Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole (Contribution to the study of elderly people's voice variability in automatic speech recognition) [in French]

Frédéric Aman, Michel Vacher, Solange Rossato, François Portet

Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole

Frédéric Aman, Michel Vacher, Solange Rossato, François Portet
Laboratoire d'Informatique de Grenoble (UMR 5217), équipe GETALP
41 avenue des Mathématiques,
BP 53 - 38041 Grenoble Cedex 9 - France
{frederic.aman, michel.vacher, solange.rossato, francois.portet}@imag.fr

RÉSUMÉ

L'utilisation de la reconnaissance vocale pour l'assistance à la vie autonome se heurte à la difficulté d'utilisation des systèmes de RAP qui ne sont pas prévus à la base pour la voix âgée. Pour caractériser les différences de comportement d'un système de reconnaissance entre les personnes âgées et non-âgées, nous avons étudié quels sont les phonèmes les moins bien reconnus en nous basant sur le corpus AD80 que nous avons enregistré. Les résultats montrent que certains phonèmes tels que les plosives sont plus spécifiquement affectés par l'âge. De plus nous avons recueilli le corpus spécifique ERES38 afin d'adapter les modèles acoustiques, avec pour résultat une diminution du taux d'erreur de mot de 15%. Malgré la grande variabilité des performances, nous avons caractérisé comment la baisse des performances du système de reconnaissance automatique de la parole peut être corrélée avec la baisse d'autonomie des personnes âgées.

ABSTRACT

Contribution to the study of elderly people's voice variability in automatic speech recognition

Using speech recognition to support ambient assisted living is impeded by the difficulty of using ASR systems that are not provided for the elderly voice. To characterize these differences in speech recognition performance, we studied phoneme categories which lead to the lowest recognition rate in the elderly speakers with respect to the younger ones based on the AD80 corpus that we recorded. The results showed that some phonemes (such as plosives) are more specifically affected by age than others. Moreover, we collected the specific ERES38 corpus to adapt the ASR acoustic model to the elderly population which resulted in a 15% decrease of the word error rate. Despite a great variability of performances, we characterized how lower performance of ASR systems can be correlated to the autonomy degradation of elderly people.

MOTS-CLÉS : reconnaissance automatique de parole, voix des personnes âgées, adaptation acoustique, assistance à la vie autonome.

KEYWORDS: automatic speech recognition, ageing voice, acoustic adaptation, ambient assisted living.

JEP-TALN-RECITAL 2012, Atelier ILADI 2012: Interactions Langagières pour personnes Agées Dans les habitats Intelligents, pages 49–59,
Grenoble, 4 au 8 juin 2012.

1 Introduction

L'assistance à la vie autonome (ou *Ambient Assisted Living - AAL*) est devenu un enjeu important étant donné la part croissante de la population âgée dans les pays industrialisés. Par contre, il reste encore à l'heure actuelle assez peu de projets qui prennent en compte une interaction vocale de la part de l'utilisateur, par exemple une commande vocale en domotique. Les personnes qui bénéficieraient le plus de ces technologies seraient les personnes en perte d'autonomie, étant donné que le langage naturel est le moyen le plus spontané de communication.

Dans ce contexte, le projet CIRDO [1] auquel participe le LIG vise à favoriser l'autonomie et la prise en charge des personnes âgées par les aidants au travers d'*e-lio*, un produit de télélien social augmenté et automatisé. *e-lio* est un système de communication en visiophonie s'adaptant au degré d'autonomie de son utilisateur. *e-lio* permet l'accès à de nombreux services interactifs : visiophonie, téléphonie, messages, partage de photos, rappels automatiques, appels d'urgence, agenda partagé, plateforme domotique, entraînement de la mémoire et de l'attention, etc. L'objectif du projet CIRDO est d'y intégrer un système de Reconnaissance Automatique de la Parole (RAP) qui inclura une détection des signaux de détresse ainsi que des commandes vocales en complément de la télécommande.

