

Reconnaissance de la parole dans le cadre de très grands vocabulaires

B. JACOB et R. ANDRE-OBRECHT

Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, France

Abstract :

This paper describes a new strategy for very large vocabulary speech recognition. The main problem is to reduce the lexical access without pruning the correct candidate. We propose to exploit the branching structure of BDLEX and the description of each word into root and flexional ending. More we use the notion of phonetic classes to decompose the dictionary into sub-dictionaries. We develop a two-stage recognition algorithm :

— Each dictionary which is considered as a sequence of phonetics classes is modeled by a HMM where the elementary units are these phonetics classes.

— Each word is modeled by a classical HMM where the elementary unit is the pseudo-diphone.

For a unknown word utterance, a first recognition gives the best dictionary to which it belongs, the Viterbi algorithm applied to the network of the best dictionary words, gives the word with the most likelihood.

Experiments are carried out with telephonic database.

1. INTRODUCTION.

L'accès lexical, dans les systèmes de RAP, dans le cadre de très grands vocabulaires, est un problème difficile. Lorsque la taille du vocabulaire augmente (> 1000 mots), il devient indispensable d'optimiser cet accès. Plusieurs tendances sont actuellement observées ; le lexique peut être organisé sous forme arborescente [1] ou sous forme de tables de règles de grammaire [2] afin de réduire l'espace de recherche. Nous nous intéressons à une nouvelle approche basée sur la recherche probabiliste de sous-dictionnaires, eux-mêmes structurés en arbres : le vocabulaire de base est considéré comme une réunion de sous-dictionnaires, chacun étant modélisé par un sous-modèle de Markov ; chaque sous-dictionnaire est lui-même un ensemble de mots et chaque mot correspond à un modèle de Markov caché. Les modèles de Markov cachés sont construits de manière hiérarchique à partir d'unités de base : pour la modélisation des dictionnaires, est introduite la notion de classes majeures et pour celle des mots, la notion de pseudo-diphones. Intuitivement cette décomposition correspond à la représentation des mots en classes phonétiques puis en phonèmes. Nous l'appliquons, dans un premier temps, à la reconnaissance de mots isolés.

La mise en oeuvre de cette stratégie n'a pu être faite sans un lexique clairement étudié et structuré, et sans un outil informatique puissant permettant la création de MMC, et leur compilation éventuelle à l'aide de règles phonologiques. C'est pourquoi nous consacrons deux paragraphes à la

description de ces outils, avant de détailler le système de reconnaissance proprement dit et les premières expérimentations.

2. BDLEX : un lexique structuré.

Le lexique de base utilisé est celui de BDLEX version 1. Il contient 23.000 entrées lexicales, chacune d'elles étant accompagnée d'une description linguistique centrée sur les niveaux morpho-syntaxique et phonologique du français [3]. Chaque entité dispose d'une description phonétique classique à laquelle est adjointe une description en classes majeures. Une classe majeure regroupe les phonèmes qui ont un comportement acoustique proche (tableau 1).

voyelles	A	a	plosives	T	p, t, k
	E	ε, e		R\$	r (fin de mot)
	I	i, u, Y		TR\$	tr (fin de mot)
	e	œ, ø		TR	tr
	O	o		TL	pl
voyelles nasales	%	ã, õ	fricatives	S	f, s, ʃ
	IN	ẽ		Z	v, Z, ʒ
nasales	N	m, n, η	liquides	R	r
semi-voyelles	J	w, j		L	l
plosives voisées	D	b, d, g	silences	#	
				\$	

A partir du lexique et en utilisant un ensemble de règles flexionnelles, on obtient un corpus de 270.000 formes fléchies. Ces règles sont basées sur l'association racine/désinence où la racine est la partie caractéristique du mot et la désinence celle qui porte les marques flexionnelles (pluriel, féminin, temps, personne, mode) [4].

Pour minimiser le problème d'accès posé par la reconnaissance grands vocabulaires, ce corpus est structuré de manière arborescente. L'arborescence principale regroupe toutes les racines présentes dans ce corpus; elle est liée aux sous arbres des désinences. Pour optimiser le traitement et réduire la taille des graphes, nous prenons en compte la représentation phonologique en classes majeures des racines et des désinences. Les formes qui ont la même représentation phonologique en classes majeures sont factorisées, les branches de l'arbre correspondant sont regroupées en une seule; finalement une branche de l'arbre des racines reliée à une branche d'un des arbres de désinence, dans le cas où cette liaison est possible, permet de définir un sous-dictionnaire.

