



IRISA • Campus universitaire de Beaulieu • 35042 Rennes Cedex France • Tél.: +33 2 99 84 71 00 • Télécopie : +33 2 99 84 71 71 • Internet : www.irisa.fr

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

Rapport d'activité 2001

Projet CORDIAL

*Communication multimodale personne-machine à composantes
orales : méthodes et modèles*

Lannion

_____ THÈME 3A _____



Table des matières

1	Composition de l'équipe	3
2	Présentation et objectifs généraux	3
3	Fondements scientifiques	5
3.1	Panorama	5
3.2	Dialogue et modélisation	5
3.2.1	Extraction des actes de dialogue	7
3.2.2	Modèles de systèmes	7
3.2.3	Erreurs de communication	7
3.2.4	Modélisation de l'application	7
3.3	Systèmes et multimodalité	8
3.4	Apprentissage pour le traitement de l'information sonore	9
3.4.1	Apprentissage d'outils syntaxiques	10
3.4.2	Apprentissage de courbes prosodiques décrites par des structures d'arbres	10
3.4.3	Méthodologies de synthèse de parole	11
3.5	Etudes en traitement de la parole pour l'apprentissage des langues	13
4	Domaines d'applications	13
5	Logiciels	14
5.1	Ordictée	14
5.2	Dictionnaire vocal	14
5.3	Epigram	15
6	Résultats nouveaux	15
6.1	Dialogue et modélisation	15
6.1.1	Représentation logique pour le dialogue	16
6.1.2	Évaluation de systèmes de dialogue	16
6.2	Systèmes et multimodalité	16
6.2.1	Géoral tactile et référence	17
6.2.2	Ordictée	18
6.3	Apprentissage et traitement de l'information sonore	18
6.3.1	Inférence de grammaires pour le dialogue oral	18
6.3.2	Apprentissage de courbes prosodiques décrites par des structures d'arbres	19
6.3.3	Synthèse flexible de la parole	20
6.3.4	Segmentation de parole en phones, inférence phonologique	21
6.4	Intégration des technologies de la parole pour l'enseignement des langues	22
6.4.1	Dictionnaire vocal	22
6.4.2	Synthèse du breton à partir du texte	22
6.4.3	Correcteur de prosodie	22

7 Contrats industriels (nationaux, européens et internationaux)	22
7.1 Synthèse de la parole	23
8 Actions régionales, nationales et internationales	23
8.1 Réseaux et groupes de travail internationaux	23
9 Diffusion de résultats	23
9.1 Animation de la communauté scientifique	23
9.2 Enseignement universitaire	24
9.3 Participation à des colloques, séminaires, invitations	24
9.4 Accueil d'étudiants et de stagiaires	24
10 Bibliographie	24

1 Composition de l'équipe

Responsable scientifique

Jacques Siroux [professeur, IUT, (jusqu'au 15 octobre)]

Laurent Miclet [professeur, Enssat, (à compter du 15 octobre)]

Personnel Université Rennes 1

Yolande Anglade [maître de conférences, IUT¹]

Olivier Boëffard [maître de conférences, Enssat]

Arnaud Delhay [maître de conférences, IUT, depuis le 1^{er} octobre]

Marc Guyomard [professeur, Enssat]

Guy Mercier [professeur associé, Enssat]

Jean-Christophe Pettier [maître de conférences, Enssat]

Chercheurs doctorants

Sylvain Camus [bourse MENRT depuis le 1^{er} octobre]

Jacques Chodorowski [jusqu'au 30 juin]

Laurent Blin [ATER depuis le 1^{er} octobre]

Hélène François [ingénieur sur contrat Université-FT R&D]

Samir Nefti [bourse FT R&D]

Ingénieur

Yves Aubry [Stage DRT en informatique de l'université du Maine]

2 Présentation et objectifs généraux

La conception et la réalisation de systèmes informatiques, tels que par exemple des serveurs ou des gestionnaires d'information, destinés à des usagers professionnels ou occasionnels, doivent intégrer de façon explicite et intentionnelle le traitement de tous les aspects de la communication personne-machine. En effet, nombreux sont les exemples d'échecs de systèmes, aux conséquences économiques souvent lourdes, dûs à une conception superficielle de l'interface gérant les interactions entre les usagers et le système. La prise en compte des nombreuses facettes de la communication personne-machine nécessite différents modèles (de l'usager, du dialogue, de connaissances...) à définir ou à adapter à partir de résultats existants.

Parmi les moyens de communication d'un système, la parole se révèle l'un des plus intéressants. Du point de vue de l'utilisateur, l'usage de la parole avec celui naturellement associé de la langue naturelle facilite la manipulation des informations fournies ou reçues du système. Des points de vue de l'utilisateur et du système, la parole s'est avérée être le plus performant des média lors de nombreuses expériences effectuées en laboratoire ou en situation réelle. Ces caractéristiques justifient les recherches effectuées sur le traitement de la parole et la progression, lente mais régulière, du nombre de systèmes commercialisés et mis à la disposition du public. Cependant les résultats encore peu fiables (sans restriction d'environnement) des techniques de reconnaissance de la parole et les difficultés de compréhension du langage naturel en parole spontanée constituent des obstacles qui limitent la mise en place d'un plus grand nombre de systèmes oraux. Il apparaît alors intéressant de rechercher des méthodes et des moyens pour pallier les problèmes posés par la reconnaissance et la compréhension

¹En congé de longue maladie.

de la parole. Pour nous, les capacités de dialogue du système, l'ajout de moyens de communication supplémentaires (multimodalité) et l'architecture même du système forment une bonne partie de ces méthodes et moyens.

Dans ce domaine, nos objectifs et activités se déclinent naturellement sur deux dimensions complémentaires : théorique et pratique. Le point de vue théorique s'attache à comprendre et modéliser les fondements de l'interaction afin d'une part de prendre en compte le maximum de phénomènes interactionnels (qu'ils soient de nature cognitive ou comportementale) et, d'autre part, de faciliter la mise au point de nouveaux systèmes. L'un des premiers objectifs est de faire en sorte qu'à partir de l'acte d'énonciation des utilisateurs on puisse faire émerger la signification contextuelle complète de l'énoncé, c'est-à-dire extraire de l'énonciation non seulement les aspects structurels de l'énoncé mais aussi le sens (sémantique) ainsi que l'intention finale de l'utilisateur.

Les obstacles à ces ambitions sont nombreux : comme on l'a déjà signalé, la reconnaissance de la parole ne fournit pas de manière fiable les messages prononcés (de plus ceux-ci peuvent faire l'objet d'altérations lors de leur émission), la structure et le sens des messages sont soumis à des déviations par rapport aux normes généralement admises et enfin, fréquemment, les intentions véhiculées n'apparaissent pas explicitement dans les énoncés. Nous cherchons donc à pallier ces problèmes en utilisant différentes approches et techniques et en nous appuyant sur les concepts intégrateurs d'actes de langage et de planification. Ces concepts servent dès à présent de fondements pour modéliser quelques-uns des principes essentiels du dialogue tels que le suivi des activités des utilisateurs sous-jacentes à l'interaction ou encore la gestion des phases principales du dialogue. Il reste néanmoins plusieurs points à développer sur et autour de ces fondements.

Il est d'abord nécessaire d'enrichir certains aspects de la modélisation afin de traiter les *incidents* de dialogue dûs aux comportements et connaissances des interlocuteurs (utilisateur et système) qui ne sont pas toujours en harmonie, ainsi que la gestion du dialogue qui ne dépend pas directement de la tâche à réaliser (gestion dite phatique). Si acte de langage et planification constituent des fondements scientifiques particulièrement appropriés et centraux pour la communication, il reste cependant à éclaircir les relations qu'ils entretiennent avec l'extérieur (*i.e.* l'application et les énonciations des utilisateurs). Il s'agit d'une part de déterminer les différents composants des actes de langage de l'énonciation dans le contexte de la reconnaissance de la parole. Pour cela, nous suivons deux approches ; la première est fondée sur les connaissances linguistiques classiques (lexicale, syntaxique, sémantique...), la prosodie et la pragmatique (notion de présupposé). La seconde vise l'utilisation de techniques d'apprentissage automatique pour faire apprendre au système à partir de corpus de phrases l'extraction des éléments pertinents ; l'avantage escompté de cette approche réside dans le fait qu'il sera possible d'adapter plus facilement le système à de nouvelles applications.

