RECONNAISSANCE SEGMENTALE MULTILOCUTEUR DE MOTS ISOLES *

SEGMENT BASED SPEAKER INDEPENDENT ISOLATED WORD RECOGNITION

L. SAUTER **

Résumé : La création des ensembles de référence pour les systèmes de reconnaissance multilocuteur de mots isolés tels que SYRIL (Flocon 84) représente un travail considérable. Dans les cas où le vocabulaire est spécifique à une application une approche segmentale permettrait de créer ou de modifier l'ensemble de référence d'une manière automatique. Nous examinons ici une telle approche, en donnant quelques résultats expérimentaux concernant la construction de références globales à partir de segments.

Abstract : Building reference sets for speaker independent isolated word recognition systems such as SYRIL (Flocon 84) implies a considerable effort. Whenever the vocabulary is specific to some domain a segment based approach should allow to create or change the reference set automatically. We examine such an approach and give experimental results concerning the use of segments for creating whole-word references.

Mots clés : Reconnaissance de la parole, mots-isolés, multilocuteur, segment, demi-syllabes.

Key words : Speech recognition, isolated-words, speaker-independent, segment, demi-syllables.

1. Introduction :

Le projet japonais "ordinateurs de la 5ème génération" comprend une unité de reconnaissance de mots isolés monolocuteur, c'est-à-dire avec apprentisage. Dans l'état actuel de la recherche, il semble possible d'offrir à l'utilisateur occasionnel d'une telle machine la possibilité de donner des commandes vocales sans avoir à effectuer un apprentissage préalable. Pour cela, il faut que le système de reconnaissance soit indépendant du locuteur. Cette indépendance peut être obtenue soit par l'adaptation au locuteur, soit par l'utilisation de référence multiples pour chaque mot du vocabulaire. Cette dernière solution a déjà été utilisée dans plusieurs systèmes de reconnaissance, dont SYRIL (Flocon 84).

Les références multiples sont obtenues à partir de répétitions de chaque mot par un nombre suffisant de personnes (de l'ordre de quelques dizaines, voire d'une centaine). C'est pour cette raison que la création d'un ensemble de référence pour un nouveau vocabulaire est d'un coût très élevé.

Si le vocabulaire à reconnaître est d'un intérêt général (c'est par exemple le cas des chiffres), cela n'a pas beaucoup d'importance. Par contre, si le vocabulaire est spécifique à une application ou encore si le vocabulaire doit pouvoir évoluer (noms d'utilisateurs, d'abonnés...) ce coût devient un handicap.

---

* Cette étude a été partiellement financée par un contrat ESPRIT.

** Laboratoires de Marcoussis, Centre de recherches de la CGE, route de Nozay - 91460 Marcoussis.

Il semble donc être intéressant de développer des techniques permettant de créer automatiquement un apprentissage multilocuteur pour un nouveau vocabulaire : à partir d'une base de données établie une fois pour toutes, créer l'ensemble de référence pour un nouveau vocabulaire en le spécifiant simplement par la transcription phonétique des mots (qui peut être obtenue elle-même automatiquement à partir du texte).

C'est une approche de ce problème que nous exposerons ici : la création d'ensembles de référence multilocuteurs à partir d'une base de données de segments.

2. Reconnaissance de mots isolés par une approche segmentale.

L'approche segmentale consiste tout d'abord à constater que l'ensemble des mots d'une langue quelconque est formé avec un nombre assez faible de sons différents. La solution idéale consisterait donc à savoir reconnaître ces sons et d'en déduire le mot qui a été prononcé. Malheureusement, les sons élémentaires du langage sont très variables. Ils dépendent de nombreux facteurs tels que le contexte (le son "k" est très différent selon qu'il précède par exemple un "i" ou un "a"). Les lois qui régissent cette variabilité sont connues empiriquement par les phonéticiens mais sont difficiles sinon impossibles à inclure dans un traitement automatique. (L'utilisation des systèmes experts le permettra vraisemblablement dans l'avenir).

