Un système de dialogue vocal pour les seniors : évaluation expérimentale

Pierrick Milhorat, Jérôme Boudy, Gérard Chollet

▶ **To cite this version:**

Pierrick Milhorat, Jérôme Boudy, Gérard Chollet. Un système de dialogue vocal pour les seniors : évaluation expérimentale. JETSAN 2015 : 5ème colloque des Journées d'Etudes sur la TéléSanté, May 2015, Compiègne, France. pp.1 - 4. hal-01273050

HAL Id: hal-01273050
https://hal.science/hal-01273050

Submitted on 11 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un Système de Dialogue Vocal pour les Seniors: Evaluation expérimentale

Pierrick Milhorat[1], Jérôme Boudy[2], Gérard Chollet[1]
[1]Télécom ParisTech, Paris, France
[2]Télécom SudParis, Evry, France
milhorat@telecom-paristech.fr, boudy@telecom-sudparis.eu, chollet@telecom-paristech.fr

Abstract – vAssist (Voice Assistive Care and Communication Services for the Home) est un projet européen dans le cadre duquel plusieurs instituts de recherche et entreprises ont collaboré. L'objectif premier est de développer des interfaces vocales adaptées pour l'accès à des services de soins à domicile et de communication. L'approche utilisée, centrée sur l'utilisateur, a mené à la mise en place de plusieurs sessions expérimentales. Cet article présente le système dans sa version finale et les analyses des études conduites en France et en Autriche

Keywords: Système de dialogue vocal, expérience utilisateurs, usage.

I. INTRODUCTION

La dépendance se définit comme un état durable de la personne entraînant des incapacités et requérant des aides pour réaliser des actes de la vie quotidienne. Au 1er janvier 2012, en France, on recensait 1,17 millions de personnes âgées dépendantes. Selon les estimations, ce nombre s'élèvera à 2,3 millions à l'horizon 2060 [1,2]. La dépendance, qui croît avec l'âge, est en augmentation proportionnelle avec le vieillissement de la population. Cette évolution, inégale selon les régions mais dont la tendance est commune, entraîne la nécessité de créer de nouvelles structures d'accueil ou d'introduire d'autres modalités de prise en charge. En conséquence, la question de la capacité des familles et des services publics à supporter financièrement ces nouveaux recours contre la dépendance est au cœur du débat politique.

On distingue deux grands types de solution pour l'accompagnement des personnes dépendantes : le maintien à domicile avec prise en charge par des professionnels mobiles et l'hébergement en établissement de soins. Quelques alternatives sont envisagées, notamment la responsabilisation des aidants familiaux pour lesquels une formation peut être proposée. Les avancées scientifiques et technologiques mettent en lumière des services de maintien à domicile automatisés tels que les pistes exploitées dans le domaine de la domotique pour la sécurité, le confort, la gestion des énergies et la communication ou plus généralement l'habitat intelligent.

vAssist est un projet européen dans le cadre du programme Ambient Assisted Living. Il a pour but la conception et l'implémentation d'une plate-forme alliant des services de communication et de soins à domicile accessibles par l'intermédiaire d'un dialogue vocal à destination des personnes âgées présentant des restrictions de la motricité fine et/ou sujets à des maladies chroniques [3,4,5].

II. DESCRIPTION DU SYSTEME DE DIALOGUE

A. *Architecture générale*

La figure 1 montre la structure générale du système de dialogue vocal. Les entrées vocales de l'utilisateur sont dans un premier temps converties en hypothèses sur leur contenu orthographique par le module de Reconnaissance Automatique de la Parole (RAP). Plusieurs procédés successifs sont nécessaires à l'analyse de ces hypothèses pour en extraire la signification et produire une ou plusieurs trames sémantiques. Ces dernières sont des structures de données composées d'un but (intention de l'utilisateur) et de zéro ou plusieurs slots auxquels une valeur est assignée. A partir des trames sémantiques et suivant le contexte de dialogue dans lequel celles-ci ont été émises, le Gestionnaire du Dialogue (GD) agit sur les applications auxquelles il est connecté et guide l'interaction selon les modèles implémentés. Pour communiquer avec l'utilisateur, le GD produit également des trames sémantiques qui, à travers les deux modules que sont le générateur de langage naturel et le synthétiseur de la parole, sont converties en signal de parole contenant le message à transmettre. Chacun de ces composants sera décrit précisément dans la suite de cette section.

