$\mathsf{ETC}_{\mathsf{vérif}}$, UN ENVIRONNEMENT MULTI-AGENTS DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE EN CONTINU

Jean-Luc COCHARD Murielle VIAL

IDIAP, Case postale 592, CH-1920 Martigny

Tél: (41) 26 22 76 64 - Fax: (41) 26 22 78 18 - e-mail: [cochard vial]@idiap.ch

Abstract

$\mathsf{ETC}_{\mathsf{vérif}}$ is a prototype of a cooperative automatic speech recognition system. This work stems from the strong intuition that a probable solution to the general problem of speech understanding lies in the development of a system able to deal with a large set of distinct, partial and even unreliable problem solvers, namely HMMs (Hidden Markov Models), GTP (Graphemes To Phonemes) agents, prosodic analysers and even higher order agents processing syntactic and semantic knowledge. Up to now, these solvers allow us to work with words and phonemes. This paper describes how a composite solution is computed from all these partial solutions.

1 INTRODUCTION

$\mathsf{ETC}_{\mathsf{vérif}}$ [CO95, CF95] est un système multi-agents, basé sur la plate-forme ETC (Environnement de Traitement Coopératif) qui fournit différents services:

1. Un modèle d'Ω-*agents* constituant les différentes sources de connaissances du système;
2. Un modèle de μ-*agents* qui sont les hypothèses fournies par les Ω-*agents*;
3. Un moteur d'évolution du système.

Ce projet, réalisé conjointement à un autre projet de collection de plusieurs bases de données du français parlé en Suisse romande, sert de vérification automatique partielle de la cohérence du contenu de bases de données.

Ses données d'entrée sont de deux sortes : le signal de parole et le texte que doit prononcer le locuteur. La sortie indique si c'est effectivement ce texte qui a été dit.

2 ARCHITECTURE DU SYSTEME

Le prototype $\mathsf{ETC}_{\mathsf{vérif}}$ s'inspire des travaux effectués par Susan E. Lander [Lan94], proposant notamment une architecture modulaire et flexible dans laquelle la résolution de problèmes est basée sur un ensemble d'agents (voir aussi [HCA94]) et la recherche d'une solution globale est orientée par un protocole de négociation générique. La spécificité de notre système est de contenir deux niveaux d'agents (voir figure 1 et [CO95]).

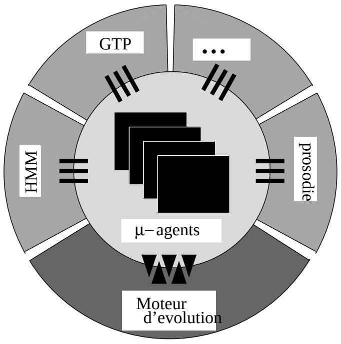

Figure 1: Architecture de $\mathsf{ETC}_{\mathsf{vérif}}$.

2.1 Les Ω-agents

Les Ω-*agents* ont chacun leur domaine de connaissance et apportent une solution locale au problème de la reconnaissance de la parole.
Deux fonctions leur sont attribuées:

1. Initialisation d'une solution: l'agent accepte en entrée la donnée complète d'un problème qu'il sait traiter localement et il transmet en sortie une solution satisfaisant au mieux ses contraintes locales;
2. Révision localisée: L'agent accepte en entrée une référence à une donnée sur laquelle il

a déjà fourni une solution initiale. Cette référence est assortie d'une indication de la zone de la donnée concernée par cette demande. L'agent transmet en sortie une nouvelle interprétation de cette zone en tenant compte des solutions qu'il a déjà fournies.

Les Ω-*agents*, actuellement utilisés dans notre système, sont des HMMs (Hidden Markov Models) [Tor94] et un agent GTP (Graphemes To Phonemes). L'agent GTP fournit une séquence de phonèmes à partir d'une phrase en codes ASCII. Cette séquence est ensuite utilisée comme grammaire de l'agent HMM basé sur des modèles de phonèmes afin de connaître la position temporelle de ces derniers. Deux agents HMM sont disponibles: l'un produit des phonèmes, l'autre des mots. Ce dernier peut également fournir des phonèmes en collaborant avec l'agent GTP: son entrée est le signal de parole, une sortie intermédiaire est une séquence de mots, cette séquence devient l'entrée de l'agent GTP et le résultat final est une transcription phonétique de la séquence de mots qui deviendra, une nouvelle fois, la grammaire de l'agent HMM basé sur des modèles de phonèmes.

