Reconnaissance automatique de la parole distante dans un habitat intelligent : méthodes multi-sources en conditions réalistes (Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions) [in French]

Benjamin Lecouteux, Michel Vacher, François Portet

Reconnaissance automatique de la parole distante dans un habitat intelligent : méthodes multi-sources en conditions réalistes

Benjamin Lecouteux, Michel Vacher, François Portet
Laboratoire Informatique de Grenoble, équipe GETALP
prénom.nom@imag.fr

RÉSUMÉ

Le domaine des maisons intelligentes s'est développé dans le but d'améliorer l'assistance aux personnes en perte d'autonomie. La reconnaissance automatique de la parole (RAP) commence à être utilisée, mais reste en retrait par rapport à d'autres technologies. Nous présentons le projet Sweet-Home ayant pour objectif le contrôle de l'environnement domestique par la voix. Plusieurs approches, état de l'art et nouvelles, sont évaluées sur des données enregistrées en conditions réalistes. Le corpus de parole distante, enregistré auprès de 21 locuteurs simule des scénarios intégrant des activités journalières dans un appartement équipé de plusieurs microphones. Les techniques opérant au cours du décodage et utilisant des connaissances *a priori* permettent d'obtenir des résultats très intéressants par rapport à un système RAP classique.

ABSTRACT

Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions

While the smart home domain has become a major field of application of ICT to improve support and wellness of people in loss of autonomy, speech technology in smart home has, comparatively to other ICTs, received limited attention. This paper presents the SWEET-HOME project whose aim is to make it possible for frail persons to control their domestic environment through voice interfaces. Several state-of-the-art and novel ASR techniques were evaluated on realistic data acquired in a multiroom smart home. This distant speech French corpus was recorded with 21 speakers playing scenarios including activities of daily living in a smart home equipped with several microphones. Techniques acting at the decoding stage and using a priori knowledge such as DDA give better results than the baseline and other approaches (Lecouteux *et al.*, 2011).

MOTS-CLÉS : domotique, parole distance, habitat intelligent, SRAP multisource.

KEYWORDS: home automation, smart home, distant speech, multisource ASRs.

1 Introduction

Les récentes avancées dans les systèmes ubiquitaires ont fait apparaître de nouveaux concepts d'environnement domotiques : les maisons intelligentes. Ce sont des habitations équipées de

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 657–664,
Grenoble, 4 au 8 juin 2012.

capteurs, d'actionneurs et dispositifs automatisés, régulés par des logiciels. Ainsi le contrôle automatisé de l'habitat permet d'y régler la luminosité, les volets électriques mais aussi la Hi-Fi, les PC, les alarmes etc. Ces maisons intelligentes représentent une solution pour l'aide aux personnes isolées, en perte d'autonomie afin qu'elles aient la possibilité de rester chez elles et de conserver une certaine indépendance. Parmi toutes les technologies d'interaction homme-machine, la reconnaissance automatique de la parole semble offrir le plus de potentiel : cette modalité est adaptée à des personnes âgées qui ont des difficultés de déplacement ou de vision. Par exemple, une interface tactile (télécommande) nécessitera à la fois des interactions visuelles et physiques (Vovos *et al.*, 2005; Hamill *et al.*, 2009). De plus les commandes vocales sont particulièrement adaptées dans les situations de détresse : une personne ne pouvant plus bouger après une chute aura toujours la possibilité d'appeler de l'aide. Malgré ces aspects la reconnaissance automatique de la parole (RAP) a rarement été utilisée dans ce cadre (Vovos *et al.*, 2005; Hamill *et al.*, 2009). Ceci est en partie dû à la difficulté de mettre en oeuvre un système de RAP dans un environnement réel (Vacher *et al.*, 2011).

Le projet Sweet-Home [1] a débuté courant 2010 et relève plusieurs défis. L'un d'eux est l'utilisation de technologies liées à la RAP dans des environments bruités (Vacher *et al.*, 2011) : les SRAP obtiennent des résultats corrects lorsque les locuteurs sont proches des micros, mais les performances se dégradent rapidement dès qu'ils s'en éloignent. En conditions réelles cette dégradation est accentuée par d'autres effets (Vacher *et al.*, 2008) tels que les réverbérations, les bruits de fond (TV, radio, travaux...) etc. Ces problèmes liés à la RAP distante doivent donc être abordés dans le contexte d'une habitation (Wölfel et McDonough, 2009). Tandis que les préférences linguistiques et les interactions vocales en fonction de l'âge ont été étudiées ces dernières décennies (Vovos *et al.*, 2005; Hamill *et al.*, 2009; Vipperla *et al.*, 2009), la parole distante dans les maisons intelligentes commence tout juste à être abordée dans la communauté (Barker *et al.*, 2011).

