IDIAP COMMUNICATION

IDIAP

Martigny - Valais - Suisse

DEVELOPEMENT D'UN SYSTEME DE DEMANDE INTERACTIF VIA LE TÉLÉPHONE (INFOVOX)

Thierry Collado

IDIAP-Com 01-08

IDIAP, LE 20 DÉCEMBRE 2001

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

Abstract

The goal of this project is to develop an interactive voice server to deal with restaurants informations. The complete system will be develop as four different modules. The telephonic interface, the recognition system, the dialog management and the database. Data from the database will be accessible via either the telephone (regular or GSM) by a natural language dialog or the web.

Résumé

Le but de ce projet est de développer un serveur vocal interactif pour obtenir des informations sur des restaurants. Le système complet sera développé en quatre différents modules. Une interface téléphonique, le système de reconnaissance, un dialogue manager et une base de données. Les données de cette dernière seront accessibles soit via le téléphone (fixe ou mobile) soit via le web.

Table des matières

Chapitre 1 Introduction ..... page 4
1.1 Description du projet ..... page 4
1.2 Partenaires ..... page 4
Chapitre 2 Les modules, généralités ..... page 5
2.1 Vue d'ensemble ..... page 5
2.2 Téléphone API ..... page 5
2.3 Reconnaissance ..... page 8
2.4 Dialogue ..... page 8
2.5 Base de données ..... page 9
Chapitre 3 Démonstrateur alpha ..... page 10
3.1 Caractéristiques ..... page 10
3.2 Résultats ..... page 11
Chapitre 4 Version Béta ..... page 11
4.1 Caractéristiques ..... page 11
4.2 Résultats ..... page 12
Chapitre 5 Travaux futurs ..... page 12
Chapitre 6 Conclusion ..... page 13
6.1 Technique ..... page 13
6.2 Personnelle ..... page 13
Bibliographie ..... page 13

Tables des figures

Figure 1 Architecture du projet ..... page 5
Figure 2 Machine d'états interface téléphonique ..... page 6
Figure 3 Echange d'informations entre les modules ..... page 7
Figure 4 Système de reconnaissance ..... page 8
Figure 5 Implémentation du dialogue sur CSLU RAD toolkit ..... page 9
Figure 6 Tables de base de données ..... page 10
Figure 7 Etat final ..... page 12
Tableau 1 Taux de reconnaissance ..... page 11

Chapitre 1 Introduction

1.1 Description du projet

Ce travail fait l'objet d'un projet CTI (CTI 4247.1) du nom d'Infovox (Interactive Voice Servers for Advanced Computer Telephony Applications) géré par la compagnie start-up de l'IDIAP ; VOXaccess (§1.2). Les buts de ce projet sont:

- Du point de vue scientifique, approfondir la recherche et le développement d'IVR (Interactive Voice Response), avec applications pour des serveurs téléphoniques pilotés par ordinateur. En règle générale, ce projet revient à tester les outils "states-of-the-art" actuels concernant la reconnaissance automatique de la parole et le traitement naturel du langage (dialogue) pour accéder à des informations situées dans de larges et complexes banques de données, et d'intégrer cette technologie dans une application téléphonique gérée par ordinateur. Plus précisément, il s'agit de mettre en place un système d'information sur les restaurants de Martigny par téléphone (portable ou fixe) ou via une interface web.

- Du point de vue technique il s'agit ici d'aider la start-up VOXaccess de mettre en place ces outils avec un souci de réutilisation pour la recherche, le développement et les programmes écrits, afin de pouvoir rapidement développer d'autres applications sur cette même technologie (ex.: informations sur les cinémas, les horaires de train,...)