D'autre part, les mécanismes généraux liés à la planification ne permettent pas de prendre en compte de manière satisfaisante les aspects de la coopération qui proviennent du mode opératoire de celle-ci (par exemple la substitution de paramètres ou encore la modification raisonnée de valeurs de paramètres pour atteindre l'information demandée). Le but est donc de proposer une modélisation de l'application qui autorisera une exploitation optimale de ses caractéristiques dans le cadre du dialogue. Nous avons montré que l'ajout d'une nouvelle modalité (le geste par l'intermédiaire d'un écran tactile) à un système de dialogue oral améliore la qualité de l'interaction en augmentant la compétence du système. Ce constat a été effectué sur un système prototype limité ; il reste à étudier de manière plus fine le comportement des utilisateurs face à un tel système. De façon plus précise, nous étudions les moyens utilisés pour référencer des éléments de l'application, que ceux-ci soient présents ou non dans

le contexte visuel. Ces études qui doivent aboutir à une modélisation et à des propositions d'architecture de traitement, portent sur les aspects linguistiques ainsi que gestuels liés aux types des éléments désignés.

La dimension pratique recouvre trois préoccupations principales. Il s'agit en premier lieu de valoriser les activités de recherche plus fondamentales en intégrant les résultats dans des systèmes opérationnels. Cette intégration pose également des problèmes de conception et d'architecture de système. Il est en effet nécessaire de faire coexister de manière souple et efficace les modules construits sur des bases hétérogènes.

Une deuxième préoccupation concerne la promotion de l'oral et des technologies vocales. Nous visons ainsi de nouvelles applications (logiciels éducatifs par exemple) qui illustrent les avantages de l'usage de la voix. Ceci nécessite en amont des activités de recherche et de développement.

Enfin, la mise au point de systèmes oraux permet aussi de viser deux buts complémentaires : le recueil de corpus réels d'interaction personne-machine qui sont nécessaires pour affiner les connaissances sur ce sujet, et l'évaluation de système qui est importante d'un point de vue technologique et qui ne peut être étudiée que si l'on dispose d'un système.

3 Fondements scientifiques

3.1 Panorama

Résumé : *Les activités se rattachent à quatre domaines complémentaires par leurs objets d'études et leurs méthodes. Le premier domaine, dialogue et modélisation, s'intéresse au code et à la structure de l'interaction ainsi qu'aux champs d'application des systèmes. Le second concerne la multimodalité (avec une emphase pour l'oral) et les prototypes de systèmes (architecture et évaluation). Le troisième porte sur les techniques d'apprentissage et leurs applications aux données sonores. Enfin, le dernier domaine est l'utilisation des technologies vocales pour l'apprentissage des langues, avec l'application à la langue bretonne.*

3.2 Dialogue et modélisation

Mots clés : acte de langage, planification, reconnaissance de plan.

Glossaire :

acte de langage dans la théorie des actes de langage fondée par Austin^[Aus70] et développée par Searle^[Sea82], le postulat de base affirme que l'émission d'un énoncé s'assimile à l'accomplissement d'actions qui modifient les états mentaux des interlocuteurs.

plan séquence d'actions destinées à réaliser un but intentionnel.

reconnaissance de plan reconnaître un plan à partir d'une séquence d'actions observées consiste à déterminer les relations qu'entretiennent ces actions afin de déterminer les buts et suites possibles du plan en cours.

[Aus70] J. AUSTIN, *Quand dire c'est faire*, Editions du seuil, Paris, 1970.

[Sea82] J. SEARLE, *Sens et expression*, Les éditions de minuit, 1982.

Résumé : *Le projet utilise une modélisation fondée sur la notion de plan d'actes de langage. Cette modélisation prend en charge le cadre général de la communication et facilite la mise en œuvre informatique, mais ne résout pas certains problèmes comme ceux de l'extraction des actes de langages à partir des énoncés observés, de l'intégration des différentes sources d'informations et d'une mauvaise communication entre les interlocuteurs.*

L'interaction personne-machine peut être considérée comme une succession d'actions particulières – les actes de langage ^[Aus70,Sea82] (nommés dans notre contexte : actes de dialogue) – qui portent la fonction de l'action dans le dialogue (exemple : une interrogation, un ordre...) ainsi qu'un contenu propositionnel (exemple : le thème de l'interrogation). Ces actes sont aussi caractérisés par leurs conditions d'utilisation qui concernent les états mentaux des participants à l'interaction (leurs intentions, connaissances et croyances). La modélisation informatique la plus pertinente est celle d'un opérateur de plan ^[All87,Lit85] dans lequel on peut faire figurer des préconditions et contraintes d'utilisation ainsi que l'effet de l'acte.

Par exemple, l'acte de demander à autrui l'exécution d'une action peut se modéliser sous la forme :

Requérir(Locuteur, Auditeur, Action(A))
 précondition-intention : *Veut(Locuteur, Requérir(Locuteur, Auditeur, Action(A)))*
 précondition-préparatoire : *Veut(Locuteur, Action(A))*
 corps : *Croyance-mutuelle(Auditeur, Locuteur, Veut(Locuteur, Action(A)))*
 effet: *Veut(Auditeur, Action(A))*

qui peut se paraphraser par : lorsqu'un agent veut que son auditeur réalise une action *A*, il peut employer l'action étiquetée *Requérir* qui consiste à établir un consensus entre les interlocuteurs pour exécuter *A*. La réalisation de ce consensus est confiée à une autre action non décrite ici. L'ensemble des actions nécessaires à la réalisation d'un but s'appelle un plan et cette approche fait l'hypothèse que chacun des deux interlocuteurs participe à la réalisation du plan de l'autre.

Cette modélisation des actes de dialogue permet d'envisager plusieurs types de raisonnements automatiques nécessaires à la conduite d'un dialogue. Le premier concerne la compréhension contextuelle des énoncés de l'interlocuteur par un mécanisme appelé reconnaissance de plans. Cela consiste à reconstruire une partie du plan de l'interlocuteur ; cette partie, si elle est correctement identifiée, permet d'explicitier les motivations de l'interlocuteur et ses croyances. Un second traitement est destiné à calculer une réponse convenable, par un mécanisme de planification, qui tient compte, par la nature même de la modélisation, des informations déjà connues et des malentendus éventuels. Ce type de modélisation rend possible une mise en œuvre informatique dans certains cas simples, mais laisse encore ouverts des problèmes importants d'ordres divers.

[Aus70] J. AUSTIN, *Quand dire c'est faire*, Editions du seuil, Paris, 1970.

[Sea82] J. SEARLE, *Sens et expression*, Les éditions de minuit, 1982.

[All87] J. ALLEN, *Natural Language Understanding*, Benjamin/Cummings Menlo Park, 1987.

[Lit85] D. J. LITMAN, *Plan Recognition and Discourse Analysis : An Integrated Approach for Understanding Dialogues*, thèse de doctorat, University of Rochester, TR 170, 1985.

3.2.1 Extraction des actes de dialogue

Le premier problème est celui du passage des énoncés prononcés par l'utilisateur à l'acte de dialogue. Ce passage n'est pas un simple problème de transcodage. Il est en effet nécessaire de prendre en compte de manière intégrée une grande variété de connaissances (états mentaux, présuppositions, prosodie...) ainsi que des indices présents dans l'énonciation elle-même (structure syntaxique, éléments lexicaux). De plus, la forme de surface des énoncés oraux présente de nombreuses irrégularités (problèmes de performance) qui compliquent la tâche de la reconnaissance de parole ainsi que celles de la compréhension et de l'interprétation de l'énoncé. Le projet CORDIAL ne travaille pas en ce moment sur ce sujet.

3.2.2 Modèles de systèmes

Le second problème réside dans l'utilisation du formalisme de plan [3, 6] pour associer trois points de vue : celui de l'application, celui du dialogue «principal» (qui concerne les intentions de l'utilisateur vis-à-vis de l'application) et celui de la gestion du dialogue lui-même (méta dialogue et dialogue phatique). Des solutions partielles ont été proposées [Lit85], mais elles résistent mal à des applications de type manipulation d'informations (interrogation de bases de données) ou qui comportent plusieurs tâches en parallèle, ainsi qu'au traitement de certaines fonctions de gestion de communication. Une approche possible de résolution de ces problèmes peut être une modélisation multi-agents. En effet, ce cadre conceptuel permet de combiner des modèles et contextes de dialogue a priori exclusifs afin d'accroître la couverture dialogique. La problématique se déplace ainsi en partie de la modélisation du dialogue vers la modélisation de l'intégration.

3.2.3 Erreurs de communication

Le troisième problème se pose fréquemment dans toute interaction : celui d'une mauvaise communication. Chacun des deux intervenants (*i.e.* l'utilisateur humain et le système) peut en effet posséder des connaissances erronées sur l'application, sur les compétences de l'autre et sur les éléments du dialogue lui-même tels que les références employées pour désigner les objets mis en cause durant le dialogue. Une erreur concernant ces informations peut, à plus ou moins longue échéance, entraîner un échec, c'est-à-dire une impossibilité pour la machine de satisfaire l'interlocuteur. La détection et le traitement de ces erreurs nécessitent en amont une tâche de caractérisation, puis une modélisation dans le cadre de la planification.

