Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat

Michel Vacher, Benjamin Lecouteux, Frédéric Aman, François Portet, Solange Rossato

Acquisition et reconnaissance automatique d'expressions et d'appels vocaux dans un habitat.

Michel Vacher[1] Benjamin Lecouteux[2] Frédéric Aman[1]
François Portet[2] Solange Rossato[2]
(1) CNRS, LIG, F-38000 Grenoble, France
(2) Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
Michel.Vacher@imag.fr, Benjamin Lecouteux@imag.fr, Frederic.Aman@imag.fr,
francois.Portet@imag.fr, Solange.Rossato@imag.fr

RÉSUMÉ
Cet article présente un système capable de reconnaître les appels à l'aide de personnes âgées vivant à domicile afin de leur fournir une assistance. Le système utilise une technologie de Reconnaissance Automatique de la Parole (RAP) qui doit fonctionner en conditions de parole distante et avec de la parole expressive. Pour garantir l'intimité, le système s'exécute localement et ne reconnaît que des phrases prédéfinies. Le système a été évalué par 17 participants jouant des scénarios incluant des chutes dans un Living lab reproduisant un salon. Le taux d'erreur de détection obtenu, 29%, est encourageant et souligne les défis à surmonter pour cette tâche.

ABSTRACT
Acquisition and recognition of expressions and vocal calls in a smart home

This paper presents a system to recognize calls for help in the home of seniors to provide reassurance and assistance. The system is using an ASR which must operate with distant and expressive speech. Moreover, privacy is ensured by running the decoding on-site and not on a remote server. Furthermore the system was biased to recognize only set of sentences. The system has been evaluated in a smart space reproducing a typical living room where 17 participants played scenarios including falls. The results showed a promising error rate, 29%, while emphasizing the challenges of the task.

MOTS-CLÉS : traitement automatique de la parole, parole expressive, habitat intelligent.

KEYWORDS: automatic speech analysis, expressive speech, smart home.

1 Introduction

Le vieillissement rapide de la population des pays industrialisés représente l'un des défis majeurs du 21e siècle. En effet, en France, on estime que le nombre de Personnes Âgées (PA) de plus de 60 ans représentera 28,4% de la population en 2020 (9,4% auront plus de 75 ans) et 32,6% en 2060 (16,2% auront alors plus de 75 ans) (Blanpain & Chardon, 2010). Ceci est à mettre en relation avec l'espérance de vie sans incapacité qui a diminué pour être 61,9 ans en France en 2010, ce qui reste dans la moyenne de l'UE (INED, 2012). Lorsqu'une PA perd son autonomie, l'assistance d'un tiers devient nécessaire, celui-ci est généralement désigné sous le terme d'« aidant ». Ce rôle est en fait souvent tenu par un ensemble d'acteurs représentés majoritairement par la famille proche, en général le conjoint ou les enfants, qui doivent assumer les soins. Par ailleurs, cette augmentation du nombre de retraités aura un impact très important sur la société et les finances publiques à travers les régimes de retraite et la sécurité sociale (EPC, 2003). Le défi posé dès maintenant à notre société est de permettre à nos aînés de pouvoir vivre de façon autonome et confortable aussi longtemps que possible en toute quiétude alors même que le nombre de jeunes pouvant contribuer à leur support sera en constante

diminution.

Avec l'âge, le risque de chute croît. Environ 30% des personnes de plus de 65 chutent chaque année et ces chutes peuvent engendrer un fort taux de mortalité, morbidité et de souffrance pour les personnes âgées et leurs familles, elles augmentent par ailleurs la pression sur le coût social induit par une hospitalisation ou les soins médicaux (WHO Regional Office for Europe, 2004). D'autres phénomènes tels que les risques de crise cardiaques, arthrose etc. peuvent être des facteurs réduisant l'autonomie des personnes âgées et conduisent à un placement en institut spécialisé alors même que les personnes concernées peuvent encore posséder un degré d'autonomie suffisant pour vivre chez elles. L'une des principales raisons de ce placement restant la crainte d'un accident non signalé lorsque la personne est seule chez elle.