1.2 Partenaires

Le projet Infovox c'est développé au sein d'un consortium composé de 5 partenaires:

- IDIAP: Institut Dalle Molle d'Intelligence Artificielle Perceptive. Créé en 1991, c'est institut basé à Martigny a vu sa reconnaissance nationale comme internationale grandir sans cesse. Cette année il a d'ailleurs été choisi comme pôle national de recherche. Ces domaines d'activités sont la reconnaissance de la parole et vérification du locuteur, reconnaissance de forme et analyse de mouvement, l'apprentissage automatique et l'analyse de données.
- EPFL (LIA): Laboratoire d'Intelligence Artificielle de l'Ecole Polytechnique Fédérale de Lausanne. Recherche et développement dans le domaine d'interfaces à langage naturel.
- Swisscom: Entreprise de télécommunication. Travaille dans le développement de solution pour les services digitaux intégrés (ISDN, Integrated Service Digital Network) ainsi que des systèmes a reconnaissance de la parole robuste.
- VOXaccess SA: Compagnie start-up de l'IDIAP, travaille essentiellement dans l'intégration de reconnaisseur de la parole ainsi que de services téléphoniques automatiques.
- Omédia SA : Société basée à Martigny active dans les services via internet. Possède une bonne expertise dans les bases de données connectées à internet.

Chapitre 2 Les Modules, généralités

2.1 Vue d'ensemble

Dans une approche méthodique ce projet se décompose en modules : une interface téléphonique qui prendra en charge l'appel de l'utilisateur, un module de reconnaissance afin de comprendre sa requête, un module de dialogue pour diriger les questions devant être posées à l'utilisateur et une base de données pour y rechercher les informations nécessaires.

Dû à la complexité de la tâche, le projet a été partagé en trois groupes de travail:

1. WG-Reco : responsable de la partie reconnaissance de la parole et interface téléphonique
2. WG-HMI : responsable de l'interaction Homme-Machine (dialogue, évaluation, etc,..)
3. WG-DB : responsable du développement des bases de données et des outils pour leurs mises à jour/consultation

La figure 1 montre l'architecture du projet avec l'assignation des tâches aux différents groupes de travail.

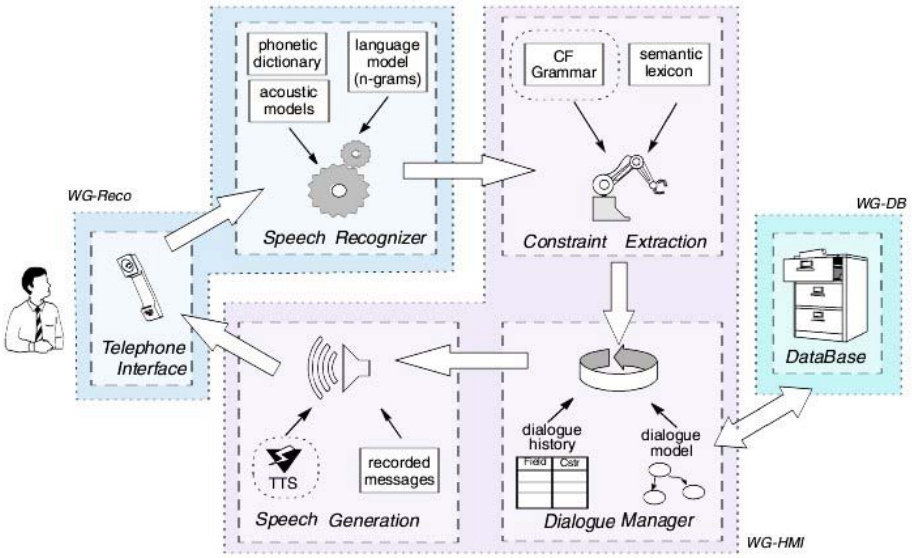

Figure 1 Architecture du projet

2.2 Téléphone API

L'application téléphonique utilise une carte analogique de la marque Dialogic sur plate-forme Windows NT. Le choix de ce matériel a été fait après discutions avec Swisscom et en comparant avec d'autres solutions proposées sur le marché. En effet, bien que toutes les applications utilisées jusqu'alors à l'IDIAP étaient sous système SUN, le nombre de solutions nouvelles pour ce système s'est révélé insuffisant, c'est pourquoi le choix d'utiliser le système NT s'est imposé. Le choix d'une carte analogique au lieu d'un modèle digital (ISDN) provient du niveau budgétaire, la seule carte ISDN proposée par Dialogic au début de projet était une carte d'accès primaire (32 canaux) qui dépassait largement les besoins nécessaires.

