# Prévention de la perte d'autonomie chez les personnes âgées : reconnaissance vocale assistée par IA avec des questionnaires d'évaluation gériatrique standardisé

*Dona Elisa Bou Zeidan[1,2] , Abir Noun[1,2] , Mohamad Nassereddine[2] , Jamal Charara[2] , Aly Chkeir[1]*
[1]Université de Technologie de Troyes, [2]Université Libanaise,
{dona_elisa.bou_zeidan, abir.noun, aly.chkeir}@utt.fr;
{jcharara, mohamad.nassereddine}@ul.edu.lb

***Abstract - Le déclin fonctionnel est un syndrome grave qui touche les personnes âgées et qui peut être retardé ou évité par une détection précoce des risques. Dans cette étude, la reconnaissance vocale est utilisée pour aider à l'auto-évaluation des tests gériatriques utilisés à cette fin. Nous étudions l'effet de la dépendance de locuteur, de la longueur de trame, du sexe du locuteur et de la taille du vocabulaire en termes de taux de reconnaissance des mots (WRR) pour trois classificateurs différents : RF, KNN et SVM. Les résultats de nos données montrent que les trois classificateurs sont appropriés pour la reconnaissance vocale. Cependant, pour le mode dépendant du locuteur, le RF donne la performance maximale en séparant les genres (pour les femmes 98,6% et pour les hommes 98,1%) et en utilisant une longueur de trame de 25 ms avec un chevauchement de 10 ms. D'autre part, pour le mode indépendant du locuteur, la RF a également donné la meilleure performance pour 20 ms avec un chevauchement de 10 ms, pour les chiffres seuls WRR=98,1% et la classification binaire seule WRR=98%.***

***Keywords: Déclin fonctionnel, Comprehensive Geriatric Assessment CGA, reconnaissance vocale, personnes âgées, WRR enhancement, Random Forest, K-nearest neighbors, Support Vector Machine, speaker-dependent, speaker-independent.***

## I. INTRODUCTION

Le déclin fonctionnel (DF) est un sujet de recherche populaire dans le domaine des soins de santé aux personnes âgées. Il s'agit d'un syndrome dévastateur affectant les personnes âgées [1], qui commencent à perdre leurs activités de la vie quotidienne (ADL) et/ou leurs ADL instrumentales [2,3]. Les conséquences graves du déclin fonctionnel [4,5] soulignent l'importance d'une détection précoce des risques. Les questionnaires d'évaluation du déclin fonctionnel gériatrique tels que BRIGHT [6], VIDA [7], l'indice de dépendance aux ADL [8], etc. peuvent être auto-évalués en continu à domicile avec l'aide de QuestIO. QuestIO est un appareil, développé à l'Université de Technologie de Troyes, qui pose ces questions, perçoit les réponses à l'aide de la reconnaissance automatique de la parole ASR, et enfin, note les tests pour détecter la présence d'un risque de maladie cardio-vasculaire. Les scores et les alertes sont envoyés en continu à un gériatre par l'intermédiaire d'un système de santé sécurisé. De nombreux dispositifs basés sur des programmes conversationnels sont utilisés avec des personnes âgées [9,10], mais à notre connaissance, aucun ou peu sont utilisés pour l'auto-évaluation continue du déclin fonctionnel à domicile.

Ces questionnaires ont des réponses fermées. La plupart des questions peuvent être répondues par oui ou par non, ou par un choix parmi plusieurs. Les mots candidats sont donc oui, non, et les chiffres renvoient à l'un des choix. Par conséquent, en tant que RAS basé sur la reconnaissance de formes de mots isolés, notre base de données comprend oui, non et des chiffres allant de zéro à dix, ce qui donne un total de treize mots.