La détection des situations de détresse telles que les chutes, les immobilisations involontaires ou encore les évanouissements est un enjeu important pour soutenir la vie autonome des personnes âgées. Une solution communément proposée s'articule autour de capteurs cinématiques portés par la personne, ce qui constitue une contrainte dans la vie de tous les jours, avec des risques d'oubli voire même le refus de les porter car il sont jugés stigmatisants ou sans bénéfice immédiat (Bloch *et al.*, 2011). L'interface vocale peut constituer une alternative car ce type de technologie a atteint la maturité suffisante pour être utilisé ici tout en libérant la contrainte du port permanent des capteurs sur soi (Portet *et al.*, 2013). Un exemple type est celui d'un système de dialogue adapté à la personne. Par exemple, (Hamill *et al.*, 2009) ont développé un système PERS (*Personal Emergency Response System*) de dialogue vocal pour réagir à un appel spécifique d'urgence et décider de la réponse à apporter par une série de réponses fermées (« oui » et « non »). Un autre exemple est apporté par le projet ROBOCARE (Bahadori *et al.*, 2004) dans lequel un robot roulant comportant un écran d'ordinateur affichant un visage animé (avatar) a été conçu pour interagir spontanément avec l'utilisateur afin de lui signaler un danger ou répondre à une question.

Dans cet article nous présentons une approche permettant d'assister des personnes âgées dans leur maison par l'identification automatique d'appels vocaux, notamment dans le cas de situations de détresse où l'usage de la parole est encore possible. Une partie des résultats sont présentés dans (Vacher *et al.*, 2015c) ; la seconde partie résulte d'expériences effectuées sur un nouveau corpus. Les défis à relever pour rendre une telle application possible sont nombreux (Vacher *et al.*, 2011). Premièrement le cas d'application consiste à reconnaître de la parole distante puisque les microphones ne sont pas portés par la personne. Ceci implique de la parole atténuée, potentiellement réverbérée voire bruitée par d'autres sources sonores. Deuxièmement, la voix d'intérêt dans cette application présente de nombreuses différences avec la voix typique qui est celle des applications de Reconnaissance Automatique de la Parole (RAP) grand public, ces différences sont introduites par les modifications dues aux effets de l'âge du locuteur, à la situation fortement émotive dans laquelle il se trouve ainsi que l'occurrence de certaines pathologies. À ceci s'ajoute le manque de connaissance et de corpus réels sur l'appel de détresse. Les quelques études sur le sujet ont porté sur les urgences téléphoniques (Lefter *et al.*, 2011; Demenko & Jastrzebska, 2012) qui n'incluent pas nécessairement de personnes âgées.

Dans notre étude nous nous intéressons au cas particulier de la parole distante et expressive. L'objectif de l'étude est de mettre en place un Système de Reconnaissance Automatique de la Parole (SRAP) à l'état de l'art et de tester ses performances sur ce type de parole. Nous nous plaçons dans la situation où une PA est seule chez elle et possède un système de télé-lien social utilisant un microphone sans qu'aucun capteur ne soit porté par cette personne. Afin de recueillir un corpus d'étude réaliste, nous avons enregistré un ensemble de participants représentatifs de notre cible (au niveau de l'âge) simulant des situations de détresses dans le Living lab du laboratoire. Ces situations ont été observées lors d'une étude socio-ethnographique. En effet, enregistrer de la véritable voix de détresse dans un habitat utilisable pour une expérience reste un défi méthodologique et éthique. Dans un souci de respect de la vie privée, nous privilégions une approche *in situ* afin que la parole ne soit pas traitée par un serveur distant. Par ailleurs, l'utilisation d'un vocabulaire restreint assure la certitude que l'identification se limite aux appels vocaux. Cet article est organisé comme suit : la section 2 introduit les méthodes d'acquisition et de reconnaissance de la parole au sein de l'habitat ainsi que la méthode de détection

des appels à l'aide. La section 3 présente les expériences et finalement les résultats sont discutés dans la section 4.