Cet article présente des résultats état de l'art et de nouvelles techniques utilisant la RAP sur des données enregistrées en conditions réalistes. La section 2 présente le projet et le corpus associé. La section 3 présente les différentes techniques exploitées. Ensuite, la section 5 expose le cadre expérimental et les expériences réalisées accompagnées de leurs résultats et finalement, nous concluons et proposons quelques perspectives.

2 Le projet Sweet-home et son corpus

Le projet Sweet-Home (sweet-home.imag.fr) propose de développer une maison intelligente basée sur un SRAP. Ce projet se focalise sur trois aspects : fournir une assistance utilisant une interaction homme-machine naturelle (commandes vocales et tactiles), la capacité d'être utilisé par tout à chacun et la détection de situations de détresse. L'objectif est donc que l'utilisateur puisse piloter son environnement à tout instant depuis n'importe quel lieu de sa maison, et ce le plus naturellement possible. L'environnement intelligent visé utilise un SRAP opérant à travers toutes les pièces via des micros placés dans les plafonds. Cette configuration soulève des problèmes liés à la parole distante où les micros sont éloignés du locuteur et enregistrent des sons extérieurs très variés. Les travaux effectués dans ce domaine se sont focalisés sur une seule

1. Cette étude a été financée par l'Agence Nationale de la Recherche dans le cadre du projet Sweet-Home (ANR-2009-VERS-011). Nous remercions particulièrement les différentes personnes qui ont accepté de participer aux enregistrements.

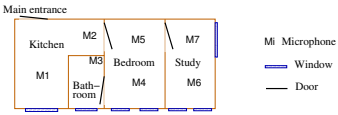

FIGURE 1 – Position des 7 micros dans l'appartement DOMUS

pièce (Vovos *et al.*, 2005) ou un nombre de micros non spécifié (Hamill *et al.*, 2009; Vipperla *et al.*, 2009).

Pour développer ce projet, un appartement témoin (projet DOMUS) a été équipé afin d'acquérir un corpus réaliste et d'expérimenter différentes techniques. Cet habitat intelligent a été mis en place par l'équipe Multicom du LIG, partenaire de ce projet. L'appartement fait environ 34 m^2 de plein pied. Il inclut une salle de bains, une cuisine, une chambre et un salon. Chaque pièce est équipée de détecteurs de présence, caméras (utilisées uniquement pour l'annotation) etc. De plus, 7 micros ont été placés dans les plafonds (Figure 1).

Une expérience a été menée afin d'acquérir un corpus parlé composé de phrases se divisant en plusieurs catégories : domotiques, appels de détresse et des phrases de la vie courante. 21 personnes dont 7 femmes ont participé aux enregistrements , en jouant des activités de la vie quotidienne. L'âge moyen des participants est de 38.5 (écart type : ± 13) ans. Afin d'assurer des enregistrements proches de la vie courante, il a été demandé aux participants d'avoir des activités dans l'appartement (se lever, s'habiller, faire la vaisselle...). Une visite préalable a été organisée afin de familiariser les participants avec leur environnement. Durant les enregistrements, aucune instruction n'a été donnée sur la manière de parler ou de s'orienter. Il en résulte que les participants n'ont pas parlé forcément en direction des micros et qu'ils pouvaient se déplacer lorsqu'ils parlaient. La distance la plus courte entre un micro et le locuteur est d'environ deux mètres. Les sons ont ainsi été enregistrés en temps réel sur 7 voies à l'aide d'une machine dédiée et disposant d'une carte son 8 voies (Vacher *et al.*, 2011).

La première phase (P1) a consisté à dérouler un scénario d'activités librement et sans contrainte de temps (prendre un déjeuner ou une douche, faire une sieste, passer l'aspirateur...). Au cours de cette phase, les participants ont prononcé 40 phrases prédéfinies de la vie courante (ex : *allô, j'ai eu du mal à dormir*), avec la liberté de les prononcer là où ils le souhaitaient. La seconde phase (P2) s'est articulée autour de la lecture de 44 phrases dont 9 issues de situations de détresse (ex : *appelez un docteur, j'ai mal, à l'aide*) et 3 des ordres domotiques (ex : *allumez la lumière, allumer ordinateur*). Dans cet article, les résultats sont restreints à la partie du corpus non bruitée (sans TV, radio ou aspirateur).

