International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 3 No. 4 Aug. 2013, pp. 893-907

http://www.issr-journals.org/ijias/

Un mécanisme de traces pour interactions téléphoniques utilisant VoiceXML

[A trace mechanism for telephone interactions using VoiceXML]

José Rouillard

LIFL Laboratory,
University of Lille 1,
59655, Villeneuve d'Ascq Cedex, France

ABSTRACT: The research that we present here are related to a study for the design and implementation of a follow-up survey of students via an interactive voice response (IVR) using VoiceXML, a W3C standard language. We present a corpus of questions and answers obtained in natural language, and we validate scientific hypotheses concerning the use of modes of interaction (voice versus direct manipulation). Then, we explain how we passed from a mechanism of exogenous traces (with a monitoring system performed by external tools recordings) to an endogenous mechanism (with a monitoring system made from within the IVR) to provide tools and instruments more adapted to the evaluation of multimodal applications that use speech and gesture (telephone keypad or mouse click on hyperlink). The trace mechanism for telephone interactions using VoiceXML presented here increases the quality of the evaluation of human-machine telephone interactions, because these traces are automatically recorded and reusable. Furthermore, we show that it is possible to get instant statistics (histograms and graphs made in real time, in PHP) using the method presented here. Thus, we have shown that pedagogical surveys, which traditionally are laborious, complex to implement and very time consuming can be facilitated through the methods and tools we recommend.

KEYWORDS: VoiceXML, Interactive Voice Response, traces, corpus, audio recording.

RESUME: Les travaux de recherche que nous exposons ici sont relatifs à une étude pour la conception et la réalisation d'une enquête de suivi d'étudiants, via un serveur vocal interactif (SVI) utilisant le langage VoiceXML, standard du W3C. Nous présentons le corpus de questions et de réponses en langue naturelle obtenu, et nous validons des hypothèses scientifiques, quant à l'usage des modes d'interactions (vocal versus manipulation directe). Puis, nous expliquons comment nous sommes passés d'un mécanisme de traces exogène (avec un système d'écoute effectué par des outils d'enregistrements externes) à un mécanisme endogène (avec un système d'écoute effectué au sein même du serveur vocal interactif), afin de fournir des outils et instruments mieux adaptés à l'évaluation d'applications multimodales permettant un usage de la parole et du geste (touche clavier téléphonique ou clic souris sur hyperlien). Le mécanisme de traces pour interactions téléphoniques utilisant le langage VoiceXML que nous exposons ici permet d'augmenter la qualité de l'évaluation des interactions hommes-machines téléphoniques, du fait que ces traces sont automatiquement enregistrées et réutilisables. Par ailleurs, nous montrons qu'il est ainsi possible d'obtenir des données statistiques instantanées (histogrammes et représentations graphiques effectuées en temps réel, en langage PHP) grâce à la méthode que nous présentons. Ainsi, nous avons montré que des enquêtes de suivi pédagogique, qui traditionnellement sont laborieuses, complexes à mettre en œuvre et très chronophages peuvent être facilitées grâce aux méthodes et outils que nous préconisons.

MOTS-CLEFS: VoiceXML, Serveur Vocal Interactif, traces, corpus, enregistrement audio.

Corresponding Author: José Rouillard (jose.rouillard@univ-lille1.fr)

1 INTRODUCTION

Depuis quelques temps, nous assistons à l'émergence de nouveaux outils de communication mobiles et ubiquitaires. Les usages des blogs, wikis, téléphones portables, smartphones et tablettes sont de plus en plus étudiés par la communauté scientifique. C'est le cas, par exemple, pour l'usage des « blogs mobiles », permettant de consulter ou de déposer une information sur un blog, grâce notamment à un téléphone mobile, en plus des autres moyens traditionnels d'accès à Internet [7]. Cela permet par exemple, pour un étudiant, de rester en contact avec l'équipe enseignante (son tuteur de stage notamment) lorsque l'usage des outils classiques (ordinateur relié à Internet) est temporairement impossible.

De ce fait, la récolte et l'analyse des traces d'interaction avec un environnement d'apprentissage est un thème de recherche en forte évolution. Tracer l'usage des outils employés permet essentiellement de pouvoir croiser différents critères permettant une meilleure adaptation des outils aux utilisateurs. Il s'agit de mieux cerner les populations d'usagers (apprenant ou groupe d'apprenants, tuteur, concepteur, administrateur, agents virtuels…), de faciliter les manipulations des larges volumes d'informations numériques recueillies et de permettre une meilleure étude des modalités de communication avec les systèmes informatiques supportant l'interaction avec les utilisateurs.

Nous exposons dans cet article la notion de traces numériques dont nous avons besoin pour tester et évaluer des systèmes informatiques permettant à des utilisateurs de dialoguer avec un automate en langue naturelle. Nos études abordent la flexibilité des serveurs vocaux interactifs (SVI) ainsi que la traçabilité des interactions lors de dialogues homme-machine téléphoniques. En effet, devant le développement rapide de nouvelles formes d'usages de l'informatique et des réseaux (mobilité, informatique ubiquitaire[1]), il devient difficile de réaliser des expérimentations valides, *in situ*, afin de déterminer la qualité des systèmes produits du point de vue de l'efficacité, de l'utilisabilité et de l'acceptation des solutions envisagées. C'est particulièrement vrai lorsque les utilisateurs peuvent interagir « sans contraintes », grâce au langage naturel ou avec des gestes, par exemple, et cela n'importe où, n'importe quand, avec une grande variété de modalités d'interaction et de multiples canaux d'accès aux systèmes interactifs. Il est alors quasiment impossible d'utiliser des méthodes et outils classiques d'évaluations des systèmes étudiés, notamment pour l'observation et la capture de traces d'interactions.

Dans cet article nous présentons dans un premier temps nos travaux relatifs à une enquête de suivi des étudiants du CUEEP[2] de Villeneuve d'Ascq, en France, réalisée au sein du laboratoire d'informatique de Lille. Nous expliquons tout d'abord le contexte et la motivation quant à ces travaux de recherche, puis nous présentons le système que nous avons mis en œuvre pour réaliser des enquêtes de suivi d'étudiants-stagiaires : il s'agit d'un serveur vocal compatible avec le langage VoiceXML [1], [14] et capable d'interagir avec l'utilisateur sur la base d'un réel dialogue homme-machine (DHM). Le protocole d'expérimentation est ensuite exposé, ainsi que les principaux résultats obtenus.

Dans un second temps, nous nous basons sur ces premiers résultats pour montrer que les traces ainsi recueillies ne sont pas suffisantes pour évaluer convenablement les systèmes de dialogue étudiés, principalement parce qu'il n'était pas possible jusqu'ici d'enregistrer de manière automatique et endogène (c'est-à-dire grâce à un mécanisme interne propre au système) ce que l'utilisateur prononçait réellement lors des différents tours de parole ; la machine enregistrait ce qu'elle croyait avoir compris, et non pas ce qu'y avait réellement été dit par l'utilisateur. Nous montrons ensuite que nous avons réussi à mettre en œuvre un système permettant de récolter des traces provenant de différentes modalités d'interactions au sein d'une même application multimodale (parole, appui d'une touche du clavier téléphonique, clic souris sur un hyperlien dans une page web).