3. Le générateur de réseaux.

La plupart des applications en reconnaissance de la parole utilisent des modèles de Markov cachés (MMC), construits de manière hiérarchique à partir de réseaux élémentaires. De nombreux outils sont actuellement disponibles tels que le compilateur PHIL 90 [5], le logiciel HTK [6]. Malheureusement ils sont inadaptés à la création et à la gestion de graphes et arbres, tels que ceux présentés au paragraphe précédent.

L'outil que nous développons, est un compilateur de réseaux; à partir de réseaux élémentaires de type MMC et d'instructions de base, un réseau hiérarchique structuré en niveaux est obtenu:

- le niveau le plus bas représente le réseau actif; les transitions et états de ce niveau sont dits actifs, en particulier les transitions porteuses des lois d'observations.
- les niveaux supérieurs représentent les étapes successives de la construction du réseau actif.

Dans l'ensemble des instructions de compilation, on peut distinguer trois grandes familles de rubriques:

- la première famille consiste à caractériser un réseau simple d'un seul niveau en décrivant ses états initiaux et finaux, et ses transitions. Le nom d'un état est formé d'une racine et d'une extension et on peut définir des classes d'états en les repérant par la même racine. Les transitions peuvent supporter des lois discrètes ou continues. Elles peuvent être éventuellement vides, c'est à dire ne

supportant aucune loi. A chaque transition correspond un chaînage arrière. On peut ainsi connaître à tout instant les prédécesseurs d'un état dans le graphe et, bien entendu, ses suivants.

- la deuxième famille d'imbrications est utilisée pour "imbriquer" des réseaux. Ceci consiste à relier des transitions appartenant à un réseau global à des sous-réseaux. Cette nouvelle liaison peut être obligatoire ("Clusters"), c'est à dire détruire la transition initiale, ou facultative ("Coarticulation"), et dans ce cas la transition initiale est doublée. On peut en outre relier une classe d'états par un sous-réseau ("Lien"). Les sous-réseaux peuvent être locaux, c'est à dire décrits en même temps que le réseau global dans le fichier texte, ou bien être déjà créés en mémoire.

Il s'établit entre l'état instancié, dans le cas de "Liens", — ou l'état de départ de la transition instanciée, dans le cas des "Clusters" et "Coarticulations" — et le sous-réseau qui le remplace, un lien de parenté du type père/fils : tous les états appartenant au sous-réseau verront la racine de leur arbre généalogique augmenté d'un ancêtre ; réciproquement l'état substitué reconnaîtra comme ses fils les plus lointains ancêtres de l'ensemble des états du sous-réseau.

L'état qui a été instancié, ainsi que les flots entrant et sortant de ses transitions, sont rendus inactifs. Les états inactifs sont les ancêtres des états du réseau actifs et sont liés entre eux par des relations inactives. Chaque génération d'états forme ainsi un niveau du réseau global. Il est possible de connaître à tout instant dans le parcours d'un tel graphe les ancêtres et les descendants d'un état.

- les lois de probabilité du réseau global sont créées avec des valeurs données par défaut. La troisième famille de rubriques permet de modifier ces valeurs en les initialisant par les données de l'utilisateur. Il est possible d'initialiser soit la totalité des lois du réseau, soit seulement celles relatives à des transitions dont les états appartiennent à une classe d'ancêtres.

Une fois les instructions compilées et le réseau final obtenu, celui-ci est optimisé en supprimant en particulier les transitions vides. Les algorithmes d'apprentissage et de reconnaissance à l'aide d'un tel réseau, basés sur l'algorithme de Viterbi, sont adaptés à cette structure de données.

4. Le système de reconnaissance.

Le système de reconnaissance automatique de mots isolés que nous abordons se décompose en trois modules principaux qui sont :

- le pré-traitement acoustique,
- la recherche du meilleur sous-dictionnaire,
- la recherche du mot le plus probable.

4.1 Le pré-traitement acoustique :

Ce pré-traitement s'effectue en deux phases :

-un algorithme de segmentation statistique, l'algorithme de divergence "forward-backward"[7], est appliqué au signal de parole et permet d'obtenir une suite de segments de taille variable correspondant dans une première approximation aux zones stables et transitoires du continuum de parole.

-sur chaque segment est ensuite effectuée une analyse spectrale. Sont ainsi extraits 8 coefficients cepstraux, l'énergie, les dérivées premières et secondes de ces paramètres ; leur est adjoint la durée du segment en centisecondes pour former un vecteur d'observation.