Deux types de cartes ont été acquises pour mener à bien le projet :

- Deux cartes deux canaux D-21H. (une pour l'EPFL, l'autre pour l'IDIAP)

- Une carte quatre canaux full duplex D-41ESC. La particularité de cette carte est de pouvoir associer à un canal la capacité de traitement (DSP, Digital Signal Processing) d'un autre, ce qui lui confère sa caractéristique full duplex. Cette particularité est due à la présence sur la carte d'un bus Dialogic dédié appelé SC-Bus et les routines soft pouvant le gérer.

Le code développé pour cette API est en C et utilise largement les fonctions proposées dans les librairies Dialogic (libdxxmt.lib, libsrlmt.lib), afin de tirer meilleur profit du matériel. Le programme ainsi développé s'apparente à une machine d'états, suivant pas a pas une conversation: enregistrement d'un utilisateur, diffusion d'une réponse, analyse de fin de dialogue,… La voici imagée (figure 2):

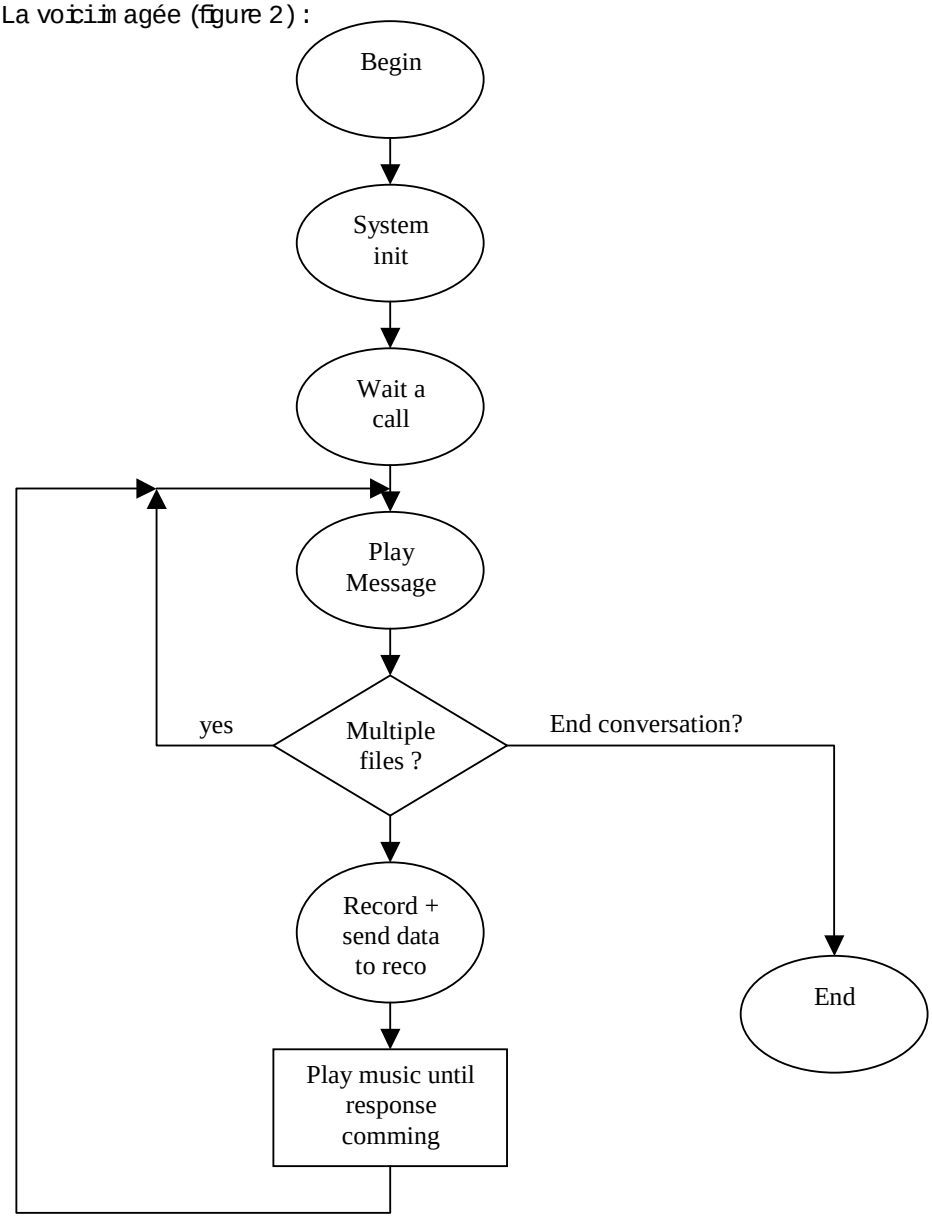

Figure 2 : Machines d'états interface téléphonique

De plus, en tant que point d'entrée du système, c'est dans le code de cette API qu'est effectuée la synchronisation avec les modules reco et dialogue, par l'échange de fichiers (données/acquittements) mais aussi de pipe vers la reco dans la dernière version. La figure suivante (3) montre cet échange d'informations :