3.2.4 Modélisation de l'application

L'application, dans un système interactif, doit se comporter comme un élément actif. Dans les systèmes actuels, la modélisation de l'application présente deux types de défaut majeurs : soit les modèles de tâche s'avèrent trop figés (plans dans les systèmes de transfert d'informations), contraignant ainsi trop fortement l'initiative de l'utilisateur, soit ils sont fondés sur des contraintes (comme dans les applications de CAO), ce qui permet une activité de l'utilisateur plus libre mais peut aussi entraîner un

manque de coopérativité pour l'aider s'il ne connaît pas la suite d'actions à accomplir pour atteindre son but. Nous pensons que la modélisation de l'application doit comporter les éléments suivants : les données et leur ontologie, les connaissances sur l'utilisation des données (modes opératoires) et l'interface avec le reste du système. Enfin, cette modélisation doit être envisagée de façon à pouvoir changer facilement d'application.

3.3 Systèmes et multimodalité

Mots clés : multimodalité, référence.

Résumé : *Pour pallier certains des problèmes dus à l'utilisation de la parole, nous étudions une modalité supplémentaire de communication, un écran tactile. Les problèmes à traiter concernent l'intégration des messages provenant des différents canaux, le traitement de la référence ainsi que l'évaluation des systèmes.*

L'utilisation des techniques vocales actuelles dans les systèmes interactifs a pour conséquence l'apparition de nouveaux problèmes et difficultés qui vont de la réalisation de logiciels complets (y compris la recherche de l'application) à la spécification d'architecture, en passant par l'amélioration de la synthèse de parole et l'introduction de la multimodalité.

La communication entre personnes est rarement monomodale : le geste et la parole sont souvent utilisés conjointement pour des raisons fonctionnelles (désignation d'éléments, fiabilité de la communication). Dans un environnement de parole, l'introduction d'une modalité supplémentaire – le geste par l'intermédiaire d'un écran tactile dans notre cas – est d'autant plus intéressante qu'elle permet de pallier les erreurs de reconnaissance de parole.

Cette adjonction fait surgir très rapidement de nouvelles difficultés. La première concerne la façon dont doivent être traitées les informations qui proviennent des différents canaux de communication : leur intégration doit-elle se faire à un niveau syntaxique, sémantique ou pragmatique ? Quel type de modélisation faut-il utiliser ? Il existe encore peu de réponses satisfaisantes à ces questions. Nous avons choisi de nous appuyer sur les travaux de M. Maybury ^[May90] destinés à un contexte différent (production d'actes communicatifs en sortie d'un système). Maybury propose plusieurs niveaux d'actes de communication (avatars d'actes de langage), ce qui permet d'intégrer à chaque niveau des informations provenant de modalités différentes. Nous reprenons ce principe (qui est cohérent avec notre modélisation du dialogue) mais en reconnaissance d'actes : les modalités tactile et parole sont traitées séparément pour fournir des actes communicatifs qui sont ensuite fusionnés pour donner des actes de langage.

La seconde difficulté porte sur le traitement de la référence, plus particulièrement dans le cadre de l'application choisie (interrogation d'une base de données géographiques et touristiques). La désignation des objets intéressants du dialogue s'effectue à l'aide du langage et du geste (pointé et tracé de zones) et tient compte du contexte applicatif (l'utilisateur peut suivre un contour d'un objet cartographique).

[May90] M. MAYBURY, « Communicative Acts for Explanation Generation », *International Journal of Man-machine studies* 37(2), 1990, p. 135–172.

Les études dans ce domaine se font en linguistique et en intelligence artificielle (représentation des connaissances). Certains linguistes ^[Van86] proposent des études très fines sur les conditions d'utilisation (approche fonctionnelle) des prépositions utilisées dans la désignation des objets. Nous pensons qu'il s'agit là de résultats intéressants que nous avons adaptés pour notre traitement syntaxique des énoncés. Du côté de l'intelligence artificielle, plusieurs modélisations des relations spatiales ont été proposées. Nous utilisons celle suggérée par l'IRIT (Toulouse) ^[Vie91] pour vérifier la cohérence sémantique des expressions référentielles dans le cadre de notre application. Ce modèle est fondé sur certaines caractéristiques (dimension, morphologie...) des éléments qui régissent l'utilisation des termes linguistiques dans les expressions référentielles (par exemple, le mot bord ne peut être utilisé qu'associé à un objet possédant deux dimensions).

L'ambition de mettre sur le marché des systèmes de dialogue doit s'accompagner d'exigences sur la qualité de l'interaction. Il faut pouvoir évaluer et comparer, dans le cadre d'applications équivalentes, différents systèmes selon plusieurs points de vue (performances de la reconnaissance, efficacité du dialogue, capacités dialogiques et langagières...) et éventuellement pour un même système, évaluer des approches différentes. Un certain nombre de métriques ont déjà été proposées^[Sun93,CS94] (exemple : longueur du dialogue, nombre de tours de parole pour la récupération d'erreurs de reconnaissance...), mais elles ne rendent pas compte de toutes les dimensions d'un système interactif. De nouvelles pistes sont actuellement envisagées dans la communauté scientifique : elles sont fondées (comme au laboratoire Clips de Grenoble) sur des aspects pragmatiques tels que la pertinence, ou bien sur un concept d'autoévaluation de système (pour tester ces différentes capacités) qui consiste à faire traiter par le système, ou par une partie de celui-ci, des fragments de dialogue qui présentent les particularités à tester tout en fournissant tous les éléments contextuels nécessaires.

3.4 Apprentissage pour le traitement de l'information sonore

Mots clés : apprentissage automatique, inférence grammaticale, synthèse de la parole, base de données sonores.

Résumé : *Cette étude a pour objectif d'élaborer des techniques d'apprentissage pour améliorer les parties amont et aval du dialogue oral : le traitement de requêtes orales et la synthèse de la parole.*

[Van86] C. VANDELOISE, *L'espace en français*, Éditions du seuil, Paris, 1986.

[Vie91] L. VIEU, *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en langage naturel*, thèse de doctorat, Université Paul Sabatier, Toulouse, 1991.

[Sun93] SUNDIAL, « SUNDIAL, Prototype performance evaluation report », *Deliverable n° D3WP8*, projet Sundial P2218, septembre 1993.

[CS94] A. COZANNET, J. SIROUX, « Strategies for Oral Dialogue Control », in : *Proceedings of International Conference on Spoken Language Processing (ICSLP) 94*, 2, p. 963–966, Yokohama, Japon, 1994.

3.4.1 Apprentissage d'outils syntaxiques

Dans la partie amont d'un système de dialogue oral, la parole est traitée par un outil de reconnaissance qui produit en général un *treillis* de mots, c'est-à-dire le tableau des hypothèses lexicales entre deux instants de la phrase. Il faut ensuite utiliser une syntaxe pour en extraire une suite unique de mots, celle dont la vraisemblance à la fois acoustique et syntaxique est la plus forte.

Cette analyse syntaxique est effectuée en général soit par un modèle formel fourni a priori par le concepteur du système, soit par un modèle statistique simple fondé sur l'enchaînement des classes grammaticales (*bi-gram*, *n-gram*), dont les paramètres sont fixés par apprentissage sur un corpus.

Il est intéressant de chercher à combiner ces deux approches, en extrayant du corpus d'apprentissage un ensemble de règles de grammaire (éventuellement probabilisées) : on profite alors des avantages de la seconde («coller» aux données d'apprentissage) et de la première (autoriser des dépendances à long terme qui reflètent une véritable structure).

Nous cherchons donc à apprendre des structures syntaxiques à partir d'exemples de phrases regroupées dans un corpus d'apprentissage. Nous nous plaçons dans le cadre du dialogue oral.

Nous nous fondons sur l'emploi des techniques d'*inférence grammaticale* [2], permettant d'extraire d'un ensemble de phrases une grammaire régulière, éventuellement probabilisée. Une étape préliminaire d'apprentissage est nécessaire : extraire du corpus des exemples (ici une liste de phrases) un ensemble de catégories grammaticales. Ensuite, il faut adapter des méthodes qui ont été en général développées en dehors d'applications comme celles que nous avons à traiter.

3.4.2 Apprentissage de courbes prosodiques décrites par des structures d'arbres

Toute phrase écrite ou orale d'un langage peut se voir sous deux aspects : un premier aspect séquentiel, comme la suite des mots qui la composent, et un second aspect hiérarchique, dans son organisation syntaxique. Pour la synthèse de la parole en général, et la prédiction de la prosodie en particulier, profiter des propriétés hiérarchiques qu'apporte implicitement la seconde structure apparaît essentiel.

