Étude de la performance des modèles acoustiques pour des voix de personnes âgées en vue de l'adaptation des systèmes de RAP

Frédéric Aman, Michel Vacher, Solange Rossato, Remus Dugheanu,
François Portet, Juline le Grand, Yuko Sasa
Laboratoire d'Informatique de Grenoble (UMR 5217), équipe GETALP
41 avenue des Mathématiques,
BP 53 - 38041 Grenoble Cedex 9 - France
Frederic.Aman@imag.fr, Michel.Vacher@imag.fr, Solange.Rossato@imag.fr,
Francois.Portet@imag.fr

RÉSUMÉ

Notre étude s'inscrit dans le cadre de l'intégration d'un système de reconnaissance de la parole pour un produit de télélien social pour personnes âgées. Du fait de l'évolution des caractéristiques acoustiques de la voix en fonction de l'âge, les taux d'erreurs de mots des systèmes de reconnaissance automatique de la parole sont plus élevés lors du décodage de parole pour des personnes âgées que non-âgées. Notre étude consiste à caractériser les différences de comportement d'un système de reconnaissance pour les personnes âgées et non-âgées, définir les phonèmes les moins bien reconnus, et recueillir un corpus spécifique pour permettre l'adaptation des modèles acoustiques à la voix âgée. Les résultats montrent que certains phonèmes tels que les plosives sont plus spécifiquement affectés par l'âge, et que le recueil des données ciblées permet de procéder à une adaptation à la voix âgée qui diminue de 5% le taux d'erreurs de mots.

ABSTRACT

Assessment of the acoustic models performance in the ageing voice case for ASR system adaptation

Our study concerns the integration of an automatic speech recognition system in a social inclusion product designed for elderly people. Due to voice change with age, speech recognition systems present higher word error rate when speech is uttered by elderly speakers compared to when non-aged voice is considered. To characterise these differences in speech recognition performance, we studied which phonemes lead to the lowest recognition rate in the elderly speakers with respect to the younger ones and we collected a specific corpus to make the adaptation of the acoustic models possible. The results show that some phonemes (such as plosives) are more specifically affected by age than others. Finally, the corpus was used to adapt the ASR to the elderly population which resulted in a 5% decrease of the word error rate.

MOTS-CLÉS : reconnaissance automatique de parole, voix des personnes âgées, adaptation acoustique, régression linéaire du maximum de vraisemblance.

KEYWORDS: automatic speech recognition, ageing voice, acoustic adaptation, maximum likelihood linear regression.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 707–714,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Grâce aux progrès de la médecine, l'espérance de vie s'est allongée. Cependant, ce phénomène couplé à une baisse de la natalité a conduit à un vieillissement de la population. Pour aider les personnes âgées à vivre le plus longtemps possible à domicile, des solutions ont été développées en s'appuyant sur la robotique, la domotique, les sciences cognitives et les réseaux informatiques. Ces solutions permettent de compenser leurs pertes physiques et mentales afin de conserver leur autonomie. Le but est aussi de leur fournir si nécessaire une aide grâce à une surveillance permettant la détection des situations de détresse et des chutes. Un tel système doit permettre l'indépendance de la personne âgée tout en facilitant le contact social, avec un impact majeur sur son bien-être et sa santé. De plus, il aide les soignants et permet de rassurer les proches. Cependant, les solutions technologiques doivent s'adapter aux besoins et capacités spécifiques de cette catégorie de la population. En effet, les personnes âgées sont souvent désarmées devant les interfaces complexes. C'est pourquoi, les interfaces habituelles (télécommandes, souris, claviers) doivent être complétées par des interfaces plus accessibles et naturelles, telles qu'un système de Reconnaissance Automatique de la Parole (RAP).

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 à travers un produit de télélien social augmenté et automatisé. L'objectif de ce projet est d'y intégrer un système de RAP incluant une détection des signaux de détresse et des commandes vocales.

Du fait de certaines caractéristiques spécifiques de la voix âgée, un travail d'adaptation des systèmes de RAP a dû être réalisé. En effet, 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). Étant donné que les modèles acoustiques de systèmes de RAP sont appris majoritairement sur de la voix non-âgée, ils ne sont pas adaptés à la voix de la population âgée, ce qui se traduit par une baisse des performances des systèmes de RAP classiques (Baba *et al.*, 2004; Vipperla *et al.*, 2008).