B. *Reconnaissance automatique de la parole*

Ce qui fait la spécificité d'un Système de Dialogue Vocal (SDV), par comparaison avec les autres types de système de dialogue, est que l'unique modalité d'interaction est la parole. La première tâche d'un programme appartenant à cette catégorie est donc de décoder le message contenu dans un signal de parole. Malgré quelques décennies de recherche sur l'analyse de tels signaux, la conversion du signal en hypothèses textuelles est toujours hautement sujette aux erreurs.

L'une des méthodes de décodage est basée sur l'algorithme de Viterbi qui permet de trouver la séquence d'états cachés la plus probable dans un réseau de modèles de Markov cachés générant une séquence d'observations. Les mel-frequency cepstrum coefficients sont actuellement les descripteurs les plus utilisés pour la représentation paramétrique d'un segment de signal de parole. La distribution des séquences de tels

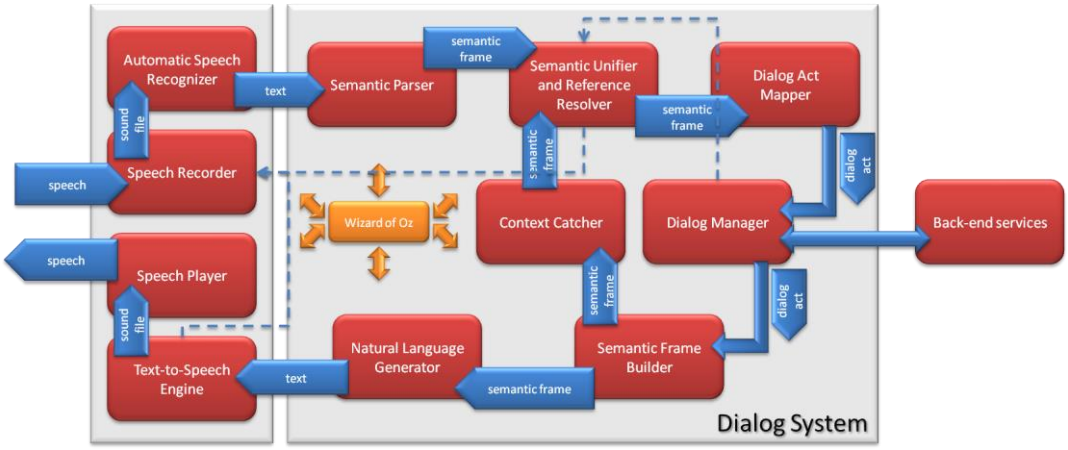

Figure 1: Vue d'ensemble du système de dialogue vocal

vecteurs de coefficients correspondante a un phonème (la plus petite unité acoustique) est modélisée par un modèle de Markov caché [6]. Les combinaisons de phonèmes pour former les mots composant un langage sont répertoriées dans un dictionnaire de prononciation. Enfin, les contraintes linguistiques sur la composition des phrases sont contenues dans le modèle de langage, la tendance actuelle pour ces derniers étant basée sur les N-grams [7].

Ces trois sources d'information (modèles acoustiques, dictionnaire et modèle de langage) permettent de délivrer un score pour chaque combinaison de mots et de produire des hypothèses sur la transcription la plus probable du signal de parole.

Nous utilisons le moteur de reconnaissance Julius, développé au Kawahara Lab de l'université de Kyoto [8]. Il est possible, en utilisant notre configuration actuelle, de reconnaître l'anglais, le français, l'espagnol, l'allemand et le hollandais. Julius a été notre choix depuis le début du projet mais nous souhaitons tout de même faire remarquer que n'importe quel système de RAP peut se substituer à celui-ci, dès lors que celui-ci produit une ou plusieurs hypothèses de transcription.