Pourtant, les techniques d'apprentissage utilisées jusqu'ici pour la génération de prosodie (principalement les réseaux connexionnistes et les arbres de décision) ne permettent pas la manipulation directe de cette hiérarchie. Elles nécessitent dans un premier temps la « mise à plat » de ces informations hiérarchiques, puis leur étude en séquence, ce qui entraîne la perte des informations qu'induisait l'arborescence. Maintenir ces structures intactes au cours d'un apprentissage permettrait la conservation et l'utilisation de leurs propriétés. Pour cela, l'adaptation ou la création de nouvelles techniques d'apprentissage est nécessaire.

Nous nous basons sur deux types d'apprentissage, par *plus proches voisins* et par *analogie*, qui s'appuient sur la notion de *distances entre arbres*. L'application précise est l'apprentissage automatique de la prosodie pour la synthèse vocale de phrases du français dans le contexte du dialogue. Les données y sont des phrases, représentées principalement sous forme arborescente (structure syntaxique pure ou agrémentée d'informations phonologiques et pragmatiques). Aux feuilles des arbres sont positionnés les différents phonèmes avec leurs caractéristiques prosodiques.

3.4.3 Méthodologies de synthèse de parole

Description à long terme La partie aval d'un système de dialogue oral est constituée d'un générateur de texte produisant une suite de mots correspondant à l'énoncé du message à diffuser. Cet énoncé textuel est ensuite converti en énoncé oral au moyen d'un système de synthèse de la parole à partir du texte [BBC⁺93,TB98,Tay00,DW95].

Dans ce cadre, on peut distinguer plusieurs pistes de recherche permettant d'une part de produire une parole de synthèse la plus naturelle possible et d'autre part de diversifier les styles de voix.

Un premier axe d'étude consiste à remettre en cause les hypothèses de traitement de la matière acoustique dans un système de synthèse de la parole. La plupart des systèmes actuels juxtaposent des unités acoustiques correspondant à des unités linguistiques bien définies (par exemple, des diphones). Pour créer cette matière sonore, nous proposons de rechercher, au fil de la synthèse, des segments acoustiques pris dans une base de parole continue.

Dans cette optique, il n'y a plus de notion d'unités définies a priori sur des critères linguistiques, mais une recherche constante de la meilleure unité à retenir. La recherche des segments se fait au moment de la synthèse en s'appuyant sur un modèle acoustique conduit par la chaîne phonétique du message à prononcer. Les paramètres de ce modèle sont obtenus par apprentissage automatique.

Un deuxième axe d'étude cherche à réduire la distance qu'il y a entre le générateur de texte du système de dialogue et le signal à la sortie du système de synthèse. Nous proposons de traiter par le système de synthèse toutes les informations connues du système de dialogue comme par exemple les informations syntaxiques et grammaticales, les informations sémantiques, voire certaines informations pragmatiques. En tenant compte de ces données, il est alors possible de générer des façons de parler adaptées à certaines situations particulières du système de dialogue (information, répétition, insistance,...).

Enfin, un dernier axe d'étude consiste à diversifier les voix de synthèse tant sur le plan du timbre que sur le plan de l'énonciation en caractérisant ce problème comme un problème de conversion de voix. À partir d'une ou de plusieurs voix de synthèse de référence et à partir de la signature vocale d'une voix cible, les caractéristiques segmentales et prosodiques de la voix de référence sont transformées pour ressembler sur le plan de la perception à celles de la voix cible.

Description à court terme Deux thèses, entamées en septembre 1999 et janvier 2000, ont permis de d'accentuer nos efforts sur les deux premiers axes décrits précédemment.

Une première thèse en collaboration avec France-Télécom R&D s'intéresse au problème de l'obtention de la matière acoustique de synthèse. L'objectif d'un système de synthèse de la parole à partir

[BBC⁺93] D. BIGORGNE, O. BOËFFARD, B. CHERBONNEL, F. EMERARD, D. LARREUR, J. L. L. SAINT-MILON, I. MÉTAYER, C. SORIN, S. WHITE, « Multilingual PSOLA Text-to-Speech system », *IEEE International Conference on Acoustics, Speech, and Signal Processing 2*, 1993, p. 187–190.

[TB98] P. TAYLOR, A. BLACK, « The architecture of the Festival speech synthesis system », in : *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 1998.

[Tay00] P. TAYLOR, « Concept-to-speech by phonological structure matches », in : *Philosophical Transactions of the Royal Society Series A*, 2000.

[DW95] R. E. DONOVAN, P. C. WOODLAND, « Improvements in an HMM-based speech synthesiser », in : *Proceedings of the Eurospeech Conference*, 1995.

du texte consiste en effet à créer un signal de parole correspondant à la prononciation d'un message écrit. Actuellement, l'état de l'art technologique d'un système de synthèse consiste à assembler des unités de parole élémentaires pour créer la matière sonore. L'ensemble des unités acoustiques, connu pour la plupart des systèmes comme un ensemble de diphones (suite de deux demi-phones), est fixe et déterminé par expertise phonétique et acoustique quelque soit les phrases de synthèse à créer [Meu96]. L'objectif de cette thèse consiste à reformuler cette hypothèse sur la nature des unités acoustiques. Pour cela on considère que le continuum acoustique de synthèse est fait par assemblage de segments acoustiques non définis a priori et soumis à des conditions contextuelles. Le support acoustique d'où sont extraits les segments à assembler est une base de donnée de parole continue suffisamment longue (au moins deux heures d'enregistrement) pour pouvoir mettre en œuvre ces choix contextuels de segments. Une solution, développée au cours de cette thèse, consiste à modéliser la génération d'une phrase de parole de synthèse par une séquence de modèles de Markov cachés; des modèles du niveau phrase étant constitués par assemblage de modèles élémentaires d'un niveau allophonique. Les critères de sélection d'une séquence de segments acoustiques seront de nature acoustique [DW95], prosodique [TB98], et syntaxique.

Une seconde thèse en collaboration avec France-Télécom R&D s'intéresse à un problème transverse concernant la segmentation automatique de la parole spontanée en phones. Cette thèse s'inscrit dans le cadre du développement d'une nouvelle génération de systèmes de synthèse de parole intelligible et naturelle. Les systèmes de synthèse de parole à partir du texte par concaténation d'unités acoustiques reposent sur l'utilisation de bases de données de sons appelées dictionnaires acoustiques. Ces dictionnaires sont construits à partir de corpus de parole naturelle segmentée en éléments acoustiques unitaires (phones, diphones, mots). Un enjeu important concerne la production automatique de dictionnaires d'unités acoustiques de qualité et de bases de données pour l'étude des phénomènes prosodiques. Bien qu'il soit possible de segmenter manuellement un corpus de parole, une telle tâche nécessite un temps important et doit être réalisée par un expert. Les approches markoviennes de segmentation d'un signal de parole en phones ont montré leur efficacité à condition de disposer d'une transcription phonétique correspondant exactement à la phrase énoncée par le locuteur. L'originalité de cette étude consiste à reformuler l'hypothèse de segmentation à un niveau plus abstrait: le système de segmentation du signal de parole en phones ne dispose pas de la transcription phonétique exacte du locuteur mais de celle provenant d'un système de transcription orthographique/phonétique automatique. L'approche que nous souhaitons valider consiste à relâcher les contraintes du treillis de phonétisation d'une phrase à segmenter par des informations de niveau phonologique calculées à partir du signal effectivement prononcé. Il peut s'agir, par exemple, de la présence d'une pause entre deux phonèmes ou bien de la réalisation d'un e muet. A partir des quelques chemins extraits du treillis phonétique par l'approche d'inférence phonologique pré-citée, une segmentation du signal de parole en phones par modèles de Markov cachés sera mise en œuvre.

[Meu96] P.-Y. L. MEUR, *Synthèse de la parole par unités de taille variable*, thèse de doctorat, Ecole nationale supérieure des télécommunications, 1996.

[DW95] R. E. DONOVAN, P. C. WOODLAND, « Improvements in an HMM-based speech synthesiser », in : *Proceedings of the Eurospeech Conference*, 1995.

[TB98] P. TAYLOR, A. BLACK, « The architecture of the Festival speech synthesis system », in : *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 1998.

3.5 Etudes en traitement de la parole pour l'apprentissage des langues

Mots clés : technologies de la parole, logiciels éducatifs, enseignement et apprentissage des langues.

