Le projet MTM – Reconnaissance de la parole et du locuteur sur une plateforme embarquée

Loïc Lefort, Teva Merlin, Jean-François Bonastre, Pascal Nocera

Laboratoire Informatique d'Avignon
339, ch. des Meinajariés – BP 1228 Agroparc
84000 AVIGNON Cedex 9
{loic.lefort,teva.merlin,jfb,pascal.nocera}@lia.univ-avignon.fr

RÉSUMÉ

This paper presents integration of speech technologies into an embedded platform. This work is part of the MTM project, funded by the European Community, which consists in designing a new Personal Digital Assistant offering UMTS connectivity and extended multimedia capabilities. Among the project goals is the ability for the applications to feature speech recognition and speaker recognition as part of the human interface. Speech and speaker recognition systems have been developed, capable of functioning in both local (on the PDA) and remote (client/server) modes. Software interfaces have been developed to offer access to these technologies for easy integration into the PDA applications.

1. INTRODUCTION : LE PROJET MTM

1.1. Objectifs

Le projet MTM [1], financé par la Communauté Européenne dans le cadre du programme IST[1], implique plusieurs partenaires industriels et universitaires issus de cinq pays européens. L'objectif annoncé est de définir et produire un terminal multimédia combinant les fonctions d'un assistant numérique personnel (PDA) et d'un téléphone sans fil, intégrant une caméra vidéo. Ce terminal représente une plateforme de démonstration de nouvelles applications basées sur des technologies émergentes comme la voix et la vidéo sur IP, l'accès Internet sans fil à haut débit (UMTS) et la reconnaissance vocale.

1.2. Plateforme actuelle

À l'heure actuelle, les développements se font sur un PDA déjà disponible sur le marché, le Compaq iPAQ H3600 [2]. Cet appareil est doté d'un processeur Intel StrongArm SA-1110 à 206 MHz, de 16 Mo de mémoire flash et de 32 Mo de RAM. L'iPAQ peut être étendu via une carte CompactFlash ou un port PCMCIA. Une carte réseau sans fil (802.11) PCMCIA a été utilisée pour simuler la liaison UMTS. Le projet MTM inclut le développement d'un module d'extension intégrant deux ports PCMCIA, un lecteur de cartes magnétiques et des boutons supplémentaires.

L'iPAQ est vendu avec Windows CE de Microsoft, mais ce dernier a été remplacé par un système Linux, offrant plus de souplesse.

[1]Projet MTM IST 1999-11100

1.3. Applications

La plateforme MTM doit être aussi ouverte que possible. C'est cet objectif qui a guidé le choix d'un système d'exploitation Linux. Tout utilisateur doit pouvoir ajouter ses propres applications au terminal.

Le projet MTM comprend le développement de plusieurs applications :

- **Télé-médecine :** L'application de télé-médecine *Mobile Chili* [5] a été développée à destination des radiologues, dans le but d'assurer une meilleure communication entre plusieurs spécialistes. Par exemple : un radiologue pourrra, en cas d'urgence, recevoir à distance des images en provenance de l'hôpital. Il sera en mesure de visualiser ces données et d'en faire une première analyse. Le radiologue pourra répondre, renvoyer un rapport, un diagnostique ou des instructions pour l'équipe médicale.
  Cette application prévoit également une assistance aux médecins lors des visites aux patients.
- **"Easy city guide" :** L'objectif est la fourniture d'un service d'information, utilisable tant par les résidents que par les touristes, optimisant et facilitant l'accès aux informations et services fournis par la municipalité (plans, visites guidées de musées, etc.).
- **Apprentissage à distance :** Cette application permet aux personnes éloignées de poursuivre leur formation scolaire ou universitaire. Elle fournit également un système de visioconférence.

2. LES TECHNOLOGIES VOCALES DANS LE CADRE DU PROJET MTM

Les technologies vocales trouvent une utilisation au sein de toutes les applications du projet MTM, contribuant à améliorer leur interface utilisateur. *Chili*, l'application de télé-radiologie, tire profit à la fois de la reconnaissance de la parole, pour commander l'interface, et de la reconnaissance du locuteur, comme système d'authentification supplémentaire. L'application "easy city guide" et l'apprentissage à distance, dont la partie client est basée sur un navigateur web, bénéficient également de l'intégration de commandes vocales. De plus, l'apprentissage à distance peut offrir une sélection automatique de profil d'utilisateur via la reconnaissance du locuteur.