2 Méthode

Pour mettre en place un système de reconnaissance d'appels vocaux en situation de détresse, la première étape a consisté à étudier ces situation sur le terrain et à définir leur environnement typique ainsi que le lexique utilisé. Cette étude est utile pour spécifier le système mais aussi pour mettre en place des scénarios de collecte de corpus. Le système global de reconnaissance d'appels vocaux est présenté avant de développer la modélisation acoustique et la méthode de reconnaissance.

2.1 Situations de détresse domestiques

La reconnaissance d'appels à l'aide s'effectue dans le contexte d'une maison équipée du dispositif *e-lio* [1], système de télé-lien social permettant de déclencher des appels (audio, vidéo ou urgence) entre les personnes âgées et leurs proches à l'aide d'une télécommande. Un objectif du projet était d'ajouter une commande vocale à ce système. Les paramètres de l'expérience ainsi que les situations de détresse ont été élaborés à partir d'une étude sociologique du laboratoire GRePS (Bobillier Chaumon *et al.*, 2013) auprès d'un grand nombre de personnes âgées qui a montré que cet équipement doit être installé sur une table de la pièce de vie, face au canapé et à la télévision. Ainsi une alerte peut facilement être lancée si la personne tombe à cause du tapis ou n'arrive pas à se lever du canapé. Plus de détails sont donnés dans l'article (Bouakaz *et al.*, 2014).

Ces études ont permis de spécifier le contexte des chutes, les mouvements pendant les chutes ainsi que la réaction des personnes une fois au sol. Les phrases prononcées pendant et après la chute ont également été identifiées : "*Ah, zut, qu'est-ce qui m'arrive ? Oh merde, merde !*".

2.2 Système d'analyse sonore en ligne

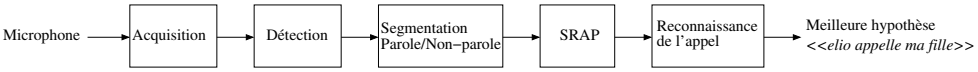

FIGURE 1: Architecture du système d'analyse sonore CIRDOX

Le traitement de l'audio est réalisé par le logiciel CIRDOX(Aman *et al.*, 2016). La Figure 1 décrit le pipeline de traitement. Le flux audio est acquis en continu et les événements sonores sont détectés à la volée en utilisant une décomposition en ondelettes associée à un seuillage adaptatif (Vacher *et al.*, 2004). Les événements sonores sont alors classifiés comme non-parole ou parole et dans ce cas envoyés au SRAP qui les transmet au système de reconnaissance des appels à l'aide.

Dans cet article, nous nous focalisons sur le SRAP et présentons différentes stratégies visant à améliorer la reconnaissance des appels à l'aide, la modélisation acoustique ainsi que l'étape de reconnaissance.

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

2.3 Modélisation langagière

La modélisation acoustique est basée sur des Modèles de Markov Cachés (MMC) gauche-droite à trois états dépendants du contexte. Les paramètres acoustiques sont basés sur des MFCC à 40 paramètres (13 coefficients + delta + delta delta + énergie). Au final les modèles acoustiques comportent 11000 états partagés dépendants du contextes avec un total de 150000 gaussiennes. Par ailleurs une adaptation SAT+fMLLR est appliquée (Povey *et al.*, 2011b).

Les modèles ont été estimés sur 500 heures de français annotées, issues d'émissions radiophonique ou TV (ESTER 1 + 2 + REPERE) (Gravier *et al.*, 2004) et 7 heures précédemment enregistrées dans un appartement intelligent (Vacher *et al.*, 2014) qui regroupe 60 locuteurs interagissant avec l'appartement et 28 minutes du corpus "voix-détresse" (Aman, 2014) composé d'enregistrements de locuteurs énonçant des appels de détresse.

2.3.1 Modèles acoustiques « Subspace GMM »

Les modèles à base de mélanges de gaussiennes et les modèles à sous-espaces sont utilisés pour le calcul d'émission des probabilités au sein des états du MMC, avec la particularité pour les SGMM d'avoir des moyennes et poids générés à partir de sous-espaces de mélanges de gaussiennes, via une projection pondérée.