L'utilisation de la reconnaissance vocale pour l'assistance à la vie autonome se heurte à la difficulté d'utilisation des systèmes de RAP qui ne sont pas prévus à la base pour la voix âgée. En effet, du fait de certaines caractéristiques spécifiques de la voix âgée, un travail d'adaptation des systèmes de RAP a dû être réalisé. De fait, la parole âgée se caractérise notamment par des tremblements de la voix, une production imprécise des consonnes, et une articulation plus lente (Ryan et Burk, 1974). Du point de vue anatomique, des études ont montré des dégénérescences liées à l'âge avec une atrophie des cordes vocales, une calcification des cartilages du larynx, et des changements dans la musculature du larynx (Takeda *et al.*, 2000; Mueller *et al.*, 1984). De plus, des changements dans les capacités de contrôle moteur de la voix au niveau cognitif modifient la production de la parole tout au long de la vie (Hooper et Cralidis, 2009). D'autres études (Georgila *et al.*, 2008) ont montré que lors de l'interaction avec un système de dialogue - incluant une RAP et une synthèse vocale - les personnes âgées utilisent, par rapport aux personnes jeunes, un vocabulaire plus riche et des phrases plus longues, et emploient plus fréquement des expressions d'interaction sociale telles que "au revoir" ou "merci", comme s'il s'agissait d'une interaction humain/humain. Du fait que les modèles acoustiques des systèmes de RAP sont appris majoritairement sur de la voix non-âgée, on observe donc une augmentation significative du taux d'erreurs de mots pour la voix des personnes âgées par rapport à la voix des adultes non-âgés (Baba *et al.*, 2004; Vipperla *et al.*, 2008, 2010; Aman *et al.*, 2012).

Afin d'améliorer le module de décodage acoustico-phonétique dans un système de RAP et de l'adapter à la voix des personnes âgées, une première analyse a consisté à étudier les phonèmes qui étaient mal reconnus pour les personnes âgées. Cette analyse, présentée dans la section 2, a permis d'extraire les phonèmes qui semblent plus problématiques à reconnaître que d'autres lors du décodage acoustico-phonétique. Nous avons réalisé une adaptation du modèle acoustique, détaillée en section 3. Nous montrons en section 4 que l'âge n'est pas le facteur déterminant sur la baisse des performances du système de RAP, mais que le niveau de dépendance de la personne âgée joue un rôle important. Nous concluons et présentons les perspectives de recherche en section 5.

1. `http://liris.cnrs.fr/cirdo/`

2 Détermination des phonèmes difficiles à reconnaître

2.1 Le corpus de test AD80

Afin d'évaluer le comportement du système de RAP sur la voix âgée, nous avons utilisé le corpus *Anondin-Détresse 80* (*AD80*). Ce corpus, enregistré par le laboratoire LIG, est spécifique au domaine de la domotique et à la détection d'appels de détresse. Ce corpus est constitué de l'enregistrement de 57 locuteurs (22 hommes et 35 femmes) âgés de 20 à 94 ans. Il a été demandé aux participants de lire une liste de 126 énoncés courts de la vie quotidienne ou caractéristiques d'un appel de détresse (ex : "Il fait chaud" ou "Aidez-moi"). Dans le cadre de l'application envisagée, nous cherchons essentiellement à reconnaître ce type d'énoncés correspondant à des ordres domotiques ou des appels à l'aide.

Le corpus *AD80* est constitué de deux groupes :
- Le groupe *voix non-âgées*, constitué de 21 locuteurs âgés de 20 à 65 ans enregistrés à Grenoble dans notre laboratoire en 2004. Tous les locuteurs étaient actifs professionnellement ou étudiants. Cette première partie a été enregistrée lors d'études relatives à la reconnaissance d'appels de détresse dans un Habitat Intelligent pour la Santé (HIS) (Vacher *et al.*, 2006), ce qui explique pourquoi il s'agit d'énoncés courts. Le texte de ces énoncés a ensuite été utilisé pour des évaluations en milieu réel en parole distante dans un appartement équipé de microphones (Vacher *et al.*, 2008).
- Le groupe *voix âgées*, constitué de 36 locuteurs âgés de 62 à 94 ans enregistrés dans un hôpital et à domicile à Grenoble en 2010 ainsi que dans un centre de rétablissement et dans une maison de retraite dans le département du Gard en 2012. Les locuteurs étaient à la retraite, et pour certains en situation de dépendance.