Notre rôle dans le projet MTM a consisté à fournir une implémentation des technologies de reconnaissance de la parole et de reconnaissance du locuteur sur la plateforme MTM, ainsi qu'à définir une interface de programmation (API) offrant un accès simple à ces technologies.

Les besoins exprimés pour la reconnaissance de la parole consistaient en un reconnaisseur de mots isolés, basé sur des modèles indépendants du locuteur et de la langue. D'autre part, toute application devait être capable de modifier dynamiquement le vocabulaire utilisé par ce reconnaisseur, y compris par l'ajout de nouveaux mots.

Pour sa part, le système de reconnaissance du locuteur devait remplir des fonctions correspondant aux tâches classiques de *vérification* et d'*identification* du locuteur. La première, consistant à contrôler la similarité de la voix de l'utilisateur et de celle d'un locuteur précédemment enregistré, répond au besoin d'un système d'authentification complémentaire. L'identification du locuteur, permettant de retrouver l'identité de l'utilisateur dans une base de locuteurs connus, offre aux développeurs d'augmenter le confort de leurs applications en y intégrant une détection automatique de l'identité de l'utilisateur.

3. SOLUTIONS DÉVELOPPÉES POUR LE PROJET MTM

3.1. Bibliothèque de reconnaissance de la parole et du locuteur

Cette bibliothèque met à disposition des développeurs d'applications des outils de reconnaissance de la parole et du locuteur.

Pour la partie reconnaissance de la parole :
- un système de reconnaissance de mots isolés ; les développeurs peuvent définir leur propre vocabulaire et le changer dynamiquement durant l'exécution de l'application (les développeurs peuvent ainsi offrir à l'utilisateur la possibilité de définir son propre vocabulaire) ;
- un outil automatique de transcription phonétique à partir du signal de parole afin de faciliter l'ajout de nouveaux mots par le développeur et/ou l'utilisateur.

Pour la partie reconnaissance du locuteur :
- possibilité d'apprendre de nouveaux modèles de locuteurs ; les modèles peuvent être améliorés par la suite lors de nouvelles séances d'apprentissage – ceci permet une première session d'apprentissage plus courte et plus conviviale ;
- vérification du locuteur ;
- identification du locuteur ; la bibliothèque permet une identification en milieu ouvert – i.e. le locuteur courant peut ne correspondre à aucun locuteur de la base des locuteurs enregistrés ;
- des fonctions simples pour manipuler les bases de locuteurs et faciliter ainsi l'implémentation de l'identification du locuteur.

Deux versions de la bibliothèque sont disponibles. La première est une version "locale", où les systèmes de reconnaissance s'exécutent sur la même machine que l'application. La deuxième offre un service d'exécution des processus de reconnaissance sur une machine distante, permettant ainsi aux applications de fonctionner sur des machines moins puissantes.

3.2. Reconnaissance de la parole

L'approche retenue pour le système de reconnaissance de la parole est basée sur des HMM avec des modèles de phonèmes. Ce choix a été préféré à des approches plus classiques dans le cadre d'un système de reconnaissance de mots isolés à petit vocabulaire, basées sur l'algorithme de DTW (*Dynamic Time Warping*) ou sur des HMM avec des modèles de mots, car celles-ci se révèlent inadaptées aux fonctionnalités requises par les utilisateurs, qui imposent des modèles acoustiques indépendants du locuteur, de la langue et du vocabulaire. Les HMM avec modèles de phonèmes offrent la souplesse nécessaire, notamment pour l'aspect dynamique du vocabulaire.

Nous avons choisi pour la partie décodage l'algorithme de Viterbi synchrone avec *beam pruning*, suffisant pour un système de reconnaissance de mots isolés à petit vocabulaire. Nous avons en revanche conservé la modélisation acoustique de notre système de reconnaissance de la parole continue grand vocabulaire existant, *Speeral*. Elle utilise des HMM avec des distributions de probabilité continues et une topologie de Bakis standard avec 3 états par phonème. Néanmoins, contrairement à *Speeral*, le reconnaisseur MTM n'utilise pas de phonèmes contextuels. La modélisation des phonèmes est réalisée avec des mixtures de gaussiennes diagonales avec 64 gaussiennes par mixture.