Au final, le corpus Sweet-home comporte 862 phrases (38mn46s par canal, le même enregistrement étant fait sur plusieurs canaux) pour P1 et 917 phrases (40mn27s par canal) pour P2. Chaque phrase a été enregistrée sur tous les canaux et annotée manuellement. Le meilleur Ratio Signal Bruit (RSB, en sélectionnant le meilleur canal) est en moyenne de 20.3 db, condition acceptable pour faire de la RAP. Cependant, dans nos expérimentations, nous avons exploité l'ensemble des 7 microphones.

3 Approches proposées

La détection des ordres domotiques dans le contexte de Sweet-Home s'articule autour d'une stratégie en trois étapes. La première consiste en la détection des activités audio et leur classification : parole ou bruit. La seconde extrait les phrases des sons de type parole en utilisant un SRAP. Enfin la dernière étape reconnaît les ordres domotiques ou des situations de détresse. Cet article se focalise sur les deux dernières étapes. La première est quant à elle décrite dans (Vacher *et al.*, 2008).

Pour aborder les problèmes liés au contexte (bruits, distance) tout en bénéficiant des conditions d'enregistrement (plusieurs micros enregistrant en continu), nous proposons de tester l'impact de techniques état de l'art qui permettent de fusionner des flux d'information à différents niveaux du traitement automatique de la parole : acoustique, décodage de la parole et à la sortie du SRAP. La prochaine section présente les différentes techniques envisagées pour obtenir un SRAP robuste.

3.1 Fusion des flux acoustiques

Au niveau acoustique, il peut être intéressant de fusionner les différents canaux afin d'améliorer le signal. Cependant une simple somme des signaux résulterait en un signal de qualité médiocre avec échos et bruits amplifiés. C'est la raison pour laquelle nous nous sommes intéressés à l'utilisation d'un algorithme dit de *beamforming* (Anguera *et al.*, 2007) conçu pour fusionner correctement différents canaux enregistrant une même source à différentes distances. Cette méthode demande des calculs raisonnables tout en permettant une combinaison efficace de plusieurs flux.

L'algorithme utilisé ici est basé sur la théorie de la pondération et sommation de canaux. Etant donnés M microphones, le signal de sortie $y[t]$ est calculé par l'équation suivante : $y[t] = \sum_{m=1}^{M} W_m[t] x_m[t - D^{(m,ref)}[t]$ Où $W_m[t]$ est le poids accordé au microphone m à un instant t t, sachant que $\sum_{m=1}^{M} W_m[t] = 1$; le signal du m^{th} canal est $x_m[t]$ et $D^{(m,ref)}[t]$ le délai entre le m^{th} canal et le canal choisi comme référence. Dans notre cas, le canal de référence est celui de meilleur RSB (il peut donc varier). Les 7 canaux ont ainsi été combinés pour chaque locuteur : une fois le signal y calculé, il peut être utilisé comme un signal monosource classique.

3.2 Décodage guidé

Au niveau du décodage, nous avons proposé une nouvelle version du décodage guidé (Driven Decoding Algorithm, DDA) qui permet d'aligner et de corriger à la volée des transcriptions auxiliaires en utilisant un SRAP (Lecouteux *et al.*, 2006). Cet algorithme améliore la qualité du système primaire en s'appuyant sur la disponibilité de transcriptions auxiliaires.

Le DDA agit sur chaque nouvelle hypothèse générée par le SRAP : elle est alignée à la volée avec la transcription auxiliaire (issue d'un décodage précédent). Dès lors, un score de similarité α est calculé pour pondérer le modèle de langage (Lecouteux *et al.*, 2006) : $\tilde{P}(w_i|w_{i-1}, w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1}, w_{i-2})$ où $\tilde{P}(w_i|w_{i-1}, w_{i-2})$ est la probabilité pondérée du mot w_i sachant son historique w_{i-2}, w_{i-3}, et $P(w_i|w_{i-1}, w_{i-2})$ est la probabilité initiale du trigramme.

