying improved spectral modeling for high quality voice conversion », *ICASSP*, Taipei, Taiwan, p. 4285-4288, 2009.

Xiao Z., Classification of emotions in audio signals, PhD thesis, Ecole Centrale de Lyon, 2008.

Zhang Z., Weninger F., Wöllmer M., Schuller B., « Unsupervised learning in cross-corpus acoustic emotion recognition », *ASRU*, Honolulu, Hawaï, December, 2011.