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. Un protocole de recueil de corpus a été mis en place pour enregistrer des personnes âgées, décrit en section 3. Ces données ont été annotées et ont été utilisées pour adapter le modèle acoustique tel que détaillé en section 4. 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 Les corpus de test Anodin-Détresse et Voice-Age

Deux corpus ont été utilisés pour l'évaluation du système de RAP.

Le corpus *Anodin-Détresse* (*AD*) a été enregistré au laboratoire CLIPS de Grenoble. Il fut constitué en 2004 pour l'évaluation d'un système de RAP pour une application de télémédecine en environnement réel avec détection d'appels de détresse (Vacher *et al.*, 2008). Ce corpus a été enregistré auprès de 21 locuteurs (11 hommes et 10 femmes) âgés de 20 à 65 ans. Il est constitué de 126 phrases courtes de la vie quotidienne et de détresse qui ont été lues par chaque participant, soit un total de 2 646 phrases audio annotées pour une durée de 38 minutes.

Le corpus *Voice-Age* (*VA*) est un corpus de voix âgées enregistré en 2010 par le laboratoire LIG en vue d'une exploration préliminaire de la RAP adaptée à la voix des personnes âgées, en français. Du fait des difficultés rencontrées lors de la constitution d'un tel corpus, le nombre de locuteurs de *VA* est restreint, soit sept locuteurs (3 hommes/4 femmes) âgés de 70 à 89 ans (âge moyen de 77 ans). Deux locuteurs ont été enregistrés dans le service de gérontologie du CHU de Grenoble, et cinq locuteurs à leur domicile. Le corpus *VA* est constitué de phrases longues extraites de journaux ou magazines, et des mêmes phrases courtes que le corpus *AD*. Au total, 5 441 phrases ont été prononcées, soit une durée de 4 heures et 8 minutes d'enregistrement.

Nous avons constitué deux groupes d'étude à partir de ces corpus : le groupe *voix non-âgées* contient les lectures des 21 locuteurs de *AD*, et le groupe *voix âgées* contient les lectures des 7 locuteurs de *VA*. Seules les phrases communes aux deux corpus *AD* et *VA* portant sur la vie quotidienne et la détresse ont été utilisées dans ces groupes, soit 2646 phrases (38 minutes) pour le groupe *voix non-âgées*, et 591 phrases (14 minutes) pour groupe *voix âgées*.

2.2 Le système de RAP

Afin de comparer l'influence des groupes *voix âgées* et *voix non-âgées* sur les systèmes de RAP, nous avons procédé à un décodage sur chaque groupe. Le moteur de RAP employé pour le décodage 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. Ce modèle acoustique 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*.

Le modèle de langage et le lexique choisis sont de type spécialisé, pour répondre au contexte de commandes vocales domotiques. Le modèle de langage a été entraîné avec les transcriptions des phrases des groupes *voix non-âgées* et *voix âgées*. Le résultat est un modèle de langage très restreint, de type trigramme, sur un vocabulaire d'environ 160 mots. Ce modèle de langage très contraint et adapté à la tâche nous permet de réduire les erreurs de reconnaissance dues au modèle de langage et de nous concentrer sur l'analyse des erreurs de l'étape de décodage acoustico-phonétique.

De plus, nous avons réalisé des alignements forcés sur les groupes *voix non-âgées* et *voix âgées* afin

de caractériser quels sont les phonèmes les plus mal reconnus par le *modèle acoustique générique*. L'alignement forcé consiste à convertir les transcriptions de référence en suites de phonèmes calés sur les données audio en utilisant un dictionnaire phonétique. Le modèle acoustique utilise l'algorithme de Viterbi pour calculer les intervalles temporels les plus probables pour tous les segments audio sur les phonèmes correspondants. L'alignement forcé a été réalisé avec Sphinx3 à partir du *modèle acoustique générique*.

2.3 Analyse des erreurs : WER et scores d'alignement forcé

Le décodage avec Sphinx3 génère une transcription orthographique à partir des paramètres MFCC du signal audio de parole. À partir des références orthographiques, Sphinx3 fournit des taux d'erreurs de mots (ou *Word Error Rate - WER*) permettant d'évaluer la qualité du décodage, qui ont été comparés entre les groupes *voix non-âgées* et *voix âgées*.