Enfin, nous tirons des leçons de ces études et exprimons un ensemble de règles à suivre, utiles, selon nous, pour améliorer la qualité des corpus recueillis selon ce mode de communication avec l'utilisateur (vocal et/ou touches du clavier téléphonique).

[1] L'informatique ubiquitaire, telle qu'elle a été décrite il y a 15 ans par Mark Weiser, postule un monde où les individus sont entourés de terminaux informatiques interconnectés via des réseaux qui les aident dans tout ce qu'ils entreprennent.
[2] CUEEP : Centre Université - Economie d' Education Permanente

1.1 LES ENQUETES A SIX MOIS : UN TRAVAIL LABORIEUX COMPLEXE ET CHRONOPHAGE

Le CUEEP est un institut pédagogique de l'Université des Sciences et Technologies de Lille. En tant que prestataire de commande de la Région Nord Pas-de-Calais, il effectue le suivi de ses anciens étudiants-stagiaires, six mois après qu'ils aient terminé leur formation. Cela représente un travail important, dont l'expérience a montré qu'il est laborieux, complexe et chronophage (coûteux en temps de travail) :

- laborieux : après avoir effectué un premier repérage dans les listes régions de celles et ceux qui ne sont plus inscrit(e)s au CUEEP depuis 6 mois (diplôme obtenu ou non), il faut appeler ces anciens étudiants-stagiaires grâce aux coordonnées qu'ils ont laissé (numéro de téléphones fixes ou mobiles, personnel ou des parents, etc.) ;

- complexe : il y a toute une organisation à mettre en place dès l'instant où la personne visée n'a pas répondu (changement de numéro de téléphone, absence avec répondeur ; absence sans répondeur : quand rappeler, combien de fois ? …) ; complexe également par l'exploitation des résultats, même quand le questionnaire comporte assez peu de question, car cela demande un travail de retranscription, d'analyse, de statistiques … ;

- chronophage : ce travail, même bien organisé, prend du temps de secrétariat et suppose un relationnel de qualité que la fatigue et l'agacement risque de compromettre, d'autant que pour joindre efficacement les anciens stagiaires il faut très souvent les appeler en dehors des horaires de bureaux.

Nous pensons que ce travail pourrait être mené à bien grâce à un serveur vocal interactif supportant le langage VoiceXML.

1.2 LE LANGAGE VOICEXML

Il existe sur la planète beaucoup plus de téléphones que d'ordinateurs ! C'est à partir de cette constatation que les premiers projets de recherche d'accès à Internet par téléphone ont débuté dans les années quatre-vingt-dix. L'idée consiste à proposer l'accès au réseau Internet par le moyen d'un DHM en langue naturelle. Cette tendance se confirme aujourd'hui avec l'essor considérable que connaît depuis quelques années la téléphonie mobile. Les efforts de standardisation vers un langage non propriétaire, permettant non seulement la gestion des aspects téléphoniques mais également, pour partie, du DHM (relance, aide, reformulation, sous-dialogue, etc.) ont amené à la spécification du langage VoiceXML (pour *Voice Extensible Markup Language*).

Le VoiceXML [16] est donc un langage standard, à balises, permettant d'accéder à Internet grâce à un navigateur vocal via un téléphone fixe ou portable. Il est basé sur XML et s'articule autour de la reconnaissance et de la synthèse vocale, des grammaires de dialogue, et facilite la mise en œuvre de DHM sur serveurs vocaux interactifs. La version actuellement en vigueur est la version 2.1 [17]. D'ici quelques temps, le langage VoiceXML devrait pouvoir supporter les interactions multimodales, selon [2]. Le W3C travaille d'ailleurs sur d'autres langages comme EMMA[3] [3] ou X+V[4] [18] qui partagent certaines balises avec le langage VoiceXML.

Nous présentons ci-après le système que nous avons développé en VoiceXML, afin de tenter d'apporter une solution aux problèmes évoqués précédemment, concernant le suivi des étudiants, six mois après leur départ du CUEEP. Nous étudions également par ce biais les avantages et les inconvénients du VoiceXML, et faisons apparaître, en se basant sur des usages réels, des points à améliorer, comme par exemple, la notion de traces, ou bien encore l'écoute discrète de la part d'un tiers.

2 UN SYSTÈME DE QUESTION/RÉPONSES PAR SERVEUR VOCAL

Le recours à ce que l'on appelle communément un « corpus » s'est généralisé dans de nombreux domaines scientifiques (sciences humaines et sociales, sciences du langage, de l'information et de la communication, etc.). Cependant, le terme « corpus » peut être employé pour désigner deux types de données : ce peut être soit un ensemble de documents, sous forme écrite, orale ou audiovisuelle brute, soit, la somme des informations élaborée à partir de ces sources brutes que l'on vient de citer.

[3] EMMA : Extensible MultiModal Annotation markup language
[4] X+V : XHTML+Voice Profile

Comme le font remarquer [4], à propos de leur projet RITEL[5], « *Un projet riche et complexe doit faire face à plusieurs points épineux. Les plus évidents sont : la reconnaissance de la parole qui doit être à grand vocabulaire et sur laquelle une contrainte temps réel s'applique, la gestion d'un dialogue en domaine ouvert, la communication et l'échange d'informations entre un système de question-réponse et le dialogue, la génération de la réponse.* »

Même si nous avons déjà travaillé, par le passé, sur des problématiques de recherche documentaire de manière électronique, et sur la base de DHM en langue naturelle [12], ou grâce à l'interaction avec des agents animés [13] (voir par exemple les travaux du Groupe de Travail sur les Agents Animés Conversationnels à ce sujet [5]), notre approche est sensiblement différente des travaux classiques de dialogues pour de la recherche d'information dans des bases de données (type renseignements SNCF [10], accès à des base de données médicales ou documentaires (système COALA [8], système AMI [9]) par exemple), dans le sens où, ici, ce n'est pas l'utilisateur qui pose une question relativement ouverte à la machine, mais le contraire. L'effort est moins porté sur la gestion du dialogue et de son fonctionnement, à l'inverse de travaux comme ceux de [6] pour l'application ARISE du LIMSI afin de délivrer des informations ferroviaires via un serveur vocal interactif.

En revanche, ces travaux s'inscrivent plus dans une optique d'usages en contexte, de traces et d'évaluations des IHM, que nous menons dans notre laboratoire.

Nous avons également travaillé sur des dialogues de type pédagogique, dans lesquels la machine interroge l'apprenant dans différentes situations : exercices générés aléatoirement (révision en mathématiques, par exemple), tutorat (la machine accompagne l'étudiant et l'aide en lui fournissant des indices si la réponse donnée est fausse ou incomplète), contrôle des connaissances (par QCM[6] notamment). Les constatations scientifiques demeurent les mêmes : les corpus ainsi que les moyens d'observer les usages *in situ* sont peu nombreux. Il faut donc trouver de nouveaux outils pour modéliser, concevoir et réaliser des systèmes informatiques capables de fournir des traces (à la fois bas niveau et d'un degré de complexité plus élevé) au moment de leur utilisation. Il sera alors très pertinent de capturer également des informations relatives aux contextes d'usages (état des réseaux, niveau de stress de l'usager, niveau sonore ambiant, luminosité, etc.), de manière à pouvoir reproduire ces situations, plus tard, si l'on veut comprendre ce qu'il s'est passé au moment de l'interaction.