Le choix de l'ensemble de phonèmes à modéliser s'est aussi posé car le système devait proposer des modèles acoustiques indépendants des vocabulaires, ceux-ci pouvant contenir des mots de plusieurs langues : anglais, français, italien, allemand et espagnol. Nous avions besoin d'un ensemble de phonèmes couvrant toutes ces langues. Notre choix s'est en fait porté sur le jeu de phonèmes français car il semblait couvrir en grande partie nos besoins, et de plus ne disposions pas de corpus suffisant pour les autres langues. Nous verrons dans la section 4.1 quelques résultats obtenus avec ces modèles sur un vocabulaire anglais.

L'apprentissage des modèles acoustiques a été réalisé à l'aide du corpus français de parole lue BREF [8]. Plusieurs méthodes ont été utilisées pour compenser la perte de performances due aux différences de microphone entre le corpus d'apprentissage et le PDA. La première étape a été de réduire les vecteurs acoustiques à une moyenne nulle et à une variance unitaire. Ces paramètres centrés/réduits (CR) ont augmenté de manière significative les performances du système. Plusieurs étapes d'adaptation au microphone du PDA ont ensuite été appliquées, en utilisant MAP [7] et MLLR [9] sur un corpus plus réduit enregistré sur le PDA. Les résultats obtenus sont décrits dans la section 4.1.

3.3. Reconnaissance du locuteur

La technologie de reconnaissance du locuteur intégrée au projet MTM est issue du savoir-faire acquis dans ce domaine par le LIA lors du développement du système AMIRAL de reconnaissance et d'indexation selon le locuteur [6].

Le système utilisé ici est indépendant du texte et basé sur des modèles GMM [14] calculés, de manière classique, sur des vecteurs cepstraux.

Un modèle du monde a été calculé en exécutant l'algorithme EM [3] sur environ une heure de parole enregistrée sur le PDA, issue de 20 voix féminines et masculines. Ce modèle du monde sert de base au calcul des modèles de locuteurs, utilisant une variante de la technique d'adaptation

MAP, décrite dans [10]. Le problème du choix du nombre de composantes des modèles, paramètre clé des performances du système, tant en termes de précision qu'en termes de vitesse, est traité dans la section 4.2.

Pour la phase de test, un classique rapport entre les vraisemblances par rapport aux modèles de locuteur et du monde est calculé pour chaque trame du signal. Le score du signal complet est fourni par le rapport de vraisemblances moyen. Aucune autre forme de normalisation n'est appliquée au score, du fait de problèmes de complexité de calcul ainsi que du manque de données enregistrées sur le PDA.

4. ÉVALUATION DU SYSTÈME

4.1. Reconnaissance de la parole

Des tests ont été menés pour vérifier la qualité des modèles acoustiques après leur adaptation au PDA. Ces tests ont été réalisés à la fois avec le système de reconnaissance mots isolés du projet MTM et le système de reconnaissance de parole continue grand vocabulaire *Speeral*.

Deux nouveaux corpus ont été enregistrés sur le PDA, en anglais et en français, par 5 locuteurs français. Le corpus anglais contient 135 mots issus d'un vocabulaire de 9 mots et le corpus français contient 375 mots issus d'un vocabulaire de 25 mots. Les taux d'erreurs obtenus sont de 12% pour l'anglais et de 2% pour le français. Les résultats font apparaître un net gain entre l'utilisation des vecteurs acoustiques bruts et leur utilisation après les avoir réduits à une moyenne nulle et une variance unitaire. Ils ne montrent en revanche pas de différence significative entre les modèles acoustiques basés directement sur ces vecteurs réduits et ceux obtenus après adaptation par MAP ou par MLLR. Ce résultat peut être expliqué par la simplicité de la tâche, mais également par les vocabulaires choisis pour ce test qui ne contenaient pas de mots phonétiquement voisins.

Une comparaison des différents types de modèles acoustiques a donc été réalisée sur une tâche plus complexe, en utilisant cette fois le système *Speeral*. Le corpus de test que nous avons utilisé ici est un sous-ensemble de 180 phrases issues de la tâche d'évaluation en français ARC B1 [4], lues par 11 locuteurs et enregistrées sur le PDA. Notre référence pour le système *Speeral* avec des modèles acoustiques non contextuels sur la tâche ARC B1 est de 22% de *Word Error Rate* (WER). Les résultats obtenus font cette fois apparaître l'intérêt de l'adaptation des modèles acoustiques au microphone du PDA. Ces résultats sont résumés dans le tableau 1.