Nous proposons ensuite une variante du DDA où la sortie d'un premier microphone est utilisée pour guider la sortie d'un autre microphone (Figure 2). Cette approche présente deux avantages :

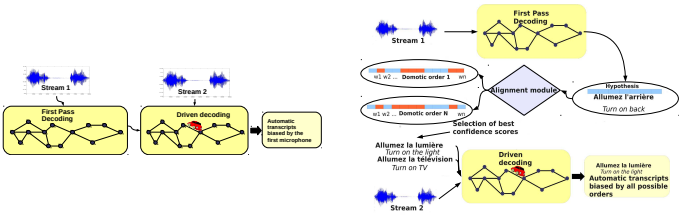

FIGURE 2 – Décodage guidé par un micro (à gauche) puis avec une information a priori (à droite)

– la vitesse du second SRAP est augmentée grâce à la présence de la transcription auxiliaire (seulement 0.1x le temps réel (TR)),
– le DDA permet de fusionner efficacement l'information issue de deux flux là où une stratégie de vote (telle que ROVER) ne peut fonctionner sans mesures de confiance.
La stratégie basée sur DDA est dynamique et utilisée pour chaque phrase décodée. Le premier décodage est effectué sur le canal de meilleur RSB et le DDA est appliqué sur le second.

Cette approche a été étendue pour prendre en considération les connaissances *a priori* sur les phrases attendues. Le SRAP est alors guidé par des patrons identifiés sur le premier micro. Cette méthode nommée DDA à deux niveaux projette les segments reconnus lors de la première passe dans un réseau de confusion (RC) comprenant les trois meilleures hypothèses de phrases attendues. Ce RC est alors utilisé pour guider le SRAP (Figure 2).

3.3 Vote par consensus

Pour effectuer une combinaison post-SRAP, nous avons utilisé une méthode ROVER (Fiscus, 1997) qui permet d'améliorer la qualité de plusieurs sorties de SRAP en effectuant un vote par consensus au niveau mot. Le principe consiste à fusionner les sorties en RC où chaque mot est pondéré en fonction de sa présence dans les différents systèmes. Le mot de meilleur score est alors sélectionné. Cette approche demande une forte charge de calculs, étant donné que chaque canal doit être préalablement décodé avec un SRAP (dans notre cas, 7 SRAP).

Notre système de référence ROVER utilise tous les canaux disponibles sans connaissance *a priori*. Ensuite, nous avons introduit une mesure de confiance basée sur le RSB : pour chaque segment décodé s_i issu du i^{eme} SRAP, la mesure de confiance associée $\phi(s_i)$ a été calculée ainsi : $\phi(s_i) = 2^{R(s_i)}/\sum_{j=1}^{7} 2^{R(s_j)}$ où $R()$ est la fonction calculant le RSB d'un segment et s_i est le segment généré par le i^{eme} SRAP. Pour chaque phrase un silence de durée I_{sil} a été rajouté au début et à la fin du signal de parole I_{speech}. Le RSB est alors calculé comme suit : $R(S) = 10 * log(\frac{\sum_{n \in I_{parole}} S[n]^2}{|I_{parole}|} / \frac{\sum_{n \in I_{sil}} S[n]^2}{|I_{sil}|})$.

Finalement, un ROVER utilisant seulement les deux meilleurs canaux a été expérimenté afin d'évaluer le degré de redondance entre les différents canaux. Ce ROVER 2 canaux permet également d'obtenir des résultats corrects avec une quantité de calculs raisonnable.

4 Détection des ordres domotiques et des phrases de détresse

Nous proposons de phonétiser automatiquement chaque phrase cible. Ainsi toutes les phrases attendues sont représentées sous la forme d'un graphe de phonèmes (avec variantes de prononciation). Le nombre de phrases à détecter est de 12 (3 ordres domotiques et 9 phrases de détresse). Les transcriptions automatiques ont également été phonétisées sur le même principe.

Pour détecter des ordres domotiques au sein des trancriptions automatiques T de taille m, chaque phrase de taille n est alignée sur T en utilisant une distance d'édition phonétique. Les coûts de suppression, insertion, substitution sont calculés empiriquement. La distance cumulée $\gamma(i,j)$ entre H_j et T_i est alors calculée : $\gamma(i,j) = d(T_i,H_j) + min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\}$

Chaque ordre domotique est aligné puis associé à un score d'alignement correspondant au pourcentage de symboles correctement alignés. L'ordre domotique de meilleur score est alors sélectionné pour prendre une décision, avec un seuil de déclenchement. Cette approche prend en compte certaines erreurs de reconnaissance ou de prononciation en se basant sur une proximité phonétique.