Dans cet article, nous comparons les performances de nos données avec trois classificateurs différents, Random Forest RF, K-nearest neighbors KNN, et Support Vector Machine SVM. Nous sélectionnons les caractéristiques optimales pour chaque classificateur en utilisant trois longueurs de trame différentes. Ensuite, en utilisant les caractéristiques sélectionnées, nous étudions l'effet de plusieurs facteurs sur le taux de reconnaissance des mots de notre système : les modèles dépendants et indépendants du locuteur, le sexe et la taille du vocabulaire. D'autres chercheurs, comme Eringis et al. [11], ont étudié l'effet de la longueur de trame et de la longueur de chevauchement sur les hommes seuls, les femmes seules et les deux. Ils ont montré qu'en utilisant le reconnaissant DTW (Dynamic Time Warping) sur le Mel Frequency Cepstral Coefficient (MFCC), la prédiction linéaire perceptuelle (PLPC) et les Linear Prediction Cepstral Coefficients (LPCC) séparément, la longueur optimale de la trame d'analyse est de 10 ms avec un décalage de trame entre 7,5 et 10 ms. Si elle est séparée, cette trame est plus courte pour les femmes (10-12,5 ms) que pour les hommes (10-17,5 ms). Cependant, dans notre étude, nous avons examiné d'autres outils de reconnaissance et des caractéristiques supplémentaires (détaillées dans la section 2.1.). S. Sunny et al [12] ont comparé les performances des classificateurs SVM, ANN et Naïve Bayes sur une longueur de fenêtre fixe. Melvyn et al. [13] ont comparé les modèles dépendants et indépendants du locuteur pour la parole masculine non dégradée, bruyante et titrée spectralement et ont prouvé que

les modèles indépendants du locuteur ont un taux d'erreur plus élevé. Dans cette étude, nous construisons de la même manière des modèles dépendants et indépendants du locuteur, mais nous comparons plusieurs effets en utilisant les mêmes signaux vocaux. Mporas et al [14] ont comparé différentes techniques de caractéristiques proposées (transformée de Fourier discrète (DFT) et transformée en paquets d'ondelettes discrètes (DWPT) - méthodes d'extraction de caractéristiques vocales comparées au MFCC et au PLPc). Cependant, notre étude combine tous les paramètres ci-dessus afin d'étudier leur effet et de comparer leur précision. Nous considérons le meilleur ensemble de caractéristiques pour chaque taille de fenêtre et chaque classificateur et nous étudions comment la précision peut être améliorée en appliquant l'effet de genre, la taille du vocabulaire et les modèles dépendants et indépendants du locuteur. Dans la section suivante, nous expliquons et détaillons notre méthodologie, puis nous exposons et discutons les résultats.

## II. METHODOLGIE

Les données ont été recueillies auprès de 25 volontaires, 14 femmes et 11 hommes, avec une fréquence d'échantillonnage de 48 kHz respectant la règle de Shannon [15]. Le protocole expérimental a été expliqué en détail à chaque participant. Ils ont ensuite signé un consentement indiquant qu'ils acceptaient d'enregistrer leur voix et de fournir les informations requises. L'anonymat des données a été entièrement respecté. Chacun des treize mots a été répété cinq fois afin de développer un système dépendant du locuteur, qui peut également être testé indépendamment du locuteur en isolant certains locuteurs pour la phase de test. Un total de 1625 audios de 1,5 seconde a été obtenu.

### *II.a Blocage des trames et extraction des caractéristiques*

Les signaux vocaux sont des signaux non stationnaires. Cependant, les courts segments de parole sont supposés être stationnaires. Par conséquent, le blocage des trames est une étape essentielle du prétraitement de ces signaux. L'approche du blocage de trame consiste essentiellement à segmenter le signal vocal en courtes trames de N échantillons qui se chevauchent de M échantillons. En général, on utilise des trames de 20 à 30 ms [17,18] avec un chevauchement de ⅓ à ½ de la taille de la trame. Par exemple, la trame de 30 ms (1440 échantillons) représente 3% d'un mot prononcé en 1 sec (48000 échantillons), et 6% du même mot prononcé en 0.5 sec (24000 échantillons). Dans notre étude, nous avons étudié l'effet de la taille des trames. Nous avons donc pris des trames de 20 ms (modèle n° 1 : 20-10), 25 ms (modèle n° 2 : 25-10) et 30 ms (modèle n° 3 : 30-10) avec un chevauchement de 10 ms. Ensuite, sur chaque segment, les coefficients du domaine fréquentiel et du domaine cepstral ont été extraits comme suit :

