Parole et traduction automatique : le module de reconnaissance RAPHAEL

Mohammad AKBAR
GEOD, CLIPS/IMAG
Université Joseph Fourier, BP. 53
38041 Grenoble cedex 9, France
Mohammad.Akbar@imag.fr

Jean CAELEN
GEOD, CLIPS/IMAG
Université Joseph Fourier, BP. 53
38041 Grenoble cedex 9, France
Jean.Caelen@imag.fr

Résumé

Pour la traduction de parole, il est nécessaire de disposer d'un système de reconnaissance de la parole spontanée grand vocabulaire, tournant en temps réel. Le module RAPHAEL a été conçu sur la plate-forme logicielle de JANUS-III développée au laboratoire ISL (Interactive Systems Laboratory) des universités Karlsruhe et Carnegie Mellon. Le corpus BREF-80 (textes lus extraits du Journal *Le Monde*) a été utilisé pour le développement, l'apprentissage et l'évaluation du module. Les résultats obtenus sont de l'ordre de 91 % de bonne reconnaissance de mots. L'article décrit l'architecture du module de reconnaissance et son intégration à un module de traduction automatique.

Introduction

La traduction des documents écrits a fait de réels progrès pendant ces dernières années. Nous pouvons constater l'émergence de nouveaux systèmes de traduction de textes qui proposent une traduction soignée en différentes langues [1]. Il semble envisageable de les adapter pour la traduction de l'oral, à condition d'en améliorer le temps de réponse et la robustesse : c'est le « challenge » posé à ces systèmes mais aussi au module de reconnaissance de la parole. Un système de traduction de l'oral repose sur l'intégration des modules de reconnaissance et de synthèse de la parole et des modules de traduction, pour obtenir une boucle complète d'analyse et de synthèse entre les deux interlocuteurs [**Fig. 1**]. Le projet CSTAR-II [3] est un projet international dans lequel toutes les équipes travaillent sur tous les aspects de ce modèle.

Pour permettre à deux personnes de communiquer, il faut deux séries de processus symétriques dans les deux langues : un module de reconnaissance pour acquérir et transcrire les énoncés dits par un locuteur dans sa langue puis un module de traduction qui traduit la transcription dans la langue du destinateur ou dans un format d'échange standard (IF = Interchange Format) et enfin un module de synthèse de la parole (et de génération si on utilise le format IF) dans la langue cible du

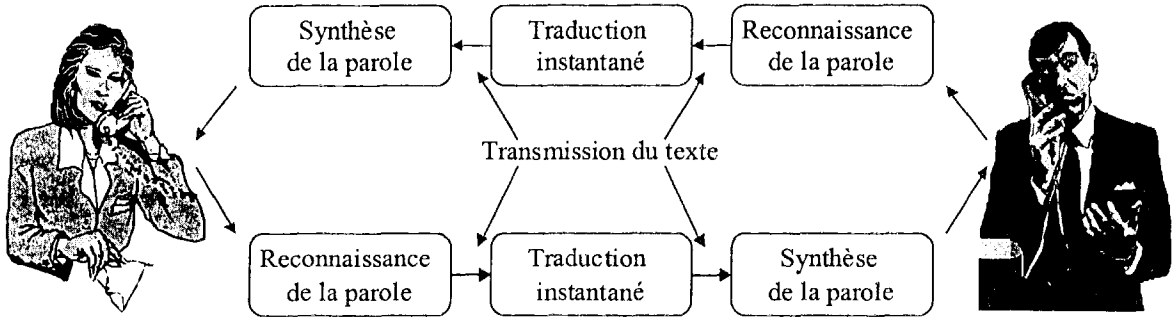

Fig. 1. L'architecture d'un système de traduction instantanée.

destinateur. Dans le cadre du projet C-STAR II nous avons en charge la conception et la réalisation du module de reconnaissance de la parole continue à grand vocabulaire pour le français. Nous collaborons avec l'équipe GETA du laboratoire CLIPS-IMAG et le laboratoire LATL pour la traduction automatique et le laboratoire LAIP pour la synthèse de la parole. Ce consortium s'est fixé l'objectif de réaliser un système de traduction de l'oral pour le français.
Dans cet article nous allons tout d'abord présenter l'architecture du système de traduction et la plate-forme de développement JANUS-III [2], puis les différentes étapes du développement du module RAPHAEL et enfin, les premiers résultats obtenus.