Les modèles SGMM introduits par (Povey *et al.*, 2011a) sont décrits par l'équation suivante :

$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i)$ avec $\mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}$ et $w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}$.

où $\mathbf{x}$ correspond au vecteur de paramètres, $j \in \{1..J\}$ est un état du MMC, i est l'index de la gaussienne, m est un sous-état de j et c_{jm} le poids du sous-état m. À chaque état j sont associés un vecteur $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S est la dimension du sous-espace phonétique), les moyennes μ_{jmi}, la pondération des mélanges w_{jmi} et un nombre de Gaussiennes partagé I. Le sous-espace phonétique $\mathbf{M}_i$, le poids de projection $\mathbf{w}_i^T$ et la matrice de covariance $\mathbf{\Sigma}_i$, c.-à-d. les paramètres globaux $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$, sont communs à tous les états. Ces paramètres peuvent être partagés et estimés à travers différentes conditions d'enregistrement.

Un mélange générique de gaussiennes I appelé *Universal Background Model (UBM)* est estimé sur toutes les données d'apprentissage, et sert à l'initialisation du SGMM.

Dans nos expériences nous avons estimé séparément trois UBM contenant 1K gaussiennes à partir des données SWEET-HOME (7h), Voix-détresse (28mn) et ESTER+REPERE (500h). Ces trois UBM ont été fusionnés en un seul (en fusionnant les paires les plus proches (Zouari & Chollet, 2006)) ; le modèle final est alors spécifiquement biaisé par les enregistrements dans la maison ainsi que par la parole expressive.

2.3.2 Modèles de langage

Cet article est focalisé sur la reconnaissance d'appels à l'aide. Pour ce faire, les phrases du corpus AD80 (Aman *et al.*, 2013) ont servi de base pour développer le modèle de langage spécialisé du système. Ce corpus a été enregistré par 43 personnes âgées et 52 non âgées dans notre laboratoire ainsi que dans une maison de retraite, avec pour objectif d'étudier la reconnaissance de la parole pour les personnes âgées. Ce corpus est composé pour chaque locuteur de 81 phrases classiques, 31 commandes vocales pour la domotique et 58 phrases d'appel à l'aide.

Quelques unes de ces phrases sont données dans la table 1. Les phrases d'appel identifiées dans l'étude reportée à la section 2.1 ont été incluses dans AD80.

2.4 Reconnaissance des appels à l'aide

Comme dit précédemment, la voix en situation de détresse peut être assez atypique et perturber le SRAP. Cependant, dans de nombreuses situations, un groupe de mots mal décodé est phonétiquement proche d'un segment "bon candidat". C'est pourquoi notre approche consiste à transcrire chaque appels possibles et l'hypothèse du SRAP en un graphe de phonèmes dans lequel chaque chemin correspond à une variante phonétique. Cette approche permet de prendre en considération un certain nombre d'erreurs de reconnaissance comme des fins de mots erronées ou des variations légères. Ainsi, chaque hypothèse du SRAP H_i est phonétisée et chaque commande vocale T_j est alignée à l'hypothèse H_i en utilisant une distance de Levenshtein. Le nombre de phrases prédéfinies est de 50 et correspondent à 11 actions différentes.

La décision de sélectionner ou non une séquence détectée est alors prise en fonction d'un seuil défini empiriquement dépendant du nombre de phonèmes alignés pour chaque candidat. À partir de là, le taux d'erreur domotique (TED) est défini par la formule : $\text{TED} = \frac{\text{Nombre d'appels manqués+Nombre de fausses alarmes}}{\text{Nombres d'appels}}$