- Coefficient cepstral de fréquence Mel (MFCC) : 13 MFCC
- Prédiction linéaire relative spectrale-perceptuelle (RASTA-PLP) : 9 coefficients cepstraux Rasta-PLPc et 28 coefficients spectraux Rasta-PLP.
- Prédiction linéaire perceptuelle (PLP) : 13 coefficients cepstraux PLPc et 28 coefficients spectraux PLP.
- Coefficients cepstraux de prédiction linéaire (LPCC) : 9 LPCC

Ces caractéristiques sont largement utilisées dans la reconnaissance automatique de la parole [19]. Consultez Sharma et al. [20] pour plus de détails sur les caractéristiques.

### *II.b Classification et sélection des caractéristiques*

La première étape de cette étude consiste à sélectionner les caractéristiques extraites à l'aide de RF, KNN et SVM. La sélection des caractéristiques se fait en fonction du taux de reconnaissance de mots le plus élevé (WRR). Une fois les caractéristiques sélectionnées pour chaque modèle et classificateur, toutes les étapes suivantes utiliseront les caractéristiques sélectionnées correspondantes.

### *II.c Paramètres d'analyse*

Pour chaque taille de trame et chaque classificateur, les effets des modèles dépendants et indépendants du locuteur, de la taille du vocabulaire et du sexe du locuteur sur le WRR sont étudiés. Tout d'abord, un modèle dépendant du locuteur est construit en incluant tous les locuteurs dans les ensembles de d'apprentissage (75 %) et de test (25 %). Toutefois, le modèle indépendant du locuteur est construit en prenant 25 % des locuteurs comme ensemble de test et le reste comme ensemble de formation. Lorsque l'on travaille sur l'effet du sexe du locuteur, la base de données est organisée de manière à inclure l'un des deux sexes par rapport à l'autre sexe et au modèle mixte.

D'autre part, en travaillant sur l'effet de la taille du vocabulaire, la base de données a été divisée entre oui et non ensemble et les chiffres seuls. Enfin, pour chaque taille de trame et chaque classificateur, le WRR de tous les sujets (tous les mots contre oui/non contre les chiffres), des femmes (tous les mots contre oui/non contre les chiffres) et des hommes (tous les mots contre oui/non contre les chiffres) est calculé et comparé afin de choisir le meilleur modèle pour le QuestIO dépendant du locuteur et le QuestIO indépendant du locuteur.

## III. RÉSULTATS ET DISCUSSION

### *III.a Caractéristiques sélectionnées pour un modèle dépendant du locuteur*

En utilisant la sélection directe sur les données collectées, la combinaison des caractéristiques sélectionnées pour chaque trame et les classificateurs sont énumérés dans le tableau ci-dessous.

**Tableau 1: Tableau présentant les caractéristiques sélectionnées et le WRR pour chaque modèle**

| | | |
|---|---|---|
| *Modèle 1* | **RF** | RASTA-PLPc, MFCC, PLPc, RASTA-PLP |
| | | WRR1=92,55%. |
| | **KNN** | RASTA-PLPc, LPCC, PLPc |
| | | WRR1= 92,30%. |
| | **SVM** | MFCC, LPCC |
| | | WRR1= 93,01%. |
| *Modèle 2* | **RF** | RASTA-PLPc, MFCC |
| | | WRR2=97,97%. |
| | **KNN** | RASTA-PLPc, LPCC, MFCC, RASTA-PLPs |
| | | WRR2=95,87%. |
| | **SVM** | MFCC, RASTA-PLPs, PLPc |
| | | WRR2=95,87%. |
| *Modèle 3* | **RF** | RASTA-PLPc, PLPc |
| | | WRR3=92,03%. |
| | **KNN** | RASTA-PLPc, LPCC, PLPc |
| | | WRR3=94,05%. |
| | **SVM** | MFCC, RASTA-PLPc, LPCC, PLPc |
| | | WRR3=94,75%. |