Résumé : *L'objectif de cette étude est de concevoir et de développer des logiciels éducatifs d'aide à l'enseignement et à l'apprentissage de langues (plus particulièrement d'apprentissage de la prosodie), centrés autour d'exercices pédagogiques d'écoute, de perception et de production intégrant les techniques de traitement de la parole. Ces outils éducatifs sont destinés à l'enseignement et à l'apprentissage du breton mais, dans une large mesure, ils sont directement utilisables pour d'autres langues.*

Les logiciels éducatifs d'enseignement et d'apprentissage de la prosodie exigent des outils de traitement de la parole de très bonne qualité aussi bien au niveau de l'analyse (détection de la période fondamentale) qu'au niveau de la synthèse (les phrases synthétisées peuvent dans ce cas servir de modèles) Il faut aussi pouvoir modifier le signal de l'apprenant (ou du maître) pour le rapprocher du modèle ou pour exagérer les erreurs et savoir déplacer les cibles à atteindre. Il faut également disposer de bons outils de segmentation de la parole pour que les paramètres prosodiques segmentaux du signal de référence et du signal de l'apprenant puissent être comparés et pour que les erreurs sur le rythme, la durée ou sur les accents soient correctement localisés.

Le premier axe de recherche concerne donc la synthèse automatique de la parole. Dans ce cadre, nos études portent principalement sur la synthèse par concaténation d'unités de taille variable, sur l'analyse linguistique (mise au point de règles de décomposition morphologique et analyse grammaticale) et sur la recherche d'un modèle prosodique bien adapté à la synthèse des textes des exercices d'apprentissage de la prosodie.

Le deuxième volet concerne la mise au point d'outils spécifiques d'alignement et de segmentation automatique des enregistrements du maître et de l'apprenant : comparaison dynamique, synthèse et décodage de Viterbi sont les techniques de base sur lesquelles reposent cette segmentation.

Le dernier volet concerne la mise au point de techniques de localisation et d'interprétation des erreurs de réalisation prosodique. Cette localisation procède par comparaison des paramètres prosodiques aux paramètres du modèle tuteur ou d'un modèle théorique de la langue à apprendre.

4 Domaines d'applications

Résumé : *Les domaines d'application du projet sont nombreux. Ce sont essentiellement des domaines d'activité humaine réunissant plusieurs personnes, où la communication orale joue un rôle important. En particulier, on peut citer les activités telles que les services d'informations, de réservation par téléphone et également les industries de la langue avec en particulier l'enseignement assisté par ordinateur. La motivation du projet est de fournir un substitut automatique à l'être humain dans cette tâche d'assistance lorsque le travail à accomplir est particulièrement contraignant.*

Les domaines d'application spécifiques des techniques de traitement de la parole sont ceux où l'usage de l'oral apporte un confort d'utilisation incontestable (télématique vocale, domotique) ou

bien s'avère indispensable (consultation de bases de données par téléphone, autres moyens de communications inopérants ou occupés). L'usage de la langue naturelle dans la communication orale permet aussi de placer les applications dans le domaine des industries de la langue. Les logiciels et produits (CD-Roms) que nous développons se situent dans ce domaine et plus particulièrement dans le sous-domaine des logiciels éducatifs. Les techniques et savoir-faire utilisés nous permettent d'envisager de nouvelles applications proches ; par exemple, le logiciel Ordictée (cf 5.1) pourra être utilisé pour l'apprentissage de la langue bretonne, ou bien encore une partie du logiciel de synthèse du breton sera utilisable pour une autre langue.

Les techniques d'apprentissage automatique peuvent être utilisées pour le développement d'interfaces homme-machine pour les applications utilisant le langage naturel. Dans ce type d'application, la variabilité des énoncés des utilisateurs ainsi que la nécessaire robustesse de la compréhension du système rendent pertinent le recours à des techniques d'ingénierie de conception à base d'apprentissage.

5 Logiciels

Résumé : *Nos activités ont permis la production de logiciels qui concernent deux types d'usages différents : didactique et support pour le développement. Pour le premier usage, deux logiciels ont été développés : Ordictée, pour l'apprentissage de l'orthographe et Argeriadur a gomz qui est un dictionnaire vocal français-breton. Pour le second usage, nous avons produit Epigram qui est un ensemble de fonctions relatives à l'inférence grammaticale permettant le développement d'applications.*

5.1 Ordictée

Participants : Marc Guyomard [*correspondant*], Olivier Boëffard.

Ordictée est un logiciel permettant à un élève de réaliser une dictée de manière autonome. Le logiciel est constitué de trois modules : le module « élève » qui effectue la dictée proprement dite, le module « tuteur » qui permet à l'instituteur de créer ses propres dictées et le module « concepteur » qui permet la gestion et le paramétrage de l'ensemble. La fonction de suivi de frappe, c'est-à-dire l'adaptation du rythme de la lecture à la vitesse de frappe de l'élève, est fondée d'une part sur l'hypothèse que les fautes conservent en général la prononciation, et d'autre part sur la proximité phonétique des deux textes considérés (le texte de l'élève et celui du tuteur).

Une version accessible via le Web est envisagée. Les premiers essais de faisabilité (contraintes temps-réel sur la synthèse de texte et sur la frappe côté client ; traitement d'un caractère par le serveur) sont en cours.

5.2 Dictionnaire vocal

Participant : Guy Mercier [*correspondant*].

La version 2.0 du logiciel « gervogal », dictionnaire vocal multimédia bilingue (français-breton et

breton-français) est en cours de test et devrait être diffusée et commercialisée par la maison d'édition Skol Vreizh. Ce dictionnaire intégrant textes, sons, jeux et images, à usage pédagogique, a été réalisé en collaboration avec le CRDP de Bretagne (TES) et l'université de Rennes II. Dans la nouvelle version, le dictionnaire a été converti en base de données (ACCESS), ce qui a permis d'accélérer la recherche des mots et de mettre au point une recherche approximative pour retrouver les mots voisins. Cette version bénéficie d'une interface graphique améliorée et de nouvelles fonctionnalités telles que la possibilité de synthétiser et d'écouter tout exemple d'usage d'un mot, présent dans la définition de ce mot, ou encore la possibilité pour un utilisateur de personnaliser son dictionnaire en y ajoutant de nouveaux mots ou en modifiant la phonétique ou les définitions des mots existants.

5.3 Epigram

Participants : Laurent Miclet [*correspondant*], Jacques Chodorowski.

Afin d'accélérer et d'uniformiser le développement d'applications pour l'inférence grammaticale, nous avons développé un outil sous forme d'une bibliothèque de classes C++. Cette bibliothèque de fonctionnalités de haut niveau a été conçue selon deux principes : la généralité et l'indépendance de la stratégie d'implantation. La première permet à l'utilisateur de la bibliothèque d'adapter facilement des fonctions implantées à son problème, et de raccourcir ainsi le temps de programmation. La seconde laisse libre le choix du style de programmation (statique, dynamique) tout en fixant un cadre formel par le biais de mécanismes des classes abstraites.

Cet outil, appelé EPIGRAM (Environnement de Programmation pour l'Inférence GRAMmaticale), a été développé à partir de 1997. Actuellement il possède deux implémentations : dynamique — en tant qu'une surcouche de la bibliothèque C++ LEDA — (bibliothèque de types de données et d'algorithmes de programmation combinatoire, développée à Max-Planck-Institut für Informatik, Saarbrück, Allemagne) et statique (développée par J. Chodorowski). Tous les programmes produits par l'équipe dans le cadre de recherches ayant pour thème l'inférence grammaticale sont désormais implantés dans EPIGRAM. Le développement de cette bibliothèque est poursuivi en collaboration avec l'équipe de l'université de Saint-Étienne (Colin de la Higuera, Franck Thollard) ainsi qu'avec le projet Aïda (Jacques Nicolas, François Coste) de l'Irisa, dans le cadre du Contrat FT R&D CTI 97 1B 004 (échu en Mars 2000).

6 Résultats nouveaux

6.1 Dialogue et modélisation

6.1.1 Représentation logique pour le dialogue

Participant : Jean-Christophe Pettier.

Après avoir défini un formalisme de modélisation de l'application adapté aux situations de dialogue [PG00], nous avons abordé l'étude des représentations logiques support. Notre étude s'est intéressée à la représentation du contexte physique de l'application en cherchant à concilier pouvoir d'expression et traitabilité des calculs.

Cette étude a conduit à proposer un langage logique du 1^{er} ordre pour le cadre fini et à en dériver un solveur de modèles, ce qui permet d'obtenir des tests de complétude et de conséquence logique. L'intérêt d'un tel solveur pour le dialogue est donc double : il évalue la couverture de la description du contexte physique de l'application et en cas d'absence d'informations, il autorise néanmoins l'évaluation logique et de ce fait le déroulement du dialogue. En effet, le statut de vérité *possible* devient accessible, ce qui permet de déclencher à la demande des actes de dialogue pour statuer par *vrai/faux* vis-à-vis du monde physique connu. Une implémentation d'une première version de ce solveur (garantissant la décidabilité mais pas la traitabilité) est en cours de validation.