Dans l'étude que nous présentons ci-après, l'objectif était double :

1) Nos voulions, dans un premier temps mettre en place un système de DHM permettant d'obtenir de manière semi-automatique des réponses de la part des personnes contactées, afin de livrer les réponses obtenues (ainsi les statistiques qui les accompagnent) à la région Nord Pas-de-Calais, commanditaire de cette étude.

L'ensemble des questions et des réponses possibles était connu à l'avance. Les questions posées n'étaient pas forcément les mêmes pour tous les personnes, car l'ordre et l'enchaînement des questions dépendait des réponses aux questions précédentes. L'automate suivait donc un algorithme particulier, pour ne poser que les questions adéquates à la situation. Par exemple, il ne fallait pas poser la question « *Suivez-vous vos cours en école ou en université ?* » si la personne venait juste de répondre, à la question précédente, qu'elle avait arrêté ses études.

L'ensemble des réponses possibles était également connu à l'avance, puisque sans cela, il ne serait quasiment pas possible de récolter les réponses des personnes interrogées. En effet, en VoiceXML, l'utilisateur peut répondre à une question de deux manières possibles : soit en prononçant un mot ou une phrase prévue par le concepteur du système (on parle alors de grammaire vocale), soit en utilisant les touches de son clavier téléphoniques (on parle alors de grammaire DTMF[7]).

Dans un cas comme dans l'autre, si l'on n'indique pas à l'avance à la machine ce que l'utilisateur est susceptible de répondre, la poursuite du dialogue est difficilement réalisable, sauf si l'on enregistre ce que dit l'utilisateur, et qu'on le soumet à un système de reconnaissance vocale indépendant de celui utilisé sur la plate-forme VoiceXML. Dans d'autres travaux, nous avons étudié cette dernière approche [15], mais pour l'instant, les temps de réponses ainsi que les résultats obtenus sont encore insuffisants pour pouvoir réellement utiliser des systèmes se passant totalement de grammaires vocales.

[5] L'objectif du projet RITEL est de réaliser un système de dialogue homme-machine permettant à un utilisateur de poser oralement des questions, et de dialoguer avec un système de recherche d'information généraliste.
[6] QCM : Question à Choix Multiple.
[7] DTMF : Dual Tone Multi Frequency

2) Dans un second temps, nous souhaitions étudier le corpus de données ainsi recueilli et l'analyser pour valider des hypothèses de travail.

Nous formulons les hypothèses suivantes :

H1 : l'interaction vocale sera plus utilisée que la manipulation directe, lorsque les deux usages seront possibles, car selon la littérature scientifique relative à la multimodalité, ce mode est en général, plus naturel, plus facile à utiliser et reste moins contraignant que tout autre mode [11].

H2 : le type de téléphone (fixe ou portable) qui sera utilisé par les personnes appelées n'aura pas d'influence sur la qualité audio de la communication, et plus particulièrement sur la qualité de la reconnaissance vocale.

H3 : le type de synthèse vocale générée (Homme versus Femme) n'influencera pas les résultats obtenus. C'est-à-dire que nous envisageons le même taux de « bonnes » interactions, quel que soit le type de voix utilisé pour la génération des phrases que devra prononcer le serveur vocal en TTS (Text To Speech).

Enfin, nous voulons, sur la base des résultats obtenus, proposer à la communauté scientifique un guide et/ou des règles ergonomiques facilitant la mise en œuvre de telles études sur le canal téléphonique (sondage d'opinions, avis de consommateurs, QCM pour des apprenants, etc.).

2.1 PRINCIPE ET RÉALISATION

Un ordinateur équipé d'une carte téléphonique et supportant le langage VoiceXML peut dialoguer avec un interlocuteur humain, afin de recueillir des données le concernant (ici : diplôme obtenu, activité depuis six mois : salarié dans quel secteur, pour quel emploi, demandeur d'emploi, en formation, ou encore quelle tranche d'âge, etc.). Pour cela, il faut disposer d'une base de données comportant toutes les informations que l'on possède déjà, à propos des personnes concernées (nom, prénom, adresse, numéro de téléphone...), ainsi que celles que l'on souhaite obtenir lors de « l'entretien téléphonique automatisé ».

Nous avons mené une expérimentation au sein de notre laboratoire, pour tester la faisabilité technique, les contraintes, ainsi que les avantages et les inconvénients d'une telle démarche. Il s'agissait donc d'étudier, de concevoir et de réaliser une application vocale permettant de questionner, via un téléphone, des anciens étudiants-stagiaires du CUEEP, afin d'obtenir de manière automatique des informations précises les concernant. Les différentes étapes du projet furent les suivantes :

1. Etudier les données à analyser (quelles sont les données déjà connues ? celles à recueillir lors de la conversation téléphonique homme-machine ? quels sont ces types de questions, ouvertes ou fermées ?, etc.

2. Etudier la base de données existantes et déterminer quelles étaient les améliorations ou modifications nécessaires à lui apporter.

3. Concevoir l'algorithme d'enchaînement des étapes sur le serveur vocal : salutations, explication du mode de fonctionnement du système, séries de questions/réponses, obtention de l'accord (de la part de la personne interrogée) pour enregistrer les résultats recueillis, formule de politesse et salutations finales. Que faire en cas d'erreurs ou d'incompréhensions ? Comment reformuler la question ?

4. Réaliser l'application vocale en langage VoiceXML ; avec connexion à un langage dynamique (PHP) pour accéder à la base de données (MySQL).

5. Mise en œuvre de l'application sur le serveur vocal de notre laboratoire.

6. Tests et évaluations (temps de réponses, intelligibilités de la synthèse vocale, ambiguïtés lors de la reconnaissance vocale, préférences des entrées DTMF – clavier téléphoniques – pour les données sensibles, etc.).

7. Prévenir les futurs sortants du dispositif de formation de l'appel d'un serveur vocal, et, au besoin, exécuter une démonstration en public.

8. Exploiter les informations recueillies sur la base de données.

9. Transférer les informations requises aux commanditaires, de manière électronique.

En résumé, afin d'alléger cette tâche, anciennement exécuté manuellement par des secrétaires, l'étude devait amener des éléments permettant de juger de la pertinence et de la réalité des arguments prônant l'utilisation d'un serveur vocal, à savoir, qu'il ne se fatigue pas, ne s'énerve pas, peut rappeler plusieurs fois les personnes absentes à différents moments de la journée et il génère automatiquement les statistiques associées aux informations recueillies dans des base de données.