C. Compréhension du langage naturel

Bien que la plupart des modules de compréhension du langage naturel pour les SDV uni-modaux acceptent une modalité d'entrée identique, la représentation de la signification d'un message vocal diffère entre eux [9]. En d'autres termes, la structure de données produite dépend des fonctionnalités du système dans son intégralité et/ou du formalisme du GD.

Les trames sémantiques ont été sélectionnées pour leur diversité. Ces trames consistent en un but, paramétré par un ensemble de paires (nom, valeur), ce sont les slots. A la présentation d'une entrée textuelle en sortie de la RAP, le module de compréhension du langage naturel a pour rôle de convertir celle-ci en trame compréhensible pour un programme informatique. Jurcicek et al. proposent un parser basé sur une succession de transformations selon des règles apprises sur un corpus d'exemples [10]. Ce corpus d'apprentissage peut être collecté, par exemple, par l'intermédiaire d'expériences de Magicien d'Oz [11].

D. Unification sémantique et résolution des références

Le module d'unification sémantique et de résolution de référence tire son origine de la nécessité d'établir une correspondance entre l'espace sémantique dans lequel la compréhension du langage naturel produit des trames et l'espace sémantique dans lequel le GD est capable de prendre des décisions quant à la suite du dialogue la plus appropriée. Le cœur du composant est un algorithme de recherche appliqué à une forêt d'arbres de réécriture, implémenté en SWI-Prolog. L'algorithme cherche à convertir les slots contenus dans une trame sémantique en un ensemble de slots à la racine des arbres. Une recherche récursive est effectuée pour obtenir la conversion. Dans le cas où une telle conversion n'est pas possible, l'algorithme produit une erreur.

E. Interfaces

Les structures de données en entrée et sortie du GD sont différentes de celles reconnues par les modules de compréhension et de génération du langage (i.e. les trames sémantiques), pour cette raison le cœur décisionnel du système est entouré par deux interfaces qui formatent les trames sémantiques en des actes de dialogues et vice-versa. De plus, l'interface d'entrée fournit une correspondance dynamique des trames sémantiques vers les actes de dialogue qui dépend de l'état courant de l'interaction.

F. Gestion du dialogue

Plusieurs algorithmes stochastiques de gestion du dialogue ont été proposés. Ce paradigme présente actuellement deux principaux défauts : il nécessite une grande quantité de données d'apprentissage et les applications sont limitées dans

le nombre de slots, de sujets de dialogue et de possibilité d'actions pour le système.

Dans notre prototype, nous avons opté pour Disco [12], un GD déterministe dont les modèles de dialogue respectent le standard ANSI/CEA-2018, qui établit une décomposition récursive des tâches en sous-tâches et en atomes

G. *Génération du langage naturel*

La qualité de la génération du langage naturel est un facteur important pour l'acceptabilité d'un SDV, ce n'est cependant pas notre intérêt de recherche. De fait, la génération est basé sur des templates qui sont filtrés en fonction de l'intention de communication puis aléatoirement sélectionnées pour enfin être instanciés et transmis au module suivant du SDV.

H. *Synthèse de la parole*

OpenMary, un système de synthèse de la parole open-source développé par DFKI, permet, entre autres, de synthétiser le français, l'allemand et l'italien [13].

III. PROTOCOLE EXPERIMENTAL

A. *Scénarios*

L'ensemble des scénarios présentés ici utilisent une unique modalité d'entrée qui est la parole. Celle-ci est continue du point de vue de l'utilisateur puisque le système se charge d'effectuer la segmentation et le rejet/la validation des segments de parole.

Le système dispose de deux biais d'interaction : toute réponse du système est vocale et peut être accompagnée d'un stimulis visuel à travers l'application sur smartphone. Celle-ci permet de confirmer l'entrée de l'utilisateur et de situer le dialogue.