Il existe néanmoins une solution à ce problème. Il suffit de donner au système un exemple de chaque son élémentaire dans tous les contextes possibles. C'est ce que l'on fait quand on utilise des diphones (concaténation de deux sons élémentaires). Pour tenir encore mieux compte du contexte et pour faciliter la segmentation de la parole, des segments plus longs sont préférables, tels que des syllabes ou des demi-syllabes. Les syllabes présentent néanmoins le problème de leur nombre (une dizaine de milliers). Les demi-syllabes constituent un bon compromis entre le nombre (un millier environ) et la qualité de la représentation du contexte. (Une demi-syllabe est le segment de parole comprenant la moitié d'une voyelle précédée (ou suivie) d'un groupe de consonnes). Toute syllabe peut être obtenue en concaténant deux demi-syllabes.

Une fois qu'on dispose d'une base de données multilocuteur de segments il faut encore déterminer comment l'utiliser pour la reconnaissance de mots isolés. A priori, il existe au moins trois méthodes :

(1) la construction d'empreintes globales à partir des segments. Cette méthode a déjà été utilisée dans le cas monolocuteur (Rosenberg 83).

(2) l'utilisation d'algorithmes permettant de chercher la suite de segments correspondant au mieux au mot prononcé.

(3) le découpage du mot prononcé en segments, suivi de l'identification de chaque segment, puis enfin de la détermination du mot prononcé.

Dans la suite, nous n'examinerons en détail que la première méthode.

3. Test préliminaire

Pour pouvoir créer des références multilocuteurs pour un mot quelconque du français il faudra disposer d'un corpus exhaustif de demi-syllabes extraites de parole enregistrée et prononcée par un nombre assez important de personnes. Cela représente un travail considérable. Néanmoins, la validation des algorithmes et des techniques utilisées pour la reconnaissance peut se faire à l'aide d'un corpus de test. Nous décrivons maintenant celui-ci.

3.1 Corpus de test

Nous utilisons l'apprentissage du système SYRIL qui comprend l'enregistrement de 35 mots par 20 locuteurs. En découpant ces 35 mots en demi-syllabes, on en obtient 87 différentes. Nous disposons donc d'un corpus potentiel de 87 demi-syllabes prononcées par 20 locuteurs. Il est bien sûr nécessaire de déterminer les frontières exactes de chaque segment à l'intérieur de son mot d'origine. Pour cela un programme interactif a été réalisé. Il nous a permis de construire le corpus de test qui comporte donc 1740 segments.

3.2. Vocabulaire test

En juxtaposant de différentes manières les 87 demi-syllabes disponibles, on obtient un ensemble de mots (qui contient évidemment les 35 mots dont sont issues les demi-syllabes). Par exemple, si nous disposons entre autres des segments [ti], [it] et [R*] alors cet ensemble contiendra le mot "titre". (Les transcriptions phonétiques sont données avec l'alphabet phonétique proposé dans (Lennig 84).)

Nous avons choisi dans cet ensemble 18 nouveaux mots pour faire les essais de reconnaissance.

Ces mots ont été choisis de manière à ce que les segments ne soient pas dans le même contexte que dans leur mot d'origine. Ainsi le [vi] utilisé pour "vitre" provient du mot "diviser", où il était dans un contexte très différent. Ceci permettra d'étudier ultérieurement des règles de lissage et de durée.

Plusieurs mots ne diffèrent que par un son (paires minimales) : "vitre", "mitre" et "titre" par exemple. Le vocabulaire de test est donc très difficile à reconnaître. Ceci a l'avantage de faciliter la comparaison de différentes méthodes. En effet, un nombre d'erreurs faible rend difficile (pour des raisons statistiques) l'évaluation des améliorations apportées au système.