Comme nous pouvons le constater, chaque classificateur dispose d'un ensemble spécifique de caractéristiques au sein de chaque modèle qui lui permet d'obtenir de meilleures performances. Dans la comparaison des caractéristiques sélectionnées dans chaque classificateur, la modification de la longueur de la fenêtre entraîne la sélection de caractéristiques différentes et conduit à des performances différentes. Toutefois, nous pouvons observer des caractéristiques communes entre les modèles. Dans les modèles 1 et 3, les mêmes caractéristiques ont été sélectionnées pour le classificateur KNN mais ont donné des WRR différents. Les résultats obtenus ici montrent que les performances de chaque classificateur dépendent fortement de la longueur de la fenêtre appliquée au signal vocal et des caractéristiques extraites de ces segments. Par conséquent, il convient d'inspecter la longueur de fenêtre et les caractéristiques optimales pour chaque classificateur choisi.

### *III.b Analyse des paramètres dans le modèle dépendant du locuteur*

En mode dépendant du locuteur, puisque le WRR le plus élevé est obtenu en utilisant des trames de parole de 25 ms avec un chevauchement de 10 ms, Tableau 2 montre strictement le $WRR_2$ du modèle #2 en considérant tous les sujets ainsi que chaque sexe seul, tous les mots et les chiffres seuls vs oui/non. Ces résultats sont en désaccord avec ceux d'Eringis et al. [11] qui ont montré que des trames d'analyse plus petites sont meilleures, sans oublier la différence de caractéristiques et d'outils de reconnaissance utilisés dans les deux études.

**Tableau 2: WRR pour le modèle dépendant du locuteur utilisant 25 ms avec 10 ms de chevauchement.**

| **RF** | **Tous les mots** | **Chiffres** | **Oui+ Non** |
|---|---|---|---|
| Tous les sujets | 97.97% | 96.56% | 99.16% |
| Femmes | 98.57% | 99.12% | 99.25% |
| Hommes | 98.09% | 96.13% | 100% |
| **KNN** | **Tous les mots** | **Chiffres** | **Oui+ Non** |
| Tous les sujets | 95.87% | 96.72% | 100% |
| Femmes | 98.46% | 98.67% | 100% |
| Hommes | 95.09% | 93.98% | 100% |
| **SVM** | **Tous les mots** | **Chiffres** | **Oui+ Non** |
| Tous les sujets | 95.87% | 95.31% | 98.23% |
| Femmes | 96.91% | 97.17% | 96.91% |
| Hommes | 91.64% | 91.39% | 100% |

Pour chaque combinaison présentée dans le Tableau 2le $WRR_1$ et le $WRR_3$ étaient significativement inférieurs au $WRR_2$ . Par exemple, dans RF, pour tous les sujets et tous les mots, $WRR_1$ =92,55% et $WRR_3$ =92,03%. De même, pour les femmes uniquement, $WRR_1$ =95,02%, $WRR_3$ =95,92% et pour les hommes $WRR_1$ =92,94%, $WRR_3$ =92,30%.

En utilisant les longueurs de trame du modèle n° 2, nous avons étudié l'effet du sexe et de la taille du vocabulaire sur le taux de reconnaissance. Pour le RF, la séparation des hommes et des femmes et la conservation de tous les mots ensemble ont amélioré le taux de reconnaissance. Cependant, la séparation des mots n'améliore pas toujours le WRR. En revanche, en utilisant le KNN, la séparation des mots a amélioré le TQR de la même manière que si l'on considérait chaque sexe séparément. Pour les femmes, le WRR a augmenté de manière significative et a continué à s'améliorer lors de la séparation des mots. Cependant, le WRR de tous les mots est resté à peu près le même pour les hommes et a diminué pour les chiffres seuls. En outre, pour le SVM, le WRR est amélioré lors de la séparation des mots et du genre, sauf pour les hommes où il diminue de manière significative, sauf pour le oui/non. En conclusion, le RF a donné les meilleurs résultats pour un modèle dépendant du locuteur créé avec nos données. Le choix entre la division des mots et la séparation des genres dépend de l'application et de la majorité des types de questions.