TAB. 1: Résultats obtenus avec *Speeral* sur de la parole enregistrée sur le PDA (CR signifie que les vecteurs acoustiques ont été réduits à une moyenne nulle et une variance unitaire)

Modèle acoustique	WER
normal	$66,2\%$
CR	$52,2\%$
CR+MLLR	$50,3\%$
CR+MAP	$43,0\%$

$$\sum \sqrt[n]{\frac{a+b}{a+c}} \qquad (1)$$

4.2. Reconnaissance du locuteur

La reconnaissance du locuteur a été testée sur les données de la campagne d'évaluation NIST 2001 [12]. Au vu du manque de données disponibles enregistrées directement sur le PDA, il a été jugé préférable de baser les tests sur une campagne d'évaluation de systèmes de reconnaissance du locuteur internationalement reconnue, utilisant une large base de données, et ce malgré la différence du type de données par rapport au PDA. Les évaluations NIST sont basées sur de la parole téléphonique spontanée, issue de conversations enregistrées en conditions réelles. Cela ne devrait pas être très éloigné de ce que les applications MTM auront à traiter.

La variation de deux paramètres a été étudiée dans le but d'établir un compromis entre les besoins de vitesse et de précision : le nombre de composantes des modèles (de locuteurs et du monde) et le pourcentage de trames effectivement traitées lors des tests [11]. Tous deux ont été réduits par rapport au système de base (GMM à 128 composantes et prise en compte de la totalité des trames lors du test). Les résultats correspondant à quelques variantes sont fournis dans les tableaux 2 et 3. Ils sont exprimés en termes de taux d'égale erreur (EER – equal error rate) obtenu pour la tâche "one speaker, different numbers, same handset" de l'évaluation NIST (tâche se rapprochant le plus des conditions trouvées dans le cadre du projet MTM).

TAB. 2: Reconnaissance du locuteur : résultats obtenus pour l'évaluation NIST 2001, en fonction de la taille de modèle utilisée.

Taille des modèles	16	32	64	128
EER (%)	16,68	15,17	13,87	13,40

TAB. 3: Reconnaissance du locuteur : résultats obtenus avec des mixtures de 64 gaussiennes, en fonction du pourcentage de trames utilisées lors des tests.

Trames utilisées	Toutes	1/4	1/8	1/12
EER (%)	13,87	13,95	14,03	14,77

Au vu de ces résultats, un système basé sur des GMM à 64 composantes, ne traitant qu'une trame sur 8 lors du test, paraît être le compromis le plus satisfaisant. Ce choix n'induit qu'une légère hausse du taux d'erreur tout en réduisant la complexité de calcul d'un facteur 16 par rapport au système de base.

5. EXEMPLE D'APPLICATION – NAVIGATEUR INTERNET

Comme plusieurs applications du projet MTM ("Easy city guide", apprentissage à distance) utilisent un navigateur internet comme interface utilisateur, nous avons choisi de développer un navigateur commandé à la voix comme exemple d'utilisation de la bibliothèque de reconnaissance de la parole MTM.

Ce navigateur peut être commandé par un ensemble limité de mots clés pour la navigation ("haut", "bas", "page précédente", ...). Il est aussi possible de définir dynamiquement des mots clés dans la page web. Les pages web contenant des mots clés vocaux sont des pages HTML

standard, dans lesquelles les mots clés sont définis à l'intérieur d'une balise `EMBED`. Ces pages peuvent donc être utilisées à la fois dans un navigateur standard et dans le navigateur "vocal". La balise utilisée contient pour chaque mot clé une transcription phonétique (incluant les alternatives) et l'adresse du lien à activer. La reconnaissance est réalisée sur un vocabulaire composé des mots clés statiques et dynamiques.

Un *plug-in* pour le navigateur *Netscape* a été développé et un navigateur spécifique pour PDA (*Konqueror Embedded* [13]) a été modifié pour reconnaître et implémenter la balise vocale.

6. CONCLUSION

Nous avons développé dans le cadre du projet MTM des systèmes de reconnaissance de la parole et du locuteur ainsi qu'une bibliothèque permettant une intégration facile de ces technologies au sein des application. Cette bibliothèque a été conçue pour répondre aux besoins exprimés par les différents partenaires du projet.

Le système de reconnaissance de la parole développé permet à l'utilisateur de contrôler les applications à l'aide d'un nombre limité de mots clés. Il est possibile d'ajouter de nouveaux mots à la volée en utilisant une transcription phonétique indépendante du locuteur et de la langue. De plus, un système de transcription automatique a été développé pour faciliter cette tâche.