4. Reconnaissance par reconstruction de références globales

Dans la plupart des systèmes de reconnaissance de la parole, le signal acoustique est représenté par une suite de vecteurs. Ceux-ci sont calculés toutes les 10 à 20 millisecondes, et caractérisent le spectre du signal. L'empreinte globale d'un mot est la suite de vecteurs obtenus par l'analyse du signal acoustique correspondant à la prononciation de ce mot. Afin de créer des références globales "synthétiques" il est possible de concaténer les vecteurs correspondant à l'analyse des différents segments qui composent le mot. Ainsi, si nous désirons créer une empreinte globale pour le mot "titre", nous mettrons bout-à-bout les vecteurs obtenus par l'analyse du signal correspondant aux segments [ti], [it] et [R*]. Le signal correspondant au segment [ti] peut provenir par exemple du découpage d'un enregistrement du mot "multiplier".

Comme nous nous intéressons au cas multilocuteur, nous devrons créer pour chaque mot plusieurs empreintes globales en utilisant des segments provenant de différents locuteurs. Nous examinons maintenant plus en détail les différentes alternatives pour construire l'ensemble de référence.

4.1.Méthodes de construction de l'ensemble de référence.

Nous avons envisagé plusieurs méthodes pour construire un ensemble d'apprentissage.

(1) On utilise ensemble uniquement des segments provenant d'un même locuteur. On obtient ainsi 20 exemplaires de chaque mot (un par locuteur).

(2) Panachage : les segments provenant de différents locuteurs sont mélangés. Il est ainsi possible de reconstruire un nombre élevé d'exemplaires de chaque mot (si m est le nombre de locuteurs et si un mot comprend n segments, on peut obtenir m à la puissance n empreintes distinctes).

Nous avons testé cette méthode avec soit 20 soit 100 empreintes pour chaque mot du vocabulaire.

4.2. Choix d'un nombre restreint de représentants.

SYRIL utilise cinq références pour chaque mot du vocabulaire. Ces références sont calculées à partir de l'ensemble des empreintes disponibles. Nous avons essayé différents algorithmes de classification et différentes méthodes pour calculer les 5 représentants pour chaque mot.

4.2.1. Méthodes de classification

La classification automatique est utilisée pour regrouper en classes les empreintes semblables. Chaque classe pourra alors être représentée par une seule référence. A priori, la qualité de la classification peut être mesurée par la distance moyenne entre un objet quelconque et le centre de sa classe.

(1) Méthode des nuées dynamiques

Cette méthode consiste alternativement à attribuer chaque objet à la classe dont le centre est le plus proche, puis à calculer les nouveaux centres de chaque classe. Cette méthode converge vers un optimum local qui dépend du choix initial (et arbitraire) des centres des classes. En répétant de nombreuses fois ce procédé avec un choix initial aléatoire, on peut choisir le partitionnement donnant le meilleur optimum local.

Plusieurs variantes ont été utilisées pour le calcul du centre de chaque classe :

- l'objet minimisant la distance cumulée aux objets de la classe
- l'objet minimisant la distance maximum aux objets de la classe (minimax).
- une moyenne dynamique des mots de la classe. (mot calculé à l'aide de l'algorithme de programmation dynamique) (Flocon 84).

(2) Méthode de classification utilisée dans le système SYRIL.

Après avoir calculé le centre (au sens des minimax) de l'ensemble des objets, on affecte à une classe tous les objets dont la distance à ce centre est inférieur à un seuil donné. On recommence ce procédé avec les objets non encore affectés. Le seuil est fixé pour obtenir le nombre de classes désiré.

4.2.2. Calcul du représentant de chaque classe

Le représentant de chaque classe peut être soit un objet soit la moyenne dynamique des objets de la classe (calcul analogue à celui du centre de la classe décrit plus haut).

4.3. Résultats des tests.

Nous donnons ici quelques résultats de reconnaissance obtenus avec les variantes décrites plus haut. Le test de reconnaissance contient 180 mots soit le vocabulaire test de 18 mots prononcé par 10 locuteurs n'appartenant pas à l'apprentissage initial.