5 Expériences et résultats

Dans toutes les expériences, P1 est utilisé pour le développement et l'apprentissage. P2 est utilisé pour l'évaluation des méthodes. Cette section présente le SRAP utilisé et les expériences s'appuyant sur les méthodes décrites.

5.1 Le SRAP Speeral

Le Laboratoire Informatique d'Avignon (LIA) a développé son propre SRAP : Speeral (Linarès *et al.*, 2007). Ce dernier a été utilisé tout au long des travaux présentés ici. Un des principaux avantage de Speeral, est qu'il implémente le DDA. Speeral repose sur un décodeur A^*, des modèles acoustiques MMC (Modèle de Markov Caché) dépendants du contexte et un modèle de langage trigramme. Les vecteurs acoustiques sont composés de 12 coefficients PLP (Perceptual Linear Predictive), de l'énergie ainsi que des dérivées premières et secondes de ces 13 paramètres.

Les modèles acoustiques ont été appris sur 80 heures de parole annotée. Dans le cadre du projet Sweet-Home nous avons utilisé une version 1x le temps réel, qui applique de nombreuses coupures lors du décodage. Les modèles acoustiques ont été adaptés aux 21 locuteurs en utilisant une régression linéaire par maximum de vraisemblance (MLLR) utilisant les données de P1. L'adaptation MLLR représente un bon compromis quant à la quantité de données annotées restreinte.

Un modèle de langage (ML) 3-grammes a été utilisé avec un lexique de 10K mots. Ce modèle de langage est interpolé entre un modèle générique (10%) et un modèle spécialisé (90%). Le ML générique a été estimé sur environ 100M mots extraits du Journal Le Monde et de Gigaword. Le modèle spécialisé a été estimé sur les ordres domotiques ou phrases de détresse attendus.

5.2 Résultats

Les résultats des différentes approches sont présentés dans le tableau 1. La RAP est évaluée via le Taux d'Erreur Mot (TEM), tandis que la détection (classification) des ordres domotiques est évaluée en terme de précision/rappel/F-mesure : le nombre d'ordre domotiques est d'environ

10 et ils sont manuellement annotés pour chaque phrase. Au cours de la détection, si un ordre marqué en tant que tel est détecté : il est considéré comme détecté. Dans tous les autres cas un ordre détecté est considéré comme une fausse alarme. Les résultats sont présentés pour l'ensemble des 21 locuteurs (avec l'écart type associé pour le TEM). Le système de référence est basé sur la sélection de la sortie proposant le meilleurs RSB (parmis les 7 canaux).

Méthode	TEM ± SD	Rappel	Précision	F-mesure
Référence	18.3 ± 12.1	88.0	90.5	89.2
beamforming	16.8 ± 8.3	89.0	92.6	90.8
DDA +RSB	11.4 ± 5.6	93.3	97.3	95.3
DDA 2 lev.+RSB	**8.8** ± 3.7	**95.6**	**98.1**	**96.8**
ROVER	20.6 ± 8.5	85.0	90.0	87.4
ROVER 2c+RSB	13.0 ± 6.6	91.3	95.3	93.3
ROVER +RSB	**12.2** ± 6.1	**92.7**	**97.4**	**95.0**
ROVER Oracle	7.8 ± 2.7	99.4	98.9	99.1

TABLE 1 – TEM et détection des ordres domotiques en fonction des approches

Le système de référence permet d'obtenir 18.3% de TEM (meilleur canal en se basant sur le RSB). Les approches basées sur le RSB présentent une nette amélioration. Le *beamforming* permet un gain relatif de 8.1%. Ce résultat montre que la combinaison des flux au niveau acoustique améliore la robustesse du SRAP. La méthode basée sur le DDA montre un gain relatif de 37.8% en utilisant le RSB. L'approche DDA à deux niveaux présente 52% de gain relatif avec une stabilité très intéressante (écart type de 3.7) : ce gain s'explique facilement par l'introduction de connaissances *a priori* retrouvées au cours du premier décodage. Finalement le ROVER basé sur le RSB permet une amélioration relative de 33.4%.