### *III.c Modèle indépendant du locuteur*

Le modèle indépendant du locuteur est créé en isolant 25 % des locuteurs pour la phase de test et en entraînant le modèle avec les locuteurs restants. C'est ainsi que le modèle sera entraîné sur des locuteurs spécifiques mais pourra reconnaître les mots prononcés par de nouveaux locuteurs. **Tableau 3** présente le WRR du modèle indépendant du locuteur en tenant compte de chacun des paramètres analysés. Nous pouvons clairement voir que la précision de la variété des modèles créés diminue par rapport au modèle dépendant du locuteur. Cela était attendu, d'après l'étude de Huang, X. et Lee, K.F [16] qui a montré que l'erreur moyenne dans un modèle indépendant du locuteur était plus élevée. Cela montre que les caractéristiques du locuteur

dans la parole affectent notamment le comportement du système de RPA. En outre, chaque classificateur a une longueur de fenêtre différente qui donne les meilleures performances. Pour commencer avec la RF, le modèle n°1 a le WRR1 le plus élevé = 90,46 %, mais lorsqu'il est divisé en chiffres et en oui/non, le WRR 2 a été amélioré pour atteindre environ 98 %. De même, KNN performe mieux avec un temps de 20 à 10ms (Modèle n°1)

**Tableau 3: WRR pour le modèle indépendant du locuteur.**

| **RF** | **Modèle** | **Tous les sujets** | **Femmes** | **Hommes** |
|---|---|---|---|---|
| Tous les mots | #1 | 90.46 % | 93.33 % | 70 % |
| | #2 | 88.92 % | 89.23 % | 61.53 % |
| | #3 | 89.53 % | 91.79 % | 69.23 % |
| Chiffres | #1 | 90.54 % | 91.51 % | 64.54 % |
| | #2 | 98.09 % | 89.69 % | 66.36 % |
| | #3 | 91.27 % | 90.30 % | 68.18 % |
| Oui+Non | #1 | 100 % | 100 % | 100 % |
| | #2 | 98 % | 100 % | 95 % |
| | #3 | 100 % | 96.66 % | 100 % |
| **KNN** | **Modèle** | **Tous les sujets** | **Femmes** | **Hommes** |
| Tous les mots | #1 | 85.53 % | 87.17 % | 59.23 % |
| | #2 | 85.23 % | 88.20 % | 67.69 % |
| | #3 | 84% | 89.23 % | 59.23 % |
| Chiffres | #1 | 89.09% | 88.48 % | 62.72 % |
| | #2 | 85.81% | 89.09 % | 69.09 % |
| | #3 | 86.90 % | 87.27 % | 62.72% |
| Oui+Non | #1 | 96% | 100 % | 100 % |
| | #2 | 100 % | 93.33 % | 100 % |
| | #3 | 98 % | 96.66 % | 100 % |
| **SVM** | **Modèle** | **Tous les sujets** | **Femmes** | **Hommes** |
| Tous les mots | #1 | 84.30 % | 82.56 % | 62.30 % |
| | #2 | 84 % | 89.74 % | 62.30 % |
| | #3 | 86.15 % | 85.64 % | 60.76 % |
| Chiffres | #1 | 84.72 % | 81.81 % | 67.27 % |
| | #2 | 85.09 % | 89.09 % | 64.45 % |
| | #3 | 87.27 % | 85.45 % | 64.54 % |
| Oui+Non | #1 | 100 % | 100 % | 100 % |
| | #2 | 100 % | 100 % | 100 % |
| | #3 | 100 % | 100 % | 100 % |

Cependant, le $WRR_1$ augmente également lorsque l'on sépare les mots candidats. De même, la division des mots pour le SVM donne les meilleures performances lors de l'utilisation du modèle n° 3. Une diminution significative des performances de reconnaissance vocale est observée pour les modèles créés avec des locuteurs masculins uniquement, sauf dans le cas de la classification binaire. Par exemple, le $WRR_2$ de RF passe de 88,92% à 61,53%. Le WRR moyen pour les hommes seuls est d'environ 60 %. Cela peut être dû à une énorme différence dans les caractéristiques vocales entre les locuteurs des ensembles de d'apprentissage et de test. Il se peut que la bande de fréquences des locuteurs de l'ensemble de test soit très différente, en raison de la différence d'âge ou de l'effet de la fumée sur les cordes vocales.