D'autre part, l'alignement forcé a permis d'obtenir les scores d'alignement forcé par phonème. Les scores d'alignement forcé sont des scores de vraisemblance d'appartenance au phonème normalement prononcé pour la portion de signal considérée. Ce score a été normalisé pour tenir compte du nombre de trames, et peut être interprété comme une proximité avec la prononciation "standard", modélisée par le *modèle acoustique générique*. Le score est inférieur ou égal à zéro, et plus il est faible, plus le phonème associé est éloigné du modèle acoustique. Les écarts de score les plus importants par catégories phonémiques entre les groupes *voix non-âgées* et *voix âgées* ont permis de caractériser quels sont les phonèmes posant le plus de problèmes pour la RAP des voix âgées.

Résultats : Avec le *modèle acoustique générique*, nous obtenons un WER de 7,33% pour le décodage sur le groupe *voix non-âgées*, et un WER de 12,28% pour le décodage sur le groupe *voix âgées*. Ainsi, nous observons une dégradation importante des performances de la RAP pour la voix âgée, avec une différence absolue de 4,95%, soit une différence relative de 67,53%.

Les scores d'alignement forcé calculés avec le *modèle acoustique générique* sont présentés Figure 1 par groupe phonémique. Ils permettent d'observer des comportements différents entre les groupes *voix non-âgées* et *voix âgées*.

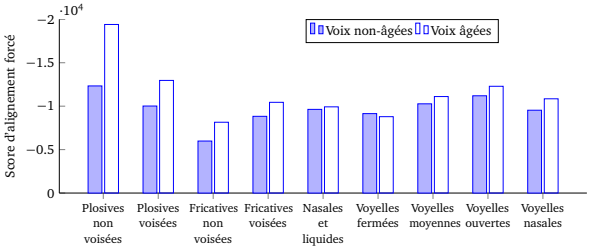

FIGURE 1 – Scores d'alignement forcé par catégorie phonémique avec le *modèle acoustique générique* pour les groupes *voix non-âgées* et *voix âgées*

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 des modèles acoustiques : les fricatives.

Pour le groupe *voix âgées*, les scores d'alignement sont globalement plus faibles que ceux obtenus pour le groupe *voix non-âgées*, et cela de façon très marquée pour les plosives. Les différences relatives de scores observées entre les deux groupes ont été calculées. Les catégories phonémiques sont par ordre descendant de différence : consonnes plosives non voisées (-57,37%), consonnes fricatives non voisées (-36,16%), consonnes plosives voisées (-29,43%), consonnes fricatives voisées (-18,25%), voyelles nasales (-13,79%), voyelles ouvertes (-9,77%), voyelles moyennes (-8,15%), consonnes nasales et liquides (-3,03%), et voyelles fermées (3,85%). Ainsi, on peut remarquer que ce sont les consonnes qui sont globalement les plus touchées. De plus, l'absence de voisement est le principal facteur de dégradation, suivie par la modalité de réalisation plosive ou fricative. Ainsi, il serait possible que les consonnes non voisées des personnes âgées soient plus proches des consonnes voisées. Enfin, il semble que le groupe le plus proche du *modèle acoustique générique* est celui des voyelles fermées, qui sont caractérisées par une ouverture minimale de la bouche.

3 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.

Le corpus constitué est un ensemble d'entretiens. Chaque entrevue met en relation une personne âgée avec deux expérimentateurs dont l'un se fait l'interlocuteur privilégié. 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 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.

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*).

4 Adaptation acoustique MLLR

La méthode d'adaptation de régression linéaire du maximum de vraisemblance (*Maximum Likelihood Linear Regression - MLLR*) a été utilisée pour adapter le *modèle acoustique générique*, appris sur *BREF120*, à la voix des personnes âgées. Le but était de voir dans quelle mesure le décodage avec modèle acoustique à adaptation MLLR diminue le WER pour le groupe *voix âgées*, avec l'hypothèse qu'il se rapprocherait du WER de 7,33% du groupe *voix non-âgées* avec le *modèle acoustique générique*. Ainsi, nous avons réalisé des adaptations MLLR selon trois méthodes différentes. Outre le décodage de référence sur le groupe *voix âgées* en utilisant le *modèle acoustique générique* pour lequel nous avions trouvé un WER total de 12,28%, nous avons réalisé trois décodages différents avec trois modèles adaptés par MLLR.

Le premier décodage a été effectué sur le groupe *voix âgées* avec un modèle acoustique dont l'adaptation MLLR a été apprise de façon globale à partir des lectures *ERES38*. L'adaptation globale est donc réalisée à partir de locuteurs (corpus *ERES38*) différents de ceux du décodage (corpus *VA*). On considère ainsi que la parole des locuteurs du corpus *ERES38* représente les caractéristiques globales de la parole âgée.