Dans toutes les configurations, la précision de la reconnaissance des ordres est bonne : le système de référence présente une F-mesure de 89.2%. Nous observons également une corrélation entre le TEM et la tâche de reconnaissance des ordres. Cependant le ROVER et les méthodes basées sur le DDA améliorerent significativement la F-mesure d'environ 7% absolus. Le gain apporté par le *beamforming* n'est pas significatif. Nous notons également que le ROVER permet d'obtenir des résultats similaires à ceux du DDA, mais nécessite un coût de calcul colossal (décodage de tous les canaux). Finalement, la meilleure configuration se base sur le DDA à deux niveaux, qui permet d'atteindre une F-mesure de 96.8%.

6 Conclusion

Nous avons présenté plusieurs approches détectant des ordres vocaux dans le cadre d'un appartement intelligent où les sons sont capturés par un ensemble de micros distants. Les approches se situent à trois niveaux différents du processus de décodage de la parole : l'acoustique, le décodage et la sélection *a posteriori* d'hypothèses. Nous avons également présenté une méthode introduisant directement dans le décodage des connaissances *a priori* telles que le RSB ou des patrons d'ordres prédéfinis.

Les résultats expérimentaux confirment que l'utilisation de tous les micros augmente la qualité du SRAP. Le *beamforming* améliore le WER (16.8%) mais reste comparativement aux autres méthodes proche du système de référence (18.3%). Ceci est sans doute causé par l'éloignement des micros entre eux, n'apportant pas sufisamment de redondance pour améliorer le signal. Le DDA permet d'obtenir les meilleures performances avec un TEM de 11.4% et une F-mesure de 95.3% pour la

classification des ordres vocaux. Les résultats obtenus par le DDA sont très légèrement meilleurs à ceux du ROVER, mais leur coût calculatoire est bien inférieur (décodage de 7 canaux avec le ROVER). Par ailleurs, nous avons proposé un DDA à deux niveaux, introduisant au sein du décodage des connaissances *a priori*. Cette méthode améliore à la fois le TEM (8.8%) et la F-mesure qui devient plus stable que celle du système de référence. Cependant cette amélioration concerne essentiellement les ordres domotiques contenus dans les données de test ; dans le cadre de l'application, seuls ces ordres doivent être reconnus et cela représente un avantage du point de vue de l'acceptation du système par les utilisateurs. Cette étude a également montré que l'utilisation de plusieurs flux permet systématiquement d'améliorer la qualité du décodage, quelle que soit la stratégie. Nous envisageons d'adapter ces méthodes à des conditions plus difficiles (bruitées), en appliquant des techniques de séparation de source, afin de filtrer les bruits issus de la vie courante.

Références

ANGUERA, X., WOOTERS, C. et HERNANDO, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022.

BARKER, J., CHRISTENSEN, H., MA, N., GREEN, P. et VINCENT, E. (2011). The PASCAL 'CHiME' Speech Separation and Recognition Challenge. *In InterSpeech 2011*. (to appear).

FISCUS, J. G. (1997). A post-processing system to yield reduced word error rates : Recognizer Output Voting Error Reduction (ROVER). *In Proc. IEEE Workshop ASRU*, pages 347–354.

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

LECOUTEUX, B., LINARÈS, G., BONASTRE, J. et NOCÉRA, P. (2006). Imperfect transcript driven speech recognition. *In Proc. InterSpeech'06*, pages 1626–1629.

LECOUTEUX, B., VACHER, M. et PORTET, F. (2011). Distant speech recognition in a smart home : Comparison of several multisource asrs in realistic conditions. *In Interspeech 2011*, pages 2273–2276.

LINARÈS, G., NOCÉRA, P., MASSONIÉ, D. et MATROUF, D. (2007). The LIA speech recognition system : from 10xRT to 1xRT. *In Proc. TSD'07*, pages 302–308.

VACHER, M., FLEURY, A., SERIGNAT, J.-F., NOURY, N. et GLASSON, H. (2008). Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment. *In Proc. InterSpeech 2008*, pages 496–499.

VACHER, M., PORTET, F., FLEURY, A. et 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.

VIPPERLA, R. C., WOLTERS, M., GEORGILA, K. et RENALS, S. (2009). Speech input from older users in smart environments : Challenges and perspectives. *In HCI International : Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*.

VOVOS, A., KLADIS, B. et FAKOTAKIS, N. (2005). Speech operated smart-home control system for users with special needs. *In Proc. InterSpeech 2005*, pages 193–196.

WÖLFEL, M. et MCDONOUGH, J. (2009). *Distant Speech Recognition*. Published by Wiley.