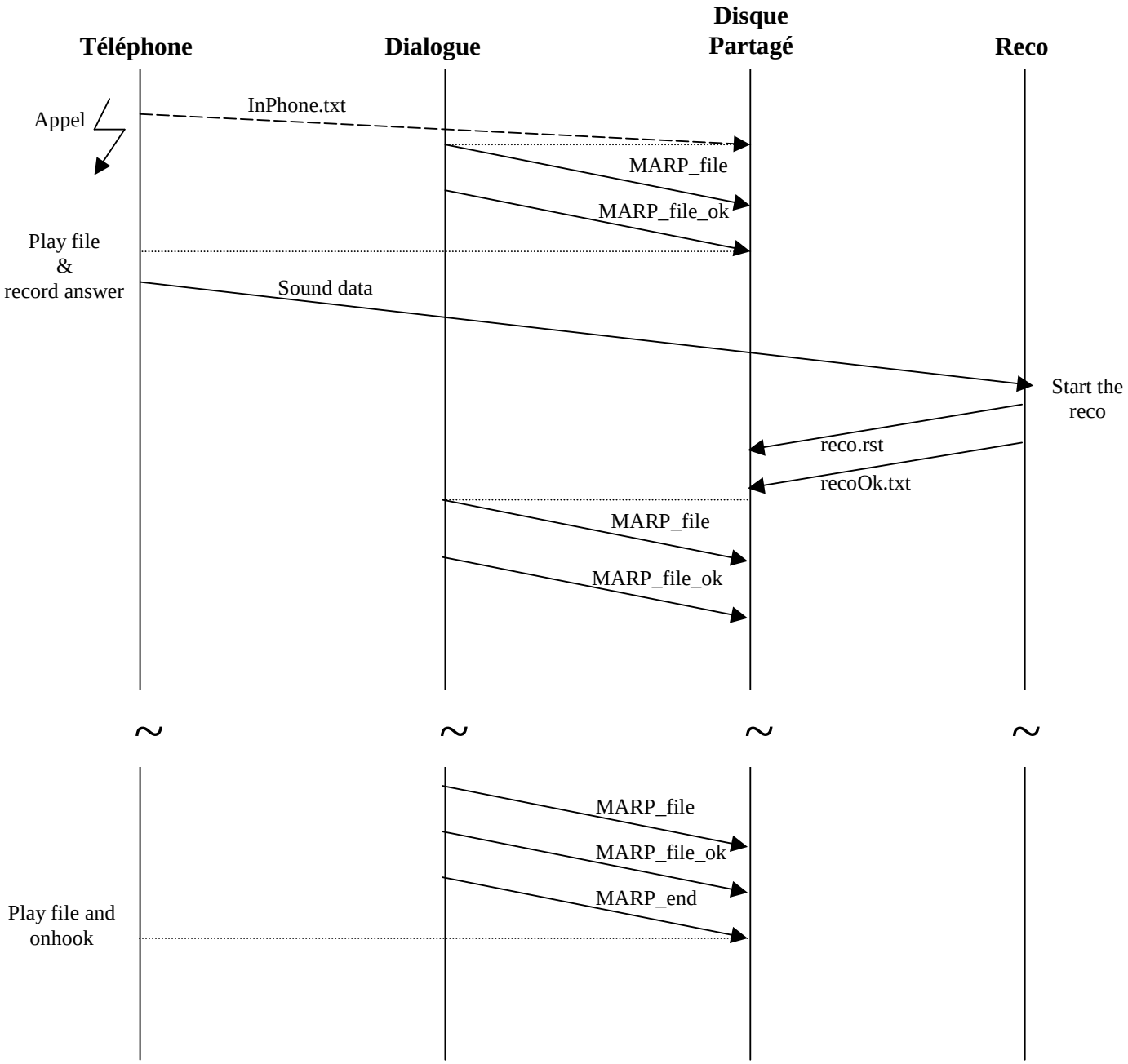

Figure 3 : Echange d'informations entre les modules

Sur cette précédente image nous remarquons qu'après chaque envoi de fichiers de données, un fichier d'acquittement est déposé sur le disque. Dès la détection de ce fichier faite par le module concerné, celui-ci traite le fichier d'informations et efface les fichiers sur le disque. Nous remarquons aussi que les données pour la reco sont envoyées directement vers le module de reconnaissance sans passé par fichiers, mais directement par des buffers.

Les particularités mises en place pour l'enregistrement des utilisateurs sont:

- détection de silence (arrête l'enregistrement après x-sec de silence)
- DTMF, touche 1 pour arrêter l'enregistrement spontanément
- DTMF, touche # pour arrêter la diffusion d'un message et commencer l'enregistrement d'une réponse/demande. Cette touche peut être utile pour des utilisateurs expérimentés.

2.3 Reconnaissance

La partie reconnaissance se base sur un système de reconnaissance à vocabulaire moyen (Suisse-Français), parole continue et indépendant du locuteur. Ce reconnaisseur est basé sur un système hybride HMM/ANN (Hidden Markov Model, Artificial Neural Network). Le système employé est visible à la figure 4 :

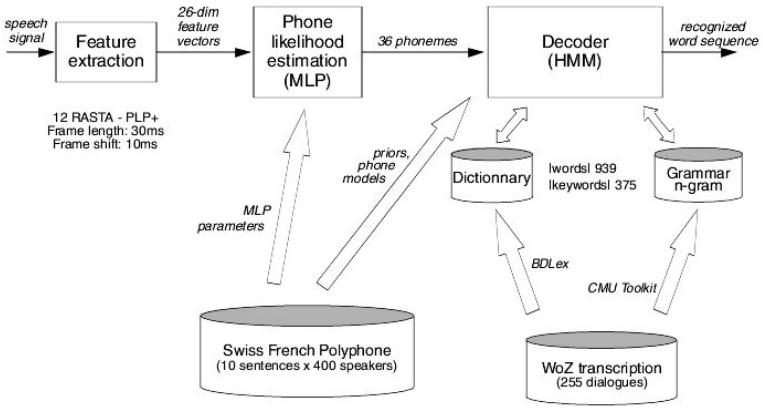

Figure 4 Système de reconnaissance