Ainsi, le système dispose de trois séquences de phonèmes obtenues par les Ω-*agents* GTP, HMM s'appuyant sur des modèles de phonèmes et HMM s'appuyant sur des modèles de mots combinés avec l'agent GTP. Par ailleurs, il dispose aussi de deux séquences de mots qui sont le texte devant être prononcé par le locuteur (i.e. l'entrée de l'agent GTP) et le résultat fourni par l'agent HMM s'appuyant sur des modèles de mots.

2.2 Les μ-agents

Les μ-*agents* sont les résultats fournis par les Ω-*agents*. Ils peut s'agir de phonèmes ou de mots. Cinq informations sont stockées dans ces agents:

1. Borne_inf: borne temporelle inférieure;
2. Borne_sup: borne temporelle supérieure;
3. Id: l'identificateur du μ-*agent* (par exemple, une étiquette phonétique);
4. Les liens de voisinage
5. La confiance locale

2.3 Fonctionnement

Les Ω-*agents* produisent une première solution décomposée en une séquence de μ-*agents* qui sont des mots et des phonèmes. Ces derniers doivent alors estimer leur fiabilité (voir section 3.2). Le "moteur d'évolution" sélectionne ensuite la meilleure solution courante, c'est-à-dire la séquence de μ-*agents* situés sur le chemin optimal (voir section 3.3). Enfin, ce chemin est parcouru afin de trouver les zones temporelles de faible fiabilité dans lesquelles les Ω-*agents* devront fournir de nouvelles solutions (voir section 3.4).

3 IMPLEMENTATION

Les différentes étapes de calcul, détaillées ci-dessous, interviennent dans un système dynamique dont la stratégie de résolution est menée par des événements: l'adjonction de μ-*agents* ou la modification de leurs informations locales.

3.1 Initialisation d'une solution

Les Ω-*agents* sont des entités qui attendent une entrée spécifique afin de produire une première solution. Celle-ci est fournie sous forme d'une séquence de μ-*agents* (phonèmes, mots) qui sont des entités autonomes capables de déterminer leur fiabilité en se mettant en correspondance les unes avec les autre. Ce facteur détermine le taux d'intégration d'une hypothèse codée par un μ-*agent* dans son environnement.

Les relations établies entre deux μ-*agents* sont de deux types:

1. Les relations horizontales – Ce type de relation s'établit entre μ-*agents* de même classe, c'est-à-dire qui contiennent des informations comparables. Ainsi, les relations suivantes peuvent être établies entre deux μ-*agents* A et B:

   - A préc B: A a une valeur de borne temporelle supérieure qui coïncide avec la borne inférieure de B;
   - A suit B: A a une valeur de borne temporelle inférieure qui coïncide avec la borne supérieure de B;
   - A équiv B: les bornes temporelles inférieures et supérieures de A et B sont très semblables et leurs étiquettes E(A) et E(B) font partie d'une même classe phonétique. A et B sont alors des μ-*agents frères*.

   Quand l'une de ces propriétés est vérifiée, deux μ-*agents* deviennent voisins.

2. Les relations verticales – Ce type de relation décrit les liens entre μ-*agents* de classes différentes, par exemple, entre un phonème et un mot. En effet, chaque mot connaît sa décomposition phonétique grâce à l'agent GTP et est

capable de savoir si une suite de phonèmes du système y correspond.

Les calculs suivants (fiabilité et recherche du chemin optimal) ne sont effectués que par les phonèmes. En effet, d'abord il ne serait pas très judicieux de les faire pour les mots car on ne dispose que de deux séquences de ceux-ci (voir section 2.1): par conséquent, s'il existe deux mots différents sur une même zone temporelle, aucun indice ne peut indiquer lequel est le plus fiable. Ensuite les phonèmes (sauf ceux générés par l'Ω-*agent* HMM s'appuyant sur des modèles de phonèmes) sont obtenus par phonétisation (grâce à l'Ω-*agent* GTP) de ces mots. Ainsi, si un mot a une forte fiabilité, cela a une répercussion directe sur les phonèmes. De ce fait, les mots vont servir à renforcer et améliorer la solution finale, composée de phonèmes. Si dans une certaine zone temporelle, une suite de phonèmes, contenue dans cette solution, correspond partiellement à un mot proposé par le système à l'intérieur de cette zone, il sera possible de rechercher les phonèmes manquants et, s'ils ont une fiabilité suffisamment élevée, de les intégrer dans la solution finale. Le système sera alors en mesure de donner la séquence de mots qui a été prononcée dans les zones de forte fiabilité tandis que, dans celles de faible fiabilité, les Ω-*agents* devront fournir de nouvelles solutions et recommencer à négocier.