2.2 PROTOCOLE DE RECUEIL DE DONNÉES POUR LE CORPUS SEC

Le protocole suivi pour l'enregistrement du corpus SEC (Suivi des Etudiants du CUEEP) fut le suivant :

1. Activer la fiche de l'étudiant n° X, en appelant une page web du type :

http://svr/suivi_etudiants/ questionnaire_pour_une_personne.php?numero=X

Cela permettait au système de générer pour la personne indiquée (clé primaire numéro X dans la base de données) un dialogue VoiceXML personnalisé, à partir d'un patron. Par exemple, la première question devenait : « Etes-vous bien Monsieur Patrice Martin ? », au lieu de « Etes-vous bien <Titre> <Prénom> <Nom> ? »

2. Lancer l'enregistrement audio. Cela consistait à mettre en route l'enregistrement vocal de l'audioconférence, à partir d'un microphone posé à côté du téléphone de l'expérimentateur. L'intégralité des conversations était enregistrée sur un PC, dans un fichier audio au format WAV.

3. Appeler l'étudiant. Cette opération manuelle, était donc effectuée par l'expérimentateur.

4. Lui expliquer la démarche (sondage d'opinion qui ne durera que quelques minutes, grâce à une machine, avec reconnaissance vocale et/ou DTMF, etc.).

5. Sans raccrocher, appeler le serveur vocal grâce à un numéro interne du laboratoire (le 31.02).

6. L'expérimentateur doit alors appuyer sur la touche « conférence » de son téléphone et rester discret. Le mode audioconférence est activé, et la personne appelée entend la voix de synthèse du serveur vocal. Le dialogue entre eux commence.

7. Une fois la conversation terminée, sauvegarder le fichier audio ainsi obtenu.

8. Faire une copie de sauvegarder de la base de données mise à jour avec les résultats de la dernière personne interrogée.

Il est à noter que ce protocole était volontairement bridé à une seule conversation en même temps, car nous enregistrions les conversations les unes après les autres, mais en pratique, rien n'empêche le serveur vocal de supporter plusieurs appels simultanément.

2.3 LE QUESTIONNAIRE

Nous avons conçu un système de Questions/Réponses de manière générique, de sorte qu'il soit facilement modifiable pour tout autre questionnaire. Pour cela, les phrases à synthétiser ainsi que les grammaires vocales nécessaires pour que les utilisateurs expriment leurs réponses ne sont pas codés « en dur » dans le serveur vocal, mais plutôt, déployé, à la volée, en fonction de fichier texte, faciles à éditer et à modifier, comme le montre la Figure 1. On y voit les étiquettes grâce auxquelles les champs (*<field>* en VoiceXML) sont créés, et qui permettent d'effectuer des sauts d'une question à une autre, en fonction des réponses obtenues.

```
>presentation
Bonjour, je suis l'automate vocal du CUEEP. Je vais vous poser quelques questions afin d'assurer le suivi des anciens inscrits au cuèpe[8].:info_touche
>info_touche
Vous pourrez répondre vocalement, ou à l'aide des touches de votre téléphone à chacune des questions qui va vous être posée.:verification
>verification
Êtes-vous bien %prenom% %nom% ?
-oui:temps
-non:espoir
>temps
Avez-vous quelques minutes à m'accorder ?
-oui:info_navigation
-non:rappeler
>espoir
Pouvez-vous me passer %prenom% %nom% ?
-oui:verification
-non:erreur
[…]
>info_navigation
A tout moment, vous pourrez obtenir de l'aide en prononçant le mot aide, ou en appuyant sur la touche étoile. Pour faire répéter une question, dites répétez. Si vous voulez modifier la réponse que vous venez de donner vous pouvez dire annuler.:q1_statut
>q1_statut
Etes-vous en activité, à la recherche d'un emploi, en formation, ou en contrat de travail ?
-en activité:q2_salarie
-en recherche d'un emploi:q4_domaine
-en formation:q5_formation
-en contrat de travail:q12_formation_contrat
```

Fig. 1. Exemple de fichier permettant la génération automatique d'un fichier VoiceXML pour DHM sur serveur vocal

L'utilisateur interagit donc avec un automate qui lui fournit des répliques personnalisées de manière dynamique, en fonction des résultats obtenus au fur et à mesure de l'avancement du dialogue. Nous présentons ci-dessous les principaux éléments obtenus.

3 RÉSULTATS

Les résultats qui nous semblent les plus pertinents sont ici présentés sous la forme de deux catégories. Il s'agit d'une part d'un corpus de Questions/Réponses, et d'autre part du corpus de fichiers audio.

3.1 LE CORPUS QUESTIONS/ RÉPONSES SEC

Le serveur vocal utilisé pour le recueil du corpus SEC (Suivi des Etudiants du CUEEP) était le système Phonic, d'Idylic[9], supportant le langage VoiceXML version 1.0, avec une reconnaissance vocale de Philsoft de *Telisma* et une synthèse vocale *Tempo* d'*Elan Informatique*, et DTMF (touches du téléphone). Trois pré-tests ont été effectués afin de vérifier le bon fonctionnement du matériel et des logiciels. Cela portait notamment sur la génération automatique du fichier VoiceXML contenant les informations relatives à la personne à contacter, l'aboutement[10] téléphonique vers des téléphones fixes et mobiles, l'enregistrement audio des conversations Homme-Machine, et la sauvegarde des réponses obtenues dans la base de données.

[8] Cette forme phonétique a été utilisée car sinon, le mot CUEEP était prononcé « CUPE » par la machine.
[9] http://www.idylic.com/
[10] Dans les spécifications du langage VoiceXML, il est possible de paramétrer le type de transfert à l'aide de l'attribut *bridge* de la balise *transfer*. S'il s'agit du mode *blind*, l'aboutement s'effectue sur le réseau de l'opérateur, tandis que s'il s'agit du mode *bridge*, l'aboutement est réalisé localement, sur le serveur vocal interactif. De plus, si l'on aboute plus de deux lignes entre elles, on obtient une conférence téléphonique.

Le nombre de personnes à contacter dans le fichier était de 39. Le nombre de personnes effectivement contactées est de 23 (58,97 %). Sur les 16 autres contacts, 5 étaient absents à chaque tentative d'appel, 4 étaient de faux numéros (soit réattribués, soit inactifs), et 7 ont commencé le questionnaire mais ne l'ont pas terminé. La part des communications effectuées vers des téléphones fixes représente 47,83 % des appels (respectivement 52,17 % vers des téléphones portables).

L'enquête était donc destinée à obtenir de manière semi-automatique des données statistiques, comme par exemple, la situation professionnelle des personnes interrogées, au moment de l'appel. Les réponses attendues pour cet exemple de question, et pour lesquelles une grammaire vocale (et DTMF) avait été préparée, étaient les suivantes : *en activité*, *en contrat de travail*, *en formation*, *à la recherche d'un emploi*. La Figure 2, ci-après, synthétise les résultats automatiquement obtenus, grâce au questionnaire programmé sur le serveur vocal interactif, pour la question relative à la situation professionnelle des personnes interrogées.