Au final, le corpus *AD80* est formé de 6 848 phrases annotées, avec 2 heures et 3 minutes d'enregistrements audio (cf. Table 1).

Corpus AD80	Nombre locuteurs	Age min-max	Durée	Nombre phrases
Groupe voix non-âgées	21	20-65	38min	2646
Groupe voix âgées	36	62-94	1h25min	4202
Total	57	20-94	2h03min	6848

TABLE 1 – Caractéristiques du corpus AD80

2.2 Le système de RAP

Le système de RAP choisi pour notre étude est Sphinx3 (Seymore *et al.*, 1998). Ce décodeur utilise un modèle acoustique dépendant du contexte avec chaînes de Markov cachées 3 états. Les vecteurs acoustiques sont composés de 13 coefficients MFCC, le delta et le double delta de chaque coefficient. Le modèle acoustique que nous utilisons a été entraîné sur le corpus *BREF120* (Lamel *et al.*, 1991) qui est composé de 100 heures de parole annotées enregistrées auprès de 120 locuteurs français. Nous avons appelé ce modèle le *modèle acoustique générique*.

Un modèle de langage spécialisé a été utilisé. Ce modèle a été construit à partir des transcriptions

des phrases du corpus *AD80*, dont il a résulté un modèle de langage restreint, de type trigramme, avec un vocabulaire d'environ 160 mots.

2.3 Comparaison des taux d'erreurs de mots et des scores d'alignements forcés entre les groupes voix âgées et voix non-âgées

Afin d'évaluer l'effet de la voix âgée sur la performance de la RAP en utilisant le *modèle acoustique générique*, nous avons comparé le taux d'erreurs de mots (ou *Word Error Rate - WER*) entre les groupes. Nous avons obtenu un WER de 7,33% pour le décodage sur le groupe *voix non-âgées*, et un WER de 27,56% pour le décodage sur le groupe *voix âgées*. Ainsi, nous avons observé une importante dégradation du système de RAP pour la voix des personnes âgées, avec une différence absolue de 20,23%.

Pour aller plus loin dans l'analyse, nous avons réalisé un alignement forcé sur ces deux groupes. L'alignement forcé consiste à convertir les transcriptions de référence en suites de phonèmes calés sur les données audio en utilisant un dictionnaire phonétique. L'alignement forcé a permis d'obtenir les scores d'alignement par phonème. Ceux-ci sont des scores de vraisemblance d'appartenance au phonème normalement prononcé pour la portion de signal considérée. Ce score peut être interprété comme une proximité avec la prononciation "standard", modélisée par le *modèle acoustique générique*. Le score exprime le logarithme d'une vraisemblance, il est inférieur ou égal à zéro, et plus il est faible, plus le phonème associé est éloigné du modèle acoustique.

Les scores sont présentés sur la Figure 1 selon la catégorie phonémique.

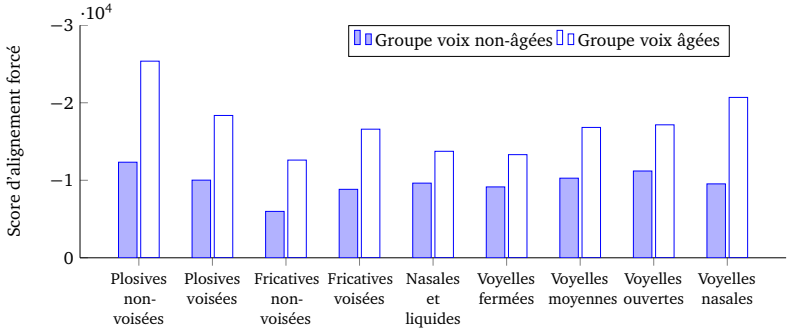

FIGURE 1 – Score d'alignement forcé par catégorie phonémique avec le *modèle acoustique générique* pour les groupes *voix non-âgées* et *voix âgées*

Pour le groupe *voix non-âgées*, certains phonèmes montrent des valeurs plus faibles du score d'alignement, tels que les plosives ou les voyelles ouvertes. D'autres sons, à l'inverse, sont plus proches des représentations de modèle acoustique : les fricatives.