La thèse que débute Sylvain Camus aborde quant à elle la représentation des états mentaux des interlocuteurs.

6.1.2 Évaluation de systèmes de dialogue

Participant : Jacques Siroux.

Cette activité était développée dans le cadre d'un financement de L'AUPEL-AUF et grâce à une collaboration entre plusieurs laboratoires. Le financement devait durer quatre ans. L'AUPEL-AUF a mis fin au financement au bout de deux ans à cause de problèmes internes et malgré les résultats intéressants obtenus. Nous avons donc temporairement arrêté les études sur ce thème tout en maintenant une veille scientifique (bibliographie, participation à un séminaire lors du congrès ACL-EACL à Toulouse en juillet 2001).

6.2 Systèmes et multimodalité

Résumé : *L'étude des phénomènes de désignation et de référence pour une version enrichie de Géoral Tactile a été menée. Nous avons également poursuivi des activités de développement pour l'amélioration du logiciel Ordictée (prise en compte des fautes d'origine phonétique, suivi de la frappe).*

[PG00] J.-C. PETTIER, M. GUYOMARD, « Action Modeling in Dialogue Context », in : *Third International Workshop on Human-Computer Conversation*, p. 136–141, Bellagio, Italie, 2000.

6.2.1 Géoral tactile et référence

Participants : Marc Guyomard, Jacques Siroux.

Les progrès en reconnaissance de la parole nous permettent d'envisager de nouveaux développements d'envergure dans le système de dialogue *Géoral Tactile*^[S⁺97]. L'accroissement du nombre de mots du vocabulaire donne la possibilité aux utilisateurs de formuler des phrases linguistiquement plus complexes. Nous utilisons ce fait pour enrichir l'univers de l'application en ajoutant de nouveaux éléments sur la carte qui sert de support aux interrogations. Dans ce nouveau cadre, plusieurs points sont à l'étude : une modélisation du contexte cartographique, les comportements linguistiques et gestuels des utilisateurs pour désigner les éléments sur la carte et enfin l'organisation même du système.

Dans un premier temps, une expérimentation a été menée afin de déterminer le comportement langagier des utilisateurs dans leur activité de désignation des éléments sur la carte. Un grand nombre de formes linguistiques ainsi que l'utilisation d'éléments construits (par exemple désignation d'un triangle à l'aide de points particuliers) ont été observés. Une nouvelle forme de geste (suivi d'une ligne) est également apparue ^[Bre98].

Nous proposons un modèle syntaxique pour analyser et filtrer les expressions référentielles dans les énoncés des utilisateurs. Ce modèle est fondé sur les travaux de Vandeloise^[Van86] et d'A. Borillo^[Bor88] qui prennent en considération les caractéristiques spatiales des éléments manipulés. Nous avons ensuite développé un modèle sémantique qui permet de filtrer plus finement les productions de l'analyseur syntaxique. Le modèle est dérivé de celui d'Aurnague^[Aur93] qui utilise des propriétés caractéristiques des éléments (telles que la dimension, la consistance, la situation...). Nous n'utilisons que trois caractéristiques (dimension, consistance et forme) mais de manière combinée, compte tenu des constructions linguistiques possibles.

Du point de vue cartographique, nous avons développé un nouveau modèle de données ainsi que des algorithmes de recherche mieux adaptés aux éléments manipulés.

Enfin, le fait de traiter en plusieurs phases des énoncés et des gestes plus complexes et celui de voir apparaître des objets qui ne sont pas présents dans la base de données nous ont amenés à faire évoluer l'architecture du système et la philosophie des traitements. Fondé sur la prééminence des activités gestuelles sur les activités orales (le contraire de ce qui se passe dans la version actuelle), le principe permet de vérifier de manière progressive et éventuellement de corriger les expressions linguistiques référentielles, de déterminer les référents potentiels sur la carte et de construire le cas échéant de nouveaux éléments dans la base. Certains des algorithmes ont été implantés mais pas intégrés au système.

Le prototypage et l'évaluation de cette approche nécessite une nouvelle plate-forme matérielle

-
- [S⁺97] J. SIROUX *et al.*, « Multimodal References in Georal Tactile », in : *Proceedings of the workshop Referring Phenomena in a multimedia Context and their Computational Treatment, SIGMEDIA and ACL/EACL*, p. 39–44, Madrid, juillet 1997.
- [Bre98] G. BRETON, « Modélisation d'un contexte cartographique et dialogique », *rapport de recherche*, DEA Informatique de Rennes 1, 1998, ENSSAT.
- [Van86] C. VANDELOISE, *L'espace en français*, Éditions du seuil, Paris, 1986.
- [Bor88] A. BORILLO, « Le lexique de l'espace : les noms et les adjectifs de localisation interne », *Cahiers de grammaire* 13, 1988, p. 1–22.
- [Aur93] M. AURNAGUE, *A unified processing of orientation for internal and external localization*, Groupe Langue, Raisonnement, Calcul, Toulouse, 1993.

et logicielle. Des études sur la mise en place d'une telle plate-forme ont débuté ; elles concernent notamment ses caractéristiques matérielles (notre volonté étant de permettre l'accueil de prototypes variés en provenance des industriels de la zone lannionnaise), la spécification des fonctionnalités des constituants (description des composants actuels de Georal, description des constituants futurs, étude des normes FIPA) ainsi que l'intégration de certains des résultats du projet Badins.

6.2.2 Ordictée

Participants : Marc Guyomard, Jacques Siroux.

Ordictée est un logiciel permettant à un élève de réaliser une dictée de manière autonome. Le logiciel est constitué de trois modules : le module «élève» qui effectue la dictée proprement dite, le module «tuteur» qui permet à l'instituteur de créer ses propres dictées et le module «concepteur» qui permet la gestion de l'ensemble.

Le principe de la fonction de suivi de frappe, c'est-à-dire l'adaptation du rythme de la lecture à la vitesse de frappe de l'élève, a été modifié. Il est présent fondé sur l'hypothèse que les fautes conservent en général la prononciation, et sur la proximité phonétique des deux textes considérés (le texte de l'élève et celui du tuteur). Des essais préliminaires dans une classe d'enseignement primaire montrent des résultats encourageants.

Le logiciel a également été adapté à la langue bretonne.

La librairie dynamique (DLL) de synthèse du breton a été intégrée au logiciel Ordictée ; les messages guidant l'élève ou le tuteur ont été traduits et sont émis par cette nouvelle DLL. Les messages visuels des diverses interfaces sont bilingues et comme pour le français, cette version dispose d'un ensemble de dictées classées par niveau de difficulté.

6.3 Apprentissage et traitement de l'information sonore

Résumé : *Les études sur l'inférence de grammaire pour le dialogue oral ont été poursuivies sur deux axes : l'amélioration de l'efficacité des méthodes n-grams et l'exploitation des méthodes d'exploration de l'espace de recherche. Le travail sur la conversion de la voix a commencé cette année. L'apprentissage automatique de la prosodie pour le dialogue oral fait l'objet d'études méthodologiques et appliquées.*

6.3.1 Inférence de grammaires pour le dialogue oral

Participants : Jacques Chodorowski, Laurent Miclet.

Jacques Chodorowski a soutenu sa thèse de doctorat de l'Université de Rennes 1 le 7 Juin 2001 : "Inférence grammaticale pour l'apprentissage de la syntaxe en reconnaissance de la parole pour le dialogue oral". La rédaction du document et la préparation de cette soutenance ont constitué pour

2001 l'essentiel des travaux sur l'inférence grammaticale (J. Chodorowski est depuis Septembre 2000 ingénieur à FT R&D).

6.3.2 Apprentissage de courbes prosodiques décrites par des structures d'arbres

Participants : Laurent Blin, Laurent Miclet.

En 2000, le travail effectué dans ce thème s'était principalement articulé autour de la définition des types de structures arborescentes utilisées, du développement d'algorithmes de calcul de distances entre arbres, de la définition d'un apprentissage par plus proches voisins sur l'application concernée, et de l'étude préliminaire du concept d'analogie [BE00a,BE00b,BM00] Ces travaux avaient été complétés et précisés lors d'un séjour de six mois effectué par L. Blin dans le laboratoire Star de SRI International². Ce séjour avait également permis la constitution d'une nouvelle base complète d'apprentissage à partir d'un corpus de langue anglaise (Boston University Speech Radio Corpus), cette base s'ajoutant à un corpus de langue française créé au début des travaux.