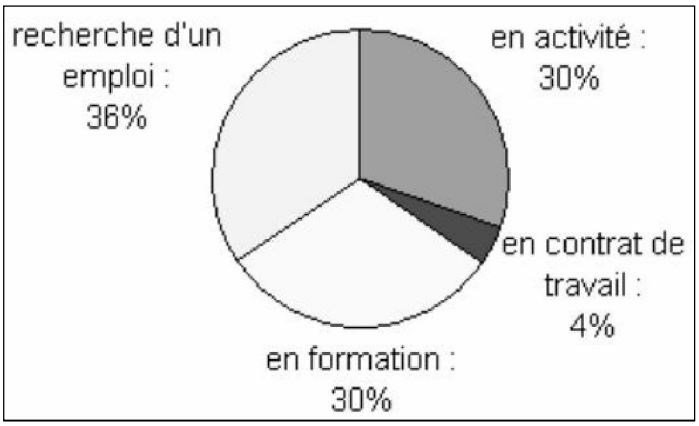

Fig. 2. Exemple de données obtenues grâce au questionnaire programmé sur le serveur vocal (ici, situation professionnelle des personnes interrogées)

L'âge des participants varie de 22 à 47 ans, avec une moyenne de 28 ans. Les autres résultats, qui ne font pas l'objet de cet article, et qui sont ici donnés à titre d'illustration, sont résumés sur la Figure 3 ci-après :

Type des activités : non salarié :1 vs salarié :6
Type des contrats : Sur le 6 salariés, 6 sont à durée indéterminé.
Domaine des activités : Pour les 16 personnes qui sont en activité, en contrat de travail ou à la recherche d'un emploi) : commercial : 5, formation : 2, santé : 3, social : 6.
Pour les 7 personnes en formation, les cours sont suivis en : école : 2 ; université : 5
Pour la seule personne en contrat de travail, les cours sont suivis en : école : 1
Pour les 5 personnes en formation en Université, les types de formation : littéraire : 2, technique et scientifique :3
Type de DAEU[11] obtenu : DAEU A : 20 ; DAEU B :3
Formation à distance : Non : 23 (100% en présentiel)
Objectif poursuivi en faisant cette formation : poursuite d'étude : 18, prétendre à un emploi : 1, satisfaction personnelle : 4.

Fig. 3. Quelques résultats propres à l'étude menée grâce au dialogue Homme-Machine vocal

La Figure 4 quant à elle, présente le nombre d'interventions vocales pour quelques questions du corpus. La rangée la plus éloignée donne le nombre d'intervention vocale au rang numéro 1, c'est-à-dire lorsque l'utilisateur a choisi ce mode pour répondre, lors de sa première intention.

[11] DAEU : diplôme d'accès aux études universitaires

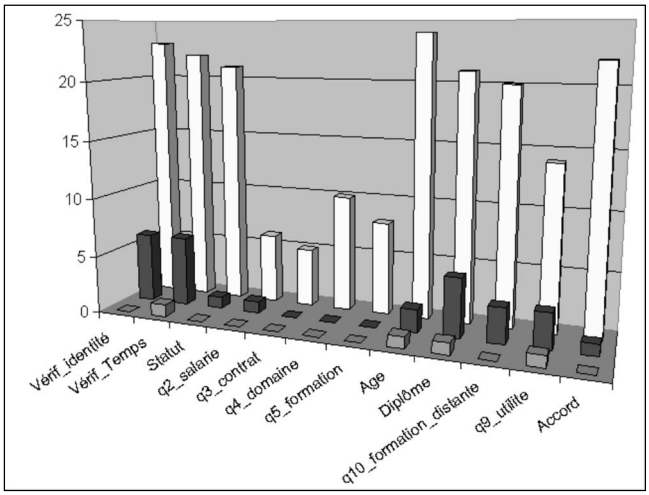

Fig. 4. Rangées V3, V2, V1, représentant les réponses vocales pour quelques questions posées par le SVI

La rangée intermédiaire représente le rang numéro 2, et la plus proche donne le rang numéro 3. Ainsi, pour la ligne 1 (étiquette verif_identité), 23 fois[12], les utilisateurs ont tenté de parler pour exprimer leur première réponse à cette question ; puis 6 fois lors d'un deuxième essai, toujours pour cette même question. On note que c'est pour donner leur âge que les utilisateurs ont le plus utilisé leur voix.

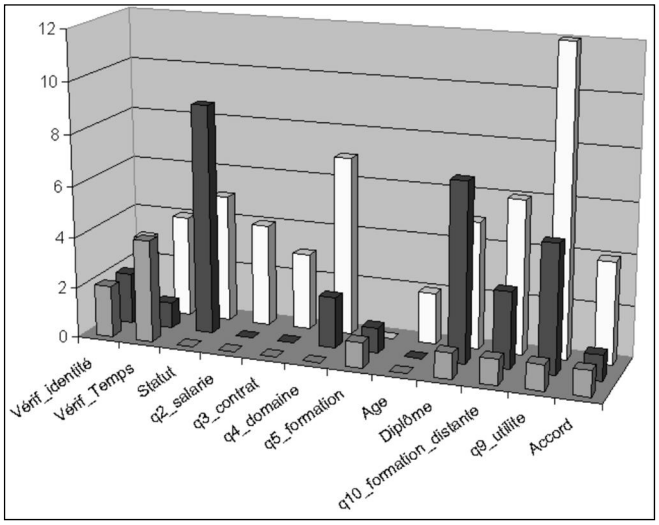

Fig. 5. Rangées D3 D2, D1, représentant les réponses par DTMF pour quelques questions posées par le SVI

Sur la Figure 5, on voit la répartition, pour les mêmes questions, mais en ce qui concerne le mode DTMF, cette fois-ci. On observe par exemple que c'est pour la question « Q9_utilite : Pourquoi avez-vous suivi cette formation ? satisfaction personnelle ? poursuite d'étude ? ou prétendre à un emploi ? » que les utilisateurs ont utilisé le plus leur clavier téléphonique.

La Figure 6, ci-après, résume les erreurs (rangée la plus éloignée), demande d'aide (rangée intermédiaire) et demande de répétition (rangée la plus proche). On voit donc que c'est en tentant de recueillir l'âge des utilisateurs que la machine a fait le plus d'erreurs, et aussi que les utilisateurs ont demandé le plus d'aide. C'est, en revanche, à propos du diplôme obtenu « DAEU A » ou « DAEU B » que les personnes interrogées ont demandé le plus souvent à la machine de répéter sa question.

[12] Le nombre d'observations est de 26 (et non pas 23), car une personne a recommencé deux fois le processus, avant de le valider, et une autre a recommencé une fois.