4.2 La recherche de sous-dictionnaires :

A partir de l'écriture en *classes majeures* du lexique BDLEX et de sa structure arborescente, le compilateur décrit ci-dessus permet d'obtenir un réseau arborescent pour représenter l'ensemble des racines et des réseaux pour ceux des désinences, puis d'effectuer les transitions entre eux. Les règles phonologiques sont interprétées à l'aide des instructions Cluster et Coarticulations. L'ensemble constitue la modélisation probabiliste de BDLEX, dans laquelle l'unité de base est la classe majeure ; les lois d'observation sont gaussiennes, les matrices de covariance sont diagonales. L'algorithme de Viterbi permet d'aligner la suite d'observations acoustiques du mot à reconnaître sur un des chemins actifs du réseau et de fournir la meilleure suite de classes majeures qui correspond, à fortiori, à un sous-dictionnaire.

4.3 La recherche du mot.

La suite de classes majeures reconnue est équivalente à un sous-dictionnaire de mots. Le compilateur de réseaux fournit pour chaque mot un modèle de Markov caché, dans lequel l'unité de base est le pseudo-diphone (parties stables des sons, transitions entre les sons...); les règles phonologiques permettent d'accéder aux différentes prononciations. Les lois d'observations sont gaussiennes de matrice de covariance diagonale. L'algorithme de Viterbi, appliqué à la même suite d'observations que précédemment, fournit le mot reconnu.

5. Expériences et conclusion.

Dans la mesure où il s'agit d'une étude préliminaire, il n'était pas envisageable, au cours des premières expériences, de modéliser et surtout d'apprendre tous les réseaux liés à BDLEX1. De plus, nous voulions nous assurer que le processus de reconnaissance en deux étapes ne dégradait pas trop les performances du système de reconnaissance par rapport à une approche plus classique.

Les premiers résultats sont obtenus sur un vocabulaire de 50 mots tirés de manière aléatoire parmi les 500 mots les plus courants du français; il en résulte qu'il s'agit essentiellement de *mots monosyllabiques et bisyllabiques*. Les enregistrements ont été faits par le CNET à travers le réseau téléphonique, la bande passante est donc réduite à 3,3kHz. Cette expérience est multi-locuteurs dans le sens où l'apprentissage est effectué à partir de 2 prononciations de chaque mot par dix locuteurs et que les tests sont faits sur une troisième répétition de chaque mot prononcé par chacun des dix locuteurs. Le taux de reconnaissance en termes de sous-dictionnaires correctement sélectionnés est de 90%, il est sensiblement égal au taux de reconnaissance en termes de mots correctement reconnus (dans le cadre de cette expérience, un sous-dictionnaire ne contient que rarement plus d'un mot).

Nous avons comparé cette approche à l'approche classique où chaque mot est directement modélisé par un MMC; l'unité de base est le pseudo-diphone, le traitement acoustique est identique. Le taux de reconnaissance est de 94%. La dégradation que nous observons est essentiellement due au manque de précautions prises lors de la modélisation des classes majeures. Nous allons poursuivre cette recherche en améliorant les modèles de Markov cachés élémentaires et en l'expérimentant sur de plus gros corpus. Alors seulement nous pourrions en tirer une réelle conclusion.

REFERENCES

- [1] V. Steinbiss, H. Ney, R. Haeb-Umbach, B.-H. Tran, U. Essen, R. Kneser, M. Oerder, H.-G. Meier, X. Aubert, C. Dugast, D. Geller, "The Philips Research System for large vocabulary continuous speech recognition", Eurospeech, Berlin 1993, pp 2125-2128.
- [2] Y. Minami, K. Shikano, T. Yamada, T. Matsuoka, "Very large vocabulary continuous speech recognition algorithm for telephone directory assistance", Eurospeech, Berlin 1993, pp 2129-2132.
- [3] G. Pérennou, "Le projet BDLEX de bases de données et de connaissances lexicales et phonologiques", 1ères Journées du GRECO-PRC Communication Homme-Machine, EC2 Editeur, Paris 1988, pp 81-111.
- [4] I. Ferrané, "Base de données et de connaissances lexicales morphosyntaxiques", Thèse de l'Université Paul Sabatier, Septembre 1991.
- [5] D. Jouvet, J. Monné, D. Dubois, A new network based speaker independent connected word recognition system. ICASSP, Tokyo 1986.
- [6] S.J. Young "HTK Hidden Markov Model Toolkit" Rapport Cambridge University Engineering Department — Speech Group — September 92.
- [7] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Trans. on ASSP, vol. 36, n°1, January 1988.