3 Expériences et résultats

3.1 Expériences actées en situation réaliste : le corpus de parole CIRDO

Les enregistrements ont été réalisé dans une pièce du Living lab du LIG équipée comme décrit précédemment. Le protocole ainsi que le corpus CIRDO obtenu sont détaillés dans (Vacher *et al.*, 2016). Les scénarios comportaient des appels à l'aide lors de chute ou en cas de blocage de hanche. Tous les enregistrements audio ont été transcrits manuellement en utilisant *Transcriber* (Barras *et al.*, 2001) et les segments audio ont alors été extraits pour analyse. Durant le scénario, certains participants soupiraient, gémissaient, toussaient, criaient ou se raclaient la gorge : ces sons n'ont pas été transcrits. Dans le même ordre d'idée, les paroles mélangées avec les bruits de chute ont été ignorées. À la fin, chaque locuteur avait prononcé entre 10 et 65 phrases ou interjections ((“*ah*”, “*oh*”, “*aïe*”, etc.) (Table 2). Certaines phrases étaient proches de celles prévues dans les scénarios (“*je peux pas me relever*”, “*e-lio appelle du secours*”, etc.), tandis que d'autres différaient (“*oh bein on est bien là tiens* ”). Dans la pratique, les participants font de longues pauses au milieu d'une phrase ou prononcent des phrases spontanées, utilisent des interjections ou des sons non-verbaux.

3.2 Reconnaissance « off-line »des appels sur le corpus CIRDO

Le SRAP utilisé est basé sur Kaldi (Povey *et al.*, 2011b). Le modèle SGMM présenté dans la section 2.3 a été utilisé comme modèle acoustique. Le modèle de langage générique a été estimé sur des corpus

Appels à l'aide	*Commandes domotiques*	*Phrases classiques*
Aïe aïe aïe ⋆	Appelle quelqu'un e-lio ⋆	Bonjour madame
Oh là ⋆	e-lio, appelle quelqu'un ⋆	Ça va très bien
Merde ⋆	e-lio tu peux appeler une ambulance	Où sont mes lunettes
Je suis tombé ⋆	e-lio tu peux téléphoner au SAMU	Le café est brûlant
Je peux pas me relever ⋆	e-lio, appelle du secours	J'ai ouvert la porte
Qu'est-ce qu'il m'arrive ⋆	e-lio appelle les secours	Je me suis endormi tout de suite
Aïe ! J'ai mal ⋆	e-lio appelle ma fille	Il fait soleil
Oh là ! Je saigne ! Je me suis blessé ⋆	e-lio appelle l'infirmière	Ce livre est intéressant
Aidez-moi	e-lio appelle le SAMU !	Je dois prendre mon médicament
Au secours	e-lio appelle les pompiers !	J'allume la lumière

TABLE 1: Exemples extraits du corpus AD80 (les ⋆ correspondent à des phrases observées dans l'étude ethnographique)

Loc.	Age	Sexe	Nb. d'interjections ou phrases courtes		Loc.	Age	Sexe	Nb. d'interjections ou phrases courtes	
			Total	*# Appel*				*Total*	*# Appel*
S01	30	M	22	14	S10	16	M	19	15
S02	-	-	-		S11	52	M	12	12
S03	24	F	16	15	S12	28	M	15	12
S04	83	F	65	53	S13	66	M	24	21
S05	29	M	24	21	S14	52	F	23	2
S06	64	F	23	19	S15	23	M	20	19
S07	61	M	23	21	S16	40	F	29	27
S08	44	M	25	15	S17	40	F	24	21
S09	16	M	32	21	S18	25	F	17	14
Total	40.76		413	341					

TABLE 2: Composition du corpus audio CIRDO

Loc.	TEM (%)		TED	Loc.	TEM (%)		TED
	Total	*# Appel*			*Total*	*# Appel*	
S01	45,0	39,1	27,8	S11	21,3	17,0	16,7
S03	41,4	44,4	40,0	S12	30,8	25,0	25,0
S04	51,9	49,6	34,0	S13	45,9	43,6	23,8
S05	19,1	15,4	14,3	S14	67,0	54,8	50,0
S06	39,2	34,3	26,3	S15	21,5	19,5	5,3
S07	21,2	20,3	28,6	S16	14,9	11,76	7,4
S08	61,8	50,8	20,0	S17	21,4	22,4	19,0
S09	49,4	41,2	33,3	S18	57,7	44,9	71,4
S10	24,5	22,4	14,3				
Total	39,3	34,0	26,8				

TABLE 3: Taux d'erreur mot et domotique pour chaque participant

de journaux ainsi que sur Gigaword. Le lexique est d'environ 13K mots. Pour réduire la variabilité linguistique, ce modèle générique a été interpolé avec un modèle de langage lié au domaine se basant sur les scénarios. Le modèle de langage final est le résultat d'une interpolation avec 10% pour le modèle générique et 90% pour le modèle spécialisé. Cette interpolation a montré des résultats corrects dans nos précédentes expérimentations (Lecouteux *et al.*, 2011). L'avantage de cette interpolation permet de biaiser le modèle pour reconnaître des situations d'appel à l'aide tout en réduisant la reconnaissance de faux positifs.