Celui-ci a été entraîné sur la base de données polyphone, enregistrée sur le réseau téléphonique suisse (analogique et digital) et regroupant un grand nombre d'utilisateurs prononçant en français toute une série d'informations comme par exemple des appels au 111.
Pour plus de détails sur ce module se référer au rapport IDIAP-COM 01-10 de Mme Hayian Wang [(1)]

2.4 Dialogue

Le but de ce module est de pouvoir diriger le dialogue avec l'utilisateur concernant la recherche d'informations sur les restaurants. Ce module observe donc quelles informations l'utilisateur lui a déjà fourni (grâce aux résultats de la reconnaissance) et pose les questions nécessaires (grâce à des messages pré enregistrés) afin d'obtenir suffisamment de renseignements pour pouvoir fournir une réponse adéquate aux vœux de l'utilisateur.

Ce dialogue s'apparente à une machine d'états reconfigurable à plusieurs sorties. Un premier test pour définir le dialogue (questions à poser, réactions des utilisateurs) à été mené (WoZ, Wizard-of-Oz). Cette expérience à permit d'enregistrer 255 dialogues et d'affiner la tactique question-réponse désirée. L'étape suivante fut d'implémenter une première version du dialogue en utilisant un programme de développement rapide spécialisé : le CSLU RAD toolkit (figure 5 : résultat de l'implémentation). Cette première version a permis de faire des tests intensifs pour tester la qualité du dialogue et améliorer la maîtrise et le naturel de ce dernier.