Pour le groupe *voix âgées*, les scores d'alignement sont moins élevés que ceux obtenus pour le groupe *voix non-âgées*. En effet, le vieillissement provoque une moins bonne maîtrise du système

articulatoire, et la prononciation des personnes âgées s'éloigne donc du modèle acoustique. Les consonnes plosives et les voyelles nasales sont les phonèmes nécessitant le plus de contrôle moteur et sont les plus difficiles à articuler. Ceci se traduit au niveau des scores d'alignement, les plus bas sont obtenus pour les consonnes plosives et les voyelles nasales.

Les écarts de scores les plus importants par catégorie phonémique sur le groupe *voix âgées* par rapport au groupe *voix non-âgées* ont permis de caractériser quels sont les phonèmes posant le plus de problèmes pour la RAP des voix âgées. Les différences relatives de scores observées entre les deux groupes ont été calculées. Les catégories phonémiques sont par ordre descendant de différence : les voyelles nasales (-117,00%), les consonnes fricatives non-voisées (-110,56%), les consonnes plosives non-voisées (-105,72%), les consonnes fricatives voisées (-87,86%), les consonnes plosives voisées (-83,29%), les voyelles moyennes (-63,74%), les voyelles ouvertes (-53,21%), les voyelles fermées (-45,52%), et les consonnes nasales et liquides (-42,65%). En se basant sur les différences relatives, les catégories phonémiques les plus affectées pour la parole âgée sont les voyelles nasales, et les consonnes fricatives et plosives. Aussi, nous pouvons noter que globalement les consonnes sont plus affectées que les voyelles. De plus, l'absence de voisement est le principal facteur de dégradation, suivie par la modalité de réalisation (plosives et fricatives). Ainsi, il serait possible que, en ce qui concerne les personnes âgées, les consonnes non voisées soient plus proches des consonnes voisées.

Ces résultats sont similaires à ceux obtenus par (Privat *et al.*, 2004) qui ont trouvé une dégradation de la RAP entre voix âgée et voix non-âgée avec une différence relative très proche de celle que nous avons obtenue pour chaque catégorie phonémique, excepté pour les consonnes nasales et liquides et pour les voyelles nasales où leur système était moins performant que le notre.

3 Adaptation acoustique

3.1 Recueil du nouveau corpus ERES38

Étant donnée la baisse de performance du système de RAP pour la voix âgées, nous avons enregistré un nouveau corpus de parole de personnes âgées en vue de l'amélioration du modèle acoustique grâce à une méthode d'adaptation acoustique. Les principales étapes de l'étude sont résumées Figure 2.

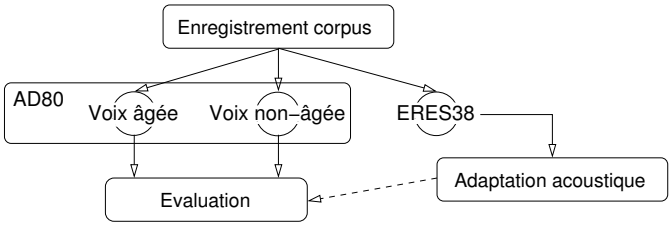

FIGURE 2 – Principales étapes de l'étude

Le corpus constitué est un ensemble d'entretiens. Chaque entrevue met en relation une per-

sonne âgée avec deux expérimentateurs dont l'un se fait l'interlocuteur privilégié. Le matériel utilisé pour les enregistrements était un enregistreur TASCAM DR-100 avec un microphone à condensateur unidirectionnel placé à proximité et en direction de la personne âgée. Une première partie introductive permet de récupérer les informations personnelles ainsi que les habitudes linguistiques du locuteur. Cette phase d'habituation avec le matériel d'enregistrement permet d'établir le passage vers une parole un peu plus informelle et spontanée pour recueillir le récit de vie de la personne, incluant une description des activités quotidiennes et de leur habitat, un récit d'accidents éventuels et des anecdotes. Une activité de lecture est également proposée lors de cet entretien. Le support choisi est un article de jardinage créé par les expérimentateurs dans le but de cibler les phonèmes problématiques. Les plosives et fricatives non voisées ont été introduites de façon à se retrouver en contexte /a/, /i/ et /u/.