Les résultats sur les données manuellement annotées sont donnés dans la table 3. Si nous considérons uniquement les appels à l'aide, le Taux d'Erreur de Mot (TEM) est 34% tandis qu'il monte à 39.3% quand toutes les interjections et phrases sont prises en considération. Malheureusement, le corpus utilisé ne permet pas de déterminer le taux de fausses alarmes car il ne contient pas de scénario comportant des appels qui ne soient pas des appels à l'aide. Nos précédentes études basées sur le corpus AD80 montrent un rappel, précision et F-mesure d'environ 88,4 %, 86,9 % and 87,2 % (Aman *et al.*, 2013). Cependant, ce corpus a été enregistré dans des conditions très différentes de celles d'un studio, comme l'avait été CIRDO. En moyenne, le TED est 26,8 %.

4 Discussion

Si nous comparons ces résultats avec ceux obtenus en utilisant le corpus de parole lue AD80, ils sont nettement inférieurs, le TEM est 26,8% contre 14,5% (Aman *et al.*, 2013). Ceci peut s'expliquer par des différences notables entre les conditions d'enregistrement des 2 corpus :

— AD80 est basé sur des enregistrements de locuteurs qui sont confortablement assis et en condition proche face au microphone, tandis que CIRDO l'a été en condition distante,

— le corpus CIRDO a été enregistré avec des participants qui tombent sur le sol ou se trouvent bloqués sur le canapé. De plus, ils ont été encouragés à se mettre en situation et parler avec l'émotion qu'ils auraient ressentie dans ce type de situation. Les enregistrements contiennent

donc une parole expressive, mais ceci ne garantit pas totalement qu'elle l'aurait été de cette manière en condition réelle.

Par contre, si nous comparons maintenant ces résultats avec ceux obtenus en ligne avec CIRDOX et un SRAP utilisant Sphinx3 (Lee *et al.*, 1990) et des modèles acoustiques de type GMM adaptés au locuteur, nous observons une amélioration notable (Vacher *et al.*, 2015a). En effet le TEM était 49,5% (contre 34%) et le TED 33% (contre 26,8%), ceci semble indiquer que des modèles plus élaborés, et qui prennent en compte plus de données d'apprentissage comme les sGMM, permettent une amélioration sensible des résultats de reconnaissance sur de la parole enregistrée en conditions difficile.

Si l'on observe le taux d'erreur au niveau des appels (TED), on observe qu'il est de 26.8% ; de plus, à l'exception d'un seul locuteur, le TED est toujours en dessous de 50% et inférieur à 20% pour 6 locuteurs. Cela suggère qu'un appel à l'aide a de meilleures chances d'être détecté si le locuteur est en capacité de le répéter deux ou trois fois. Cependant, si le système n'identifie pas le premier appel de détresse à cause de la voix altérée par l'émotion, il y a des chances que cette non-détection augmente encore l'émotion ressentie par la personne, ce qui aura pour conséquence une difficulté accrue pour le système de reconnaissance. Dans le même ordre d'idée, ce corpus a été enregistré en conditions réalistes, mais ne reproduisant pas totalement la réalité : la production vocale des personnes âgées fragiles est difficilement reproductible par des personnes jeunes. Il convient d'éviter au maximum la nécessité pour les personnes d'avoir à répéter un appel, il est donc important de poursuivre les efforts en vue de résoudre ces erreurs.