Pour la partie reconnaissance du locuteur, le système permet aux applications de vérifier l'identité des utilisateurs ou de sélectionner automatiquement le profil utilisateur en utilisant l'identification du locuteur.

Il n'est pas encore possible de faire fonctionner les moteurs de reconnaissance du locuteur et de la parole directement sur le PDA pour des raisons de performance. Cependant, le terminal multimédia MTM dispose de moyens de communications suffisants pour permettre une utilisation client-serveur en temps réel. Ce mode de fonctionnement distant a néanmoins un inconvénient majeur dû au modèle client-serveur qui semble en contradiction avec la nature indépendante d'un PDA. Dans le cas du projet MTM, ce n'est pas un problème puisque toutes les applications MTM nécessitent une connexion à un serveur de données.

Nos efforts vont maintenant se porter sur un déplacement progressif du processus de reconnaissance sur le PDA. Une étape dans cette direction sera l'extraction des paramètres acoustiques avant le transfert vers le serveur. Ceci permettra dans un premier temps de diminuer la bande passante nécessaire par un facteur d'au moins 5. Ensuite, la partie la plus importante du travail sera de transférer le reste de la reconnaissance sur le PDA.

Il est évident qu'une autre voie de développement intréressante sera la mise en place d'une intégration plus poussée entre les modules de reconnaissance de la parole et de reconnaissance du locuteur. Il est en effet possible à l'heure actuelle de détecter l'identité du locuteur en même temps que se fait le décodage de la commande qu'il a prononcée, mais cette information n'est exploitée que par les applications. Le système de reconnaissance de la parole lui-même ne l'utilise pas, étant basé sur des modèles acoustiques indépendants du locuteur (l'adaptation qui leur a été appliquée est une adaptation au microphone uniquement). Le passage à des modèles acoustiques adaptés au locuteur permettrait de tirer profit de cette information pour sélectionner à la volée les bons modèles.

Enfin, plus de tests doivent être réalisés, en particulier des tests en conditions réelles avec les utilisateurs et les applications MTM.

RÉFÉRENCES

[1] MTM Consortium. Project multimedia terminal mobile (MTM). *www.mtm-project.com/*.

[2] Compaq Computer Corporation. iPAQ Pocket PC. *www.compaq.com/products/handhelds/pocketpc/*.

[3] D. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via EM algorithm. *J. Roy. Stat. Soc.*, 39 :1–38, 1977.

[4] J.-M. Dolmazon, F. Bimbot, G. Adda, M. El Beze, J.-C. Caerou, J. Zeiliger, and M. Adda-Decker. ARC B1 - organisation de la première campagne AUPELF pour l'évaluation des systèmes de dictée vocale. In *JST97 FRANCIL*, Avignon, FRANCE, Apr. 1997.

[5] U. Engelmann, A. Schröter, E. Borälv, T. Schweitzer, and H.P. Meinzer. Mobile teleradiology : All images everywhere. In *CARS 2001 : Proceedings of the 15th International Congress and Exhibition*, pages 798–803, Amsterdam, 2001.

[6] C. Fredouille, J.-F. Bonastre, and T. Merlin. Amiral : a block-segmental multi-recognizer approach for automatic speaker recognition. *Digital Signal Processing*, 10(1–3) :172–197, Jan.-Apr. 2000.

[7] J.-L. Gauvain and C.-H. Lee. Maximum *a posteriori* extimation for multivariate gaussian mixture observations of Markov chains. In *IEEE transactions on speech and audio processing*, volume 2, Apr. 1994.

[8] L.F. Lamel, J.-L. Gauvain, and M. Eskenazi. BREF, a large vocabulary spoken corpus for French. In *EUROSPEECH-91*, Genoa, Italy, 1991.

[9] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech and Language*, pages 171–185, 1995.

[10] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium. Overview of the ELISA consortium research activities. In *2001, A Speaker Odyssey*, pages 67–72, Crete, Jun. 2001.

[11] J. McLaughlin and D.A. Reynolds. A study of computation speed-ups of the GMM-UBM speaker recognition system. In *EUROSPEECH-99*, volume 3, pages 1215–1218, Budapest, Hungary, 1999.

[12] National Institute of Standards and Technology. The NIST 2001 speaker recognition evaluation plan. *www.nist.gov/speech/tests/spk2001/doc/2001-spkrec-evalplan-v53.pdf*, Mar. 2000.

[13] The KDE Project. The Konqueror/Embedded web browser. *www.konqueror.org/embedded*.

[14] D.A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, pages 91–108, Aug. 1995.