1 RAPHAEL pour la Traduction

L'architecture du système de traduction de parole est composée de trois modules essentiels (la reconnaissance, la traduction et la synthèse de la parole) [Fig. 2]. Dans ce projet nous utilisons ARIANE et GB [3] pour la traduction et LAIP-TTS [4] pour la synthèse. Le

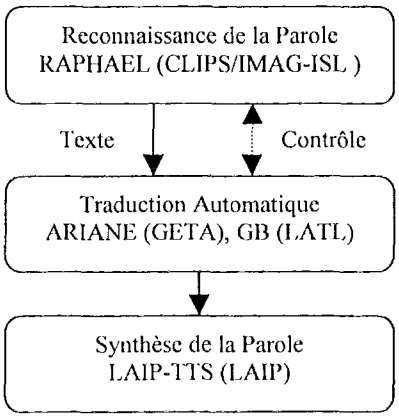

Fig. 2. Les composants du système

développement du module de reconnaissance RAPHAEL a été effectué sur la plate-forme logicielle de JANUS-III. RAPHAEL donne en sortie un treillis de mots sous le protocole TCP/IP. Le traducteur utilise ce résultat pour en donner une version traduite. Cette version est ensuite envoyée au synthétiseur de la parole. Dans cet article nous nous intéresserons seulement au module de reconnaissance RAPHAEL. Pour l'instant la stratégie d'échange entre les modules est entièrement séquentielle. Afin d'améliorer le résultat final (surtout du point de vue de la robustesse) nous envisageons l'intégration d'une seconde couche de contrôle pour permettre le « rescoring » des hypothèses en tenant compte des taux de confiance associés aux différents mots de l'énoncé reconnu.

1.1 Plate-forme de JANUS-III

Cette plate-forme de traduction a été développée dans le laboratoire d'ISL des universités Carnegie Mellon et Karlsruhe et contient tous les composants nécessaires au développement d'un système de reconnaissance phonémique à grand vocabulaire à base de Chaînes de Markov Cachées (**CMC**) et de réseaux de neurones. La facilité d'écrire un module de reconnaissance en langage Tcl/Tk avec JANUS-III nous permet d'adapter ses capacités selon les besoins d'application et les caractéristiques du français. De cette plate-forme, seul le moteur de reconnaissance est directement exploité. Mais le travail de préparation des bases de données, l'apprentissage des modèles de phonèmes, l'évaluation sont également effectués dans cet environnement de programmation. Le langage PERL est en grand partie utilisé parallèlement pour traitement du texte du corpus.
Les détails techniques de JANUS-III sont donnés dans [2], [5], [6]. Cependant nous en présentons brièvement quelques points ci-après.

2 Le Module RAPHAEL

L'architecture du module de reconnaissance RAPHAEL est présentée sur la [Fig. 3]. L'analyse de la parole produit une suite de vecteurs de paramètres acoustiques. Ces vecteurs sont utilisés par un moteur de recherche à base de CMC pour estimer la suite des phonèmes énoncés. Un modèle de langage stochastique à bigramme et trigramme, et un dictionnaire des variantes phonétiques sont en parallèle exploités pour restreindre le champ de recherche[1]. Au cours de la recherche le dictionnaire phonétique fournit le(s) phonème(s) suivant(s). Le modèle probabiliste de langage à base de bigramme et de trigramme est utilisé lors de la transition entre deux mots pour fournir un ensemble de mots [Fig. 4].

[1] Avec 45 phonèmes en moyenne une suite de cinq phonèmes se transforme théoriquement en un arbre de décision de 45^5 = 184,528,125 feuilles !