5 Conclusion et perspectives

Cette étude s'est focalisée sur la RAP dans le cadre de maisons intelligentes et en conditions distantes et réalistes : des conditions très différentes de celles d'un corpus enregistré assis et proche du micro. En effet, le corpus CIRDO constitué d'appels de détresse en parole distante et qui simulent des conditions réalistes (chutes, blocage sur un canapé) a été utilisé. Le TEM obtenu en sortie d'un système SRAP dédié était 36,3% au niveau des appels de détresse. Cependant, grâce à un filtrage des hypothèse au niveau phonétique, plus de 70% des appels ont pu être détectés.

Ces résultats obtenus en conditions réalistes donnent une bonne idée des performances qui peuvent être obtenues avec un SRAP état de l'art dans ces conditions spécifiques avec des utilisateurs finaux. Ces résultats on été obtenus dans des cas de situations de détresse, perturbées par les émotions ; ce type d'expérimentation pourra être étendu à d'autres situations où l'expressivité est particulièrement marquée.

Comme expliqué précédemment, les résultats obtenus doivent être améliorés afin que le système puisse être utilisé en « production ». Deux axes peuvent être explorées, le premier pourrait être une adaptation des modèles acoustiques à la prosodie de la parole expressive. L'enregistrement du type de corpus nécessaire à cette adaptation est délicat car il doit se faire en conditions réelles, ce qui représente un inconvénient majeur. Une autre piste serait la reconnaissance de répétitions à intervalles réguliers d'évènements phonétiquement similaires : même si la reconnaissance de la parole en tant que telle n'est pas bonne, cela peut être signe d'un problème à régler rapidement.

Ces travaux montrent la nécessité de procéder à des analyses basées sur des corpus enregistrés en conditions écologiques. Il sont complémentaires à d'autres travaux sur la commande vocale de la domotique (Vacher *et al.*, 2015b) pour lesquels l'aspect parole expressive n'était pas prédominant mais qui ont montré la nécessité d'une adaptation au vocabulaire et aux tournures propres de chaque utilisateur au moyen d'une analyse lexicale.

Références

AMAN F. (2014). *Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile*. PhD thesis, Université de Grenoble, Ecole doctorale MSTII.

AMAN F., VACHER M., PORTET F., DUCLOT W. & LECOUTEUX B. (2016). CirdoX : an On/Off-line Multisource Speech and Sound Analysis Software. In *LREC 2016*. (Accepted Paper).

AMAN F., VACHER M., ROSSATO S. & PORTET F. (2013). Speech Recognition of Aged Voices in the AAL Context : Detection of Distress Sentences. In *The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013*, p. 177–184, Cluj-Napoca, Romania.

BAHADORI S., CESTA A., GRISETTI G., IOCCHI L., LEONE R., NARDI D., ODDI A., PECORA F. & RASCONI R. (2004). RoboCare : Pervasive Intelligence for the Domestic Care of the Elderly. *Intelligenza Artificiale*, **1**(1), 16–21.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1-2), 5–22.

BLANPAIN N. & CHARDON O. (2010). Projections de population à l'horizon 2060 : Un tiers de la population âgé de plus de 60 ans. Institut national de la statistique et des etudes economiques (France). [in French].

BLOCH F., GAUTIER V., NOURY N., LUNDY J., POUJAUD J., CLAESSENS Y. & RIGAUD A. (2011). Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects. *Annals of Physical and Rehabilitation Medicine*, **54**, 391–398.

BOBILLIER CHAUMON M., CROS F., CUVILLIER B., HEM C. & CODREANU E. (2013). Concevoir une technologie pervasive pour le maintien à domicile des personnes âgées : la détection de chutes dans les activités quotidiennes. In *Activités Humaines, Technologies et bien-être, Congrès EPIQUE (Psychologie Ergonomique)*, p. 189–199, Belgique - Bruxelles.

BOUAKAZ S., VACHER M., BOBILLIER-CHAUMON M.-E., AMAN F., BEKKADJA S., PORTET F., GUILLOU E., ROSSATO S., DESSERÉE E., TRAINEAU P., VIMON J.-P. & CHEVALIER T. (2014). CIRDO : Smart companion for helping elderly to live at home for longer. *Innovation and Research in BioMedical engineering (IRBM)*, **35**(2), 101–108.