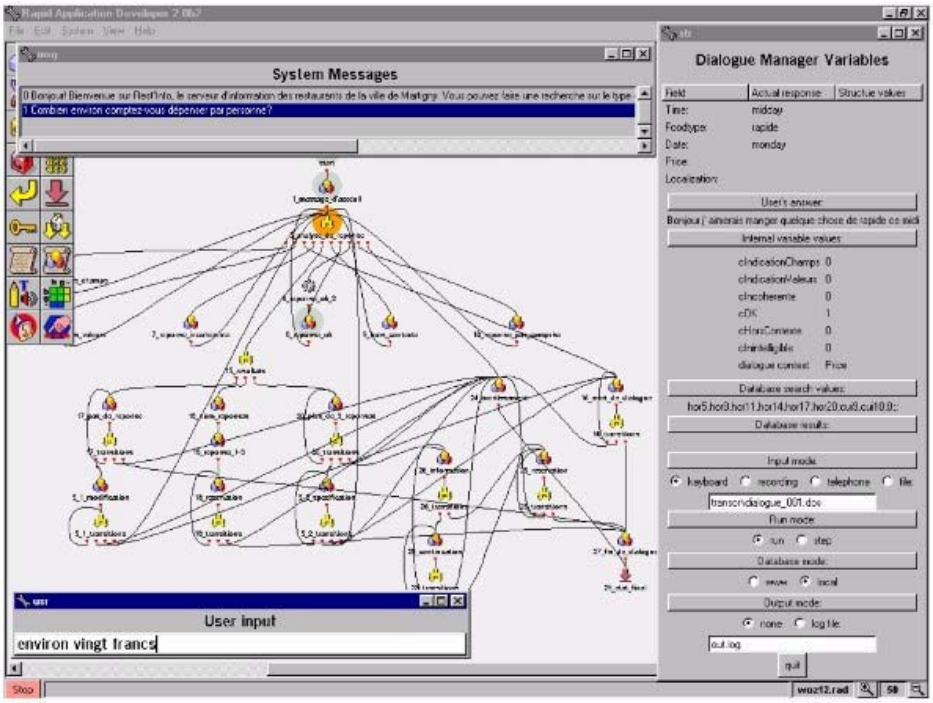

Figure 5 Implémentation du dialogue sur CSLU RAD toolkit

Après ce prototypage rapide, le dialogue a été implémenté en quasi-temps réel en langage C++.

2.5 Bases de données

La base de donnée doit dans le cadre de ce projet pouvoir répondre aux contraintes suivantes: mises à jour régulière et aisée, accès simultané par support différent (voix/web), mécanisme de stockage unifié pour les données internes et externes.

De pars sa simplicité de mise en œuvre et faisant partie d'un standard dans le domaine, la base de donnée MySQL[2] a été choisie. Une API web utilisant le langage PHP[3] a été faite pour permettre les mises à jours et les consultations de la base.

Les informations rentrées dans ces bases regroupent la totalités des informations sur les restaurants de la ville de Martigny (51) telque type de cuisine, horaires, prix moyen d'un repas, localisation dans la ville, adresse, numéro de téléphone...

L'image ci-dessous (figure 6) représente une partie des tables implémentées dans la base de données :

Figure 6 Tables de base de donnée

Chapitre 3 Démonstrateur alpha

3.1 Caractéristiques

Le premier démonstrateur a tout d'abord permis aux différents groupes de travail de se familiariser plus amplement avec le travail à faire et les outils pour le réaliser. Il a fallu dans un premier temps procéder à l'installation du hardware dialogic sur une station Windows NT et les librairies STRUT[4] pour le module de reconnaissance dans une station SUN (plates-formes sous laquelle les outils de reconnaissance développé a l'IDIAP fonctionne)

Comme mentionné, le démonstrateur est partagé entre deux machines :

- une station NT qui gère l'interface téléphonique
- une station SUN qui gère le module de reconnaissance et celui du dialogue
- la base de données est indépendante des plates-formes

Afin que les modules puissent se synchroniser et échanger leurs informations (fichiers son de l'utilisateur, résultat de la reconnaissance, fichier à diffuser), un disque "partagé" entre les systèmes SUN-PC (accès en lectures et écritures) à été mis en place. En fait il s'agit d'un répertoire disque sur station SUN que l'on rend visible par NT avec des droits utilisateurs forcés en écriture pour tous grâce au logiciel Samba, qui émule le protocole de disque SMB de Windows sur les disques SUN.