Le corpus est constitué de 17 heures et 44 minutes d'enregistrements (cf. Table 2) avec 24 locuteurs (16 femmes et 8 hommes) dont l'âge varie de 68 à 98 ans, incluant 48 minutes de lectures par 22 locuteurs. Ces locuteurs sont issus de structures spécifiques pour personnes âgées, foyers logements ou maisons de retraite. Les entretiens ont été effectués avec des personnes plus ou moins autonomes, sans déficience cognitive, parfois avec de sérieuses difficultés motrices, mais sans handicap lourd.

Corpus ERES38	Nombre locuteurs	Age min-max	Durée	Nombre phrases
Lecture de texte	22	68-98	16h56min	300
Parole spontanée	24	68-98	48min	7300
Total	24	68-98	17h44min	7600

TABLE 2 – Caractéristiques du corpus ERES38

Les enregistrements des entretiens ont commencé à être transcrits, et toutes les lectures ont été transcrites et vérifiées. Ces données annotées et structurées constituent le corpus *Entretiens RESidences 38* (*ERES38*).

3.2 Adaptation MLLR

La méthode d'adaptation de régression linéaire du maximum de vraisemblance (*Maximum Likelihood Linear Regression* ou *MLLR*) a été utilisée pour adapter le *modèle acoustique générique*, appris sur *BREF120*, à la voix des personnes âgées. L'adaptation a été faite globalement avec l'ensemble des lectures du corpus *ERES38*. Nous avons ainsi obtenu un nouveau modèle acoustique appelé *modèle acoustique adapté par MLLR*.

A partir du corpus *AD80*, un premier décodage a été fait sur le groupe *voix âgées* en utilisant le *modèle acoustique générique*. Puis un second décodage a été effectué sur ce même groupe avec le *modèle acoustique adapté par MLLR*. Le but était de voir dans quelle mesure est la différence de WER avec l'utilisation de l'un ou l'autre des modèles, avec l'hypothèse que le WER avec le groupe *voix âgées* issu du décodage avec le *modèle acoustique adapté par MLLR* serait proche du WER avec le groupe *voix non-âgées* issu du décodage avec le *modèle acoustique générique*. En outre, nous avons réalisé un décodage avec le *modèle acoustique adapté par MLLR* sur le groupe *voix non-âgées* afin de tester la spécificité de l'adaptation.

La Figure 3 montre que l'utilisation du *modèle acoustique adapté par MLLR* a permis de réduire le WER pour tous les locuteurs du groupe *voix âgées*. Avec l'adaptation MLLR globale, le WER est de 11,95%. Comparé au WER de 27,56% sans adaptation, la différence absolue est de -15,61% (différence relative de -56,65%). D'un point de vue applicatif, cela montre que l'on peut utiliser une base de parole âgée pour l'adaptation MLLR dont les locuteurs sont différents de ceux de la base de test. Cela démontre que les voix des personnes âgées ont des caractéristiques propres communes. De plus, nous voyons que l'utilisation d'un corpus de petite taille (48 minutes de lecture par 22 locuteurs du corpus *ERES38*) pour l'adaptation MLLR globale est suffisante pour donner une amélioration significative avec un WER de 11,95%, proche du WER de 7,33% trouvé dans le cas du décodage de la parole non-âgée.

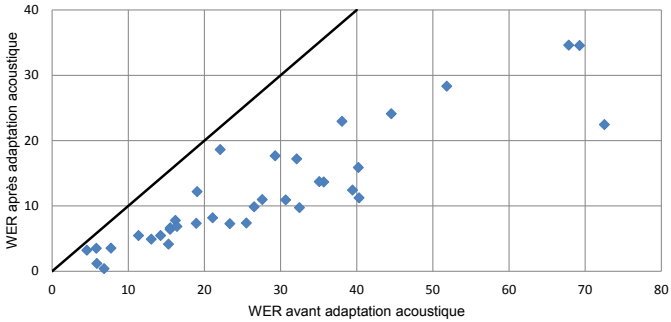

FIGURE 3 – WER avec adaptation acoustique MLLR en fonction du WER sans adaptation. La droite représente la fonction identité

De plus, nous avons obtenu un WER de 10,39% lors du décodage sur le groupe *voix non-âgées* avec le *modèle acoustique adapté par MLLR*. Cela représente une dégradation relative de 43,75% comparée au WER de 7,33% pour le décodage avec le *modèle acoustique générique* sur le même groupe. Il est donc convenu que le *modèle acoustique adapté par MLLR* est spécifique à la RAP sur la population âgée.