DEMENKO G. & JASTRZEBSKA M. (2012). Analysis of voice stress in call centers conversations. In *Speech Prosody*.

EPC (2003). The impact of ageing populations on public finances : overview of analysis carried out at EU level and proposals for a future work programme. Economic Policy Committee, Union Européenne.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC* : European Language Resources Association.

HAMILL M., YOUNG V., BOGER J. & MIHAILIDIS A. (2009). Development of an automated speech recognition interface for personal emergency response systems. *Journal of NeuroEngineering and Rehabilitation*, **6**.

INED (2012). Dernières données sur l'espérance de vie sans incapacité des 27 pays de l'ue. Communiqué de Presse de l'Ined.

LECOUTEUX B., VACHER M. & PORTET F. (2011). Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. In *Proc. InterSpeech*, p. 2273–2276.

LEE K.-F., HON H.-W. & REDDY R. (1990). An overview of the SPHINX speech recognition system. *IEEE TASSP*, **38**(1), 35–45.

LEFTER I., ROTHKRANTZ L. J., VAN LEEUWEN D. A. & WIGGERS P. (2011). Automatic stress detection in emergency (telephone) calls. *International Journal of Intelligent Defence Support Systems*, **4**(2), 148–168.

PORTET F., VACHER M., GOLANSKI C., ROUX C. & MEILLON B. (2013). Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects. *Personal and Ubiquitous Computing*, **17**(1), 127–144.

POVEY D., BURGET L., AGARWAL M., AKYAZI P., KAI F., GHOSHAL A., GLEMBEK O., GOEL N., KARAFIÁT M., RASTROW A., ROSE R. C., SCHWARZ P. & THOMAS S. (2011a). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, **25**(2), 404 – 439.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011b). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

VACHER M., AMAN F., ROSSATO S. & PORTET F. (2015a). Development of Automatic Speech Recognition Techniques for Elderly Home Support : Applications and Challenges. In *HCII, ITAP 2015*, volume Part II of *LNCS 9194*, p. 341–353, Los Angeles, CA, United States.

VACHER M., BOUAKAZ S., BOBILLIER CHAUMON M.-E., AMAN F., KHAN R. A., BEKKADJA S., PORTET F., GUILLOU E., ROSSATO S. & LECOUTEUX B. (2016). The CIRDO corpus : comprehensive audio/video database of domestic falls of elderly people. In *LREC 2016*. (Accepted Paper).

VACHER M., CAFFIAU S., PORTET F., MEILLON B., ROUX C., ELIAS E., LECOUTEUX B. & CHAHUARA P. (2015b). Evaluation of a context-aware voice interface for Ambient Assisted Living : qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing*, **7**(issue 2), 5 :1–5 :36.

VACHER M., ISTRATE D. & SERIGNAT J. (2004). Sound detection and classification through transient models using wavelet coefficient trees. In S. LTD, Ed., *Proc. 12th European Signal Processing Conference*, p. 1171–1174, Vienna, Austria.

VACHER M., LECOUTEUX B., AMAN F., ROSSATO S. & PORTET F. (2015c). Recognition of Distress Calls in Distant Speech Setting : a Preliminary Experiment in a Smart Home. In *6th Workshop on Speech and Language Processing for Assistive Technologies*, p. 1–7, Dresden, Germany : SIG-SLPAT.

VACHER M., LECOUTEUX B., CHAHUARA P., PORTET F., MEILLON B. & BONNEFOND N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, p. 4499–4506, Reykjavik, Iceland.

VACHER M., PORTET F., FLEURY A. & NOURY N. (2011). Development of Audio Sensing Technology for Ambient Assisted Living : Applications and Challenges. *International Journal of E-Health and Medical Communications*, **2**(1), 35–54.

WHO REGIONAL OFFICE FOR EUROPE (2004). What are the main risk factors for falls amongst older people and what are the most effective interventions to prevent these falls ? World Health Organisation.

ZOUARI L. & CHOLLET G. (2006). Efficient gaussian mixture for speech recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, p. 294–297.