Cinq scénarios ont été définis pour la phase d'expérimentation :

- New prescription : addition d'une nouvelle prescription dans la base de données.
- Sending a message : envoi d'un message (mail/sms)
- Reporting side effect : signalement d'effets secondaires supposés dûs à la prise d'un médicament.
- Making a phone call : appeler un correspondant.
- Taking blood measure : enregistrement des données de tension artérielle et de glycémie.

B. *Protocole*

Les expériences ont été conduites auprès de 32 utilisateurs cibles : 16 en France, 16 en Autriche. Il était demandé à chaque utilisateur de finir les 5 scenarios décrits précédemment, à l'aide de la voix uniquement. L'assistant de l'expérience pouvait intervenir dans le cas où un sujet semblait ne pas parvenir à terminer un cycle d'interaction.

IV. EVALUATION EXPERIMENTALE

A. *Facilité de la tâche*

La perception de facilité est considérée comme un facteur important influençant l'expérience de l'utilisateur. La difficulté (ou facilité) apparente de la tâche a été mesurée par un Single Ease Questionnaire (SEQ) [14]. Le questionnaire été présenté à chaque fois qu'un participant avait complété une tâche. Le SEQ mesure la difficulté d'une tâche selon une échelle de valeur allant de 1 (très difficile) à 7 (très facile).

Comme montré sur la figure 2, la plupart des tâches ont été perçues par les utilisateurs comme étant relativement aisées. « reporting side effects » a été décrite comme étant plus problèmatique : plusieurs problèmes relatifs à l'interaction vocale ont été rencontrés (transmission vers le serveur du SDV défaillante, faible couverture du module de compréhension du langage, peu de données initiales, etc), indépendamment de la tâche en elle-même.

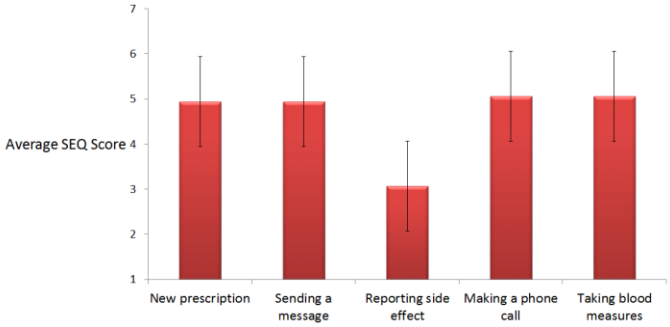

Figure 3: Résultats du SEQ

B. *Usage*

L'usage du système a été évalué selon la System Usability Scale. Les notes ou scores de cette échelle sont définis entre 0 et 100. Un score élevé traduit un meilleur usage. Des résultats inférieurs à 50 sont considérés comme non acceptable [15].

Les scores moyens en France et en Autriche sont, respectivement, 70 et 68, avec une déviation standard de 11,5 et 17,2. Il est nécessaire d'améliorer l'usage du système dans le cadre du projet.

C. *Evaluation de l'interaction vocale*

L'Evaluation subjective de l'interaction vocale a été conduite à l'aide du protocole Subjective Assesment of Speech System Interface (SASSI) [16]. Le SASSI est un outil spécialement dédié à l'évaluation d'interfaces parlantes. On présente une série d'affirmation à l'utilisateur pour lesquels il note sont accord de 1 (pas du tout d'accord) à 7 (tout à fait d'accord). L'analyse des questionnaires permet de déterminer les caractéristiques du système selon plusieurs axes tels que la facilité, l'usage, la rapidité, etc…

Les résultats pour la France et l'Autriche sont présentés sur la figure 3. Plus la valeur est élevée pour une caractéristique donnée, plus les utilisateurs ont apprécié les performances du système en regard de celle-ci.

En résumé, le confort d'usage (likeability) et la demande cognitive (cognitive demand) ont été jugés comme étant bons selon le SASSI. Les résultats moyens sont, respectivement, 5,28/7 et 5,15/7.