(1) Méthode de construction de l'ensemble d'apprentissage : le panachage dégrade la qualité de l'ensemble de référence. Toutes choses égales par ailleurs, le nombre d'erreurs passe de 30 (sans panachage) à 36 et 39 avec panachage (respectivement avec 20 et 100 exemples par mot).

(2) Le choix de l'algorithme de classification ne modifie pas les résultats avec 100 exemples par mot, par contre l'algorithme des nuées dynamiques semble mieux adapté au cas de 20 exemples. Nous rappelons que pour cet algorithme le choix des conditions initiales influe sur le résultat de la convergence. Nous avons fait varier ces conditions initiales et avons retenu le meilleur résultat. Dans tous les cas il vaut mieux utiliser comme centre l'objet qui minimise la distance cumulée aux autres objets plutôt que le minimax ou la moyenne dynamique.

(3) Les résultats sont meilleurs en choisissant comme représentant de chaque classe la moyenne dynamique des objets de la classe. Ce calcul est fait une fois alors que les classes ont été déterminées comme décrit au paragraphe précédent. Le nombre d'erreurs passe de 30 en choisissant l'objet qui minimise la distance cumulée aux autres objets à 24 en calculant la moyenne.

(4) La qualité de l'ensemble de référence est amélioré par un lissage aux frontières des segments (Kahn 84). Dans notre cas, les meilleurs résultats ont été obtenus en remplaçant quatre vecteurs de part et d'autre de la frontière par une droite (moindres carrés). Dans ce cas, le nombre d'erreurs passe de 24 à 16 (soit un taux de reconnaissance de 91.1 %) Nous pensons que les résultats pourraient être encore améliorés en utilisant des règles de durée.

Etant donné la complexité du vocabulaire, ces résultats sont tout à fait encourageants. Malheureusement, nous ne disposons pas d'un apprentissage multilocuteur pour le vocabulaire test : il ne nous est donc pas possible d'évaluer exactement la dégradation de la reconnaissance par rapport à l'approche globale. Néanmoins, les résultats peuvent être comparés à ceux obtenus par l'approche classique pour un vocabulaire différent. L'approche segmentale telle qu'elle a été décrite ci-dessus atteint un taux de reconnaissance de 91.1 %. SYRIL sur un vocabulaire à priori moins difficile reconnaît 96.3 % des mots.

5. Conclusion

Nous avons présenté une méthode permettant de créer un ensemble de référence pour un système de reconnaissance de mots isolés multilocuteur à partir d'un corpus de segments. La dégradation des performances par rapport à l'approche classique est assez faible : cette approche peut donc être intéressante dans les cas où le coût d'un apprentissage complet est trop élevé.

La méthode décrite ci-dessus pourrait vraisemblablement être étendue à d'autres types de systèmes de reconnaissance de mots isolés indépendants du locuteur, notamment aux systèmes utilisant l'approche dite des "Modèles de Markov Cachés".

REFERENCES

(Flocon 84) B. Flocon, N. Briant, "SYRIL : Système temps réel de reconnaissance de mots indépendant du locuteur", 4ème congrès AFCET RFIA, pp. 33-39, Janvier 1984.

(Kahn 84) C. Kahn, L.R. Rabiner and A. E. Rosenberg, "On duration and smoothing rules in a demisyllable-based isolated-word recognition system", J. Acoust. Soc. Am., vol. 75 (2), pp. 590-598, Février 1984.

(Lennig 84) M. Lennig et J.P. Brassard, "Machine-readable phonetic alphabet for English and French", Speech Communication, vol. 3, no. 2, pp. 165-166, Août 1984.

(Rosenberg 83) A.E. Rosenberg, L. A. Rabiner, J.G. Wilpon and D. Kahn, "A Demisyllable-based Isolated Word Recognition System", IEEE Trans. on Acoust. Speech and Sign. Proc. vol. ASSP-31, no. 3, pp. 713-726, juin 1983.