Enfin, nous devons diviser les mots pour un mode indépendant du locuteur afin d'obtenir une performance optimale. Pour utiliser RF, nous utilisons 25 ms avec un chevauchement de 10 ms, pour KNN, nous utilisons une longueur de fenêtre de 20 ms et 30 ms pour SVM. Cela confirme qu'il existe une longueur de fenêtre optimale pour chaque cas.

## IV. CONCLUSION ET PERSPECTIVES D'AVENIR

RF, KNN et SVM sont tous bons pour notre système de reconnaissance vocale, mais RF est plus performant. Le sexe, la taille du vocabulaire, la longueur des trames et le mode de locuteur affectent tous les performances du classificateur et chacun a des paramètres optimaux qui donnent les meilleures performances. Une diminution significative des modèles masculins indépendants du locuteur a ouvert de nouvelles perspectives pour la suite des travaux. Cette question doit être étudiée ultérieurement. En conclusion, QuestIO permet d'évaluer les tests de déclin fonctionnel de manière indépendante et de reconnaître les réponses des locuteurs, avec une précision de 98,33 % pour les locuteurs connus et de 98,05 % pour les nouveaux locuteurs. Une meilleure reconnaissance peut être obtenue en adaptant notre système de reconnaissance vocale à la population cible.

## V. RÉFÉRENCES


[1] R. Hébert, "Functional decline in old age," Cmaj, v.157, p. 1037–1045, 1997.
[2] B. M. Buurman, et al, "Clinical characteristics and outcomes of hospitalized older patients with distinct risk profiles for functional decline: a prospective cohort study," PloS one, vol. 7, p. e29621, 2012.
[3] J. J. Suijker, et al "A simple validated questionnaire predicted functional decline in community-dwelling older persons: prospective cohort studies," Journal of clinical epidemiology, vol. 67, p. 1121–1130, 2014.
[4] K. E. Covinsky, et al "Measuring prognosis and case mix in hospitalized elders: the importance of functional status," Journal of general internal medicine, vol. 12, p. 203–208, 1997.
[5] R. H. Fortinsky, et al, "Effects of functional status changes before and during hospitalization on nursing home admission of older adults," Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences, vol. 54, p. M521–M526, 1999.
[6] N. Kerse, M & al The BRIGHT tool,"Age and ageing, v.37,p.553–588, 2008.
[7] Martin-Lesende, et al "Design and validation of the vida questionnaire, for assessing instrumental activities of daily living in elderly people," J Gerontol Geriat Res, vol. 4, p. 2, 2015.
[8] K. E. Covinsky, et al, "Development and validation of an index to predict activity of daily living dependence in community-dwelling elders," Medical care, p. 149–157, 2006.
[9] M. Vacher et al, "Development of automatic speech recognition techniques for elderly home support: Applications and challenges," in International Conference on Human Aspects of IT for the Aged Population, 2015.
[10] A. Ismail, et al "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," Sustainability, vol. 12, p. 2403, 2020.
[11] D. Eringis et G. Tamulevičius, «Improving speech recognition rate through analysis parameters», The Scientific Journal of Riga Technical University-Electrical, Control and Communication Engineering, vol. 5, p. 61–66, 2014.
[12] S. Suuny, S. D. Peter et K. P. Jacob, «Performance of different classifiers in speech recognition», Int. J. Res. Eng. Technol, vol. 2, p. 590–597, 2013.

[13] M. J. Hunt et C. Lefebvre, «Speaker dependent and independent speech recognition experiments with an auditory model», chez ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing, 1988.
[14] I. Mporas, et al «Comparison of speech features on the speech recognition task», Journal of Computer Science, vol. 3, p. 608–616, 2007.
[15] C. E. Shannon, "A mathematical theory of communication," The Bell system technical journal, vol. 27, p. 379–423, 1948.
[16] X. Huang et al, «On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition», IEEE Transactions on Speech and Audio Processing, vol. 1, pp. 150-157, 1993.