Le module de reconnaissance est codé en tcl faisant appel à plusieurs scripts et à l'environnement STRUT. Un dictionnaire (mots avec leur transcription phonétique) spécifique à l'application à été crée ainsi qu'une analyse fréquentielle du regroupement de ces mêmes mots pour obtenir la sémantique (LM). Le son (parole de l'utilisateur) est caractérisé par 13 coefficients rasta[5] (12 de bases et 1 d'énergie) de l'environnement STRUT. L'étape suivante est l'interprétation des coefficients via un réseau de neurones pour en obtenir des phonèmes.

Ces phonèmes seront ensuite regroupés selon des probabilités issues du modèle de langage pour générer des mots, puis des phrases.

Afin de rendre le reconnaisseur plus robuste, une étude[6] sur l'effet de la dégradation du signal dû au réseau téléphonique a été faite grâce à un simulateur de lignes téléphoniques.

3.2 Résultats

L'utilisation du démonstrateur s'est révélée assez restrictive (ex. : pas de possibilité pour l'utilisateur de raccrocher à n'importe quel moment). De longs temps d'attente pour permettre au reconnaisseur de faire son travail se font aisément sentir, notamment dû à l'absence de diffusion de musique pour faire patienter l'utilisateur. On a pu relever également de nombreux bugs relatifs aux réinitialisations du système et à la re synchronisation entre les différents modules (fichiers latent sur le disque partagé…)

Pour évaluer les performances du reconnaisseur seul le nombre de mots ou mots clefs (liste de mots considérés comme important dans le contexte de l'application) reconnus a été pris en compte. Le tableau suivant présente brièvement les résultats :

mean results (%), all words					mean results (%), keywords				
good	sub	del	in	err	good	sub	del	in	err
57.4	28.6	14.0	3.7	46.4	69.5	10.7	19.8	4.2	34.7

Table 1 Taux de reconnaissance

Ces résultats expriment encore quelques difficultés du reconnaisseur à différencier certains mots. On a aussi remarqué qu'il était très sensible à la vitesse et à l'élocution de la parole. Le taux de reconnaissance obtenu, de l'ordre de 58% , peut être comparé à celui d'autres projets semblables :

- LIMSI RailTel : 82% de reconnaissance sur les mots, (base de 800 mots, incl. 58 noms de station)
- CSELT Dialogos (railinfo) : 68,2% de reconnaissance sur les mots (sur 3471 mots, incl. 2983 noms de villes)
- Italian Sundial (railinfo) : 53,4% de reconnaissance sur les mots (base de 800 mots)

Chapitre 4 Version Beta

4.1 Caractéristiques

L'amélioration significative apportée de nos jours au projet et que celui-ci tourne maintenant sur une seule plate-forme. En effet, les modules de reconnaissance et de dialogue ont été modifiés afin de pouvoir fonctionner sur une station NT (adaptions code, little-big indian, réécriture partielle ou totale du code,…), ceux-ci étant dirigés directement au sein du code de pilotage de l'interface téléphonique.

Un grand effort a été appliqué au module de reconnaissance. Bien que gardant les même caractéristiques de fonctionnement (coefficient rasta, …) celui-ci ne s'appuie plus sur le système STRUT. En résulte un gain appréciable en temps. Un effort à aussi été apporté à une meilleure modélisation du dictionnaire.

L'interface téléphonique a aussi subit plusieurs améliorations pour la rendre plus "naturelle" : diffusion de musique lors d'attente, acceptation des raccrochages intempestifs, traitement multi-thread d'informations,…

Le dialogue a subit quelques raffinements en regard aux questions posées et grâce au nombreux tests effectués dans les phases de développements.

Une interface web rudimentaire pour la recherche de restaurants sur Martigny a aussi été mise en place. Cette interface permet d'avoir accès aux données des restaurants de façon simple.