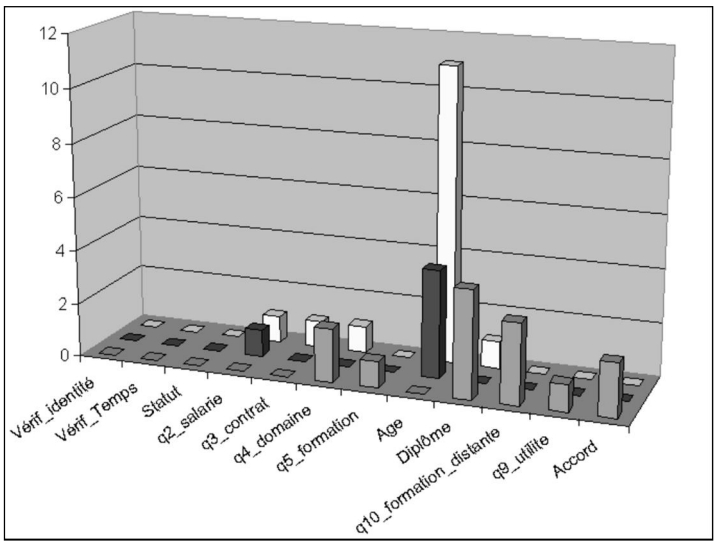

Fig. 6. Rangées Aide, Répéter et Erreur

3.2 LE CORPUS AUDIO SEC

Le corpus audio représente au total 4888 secondes d'enregistrement, soit 1h35 minutes. En moyenne chaque conversation dure 213 secondes (écart type de 90). Le dialogue le plus court dure 132 secondes, et le plus long dure 574 secondes. Ces durées ont été automatiquement calculées par l'automate qui estampillait le début et la fin de chaque conversation. Il s'agit ici des durées effectives des DHM, c'est-à-dire que l'on n'intègre pas dans ce calcul le laps de temps dans lequel l'expérimentateur explique le principe de fonctionnement à la personne appelée.

Il apparaît très nettement au vu des résultats d'interaction avec le serveur vocal, que l'hypothèse H1 que nous avions formulé était correcte, dans le cadre de cette étude. En effet, sur les 325 interactions, 228 (soit 70,15%) étaient orales, contre 97 par manipulation directe (DTMF).

Plus précisément, ces 228 interactions vocales ont été réalisées en première intention à 196 fois (étiquette notée V1 dans le corpus), en deuxième intention à 28 reprises (V2), et en troisième intention à 4 reprises (V3). Pour le DTMF, les valeurs D1, D2 et D3 sont respectivement de 55, 31 et 11.

Les résultats relatifs aux types de téléphones (11 téléphones fixes contre 12 téléphones mobiles dans l'étude), semblent aller dans le sens de l'hypothèse que nous avions formulé (H2) : il n'y a pas plus d'erreurs avec des téléphones mobiles qu'avec des téléphones fixes dans notre étude (respectivement 8 erreurs de compréhension, avec des téléphones fixes, corrigées après réécoute par l'expérimentateur, de manière asynchrone, contre 7 erreurs pour les portables).

En revanche, pour les 13 demandes d'aide, 12 émanent d'utilisateurs équipés des téléphones mobiles, contre 1 seule d'un téléphone fixe. Mais ce résultat n'est donné qu'à titre anecdotique, car rien ne permet, à ce stade, de prétendre que la demande d'aide est fortement liée au type de téléphone utilisé. Pour cela, nous devrons croiser plusieurs autres critères du corpus qui n'ont pas encore totalement été exploités.

Enfin, en ce qui concerne l'hypothèse H3, elle semble être correcte, puisque nous ne voyons pas de taux d'erreurs qui seraient plus accentuées avec l'une ou l'autre des voix de synthèse. Les taux d'erreurs sont proportionnels aux taux d'apparition des TTS respectives (les 14 utilisations de voix masculine contre 9 utilisations de voix féminine sont proportionnelles ont 11 erreurs avec des TTS masculine, et 7 avec des TTS féminine).

3.3 EXEMPLES D'ERREURS TIRÉES DU CORPUS

Voici quelques exemples, tirés du corpus SEC, que nous donnons pour illustrer les éléments que nous venons de citer, ci-dessus.

- Question : « Pourquoi avez-vous suivi cette formation ?[13] satisfaction personnelle ? poursuite d'étude ? ou prétendre à un emploi ? »

- Réponse : « Au début satisfaction personnelle, et ensuite …. »

La machine n'a pas compris cette phrase, car la grammaire vocale ne s'attendait pas à une telle amorce de la part de l'utilisateur. Sans le mécanisme de trace et d'écoute discrète que nous avons mis en œuvre, il n'aurait pas été possible, après coup, de se rendre compte d'où provenait l'erreur.

- Question : « Etes-vous en activité, à la recherche d'un emploi, en formation, ou en contrat de travail ? »

- Réponse : retraitée.

Cette réponse n'ayant pas été prévue par les commanditaires de l'étude, la machine n'a pas compris cette phrase. La même question a donc été reposée plusieurs fois à l'utilisateur, qui, après avoir insisté et changé de ton, a finalement a décidé de raccrocher, non sans avoir prononcé une dernière phrase de mécontentement : « oh, hein, ça va bien hein ».

- Question : « Etes-vous en activité, à la recherche d'un emploi, en formation, ou en contrat de travail ? »

- Réponse : « Oui ».

L'utilisateur a cru qu'on lui demandait s'il était dans l'une des situations évoquées, et n'a pas compris qu'il agissait d'énoncer la situation précise dans laquelle il se trouvait.

- Question : « Votre formation est-elle littéraire ou technique et scientifique ? »

- Réponse : « Non »

- Réponse : « Scientifique »

Ici, le découpage était d'une part « littéraire » et d'autre part « technique et scientifique ». L'utilisateur n'a pas su, à l'oreille, faire ce distinguo.

Ces quelques exemples illustrent bien le besoin de traces, d'instruments et d'outils dont les chercheurs ont besoin pour mener à bien leurs études scientifiques. Nous expliquons dans les lignes qui suivent notre approche pour proposer une solution originale à ce problème.

4 INSTRUMENTATION ET MÉCANISME DE TRACES POUR VOICEXML

A la lumière des résultats obtenus lors de cette première expérience, nous pouvions dire que le langage VoiceXML ne possédait pas, de manière interne, un mécanisme permettant de tracer les interactions avec les utilisateurs qui se connectaient sur le SVI. Jusqu'à la version 2.0 de VoiceXML, il n'était pas possible, à notre connaissance d'effectuer une trace réelle de ce qu'avait effectivement prononcé les utilisateurs au cours de leur dialogue avec le système. En revanche, il était possible de consulter la variable chargée de récupérer la valeur supposée correspondre à un élément de la grammaire vocale. Autrement dit, il n'était pas possible, jusqu'ici, d'enregistrer de manière automatique ce que l'utilisateur prononçait réellement lors des différents tours de parole.

Le processus que nous avions mis en œuvre afin d'enregistrer les conversations entre les utilisateurs et notre serveur vocal était externe au système lui-même et enregistrait également les autres bruits ambiants. Cette première étape a démontré que ce que disaient les utilisateurs n'était pas forcément ce que le système avait cru entendre. Typiquement, à la question « Quel est votre âge ? », il est arrivé que la machine enregistre la réponse « 33 » dans la base de données, alors que la véritable réponse prononcée par l'utilisateur était « 23 ». L'expérimentateur, en écoute discrète avait noté cette anomalie et a pu réécouter l'enregistrement sonore, plus tard, afin de conforter son opinion.