Le second décodage a été effectué sur le groupe *voix âgées* avec un modèle acoustique dont l'adaptation MLLR a été faite avec une adaptation pour chaque locuteur. Pour l'adaptation au locuteur, nous avons utilisé, à partir du seul corpus *VA*, une partie de l'enregistrement (les phrases longues extraites de magazines et journaux) d'un locuteur donné pour l'adaptation, et l'autre partie (les phrases du groupe *voix non-âgées*, c'est-à-dire les phrases courtes de vie quotidienne et de détresse) pour le décodage.

Le dernier décodage a été effectué sur le groupe *voix âgées* avec un modèle acoustique combinant les deux précédentes adaptations MLLR, soit une adaptation apprise de façon globale à partir des lectures *ERES38* suivie d'une adaptation au locuteur.

Locuteur	Genre	Age	$\text{WER}_{generique}$	$\text{WER}_{MLLRglobale}$	$\text{WER}_{MLLRlocuteur}$	$\text{WER}_{MLLRcombinee}$
L01	H	89	19,05%	12,17%	10,05%	9,79%
L02	F	83	22,08%	18,61%	14,89%	15,38%
L03	F	74	6,84%	0,38%	1,52%	1,52%
L04	H	70	5,88%	1,18%	1,57%	1,96%
L05	F	70	5,81%	3,49%	3,88%	3,88%
L06	F	77	13,04%	4,89%	5,98%	6,52%
L07	H	77	7,75%	3,52%	6,34%	6,34%
WER_{total} :			12,28%	7,29%	7,11%	7,25%
Différence absolue WER :			-	-4,99%	-5,17%	-5,03%
Différence relative WER :			-	-40,64%	-42,10%	-40,96%

TABLE 1 – Comparaison des WER en fonction des modèles acoustiques adaptés pour le groupe *voix âgées*

Résultats : Les locuteurs L01 et L02, enregistrés à l'hôpital, présentent des WER plus élevés par rapport aux autres locuteurs (cf. Table 1). Cela est lié à leurs âges et à leurs degrés de dépendance plus élevés que les personnes enregistrées à domicile.

De plus, nous voyons à la Table 1 que l'utilisation de modèles acoustiques adaptés par MLLR diminue significativement le WER, avec respectivement dans le cas de l'adaptation MLLR globale sur *ERES38*, de l'adaptation MLLR au locuteur et de l'adaptation combinée une baisse relative de 40,64%, 42,10% et 40,96%, et un WER de 7,29%, 7,11% et 7,25% par rapport au WER de 12,28% sans adaptation. En revanche, les différences entre les WER_{total} issus des décodages avec les différents modèles acoustiques adaptés par MLLR sont très faibles. 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, avec des résultats équivalents à un cas d'adaptation MLLR au locuteur. 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 un résultat satisfaisant avec un WER de 7,29%, similaire au WER de 7,33% trouvé dans le cas du décodage sur le groupe *voix non-âgées*.

5 Conclusion

L'article présente notre étude sur le comportement d'un système de RAP vis-à-vis de la voix âgée. Face à l'absence de corpus contenant de la voix de personnes âgées de langue française exploitable pour la création ou l'adaptation des modèles, nous avons procédé à l'enregistrement de nouveaux corpus. A partir du corpus *VA*, 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 taux d'erreurs de mots du système de RAP, avec une différence relative entre voix non-âgée et âgée de 67.53%. Ensuite, nous avons procédé à l'enregistrement du corpus *ERES38*, qui nous a permis d'adapter le *modèle acoustique générique* à la voix des personnes âgées grâce à la méthode d'adaptation MLLR. 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 similaires au cas d'une reconnaissance avec modèle acoustique générique de parole non-âgée, avec un WER de 7,29%, contre 12,28% avant adaptation, soit une amélioration relative de 40,64%.

Par la suite, la continuation de l'enregistrement de notre corpus s'avérera nécessaire afin d'approfondir notre évaluation des modèles acoustiques de RAP pour la voix âgée, et notre travail se portera sur l'analyse des substitutions, délétions et insertions pour chaque phonème. L'élargissement du corpus nous permettra aussi d'adapter les modèles de langage des systèmes de RAP au vocabulaire du produit de télélien social du projet CIRDO.

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 Claude Aynaud et Quentin Lefol pour leur contribution, ainsi que les différentes personnes âgées qui ont accepté de participer aux enregistrements.

Références

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 (Electronics)*, 87:49–57.

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.

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, Australia.

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