3.2 Calcul de la fiabilité d'un μ-agent

La fiabilité des μ-*agents* dépend de leur niveau d'intégration dans leur contexte. Afin de déterminer cette mesure, ils vont se communiquer différents résultats plutôt que de se voir attribuer des scores, de manière plus classique mais sans échange d'informations.

Les relations (horizontales) établies précédemment sont utilisées pour déterminer la fiabilité des μ-*agents*. Ainsi, ces derniers sont d'autant plus fiables que leurs voisins sont nombreux et ont, eux-mêmes, une fiabilité élevée. Cette mesure est dynamique et évolue avec l'adjonction de nouveaux μ-*agents* dans le système et la création de nouvelles relations.

Le calcul de confiance est itératif. Soit $F_i(A)$ la fiabilité du μ-*agent* A à l'itération i. On a les relations suivantes:

$$\begin{aligned} F_i(A) &= F_{i-1}(A) - \frac{1}{(2 + Fv_{i-1}(A))^i} \\ F_0(A) &= 1 \end{aligned}$$

$Fv_{i-1}(A)$ est la somme des fiabilités des voisins de A à l'itération $i-1$. Le calcul de la fiabilité de A se termine lorsque $F_i(A)$ se stabilise, c'est-à-dire lorsque:

$$\begin{aligned} F_{i-1}(A) - F_i(A) < \Delta \\ \Delta = 0.01 \end{aligned}$$

Cette valeur permet de parvenir à une bonne approximation de la valeur stable sans avoir des temps de calcul trop longs. De plus, la définition de $F_i(A)$ garantit que la confiance converge dans l'intervalle $]0,1[$ puisque $Fv_i(A) > 0$.

3.3 Recherche du chemin optimal

Le chemin optimal est la séquence de phonèmes de plus forte fiabilité. La méthode la plus simple est de choisir celui sur lequel la somme des fiabilités est la plus forte mais comme le montre la figure 2, on choisit alors le chemin comportant le plus grand nombre de phonèmes (dans notre exemple, B, C et D plutôt que A et D car 0.50 + 0.60 + 0.65 > 0.70 + 0.65), ce qui n'est pas le but recherché.

Aussi, choisit-on de calculer la somme des fiabilités pondérées par les durées des μ-*agents* afin que le chemin optimal comporte la séquence la plus fiable. Ainsi, dans la figure 2, si on considère que les larges rectangles ont une durée deux fois plus élevée que les rectangles étroits, le chemin optimal passe effectivement par les μ-*agents* de plus forte fiabilité (c'est-à-dire A et D car 2*0.70 + 0.65 > 0.50 + 0.60 + 0.65).

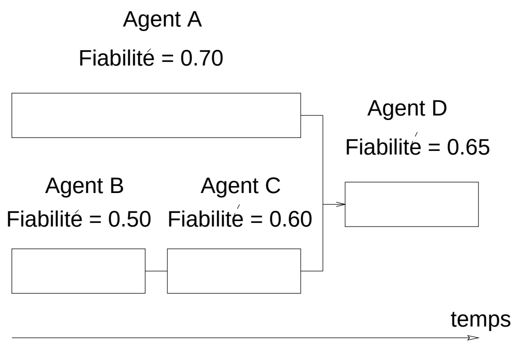

Figure 2: Recherche du chemin optimal

3.4 Le "moteur d'évolution"

Le chemin optimal (composé de phonèmes) est ensuite parcouru de façon à mettre, dans les zones de forte fiabilité, les phonèmes en relation avec les mots proposés par le système et à déterminer les zones de faible fiabilité dans lesquelles les Ω-*agents* doivent fournir de nouvelles solutions. Ceci est illustré par la figure 3. Dans les zones 1 et 2, la même séquence de phonèmes est reconnue par les trois Ω-*agents* (avec tout de même un décalage temporel pour l'Ω-*agent* HMM_PH): on se trouve donc dans des zones de forte fiabilité dans lesquelles les mots reconnus sont: cent six mille. Dans la zone 3, les Ω-*agents* HMM_PH et GTP se sont mis d'accord pour reconnaître le mot zéro, dont on retrouve la décomposition phonétique quasi complète sur le chemin optimal. On peut donc en conclure que le mot zéro a été reconnu sur la zone 3. Ensuite, sur la zone 4, le mot zéro est de nouveau reconnu. Enfin, la zone 5 est une zone de faible fiabilité où les Ω-*agents* devront fournir de nouvelles solutions.