4 Relation entre l'autonomie des personnes âgées et la RAP

Le WER issu du décodage sur le groupe *voix âgées* avec le *modèle acoustique adapté par MLLR* est représenté en fonction de l'âge sur la Figure 4. Nous observons par la dispersion des points des locuteurs représentés que l'âge est un mauvais indicateur du WER. De plus, nous avons calculé la corrélation de Pearson entre les variables "WER après adaptation acoustique" et "Age". Nous avons trouvé un score de corrélation de -0,053 ($p = 0{,}759\%$) prouvant que le WER et l'âge ne sont pas corrélés. En conséquence, nous avons cherché si d'autres paramètres relatifs à la dépendance peuvent être des indicateurs de la performance du système de RAP.

Nous avons fait l'hypothèse que la dégradation physique et psychique affecte la production de la parole et donc les performances de la RAP. Nous avons pris pour référence un test national relatif à l'autonomie des personnes âgées : la grille AGGIR (Autonomie Gérontologie Groupes

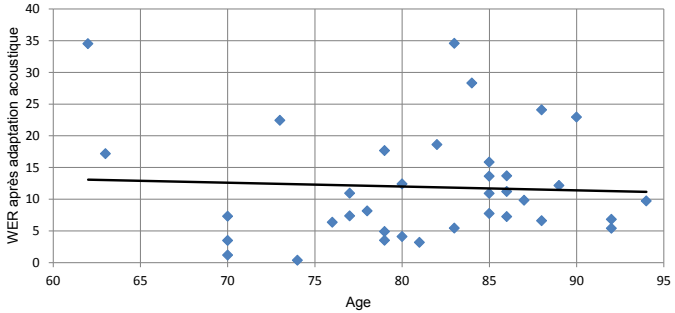

FIGURE 4 – WER après adaptation acoustique MLLR en fonction de l'âge, avec la droite correspondant à la régression linéaire

Iso-Ressources) [2]. Pour 29 locuteurs âgés en établissement que nous avions enregistrés pour la constitution du corpus *AD80*, nous avons complété pour chacun d'entre eux une grille AGGIR avec l'aide du personnel soignant.

La grille AGGIR est un outil d'évaluation du degré de perte d'autonomie et de dépendance, en terme de dégradation physique et psychique, pour l'attribution de l'Allocation Personnalisée d'Autonomie (APA), qui est une aide financière pour les personnes âgées dépendantes en France. L'évaluation est faite à partir de 17 variables. Dix variables se réfèrent à la perte d'autonomie physique et psychique : cohérence, orientation, toilette, habillage, alimentation, élimination, transferts (se lever, se coucher, s'asseoir), déplacement à l'intérieur, déplacement à l'extérieur, et communication à distance. Sept variables se rapportent à la perte d'autonomie domestique et sociale : gestion personnelle de son budget et de ses biens, cuisine, ménage, transports, achats, suivi du traitement, et activités de temps libre. Chaque variable est codée par A (fait seul), B (fait partiellement) ou C (ne fait pas). Le score GIR (Groupe Iso-Ressources) est calculé à partir des variables afin de classer les personnes âgées dans un des six groupes : de GIR 1 (dépendance totale) à GIR 6 (autonomie totale). Les personnes classées de GIR 1 à GIR 4 sont autorisées à recevoir une aide financière selon leur degré de dépendance.