Au cours de cette année, l'apprentissage par *plus proches voisins* a été finalisé pour obtenir un premier système de prédiction complet. Dans cette approche, la prédiction des caractéristiques prosodiques d'une phrase X_1 s'effectue par la recherche dans un ensemble de phrases (de caractéristiques prosodiques connues) de la phrase Y_1 la plus proche (au sens d'une des distances définies sur les caractéristiques arborescentes choisies).

L'évaluation de cette méthode a permis de mettre à jour un phénomène prévisible : la taille relativement petite des bases d'apprentissage utilisées (environ 640 phrases pour la base française et 320 phrases pour la base anglaise) n'autorise pas toutes les variétés structurales possibles. Cela se traduit en pratique au niveau des calculs de distances entre structures par des valeurs souvent relativement élevées. Devant les hypothèses posées, il devient illégitime d'essayer de générer la prosodie d'une phrase à partir d'une autre qui ne possède pas un minimum de caractéristiques semblables.

Cette constatation a conduit à développer une première solution basée sur un éclatement des structures manipulées. Étant donné un arbre représentant une phrase, celui-ci est décomposé en sous-arbres, chacun représentant les groupes logiques de plus bas niveau constituant la phrase initiale. Le but de cette décomposition est de travailler sur des structures plus petites, et donc moins diversifiées. Néanmoins, le calcul des distances entre ces nouvelles structures est à modifier légèrement pour prendre en compte le déséquenceage de ces morceaux de phrases (ne chercher à comparer un début de phrase qu'avec un début de phrase, par exemple). Dès lors, on voit bien que si le processus de découpage des structures est poursuivi plus en profondeur, les avantages de la structure hiérarchique sont de plus

²Stanford Research Institut International

-
- [BE00a] L. BLIN, M. EDGINGTON, « Prosody Prediction from Tree-Like Structure Similarities », in : *Third International Workshop on Text, Speech and Dialogue (TSD 2000)*, P. Sojka, I. Kopeček, K. Pala (éditeurs), *Lecture Notes in Artificial Intelligence LNCS/LNAI 1902*, Springer-Verlag, p. 369–374, Brno, République Tchèque, Septembre 2000.
- [BE00b] L. BLIN, M. EDGINGTON, « Prosody Prediction using a Tree-Structure Similarity Metric », in : *International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, Chine, Octobre 2000.
- [BM00] L. BLIN, L. MICLET, « Generating Synthetic Speech Prosody with Lazy Learning in Tree Structures », in : *Proceedings of the 4th Conference on Computational Language Learning*, p. 87–90, Lisbonne, Portugal, 2000.

en plus perdus, pour aboutir à l'analyse d'une séquence de morceaux élémentaires de l'arbre initial mis à plat. Le juste milieu est donc à trouver. Les évaluations menées sur ce principe, en se limitant à un découpage au plus haut niveau, présentent une amélioration substantielle de l'appariement entre structures.

À la suite de ces évaluations, le concept d'apprentissage par *analogie* sur des structures arborescentes a été finalisé et développé. Cette approche pallie les faiblesses de l'apprentissage par plus proches voisins, en particulier lorsque les structures arborescentes de la phrase X_1 et de sa plus proche voisine Y_1 ne sont pas aussi similaires que désirées, et que les liens entre les deux structures ne sont pas suffisamment porteurs d'information pour permettre de générer la prosodie de X_1 à partir de celle de Y_1 .

L'hypothèse de base de cette approche est qu'il doit être possible de trouver dans l'ensemble de phrases connues un triplet d'arbres (Z_1, Z_2, Z_3) tel que le couple (X_1, Z_1) présente la même alternance structurelle que le couple (Z_2, Z_3) , et que l'on peut alors en déduire la modification des caractéristiques prosodiques induite par cette modification des caractéristiques structurelles.

Un nouvel algorithme a été créé à ces fins. Étant donné un quadruplet d'arbres (Z_1, Z_2, Z_3, Z_4) , cet algorithme détermine s'il existe une transformation structurelle qui permette à la fois de passer de l'arbre Z_1 à l'arbre Z_2 , et de l'arbre Z_3 à l'arbre Z_4 . Cette recherche, basée sur les mêmes opérations de transformations entre arbres que celles manipulées par les algorithmes de calcul de distance, en respecte également les contraintes d'ordre entre noeuds. La mise au point de l'algorithme d'apprentissage et de génération de prosodie correspondant constitue l'étape suivante de ces travaux.

Une collaboration de six mois avec l'équipe *Interaction, Parole et Sons* du centre FT R&D de Lannion se déroule actuellement sur ces travaux, depuis le 19 novembre. Elle va permettre l'accès à des corpus de données complets, avec pour objectif l'évaluation approfondie des méthodes développées (tests perceptifs), et leur comparaison avec les systèmes de synthèse actuels de FT R&D. Cette évaluation finale conduira à la soutenance de thèse de L. Blin en 2002.

6.3.3 Synthèse flexible de la parole

Participants : Hélène François, Olivier Boëffard.

Ce sujet est traité dans le cadre d'une thèse financée par un marché d'études avec le centre FT R&D de Lannion (FT R&D/DIH/ISP). Les travaux ont débuté le 1 octobre 1999.

En marge de l'objectif énoncé précédemment (paragraphe ??), nous proposons de déterminer automatiquement un ensemble minimal de phrases à enregistrer qui serviront de bases d'exemples pour les techniques d'apprentissage des modèles de prédiction acoustique. Cet ensemble minimal doit être *optimal* au sens de la précision des modèles acoustiques. Le problème peut être formalisé comme une couverture minimale d'ensemble [?], [?]. Les travaux menés en 2001 sur ce sujet ont conduit à deux résultats importants :

- Dans un premier temps, la constitution d'une base de données textuelles composée de textes ou de transcriptions de discours variés. Cette base de données comporte environ 310 000 phrases.
- Il est évidemment inenvisageable d'enregistrer l'équivalent sonore de cette première base. Par un algorithme de résolution de couverture optimale d'ensemble nous avons déterminé une base d'environ 4000 phrases assurant 95% de la couverture allophonique de la base de référence.

Cette base a été enregistrée par un locuteur et sert de support d'étude à la seconde partie de nos travaux concernant la définition de modèles acoustiques représentés par des modèles de Markov cachés. Sur un plan algorithmique, nous avons mis au point une solution de type glouton appliquée à la réduction optimale d'une base de données textuelles annotée par des informations phonologiques. Deux programmes ont été écrits, l'un répondant à un objectif fonctionnel, l'autre à un objectif d'optimisation du temps de calcul. Le problème est de type NP-difficile et correspond à l'application de traitements sur une matrice l'ordre de 310000×30000 . Ces travaux ont fait l'objet d'une publication au congrès eurospeech 2001 [10].

Les algorithmes de type glouton peuvent s'appuyer sur une variété de fonctions de pondération. Une étude connexe consiste à mesurer *l'utilité* d'une phrase décomposable sur un vocabulaire d'unités phonétiques. Le modèle utilisé est un modèle probabiliste. On fait, en outre, l'hypothèse de ne pas connaître a priori le vocabulaire des unités composant les séquences observées. Il s'agit d'apporter à la fois une solution à un problème d'inférence et d'optimisation. La variante de l'algorithme EM proposée résout l'asynchronisme entre le processus caché et le processus observé. Ces travaux ont fait l'objet d'une publication au congrès eurospeech 2001 [9].

6.3.4 Segmentation de parole en phones, inférence phonologique

Participants : Samir Nefti, Olivier Boëffard.

Ce sujet est traité dans le cadre d'une thèse financée par un marché d'études avec le centre FT R&D de Lannion (FT R&D/DIH/ISP). Les travaux ont débuté le 1 janvier 2000.

Le sujet concerne la segmentation automatique des corpus de parole naturelle lue et spontanée pour les applications de synthèse de parole par concaténation d'unités acoustiques. Le cadre général de cette étude est citée au paragraphe ???. Le système de segmentation de parole développé cette année est entièrement fonctionnel. Les scores de segmentation que nous avons obtenus sont équivalents à ceux des systèmes "état de l'art". Pour analyser le comportement de ce système de segmentation, nous avons de plus réalisé une série d'expériences qui se situent sur deux plans : celui de la description topologique et celui de l'analyse acoustique.

Au niveau topologique, nous avons évalué des architectures de modèles de Markov cachés dérivées des connaissances phonologiques. Cette étude n'a pas apporté d'amélioration significative par rapport à l'architecture de type modèle de Bakis avec des densités d'observation représentées par des multigaussiennes.

Au niveau acoustique, nous avons évalué le rendement en qualité de segmentation des coefficients du type "paires de lignes spectrales" (Line Spectrum Pairs). Ces coefficients sont plutôt utilisés principalement en codage de parole ou en conversion de voix. Les scores obtenus ne sont pas aussi bons que ceux donnés par une analyse acoustique classique du type MFCC (Mel Frequency Cepstral Coefficients). L'ensemble de ces travaux ont fait l'objet d'une publication au congrès eurospeech 2001 [11].