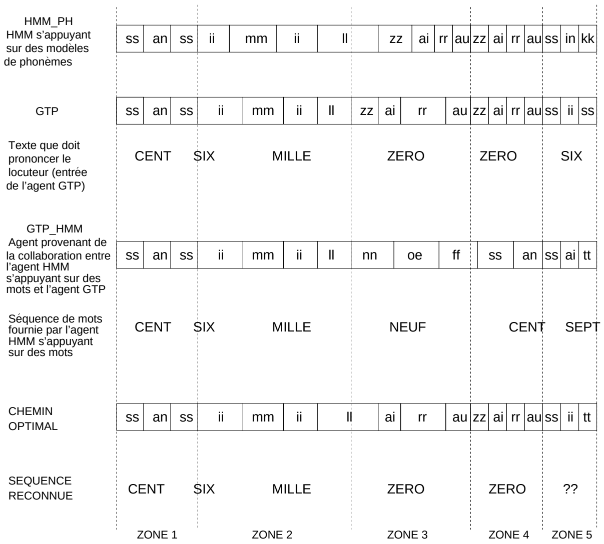

Figure 3: Recherche du texte prononcé par le locuteur

4 CONCLUSION

Il n'est pas actuellement possible de décrire les performances de $\mathsf{ETC}_{\mathsf{vérif}}$ car beaucoup de points doivent être améliorés (notamment le calcul de fiabilité et la coopération entre les phonèmes et les mots). Cependant, cette approche présente des caractéristiques intéressantes:

1. L'hétérogénéité des sources de connaissance – Actuellement, nous disposons de phonèmes et de mots. Par la suite, nous intégrererons également d'autres types d'informations comme des indices prosodiques [Lan95, LC95], des connaissances syntaxiques...

2. Un système distribué – Grâce au langage de programmation Oz, utilisé dans ce système, il est possible d'avoir des entités concurrentes (les μ-*agents*) avec des mécanismes de synchronisation.

3. L'interactivité – L'utilisateur peut devenir un Ω-*agent*: les solutions alors apportées sont prises en compte par le système (qui réagit à l'arrivée d'événements provoqués notamment par sélection de menus dans une interface graphique).

Cependant, des améliorations doivent encore être apportées sur ces différents points (comme, par exemple, la mise en relation des hypothèses) afin d'obtenir un système plus complet.

Bibliographie

[CF95] Jean-Luc Cochard and Philippe Froidevaux. Environnement multi-agents de reconnaissance automatique de la parole en continu. In *Actes des 3èmes Journées Francophones sur l'Intelligence Artificielle Distribuée et les Systèmes Multi-agents*, pages 101–110, mars 1995.

[CO95] Jean-Luc Cochard and Olivier Oppizzi. Reliability in a multi-agent spoken language recognition system. In *4th European Conference on Speech Communication and Technology*, Madrid, Spain, Sep 1995.

[HCA94] Béat Hirsbrunner, Michèle Courant, and Marc Aguilar. Parallelism and artificial intelligence. In *University of Fribourg Series in Computer Science Volume 3*, Institute of Informatics, University of Fribourg, Fribourg, Switzerland, August 1994.

[Lan94] Susan E. Lander. Distributed search and conflict management among reusable agents. Phd thesis, Dept of Computer Science, University of Massachusetts Amherst, May 1994.

[Lan95] Philippe Langlais. Traitement de la prosodie dans les systèmes de reconnaissance automatique de la parole. Thesis, Université d'Avignon et des pays du Vaucluse, France, October 11, 1995.

[LC95] Philippe Langlais and Jean-Luc Cochard. The use of prosodic agents in a cooperative automatic speech recognition system. In *International Congress of Phonetic Sciences*, Stockholm, Sweden, August 13–19 1995.

[Tor94] Kari Torkkola. New ways to use LVQ-codebooks together with hidden Markov models. In *ICASSP*, Adelaide, Australia, April 1994.

[Via95] Murielle Vial. Définition et évaluation d'un protocole de négociation dans un système multi-agents de reconnaissance de la parole. Technical report, IDIAP, Martigny, Switzerland, June, 1995.