Nous avons regardé si le score GIR est représentatif de la performance de la RAP. Le WER des 29 participants testés sont représentés en fonction de leur score GIR en Figure 5. Quatre locuteurs sont GIR 2, deux locuteurs sont GIR 3, quinze locuteurs sont GIR 4 et huit locuteurs sont GIR 6. Aucun locuteur n'est représenté en GIR 1 et GIR 5. Nous observons en Figure 5 que le score GIR pourrait avoir une influence sur le WER, avec une baisse du WER en fonction de l'augmentation du score GIR, sauf pour GIR 2.

Du fait du faible nombre de locuteurs en GIR 2 et GIR 3, nous avons rassemblé ces deux groupes en un groupe nommé GIR 2-3. Puis nous avons avons réalisé une ANOVA sur GIR 2-3, GIR 4 et GIR 6 pour vérifier si le score GIR pourrait avoir un effet significatif sur le WER. Nous avons trouvé les résultats suivants : $F(DDL, DDL_{error}) = F(2, 26) = 3{,}7$; $p < 0.05\%$, prouvant qu'il y a au moins une distribution dont la moyenne diffère des autres moyennes. Nous avons réalisé un

2. http ://vosdroits.service-public.fr/F1229.xhtml

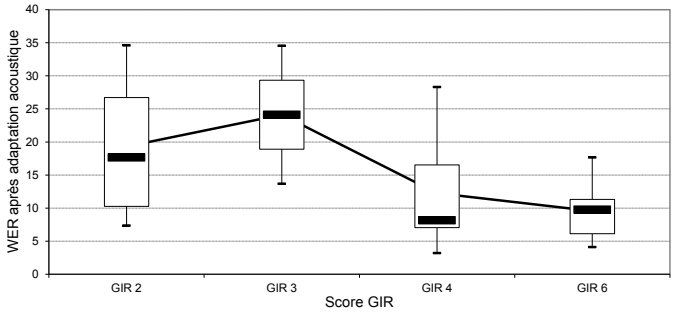

FIGURE 5 – WER avec adaptation acoustique MLLR en fonction du score GIR

test post-hoc de Bonferroni afin de caractériser quels groupes sont significativement différents de quels autres groupes. Le test post-hoc a révélé que les groupes extrêmes GIR 2-3 et GIR 6 ont un effet significativement différent sur le WER, et que GIR 4 n'a pas une influence significativement différente sur le WER par rapport aux autres groupes.

De plus, nous avons réalisé une étude préliminaire (que nous détaillerons dans un prochain article) sur les corrélations entre le WER et chacun des 17 paramètres de la grille AGGIR. Il semblerait que les paramètres concernant le contrôle moteur des membres supérieurs et la continence pourrait être les plus corrélés au WER. En effet, ces paramètres pourraient être représentatifs d'une dégradation physique généralisée avancée affectant également le contrôle de la voix, et donc diminueraient la performance du système de RAP. Une perspective pourrait être de permettre une prédiction du WER en se basant sur les caractéristiques de dépendance des personnes âgées.

5 Conclusion

Cet article présente notre étude sur le comportement d'un système de RAP vis-à-vis de la voix âgée. Étant donné l'absence de corpus contenant de la voix de personnes âgées de langue française utilisable pour la création ou l'adaptation des modèles, nous avons procédé à l'enregistrement du corpus *AD80*. A partir de ce corpus, nous avons analysé quels étaient les phonèmes pour la voix âgée posant le plus problème au système de RAP. Nous avons pu déterminer que leur éloignement par rapport à la prononciation modélisée par les modèles acoustiques provoque une augmentation du WER du système de RAP, avec une différence absolue entre voix non-âgée et âgée de 20,23%. Ensuite, nous avons procédé à l'adaptation du *modèle acoustique générique* à la voix des personnes âgées, grâce à la méthode d'adaptation MLLR, à partir du corpus *ERES38*. Le cas de l'adaptation MLLR globale est intéressante car avec moins d'une heure d'enregistrements, à partir de locuteurs différents des locuteurs de test, nous avons obtenu des taux d'erreurs de mots proches du cas d'une reconnaissance avec le *modèle acoustique générique* de parole non-âgée, avec un WER de 11,95%, contre 27,56% avant adaptation. De plus, nous avons montré que le WER n'est pas corrélé avec l'âge mais pourrait être corrélé avec le niveau de dépendance de la