6.4 Intégration des technologies de la parole pour l'enseignement des langues

Participants : Yves Aubry, Guy Mercier, Jacques Siroux.

Résumé : *Les études ont porté sur la mise en place d'une nouvelle version du dictionnaire vocal, sur l'amélioration du modèle de prosodie pour la synthèse du breton et sur les méthodes d'alignement automatique et de comparaison des enregistrements de parole « maître et élève » dans le logiciel d'apprentissage et d'enseignement de la prosodie.*

6.4.1 Dictionnaire vocal

L'interface graphique du dictionnaire a été améliorée, en particulier au niveau des définitions et des exemples d'usage qui sont présentés d'une manière plus structurée que dans la première version. Un nouveau logiciel permettant d'introduire facilement les origines dialectales des variantes de prononciation dans la définition des mots a été mis au point. Enfin, il est désormais possible d'appeler le dictionnaire comme outil de correction orthographique, dans une application de type traitement de texte. Comme on l'a vu précédemment, une version 2.0 de ce dictionnaire est en cours de test.

6.4.2 Synthèse du breton à partir du texte

Les principales améliorations concernent la modélisation prosodique. L'utilisation de nouveaux corpus de parole mis à notre disposition par l'université de Rennes I et le CRDP et la mise au point d'outils de segmentation automatique nous ont permis d'améliorer nos connaissances sur la prosodie du breton et d'améliorer les règles de calcul des paramètres prosodiques au niveau des syllabes accentuées et réduites.

6.4.3 Correcteur de prosodie

Les améliorations introduites dans ce logiciel ont porté principalement sur le module de comparaison dynamique entre la parole à segmenter et la parole « modèle ». Grâce à ce module, le tuteur peut maintenant créer ses propres exercices pédagogiques de manière plus conviviale car il peut segmenter ses enregistrements de manière quasi automatique en comparant la représentation spectrale de sa prononciation à la représentation acoustique de la même phrase produite par synthèse à partir du texte. Grâce à cette segmentation automatique, on peut synchroniser et superposer les différentes courbes prosodiques (énergie, fréquence fondamentale), syllabe par syllabe et ainsi bien visualiser les différences de production entre le maître et l'élève.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Synthèse de la parole

Un contrat entre l'université de Rennes 1 et France-Télécom, *Synthèse flexible de la parole* a été notifié en Novembre 1999 pour une durée de trois ans (marché FT 99 IB 588). Ce contrat comporte le financement de la thèse d'Hélène François. Les objectifs et les premiers résultats de cette thèse sont rappelés dans les paragraphes 3.4.3 et 6.3.3.

Un contrat entre l'université de Rennes 1 et France-Télécom, *Segmentation automatique de parole naturelle et spontanée* a été notifié en Juin 2000 pour une durée de trois ans (marché 00 IB 427). Ce contrat porte sur l'accueil de la thèse de Samir Nefti. Les objectifs de cette thèse sont rappelés au paragraphe 3.4.3 et les résultats en 2001 au paragraphe 6.3.4.

Une convention entre l'université de Rennes 1 et France-Télécom, *Apprentissage de la prosodie pour la synthèse en dialogue oral* a été notifié en octobre 2001 pour une durée de six mois (marché 00 IB 427). Ce contrat porte sur l'accueil à FT R&D de L. Blin et la mise à sa disposition de bases données prosodiques. En contrepartie, le logiciel qu'il a développé dans le cadre de sa thèse sera complété et FT R&D en disposera après l'avoir acheté. La soutenance de la thèse de L. Blin aura lieu à la fin de cette convention, vers Juin 2002.

8 Actions régionales, nationales et internationales

8.1 Réseaux et groupes de travail internationaux

Le projet Cordial fait partie du réseau d'excellence européen Elsnets (linguistique informatique) ainsi que du réseau francophone Francil (linguistique).

Dans le cadre des actions intégrées franco-allemandes (Procope 2001), un projet de collaboration entre l'équipe Cordial et l'institut de phonétique de l'université de Bonn a été accepté. Dans ce cadre, Yves Aubry et Hélène François ont présenté leurs recherches à l'institut de phonétique et Wolfgang Hess de l'université de Bonn a présenté sa méthode de synthèse par mots au cours de son séjour à l'ENSSAT.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

Jacques Siroux a fait partie des comités de programme TALN2001 et RECITAL2001. Jacques Siroux fait partie du comité de lecture de la revue InCognito.

Laurent Miclet a fait partie du comité de programme du congrès d'apprentissage *CAP 2001*. Il aura achevé début 2002 la rédaction avec A. Cornuéjols d'un livre intitulé *L'apprentissage artificiel : concepts et algorithmes*, à paraître chez Eyrolles.

9.2 Enseignement universitaire

Au cours de cette année, une antenne de la filière *Intelligence Artificielle et Images* du DEA Informatique de Rennes 1 a été mise en place à l'ENSSAT sous la coordination de L. Miclet.

O. Boëffard enseigne le cours de *Synthèse de la Parole* dans le DEA STIR, Rennes 1 (option Signal, orientation 2).

M. Guyomard et J. Siroux sont intervenus en décembre 2000 dans l'option IIIF (*Interaction Intellectuelle et Interrogation Flexible*) du DEA informatique de l'Ifsic.

M. Guyomard est responsable de l'option IIIF du DEA informatique de l'Ifsic.

M. Guyomard, G. Mercier et J. Siroux enseignent le module *Communication homme-machine*, en troisième année (option LSI) de l'Enssat Lannion.

L. Miclet enseigne le cours de *Reconnaissance des Formes* dans le séminaire Parole du DEA STIR ainsi que le module *Classification et Apprentissage* (CLAP) dans le DEA Informatique de Rennes I et en 3ème Année de l'ENSSAT.

9.3 Participation à des colloques, séminaires, invitations

L. Miclet a participé au jury de thèse de J. Chodorowski (Juin 2001) à l'ENSSAT, Université de Rennes 1

9.4 Accueil d'étudiants et de stagiaires

Le projet a accueilli les étudiants suivants :

- S. Finet, 3^{ème} année Enssat, correcteur orthographique
- J. Dupin, licence MIME du Mans, dictionnaire vocal
- M. Guérin, licence MIME du Mans, logiciel de marques dialectales
- J. Herry, 3^{ème} année Enssat, solveur logique
- C. Lebeaupin, 3^{ème} année Enssat, solveur logique

10 Bibliographie

Ouvrages et articles de référence de l'équipe

- [1] P. DUPONT, L. MICLET, E. VIDAL, « What is the search space of the regular inference ? », in : *Grammatical Inference and Applications, Lecture notes in AI 862*, Springer Verlag, septembre 1994.

-
- [2] P. DUPONT, L. MICLET, « L'inférence grammaticale régulière : fondements théoriques et principaux algorithmes », *rapport de recherche n° 3449*, INRIA, juillet 1998.
- [3] M. GUYOMARD, P. NERZIC, J. SIROUX, « Plans, métaplans et dialogue », *rapport de recherche n° 1169*, Irisa, septembre 1998.
- [4] M. GUYOMARD, J. SIROUX, *Suggestive and Corrective Answers: A Single Mechanism*, North Holland, Amsterdam, 1989.
- [5] L. MICLET, *Méthodes Structurelles pour la Reconnaissance des Formes*, Eyrolles, 1986.
- [6] P. NERZIC, M. GUYOMARD, J. SIROUX, « Reprise des échecs et erreurs dans le dialogue homme-machine », *Cahiers de linguistique sociale* 21, 1992, p. 35–46.
- [7] P. NERZIC, *Erreurs et échecs dans le dialogue oral homme-machine, détection et réparation*, thèse, université de Rennes 1, janvier 1993.
- [8] J. SIROUX, M. GUYOMARD, F. MULTON, C. RÉMONDEAU, « Oral and Gestural Activities of the users in the GÉORAL System », in : *Intelligence and Multimodality in Multimedia, Research and Applications*, John Lee (ed), AAAI Press, 1998.

Communications à des congrès, colloques, etc.

- [9] O. BOËFFARD, « Variable-length acoustic units inference for text-to-speech synthesis », in : *proceedings of the Eurospeech Conference*, 2001.
- [10] H. FRANÇOIS, O. BOËFFARD, « Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem », in : *proceedings of the Eurospeech Conference*, 2001.
- [11] S. NEFTI, O. BOËFFARD, « Acoustical and topological experiments for an HMM-based speech segmentation system », in : *proceedings of the Eurospeech Conference*, 2001.