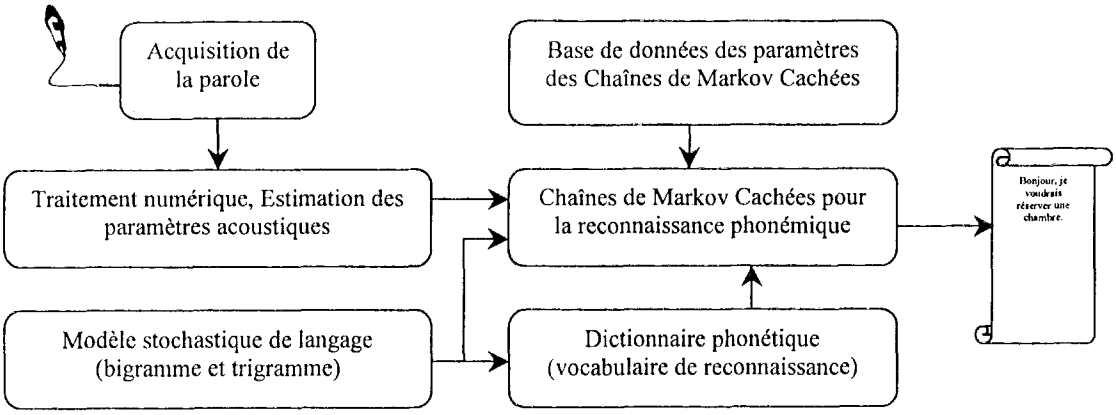

Fig. 3. Schéma du module de reconnaissance phonémique RAPHAEL.

2.1 Chaîne de Markov Cachées

Pour utiliser les CMC il faut conduire une phase d'apprentissage préalable dans laquelle on adapte les probabilités des transitions et des symboles sortis pour un phonème donné de manière à ce que la probabilité du processus associé soit maximale. Les paramètres des modèles et la transcription phonétique des énoncés du corpus sont deux éléments essentiels d'apprentissage.

RAPHAEL comporte 45 CMC représentant 42 phonèmes de base du français et 3 modèles pour le silence et le bruit. A quelques exceptions près les CMC se composent de trois états. Le vecteur de paramètres d'entrée est de dimension 12[2]. Les CMC ont 16 distributions Gaussiennes pour chaque état.

Lors de l'apprentissage nous produisons la transcription phonétique correspondante à chaque énoncé (cela se fait à l'aide du dictionnaire phonétique). Pour chaque énoncé les CMC correspondant aux phonèmes sont concaténées pour créer une longue chaîne. Ensuite l'algorithme de Viterbi [5] propose un alignement de l'énoncé avec cette chaîne. Avec cet alignement l'algorithme de Baum-Welch [5] procède à l'estimation des paramètres de chaque CMC présente dans la chaîne. Ce procédé est répété pour tous les énoncés du corpus d'apprentissage et cela plusieurs fois. La présence des différents contextes phonémiques permet à ce procédé de minimiser le taux d'erreur de reconnaissance. L'évaluation du taux d'erreur à la fin de chaque itération permet d'étudier l'avancement de l'apprentissage.

2.2 Modèle de langage stochastique

Afin de réduire le champ de recherche, un modèle de langage doit être utilisé. Bien que dans les systèmes à commande vocale qui utilisent une syntaxe réduite les grammaires finies ou récurrentes peuvent être utilisées, celles-ci ne sont pas capables de décrire tous les phénomènes de la langue parlée (ellipses, hésitations, répétitions, etc.). Pour cette raison il est souhaitable d'utiliser un modèle stochastique qui estime dans un contexte donné, la probabilité de succession des mots. Dans le modèle actuel les contextes gauches d'ordres un et deux (bigramme et trigramme) sont en même temps exploités. Le bigramme est utilisé dans la première phase de recherche pour créer un treillis de mots, puis le trigramme est utilisé pour raffiner le résultat et déterminer les N meilleurs phrases plausibles. Le modèle de langage se charge en même temps de la résolution de l'**accord** en français.

Le calcul des paramètres de ce modèle a été effectué à partir des corpus enregistrés et transcrits. Dans l'état actuel un vocabulaire de 7000 mots a été sélectionné.

[2] Les coefficients MFCC [5] d'ordre 16 sont calculés sur une trame de 16 ms de parole, avec un pas d'avancement de 10ms. La parole est échantillonnée à 16 kHz et sur 16 bits. Les MFCC, l'énergie du signal, et leurs première et seconde dérivées (51 valeurs) subissent ensuite une analyse en composantes principales (ACP) pour réduire la dimension du vecteur à 12. La matrice d'ACP est calculée avant la phase d'apprentissage, sur un grand corpus enregistré.