personne âgée du fait d'une dégradation physique générale. La continuation de notre travail sera de réaliser une adaptation au locuteur et de montrer comment les différents paramètres de la grille AGGIR sont corrélés au WER. La prédiction du comportement du système de RAP permettra de faciliter l'utilisation de ces nouvelles technologies dans la vie quotidienne des personnes âgées dépendantes. Aussi, nous allons étudier dans quelle mesure les sons non verbaux (inspirations, bruits de bouche) ainsi que les hésitations et les défauts d'articulation sont plus fréquents chez les personnes âgées, et une prochaine étape de notre travail sera de prendre en considération ces phénomènes pour la construction des modèles acoustiques et de langage.

Remerciements

Cette étude a été financée par l'Agence Nationale de la Recherche dans le cadre du projet CIRDO-Recherche Industrielle (ANR-2010-TECS-012). Nous remercions particulièrement Remus Dugheanu, Juline le Grand, Yuko Sasa, Claude Aynaud et Quentin Lefol pour leur active contribution, ainsi que les différentes personnes âgées et le personnel soignant qui ont accepté de participer aux enregistrements.

Références

Aman, F., Vacher, M., Rossato, S., Dugheanu, R., Portet, F., le Grand, J. et Sasa, Y. (2012). Étude de la performance des modèles acoustiques pour des voix de personnes âgées en vue de l'adaptation des systèmes de RAP. *In JEP*, Grenoble, France.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A. et Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2*, 87:49–57.

Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., Moore, J. et Watson, M. (2008). A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. *In Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Hooper, C. et Cralidis, A. (2009). Normal changes in the speech of older adults : You've still got what it takes ; it just takes a little longer ! *Perspectives on Gerontology*, 14:47–56.

Lamel, L., Gauvain, J. et EskEnazi, M. (1991). BREF, a large vocabulary spoken corpus for french. *In Proceedings of EUROSPEECH 91*, volume 2, pages 505–508, Geneva, Switzerland.

Mueller, P., Sweeney, R. et Baribeau, L. (1984). Acoustic and morphologic study of the senescent voice. *Ear, Nose, and Throat Journal*, 63:71–75.

Privat, R., Vigouroux, N. et Truillet, P. (2004). Etude de l'effet du vieillissement sur les productions langagières et sur les performances en reconnaissance automatique de la parole. *Revue Parole*, 31-32:281–318.

Ryan, W. et Burk, K. (1974). Perceptual and acoustic correlates in the speech of males. *Journal of Communication Disorders*, 7:181–192.

Seymore, K., Stanley, C., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R. et Thayer, E. (1998). The 1997 CMU Sphinx-3 English

broadcast news transcription system. *In DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA.

TAKEDA, N., THOMAS, G. et LUDLOW, C. (2000). Aging effects on motor units in the human thyroarytenoid muscle. *Laryngoscope*, 110:1018–1025.

VACHER, M., FLEURY, A., SERIGNAT, J., NOURY, N. et GLASSON, H. (2008). Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment. *In 9th International Conference on Speech Science and Speech Technology (InterSpeech 2008)*, volume 1, pages 496–499, Brisbane Convention & Exhibition Centre (BCEC), Brisbane (Australia). Australasian Speech Science and Technology Association (ASSTA).

VACHER, M., SERIGNAT, J.-F., CHAILLOL, S., ISTRATE, D. et POPESCU, V. (2006). Speech and Sound Use in Remote Monitoring System for Health Care. *In* Faculty of INFORMATICS, M. U., éditeur : *9th International Conference on Text, Speech and Dialogue (TSD 2006)*, volume 4148 de *LNCS - LNAI*, pages 711–718, Faculty of Informatics, Masaryk University, Brno (Czech Republic).

VIPPERLA, R., RENALS, S. et FRANKEL, J. (2008). Longitudinal study of ASR performance on ageing voices. *Interspeech*, pages 2550–2553.

VIPPERLA, R., RENALS, S. et FRANKEL, J. (2010). Ageing voices : The Effect of Changes in Voice Parameters on ASR Performance. *EURASIP Journal on Audio, Speech, and Music Processing*.