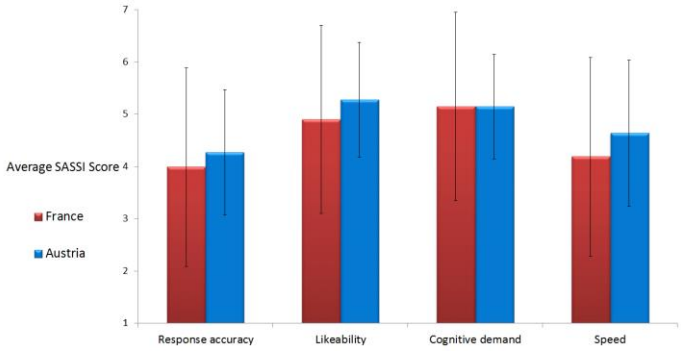

Figure 4: Résultats du SASSI

V. CONCLUSION

L'évaluation du système de dialogue vocal développé dans le cadre du projet vAssist a montré des résultats encourageant quant au déploiement de telles interfaces auprès d'une population âgée et peu encline aux nouvelles technologies.

VI. ACKNOWLEDGMENT

The research presented in this paper is conducted as part of the vAssist project (AAL-2010-3-106), which is partially funded by the European Ambient Assisted Living Joint Program and the National Funding Agencies from Austria, France and Italy.

VII. REFERENCES

[1] L. Salembier, "Projections de population dépendante à l'horizon 2020 en Ile-de-France 23600 personnes âgées potentiellement dépendantes supplémentaires d'ici 2020," Insee publications, Octobre 2009

[2] P. Boulin, C. Scribe, "Les services à la personne : un fort potentiel d'emploi, des conditions à améliorer," Insee publications, Janvier 2011

[3] G. Cordasco, M. Esposito, F. Masucci, M. T. Riviello, A. Esposito, G. Chollet, S. Schlögl, P. Milhorat, G. Pelosi, "Assessing Voice User Interfaces: The vAssist System Prototype," Cognitive Infocommunications, pp. 91-96, 2014

[4] P. Milhorat, S. Schlögl, G. Chollet, and J. Boudy, "Un système de dialogue vocal pour les seniors: études et spécifications," Journées d'Etude sur la TéléSanté, 2013

[5] P. Milhorat, G. Chollet, and J. Boudy, "Un système de dialogue vocal comme agent d'aide à la personne," Journées d'Etude sur la TéléSanté, 2014

[6] B.-H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," Technometrics, vol. 33, no. 3, pp. 251–272, 1991

[7] A. Stolcke, "SRILM – An extensible language model toolkit," International Conference on Spoken Langauage Processing, 2002

[8] L. Akinobu and T. Kawahara, "Recent Development of Open-Source Speech Recognition Engine Julius," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 131-137, 2009

[9] R. D. Mori, F. Béchet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding : A survey," Automatic Speech Recognition & Understanding, 2007

[10] F. Jurcicek, B. Thomson, and S. Young, "Reinforcement learning for parameter estimation in statistical spoken dialogue systems," Computer Speech & Language, vol. 26, no. 3, pp. 168–192, 2011

[11] S. Schlögl, G. Chollet, P. Milhorat, J. Deslis, J. Feldmar, J. Boudy, M. Garschall, M. Tscheligi, "Using Wizard of Oz To Collect Interaction Data For Voice Controlled Home Care And Communication Services," Signal Processing, Pattern Recognition and Applications, pp. 12-14, 2013

[12] C. Rich, "Building task-based user interfaces with ANSI/CEA-2018," Computer, vol. 42, no. 8, pp. 20–27, 2009

[13] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY : A tool for research, development and teaching," International Journal of Speech Technology, 2003

[14] V. Venkatesh and H. Bala, "Technology Acceptance model 3 and a Research Agenda on Interventions," Decision Sciences, vol. 39, no. 2, pp. 273–315, 2008

[15] J. Sauro and J. R. Lewis, "Quantifying the user experience: Practical statistics for user research," Elsevier, 2012

[16] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," Natural Language Engineering, vol. 6, pp. 287–303, 2000