Depuis la version 2.1 du VoiceXML du W3C, il est techniquement possible d'enregistrer ce que prononce l'utilisateur lors d'une interaction vocale. Pour cela, il faut initialiser l'attribut *recordutterance* de la balise <*property*> à la valeur « *true* ». On peut alors obtenir, grâce à certaines variables d'application, les données suivantes :

[13] Même si la phrase n'est pas très correcte, syntaxiquement, nous y avons laissé plusieurs points d'interrogations afin d'obtenir une meilleure prosodie (ton montant pour une interrogative).

application.lastresult$.confidence
Le niveau de fiabilité de l'énoncé de cette interprétation dans l'intervalle 0.0-1.0. Une valeur de "0.0" indique une fiabilité minimale et une valeur de "1.0" une fiabilité maximale.

application.lastresult$.inputmode
Le mode selon lequel l'entrée d'utilisateur a été fournie : "dtmf" ou "voice".

application.lastresult$.recording
La donnée vocale correspondant à ce que l'utilisateur a dit.

application.lastresult$.recordingduration
La durée (en msec) de la dernière reconnaissance vocale.

application.lastresult$.recordingsize
La taille (en octets) de la dernière reconnaissance vocale.

application.lastresult$.utterance

La chaîne des mots bruts qui ont été reconnus pour cette interprétation.

Dans le cas d'une grammaire DTMF, cette variable contiendra la chaîne numérique reconnue. Nous avons utilisé cette spécificité dont dispose le serveur vocal de notre laboratoire[14] pour mettre un œuvre un mécanisme d'enregistrement global de toutes les traces d'interactions homme-machine possibles en VoiceXML (voix et clavier téléphonique). Cette instrumentation permet donc de collecter des traces qui reflètent l'usage d'une application. Ces informations sont estampillées et enregistrées dans une base de données. Nous avons développé un moyen d'accès à ces traces, pour qu'un expérimentateur, non informaticien, puisse analyser les données recueillies. A partir de cette interface, un « expert » du domaine étudié peut prendre connaissance de l'enregistrement vocal (fichier .wav) correspondant à la réponse d'un utilisateur et le comparer à ce que la machine a cru comprendre durant l'interaction.

Nous avons testé ce processus de traces avec une application multimodale de commerce électronique comportant une reconnaissance et une synthèse vocale, l'usage possible du clavier téléphonique DTMF, mais aussi des clics souris sur des hyperliens d'une page Web, la visualisation des images des produits présentés, etc.

La Figure 7 ci-après indique, en première ligne par exemple, que lorsque l'action consistait à choisir une taille de vêtement, la machine a cru reconnaître la réponse « 36 ». En cliquant sur le fichier .wav de cette même ligne, l'expert peut vérifier que l'utilisateur avait bien prononcé cette information, et le cas échéant rectifier l'information. Les statistiques de bonnes/mauvaises compréhension de la part de la machine sont ainsi mises à jour et permettent d'instrumentaliser les évaluations des interfaces générées.

Num	la_date	heure	action	utilisateur	trace_utilisateur
572	Tue16May	16h40min19s	choisir taille	36	utilisateur_Tue16May_16h40min19s_vox.wav
573	Tue16May	16h40min34s	choisir couleur	bleu	utilisateur_Tue16May_16h40min34s_vox.wav
574	Tue16May	16h40min53s	choisir référence	25	utilisateur_Tue16May_16h40min53s_vox.wav
575	Tue16May	16h41min05s	ajouter au panier	non	utilisateur_Tue16May_16h41min05s_vox.wav
576	Tue16May	16h44min12s	reference client	2	utilisateur_Tue16May_16h44min12s_vox.wav
577	Tue16May	16h44min30s	choisir produit	Robe Bleue	utilisateur_Tue16May_16h44min30s_vox.wav
578	Tue16May	16h44min40s	choisir rayon	Femme	utilisateur_Tue16May_16h44min40s_vox.wav

Fig. 7. Traces d'enregistrement vocal en VoiceXML 2.1

[14] En effet, le SVI Sibilo Voice de l'entreprise App-Line supporte partiellement la version 2.1 de VoiceXML.

Le mécanisme mis en œuvre pour cette instrumentalisation est schématisé sur la Figure 8 ci-dessous.

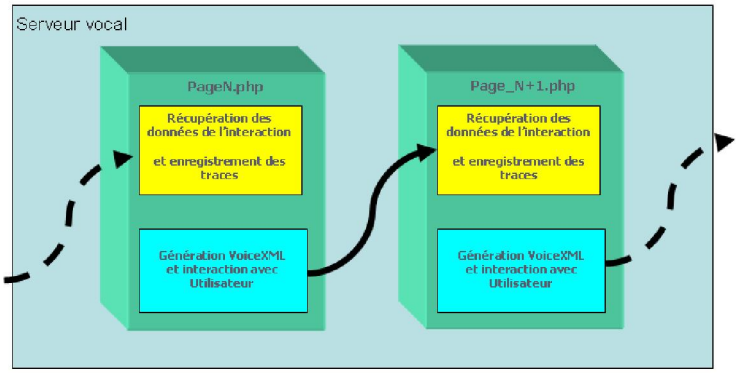

Fig. 8. Mécanisme pour recueillir des traces d'interactions sur SVI

En résumé, une page web dynamique récolte les traces du tour de parole précédent, puis sauvegarde ces informations dans une base de données ainsi que des fichiers audio, et enfin génère les fichiers VoiceXML pour l'interaction en cours.

5 LEÇONS À RETENIR POUR LA CONCEPTION DE DHM GRÂCE À VOICEXML

La conception et le développement d'une bonne interface Homme-Machine passe par diverses étapes. Nous en rappelons brièvement le cycle de vie. Il est tout d'abord nécessaire de passer par une phase d'analyse. Elle devra déterminer l'étendue de l'application, les besoins et les types d'utilisateurs identifiés, la valeur ajoutée que le service apportera à cette population d'usagers, etc. La réalisation passera par une phase de choix de la plate-forme VoiceXML, selon des critères de qualité des technologies vocales, des coûts, de la possibilité de suivre (en temps réel ou en différé) les activités des utilisateurs connectés, etc. Ensuite, les phases de développement, de tests et d'évaluation devront être menées à bien. Durant cette période, il est important de consulter régulièrement les fichiers de traces des interactions sur le serveur vocal (log files en anglais). Cela permet de vérifier le bon déroulement des différentes phases du dialogue, et l'on peut y déceler certaines anomalies ou problèmes techniques particuliers : on y découvrira, par exemple, qu'un fichier n'est pas réellement chargé car il se trouve déjà en mémoire cache de l'ordinateur. Les outils et mécanismes de traces précédemment décrits seront utiles non seulement lors des phases d'évaluations, mais également tout au long du cycle de développement d'applications monomodales (uniquement à l'oral par exemple) ou multimodales, pour lesquelles il existe très peu de moyens permettant de tracer les usages. Cela est d'autant plus vrai dans le cadre de campagne d'évalutions hors laboratoire, qui sont nécessaires pour des usages spécifiques dans le cadre de la mobilité, de l'informatique ambiante ou pervasive.