4.2 Résultats

Le taux de reconnaissance n'a pas évolué significativement, cependant le dictionnaire a été amélioré et la robustesse du module augmentée. Ceci allié à une nette augmentation de la rapidité et d'une interface téléphonique plus naturelle permet à l'utilisateur de se sentir à l'aise avec le démonstrateur et d'obtenir rapidement résultat à ses requêtes.

En règle générale, il est maintenant aisé grâce à ce système d'obtenir rapidement une information (adresse, nom,…) sur les restaurants de Martigny, cependant le système est encore trop dépendant du locuteur (certains utilisateurs obtiennent réponse a leur requêtes dans l'immédiat, alors que d'autres n'y parviennent jamais)

Chapitre 5 Travaux futurs

L'étape suivante est naturellement de passer de l'étape de prototype à celle de produit. Pour ce faire, il y a encore plusieurs point qu'il faudra améliorer :

- acquérir plus de données audio afin de rendre le lexique plus riche et améliorer le modèle de langage
- éventuellement mettre en place un système de confirmations de réponses au sein du dialogue (impacte à mesurer sur les utilisateurs)
- augmenter le contenu des bases de données (menus,…)
- mise en place d'une interface web agréable.
- acceptation de plusieurs utilisateurs simultané (plusieurs lignes téléphoniques, multiplication des processus de reconnaissance, dialogue,…)
- packaging : amélioration graphique, portabilité, mise en œuvre et installation : faciliter toutes ces taches pour un client potentiel.

Voici une représentation du projet en son état final :

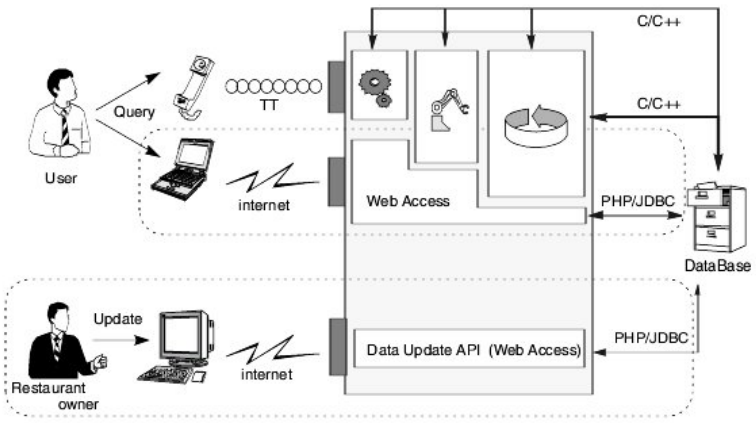

Figure 7 Etat final

Chapitre 6 Conclusion

6.1 Technique

Ce projet a permis à différents partenaires de démontrer leur savoir-faire et améliorer leurs techniques et connaissance grâce à l'apport des autres au sein des groupes de travail.

Bien que les informations sur les restaurants soient accessibles part l'interface vocale (téléphone fixe ou mobile) ou l'interface web, de nombreuses améliorations doivent encore être apportées au projet pour que celui-ci puisse être réellement appliqué dans un produit commercial.

6.2 Personnelle

Ce travail m'a permis de faire connaissance avec de nombreux domaines intéressants, académiques mais aussi appliqués tel que le travail sur la carte Dialogic.
Grâce à ce projet j'ai pu évoluer au sein d'un groupe, tout en étant indépendant et entreprenant, devant faire face à de nombreux outils nouveaux mais respecter un cahier des charges précis et un plan de travail.

Bibliographie

(1) Rapport de Haiyan Wang, IDIAP-COM 01-10, "Speech Recognition Engine for Interactive Voice Response application on Windows".
(2) www.mysql.org
(3) www.php.net
(4) STRUT User's Guide, http://tcts.fpms.ac.be/asr/strut/users-guide/html/users-guide.html
(5) RASTA Processing of speech, H. Heransky, N. Morgan, IEEE transactions on speech and audio processing, vol 2, no 4, October 1994
(6) ICSLP 00408, S. Möller, H. Bourlard, Real-Time Telephone Transmission Simulation