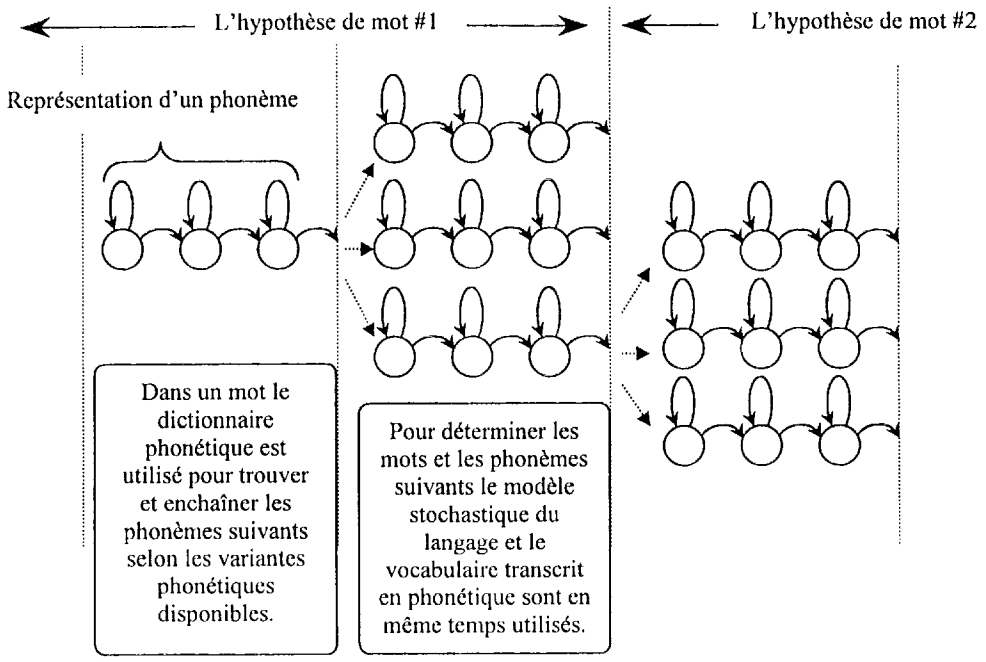

Fig. 4. Représentation de l'algorithme de recherche

2.3 Dictionnaire Phonétique

La conversion d'une chaîne d'hypothèses phonétiques en une chaîne orthographique se fait à partir d'un dictionnaire phonétique. Pour couvrir un grand nombre de prononciations différentes dues aux différents dialectes de la langue et aux habitudes des locuteurs, ce dictionnaire contient pour chaque mot un ensemble de variantes phonétiques. A chaque hypothèse de mot proposé par le modèle de langage on associe cet ensemble de variantes. Indépendamment donc de la variante utilisée dans l'énoncé, nous obtenons la même transcription orthographique. Nous utilisons spécifiquement cette technique pour couvrir les variantes produites par la **liaison**, par exemple :

*Je **suis** parti de la maison.* *(Z& **sHi** paRti ...)*
*Je **suis** allé à la maison.* *(Z& **sHiz** ale ...)*

3 L'apprentissage

Le corpus BREF-80 [8] comportant 5330 énoncés par 80 locuteurs (44 femmes et 36 hommes)[3] a été utilisé pour les phases d'apprentissage et d'évaluation. Un sous-ensemble de BREF-80 comprenant les énoncés de 4 femmes et 4 hommes a été utilisé pour l'évaluation[4]. Le vocabulaire a été transcrit soit manuellement, soit à partir du dictionnaire phonétique BDLEX-23000. Le modèle de langage a été estimé à partir de BREF-80 et un corpus de texte d'à peu près 10 millions de mots extrait du journal *Le Monde*.

Pour l'initialisation des CMC, au lieu d'utiliser les valeurs aléatoires (technique habituelle), nous avons choisi d'utiliser les modèles issus du projet GlobalPhone [7]. Pour chaque phonème de notre module nous avons manuellement choisi un phonème dans une des langues supportées par GlobalPhone (principalement allemande) et nous avons utilisé ses paramètres comme valeurs initiales de nos CMC. Ensuite ces modèles ont été adaptés au français au moyen de l'algorithme d'apprentissage décrit en 2.1. A la fin de chaque itération et ce pour 3

[3] BREF-80 contient 3747 textes différents et environ 150,000 mots.