6 CONCLUSIONS ET PERSPECTIVES

Nous avons montré dans un premier temps qu'une tâche consistant à questionner une personne avec une liste de Questions/Réponses connues peut être avantageusement effectuée par un serveur vocal interactif supportant le langage VoiceXML. Au cours de notre étude préliminaire, nous avons recueilli un corpus de réponses, pour le commanditaire de l'étude (la région Nord Pas-de-Calais), mais également, des données permettant de confronter nos résultats à des hypothèses scientifiques. Il a été vérifié, qu'effectivement, l'interaction vocale est plus utilisée que la manipulation directe, dans cette étude, lorsque l'utilisateur en à la choix (H1), que le type de téléphone utilisé (fixe ou mobile) n'influence en rien les résultats obtenus (H2), ni même le type de synthèse vocale jouée, qu'elle soit masculine ou féminine (H3).

Mais nous avons surtout expliqué que la communauté scientifique manque significativement d'outils et d'instruments fiables permettant d'effectuer des traces d'interactions de nouvelles formes de communications multimodales. La deuxième étude présentée dans cet article a montré comment nous sommes passés de traces exogènes, avec enregistrements extérieurs au système à un mécanisme de traces endogènes, grâce notamment à l'usage de la version du langage standard VoiceXML 2.1. Cela a permis de récolter des traces provenant de différentes modalités d'interactions au sein d'une même application multimodale (parole, appui d'une touche du clavier téléphonique, clic souris sur un hyperlien dans une page web).

Pour rendre le corpus que nous avons succinctement présenté, véritablement exploitable par d'autres chercheurs, des travaux sont encore nécessaires. Il faudra anonymiser le corpus audio, l'annoter avec un outil comme Praat[15], puis permettre l'interrogation de la base de données sur des critères croisés : âge, sexe, type de téléphone (fixe ou mobile) des personnes interrogées, mais aussi par exemple le type de voix de la synthèse vocale utilisé (homme versus femme), la qualité audio de la conversation, la durée totale ou pour chaque question, l'heure de l'appel, etc.

Par ailleurs, nous travaillons actuellement, en collaboration avec l'entreprise App-Line[16] à l'amélioration des serveurs vocaux qui intègrent déjà certaines fonctionnalités du langage VoiceXML 2.1, comme l'enregistrement de ce que prononce l'utilisateur afin d'effectuer des traces de manière standard. Nous sommes en train de tester certaines balises (dites propriétaires, car ne faisant pas partie de la spécification VoiceXML 2.1 officielle) permettant de débuter et de stopper une trace automatique, et cela à tout moment de l'interaction avec le SVI. Cela peut se faire, soit grâce à une console de suivi en temps réel des conversations transitant sur le SVI (l'expert clique sur un bouton pour activer/stopper l'enregistrement vocal), soit de manière logicielle (une balise <object> en VoiceXML), soit encore par le biais de Web services supportant le protocole SOAP[17].

Enfin, nous continuons nos recherches, afin de mettre en place une solution complète de « push vocal », permettant de passer efficacement d'un système semi-automatique (dans lequel l'humain effectue encore l'amorce auprès de l'interlocuteur, en lui présentant la démarche et le principe d'utilisation), à un système totalement automatique, avec possibilités d'enregistrements de plusieurs conversations téléphoniques simultanément.

REMERCIEMENTS

Ces travaux sont, pour partie, le résultat de réflexions et de discussions avec Christian Ladesou, du Centre CUEEP de Villeneuve d'Ascq. D'autre part, ces études scientifiques sont partiellement financées le contrat de plan Etat Région Nord Pas de Calais, et le FEDER (Fonds Européen de Développement Régional).

[15] http://www.fon.hum.uva.nl/praat
[16] http://www.app-line.com
[17] SOAP : Simple Object Access Protocol

REFERENCES

[1] Anderson, E. A., Breitenbach, S., Burd, T., Chidambaram, N., Houle, P., D. Newsome, D, Tang, X., Zhu, X., *Early Adopter VoiceXML*, Wrox, 2001.

[2] Dettmer, R., "It's good to talk, speech technology, for on-line services access," *IEE Review*, Vol. 49, No. 6, June 2003, pp. 30-33.

[3] EMMA : *Extensible MultiModal Annotation markup language*. [Online] Available: http://www.w3.org/TR/2005/WD-emma-20050916/ (2013)

[4] Galibert O., Illouz G., Rosset S., *RITEL : dialogue homme-machine à domaine ouvert*, TALN 2005, Dourdan, 2005.

[5] GT ACA : *Groupe de Travail sur les Agents Conversationnels animés* [Online] Available: http://www.limsi.fr/aca/ (2013)

[6] Lamel, L., Rosset, S., Gauvain, J.L., Bennacef, S., Garnier-Rizet, M. et Prouts, B., "The LIMSI ARISE System," *Speech Communication*, 31(4):339-354, 2000.

[7] Leclet, D., Leprêtre, E., Peter, Y., Quénu-Joiron, C., Talon, B., Vantroys, T., "Améliorer un dispositif pédagogique par l'intégration de nouveaux canaux de communication," *EIAIH 2007*, Lausanne, Suisse.

[8] Lehuen J., "Un modèle de dialogue dynamique et générique intégrant l'acquisition de sa compétence. Le système COALA," *Thèse de doctorat, Université de Caen*, Juin 1997.

[9] Lemeunier T., "L'intentionnalité communicative dans le dialogue homme-machine en langue naturelle," *Thèse informatique, Le mans*, 2000.

[10] Luzzati D., "Recherches sur le dialogue homme-machine : modèles linguistiques et traitements automatiques," *Thèse de l'Université de la Sorbonne*, 1989.

[11] Oviatt, S. L. & Cohen, P. R., "Spoken language in interpreted telephone dialogues," *Computer Speech and Language*, 6 (3) 277-302, 1992.

[12] Rouillard J., "Hyperdialogue sur internet. Le système HALPIN," *thèse d'informatique de l'université Joseph Fourier, Grenobl*e, 2000.

[13] Rouillard J., "Hyperdialogue avec un agent animé sur le Web," *ERGO-IHM'2000*, Biarritz, 2000.

[14] Rouillard J., *VoiceXML. Le langage d'accès à Internet par téléphone*, éditions Vuibert, ISBN : 271174826X, 197 pages, Paris, 2004.

[15] Rouillard, J., Truillet P., *Enhanced VoiceXML*, HCI International 2005, Las Vegas, 2005.

[16] VoiceXML 1.0., *W3C Recommendation*. [Online] Available: http://www.w3.org/TR/voicexml10 (2013)

[17] VoiceXML 2.1, *Working Draft*. [Online] Available: http://www.w3c.org/TR/2004/WD-voicexml21-20040323 (2013)

[18] X+V : *XHTML+Voice Profile 1.0* [Online] Available: http://www.w3.org/TR/xhtml+voice/ (2013)