[4] Les sous-corpus de l'apprentissage et de l'évaluation n'ont aucun énoncé et locuteur en commun. En réalité, nous avons enlevé tous les énoncés en communs entre ces deux sous corpus. Ainsi le sous-corpus d'apprentissage comprend 4854 énoncés et le sous-corpus d'évaluation 371 énoncés. Nous avons retiré 105 énoncés pour assurer la disjonction des deux sous-corpus.

itérations, le système a été évalué avec le sous corpus de l'évaluation.

4 Résultats

Les résultats d'évaluation en terme de taux de reconnaissance sont donnés dans le [**Tableau 1**].

Systèmes	*% mots reconnus*
Modèles issus de GlobalPhone	***29***
Première itération	***88,8***
Troisième itération	***91,1***

Tableau 1. Les résultats de l'évaluation

4.1 Commentaires

Une très bonne initialisation de certaines consonnes identiques dans des différentes langues (p, t, k, b, d, g, etc.) a rapidement permis d'obtenir un système fonctionnel.
On constate une saturation très rapide du taux de reconnaissance dès la troisième itération. Nous pouvons distinguer trois types de problème qui nous empêchent d'atteindre un meilleur taux de reconnaissance :

- Fautes de frappe dans le texte du corpus,
- Transcription erronée ou insuffisamment détaillée des énoncés,
- La couverture partielle de toutes les variantes phonétiques d'un mot.

Ces trois problèmes sont les causes d'un grand nombre d'erreurs d'alignement qui vont directement influencer le résultat final. Nous devons donc effectuer une vérification complète du corpus et du dictionnaire phonétique.
Les mots hors du vocabulaire sont à l'origine d'un pourcentage important d'erreurs. En effet, dans 371 énoncés du sous-corpus de l'évaluation nous rencontrons environ 300 mots hors vocabulaire. Ces mots représentent environ 3,5 % de la taille du vocabulaire. Il ne sont pas représentés dans le corpus d'apprentissage et leur transcription n'existe pas dans le dictionnaire phonétique.

Conclusion et perspectives

Dans cet article nous avons brièvement décrit, en termes d'avancement de projet, notre système de reconnaissance RAPHAEL à grand vocabulaire et rapporté des premiers résultats obtenus. Notre but est d'améliorer le taux de reconnaissance par l'utilisation des modèles phonétiques contextuels et d'élargir le vocabulaire utilisé à plus de 10000 mots. Pour atteindre ce but nous allons spécialiser le vocabulaire dans le domaine du tourisme et utiliser d'autres corpus de la parole spontanée dans ce domaine avec un nombre plus important de locuteurs. En même temps nous définirons un protocole d'échange plus élaboré avec le module de traduction afin de permettre la communication d'informations linguistiques et statistiques au module de traduction, toujour dans le but d'améliorer les performances de notre système.

Remerciement

Nous remercions Alex Waibel pour la mise à disposition de JANUS-III et Tanja Schultz pour son support scientifique et technique dans l'utilisation des résultats du projet GlobalPhone.

Références

1 Hutchins W. J. (1986) *Machine Translation : Past, Present, Future.* Ellis Horwood, John Wiley & Sons, Chichester, England, 382 p.

2 Finke M., Geutner P., Hild H., Kemp T., Ries K., Westphal M. (1997) : *The Karlsruhe- Verbmobil Speech Recognition Engine*, Proc. of ICASSP, Munich, Germany.

3 Boitet Ch., (1986) *GETA's MT methodology and a blueprint for its adaptation to speech translation within C-STARII*, ATR International Workshop on Speech Translation, Kyoto, Japan.

4 Keller, E. (1997). *Simplification of TTS architecture versus Operational quality*, Proceedings of EuroSpeech'97, Rhodes, Greece.

5 Rabiner L., Juang B.H. (1993), *Fundamentals of Speech Recognition*, Prentice Hall, 507 p.

6 Haton J.P., Pierrel J.M., Perennou G., Caelen J., Gauvain J.L. (1991), *Reconnaissance automatique de la parole*, BORDAS, Paris, 239 p.

7 Schultz T. Waibel A., *Fast Bootstrapping of LVCSR systems with multilingual phonem sets*, Proceedings of EuroSpeech'97, Rhodes, Greece.

8 Lamel L.F., Gauvain J.L., Eskenazi M. (1991), *BREF, a Large Vocabulary Spoken Corpus for French*, Proceedings of. EuroSpeech'91, Genoa, Italy.